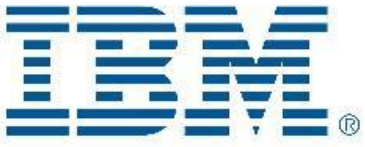


ARTIFICIAL INTELLIGENCE: STUDY MATERIAL

CLASS XI

LEVEL 2: AI INQUIRED (UNIT 6 – UNIT 10)

TEACHER INSTRUCTION MANUAL

INDEX

Unit 6

Critical and Creative Thinking

Title: Critical and Creative Thinking	Approach: Interactive/ Discussion, Team Activity

Summary: We are living in a rapidly changing complex world characterised by learning, unlearning and relearning which happens to be the new normal. 85% of the future jobs have either not been visualised or invented yet. So, how can we prepare our children for a future full of uncertainty and dramatic changes? Of the 10 skills expected to be in high demand in the future, World Economic Forum lists complex problem solving, critical thinking and creativity as the top three skills for future employment. Hence, it becomes imperative for the school or an educator to prepare and connect the students with the demands of the real world. And design thinking can be instrumental in establishing such connections which is going to be the topic of discussion in this unit.

Objectives:

1. To build focus on research, prototyping, and testing products and services so as to find new ways to improve the product, service or design.
2. Students develop the understanding that there is more to Design thinking than just hardware designing.
3. To inculcate design thinking approach to enhance the student's creative confidence.

Learning Outcomes:

1. Underlining the importance of Prototype as a solution to user challenge.
2. Recognizing empathy to be a critical factor in developing creative solutions for the end users.
3. Applying multiple brainstorming techniques to find innovative solutions.

Pre-requisites: Reasonable fluency in the English language.

Key Concepts: Design Thinking framework, Prototype, Ideate

(https://en.wikipedia.org/wiki/The_Thinker)

The illiterate of the 21st century will not be those who cannot read and write, but those who cannot learn, unlearn, and relearn." - ***Alvin Toffler, author of Future Shock.***

1. Design Thinking Framework

Activity 1

Can you please write a paragraph in not more than 10 lines on the image on the previous page?

Activity 2

Have the class form groups of 3 or 4 students each and assign them tasks (let's say to plan a party).

In round one, get everyone to start each sentence of their conversation with "Yes, BUT…….". After the first round, ask your participants how the conversation went? How did their discussion to plan the party go?

For round two, get the participants to start their conversation with "Yes, AND….". After the second round, ask the group how that round went and compare the two rounds of discussions. The differences between the two will be striking!

Purpose: Collaboration along with distinction between an open and closed mind set.

Activity 3

Divide the class into groups of 4-5 students each. Pick a random object (i.e. a paperclip, pen, notebook), and challenge each group to come up with 40 uses for the object. No repeats!

Each group will take turns in coming up with new ideas. Make sure that each group has a volunteer note-taker to capture the ideas along with the total number of ideas their group comes up with. Allow 4 mins for this challenge. When time is up, have each group share how many ideas they generated. The group with the most ideas is declared the winner!

Purpose: Thinking out-of-the box, encouraging wild idea generation

Activity 4

This is an activity that promotes pure imagination. The purpose is to think expansively around an ideal future for the school or about yourself too; it's an exercise in visioning.

The objective of this activity is to suspend all disbelief and envision a future that is so stellar that it can land you or your school on the cover of a well-known international magazine. The student must pretend as though this future has already taken place and has been reported by the mainstream media.

Purpose: Encouraging students to "think big,", Planting the seeds for a desirable future

According to Wikipedia, "Design thinking refers to the cognitive, strategic and practical processes by which design concepts (proposals for new products, buildings, machines, etc.) are developed." Design thinking is also associated with prescriptions for the innovation of products and services within business and social contexts.

Most often, the design is used to describe hardware, machine or a structure, but essentially, it is a process. It is a set of procedures and principles that employ creative and innovative techniques to solve any complex technological or social problem. It is a way of thinking and working about the potential solution to a complex problem.

Let's hear a story now…

Some years ago, an incident occurred where a truck driver tried to pass under a low bridge. But he failed, and the truck got lodged firmly under the bridge. The driver was unable to continue driving through or reverse out.

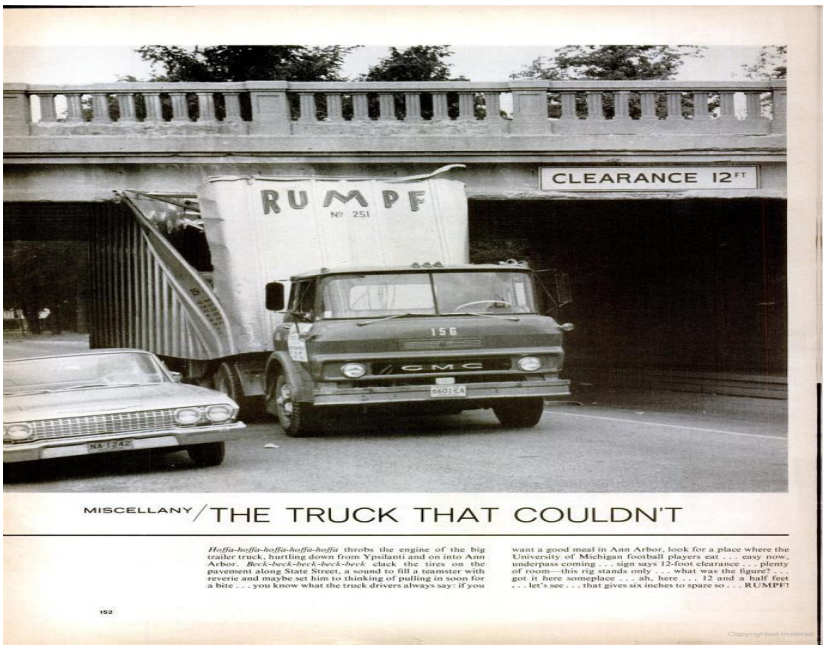
The story goes that as the truck became stuck, it caused massive traffic problems, which resulted in emergency personnel, engineers, firefighters and truck drivers gathering to devise and negotiate various solutions for dislodging the trapped vehicle.

Emergency workers were debating whether to dismantle parts of the truck or chip away at parts of the bridge. Each spoke of a solution that fits within his or her respective level of expertise.

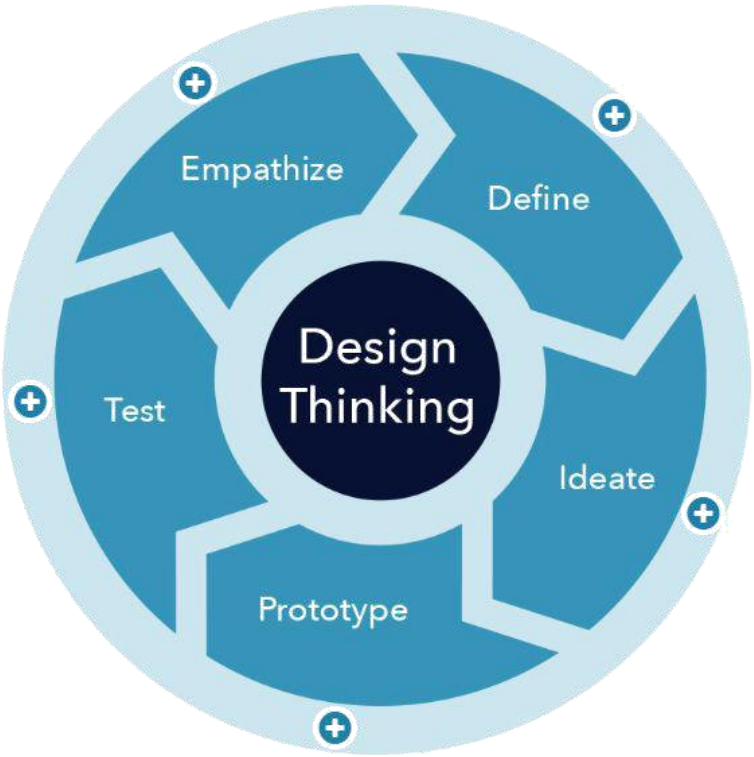
A boy walking by and witnessing the intense debate looked at the truck, at the bridge, then looked at the road and said nonchalantly, "Why not just let the air out of the tires?" to the absolute amazement of all the specialists and experts trying to unpick the problem.

When the solution was tested, the truck was able to drive free with ease, having suffered only the damage caused by its initial attempt to pass underneath the bridge. The story symbolizes the struggles we face where oftentimes the most obvious solutions are the ones hardest to come by because of the self-imposed constraints we work within.

(Source - https://www.interaction-design.org/literature/article/what-is-design-thinking-and-why-is-it-so-popular)



Now let's move on to understand the Design Thinking framework. The illustration below has the various components of the framework.



Empathize

Design thinking begins with empathy. This requires doing away with any preconceived notions and immersing oneself in the context of the problem for better understanding. In simple words, through empathy, one is able to put oneself in other people's shoes and connect with how they might be feeling about their problem, circumstance, or situation.

There is a challenge one needs to solve. How does one approach it? Empathy starts from here. As a designer of the solution to a challenge, one should always understand the problem from the end-user perspective.

This is done by observation, interaction or by imagination.

Define

In the Define stage, information collected during Empathize is used to draw insights and is instrumental in stating the problem that needs to be solved. It's an opportunity for the design thinker to define the challenge or to write the problem statement in a human-centred manner with a focus on the unmet needs of the users.

Ideate

By now the problem is obvious and it is time to brainstorm ways and methods to solve it. At this stage, numerous ideas are generated as a part of the problem-solving exercise. In short, ideation is all about idea generation. During brainstorming, one should not be concerned if the generated ideas are possible, feasible, or even viable. The only task of the thinkers is to think of as many ideas as possible for them. It requires "going wide" mentally in terms of concepts and outcomes. There are many brainstorming tools that can be used during this stage.

By this time, you are already aware of who your target users are and what your problem statement is. Now it's time to come up with as many possible solutions. This phase is all about creativity and imagination; all types of ideas are encouraged, whether stupid or wise – it hardly matters as long as the solution is imagined.

Ideation is the most invigorating stage of Design Thinking, and consists of a process where any and all ideas are welcomed, no matter how outrageous they may seem. A lot of planning and preparation goes into this stage to ensure that the results are varied and innovative. After everyone shares their ideas, specific measures are applied to evaluate the ideas without being judgmental or critical to narrow the list. It may so happen that the solution comes from the unlikeliest of ideas. So, at this point focus is on quantity over quality of ideas. The most feasible ideas are chosen for further exploration. Storyboarding, or making a visual mock-up of an idea, can also be useful during ideation.

Prototype

The prototype stage involves creating a model designed to solve consumers' problems which is tested in the next stage of the process. Creating a prototype is not a detailed process. It may include a developing simple drawing, poster, group role-playing, homemade “gadget, or a 3d printed product.” The prototypes must be quick and easy to develop and cheap. Therefore, prototypes are visualised as rudimentary forms of what a final product is expected to look like. Prototyping is intended to answer questions that get you closer to your final solution. Prototypes, though quick and simple to make, bring out useful feedback from users. Prototypes can be made with everyday materials.

Test

One of the most important parts of the design thinking process is to test the prototypes with the end users. This step is often seen going parallel to prototyping. During testing, the designers receive feedback about the prototype(s), and get another opportunity to interact and empathize with the people they are finding solutions for. Testing focuses on what can be learned about the user and the problem, as well as the potential solution.

Having understood the different stages, let us see some of the best examples of Design Thinking. You will need to identify and highlight wherever you feel design thinking has been applied.

Example 1: Toilet tank cover with faucet

Example 2: Illuminated Light Switch

Example 3:

Fuel dispensers hanging overhead, unlike what is usually seen in our gas filling stations in India

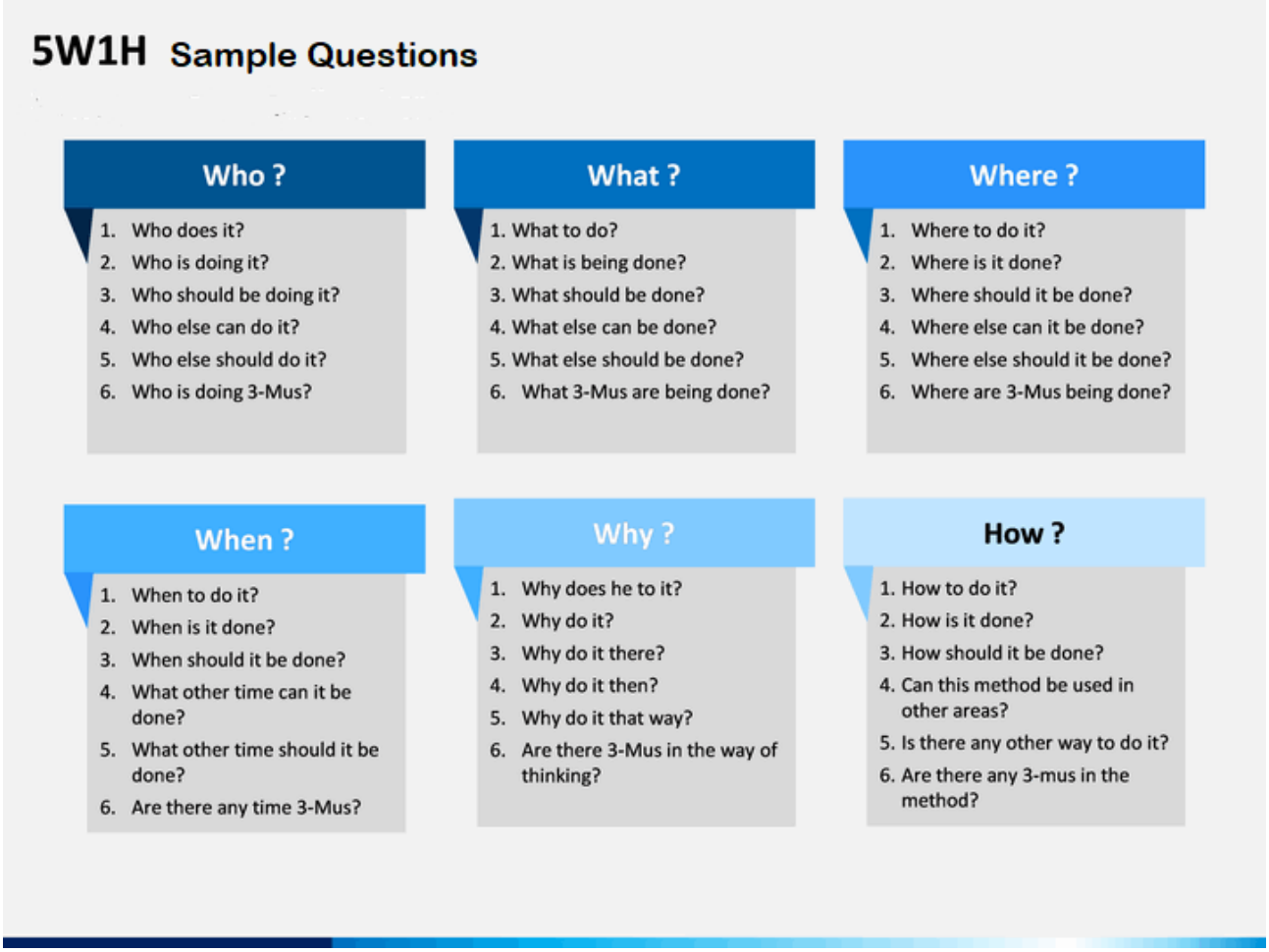
1.1 Right Questioning

Designers generally avoid asking questions to the users as they work against stringent deadlines which demand quick solutions and delivery. However, great questions often lead to better understanding of the problem which result in designing great solutions. In the process of developing solutions using design thinking framework, designers are expected to interact with customers / users very frequently to gather detailed facts about the problems and user's expectations. A detailed analysis of these facts leads to approaching the problem in best possible way.

In order to extract / gather relevant facts and information from users/customers, it is recommended to use this simple and reliable method of questioning: the 5W1H method.

(https://www.workfront.com/blog/project-management-101-the-5-ws-and-1-h-that-should-be-asked-of-every-project)

To collect facts and key information about the problem, ask and answer the 5 W's and One H question—***Who? What? When? Where? Why?*** and ***How?***



https://www.sketchbubble.com/en/presentation-5w1h-model.html

For instance, if one's car is giving inadequate gas mileage the following questions can be asked:

- Who recognized the problem or who drives the car?
- What has changed - for instance, maintenance and repairs done last, change in the gas station?
- When did the mileage start to deteriorate?
- Where are the new driving routes or distances that the car is covering?
- How the problem became noticeable? How can it be addressed?

The questions can be changed to make them pertinent to whatever problem or issue that needs to be addressed. The essential W's and H help to cover all aspects of a problem so that a comprehensive solution can be found.

Activity 1

Your best friend who had scored very high marks in the mid-term exams has surprisingly put up a poor performance in the final term exams. You decide to bring him back on track by spending time with him and try to extract facts to get to the root of the problem.

Use the below 5W1H worksheet given below to record the questions and answer with your friend -

Worksheet: 5W and 1H (for problem solving)

Five W's and One H	**Answer**
What is the Problem?	
Where is it happening?	
When is it Happening?	
Why is it happening?	
How can I help my friend overcome the problem?	
Why will I need to involve myself?	

1.2 Identifying the problem to solve

Problem solving is the act of defining a problem; determining the cause of the problem; brainstorming to generate probable solutions and selecting alternatives for the most suitable solution.

Problems are at the centre of what many people do at work every day. Whether you're solving a problem for a client (internal or external) or discovering new problems to solve - the problems you face can be large or small, simple or complex.

The problem, in the below picture may appear simple to you. Thinking every aspect from the perspective of giraffe, can you solve it for them?

It has often been found that finding or identifying a problem is more important than the solution. For example, Galileo recognised the problem of needing to know the speed of light, but did not come up with a solution. It took advances in mathematics and science to solve this measurement problem. Yet to date Galileo still receives credit for finding the problem.

Question -1: Rohan has been offered a job that he wants, but he doesn't have the facility to reach the office premises and also doesn't have enough money to buy a car."

What do you think is Rohan's main problem?

Question-2: Instructors at a large university do not show up for technology training sessions. What do you think is the problem?

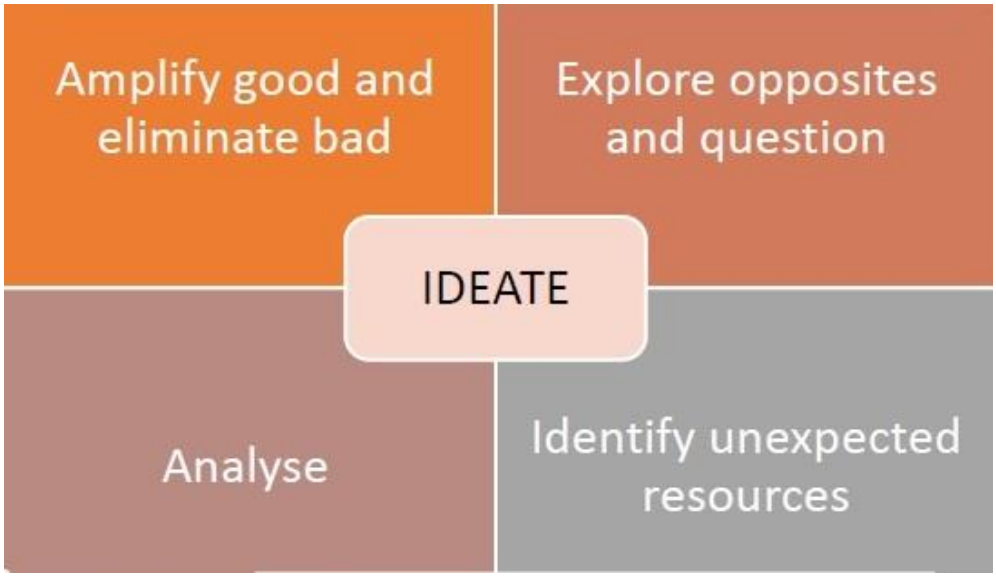
- The time frame for the training sessions does not meet the instructors' schedules.
- There is no reward for investing time in training sessions.
- The notifications for the training are sent in bulk mailings to all email accounts.

The define stage of design thinking (identify the problem) ensures you fully understand the goal of your design project. It helps you to articulate your design problem, and provides a clear-cut objective to work towards.

Without a well-defined problem statement, it's hard to know what you're aiming for. With this in mind, let's take a closer look at problem statements and how you can go about defining them.

1.3 Ideate

Ideation is the process of generating ideas and solutions through sessions such as sketching, prototyping, brainstorming etc. In the ideation stage, design thinkers generate ideas — in the form of questions and solutions — through creative and curious activities.



https://www.tutorialspoint.com/design_thinking/design_thinking_ideate_stage.htm

Ideation Will Help You:

- Ask the right questions and innovate with a strong focus on your users, their needs, and your insights about them.
- Bring together perspectives and strengths of your team members.
- Get obvious solutions out of your heads, and drive your team beyond them.

Ideation Techniques:

Here is an overview of the most essential ideation techniques employed to generate numerous ideas:

Brainstorm

During a Brainstorming session, students leverage the synergy of the group to generate new innovative ideas by building on others' ideas. Participants should be able to discuss their ideas freely without fear of criticism. A large number of ideas are collected so that different options are available for solving the challenge.

Brain dump

Brain dump is very similar to Brainstorm; however, it's done individually. It allows the concerned person to open the mind and let the thoughts be released and captured onto a piece of paper. The participants write down their ideas onto paper or post-it notes and share their ideas later with the larger group.

Brain writing

Brain writing is also very similar to a Brainstorm session and is known as 'individual brainstorming'. At times only the most confident of team members share their ideas while the introverts keep the ideas to themselves. Brainwriting gives introverted people time to write them down instead of sharing their thoughts out loud with the group. The participants write down their ideas on paper and, after a few minutes, pass on their own piece of paper to another participant who then elaborates on the first person's ideas and so forth. In this way all

participants pass their papers on to someone else and the process continues. After about 15 minutes, the papers are collected and posted for instant discussion.

Group Activity: Your class have been tasked with a responsibility - "How to redesign the classroom to better meet the students' needs without incurring any cost"?

Form groups of 4 or 5 students each. Apply the design thinking framework i.e. all five phases. Every group is supposed to submit a detailed report (not more than 10 pages) in a week's time to the teacher.

Tools supporting the design thinking process:

- https://www.google.com/keep/ (A collaborative note-taking tool that works with Google accounts.)
- https://www.sketchup.com/ (Free 3D digital design tool. Ideal for prototyping and mocking up design solutions.

A Focus on Empathy

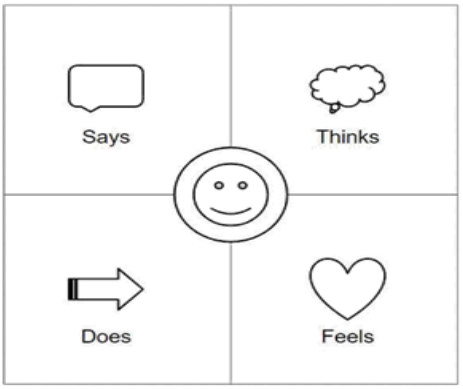
Empathy is the first step in design thinking because it allows designers to understand, empathize and share the feelings of the users. Through empathy, we can put ourselves in other people's shoes and connect with how they might be feeling about their problem, circumstance, or situation.

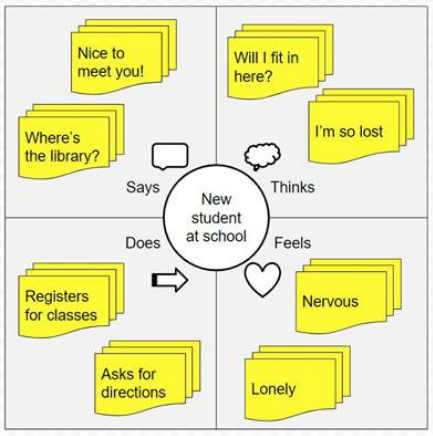
A big part of design thinking focuses on the nature of impact that innovative thinking has on individuals. Recall the students who were featured at the beginning of the module. Empathy was at the centre of their designs.

In preparation for your AI challenge, you are going to engage in an empathy map activity to practice one way of empathizing in the design process.

What's an Empathy Map?

Before we start to figure out what the problem is or try to solve it, it's always a good idea to "walk a mile in the user's shoes" and get an understanding of the user. An extremely useful tool for understanding the users' needs and gaining a deeper insight into the problem at hand is the empathy map. It also helps in deepening that understanding, gaining insight into the user's behaviour.





To create a "persona" or profile for the user, you can use the empathy map activity to create a realistic general representation of the user or users. Personas can include details about a user's education, lifestyle, interests, values, goals, needs, thoughts, desires, attitudes, and actions.

Activity on empathy map

Please look at the below links for the activity and the related video

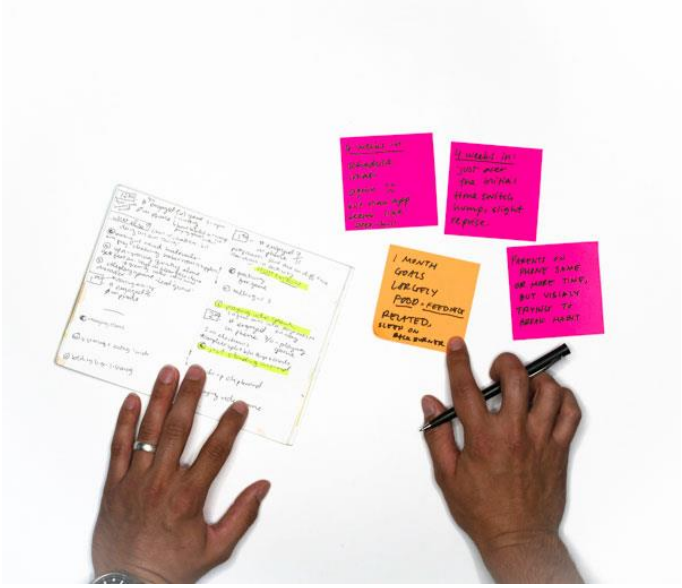
Source- https://www.ibm.com/design/thinking/page/toolkit/activity/empathy-map

Video link- https://video.ibm.com/recorded/116968874

Instructions:

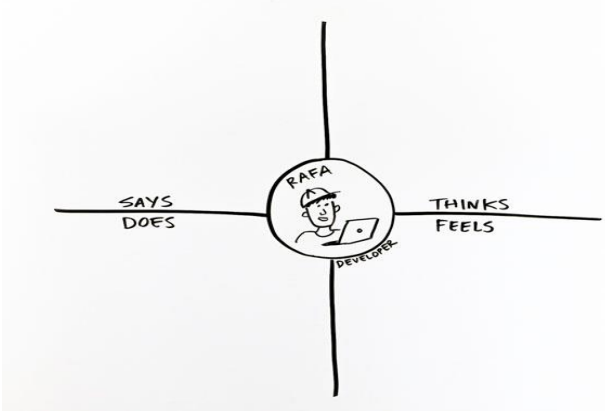
1. Come prepared with observations

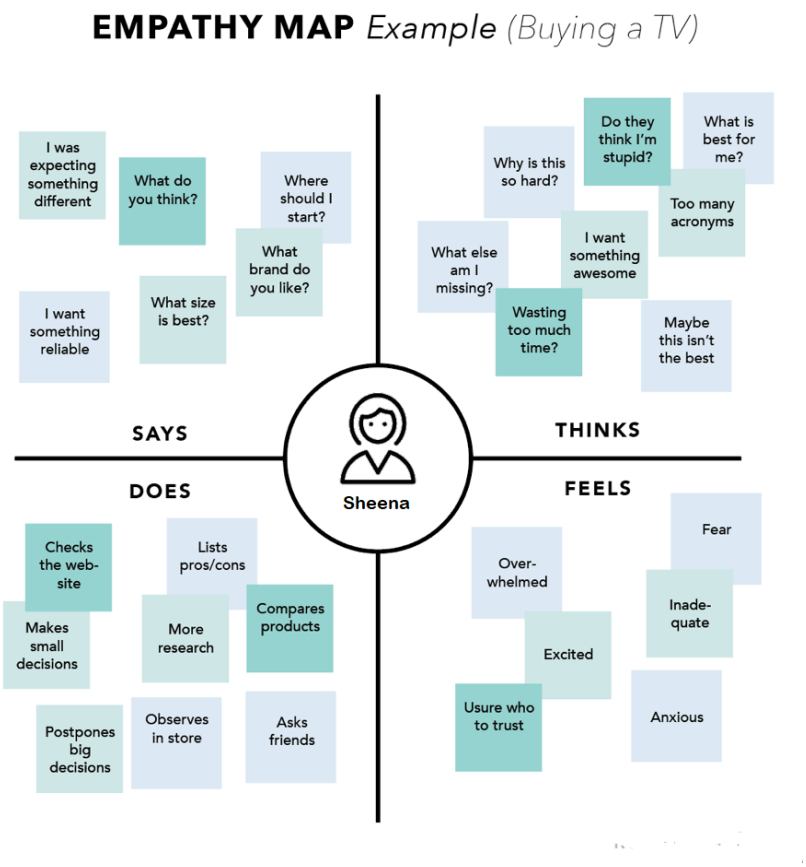
Empathy mapping is only as reliable as the data you bring to the table, so make sure you have defensible data based on real observations (for example, from an interview or contextual inquiry). When you can, invite users or Sponsor Users to participate.



2. Set up the activity

Draw a grid and label the four essential quadrants of the map: Says, Does, Thinks, and Feels. Sketch your user or stakeholder in the centre. Give them a name and brief description of who they are and what they do.

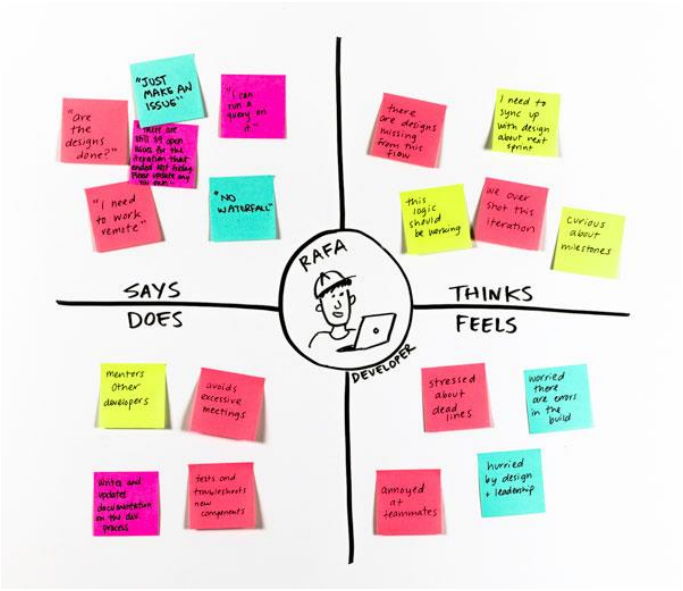




https://www.uxbooth.com/articles/empathy-mapping-a-guide-to-getting-inside-a-users-head/

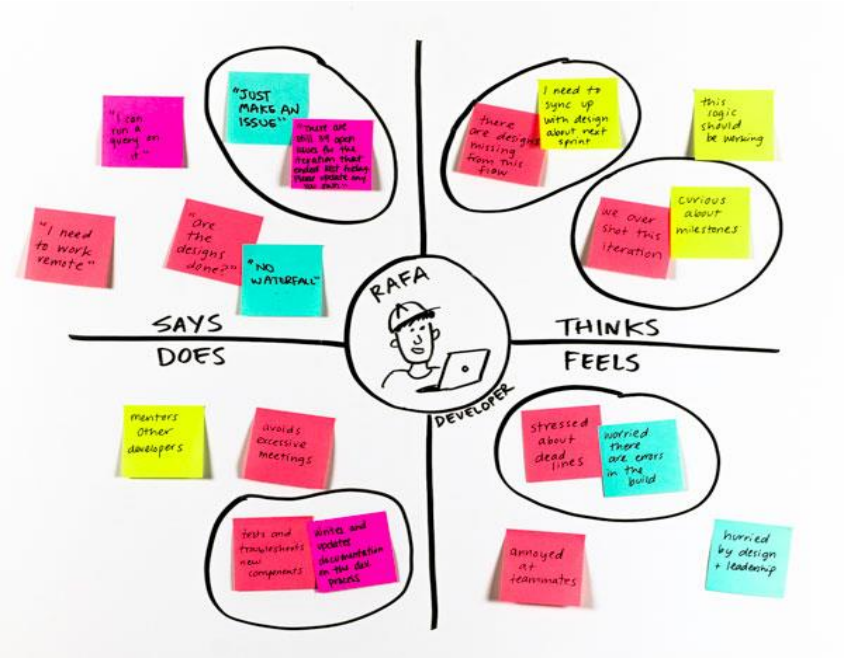
3. Capture observations

Have everyone record what they know about the user or stakeholder. Use one sticky note per observation. Place the sticky notes with the relevant answers on the appropriate quadrant of the map.



4. Find patterns and identify unknowns

Within each quadrant, look for similar or related items. If desired, move them closer together. As you do, imagine how these different aspects of your user's life really affect how they feel. Can you imagine yourself in their shoes?



5. Playback and discuss

Label anything on the map that might be an assumption or a question for later inquiry or validation. Look for interesting observations or insights. What do you all agree on? What surprised you? What's missing? Make sure to validate your observations with other participants involved in the activity.

You've been asked to build a mobile app that will help connect students and tutors.

- **Persona 1:** Neha is a high school student and is focused on maintaining a high Percentage to increase her chances of getting into her first-choice college after Class 12th. She is struggling with her Physics class and wants to find a tutor. She is looking for someone in her neighbourhood who she can meet with after school, possibly on Saturday mornings.

- **Persona 2:** Priya is a college student and an expert in Physics who would like to make a little extra money by helping students. She hopes to be a teacher one day and thinks being a tutor would help her gain experience and build her resume. She would like to offer her services to students looking for a Physics tutor.
- **Persona 3:** Mr. Jaswinder Singh is a high school teacher and has several students struggling with their Physics assignments. He would like to be able to direct his students to available tutors to help them improve their grades and catch up with the rest of the class. He also wants to be able to check the progress of his students to ensure they are taking appropriate steps to improve.

Unit 7

Data Analysis

Title: Data Analysis	**Approach: Interactive/ Discussion, Team Activity, Case studies**

Summary: In the AI age, where data is the new electricity, students need to know how to use, analyse and communicate data effectively. Data Analysis should not be limited to mathematics, statistics or economics, but should be a cross-curriculum concept.

Institutions like the World Bank to entities like the local government, organizations are becoming increasingly open about the information that they gather and are ready to share the same with the public. Those who know how to analyse and interpret data, can crunch those numbers to make predictions, identify patterns, explain historical trends, or find fault in arguments. Students who become data literate are better equipped to make sense of the information that's all around them so that they can support their arguments with reliable evidence.

Statistics is the science of data and its interpretation. In other words, statistics is a way to understand the data that is collected about us and the world; therefore, the basic understanding of statistics is important. There are statistics all around us – news, in scientific observations, sports, medicine, populations, and demographics. Understanding statistics is essential to understand research in the social sciences, science, medicine and behavioural sciences. In this unit you will learn the basics of statistics; not just how to calculate them, but also how to evaluate them. This module will also prepare you for the next unit of this Level-II.

Objectives:

1. Demonstrate an understanding of data analysis and statistical concepts.
2. Recognise the various types of structured data – string, date, etc
3. Illustrate an understanding of various statistical concepts like mean, median, mode, etc.

Learning Outcomes:

1. Comprehension and demonstration of data management skills.
2. Students will demonstrate proficiency in applying the knowledge in statistical analysis of data.

Pre-requisites: No previous knowledge is required, just an interest in methodology and data. All you need is an Internet connection.

Key Concepts: Data Analysis, Structured Data, Statistical terms and concepts

Q 1. What is your understanding of data? State with examples.

_____
_____
_____
_____
_____

Q 2. How is data collected?

_____
_____

Q 3. Why is data collected? State a few reasons you can think of.

_____
_____
_____
_____
_____

Q 4. What is the difference between data analysis and data interpretation?

_____
_____
_____
_____
_____

It is a widely known fact that Artificial Intelligence (AI) is essentially data-driven. AI involves converting large amounts of raw data into actionable information that carry practical value and is usable. Therefore, understanding the statistical concepts and principles are essential to Artificial Intelligence and Machine Learning. Statistical methods are required to find answers to the questions that we have about data. Statistics and artificial intelligence share many commonalities. Both disciplines have much to do with planning, combining evidence, and decision-making. We are aware that statistical methods are required to understand the data used to train a machine learning model and to interpret the results of testing different machine learning models.

The first section of this unit describes the different data types and how they get stored in a database. The second section of this unit deals with data representation. Data are usually collected in a raw format and thus difficult to understand. However, no matter how accurate and valid the captured data might be it would be of no use unless it is presented effectively. In the third part of the unit, we will get to learn what cases and variables are and how you can compute measures of central tendency i.e. mean, median, mode, and dispersion i.e. standard deviation and variance**.**

1. Types of Structured Data

Recalling what we learnt about structured data in Level I, we know that it is highly organised in a formatted repository and has a predefined data type and a definite structure. It fits neatly within fixed fields and columns and therefore can be easily stored and searched in a relational database management system (RDBMS). Some examples of structured data that we come across in daily life include names, dates, addresses, credit card numbers, stock information, etc. SQL (Structured Query Language) is used to manipulate structured data.

Structured data

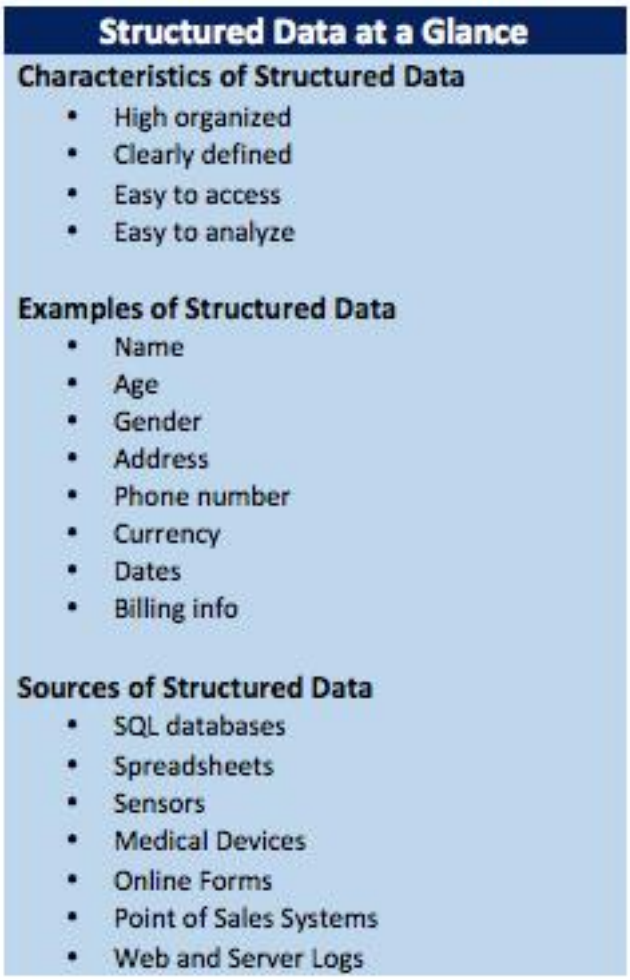
ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

Example of Structured data (https://www.researchgate.net/figure/Unstructured-semi-structured-and-structured-data_fig4_236860222)

Common sources of structured data are:

- Excel files
- SQL databases
- Medical devices Logs
- Online Forms

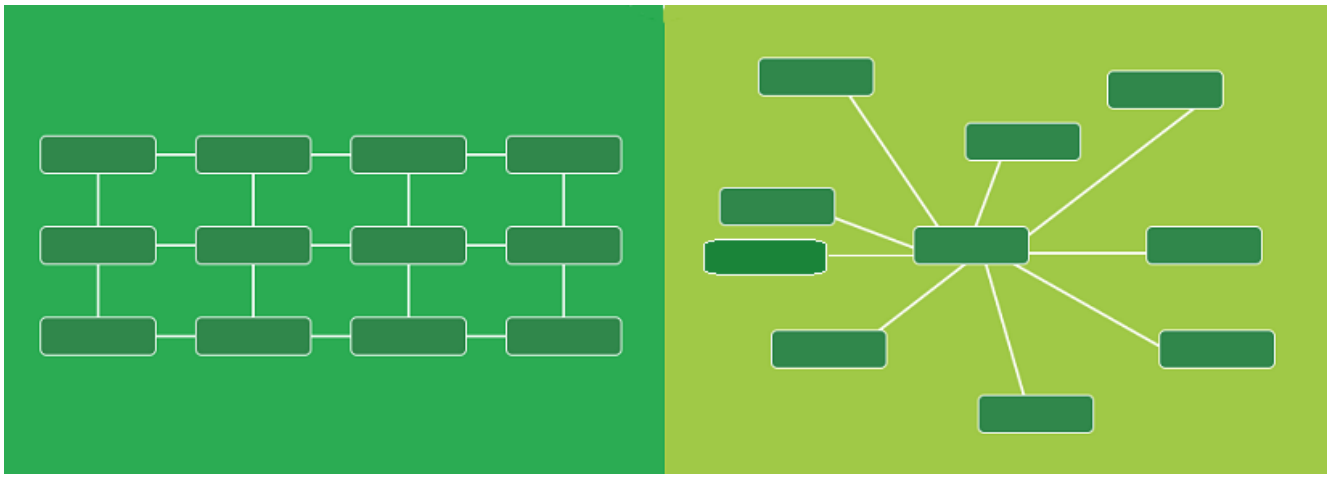
Each of these has structured rows and columns that can be sorted or manipulated. Structured data is highly organized and easily understood by machine language. The most attractive feature of the structured database is that those working within relational databases can easily input, search, and manipulate structured data.



https://www.datamation.com/big-data/structured-data.html

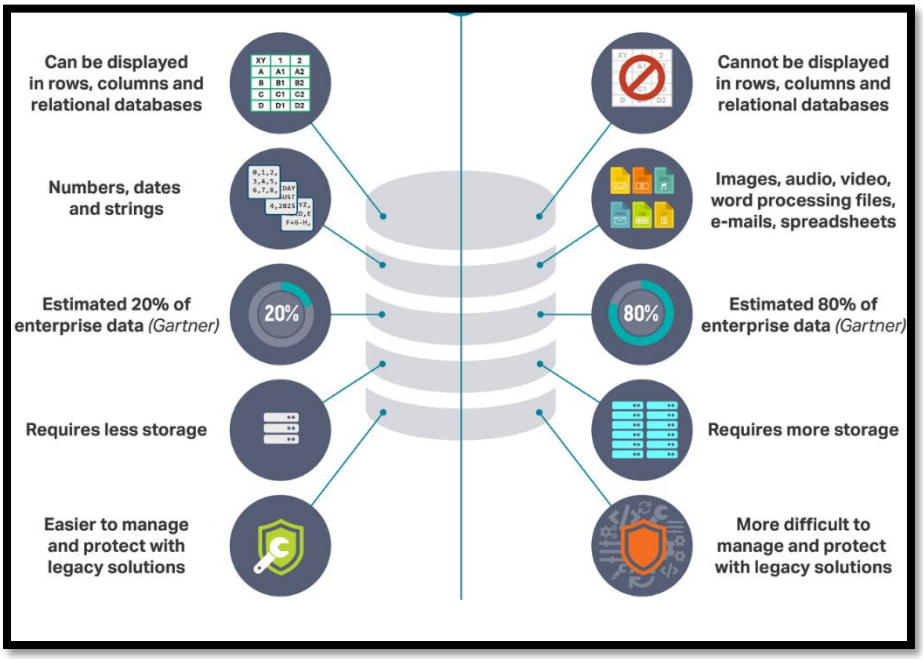
Activity 1

Tick the correct image depicting structured data depending on your understanding of the same:

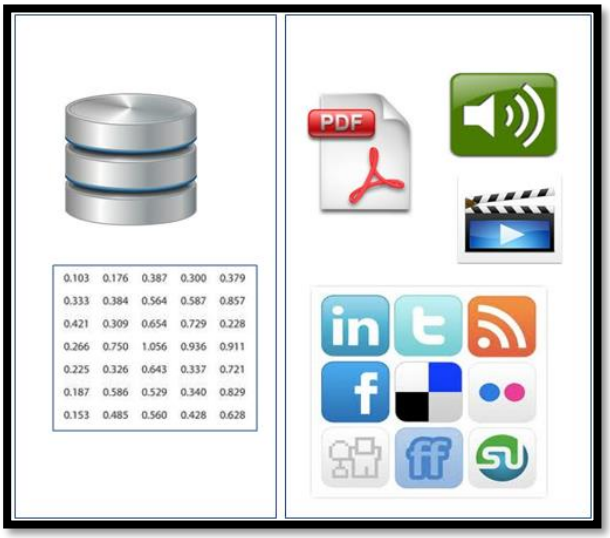


https://www.curvearro.com/blog/difference-between-structured-data-unstructured-data/

https://www.nbnminds.com/structured-data-vs-unstructured-data/



https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/



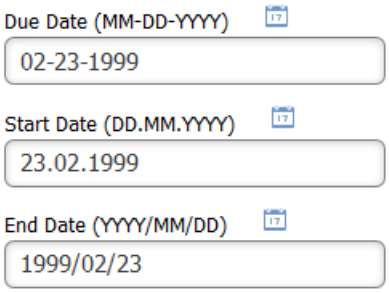
https://www.laserfiche.com/ecmblog/4-ways-to-manage-unstructured-data-with-ecm/

1.1 Date and Time Datatype

'Date and Time ' datatype is used to store values that contain both – date and time. There could be many formats, in which date-time data can be stored. Let us take one format for example

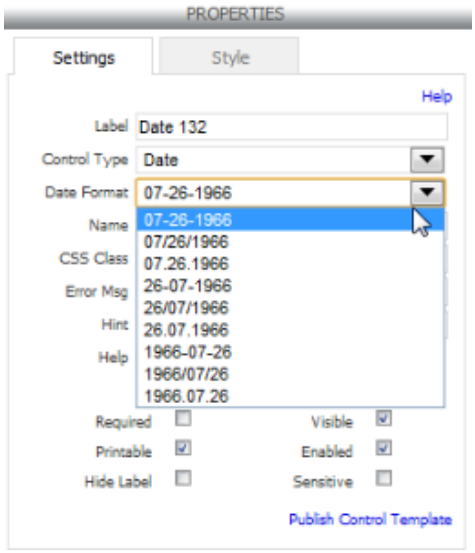
Data Type	Format
date	YYYY-MM-DD
time	HH:MM:SS
Year	YYYY

- Date data type helps us to specify the date in a particular format. Let's say if we want to store the date, 2 January 2019, then first we will give the year which would be 2 0 1 9, then the month which would be 01, and finally, the day which would be 02.
- Time data type helps us specify the time represented in a format. Let's say, we want to store the time 8:30:23 a.m. So, first, we'll specify the hour which would be 08, then the minutes which would be 30, and finally the seconds which would be 23. Year data type holds year values such as 1995 or 2011.



Sample Date-Time format (https://docs.frevvo.com/d/display/frevvo/Setting+Properties)

Some date format choices are given below:



https://docs.frevvo.com/d/display/frevvo/Setting+Properties

Activity

Write the date format used for the dates mentioned below. The first one has been solved as an example. Pay attention to the separators used in each case. You may use MM or mm to denote month, DD or dd for day and YY or yyyy for year.

a. mm-dd-yyyy - (07-26-1966)
b. _____ - (07/26/1966)
c. _____ - (07.26.1966)
d. _____ - (26-07-1966)
e. _____ - (26/07/1966)
f. _____ - (26.07.1966)
g. _____ - (1966-07-26)
h. _____ - (1966/07/26)
i. _____ - (1966.07.26)

1.2 String Data Type

A string is a structured data type and is often implemented as an array of bytes (or words) that stores a sequence of elements. A string can store alphanumeric data, which means a string can contain [A -Z], [as z], [0 -9] and [all special characters] but they are all considered as if they were text. It also contains spaces. String data must be placed within a quote (““or ' ‘).

Examples:

Address = “9th Floor, S**AS Tower, Gurgaon"**
“Hamburger”
“I ate 3 hamburgers”.

1.3 Categorical Data Types

Categorical data symbolises characteristics. Therefore, it illustrates things like a person's gender, language etc. Categorical data can also take on numerical values like 1 for female and 0 for male. However, these numbers don't have mathematical meaning. Mathematical calculations like addition and subtraction cannot be performed on the numbers. Categorical data is also looked upon as a collection of information that can be divided into groups i.e. the report cards of all students is referred to as categorical. This data is called categorical because it may be grouped depending on the variable present in the report card, like – class, subjects, sections, school-house, etc.

It means, the report card can be bundled section-wise, or class-wise, or house-wise, so section, class or house are the categories here. The easy way to determine whether the given data is categorical or numerical data is to calculate the average. If you can calculate the average, then it is considered to be a numerical data. If you cannot calculate the average, then it is considered to be a categorical data.

What flavor of ice cream would you pick?			
	Chocolate	Vanilla	Neither
Children	40	22	15
Teens	12	16	45
Adults	55	54	10
Total	107	92	70

https://study.com/academy/exam/topic/ppst-math-data-analysis.html

Eye Colour					
Hair Colour	Green	Blue	Brown	Black	Total
Blonde	4	7	2	1	14
Brown	2	4	18	2	26
Black	1	2	5	2	10
Total	7	13	25	5	50

http://www.intellspot.com/categorical-data-examples/

Question 1: Pin code of a place - Categorical data or numerical data?

Question 2: Date of birth of a person – Categorical data or numerical data?

Question 3: Refer to the table on ice cream and answer the following:

- How many belong to the group, 'Adults who like Chocolate ice creams'?
- Can you name the group which had 45 people in it?

Question 4: Refer to the table on hair colour, and answer the following:

- How many groups can be formed from this table?
- How many blondes have green eyes?

2. Representation of Data

According to Wikipedia, “Statistics is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data.” It is the science of data that transforms the observations into usable information. To achieve this task, statisticians summarize a large amount of data in a format that is compact and produces meaningful information. Without displaying values for each observation (from populations), it is possible to represent the data in brief while keeping its meaning intact using certain techniques called 'data representation'. It can also be defined as a technique for presenting large volumes of data in a manner that enables the user to interpret the important data with minimum effort and time.

Data representation techniques are broadly classified in two ways:

2.1. Non-Graphical technique: Tabular form and case form

This is the old format of data representation not suitable for large datasets. Non-graphical techniques are not so suitable when our objective is to make some decisions after analysing a set of data.

This is not the subject of our study in this unit.

2.2. Graphical Technique: Pie Chart, Bar graphs, line graphs, etc.

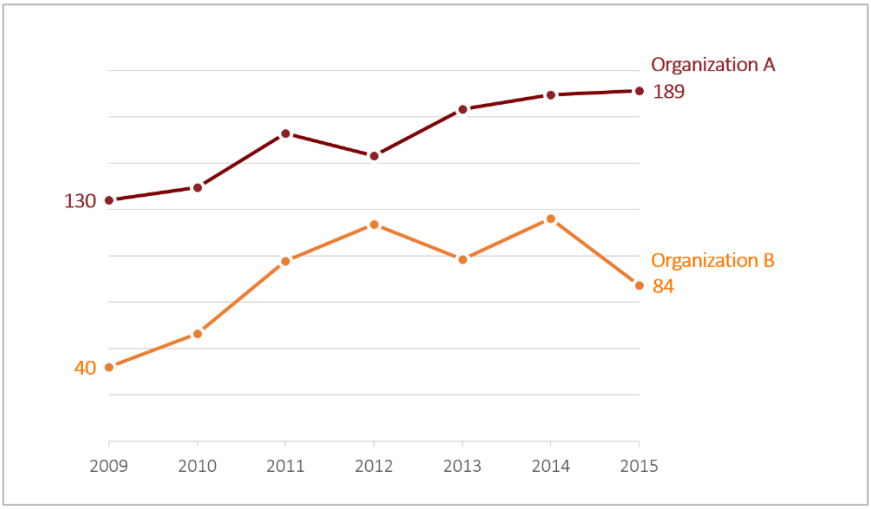
The visual display of statistical data in the form of points, lines, dots and other geometrical forms is most common. It would not be possible to discuss the methods of construction of all types of diagrams and maps primarily due to time constraint. We will, therefore, describe the most commonly used graphs and the way they are drawn.

These are:

- Line graphs
- Bar diagrams
- Pie diagram
- Scatter Plots

2.2.1 Line Graphs

A line graph also called the line chart is a graphical display of information that changes constantly over time Within a line graph, the data is connected by points which show a continuous change. The lines in a line graph can descend and ascend based on the data points it represents. We can use a line graph to represent the time series data related to temperature, rainfall, population growth, birth rates, death rates, etc.



Example of Line Graph (https://depictdatastudio.com/labeling-line-graphs/)

Rules to be followed during construction of Line Graph

(a) Simplify the data by converting it into round numbers such as the growth rate of the population as shown in the table for the years 1901 to 2001

(b) Draw an X and Y-axis. Mark the time series variables (years/months) on the X-axis and the data quantity/value to be plotted (growth of population) in percent on the Y-axis.

(c) Choose an appropriate scale and label it on Y-axis. If the data involves a negative figure then the selected scale should also show it.

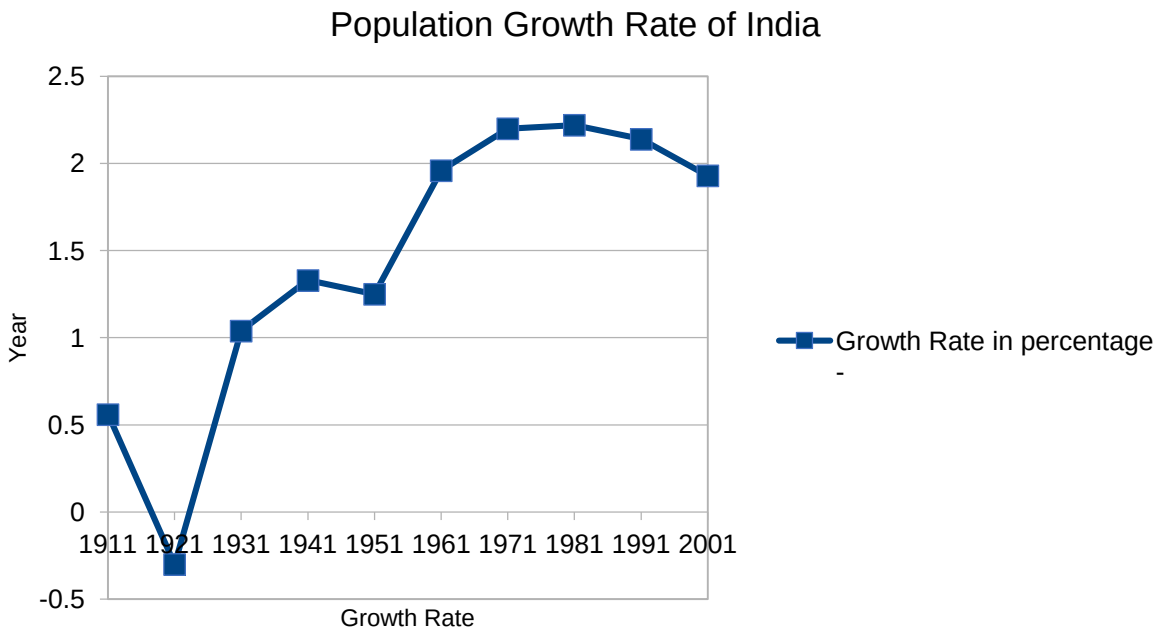
The advantages of using Line graph is that it is useful for making comparisons between different datasets, it is easy to tell the changes in both long and short term, with even small changes over time.

Let us take the following table:

SIZE AND GROWTH OF POPULATION IN INDIA: 1901 TO 2001

Census Year	*Average Annual Exponential Growth (%)*
1901	—
1911	0.56
1921	(–) 0.03
1931	1.04
1941	1.33
1951	1.25
1961	1.96
1971	2.20
1981	2.22
1991	2.14
2001	1.93

https://www.yourarticlelibrary.com/population/growth-of-population-in-india-1901-to-2001-with-statistics/39653



Activity 1: Find out the reasons for the sudden change in growth of population between 1911 and 1921 as is evident from the above graph.

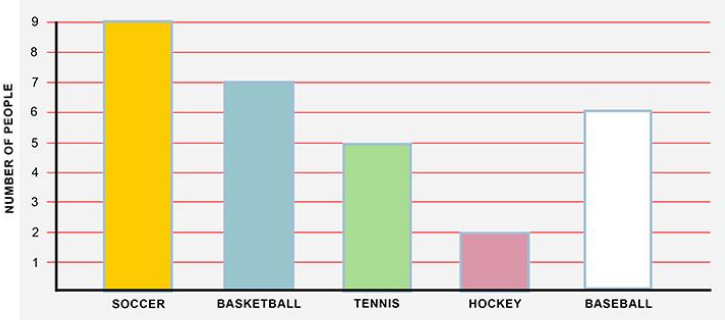
Activity 2: Between the student attendance data and student's score, which one according to you should be represented using the line graph?

Activity 3: Construct a simple line graph to represent the rainfall data of Tamil Nadu as shown in the table below

Months	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Rainfall (cm)	2.3	2.1	3.7	10.6	20.8	35.6	22.8	14.6	13.8	27.5	20.6	7.5

2.2.2 Bar Diagram

A bar graph (also known as a bar chart or bar diagram) is a visual tool in which the bars are used to compare data among categories. The length of the bar is directly proportional to the value it represents. In simple terms, the longer the bar, the greater the value it represents. The bars in the graph may run horizontally or vertically and are of equal width.



Example of Bar Graph (https://mammothmemory.net/maths/graphs/other-graphs-charts-and-diagrams/bar-graph.html)

Following rules should be observed while constructing a bar diagram:

(a) The width of all the bars or columns should be similar.

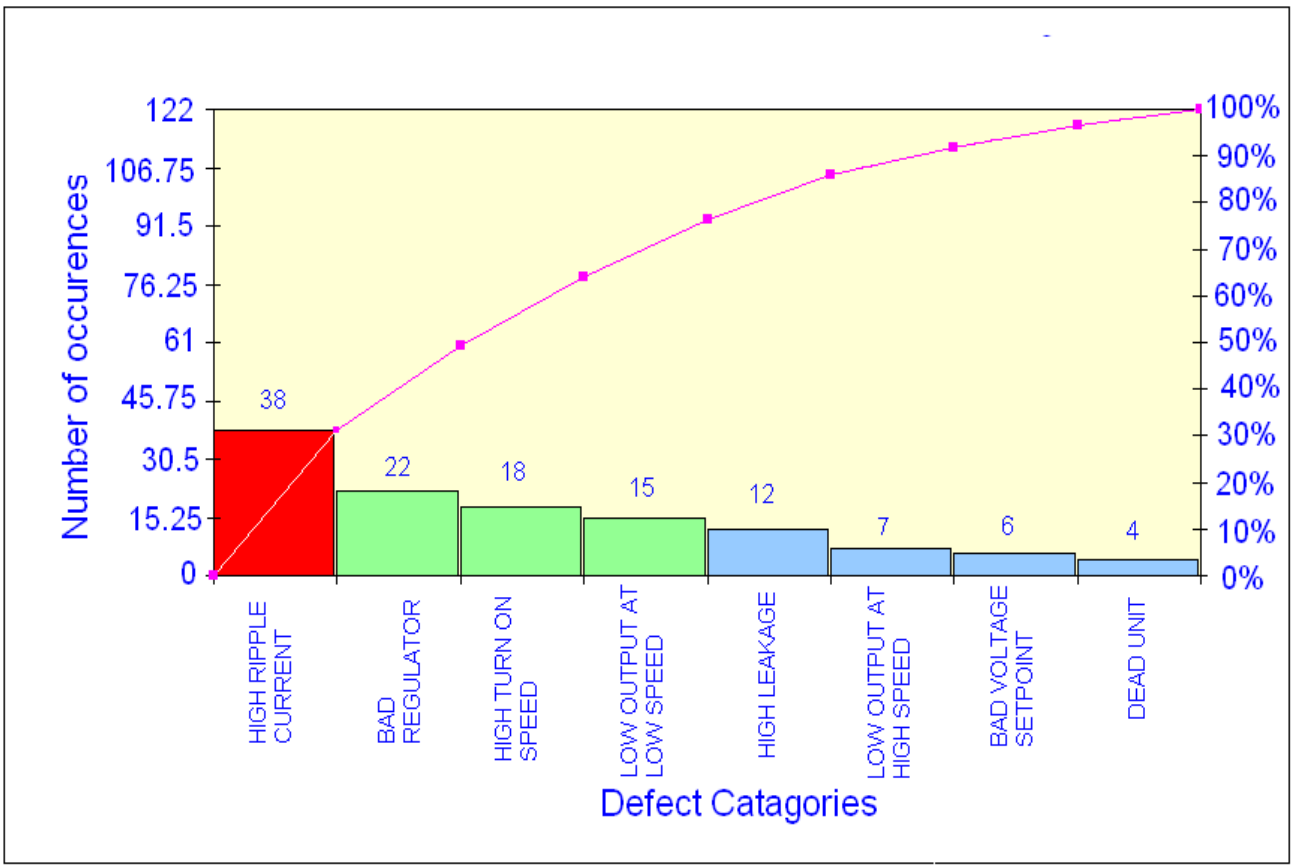
(b) All the bars should be placed on equal intervals/distance.

(c) Bars may be shared with colours or patterns to make them distinct and attractive.

The advantages of using a bar graph are many, it is useful for comparing facts, it provides a visual display for quick comparison of quantities in different categories, and they help us to ascertain relationships easily. Bar graphs also show big changes over time.

The following table shows the defects during production in a factory:

TYPE	Number of occurrences	% OF TOTAL
HIGH TURN ON SPEED	18	14.754
HIGH RIPPLE CURRENT	38	31.147
HIGH LEAKAGE	12	9.836
LOW OUTPUT AT LOW SPEED	15	12.295
LOW OUTPUT AT HIGH SPEED	7	5.737
DEAD UNIT	4	3.278
BAD REGULATOR	22	18.032
BAD VOLTAGE SETPOINT	6	4.918



Using the information being depicted in the graph above, answer the questions below:

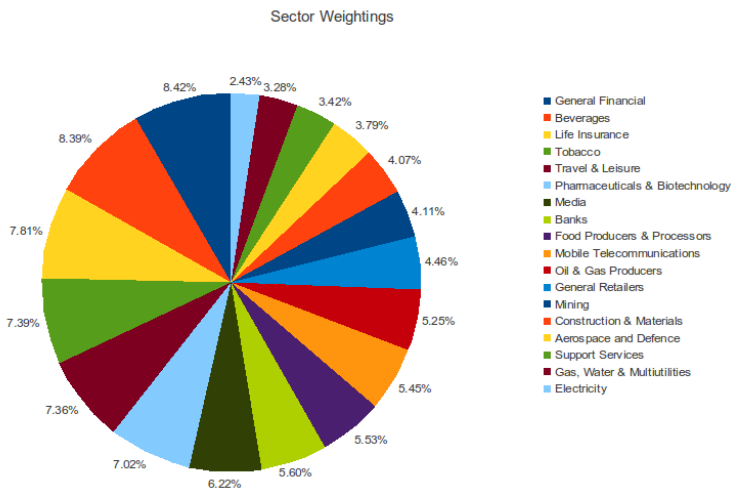
Question 1: Which type of defect has the highest occurrence?

Question 2: Name the types of faults whose occurrence is below 10.

2.2.3 Pie Chart

A pie chart is a circular graph in which the circle is divided into many segments or sections. Each division (segment/ sector) of the pie shows the relative size i.e. each category's contribution or a certain proportion or percentage of the total. The entire diagram resembles a pie and each component resembles a slice. Pie charts are a popular means to visualize data taken from a small table. It is a best practice to have not more than seven categories in a pie chart. Zero values cannot be represented in such graphs. However, such graphs are hard to interpret and difficult to compare with data from another pie chart.

Pie charts are used to for representing compositions or when trying to compare parts of a whole. They do not show changes over time. Various applications of pie charts can be found in business, school and at home. For business, pie charts can be used to show the success or failure of certain products or services. At school, pie chart applications include showing how much time is allotted to each subject. At home, pie charts can be used to see the expenses of monthly income on different goods and services.



Example of Pie Chart (https://brilliant.org/wiki/data-presentation-pie-charts/)

The advantages of a pie chart is that it is simple and easy-to-understand and provides data comparison at a glance.

Imagine, you survey your class to find what kind of books they like the most. You recorded your findings in the table for all the (40) students in the class

Step 1: Record your observation in a table

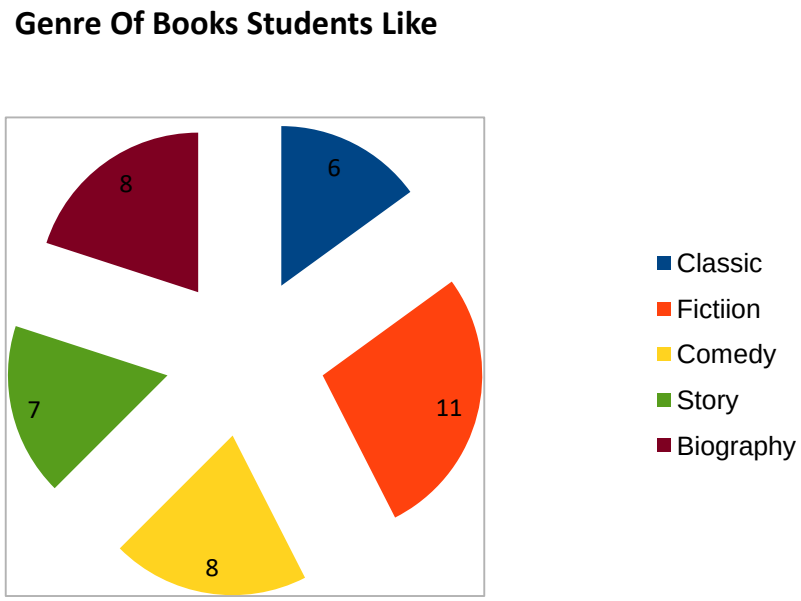
Classic	Fiction	Comedy	Story	Biography
6	11	8	7	8

Step 2: Add up the total observations to check if totals up to 40

Classic	Fiction	Comedy	Story	Biography	Total
6	11	8	7	8	40

Step 3: Next, divide each value by the total and multiply by 100 to get the percentage

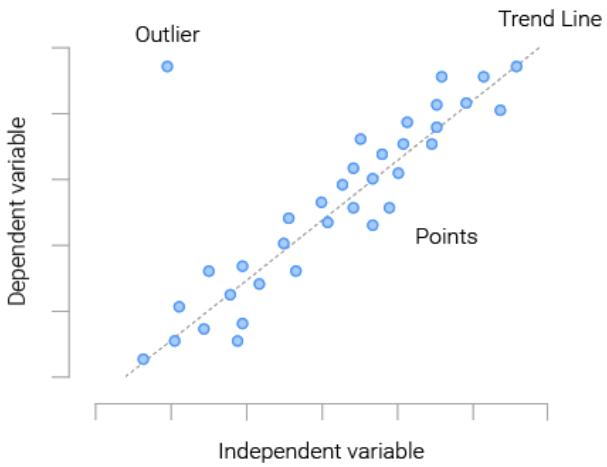
Classic	Fiction	Comedy	Story	Biography	Total
6	11	8	7	8	40
(6/40) * 100 **= 15%**	(11/40) * 100 **= 27.5 %**	(8/40) * 100 **= 20%**	(7/40) * 100 **= 17.5%**	(8/40) * 100 **= 20%**	(40/40) * 100 **= 100%**



2.2.4 Scatter Plots

Scatter plots are a set of data points plotted along the x and y axis and is used to represent the relationship between two variables (or aspects) for a set of paired data. The shape the data points assume narrates a unique story, most often revealing the correlation (positive or negative) in a large amount of data. The pattern of the scatter describes the relationship as shown in the examples below. A scatter plot is a graph of a collection of ordered pairs (XXY), with one variable on each axis.

Scatter plots are used when there is paired numerical data and when the dependent variable may have multiple values for each value of your independent variable. The advantage of scatter graph lies in its ability to portray trends, clusters, patterns, and relationships.



Example of scatter plot (https://www.learnbyexample.org/r-scatter-plot-base-graph/)

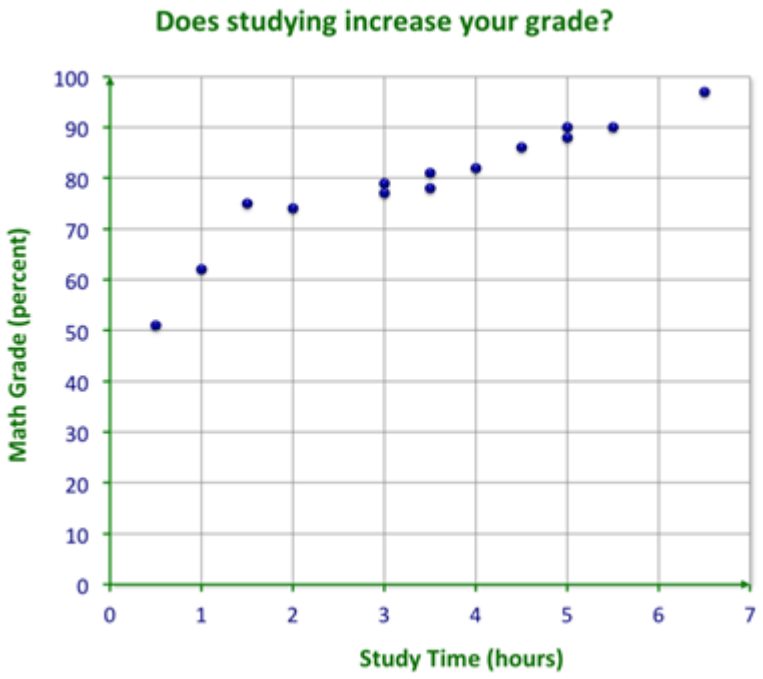
A student had a hypothesis for a science project. He believed that the more the students studied Math, the better their Math scores would be. He took a poll in which he asked students the average number of hours that

they studied per week during a given semester. He then found out the overall percentage that they received in their Math classes. His data is shown in the table below:

Study Time (Hours)	4	3.5	5	2	3	6.5	0.5	3.5	4.5	5
Maths Grade (%)	82	81	90	74	77	97	51	78	86	88

To understand this data, he decided to make a scatter plot.

The independent variable, or **input data**, is the study time because the hypothesis is that the Math grade depends on the study time. That means that the Math grade is the dependent variable, or the **output data**. The input data is plotted on the x-axis and the output data is plotted on the y-axis.



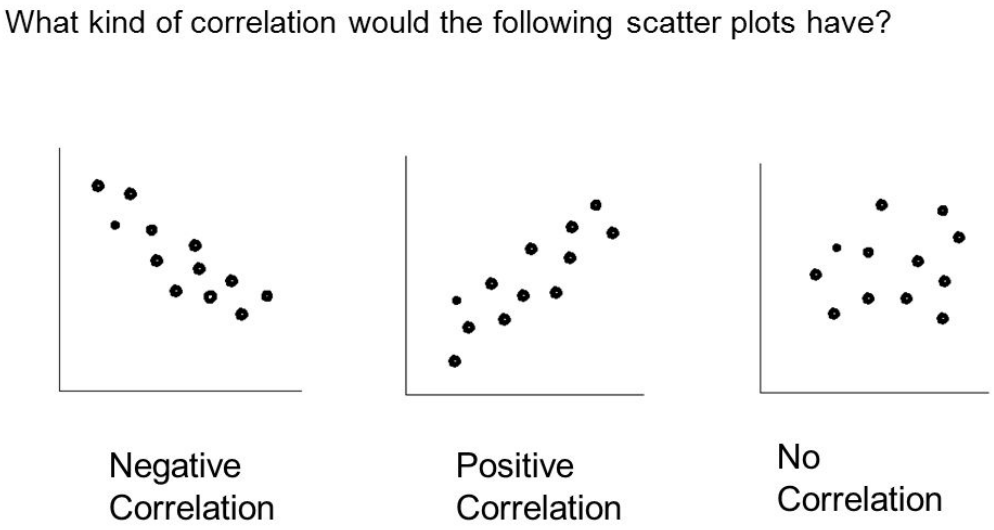
2.2.5 Types of Correlation

Positive Correlation: Both variables are seen to be moving in the same direction. In other words, with the increase in one variable, the other variable also increases. As one variable decreases, the other variable is also found to be decreasing. This means that data points along both x and y – coordinates increase and are related. E.g. Years of education and annual salary is positively correlated.

Negative Correlation: Both the variables are seen to be moving in opposite directions. While one variable increases, the other variable decreases. As one variable decreases, the other variable increases. If among the data points along the x – coordinate and the y – coordinate, one increases and the other decreases it is termed as a negative correlation.

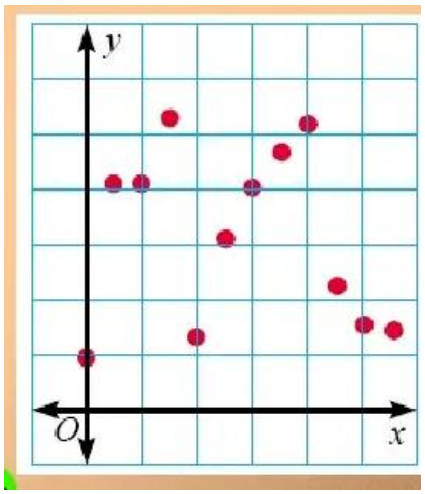
E.g. When hours spent sleeping increases hours spent awake decreases, so they are negatively correlated.

No correlation: If no relationship becomes evident between the two variables then there is no correlation. E.g. Eg: There is no correlation between the amount of tea consumed and the level of intelligence.



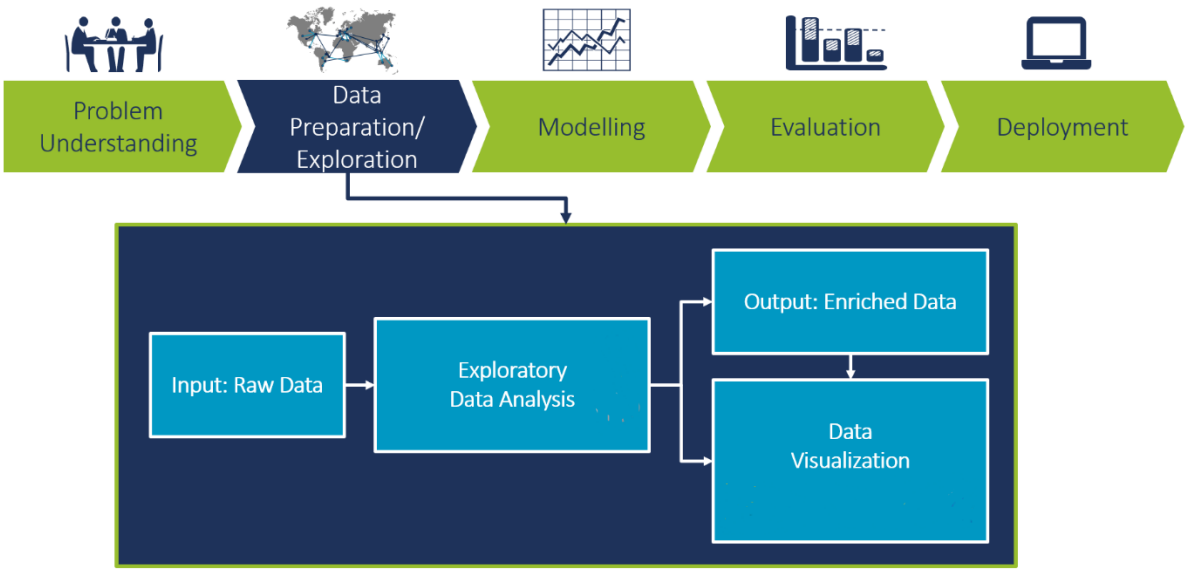
Example of scatter Plots with different correlations (https://slideplayer.com/slide/9489537/)

Activity 1: What type of correlation do see in the below graph?



3. Exploring Data

Exploring data is about "getting to know" the data: and its values - whether they are typical, unusual; centered, spread out; or whether they are extremes. More importantly, during the process of exploration one gets an opportunity to identify and correct any problems in your data that would affect the conclusions you draw in any way during analysis. This is the first step in data analysis and involves summarizing the main characteristics of a dataset, such as its size, accuracy, initial patterns in the data and other attributes.



Pictorial representation of Data Exploration (https://blog.camelot-group.com/2019/03/exploratory-data-analysis-an-important-step-in-data-science/)

3.1 Case, Variables and Levels of Measurement

3.1.1 Cases and Variables

Imagine you are very interested in cricket. You want to all the details about that game - how many matches won by a team, how many wickets were taken by a particular bowler or how many runs scored by a batsman?

The number of runs, wickets of a team can be expressed in terms of variables and case.

Before getting into the proper definition of case or variable let us try to understand its meaning in simple words:

“Variables” are the features of someone / something and "Case" is that something/ someone. So, here Cricket is the case and features of cricket like wickets, runs, win, etc are the variables.

Example 1:

Take another example, you want to know the age, height and address of your favourite cricket player.

Question-1: What are the variables here?

Question-2: What is the case here?

Example 2:

Let us take one more example where data is collected from a sample of STAT 200 students. Each student's major, quiz score, and lab assignment score is recorded.

Question -3: Name the variables here

_____

Question-4: Name the case here

_____

Example 3:

A fourth-grade teacher wants to know if students who spend more time studying at home get higher homework and exam grades.

Question 5: Name the variables

_____

Question 6: Name the case

_____

So, given the examples you came across here, you would have understood that a **dataset** contains information about a **sample**. Hence a dataset is said to consist of **cases** and cases are nothing but a collection of objects. It must now also be clear that a **variable** is a characteristic that is measured and whose value can keep changing during the program. In other words, something that can vary. This is in striking contrast to a constant which is the same for all cases in a study.

Example:

Let's say you are collecting blood samples from students in a school for a CBC test, where the following components would be measured:

- Haemoglobin level
- White Blood Cells
- Red Blood Cells
- Platelets

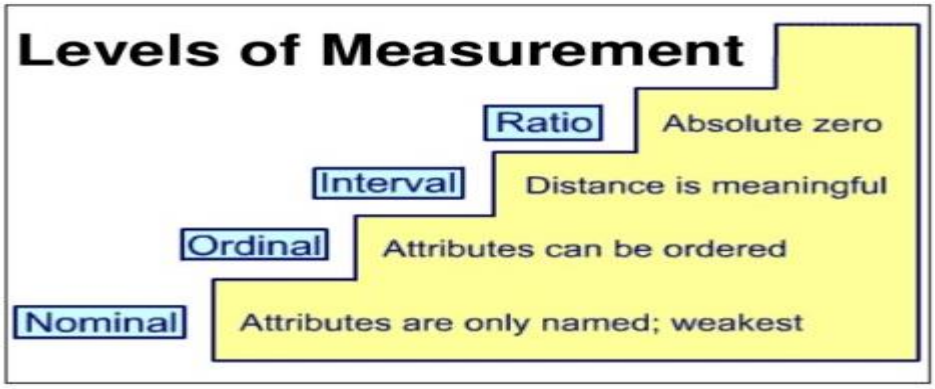
The students are the **cases** and all the components of blood are the **variables**.

Take another example, x = 10, this means that x is variable that stores the value 10 in it.

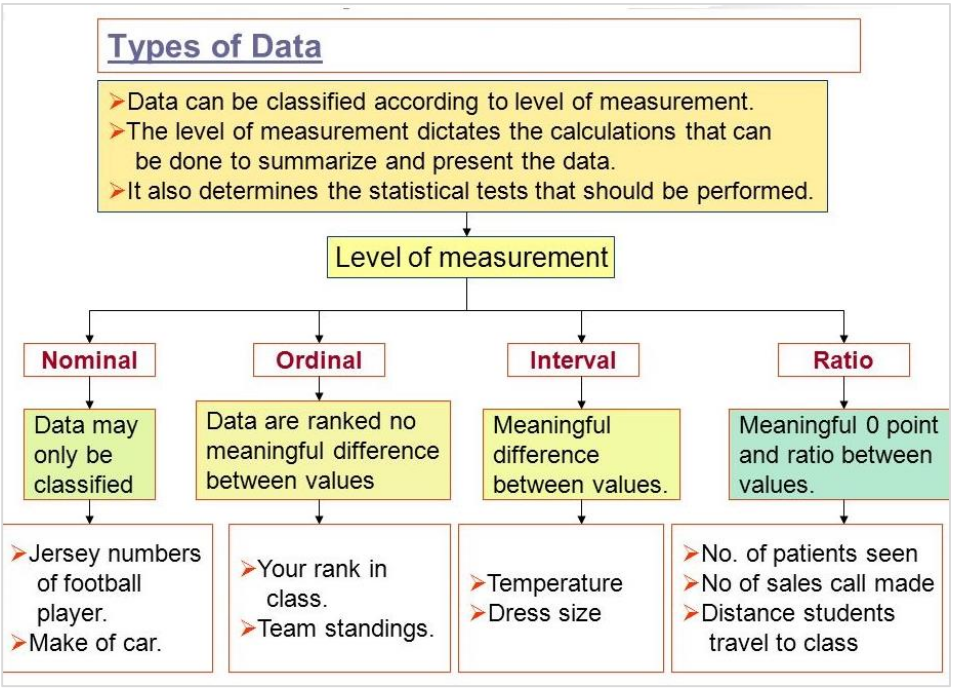
x = x + 5, name of variable is still x but its value has changed to 15 due to the addition of a constant 5

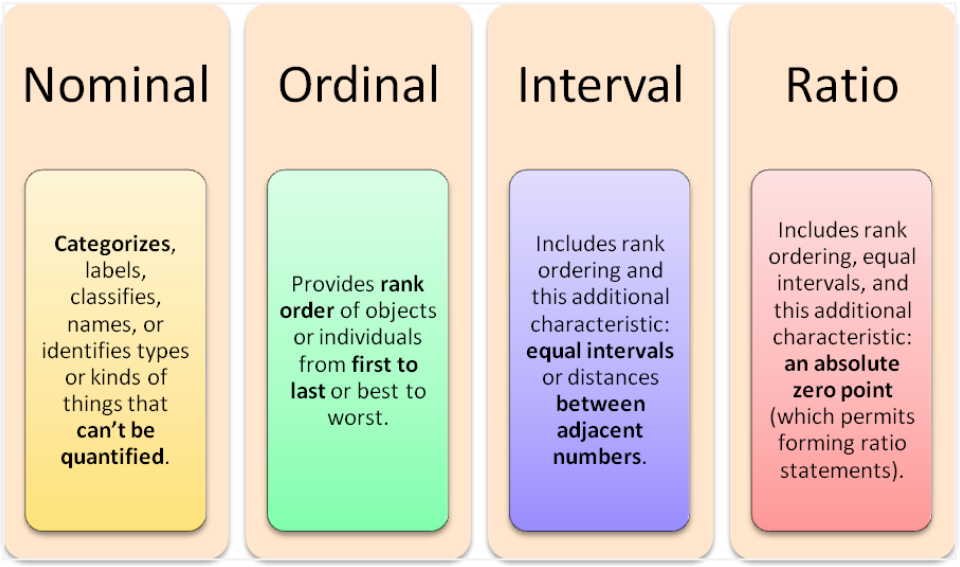
3.1.2 Levels of Measurement

The way a set of data is measured is called the **level of measurement**. Not all data can be treated equally. It makes sense to classify data sets based on different criteria. Some are quantitative, and some qualitative. Some data sets are continuous and some are discrete. Qualitative data can be nominal or ordinal. And quantitative data can be split into two groups: interval and ratio.



https://slideplayer.com/slide/8137745/





Example of data categorisation in the four Levels of Measurement

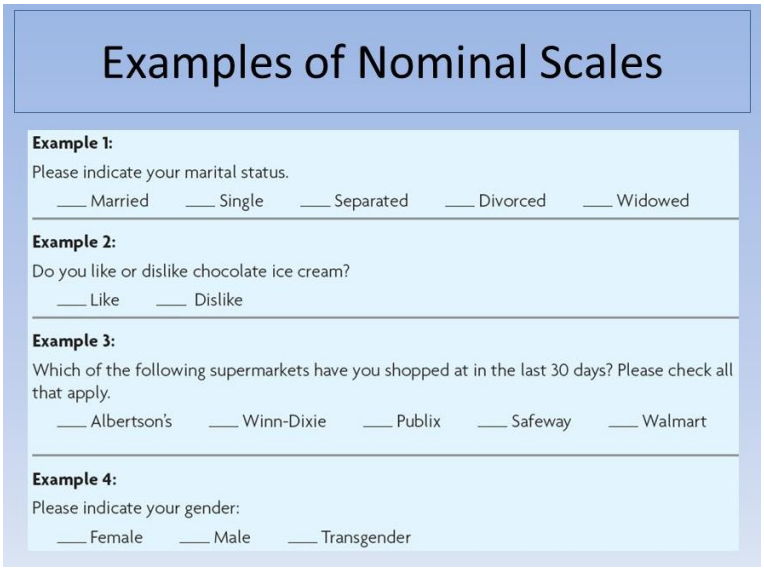
Broadly, there are four levels of measurement for the variables

1. Nominal Level

Data at the nominal level is qualitative. Nominal variables are like categories such as Mercedes, BMW or Audi, or like the four seasons – winter, spring, summer and autumn. They aren't numbers, and cannot be used in calculations and neither in any order or rank. The nominal level of measurement is the simplest or lowest of the four ways to characterize data. Nominal means "in name only".

Colours of eyes, yes or no responses to a survey, gender, smartphone companies, etc all deal with the nominal level of measurement. Even some things with numbers associated with them, such as a number on the back of a cricketer's T-shirt are nominal since they are used as "names" for individual players on the field and not for any calculation purpose.

.



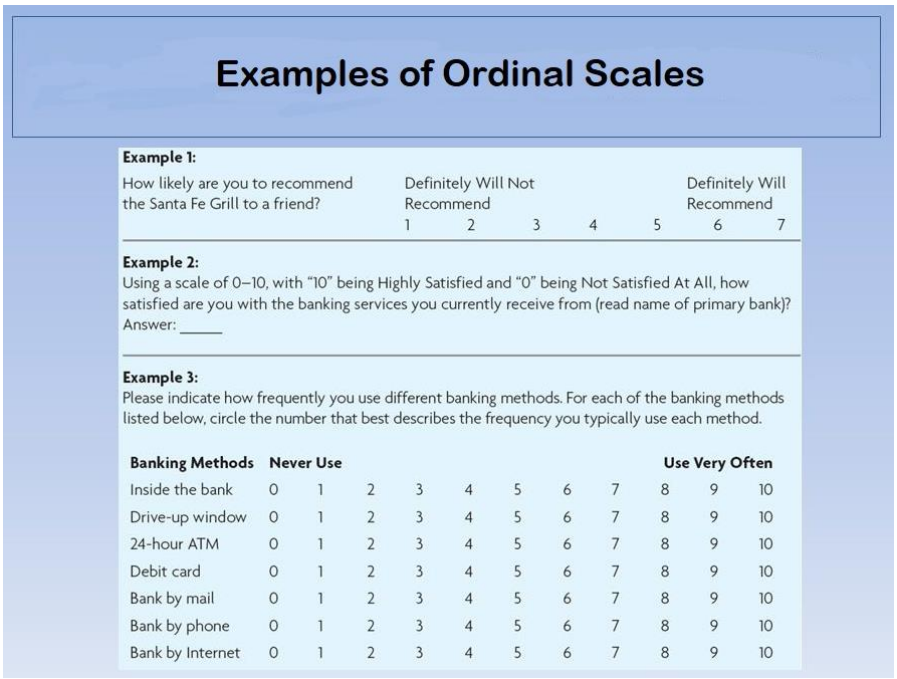
https://slideplayer.com/slide/8059841/

2. Ordinal Level

Ordinal data, is made up of groups and categories which follow a strict order. For e.g. if you have been asked to rate a meal at a restaurant and the options are: unpalatable, unappetizing, just okay, tasty, and delicious. Although the restaurant has used words not numbers to rate its food, it is clear that these preferences are ordered from negative to positive or low to high, thus the data is qualitative, ordinal. However, the difference between the data cannot be measured. Like the nominal scale data, ordinal scale data cannot be used in calculations.

A Hotel industry survey where the responses to questions about the hotels are accepted as, "excellent," "good," "satisfactory," and "unsatisfactory." These responses are ordered or ranked from the excellent service to satisfactory response to the least desired or unsatisfactory. But the differences between the two pieces of data as seen in the previous case cannot be measured.

Another common example of this is the grading system where letters are used to grade a service or good. You can order things so that A is higher than a B, but without any other information, there is no way of knowing how much better an A is from a B.



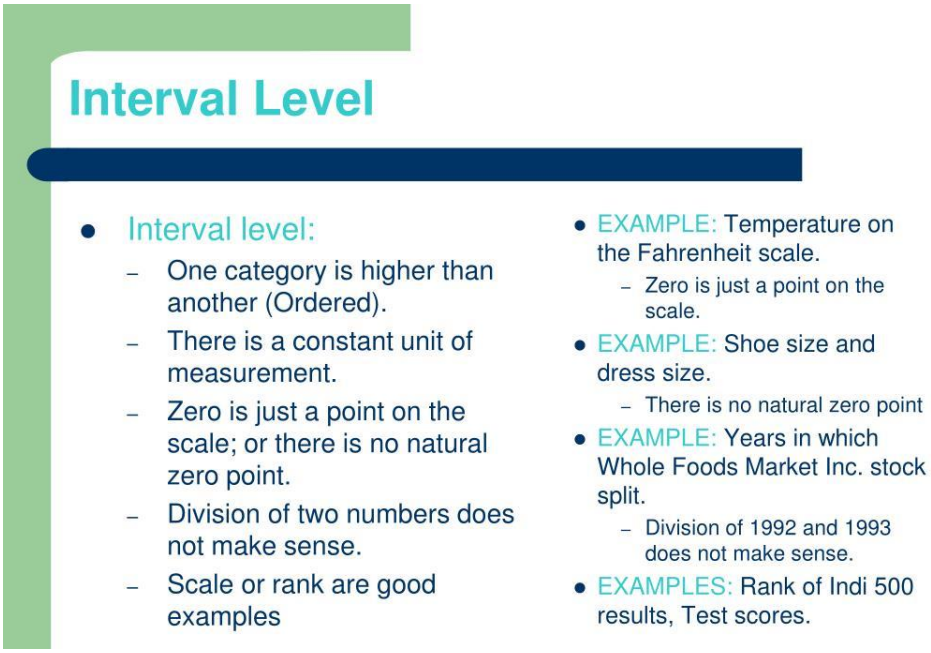
https://slideplayer.com/slide/6564103/

3. Interval Level

Data that is measured using the interval scale is similar to ordinal level data because it has a definite ordering but there is a difference between the two data. The differences between interval scale data can be measured though the data does not have a starting point i.e. zero value.

Temperature scales like Celsius (°C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements, 40° is equal to 100° minus 60°. Differences make sense. But 0 degrees does not because, in both scales, 0 is not the absolute lowest temperature. Temperatures like -20° F and -30° C exist and are colder than 0.

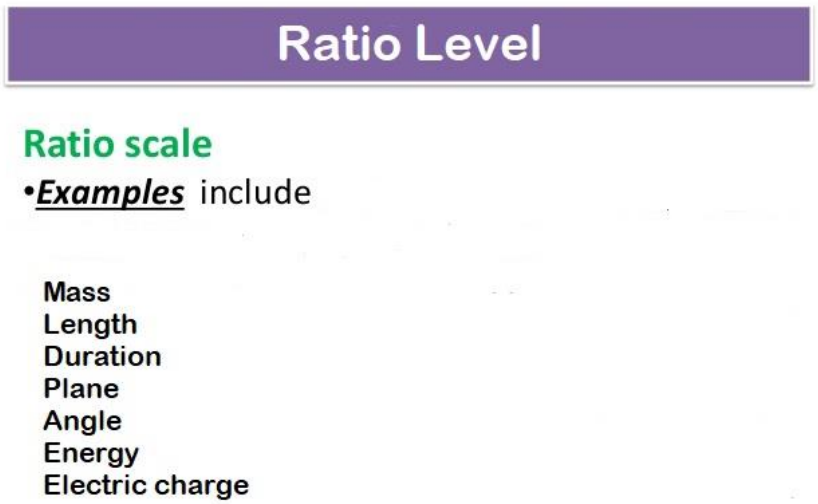
Interval level data can be used in calculations, but the comparison cannot be done. 80° C is not four times as hot as 20° C (nor is 80° F four times as hot as 20° F). There is no meaning to the ratio of 80 to 20 (or four to one).



4. Ratio Scale Level

Ratio scale data is like interval scale data, but it has a 0 point and ratios can be calculated. For example, the scores of four multiple choice statistics final exam questions were recorded as 80, 68, 20 and 92 (out of a maximum of 100 marks). The grades are computer generated. The data can be put in order from lowest to highest: 20, 68, 80, 92 or vice versa. The differences between the data have meaning. The score 92 is more than the score 68 by 24 points. Ratios can be calculated. The smallest score is 0. So, 80 is four times 20. The score of 80 is four times better than the score of 20.

So, we can add, subtract, divide and multiply the two ratio level variables. Egg: Weight of a person. It has a real zero point, i.e. zero weight means that the person has no weight. Also, we can add, subtract, multiply and divide weights at the real scale for comparisons.



Activity

1. Student Health Survey – Fill in the response and mention appropriate Level of Measurement

Query	Response	Level of Measurement
Sex (Male/ Female)		
Height (in metres)		
Weight (in kilograms)		
Rate overall health (Excellent; Good; Average; Below Average; Poor)		
Pulse rate (in BPM)		
Body temperature (in Fahrenheit)		
Country of residence		

2. State whether or not the following statements are true
   - All ordinal measurements are also nominal
   - All interval measurements are also ordinal
   - All ratio measurements are also interval

3. Indicate whether the variable is ordinal or not. If the variable is not ordinal, indicate its variable type.
   - Opinion about a new law (favour or oppose)
   - Letter grade in an English class (A, B, C, etc.)
   - Student rating of teacher on a scale of 1 – 10.

3.2 Data Matrix and Frequency Tables

If you are conducting a statistical study or research, you need to think of your data in terms of variables and cases. In this topic, you will learn how to present and order your variables and cases.

3.2.1 Data Matrix

When data is collected about any case, it is finally stored in a data matrix that contains a certain number of rows and columns. The data is usually arranged in such a way, that each row of the data matrix contains details about the case. The rows, therefore, represent the case or samples, whereas the columns represent the variable.

So, the tabular format of representation of cases and variables being used in your statistical study is known as the **Data Matrix**. Each row of a data matrix represents a case and each column represent a variable. A complete Data Matrix may contain thousands or lakhs or even more cases.

Each cell contains a single value for a particular variable (or observation).

Imagine you want to create a database of top 3 scorers in each section of every class of your school. The case you are interested in is individual students (top 3) and variables you want to capture – name, class, section, age, aggregate %, section rank, and address.

The best way to arrange all this information is to create a data matrix.

Name	Class	Section	Age	Aggregate %	Section Rank	Address
A	X	M	16	92	3	Add1
B	X	M	15	98	1	Add2
C	X	M	16	95	2	Add3
D	IX	N	14	96	1	Add4
E	IX	N	14	95	2	Add5
.....	.....	.....	.....	.....	.....	.....
Z	IV	M	9	97	1	Add10

Activity 1

Flipping a coin

Suppose you flip a coin five times and record the outcome (heads or tails) each time

1. What would be the observations?
2. What is the variable? (note there is only one)
3. Flip a coin and record the outcome in a data matrix as below

	Outcomes
1	
2	
3	
4	
5	

Activity 2

The car data set was prepared by randomly selecting 54 cars and then collecting data on various attributes of these cars. The first ten observations of the data can be seen in the data matrix below:

CAR	**TYPE**	**PRICE**	**DRIVETRAIN**	**PASSNEGERS**	**WEIGHT**
#1	Small	15.9 L	Front	5	2705
#2	Midsize	33.9 L	Front	5	3560
#3	Midsize	33.9 L	Front	6	3405
#4	Midsize	30.0 L	Rear	4	3640
#5	Midsize	15.7 L	Front	6	2880
#6	Large	20.8 L	Front	6	3470

1. What are the observations in this data matrix?
2. What are the variables in this data matrix?

Activity 3: Prepare a data matrix to record sales of different types of fruits from a grocery store. Note variables can be weight, per unit cost, total cost.

3.2.1 Frequency Tables

The frequency of a particular data value is the number of times the data value occurs (occurrences) in a given set of data. Let's say, if four players have scored 90 runs in cricket, then the score of 90 is said to have a frequency of 4. The frequency of a data value is often represented by 'f'.

A **frequency table** is constructed by arranging the collected data values in either ascending order of magnitude or descending order with their corresponding frequencies.

Marks	Tally	Frequency
1	///	3
2	///	3
3	//	2
4	//	2
5	//	2
6	~~////~~	5
7	////	4
8	~~////~~	5
9	//	2
10	//	2
Total	**30**	**30**

Example 1: The following data shows the test marks obtained by a group of students. Draw a frequency table for the data.

6	7	7	1	3	2	8	6	8	2
4	4	9	10	2	6	3	1	6	6
9	8	7	5	7	10	8	1	5	8

Go through the data; make a stroke in the tally column for each occurrence of the data. The number of strokes will be the frequency of the data.

When the set of data values are spread out, it is difficult to set up a frequency table for every data value as there will be too many rows in the table. So, we group the data into class intervals (or groups) to help us organize, interpret and analyse the data.

Example 2: The number of calls from motorists per day for roadside service was recorded for a particular month in a year. The results were as follows:

28	122	217	130	120	86	80	90	120	140
70	40	145	187	113	90	68	174	194	170
100	75	104	97	75	123	100	82	109	120
81									

Set up a frequency table for this set of data values.

Hint: Make your frequency table something like:

Class interval	Tally	Frequency
0 - 39		
40 - 79		
80 - 119		
120 - 159		
160 - 199		
200 - 239		

3.3 Graphs and Shapes of Distributions

Statisticians or machine learning engineers often want to summarize the data they have. They can do it by various available methods like data matrix, frequency tables or by graphical representation. When graphed, the data in a set is arranged to show how the points are distributed throughout the set. These distributions show the spread (dispersion, variability, scatter) of the data. The spread may be stretched (covering a wider range) or squeezed (covering a narrower range).

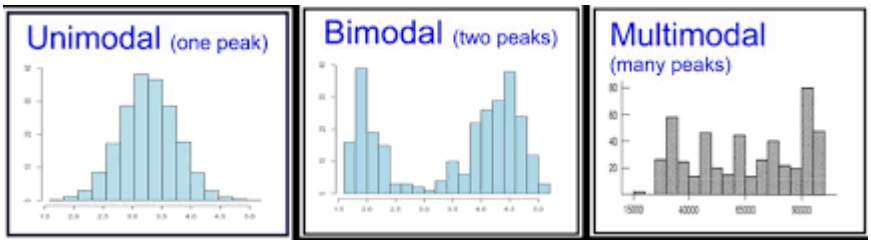
We have already studied about the graphs and various graphical representation methods in the previous chapters. Here we will learn about the spread of data and meaning of the spread i.e. distribution. The shape of a distribution is described by its number of peaks and by its possession of symmetry, its tendency to skew, or its uniformity. (Distributions that are skewed have more points plotted on one side of the graph than on the other.) The shape of the data distribution represents:

- Spread of data i.e. scatter, variability, variance etc
- Where the central tendency (i.e. mean) lies in the data spread
- What the range of data set is

Shapes of distribution are defined by different factors such as:

1. Number of Peaks

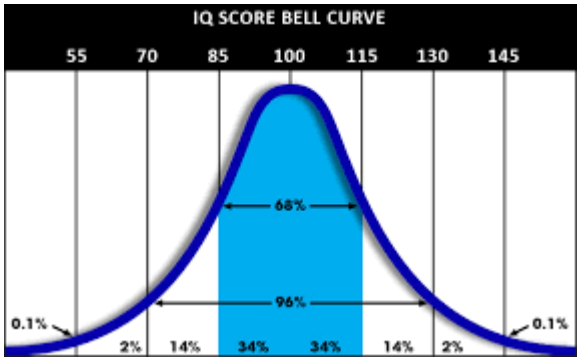
Distribution can have single peak (unimodal), also called modes, two peaks (bimodal) or (multimodal).



http://www.lynnschools.org/classrooms/english/faculty/documents

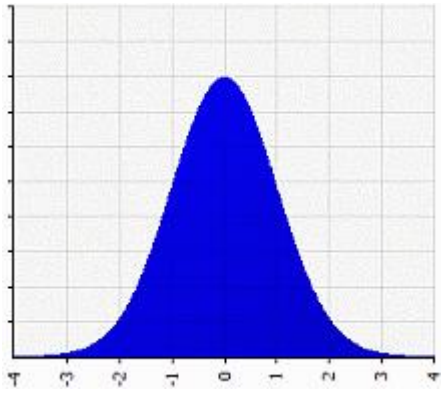
One of the most common types of unimodal distribution is normal distribution of 'bell curve' because its shape looks like bell.

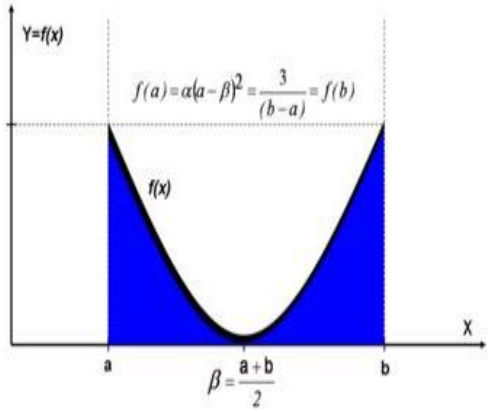
Unimodal

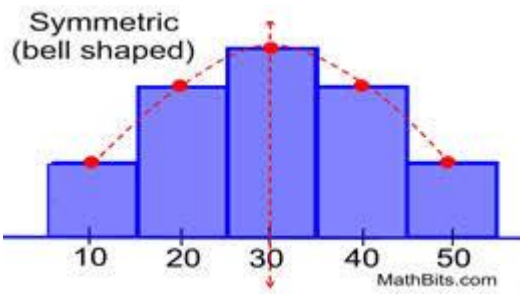


2. Symmetry

A symmetric graph when graphed and a vertical line drawn at the centre forms mirror images, with the left half of the graph being the mirror image of the right half of the graph. The normal distribution or U-distribution is an example of symmetric graphs.



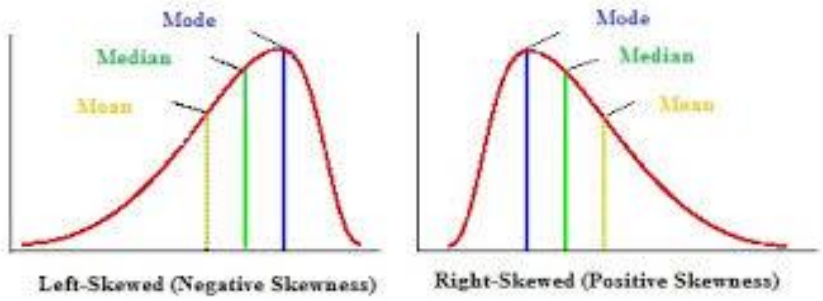




https://mathbitsnotebook.com/Algebra1/StatisticsData/STShapes.html

3. Skewness

Unsymmetrical distributions are usually skewed. They have pointed plot on one side of mean. This causes a long tail either in the negative direction on the number line (left skew) or long tail on the positive direction on the number line (positive skew or right skew).

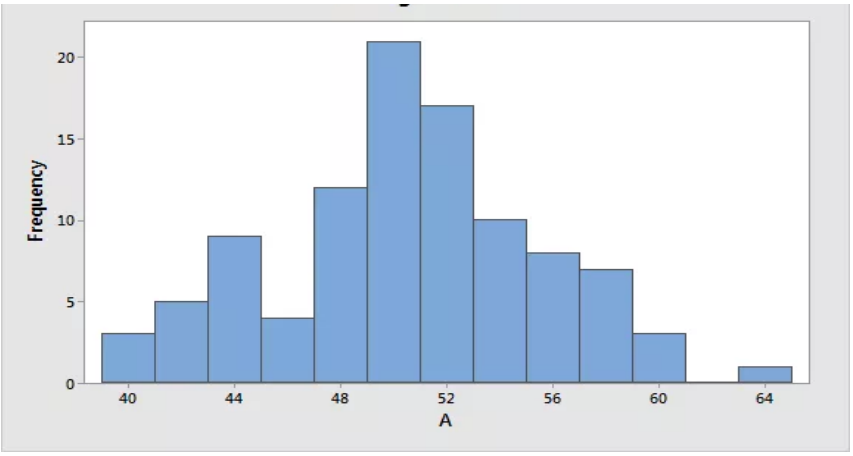


Now let us look at
some cases to check our understanding of the above

Let us further describe these shapes of distribution using histogram, one of the simplest but widely used data representation methods.

Histograms are a very common method of visualizing data, and that means that understanding how to interpret histograms is a valuable and important skill in statistics and machine learning.

Case – 1

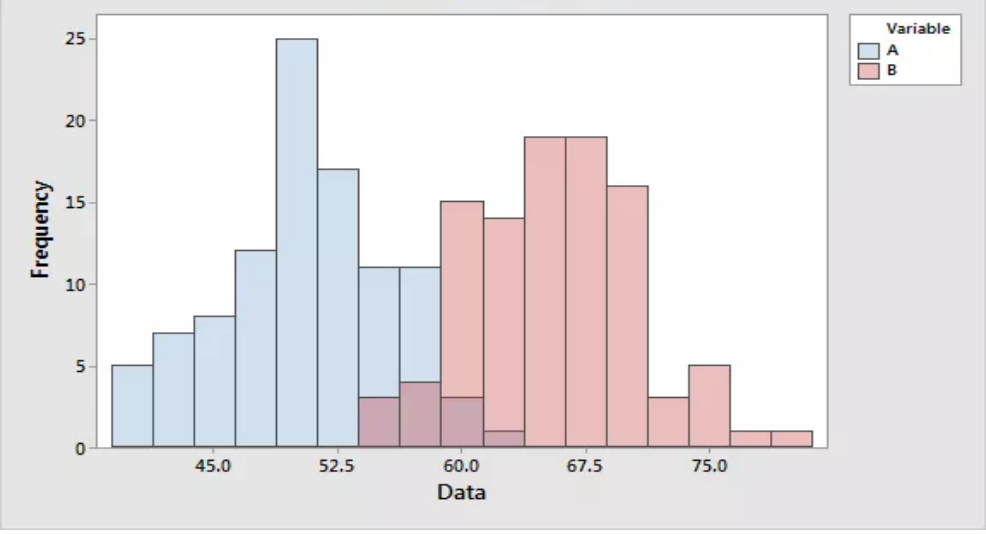


In this histogram, you can see that mean is close to 50. The shape of the graph is roughly symmetric and the values fall between 40 to 64. In some sense, value 64, looks like outlier.

Question -1: What type of distribution do you see in this graph?

_____

Case -2



This histogram has 2 means which suggests that this histogram represents two cases. One group has mean 50 and other group has a mean of 65.

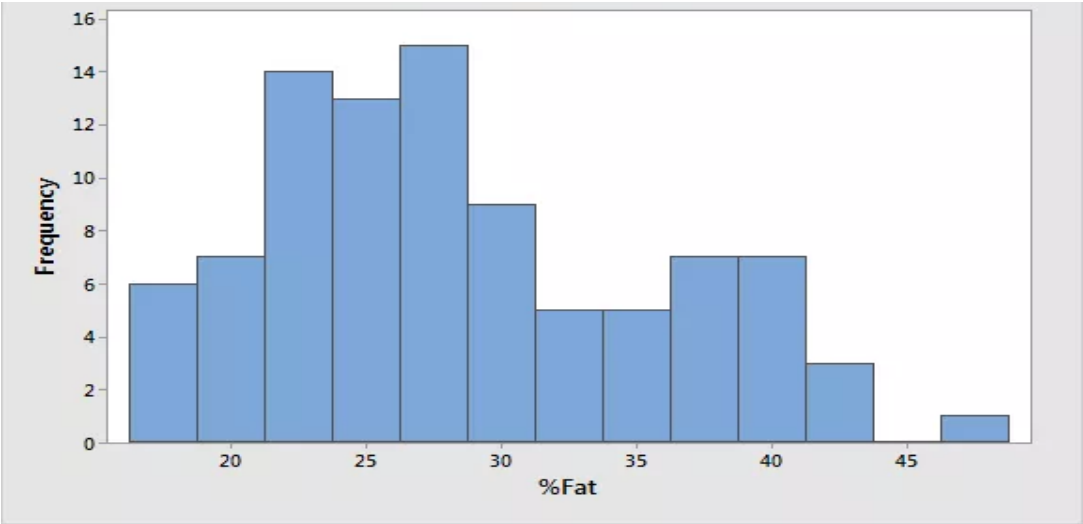
Can you think of a real-life example for such a case with two means?

Let us assume, that for the sports meet being held in your school your parents along with your grandparents were also invited. The age of the parents ranged between 40 – 55 years (in blue colour), and that of grandparents in the range of 60 – 80 years (in pink colour). During the break, both these cases (parents and grandparents) bought snacks from the snacks counter. Y – Axis shows the money they spent at the counter in buying the snacks.

Activity 1

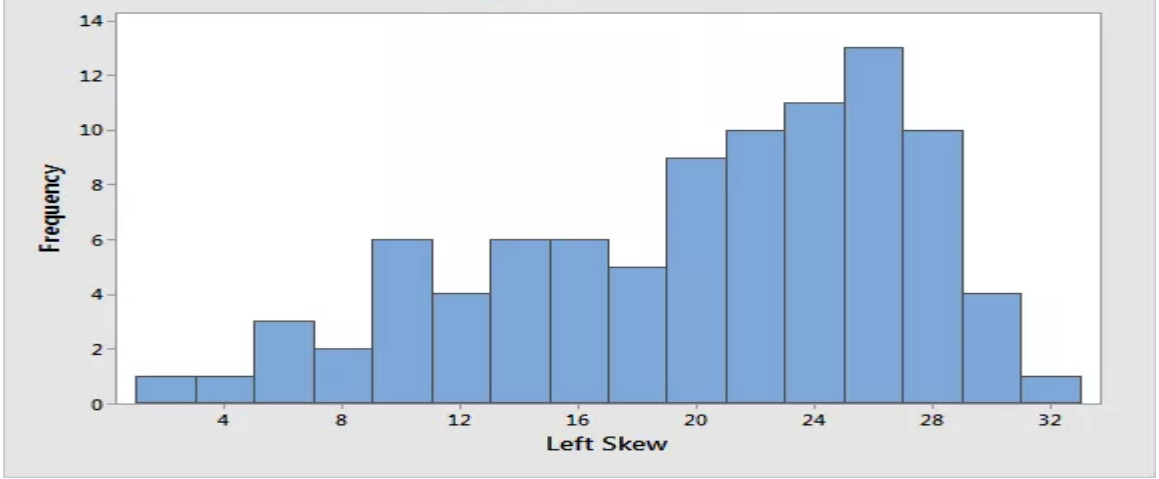
Can you think of an event(s) where you have two cases, in your classroom environment? Capture the data and plot on the histogram to have two peaks. Once done, tell your data story to the class.

Case – 3



Case 3 represents the right-skewed distribution (the direction of the skew indicates which way the longer tail extends), the long tail extends to the right while most values cluster on the left, as shown in the histogram above.

Case – 4



In **Case 4** which is a left skewed distribution, long tail extends to the left while most values cluster on the right.

3.4 Mean, Median and Mode

We have already covered the measures of central tendency (mean, median and mode) in the Level 1. In this level, we will do a quick recap citing examples from our life.

3.4.1 Mean

Think of a situation when your final results have been declared in school and you reach home with your report card. Your parents will enquire about your scores and your overall performance - "What is your average score?" In fact, they are trying to find your **MEAN** score.

The mean (or average) is the most popular and well-known measure of central tendency. The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. Therefore, in this case, mean is the sum of the total marks you scored in all the subjects divided by the number of subjects.

$M = \sum fox / n$

Where M = Mean

∑ = Sum total of the scores

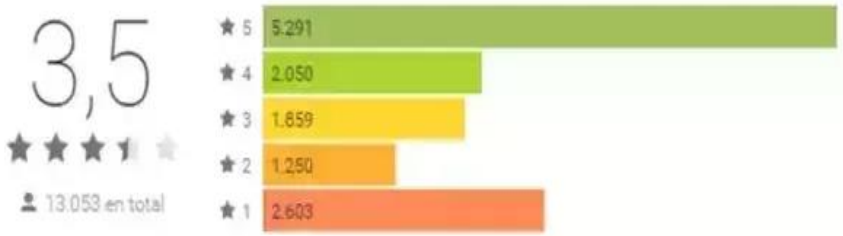
f = Frequency of the distribution

x = Scores

n = Total number of cases

Activity 1: When you try to search a game app on play store, you must be looking at the rating of the app. Can you figure out how that rating is calculated?

Activity 2: In the example below, if we add all rating given by users, it will come to 45685. Then how come rating is 3.5?



3.4.2 Median

Suppose, there are 11 students in your class and one of them is from a very wealthy family.

Pocket money received by all 11 students is as follows:

Student	1	2	3	4	5	6	7	8	9	10	11
Pocket Money	600	580	650	650	550	700	590	640	600	595	20000

Upon calculating the mean or average, it would turn out that the average pocket money received by students of the class is Rs. 7973

However, on crosschecking with the amounts mentioned in the table it does not appear to be true. Is any student getting a pocket money even close to INR 7973? The answer is No! So, should we use mean to represent the data here? No! Because of one the extreme contributions received by a student from a wealthy family has upset the mean.

So, what else can we do to so that the value should represent maximum students?

The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data. So median value here is INR 700 and this value will represent maximum students.

For a grouped data, calculation of a median in continuous series involves the following steps:

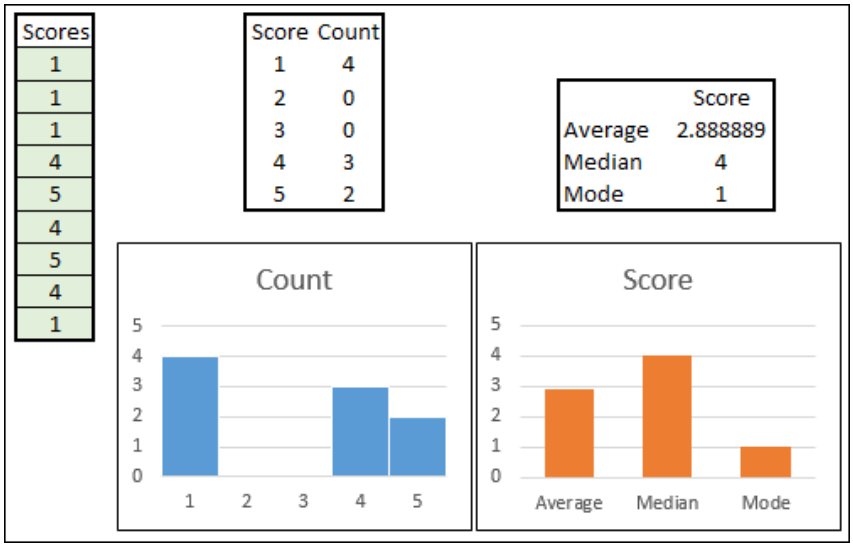
(I) The data arranged in ascending order of their class interval
(ii) Frequencies are converted into cumulative frequencies
(iii) Median class of the series is identified
(iv) Formula used to find actual median value

And the formula is: Median = $l_1 + \frac{\frac{N}{2} - c.f}{f} \times i$

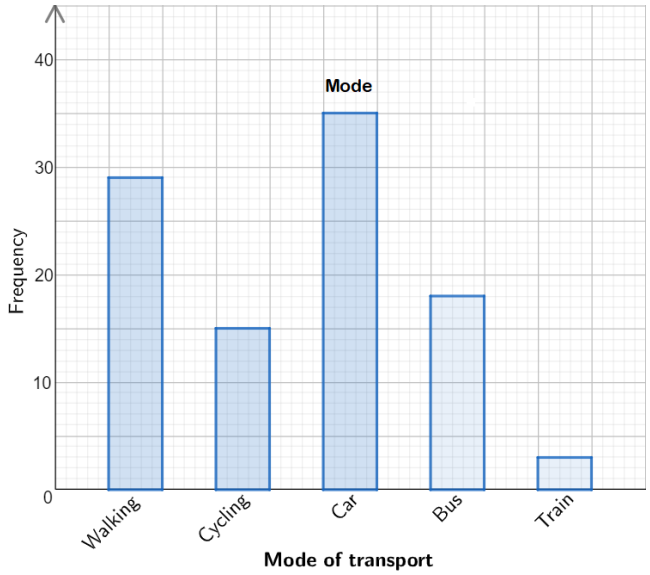
l1= Lower limit of median class
cuff= Cumulative frequency of the class preceding the median class
f= Frequency of the median class
I= Class size

3.4.3 Mode

Mode is another important measure of central tendency of tactical series. It is the value which occurs most frequently in the data series. The mode is the most frequent score in our data set. On a histogram it represents the highest bar in a bar chart or histogram. You can, therefore, sometimes consider the mode as being the most popular option.



https://contexturesblog.com/archives/2013/06/13/excel-functions-average-median-mode/



https://mathsmadeeasy.co.uk/gcse-maths-revision/bar-graphs-revision/

Let us take an example…

You need to buy a shoe. You go to the market and ask the shopkeeper 'what is the average shoe size you sell?', he will give an answer corresponding to the size that he sells maximum. That is the mode.

The arithmetic mean and median would give you figures of shoe size that don't exist and are obviously meaningless.

Now let us understand how you can use these central tendencies in Machine Learning

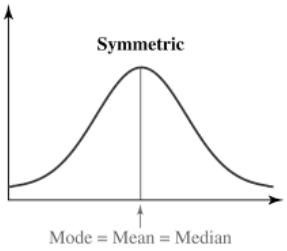
Machine Learning (or AI) is a data driven technology. If the input data is wrong, ML will produce wrong results. When working with data it is good to have an overall picture of the data. Where it's good to have an idea of how values in a given data set is distributed. The data distribution of data set can be

- Symmetric
- Skewed
  - Negatively skewed
  - Positively skewed

The skewness of data can be found either by data visualization techniques or by calculation of central tendency.

If the data is symmetrically distributed:

Mean = Median = Mode

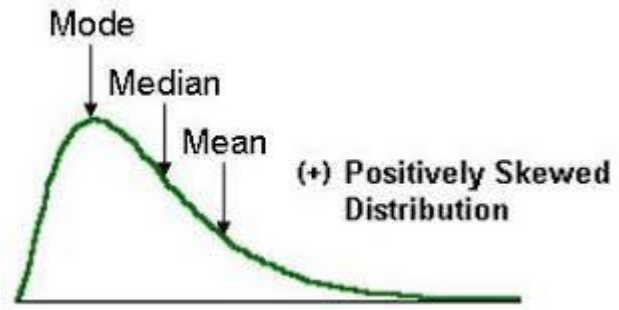


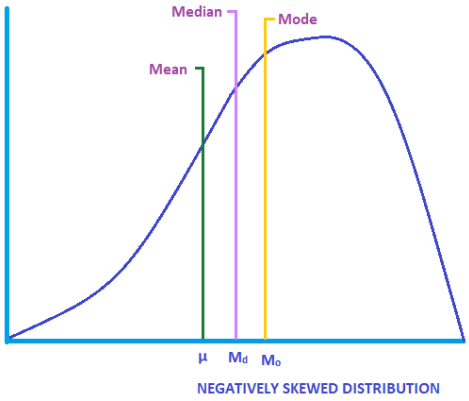
(c)

http://homepage.stat.uiowa.edu/~rdecook/stat1010/notes/Section_4.2_distribution_shapes.pdf

If the data is positively distributed

Mode<Median<Mean





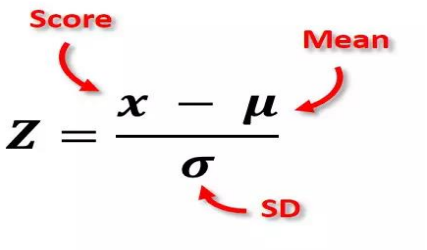
If the data is negativity distributed

Mean<Median<Mode

https://www.calculators.org/math/mean-median-mode.php

Z – score (For Advance Learners)

Z-score gives us an idea of how far our data point (in question) is from the mean. But more technically, it's a measure of how many standard deviations below or above, the data point is from the population mean. The z-score is positive if the value lies above the mean, and negative if it lies below the mean.



As the formula shows, the z-score is simply the raw score minus the sample mean, divided by the sample standard deviation.

How do we interpret a z-score?

The value of the z-score tells you how many standard deviations your data point is away from the mean. If a z-score is equal to 0, it is on the mean.

A positive z-score indicates the raw score is higher than the mean average. For example, if a z-score is equal to +1, it is 1 standard deviation above the mean.

A negative z-score reveals the raw score is below the mean average. For example, if a z-score is equal to -2, it is 2 standard deviations below the mean.

Question

There are 50 students in your class, you scored 70 out of 100 in SST exam. How well did you perform in your SST exam?

Answer

Let us re-phrase this question – In fact you need to find - "What percentage (or number) of students scored higher than you and what percentage (or number) of students scored lower than you?

This is a perfect case of z-score. To calculate z-score, you need to find the mean score of your class in SST and standard deviation.

So, let us assume, mean is 60 and standard deviation is 15.

z-score = (x – μ) / ṗ

= (70 – 60) / 15 = 10 / 15

= .6667

This means you scored .6667 standard deviations above the mean.

Practice Questions

1. Find (a) the mean (b) the median (c) the mode (d) the range of the below data set

5, 6, 2, 4, 7, 8, 3, 5, 6, 6

2. In a survey of 10 households, the number of children was found to be

4, 1, 5, 4, 3, 7, 2, 3, 4, 1

(a) State the mode

(b) Calculate

(I) the mean number of children per household

(ii) the median number of children per household.

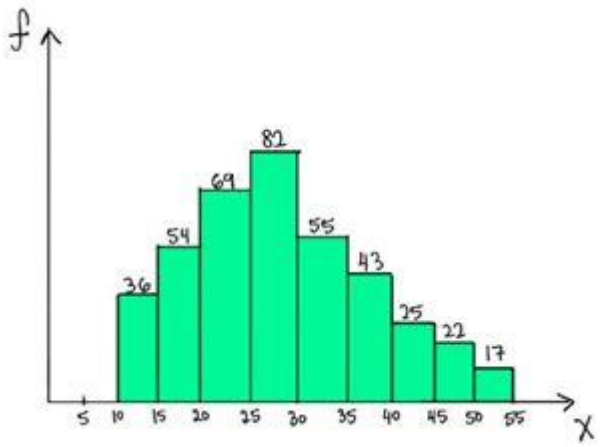
(c) A researcher says: "The mode seems to be the best average to represent the data in this survey." Give ONE reason to support this statement.

3. The mean number of students ill at a school is 3.8 per day, for the first 20 school days of a term. On the 21st day 8 students are ill. What is the mean after 21 days?

4. **If a positively skewed distribution has a median of 50, which of the following statement is true?**

A) Mean is greater than 50
B) Mean is less than 50
C) Mode is less than 50
D) Mode is greater than 50
E) Both A and C
F) Both B and D

5. **Which of the following is a possible value for the median of the below distribution?**



A) 32
B) 26
C) 17
D) 40

Unit 8

Regression

Title: Regression	Approach: Problem Solving , Discussion, Team Activity, Case studies

Summary: Artificial Intelligence / Machine Learning has become prevalent in almost every aspect of our life, society and business. People across different disciplines are trying to apply AI to be more accurate and to have better control of the future. For example, economists are using AI to predict future market prices to make a profit, doctors use AI to classify whether a tumour is malignant or benign, meteorologists use AI to predict the weather, HR recruiters use AI to check the resume of applicants to verify if the applicant meets the minimum criteria for the job, banks are using AI to check paying capacity of the customers before loan disbursement.

The AI / ML algorithm that every AI learner starts with is a linear regression (and correlation) algorithm. So let us learn the foundations of linear regression to build a solid base for the learning of AI and ML.

Linear regression is a method for modelling the relationship between one or more independent variables and a dependent variable. It is the foundation block of Machine Learning and Artificial Intelligence. It is a form of predictive modelling technique that depicts the relationship between a dependent (target) and the independent variables (predictors).

This technique is used for forecasting, time series modelling and finding the cause - effect relationship between the variables.

Objectives:

1. To understand the difference between correlation and regression.
2. To Understand the Pearson correlation coefficient (*r*) measures,
3. To understand how regression analysis is used to predict outcome.
4. To understand the main features and characteristics of the Pearson *r*.

Learning Outcomes:

1. Students should be able to estimate the correlation coefficient for a given data set
2. Students should be able to estimate the line of best fit for a given data set
3. Students should be able to determine whether a regression model is significant

Pre-requisites:

1. Students must be able to plot points on the Cartesian coordinate system
2. They should have basic understanding of statistics and central tendencies

Key Concepts: Regression, Correlation, Pearson's r

1. Regression and Correlation

Regression can be defined as a method or an algorithm in Machine Learning that models a target value based on independent predictors. It is essentially a statistical tool used in finding out the relationship between a dependent variable and an independent variable. This method comes to play in forecasting and finding out the cause and effect relationship between variables.

Regression techniques differ based on:

1. The number of independent variables
2. The type of relationship between the independent and dependent variable

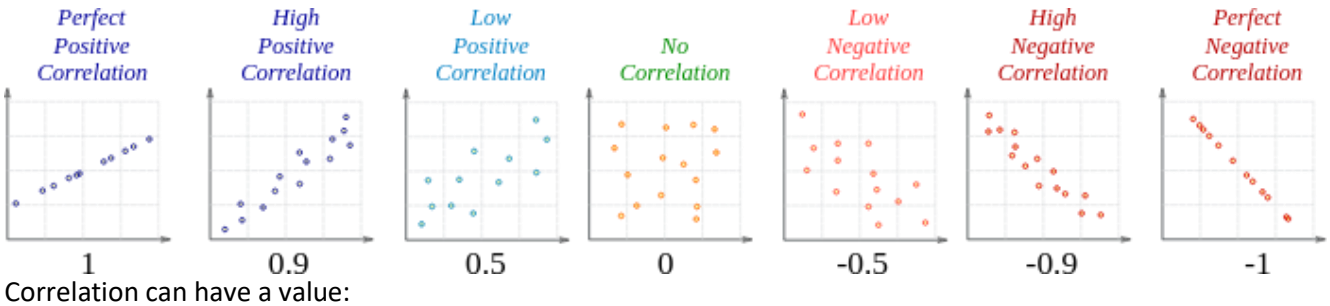
Regression is basically performed when the dependent variable is of a continuous data type. The independent variables, however, could be of any data type — continuous, nominal/categorical etc.

Regression methods find the most accurate line describing the relationship between the dependent variable and predictors with least error. In regression, the dependent variable is the function of the independent variable and the coefficient and the error term.

Correlation is a measure of the strength of a linear relationship between two quantitative variables (e.g. price, sales)

- Correlation is **positive** when the values increase together
- Correlation is **negative** when one value decreases as the other increases

A correlation is assumed to be linear i.e. following a line.



Correlation can have a value:

- 1 is a perfect positive correlation
- 0 is no correlation (the values don't seem linked at all)
- -1 is a perfect negative correlation

The value shows how good the correlation is (not how steep the line is), and if it is positive or negative.

1.1 Crosstabs and Scatterplots

1.1.1 Crosstabs

Cross tabs help us establish a relationship between two variables. This relationship is exhibited in a tabular form.

The table below is a crosstab that shows by age whether somebody has an unlisted phone number.

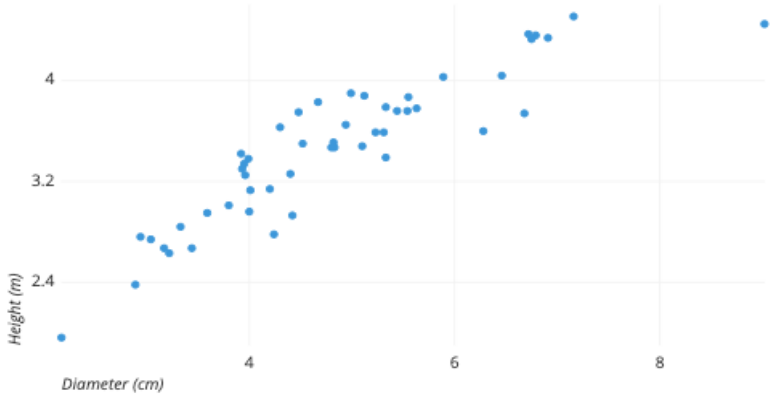
		Unlisted phone number		
Age		No	Yes	**NET**
18-34	% within column	24%	49%	**29%**
	n	185	90	**275**
35-44	% within column	20%	26%	**21%**
	n	153	48	**201**
45-54	% within column	17%	10%	**16%**
	n	133	19	**152**
55-64	% within column	17%	11%	**16%**
	n	130	21	**151**
65+	% within column	23%	3%	**19%**
	n	178	6	**184**
NET	% within column	**100%**	**100%**	**100%**
	n	**779**	**184**	**963**

- This table shows the number of observations with each combination of possible values of the two variables in each cell of the table
- We can see, for example, there are 185 people aged 18 to 34 years who do not have an unlisted phone number.
- *Column percentages* are also shown (these are percentages within the columns, so that each column's percentages add up to 100%); for example, 24% of all the people without an unlisted phone number are aged 18 to 34 years.
- The age *distribution* for people without unlisted numbers is different from that for people with unlisted numbers. In other words, the crosstab reveals a relationship between the two: people with unlisted phone numbers are more likely to be younger.
- Thus, we can also say that the variables used to create this table are correlated. If there were no relationship between these two categorical variables, we would say that they were not correlated.

In this example, the two variables can both be viewed as being ordered. Consequently, we can potentially describe the patterns as being positive or negative correlations (negative in the table shown). However, where both variables are not ordered, we can simply refer to the strength of the correlation without discussing its *direction* (i.e., whether it is positive or negative).

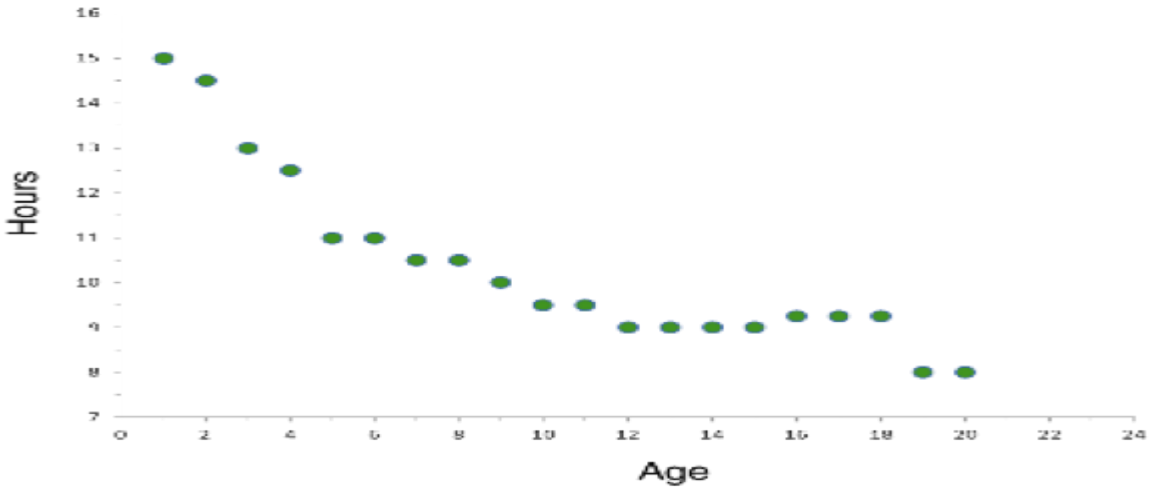
1.1.2 Scatterplots

A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.



Example

This is a scatter plot showing the amount of sleep needed per day by age.



As seen above, as you grow older, you need less sleep (but still probably more than you're currently getting).

Question: What type of correlation is shown here?

Answer: This is a negative correlation. As we move along the x-axis toward the greater numbers, the points move down which means the y-values are decreasing, making this a negative correlation.

1.2 Pearson's r

The Pearson correlation coefficient is used to measure the strength of a linear association between two variables, where the value r = 1 means a perfect positive correlation and the value r = -1 means a perfect negative correlation. So, for example, you could use this test to find out whether people's height and weight are correlated (the taller the people are, the heavier they're likely to be).

Requirements for Pearson's correlation coefficient are as follows:Scale of measurement should be interval or ratio

- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

Equation

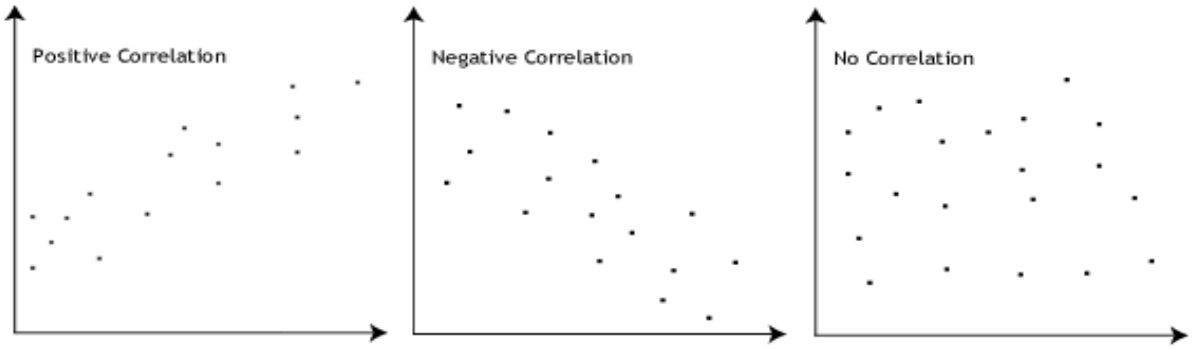
$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2} \sqrt{\sum_i (y_i - \overline{y})^2}}$$

What does this test do?

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by '*r*'. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

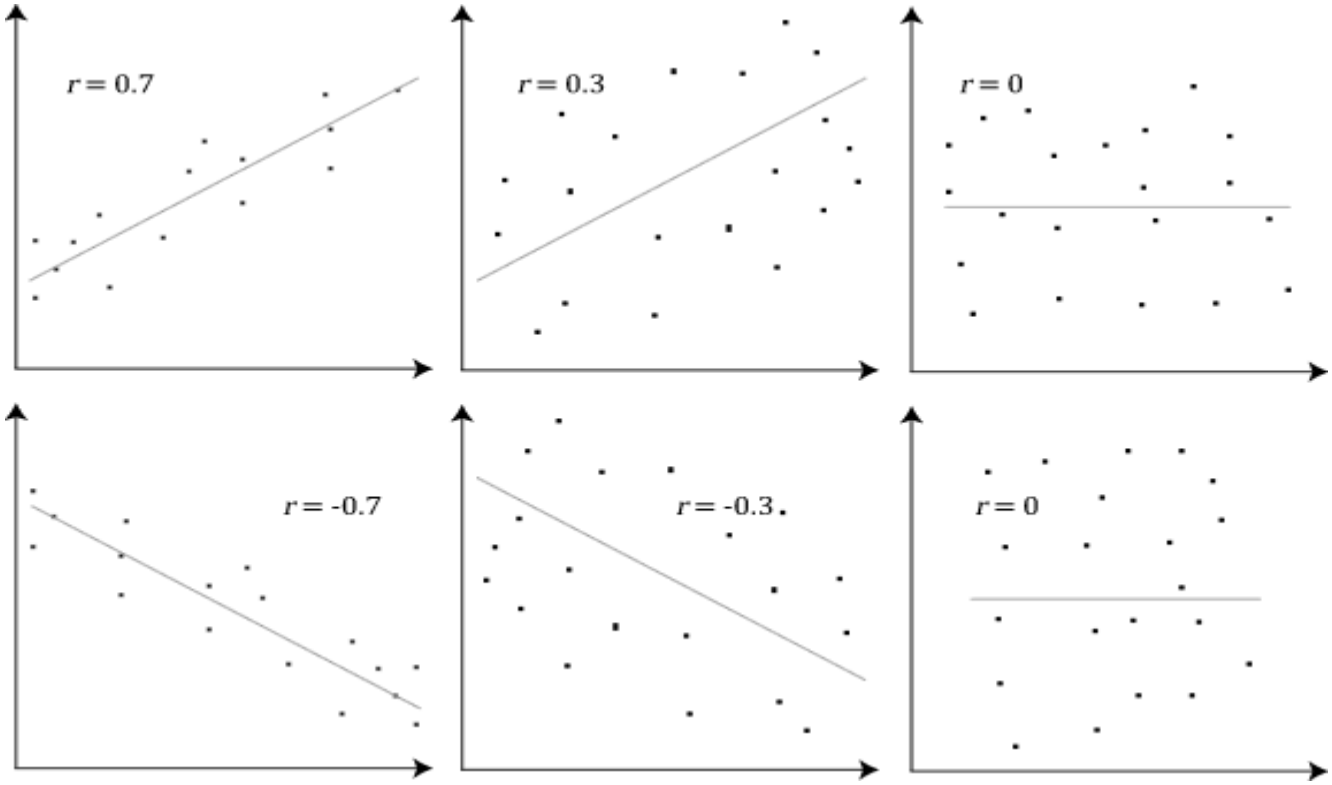
What values can the Pearson correlation coefficient take?

The Pearson correlation coefficient, *r*, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



How can we determine the strength of association based on the Pearson correlation coefficient?

The stronger the association of the two variables, the closer the Pearson correlation coefficient, *r*, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line. Values for *r* between +1 and -1 (for example, *r* = 0.8 or -0.4) indicate that there is variation around the line of best fit. The closer the value of *r* to 0 the greater the variation around the line of best fit. Different relationships and their correlation coefficients are shown in the diagram below:



Are there guidelines to interpreting the Pearson's correlation coefficient?

Yes, the following guidelines have been proposed:

	Coefficient, *r*	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to -1.0

Remember that these values are guidelines and whether an association is strong or not will also depend on what you are measuring.

Example 1

In the example below of 6 people with different age and different weight, let us try calculating the value of the Pearson r.

Sr. No	Age (x)	Weight (y)
1	40	78
2	21	70
3	25	60
4	31	55
5	38	80
6	47	66

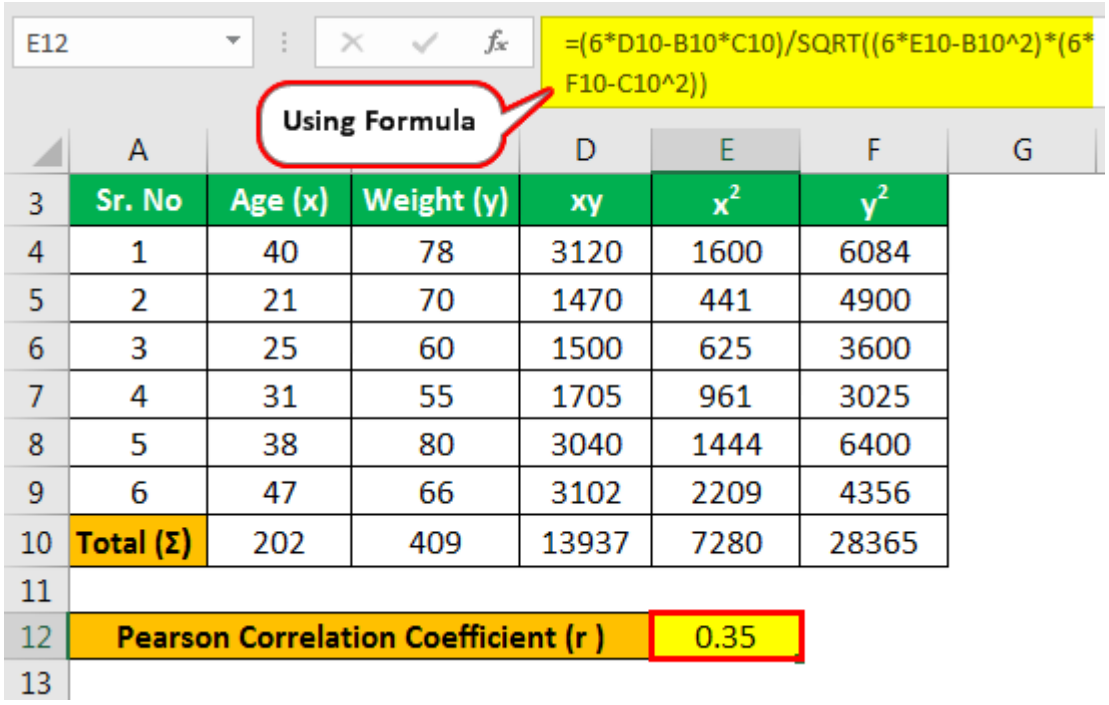
Solution:

For the Calculation of the Pearson Correlation Coefficient, we will first calculate the following values:

Sr. No	Age (x)	Weight (y)	xy	x^2	y^2
1	40	78	3120	1600	6084
2	21	70	1470	441	4900
3	25	60	1500	625	3600
4	31	55	1705	961	3025
5	38	80	3040	1444	6400
6	47	66	3102	2209	4356
Total (Σ)	202	409	13937	7280	28365

Here the total number of people is 6 so, **n=6**

Now the calculation of the Pearson R is as follows:



	A	B	C	D	E	F	G
3	Sr. No	Age (x)	Weight (y)	xy	x^2	y^2	
4	1	40	78	3120	1600	6084	
5	2	21	70	1470	441	4900	
6	3	25	60	1500	625	3600	
7	4	31	55	1705	961	3025	
8	5	38	80	3040	1444	6400	
9	6	47	66	3102	2209	4356	
10	Total (Σ)	202	409	13937	7280	28365	
11							
12	Pearson Correlation Coefficient (r)				0.35		
13							

- r = (n (∑xy)- (∑x)(∑y))/(√ [n ∑x2-(∑x)2][n ∑y2– (∑y)2)
- r = (6 * (13937)- (202)(409)) / (√ [6 *7280 -(202)2] * [6 * 28365- (409)2)
- r = (6 * (13937)- (202) * (409))/(√ [6 *7280 -(202)2] * [6 * 28365- (409)2)
- r = (83622- 82618)/(√ [43680 -40804] * [170190- 167281)
- r = 1004/(√ [2876] * [2909)
- r = 1004 / (√ 8366284)
- r = 1004 / 2892.452938
- **r = 0.35**

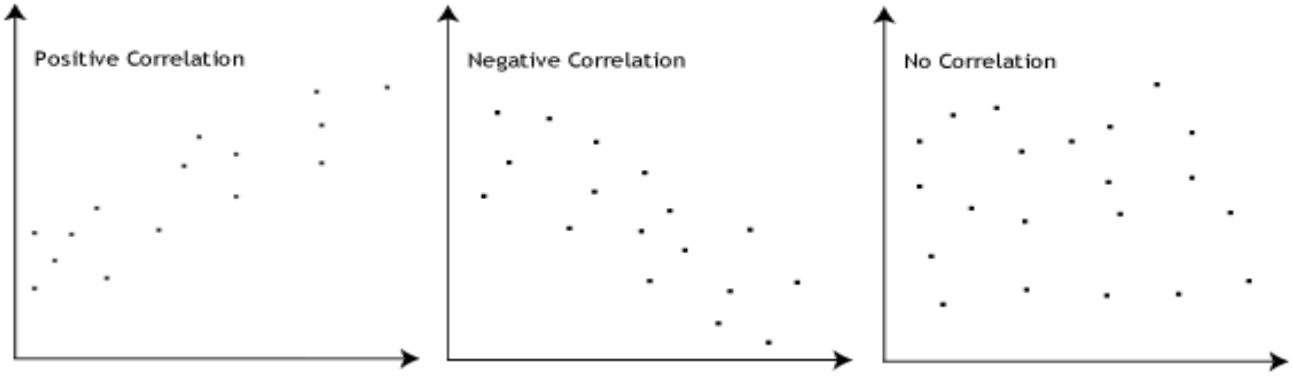
Thus the value of the Pearson correlation coefficient is **0.35**

Assumptions

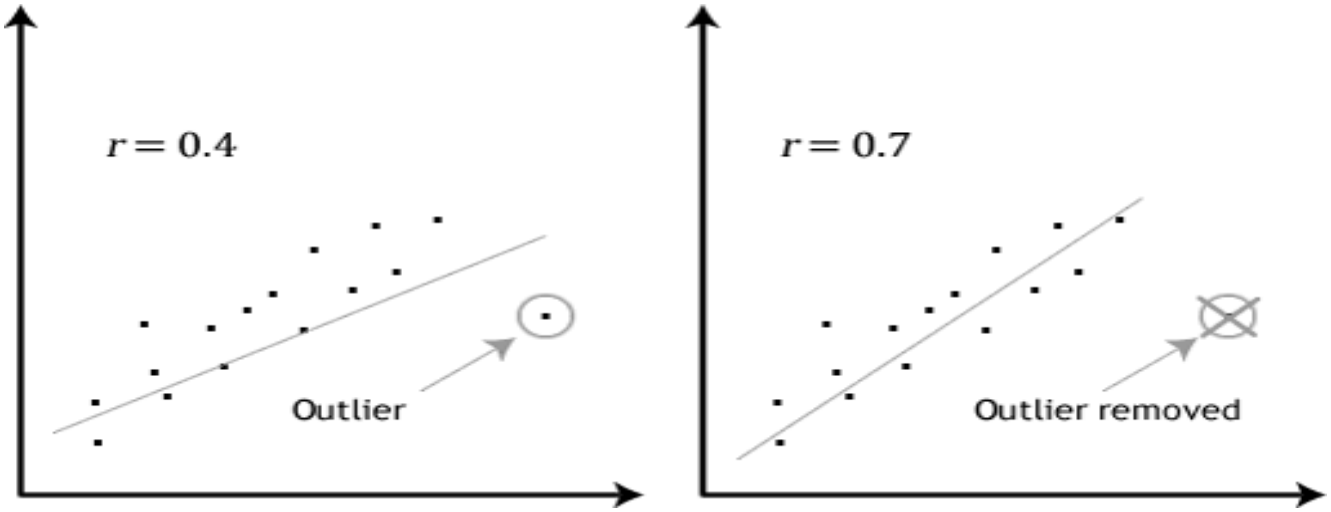
There are four "assumptions" that underpin a Pearson's correlation. If any of these four assumptions are not met, analysing your data using a Pearson's correlation might not lead to a valid result.

Assumption # 1: The two variables should be measured at the continuous level. Examples of such continuous variables include height (measured in feet and inches), temperature (measured in °C), salary (measured in dollars/INR), revision time (measured in hours), intelligence (measured using IQ score), reaction time (measured in milliseconds), test performance (measured from 0 to 100), sales (measured in number of transactions per month), and so forth.

Assumption # 2: There needs to be a linear relationship between your two variables. Whilst there are a number of ways to check whether a Pearson's correlation exists, we suggest creating a scatterplot using Stata, where you can plot your two variables against each other. You can then visually inspect the scatterplot to check for linearity. Your scatterplot may look something like one of the following:



Assumption #3: There should be no significant outliers. Outliers are simply single data points within your data that do not follow the usual pattern (e.g. in a study of 100 students' IQ scores, where the mean score was 108 with only a small variation between students, one student had a score of 156, which is very unusual, and may even put her in the top 1% of IQ scores globally). The following scatterplots highlight the potential impact of outliers:



Pearson's *r* is sensitive to outliers, which can have a great impact on the line of best fit and the Pearson correlation coefficient, leading to very difficult conclusions regarding your data. Therefore, it is best if there are no outliers or they are kept to a minimum. Fortunately, you can use Stata to detect possible outliers using scatterplots.

Assumption # 4: Your variables should be approximately normally distributed. In order to assess the statistical significance of the Pearson correlation, you need to have bivariate normality, but this assumption is difficult to assess, so a simpler method is more commonly used.

1.3 Regression – Finding The line

When we make a distribution in which there is an involvement of more than one variable, then such an analysis is called Regression Analysis. It generally focuses on finding or rather predicting the value of the variable that is dependent on the other.

Let there be two variables *x and y*. If *y* depends on *x,* then the result comes in the form of a simple regression. Furthermore, we name the variables *x* and *y* as:

y – Regression or Dependent Variable or Explained Variable
x – Independent Variable or Predictor or Explanator

Therefore, if we use a simple linear regression model where *y* depends on *x,* then the regression line of *y* on *x* is:

y = a + bx

Regression Coefficient

The two constants *a* and *b* are regression parameters. Furthermore, we denote the variable *b* as b_{yx} and we term it as *regression coefficient* of *y* on *x.*

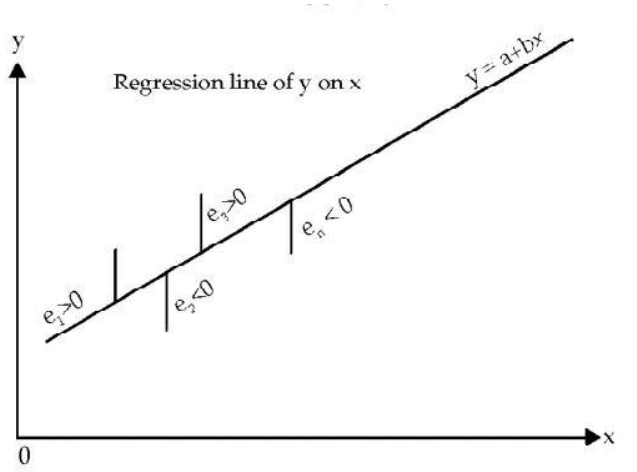
Also, we can have one more definition for the regression line of *y* on *x.* We can call it the best fit as the result comes from least squares. This method is the most suitable for finding the value of *y* on *x* i.e. the value of a dependent variable on an independent variable.

Least Squares Method

∑ ei2 = ∑ (yi – y ^ i)2 = ∑ (yi – a – bxi)2

Here:

- Variable *yi* is the actual value or the observed value
- y ^ i = a + bxi, denotes the estimated value of *yi* for a given random value of a variable of xi
- ei = Difference between observed and estimated value and is the error or residue. The regression line of y or x along with the estimation errors are as follows:



On minimizing the least squares equation, here is what we get. We refer to these equations Normal Equations.

∑yi = na + b ∑xi
∑xiyi = a ∑xi2 + b ∑xi

We get the *least squares* estimate *for a* and *b* by solving the above two equations for both *a* and *b.*

b = Cov(x,y)/Sx2
= (r.SxSy)/Sx2
= (r.Sy)/Sx

The estimate of *a,* after the estimation of *b* is:

$a = \bar{y} - b\bar{x}$

On substituting the estimates of *a* and *b is:*

$[y - \bar{y}]/Sy = r[x - \bar{x}]/Sx$

Sometimes, it might so happen that variable *x* depends on variable *y.* In such cases, the line of regression of x on y is:

x = a ^ + b^y

Regression Equation

The standard form of the regression equation of variable *x* on *y* is:

$[x - \bar{x}]/Sx = r[y - \bar{y}]/Sy$

Question

The regression equation for variables x and y are 7x – 3y – 18 = 0 and 4x – y – 11 = 0.

1.What is the AM for x and y?

2.Find the correlation coefficient in between x and y.

Solution

(i) The intersection of two lines have the same intersection point and that is [x^-, y^-]. Therefore, we replace, x and y with x^- and y^-

7x – 3y = 18
4x – y = 11

Hence, on solving these two equations we get x^- = 3 and y^- = 1.

(ii) We know,

r2 = 7/12

Therefore,

r = 712––√ (r is positive as both the coefficients are positive)
= 0.7638

1.4 Regression – Describing the line

Definition: In statistics, a regression line is a line that best describes the behaviour of a set of data. In other words, it's a line that best fits the trend of a given data.

What Does Regression Line Mean?

Regression lines are very useful for forecasting procedures. The purpose of the line is to describe the interrelation of a dependent variable (Y variable) with one or many independent variables (X variable). By using the equation obtained from the regression line an analyst can forecast future behaviours of the dependent variable by inputting different values for the independent ones. Regression lines are widely used in the financial sector and in business in general.

Financial analysts employ linear regressions to forecast stock prices, commodity prices and to perform valuations for many different securities. On the other hand, companies employ regressions for the purpose of forecasting sales, inventories and many other variables that are crucial for strategy and planning.

The regression line formula is like the following:

(Y = a + bX + u)

The multiple regression formula looks like this:

(Y = a + b1X1 + b2X2 + b3X3 + … + btXt +u.)

- Y is the dependent variable
- X is the independent ones
- a is the interception point
- b is the slope
- u is the residual regression

Example: 1

Data was collected on the "depth of dive" and the "duration of dive" of penguins. The following linear model is a fairly good summary of the data:

Where:

- *t* is the duration of the dive in minutes
- *d* is the depth of the dive in yards

The equation for the model is: **d t = + 0.015 2.915**

Interpretation of the slope: If the duration of the dive increases by 1 minute, we predict the depth of the dive will increase by approximately 2.915 yards.

Interpretation of the intercept: If the duration of the dive is 0 seconds, then we predict the depth of the dive is 0.015 yards.

Comments: The interpretation of the intercept doesn't make sense in the real world. It isn't reasonable for the duration of a dive to be near t = 0, because that's too short for a dive. If data with x-values near zero wouldn't make sense, then usually the interpretation of the intercept won't seem realistic in the real world. It is, however, acceptable (even required) to interpret this as a coefficient in the model.

Example: 2

Reinforced concrete buildings have steel frames. One of the main factors affecting the durability of these buildings is carbonation of the concrete (caused by a chemical reaction that changes the pH of the concrete), which then corrodes the steel reinforcing the building.

Data is collected on specimens of the core taken from such buildings, where the following are measured:

- Depth of the carbonation (in mm) is called *d*
- Strength of the concrete (in Mpa) is called *s*

It is found that the model is **s = 24.5 – 2.8.d**

Interpretation of the slope: If the depth of the carbonation increases by 1 mm, then the model predicts that the strength of the concrete will decrease by approximately 2.8 Mpa.

Interpretation of the intercept: If the depth of the carbonation is 0, then the model predicts that the strength of the concrete is approximately 24.5 Mpa.

Comments: Notice that it isn't necessary to fully understand the units in which the variables are measured in order to correctly interpret these coefficients. While it is good to understand data thoroughly, it is also important to understand the structure of linear models. In this model, notice that the strength decreases as the carbonation increases, which is shown by the negative slope coefficient. When you interpret a negative slope, notice that you must say that, as the explanatory variable increases, then the response variable decreases.

Example: 3

When cigarettes are burned, one by-product in the smoke is carbon monoxide. Data is collected to determine whether the carbon monoxide emission can be predicted by the nicotine level of the cigarette.

- It is determined that the relationship is approximately linear when we predict carbon monoxide, *C*, from the nicotine level, *N*
- Both variables are measured in milligrams
- The formula for the model is **C = 3.0 + 10.3.N**

Interpretation of the slope: If the amount of nicotine goes up by 1 mg, then we predict the amount of carbon monoxide in the smoke will increase by 10.3 mg.

Interpretation of the intercept: If the amount of nicotine is zero, then we predict that the amount of carbon monoxide in the smoke will be about 3.0 mg.

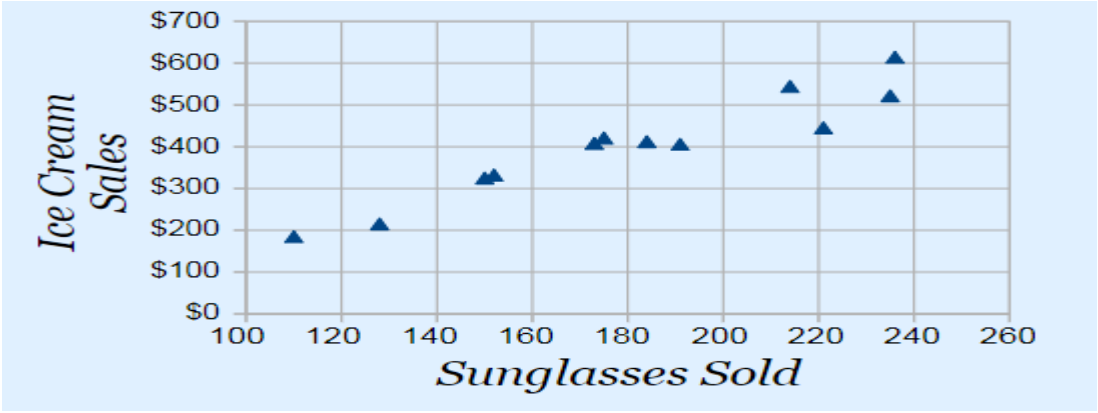
1.4 Correlation is not Causation

Correlation and causation are terms which are mostly misunderstood and often used interchangeably. Understanding both the statistical terms is very important not only to make conclusions but more importantly, making correct conclusions at the end. In this section we will understand why correlation does not imply causation.

Correlation is a statistical technique which tells us how strongly the pair of variables are linearly related and change together. It does not tell us why and how behind the relationship but it just says the relationship exists.

Example: Correlation between Ice cream sales and sunglasses sold.

As the sales of ice creams is increasing so do the sales of sunglasses.



Causation takes a step further than correlation. It says any change in the value of one variable will **cause** a change in the value of another variable, which means one variable makes the other happen. It is also referred to as cause and effect.

Two or more variables considered to be related, in a statistical context, if their values change so that as the value of one variable increases or decreases so does the value of the other variable (it may be in the same or opposite direction).

For example,

- For the two variables "hours worked" and "income earned" there is a relationship between the two such that the increase in hours worked is associated with an increase in income earned as well.
- If we consider the two variables "price" and "purchasing power", as the price of goods increases a person's ability to buy these goods decreases (assuming a constant income).

Therefore:

- Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables.
- A correlation between variables, however, does not automatically mean that the change in one variable is the cause of change in the values of the other variable.
- Causation indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events. This is also referred to as cause and effect.

Theoretically, the difference between the two types of relationships are easy to identify — an action or occurrence can cause another (e.g. smoking causes an increase in the risk of developing lung cancer), or it can correlate with another (e.g. smoking is correlated with alcoholism, but it does not cause alcoholism). In practice, however, it remains difficult to clearly establish cause and effect, compared to establishing correlation.

1.6 Contingency Tables – Examples

A **contingency table** provides a way of portraying data that can facilitate calculating probabilities. The table helps in determining conditional probabilities quite easily. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. Later on, we will use contingency tables again, but in another manner.

Example 1

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

Cell Phone User	Speeding violation in the last year	No speeding violation in the last year	Total
Yes	25	280	305
No	45	405	450
Total	**70**	**685**	**755**

The total number of people in the sample is 755. The row totals are 305 and 450. The column totals are 70 and 685. Notice that 305 + 450 = 755 and 70 + 685 = 755.

Calculate the following probabilities using the table:

1. Find P (Person is a cell phone user)

   Number of cell phone users / Total number in study = ***305 / 755***

2. Find P (person had no violation in the last year)

   Number of no violations / Total number in study = ***685 / 755***

3. Find *P* (Person had no violation in the last year AND was a cell phone user)

Number of cell phone users with no violation / Total number in study = ***280/755***

4. Find *P* (Person is a cell phone user OR person had no violation in the last year)

(305 / 755 + 685 / 755) – 280 / 755= 710 / 755

Example 2

This table shows a random sample of 100 hikers and the areas of hiking they prefer.

Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	___	45
Male	___	___	14	55
Total	___	41	___	___

1. Complete the above table

Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	11	45
Male	**16**	**25**	14	55
Total	**34**	41	**25**	**100**

2. Find the probability that a person is male given that the person prefers hiking near lakes and streams

Hint:

Let *M* = being male, and let *L* = prefers hiking near lakes and streams.

1. What word tells you this is conditional?
2. Fill in the blanks and calculate the probability: *P*(___|___) = ___.
3. Is the sample space for this problem all 100 hikers? If not, what is it?

Answer

1. The word *"given"* tells you that this is a conditional.
2. P(M|L) =2541
3. No, the sample space for this problem is the 41 hikers who prefer lakes and streams.

2. Reading

2.1 Correlation

Correlation is a measure of how closely two variables move together. Pearson's correlation coefficient is a common measure of correlation, and it ranges from +1 for two variables that are perfectly in sync with each other, to 0 when they have no correlation, to -1 when the two variables are moving opposite to each other.

For linear regression, one way of calculating the slope of the regression line uses Pearson's correlation, so it is worth understanding what correlation is.

The equation for a line is

Y = a + bx

here a = intercept and b = the slope.

How to Find the Correlation?

The correlation coefficient that indicates the strength of the relationship between two variables can be found using the following formula:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

r_{xy} – the correlation coefficient of the linear relationship between the variables x and y

x_i – the values of the x-variable in a sample

$\bar{x}$ – the mean of the values of the x-variable

y_i – the values of the y-variable in a sample

$\bar{y}$ – the mean of the values of the y-variable

In order to calculate the correlation coefficient using the formula above, you must undertake the following steps:

1. Obtain a data sample with the values of x-variable and y-variable.
2. Calculate the means (averages) $\bar{x}$ *for* the x-variable and $\bar{y}$ for the y-variable.
3. For the x-variable, subtract the mean from each value of the x-variable (let's call this new variable "a"). Do the same for the y-variable (let's call this variable "b").
4. Multiply each a-value by the corresponding b-value and find the sum of these multiplications (the final value is the numerator in the formula).
5. Square each a-value and calculate the sum of the result
6. Find the square root of the value obtained in the previous step (this is the denominator in the formula).
7. Divide the value obtained in *step 4* by the value obtained in *step 7*.

You can see that the manual calculation of the correlation coefficient is an extremely tedious process, especially if the data sample is large. However, there are many software tools that can help you save time when calculating the coefficient. 'CORREL' function of MS Excel returns the correlation coefficient of two cell range.

Example of Correlation

- X is an investor; he invests money in share market. His portfolio primarily tracks the performance of the S&P 500 (this is a stock market index in USA that measures the performance of top 500 large companies in the USA).
- X wants to add the stock of Apple Inc. Before adding Apple to his portfolio, he wants to assess the correlation between the stock and the S&P 500 to ensure that adding the stock won't increase the systematic risk of his portfolio.
- To find the coefficient, X gathers the following prices from the last five years (**Step 1**)

	S&P 500	Apple
2013	1691.75	68.96
2014	1977.80	100.11
2015	1884.09	109.06
2016	2151.13	112.18
2017	2519.36	154.12

Using the formula above, X can determine the correlation between the prices of the S&P 500 Index and Apple Inc.

- Next, X calculates the average prices of each security for the given periods (**Step 2**):

	S&P 500	Apple
2013	1691.75	68.96
2014	1977.80	100.11
2015	1884.09	109.06
2016	2151.13	112.18
2017	2519.36	154.12
Mean	**2044.83**	**108.89**

- After the calculation of the average prices, we can find the other values. A summary of the calculations is given in the table below:

			Step 3		Step 4	Step 5	
	S&P 500	Apple	a	b	a x b	a^2	b^2
2013	1691.75	68.96	- 353.08	- 39.93	14,096.91	124,662.66	1,594.09
2014	1977.80	100.11	- 67.03	- 8.78	588.22	4,492.48	77.02
2015	1884.09	109.06	- 160.74	0.17	- 27.97	25,836.07	0.03
2016	2151.13	112.18	106.30	3.29	350.16	11,300.52	10.85
2017	2519.36	154.12	474.53	45.23	21,465.08	225,182.62	2,046.11
Mean	**2044.83**	**108.89**	**Sums**		**36,472.40**	**391,474.35**	**3,728.10**

- Using the obtained numbers, X can calculate the coefficient:

$$r_{xy} = \frac{36{,}272.40}{\sqrt{391{,}474.35 \times 3{,}728.10}} = 0.95$$

The coefficient indicates that the prices of the S&P 500 and Apple Inc. have a high positive correlation. This means that their respective prices tend to move in the same direction. Therefore, adding Apple to his portfolio would, in fact, increase the level of systematic risk.

2.2 Regression

With correlation, we determined how much two sets of numbers changed together. With regression, we will to use one set of numbers to make a prediction on the value in the other set. Correlation is part of what we need for regression. But we also need to know how much each set of numbers change individually, via the standard deviation, and where we should put the line, i.e. the intercept.

The regression that we are calculating is very similar to correlation. So you might ask, why do we have both regression and correlation? It turns out that regression and correlation give related but distinct information.

- Correlation gives you a measurement that can be interpreted independently of the scale of the two variables. Correlation is always bounded by ±1. The closer the correlation is to ±1 the closer the two variables are to a perfectly linear relationship.

- The regression slope by itself does not tell you that. The regression slope tells you the expected change in the dependent variable *y* when the independent variable *x* changes one unit. That information cannot be calculated from the correlation alone.

A fallout of those two points is that correlation is a unit-less value, while the slope of the regression line has units. If for instance, you owned a large business and were doing an analysis on the amount of revenue in each region compared to the number of salespeople in that region, you would get a unit-less result with correlation, and with regression, you would get a result that was the amount of money per person.

Regression Equations

With linear regression, we are trying to solve for the equation of a line, which is shown below.

Y = a + bx

The values that we need to solve for are 'b' the slope of the line, and 'a' the intercept of the line. The hardest part of calculating the slope 'b', is finding the correlation between x and y, which we have already done. The only modification that needs to be made to that correlation is multiplying it by the ratio of the standard deviations of x and y, which we also already calculated when finding the correlation. The equation for slope is shown below

$$b = r * \frac{s_y}{s_x}$$

Once we have the slope, getting the intercept is easy. Assuming that you are using the standard equations for correlation and standard deviation, which go through the average of x and y ($\bar{x}$,$\bar{y}$), the equation for intercept is

$$a = \bar{y} - b\bar{x}$$

Simple Linear Model for Predicting Marks

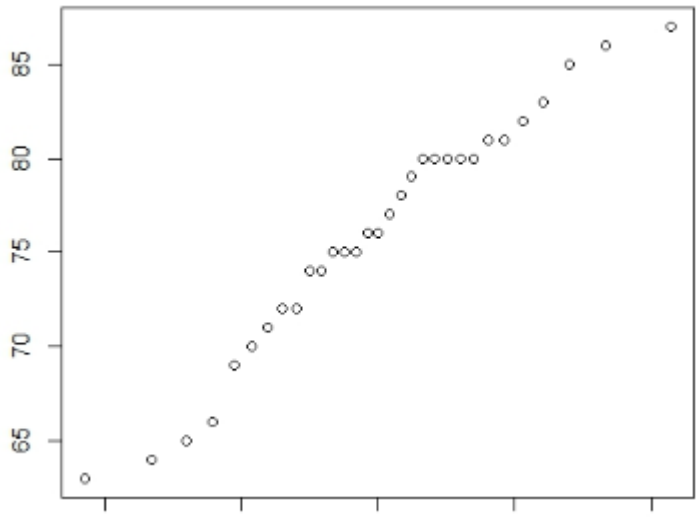
Let's consider the problem of predicting the marks of a student based on the number of hours he/she put in towards preparation. Although at the outset, it may look like a problem which can be modelled using simple linear regression, it could turn out to be a multiple linear regression problem depending on multiple input features. Alternatively, it may also turn out to be a non-linear problem. However, for the sake of example, let's consider this as a simple linear regression problem.

- Let's assume for the sake of understanding that the marks of a student (M) do depend on the number of hours (H) he/she has put in towards preparation.

The following formula can represent the model:

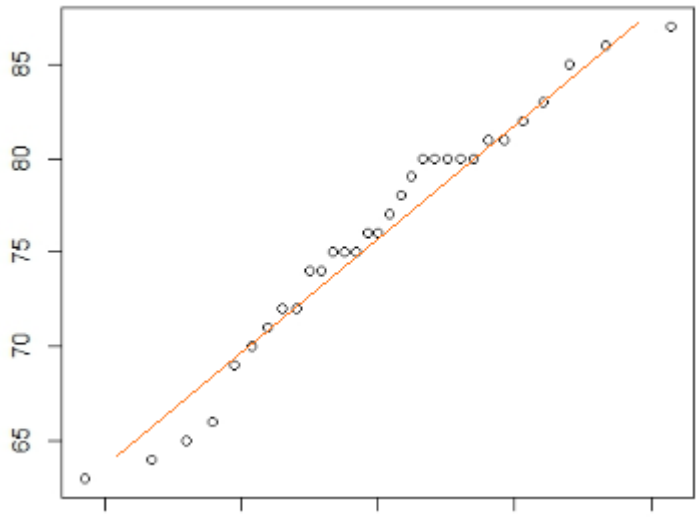
Marks = function (No. of hours)
=> Marks = m*Hours + c

The best way to determine whether it is a simple linear regression problem is to do a plot of Marks vs Hours. If the plot comes like below, it may be inferred that a linear model can be used for this problem.



Plot representing a simple linear model for predicting marks

The data represented in the above plot would be used to find out a line such as the following which represents a best-fit line. The slope of the best-fit line would be the value of "m".



Plot representing a simple linear model with a regression line

The value of m (slope of the line) can be determined using an objective function which is a combination of loss function and a regularization term. For simple linear regression, the objective function would be the summation of Mean Squared Error (MSE). MSE is the sum of squared distances between the target variable (actual marks) and the predicted values (marks calculated using the above equation). The best fit line would be obtained by minimizing the objective function (summation of mean squared error).

2.3 Practice Exercise

Problem 1

A statistics instructor at a university would like to examine the relationship (if any) between the number of optional homework problems students do during the semester and their final course grade. She randomly selects 12 students for study and asks them to keep track of the number of these problems completed during the course of the semester. At the end of the class each student's total is recorded along with their final grade. The data is available in the following table:

Final Course Grade Vs the Number of optional homework problems completed		
No. of Problems completed	**Final Course Grade**	**Problem Grade**
51	62	3162
58	68	3944
62	66	4092
65	66	4290
68	67	4556
76	72	5472
77	73	5621
78	72	5616
78	78	6084
84	73	6132
85	76	6460
91	75	6825
873	848	62254
ΣPrb	**Σgrd**	**Σprb * Grd**

1) For this setting identify the response variable

2) For this setting, identify the predictor variable

3) Compute the linear correlation coefficient – r – for this data set

4) Classify the direction and strength of the correlation

5) Test the hypothesis for a significant linear correlation

6) What is the valid prediction range for this setting?

7) Use the regression equation to predict a student's final course grade if 75 optional homework assignments are done.

8) Use the regression equation to compute the number of optional homework assignments that need to be completed if a student expects a course grade of 85

Problem 2

The following data set of the heights and weights of a random sample of 15 male students is acquired. Is there any apparent relationship between the two variables?

S.no	Height	Weight
1	5 ft 6 inch	60 kgs
2	5 ft 4 inch	55 kgs
3	5 ft 8 inch	78 kgs
4	5ft 9 inch	82 kgs
5	5 ft 4 inch	53 kgs
6	5 ft 7 inch	56 kgs
7	5 ft 3 inch	54 kgs
8	5ft 5 inch	65 kgs
9	5 ft 6 inch	74 kgs
10	5 ft 3 inch	65 kgs
11	5 ft 9 inch	76 kgs
12	5ft 10 inch	79 kgs
13	5ft 6 inch	75 kgs
14	5 ft 4 inch	63 kgs
15	5 ft 7 inch	62 kgs

Would you expect the same relationship (if any) to exist between the heights and weights of the opposite sex?

Problem 3

From the following data of hours worked in a factory (x) and output units (y), determine the regression line of y on x, the linear correlation coefficient and determine the type of correlation.

Hours (X)	80	79	83	84	78	60	82	85	79	84	80	62
Production (Y)	300	302	315	330	300	250	300	340	315	330	310	240

Problem 4

The height (in cm) and weight (in kg) of 10 basketball players on a team are as below:

Height (X)	186	189	190	192	193	193	198	201	203	205
Weight (Y)	85	85	86	90	87	91	93	103	100	101

Calculate:

i) The regression line of y on x.
ii) The coefficient of correlation.
Iii) The estimated weight of a player who measures 208 cm.

Unit 9

Classification & Clustering

Title: Classification	**Approach: Interactive/ Discussion, Team Activity, Case studies**

Summary:

Building machine learning (ML) models has traditionally required a binary choice. On one hand, you could manually prepare the features, select the algorithm, and optimize the model parameters in order to have full control over the model design and understand all the thought that went into creating it. However, this approach requires deep understanding of many ML concepts / algorithms and classification is one of them.

There are many practical business applications for machine learning classification. For example, if you want to predict whether or not a person will default on a loan, you need to determine if that person belongs to one of two classes with similar characteristics: the defaulter class or the non-defaulter class. This classification helps you understand how likely the person is to become a defaulter, and helps you adjust your risk assessment accordingly.

Objectives:

1. The main goal of this unit is to help students learn and understand classification problems
2. Define Classification and list its algorithms
3. Student should understand classification as a type of supervised learning

Learning Outcomes:

I. Describe the input and output of a classification model

II. Students should be able to differentiate the regression problem with classification problem

Pre-requisites: Concept of machine learning and artificial intelligence. Understanding of supervised and unsupervised learning and Regression Analysis.

1. Classification

Almost everyday, we deal with classification problems. Here are few interesting examples to illustrate the widespread applications of classification problems.

Case 1:

A credit card company typically receives hundreds of applications for a new credit card. It contains information regarding several different attributes such as, annual salary, outstanding debt, age etc. The problem is to categorize applications into those who have good credit, bad credit or somewhere in the middle. Categorization of the application is nothing but a classification problem.

Case 2:

You may want to own a dog but which kind of dog? This is the beginning of a classification problem. Dogs can be classified in a number of different ways. For example, they can be classified by breed (examples include beagles, hounds, Pug and countless others). they can also be classified by their role in the lives of their masters and the work they do (examples include a dog might be a family pet, a working dog, a show dog, or a hunting dog). In many cases, dogs are defined both by their breed and their role. Based on different classification criteria, you decide eventually which one you want to own.

Let us take a technical example to explain classification problem next.

Case 3:

A common example of classification comes with detecting spam emails. To write a program to filter out spam emails, a computer programmer can train a machine learning algorithm with a set of spam-like emails labeled as “spam” and regular emails labeled as “not-spam”. The idea is to make an algorithm that can learn characteristics of spam emails from this training set so that it can filter out spam emails when it encounters new emails.

Based on the above examples let us try the following activities…

Activity-1

Look at the pictures below and tell me whether the fruit seller knows the art of classification or not. Justify your answer.

In order to understand 'Classification', let us revise the concept of 'Supervised Learning', because classification is type of supervised learning.

Supervised learning as the name indicates is the presence of a supervisor as a teacher. Basically supervised learning is learning in which we teach or train the machine using data which is well labeled that means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples (data) so that supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labeled data.

For instance, suppose you are given a basket filled with different kinds of fruits. Now the first step is to train the machine with all different fruits one by one like below:

- If shape of object is rounded with a depression at top and Red in colour, then it will be labeled as – Apple.
- If shape of object is long curving cylinder and green in colour, then it will be labeled as – Banana.

Now suppose after training the data, you present a new fruit (say Banana) from basket and ask the machine to identify it. Since the machine has already learnt from previous data, it will use the learning wisely this time to classify the fruit based on its shape and color and would confirm the fruit as BANANA and place it in Banana category. Thus the machine learns the things from training data (basket containing fruits) and then applies the knowledge to test data (new fruit).

Supervised learning is further classified into two categories of algorithms:

- **Classification:** A classification problem is when the output variable is a category, such as "Red" or "blue" or "disease" and "no disease".
- **Regression:** A regression problem is when the output variable is a real value, such as "INR" or "Kilograms", "Fahrenheit" etc.

Out of these two supervised learning algorithms, the context of the current unit is – classification learning / training algorithm.

1.1 What is classification in Artificial Intelligence / Machine Learning (AI/ML)

Classification is the process of categorizing a set of data (structured data or unstructured data) into different categories or classes where we can assign label to each class.

Let's say, you live in a gated housing society and your society has separate dustbins for different types of waste: paper waste, plastic waste, food waste and so on. What you are basically doing over here is classifying the waste into different categories and then labeling each category.

In the below picture, we are assigning the labels 'paper', 'metal', 'plastic', and so on to different types of waste.

Let's say you own a shop and you want to figure out if one of your customers is going to come visit your shop again or not. The answer to that question can only be a 'Yes' or 'No'. There can't be third type of answer to such a question. These kind of problems in Machine Learning are known as classification problem. Classification problems normally have a categorical output like a 'yes' or 'no', '1' or '0', 'True' or 'false'.

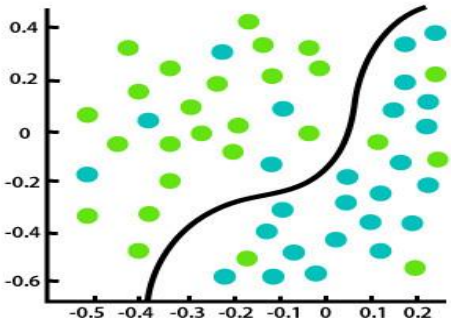
Let's look at another example:

Say you want to check if on a particular day, it will rain or not. In this case the answer is dependent on the weather condition and based on the same, the outcome can either be 'Yes' or 'No'.

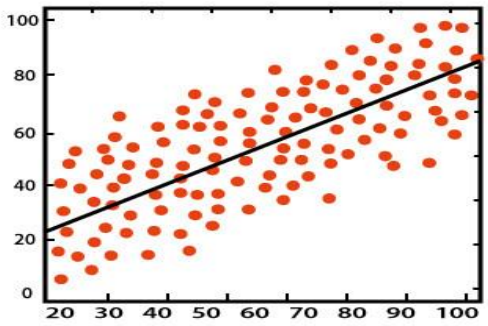
Question 1: Can you find 2 differences between classification and regression?

Question 2: Look at the two graphs below and suggest which graph represents the classification problem.

Graph 1



Graph 1



Question 3: "Predicting stock price of a company on a particular day" - is it a classification problem? Justify your answer.

Question 4: "Predicting whether India will lose or win a cricket match "- is it a regression problem? Justify your answer.

1.2 Few more examples of Classification Problems

Example 1: In the banking industry, where you would like to know whether a transaction is fraudulent or otherwise violating some regulation. That is a classification pattern because most of the time you will attempt to match against a pattern, which may not always be 100% correct.

Example 2: Speech Understanding: Given an utterance from a user, identify the specific request made by the user. A model of this problem would allow a program to understand and make an attempt to fulfill that request. Eg: Siri, Cortana, google now has this capability.

Example 3: Face Detection: Given a digital photo album of many hundreds of digital photographs, identify those photos that include a given person. A model of this decision process would allow a program to organize photos by person. Some cameras and software like Facebook, Google Photos have this capability.

Activity:

Form a group of 5 students. Each group should think and come up with one use case from the classroom environment or their home/society, where they would like to apply classification algorithm to solve the problem.

2. Types of Classification Algorithm

Classification is a type of supervised learning. It labels the examples of input data and is best used when the output has finite and discrete values.

Examples of classification problems include:

- Given an email, classify if it is spam or not.
- Given a handwritten character, classify it as one of the known characters.
- Given recent user behavior, classify as churn or not.

There are two main types of classification tasks that you may encounter, they are:

i) **Binary Classification**: Classification with only 2 distinct classes or with 2 possible outcomes

Example: Male and Female

Example: Classification of spam email and non-spam email

Example: Results of an exam: pass/fail

Example: Positive and Negative sentiment

ii) **Multi Class Classification:** Classification with more than two distinct classes.

Example: classification of types of soil

Example: classification of types of crops

Example: classification of mood/feelings in songs/music

2.1 Binary Classification

As we stated earlier, Binary Classification refers to those classification tasks that have two class labels i.e. two possible outcomes.

Typically, binary classification involves one class that is the normal state and another class that is the abnormal state. For example, "*not spam*" is the normal state and "*spam*" is the abnormal state. Another example is "*cancer not detected*" is the normal state of a task that involves a medical test and "*cancer detected*" is the abnormal state.

The class for the normal state is assigned the class label 0 and the class with the abnormal state is assigned the class label 1.

Popular algorithms that can be used for binary classification include:

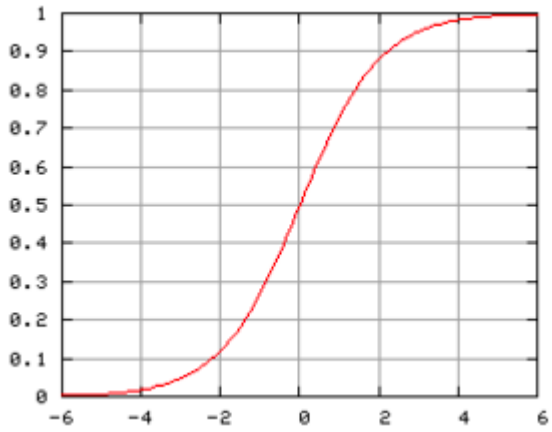
- Logistic Regression
- k-Nearest Neighbors
- Decision Trees
- Support Vector Machine

Out of these binary classification algorithms, we are going to study about 'Logistic Regression'.

2.1.1. Logistic Regression

Logistic regression is one of the binomial classification algorithm used to assign observations to a discrete set of classes. Logistic Regression works with binary data, where either the event happens (1) or the event does not happen (0).

So given some feature **x** it tries to find out whether some event **y** happens or not. So **y** can either be 0 or 1. In the case where the event happens, **y** is given the value 1. If the event does not happen, then **y** is given the value of 0. For example, if **y** represents whether a sports teams wins a match, then **y** will be 1 if they win the match or **y** will be 0 if they do not.



Example of a Logistic Curve is where the values of y cannot be less than 0 or greater than 1.

Let us look at a few more examples to understand logistic regression.

Example 1: Spam detection is a binary classification problem where we are given an email and we need to classify whether or not it is spam. If the email is spam, we label it 1; if it is not spam, we label it 0. In order to apply Logistic Regression to the spam detection problem, the following features of the email are extracted:

- Sender of the email
- Number of typos in the email
- Occurrence of words/phrases like “offer”, “prize”, “free gift”, “lottery”, “you won cash” and more

The resulting feature vector is then used to train a Logistic classifier which emits a score in the range 0 to 1. If the score is more than 0.5, we label the email as spam. Otherwise, we don’t label it as spam.

Example 2: A Logistic Regression classifier may be used to identify whether a tumor is malignant or if it is benign. Several medical imaging techniques are used to extract various features of tumors. For instance, the size of the tumor, the affected body area, etc. These features are then fed to a Logistic Regression classifier to identify if the tumor is malignant or if it is benign.

Above two problems are solved using logistic regression algorithm because the possible labels in both the cases are two only – Spam / Not spam, malignant/benign i.e. binomial classification.

2.2 True positives, true negatives, false positives and false negatives

In the field of machine learning / Artificial Intelligence, a matrix (NxN table) is used to validate how successful a classification model i.e. classifier’s predictions are, where N is the number of target classes. The confusion matrix compares the actual target values with those predicted by the classification model. This gives how well the classification model is performing and what kind of error it is making.

For a binary classification problem, we would have a 2x2 matrix.

	Prediction is Positive	Prediction is Negative
Actual Outcome is Positive	True Positive (TP)	False Negative (FN)
Actual Outcome is Negative	False Positive (FP)	True Negative (TN)

Let’s understand the matrix:

- The target variable has two values: **Positive** or **Negative**
- The **columns** represent the **actual values** of the target variable
- The **rows** represent the **predicted values** of the target variable

But wait – what’s TP, FP, FN and TN here? That’s the point we have to understand in confusion matrix. Let’s understand each term below.

True Positive (TP)

- The predicted value matches the actual value
- The actual value was positive and classification model also predicts positive
- There is no error

True Negative (TN)

- The predicted value matches the actual value
- The actual value was negative and classification model also forecasts negative
- There is no error

False Positive (FP)

- The predicted value doesn't match the actual value
- The actual value was negative but the model predicted a positive value
- This is Type 1 Error

False Negative (FN)

- The predicted value doesn't match the actual value
- The actual value was positive but the model predicted a negative value
- This is Type 2 Error

Let me give you an example from cricket to better explain this.

True Positive (TP) - Umpire gives a batsman NOT OUT when he is NOT OUT.

True Negative (TN) - Umpire gives a Batsman OUT when he is OUT.

False Positive(FP) - Umpire gives a Batsman NOT OUT when he is OUT.

False Negative(FN) - Umpire gives a Batsman OUT when he is NOT OUT.

Question 1:

Assume there are 100 images, 30 of them depict a cat, the rest do not. A machine learning model predicts the occurrence of a cat in 25 of 30 cat images. It also predicts absence of a cat in 50 of the 70 no cat images.

In this case, what are the true positive, false positive, true negative and false negative?

Solution: Assuming cat as a positive class.

Confusion Matrix:

TN	**FP**
FN	**TP**

- **True Positive (TP)**: Images which are cat and actually predicted cat i.e. 25
- **True Negative (TN)**: Images which are not-cat and actually predicted not-cat i.e. 50
- **False Positive (FP)**: Images which are not-cat and actually predicted as cat i.e. 20
- **False Negative (FN)**: Images which are cat and actually predicted as not-cat i.e. 5

Precision: TP/(TP+FP)
Recall: TP/(TP+FN)
Precision: 25/(25+20) = 0.55,
Recall: 25/(25+5) = 0.833

Confusion Matrix Example 1: Do you still remember the shepherd boy story?

"A shepherd boy used to take his herd of sheep across the fields to the lawns near the forest. One day he felt very bored. He wanted to have fun. So he cried aloud "Wolf, Wolf. The wolf is carrying away a lamb". Farmers working in the fields came running and asked, "Where is the wolf?". The boy laughed and replied "It was just for fun. Now get going all of you".
The boy played the trick for quite a number of times in the next few days. After some days, as the boy was perched on a tree, singing a song, there came a wolf. The boy cried loudly "Wolf, Wolf, the wolf is carrying a lamb away." There was no one to the rescue. The boy shouted "Help! Wolf! Help!" Still no one came to his help. The villagers thought that the boy was playing mischief again. The wolf carried a lamb away"

Let us work on arriving at a confusion matrix for the above situation:

- "Wolf" is a **positive class**.
- "No wolf" is a **negative class**.

True Positive (TP): Reality: A wolf threatened Shepherd said: "Wolf" Outcome: Shepherd is a hero	**False Positive (FP):** Reality: No wolf threatened Shepherd said: "Wolf" Outcome: Villagers are angry at shepherd for waking them up
False Negative (FN): Reality: A wolf threatened Shepherd said: "No wolf" Outcome: The wolf ate all the sheep	**True Negative (TN):** Reality: No wolf threatened Shepherd said: "No wolf" Outcome: Everyone is fine

A **true positive** is an outcome where the model *correctly* predicts the *positive* class. Similarly, a **true negative** is an outcome where the model *correctly* predicts the *negative* class.

A **false positive** is an outcome where the model *incorrectly* predicts the *positive* class. And a **false negative** is an outcome where the model *incorrectly* predicts the *negative* class.

Question 2:

Assume there are 100 images, 30 of them depict a cat, the rest do not. A machine learning model predicts the occurrence of a cat in 25 of 30 cat images. It also predicts absence of a cat in 50 of the 70 no cat images.

In this case, what are the true positive, false positive, true negative and false negative? Let's take cat as negative class.

		Predicted Class	
		Spam	No-spam
Actual Class	Spam	12	3
	No-spam	4	81

Question 3:

Below is a confusion matrix prepared for a binary classifier to detect email as Spam and Not Spam.

What is your interpretation of the above matrix?

Why do you need a Confusion matrix?

Here are the benefits of using a confusion matrix:

- It shows how any classification model is confused when it makes predictions
- Confusion matrix not only gives insight into the errors being made by the classifier but also types of errors that are being made
- This breakdown helps overcome the limitations of using classification accuracy alone
- Every column of the confusion matrix represents the instances of the predicted class
- Each row of the confusion matrix represents the instances of the actual class
- It provides insight not only into the errors which are made by a classifier but also errors that are being made in general

2.3. False Positive or False Negative in Medical Science

In medical testing, and more generally in binary classification, a false positive is an error in data reporting in which a test result improperly indicates presence of a condition, such as a disease (the result is positive), when in reality it is not present, while a false negative is an error in which a test result improperly indicates absence of a disease, when in reality it is present. These are the two kinds of errors in a binary test.

While many of today's medical tests are accurate and reliable but still there are false positives or false negatives and their implications are severe on the patients, family or society.

False positive prompts patients to take medication or treatment they don't really need. Perhaps, even more dangerous is the 'false negative' - the test that says you don't have a disease for a condition you actually have.

We most often hear about false negatives in the context of home pregnancy tests, which are more prone to giving false negatives than false positives. However, when it comes to screening for more serious conditions like HIV or cancer, a false negative can have dire repercussions.

Case 1:

Consider a health prediction case, where one wants to diagnose cancer. Imagine that detecting cancer will trigger further analysis (the patient will not be immediately treated) whereas if you don't detect cancer, the patient is sent home without further prognosis.

This case is thus asymmetric, since you definitely would like to avoid sending home a sick patient (False Negative). You can however make the patient wait a little more by asking him/her to take more tests even if the initial results show them negative for cancer (False Positive). As in this situation, you would prefer a False Positive over a False Negative.

Case 2:

Imagine a patient taking an HIV test. The impacts of a **false positive** on the patient would at first be heartbreaking; to have to deal with the trauma of facing this news and telling your family and friends. But on further examination, the doctors will find out that person in question does not have the virus. Again, this would not be a particularly pleasant experience. But not having HIV is ultimately a good thing.

On the other hand, **a false negative** would mean that the patient has HIV but the test shows a negative result. The implications of this are terrifying, the patient would be missing out on crucial treatment and runs the risk of spreading.

Without much doubt, the **false negative** here is the bigger problem. Both for the person and for society.

3. Practice exercise on simple binary classification models

Q 1: A binary classifier was evaluated using a set of 1,000 test examples in which 50% of the examples are negative. It was found that the classifier has 60 % sensitivity and 70 % accuracy. Write the confusion matrix for this case.

Q 2: The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others. In this challenge, we ask you to build a predictive model that answers the question: "what sorts of people were more likely to survive?" using passenger data (i.e. name, age, gender, socio-economic class, etc.). Please refer : https://www.kaggle.com/c/titanic

Q 3: Why can't linear regression be used in place of logistic regression for binary classification?

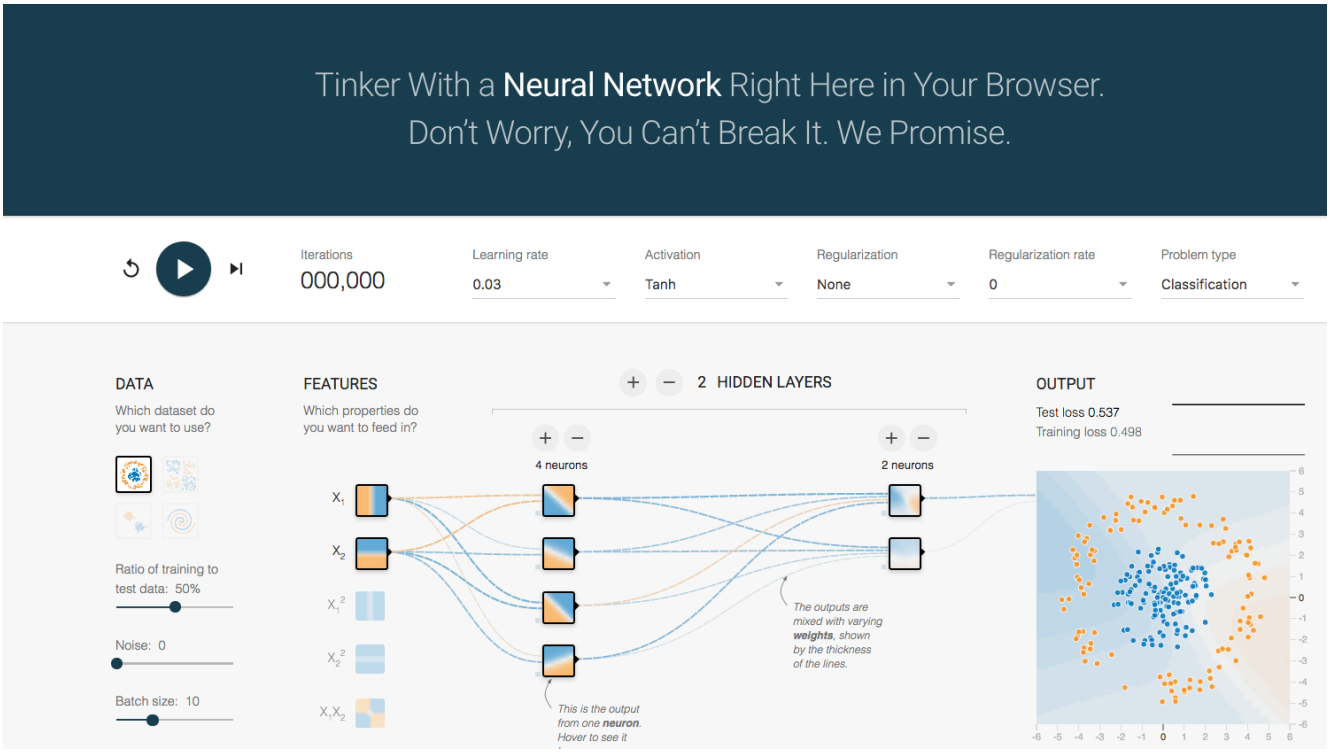
Q 4: What are false positives and false negatives?

Q 5: What is true positive rate (TPR), true negative rate (TNR), false-positive rate (FPR), and false-negative rate (FNR)?

Activity 1:

You may have heard a lot about Artificial Neural Networks (ANN), Deep Learning (DL) and Machine Learning (ML). You must have also heard about the different training algorithms like clustering, classification etc. and would like to learn more. But when you learn about the technology from a textbook, you may find yourself overwhelmed by mathematical models and formulae.

To make this easy and interesting, there's an awesome tool to help you grasp the idea of neural networks and different training algorithms like classification and clustering. This tools is called [TensorFlow Playground](), a web app written in JavaScript that lets you play with a real neural network running in your browser and click buttons and tweak parameters to see how it works.



TensorFlow Playground home screen

First, we will start with understanding some of the terms in the above picture.

I. Data

We have six different data sets Circle, Exclusive OR (XOR), Gaussian, Spiral, plane and multi Gaussian. The first four are for **classification** problems and last two are for **regression problems**. Small circles are the data points which correspond to positive one and negative one. In general, positive values are shown in blue and negative in orange.

In the hidden layers, the lines are colored by the weights of the connections between neurons. Blue shows a positive weight, which means the network is using that output of the neuron as given. An orange line shows that the network is assigning a negative weight.

In the output layer, the dots are colored orange or blue depending on their original values. The background color shows what the network is predicting for a particular area. The intensity of the color shows how confident that prediction is.

II. Features

We have seven features or inputs (X1, X2, squares, product and sine). We can turn on and off different features to see which features are more important. It is a good example of feature engineering.

III. Epoch

Epoch is one complete iteration through the data set.

IV. Learning Rate

Learning rate (alpha) is responsible for the speed at which the model learns.

V. Activation Function

We may skip this term for now but for the purpose of the activity, you may choose any one of the given 4 activation functions (Tanh, ReLu, Sigmoid and Linear). We will read about this in the next class.

VI. Regularization

The purpose of regularization L1 and L2 is to remove / reduce overfitting.

VII. Neural Network Model or Perceptron

A neural network model is a network of simple elements called neurons, which receive input, change their internal state (activation) according to that input, and produce output (0 or 1) depending on the input and activation. We have one input, one output and at least one hidden layer in the simplest neural network called shallow neural network. When the hidden layers are 3 or more then we call it a deep neural network. Each hidden layer has actual working elements called neurons that take input from features or predecessor neurons and calculate a linear activation function (z) and an output function (a).

VIII. Problem Type

We have four data sets for classification and two for regression problem. We can select the type of problem we want to study.

IX. Output

Check the model performance after training the neural network. Observe the Test loss and Training loss of the model.

Activity – Classification problem using TensorFlow playground.

Below are the steps on how to play in this neural network playground:

- Select the Exclusive OR Data Set Classification problem.
- Set Ratio of training and test data to 60% – which means we have 60% train data and 40% testing data.
- Noise is added to 5 and increase it and do some experiment with it, check how the output losses are changing and select the batch size to 10.
- First Select simple features like X1 and X2 then note down the output losses

  (Training loss: -0.004, Test loss: – 0.002, Steps: -255)

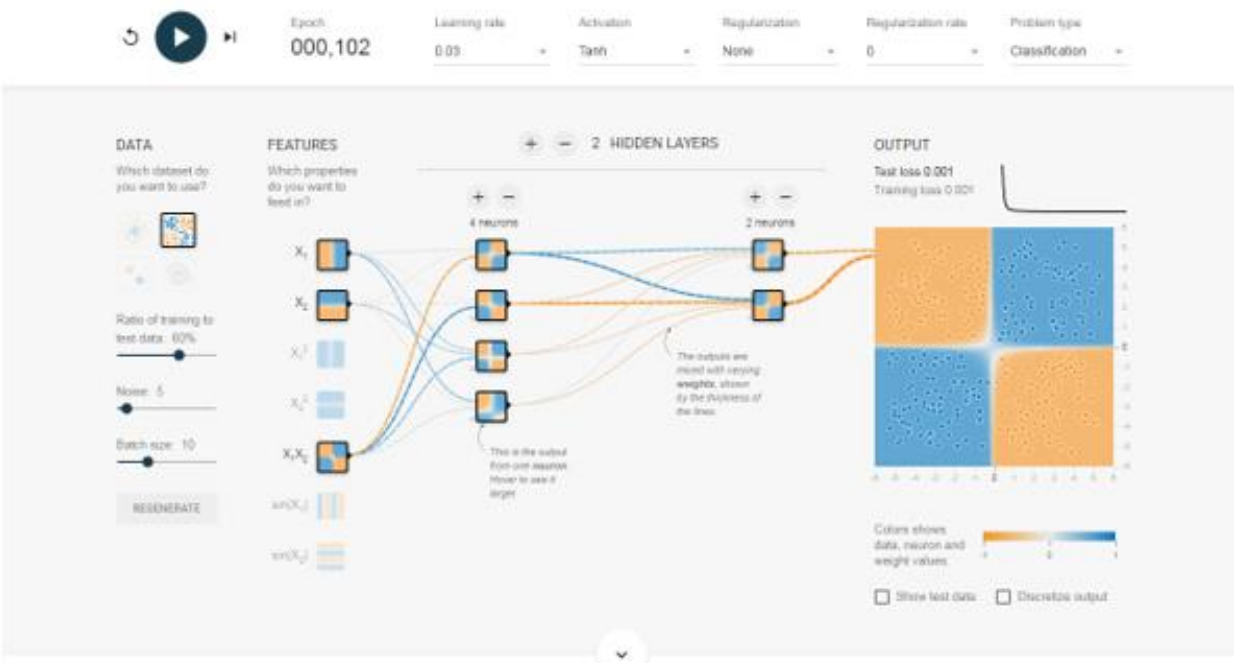
- Now add the third feature product of (X1X2) then observe the Losses

  (Training loss: -0.001, Test loss: – 0.001, Steps: -102)

- This is how you can understand the value of features, and how to get good results in minimum steps
- Set the learning rate to 0.03, also check how the learning rate plays an important role in training a neural network

Since you have already learnt about regression, you may also play with regression, so you have a clear idea about regression.

- Select 2 hidden layers, set 4 neurons for the first hidden layer and 2 neurons for the second hidden layer then followed by the output
- Starting from the first layer the weights are passed on to the first hidden layer which contains output from one neuron, second hidden layer output is mixed with different weights. Weights are represented by the thickness of the lines
- Then the final output will contain the Train and Test loss of the neural network



Title: Clustering	**Approach: Interactive/ Discussion, Team Activity, Case studies**

Summary: Data clustering is a basic problem occurring in machine learning, pattern recognition, computer vision and data compression. The goal of clustering is to categorize similar data into one cluster based on some similarity measure.

In this chapter, we will be reviewing two main components:

First, you will be learning about the purpose of clustering and how it applies to the real world. **Second**, you will get a general overview of clustering such as K-means clustering.

We will also try to understand the implementation of clustering algorithm to solve some real world problems.

Objectives:

1. The main goal of this unit is to help students learn and understand clustering problems
2. Define Clustering and list its algorithms
3. Understand clustering as a type of unsupervised learning

Learning Outcomes:

1. Describe the input and output of a clustering model
2. Students should be able to differentiate between supervised and unsupervised learning
3. Students also should be able to differentiate classification problems from clustering problems.

Pre-requisites: Understanding of supervised and unsupervised learning

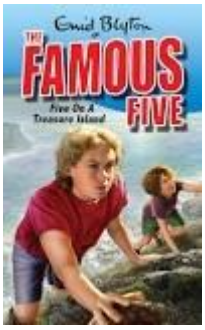
Key Concepts. Clustering algorithms in Machine learning

1. Clustering

Consider you have large collection of books that you have to arrange according to categories in a bookshelf. Since you haven't read all the books, you have no idea about the content of the titles. You start by first bringing the books with similar titles together.

For example, you would arrange books like the "Harry Potter" series in one corner and the "Famous Five" series in another.

Harry Potter Series (Cluster -1)



Famous Five series collection (Cluster – 2)

This is your first experience with clustering, where the books are clustered according to the similarity in their titles. There could be many other criteria of clustering like – clustering based on authors, genre, year publication, hardcover vs. paperback etc.

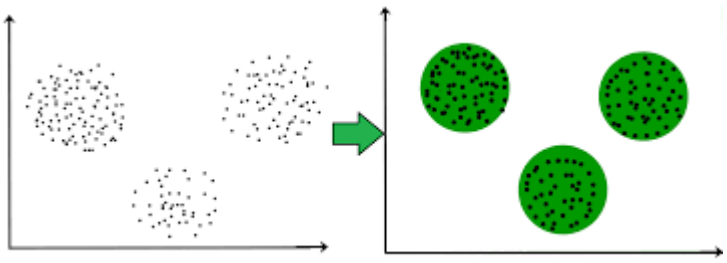
Let us take another example:

When I visit a city, I would like to walk as much as possible, but I want to optimize my time to see as many attractions as possible. While I am planning my next trip to Mumbai for four days. I have researched online and made a list of 20 places that I would like to visit, at during this trip. In order to optimize time and cover all the shortlisted places, I will need to bucket ("cluster") the places based on proximity to each other. Creating the buckets is in fact a method of clustering. Having said that, we perform the process of clustering almost every day in some way or the other.

1.1 What is Clustering

Clustering is unsupervised learning which deals with finding a pattern in the collection of unlabeled data. Having said that, clustering is a technique of grouping similar data in such a way that data/objects in a group are more similar to each other than the data/ objects in the other groups.

Let us understand this with a simple graphical example:

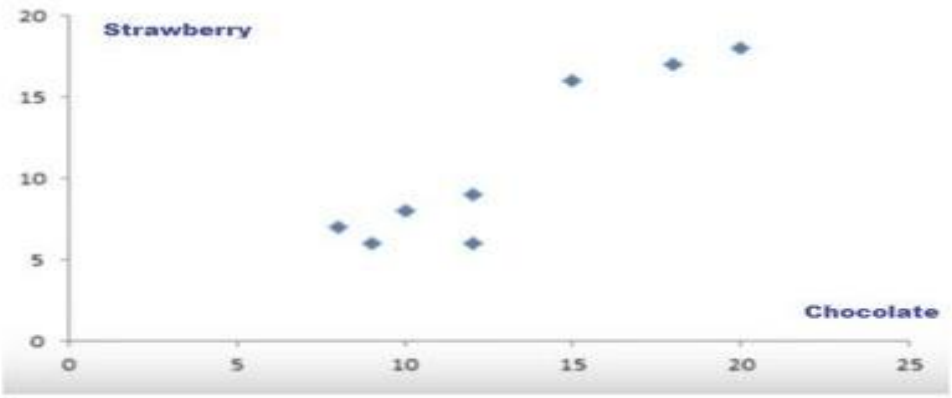


Example of Clustering technique – Grouping similar data in similar groups

Let us take another example to understand clustering. Imagine *X* owns a chain of flavored milk parlors. The parlor sells milk in 2 flavors – Strawberry (S) and Chocolate (C) across 8 outlets. In the below table, you see the sales of both strawberry and chocolate flavored milk across the eight outlets.

Outlet	Strawberry	Chocolate
Outlet 1	12	6
Outlet 2	15	16
Outlet 3	18	17
Outlet 4	10	8
Outlet 5	8	7
Outlet 6	9	6
Outlet 7	12	9
Outlet 8	20	18

In order to get a better understanding of the sales data, you can plot it on a graph. Below we have plotted the sales of both strawberry and chocolate. There are eight dots in this graph that represents the 8 stores and the Y-axis indicates the strawberry sales and the X- axis indicates the chocolate sales.



After the analysis of this graph, you will have a better insight into the sales data and see a pattern emerging with respect to two groups of stores that behave slightly different in terms of their strawberry and chocolate sales and this is essentially how clustering works.

Clustering algorithms can be applied in many fields, for instance:

- **Marketing*:*** If you are a business, it is crucial that you target the right people. Clustering algorithms are able to group together people with similar traits and likelihood to purchase your product/service. Once you have the groups identified, target your messaging to them to increase sales probability.
- **Biology*:*** Classification of plants and animals given their features

- **Libraries*:*** Book ordering
- **Insurance*:*** Identifying groups of motor insurance policy holders with a high average claim cost; Identifying frauds
- **City-planning*:*** Identifying groups of houses according to their house type, value and geographical location
- **Earthquake studies*:*** Clustering observed earthquake epicenters to identify dangerous zones
- **WWW*:*** Document classification; clustering weblog data to discover groups of similar access patterns
- **Identifying Fake News:** Fake news is being created and spread at a rapid rate due to technology innovations such as social media. But clustering algorithm is being used to identify fake news based on the news content. The way that the algorithm works is by taking in the content of the fake news article and examining the words used and then clustering them. These clusters are what help the algorithm determine which pieces are genuine and which ones are fake. Certain words are found more commonly in fake articles and once you see more such words in an article, it gives a higher probability of the material being fake news.

1.2 Clustering Workflow

In order to cluster the data, the following steps are required to be taken:

1. Prepare the data: Data preparation refers to the set of features that will be available to the clustering algorithm. For the clustering strategy to be effective, the data representation must include descriptive features in the input set (feature selection), or the new features based on the original set to be generated (feature extraction). In this stage, we normalize, scale, and transform feature data.

2. Create similarity metrics: To calculate the similarity between two data sets, you need to combine all the feature data for the two examples into a single numeric value. For instance, consider a shoe data set with only one feature – “shoe size”. You can quantify how similar two shoes are by calculating the difference between their sizes. The smaller the numerical difference between sizes, the greater the similarity between shoes. Such a handcrafted similarity measure is called a **manual similarity measure**. The similarity measure is critical to any clustering technique and it must be chosen carefully.

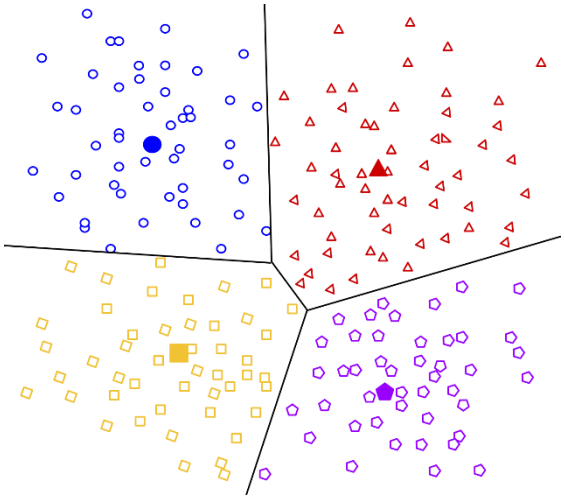
3. Run the clustering algorithm: In machine learning, you sometimes encounter datasets that can have millions of examples. ML algorithms must scale efficiently to these large datasets. However, many clustering algorithms do not scale because they need to compute the similarity between all pairs of points. There are many different approaches to clustering data. Roughly speaking, the cluster algorithms can be classified as hierarchical or partitioning for a more comprehensive taxonomy of clustering techniques.

4. Interpret the results: Because clustering is unsupervised, no “truth” is available to verify results. The absence of truth complicates assessing quality. In this situation, interpretation of results becomes crucial.

2.Types of Clustering

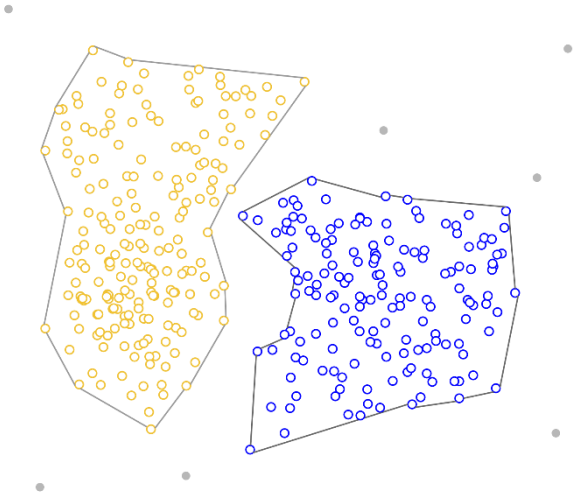
In fact, there are more than 100 clustering algorithms known. But few of the algorithms are used popularly, let's look at them in detail:

1. Centroid-based clustering organizes the data into non-hierarchical clusters, k-means is the most widely-used centroid-based clustering algorithm. Centroid-based algorithms are efficient but sensitive to initial conditions and outliers. This course focuses on k-means because it is an efficient, effective, and simple clustering algorithm.



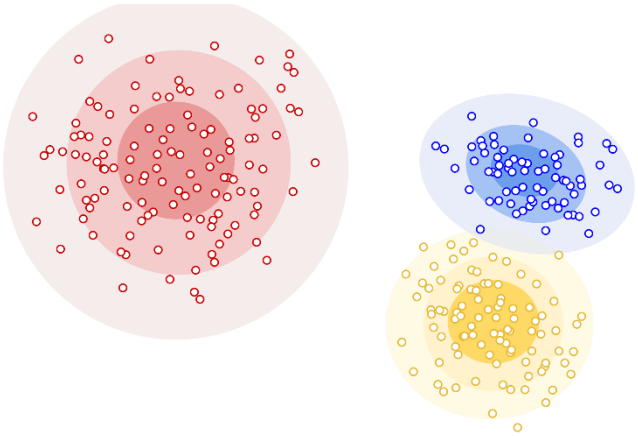
Example of centroid-based clustering

2. Density-based clustering connects areas of high example density into clusters. This allows for arbitrary-shaped distributions as long as dense areas can be connected. These algorithms have difficulty with data of varying densities and high dimensions. Further, by design, these algorithms do not assign outliers to clusters.

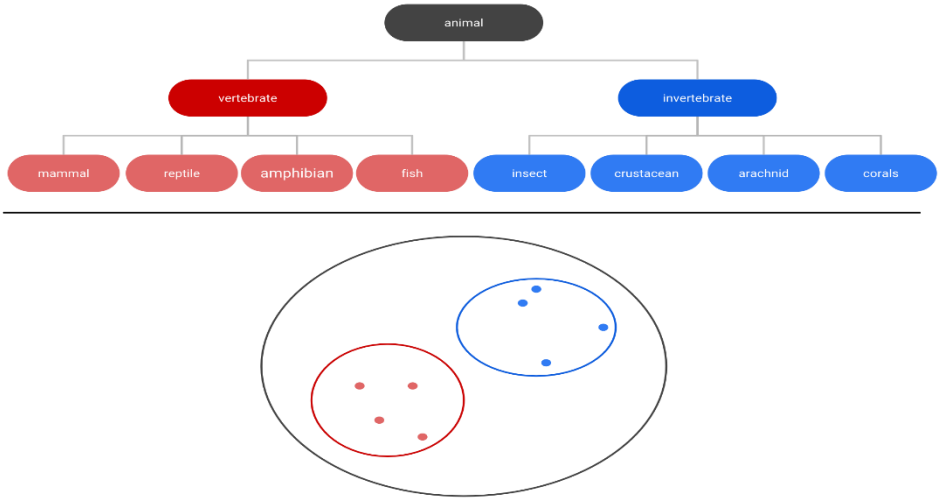


Example of density-based clustering

3.Distribution-based Clustering approach assumes data is composed of distributions, such as **Gaussian distributions**. In the below figure, the distribution-based algorithm clusters data into three Gaussian distributions. As distance from the distribution's center increases, the probability that a point belongs to the distribution decreases. The bands show that decrease in probability. When you do not know the type of distribution in your data, you should use a different algorithm.



4.Hierarchical clustering creates a tree of clusters. Hierarchical clustering, not surprisingly, is well suited to hierarchical data, such as taxonomies. See Comparison of 61 Sequenced Escherichia coli Genomes by Oksana Lukjancenko, Trudy Wassenaar & Dave Ussery for an example. In addition, another advantage is that any number of clusters can be chosen by cutting the tree at the right level.



Example of a hierarchical tree clustering animals.

Out of several approaches to clustering mentioned above, the most widely used clustering algorithm is - *"centroid-based clustering using k-means".*

2.1. K- means Clustering

K-means is an iterative clustering algorithm that aims to find local maxima in each iteration. This algorithm works in two steps:

Step 1: Cluster Assignment

In this step, the algorithm goes to each of the data points and assigns the data point to one of the cluster centroids. The assignment of data point to a particular cluster is determined by how close the data point is from the particular centroid.

Step 2: Move centroid

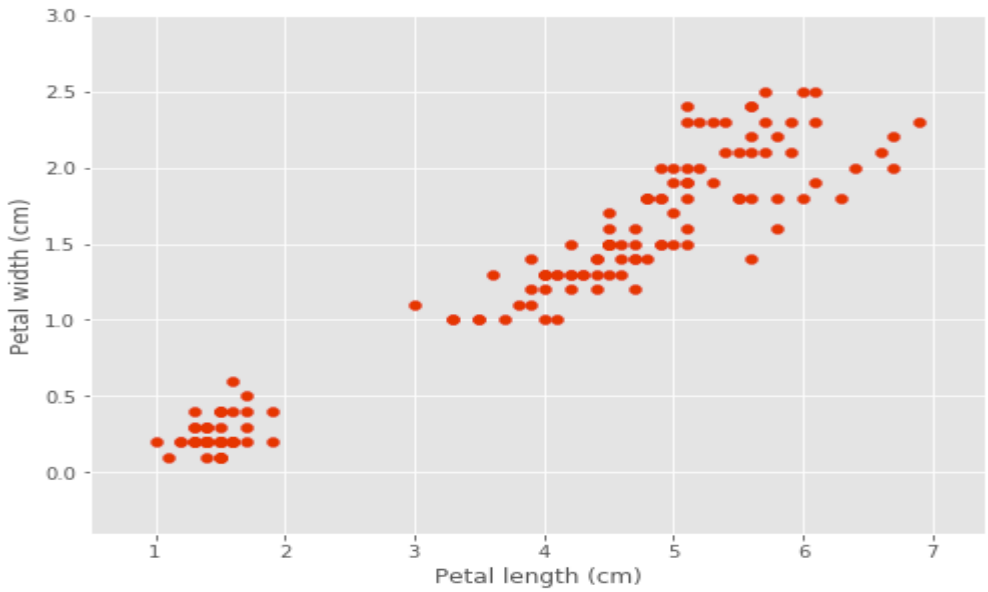
In move centroid step, K-means moves the centroids to the average of the points in a cluster. In other words, the algorithm calculates the average of all the points in a cluster and moves the centroid to that average location. This process is repeated until all data points get a cluster and hence there is no further opportunity of change in the clusters. The number of starting cluster is chosen randomly.

Example 1:

Let us see how this algorithm works using the well-known Iris flower data set - (https://archive.ics.uci.edu/ml/datasets/iris) .

This dataset contains four measurements of three different Iris flowers. The measurements are - Sepal length, Sepal width, Petal length, and Petal width. The three types of Iris are Setosa, Versicolour, and Virginica as shown below in the same order.

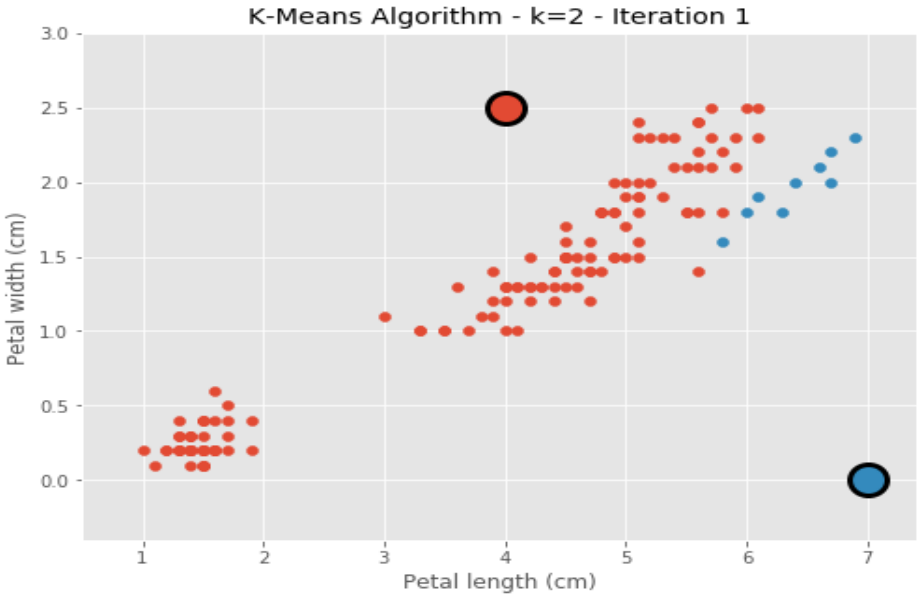
Let's first plot the values of the petals' lengths and widths against each other.



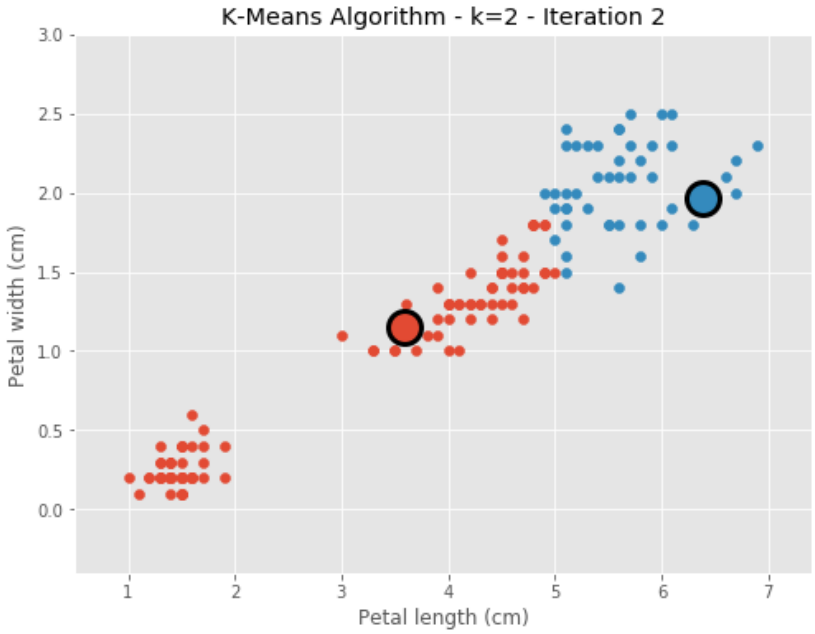
With just a quick glance, it is clear that there are at least two groups of flowers shown on the chart. Let's see how we can use a K-means algorithm to find clusters in this data.

Applying the K-Means Algorithm

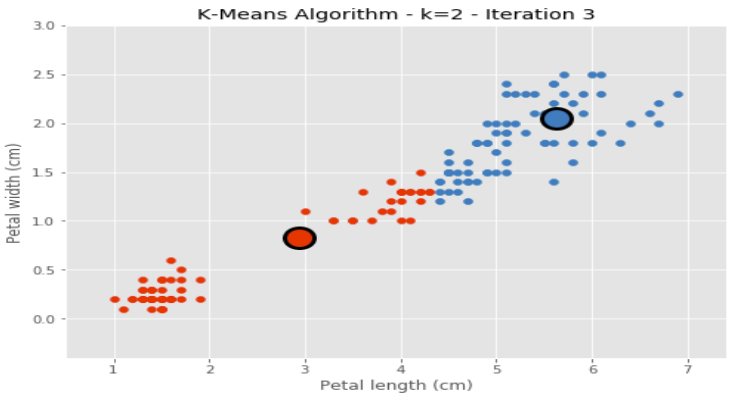
Iteration 1: First, we create two randomly generated centroids and assign each data point to the cluster of the closest centroid. In this case, because we are using two centroids, that means we want to create two clusters i.e. K=2.

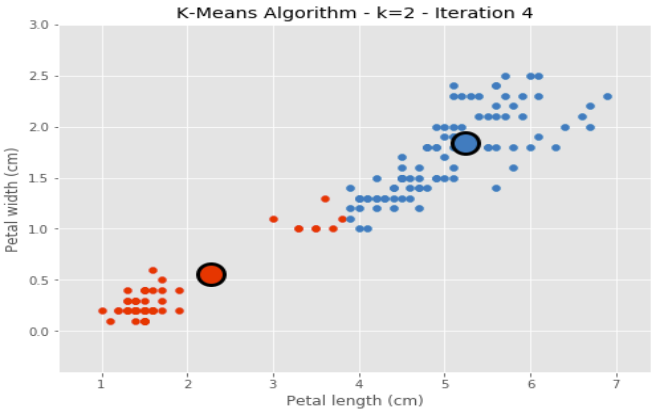


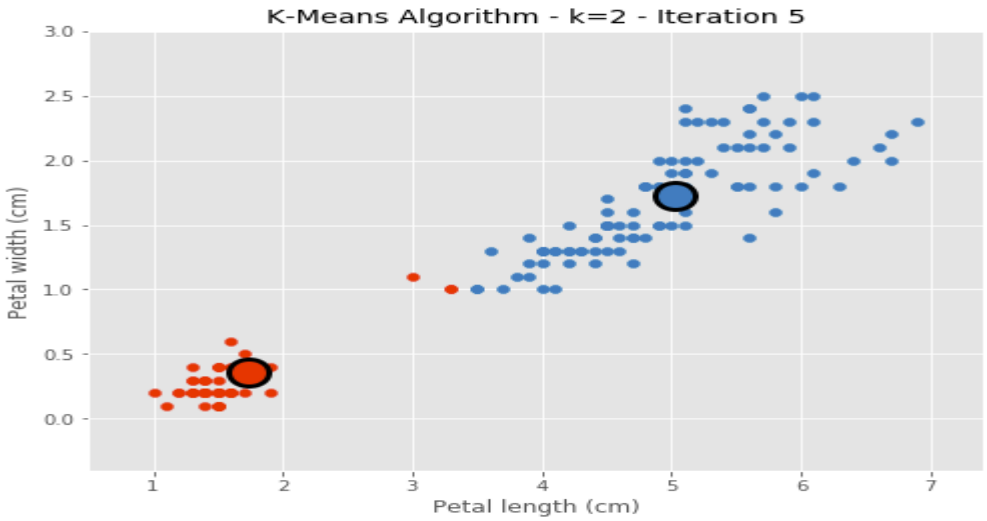
Iteration 2: As you can see above, the centroids are not evenly distributed. In the second iteration of the algorithm, the average values of each of the two clusters are found and become the new centroid values.



Iterations 3-5: We repeat the process until there is no further change in the value of the centroids.







Finally, after iteration 5, there is no further change in the clusters.

Finally, after iteration 5, there is no further change in the clusters.

2.1.1 k-Means Clustering: Advantages and Disadvantages

Advantages of k-means

- Relatively simple to implement.
- Scales to large data sets.
- Guarantees convergence.
- Can warm-start the positions of centroids.
- Easily adapts to new examples.
- Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

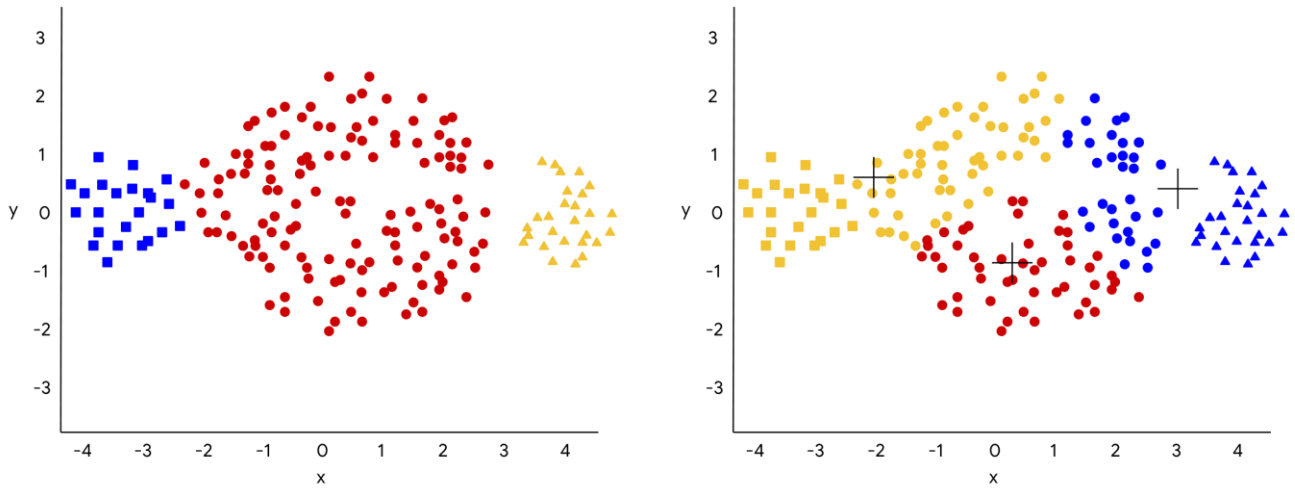
Disadvantages of k-means

- **Choosing k manually**
- **Being dependent on initial values:** For a low k, you can mitigate this dependence by running k-means several times with different initial values and picking the best result. As 'k' increases, you need advanced versions of k-means to pick better values of the initial centroids (called k-means seeding
- **Clustering data of varying sizes and density:** k-means has trouble clustering data where clusters are of varying sizes and density. To cluster such data, you need to generalize k-means as described in the Advantages section.
- **Clustering outliers:** Centroids can be dragged by outliers, or outliers might get their own cluster instead of being ignored. Consider removing or clipping outliers before clustering.
- **Scaling with number of dimensions:** As the number of dimension increases, a distance-based similarity measure converges to a constant value between any given examples. Reduce dimensionality either by using PCA on the feature data, or by using "spectral clustering" to modify the clustering algorithm as explained below.

2.1.2. k-means Generalization

What happens when clusters are of different densities and sizes? Look at Figure 1. Compare the intuitive clusters on the left side with the clusters actually found by k-means on the right side. The comparison shows how k-means can stumble on certain datasets.

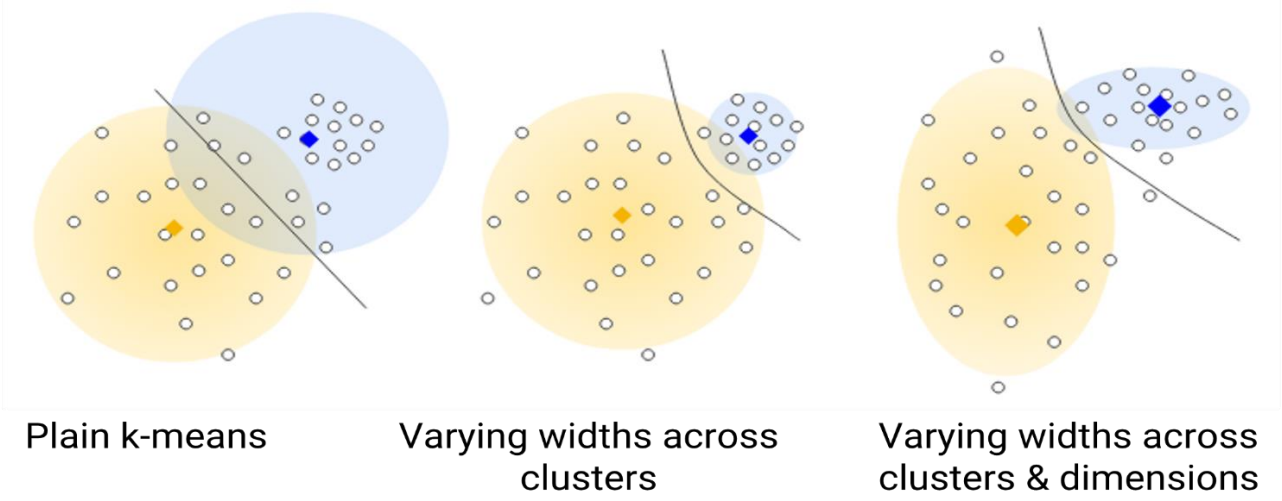
Two graphs side-by-side. The first showing a dataset with somewhat obvious clusters. The second showing an odd grouping of examples after running k-means.



Ungeneralised k-means example.

To cluster naturally imbalanced clusters like the ones shown in Figure 1, you can adapt (generalize) k-means. In Figure 2, the lines show the cluster boundaries after generalizing k-means as:

- Left plot: No generalization, resulting in a non-intuitive cluster boundary.
- Centre plot: Allow different cluster widths, resulting in more intuitive clusters of different sizes.
- Right plot: Besides different cluster widths, allow different widths per dimension, resulting in elliptical instead of spherical clusters, improving the result.



Two graphs side-by-side. The first a spherical cluster example and the second a non-spherical cluster example.

A spherical cluster example and a non-spherical cluster example.

3. Why is it Unsupervised?

In clustering, we group some data-points into several clusters. So usually clustering does not look at target/labels instead it groups the data considering the similarities in the features. Therefore, clustering employs a similarity function to measure the similarity between two data-points (e.g. k means clustering measures the Euclidean distance). And feature engineering plays a key role in clustering because the feature that you provide to the cluster decides the type of groups that you get.

For example, if you use a set of features that characterized the CPU (no. of cores, clock speed etc.) to cluster laptops, each cluster will have laptops with similar CPU power, if you add the price of the laptop as a feature you may be able to get clusters that illustrate overpriced and economical laptops based on their price and CPU specs.

How do you classify it?

The usual approach requires a set of labelled data/or a person to annotate the clusters.

4. Decide the features
5. Cluster the data
6. Use labelled data or human evaluators to annotate the clusters.

In the third step, we try to assign a label to each cluster by looking at the data-points in them. If a certain cluster has 90% of overpriced laptops (based on the labelled data or human evaluation), then we label that cluster as an overpriced laptop cluster. Such that we may get multiple overpriced laptop clusters. When we classify a new laptop, if it belongs to one of those overpriced laptop clusters then we classify that laptop as an overpriced laptop.

For Advance Learners (Optional)

Project 1: Customer Segmentation (Clustering)

Install

This project requires **Python 2.7** and the following Python libraries installed:

- NumPy
- Pandas
- matplotlib
- scikit-learn

You will also need to have software installed to run and execute an iPython Notebook

Code

Template code is provided in the notebook `customer_segments.ipynb` notebook file. Additional supporting code can be found in `renders.py`. While some code has already been implemented to get you started, you will need to implement additional functionality when requested to successfully complete the project.

(Source Code : https://github.com/ritchieng/machine-learning-nanodegree)

Getting Started

In this project, you will analyze a dataset containing data on various customers' annual spending amounts (reported in *monetary units*) of diverse product categories for internal structure. One goal of this project is to best describe the variation in the different types of customers that a wholesale distributor interacts with. Doing so would equip the distributor with insight into how to best structure their delivery service to meet the needs of each customer.

The dataset for this project can be found on the UCI Machine Learning Repository. For the purposes of this project, the features `'Channel'` and `'Region'` will be excluded in the analysis — with focus instead on the six product categories recorded for customers.

Run the code block below to load the wholesale customers dataset, along with a few of the necessary Python libraries required for this project. You will know the dataset loaded successfully if the size of the dataset is reported.

```
# Import libraries necessary for this project

import numpy as np
import pandas as pd
import renders as rs
from IPython.display import display # Allows the use of display() for DataFrames

# Show matplotlib plots inline (nicely formatted in the notebook)
%matplotlib inline

# Load the wholesale customers dataset
try:
```

```
    data = pd.read_csv("customers.csv")
    data.drop(['Region', 'Channel'], axis = 1, inplace = True)
    print "Wholesale customers dataset has {} samples with {} features
each.".format(*data.shape)
except:
    print "Dataset could not be loaded. Is the dataset missing?"
```

STEP -1: Data Exploration

Run the code block below to observe a statistical description of the dataset. Note that the dataset is composed of six important product categories: **'Fresh'**, **'Milk'**, **'Grocery'**, **'Frozen'**, **'Detergents_Paper'**, and **'Delicatessen'**. Consider what each category represents in terms of products you could purchase.

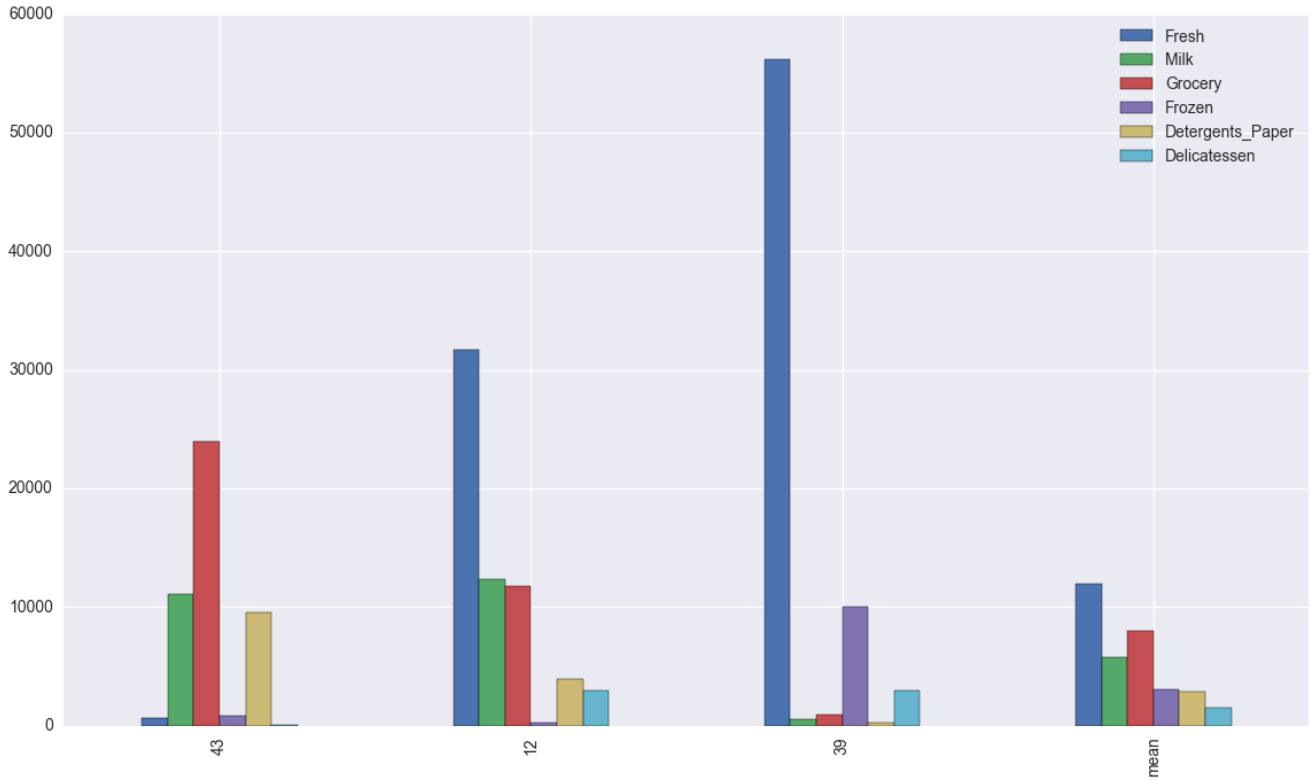
```
# Display a description of the dataset

stats = data.describe()
stats
```

OUTPUT:

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

Data Visualization code [Sample]

```
  # Import Seaborn, a very powerful library for Data Visualisation

import seaborn as sns

# Get the means
mean_data = data.describe().loc['mean', :]

# Append means to the samples' data
samples_bar = samples.append(mean_data)

# Construct indices
samples_bar.index = indices + ['mean']

# Plot bar plot
samples_bar.plot(kind='bar', figsize=(14,8))
```



STEP -2: Implementation – Features Relevance

The code is slightly big, so not writing code block here. Please refer the github link shared above.

In the code block below, you will need to implement the following:

- Assign `new_data` a copy of the data by removing a feature of your choice using the `DataFrame.drop` function.
- Use `sklearn.cross_validation.train_test_split` to split the dataset into training and testing sets.
    - Use the removed feature as your target label. Set a `test_size` of `0.25` and set a `random_state`.
- Import a decision tree regressor, set a `random_state`, and fit the learner to the training data.
- Report the prediction score of the testing set using the regressor's `score` function.

STEP – 3: Data Pre-processing

In the code block below, you will need to implement the following:

- Assign a copy of the data to `log_data` after applying a logarithm scaling. Use the `np.log` function for this.
- Assign a copy of the sample data to `log_samples` after applying a logrithm scaling. Again, use `np.log`.

```
# TODO: Scale the data using the natural logarithm
log_data = np.log(data)
```

```
# TODO: Scale the sample data using the natural logarithm
log_samples = np.log(samples)

# Produce a scatter matrix for each pair of newly-transformed features
pd.scatter_matrix(log_data, alpha = 0.3, figsize = (14,8), diagonal = 'kde');
```

STEP – 4: Outlier Detection

Detecting outliers in the data is extremely important in the data preprocessing step of any analysis. The presence of outliers can often skew results which take into consideration these data points.

In the code block below, you will need to implement the following:

- Assign the value of the 25th percentile for the given feature to `Q1`. Use `np.percentile` for this.
- Assign the value of the 75th percentile for the given feature to `Q3`. Again, use `np.percentile`.
- Assign the calculation of an outlier step for the given feature to `step`.
- Optionally remove data points from the dataset by adding indices to the `outliers` list.

NOTE: If you choose to remove any outliers, ensure that the sample data does not contain any of these points! Once you have performed this implementation, the dataset will be stored in the variable `good_data`.

```
import itertools
# Select the indices for data points you wish to remove
outliers_lst  = []

# For each feature find the data points with extreme high or low values
for feature in log_data.columns:
    # TODO: Calculate Q1 (25th percentile of the data) for the given feature
    Q1 = np.percentile(log_data.loc[:, feature], 25)

    # TODO: Calculate Q3 (75th percentile of the data) for the given feature
    Q3 = np.percentile(log_data.loc[:, feature], 75)

    # TODO: Use the interquartile range to calculate an outlier step (1.5 times
the interquartile range)
    step = 1.5 * (Q3 - Q1)

    # Display the outliers
    print "Data points considered outliers for the feature
'{}':".format(feature)

    # The tilde sign ~ means not
    # So here, we're finding any points outside of Q1 - step and Q3 + step
    outliers_rows = log_data.loc[~((log_data[feature] >= Q1 - step) &
(log_data[feature] <= Q3 + step)), :]
    # display(outliers_rows)

    outliers_lst.append(list(outliers_rows.index))

outliers = list(itertools.chain.from_iterable(outliers_lst))

# List of unique outliers
# We use set()
# Sets are lists with no duplicate entries
```

```
uniq_outliers = list(set(outliers))

# List of duplicate outliers
dup_outliers = list(set([x afor x in outliers if outliers.count(x) > 1]))

print 'Outliers list:\n', uniq_outliers
print 'Length of outliers list:\n', len(uniq_outliers)

print 'Duplicate list:\n', dup_outliers
print 'Length of duplicates list:\n', len(dup_outliers)

# Remove duplicate outliers
# Only 5 specified
good_data = log_data.drop(log_data.index[dup_outliers]).reset_index(drop = True)

# Original Data
print 'Original shape of data:\n', data.shape
# Processed Data
print 'New shape of data:\n', good_data.shape
```

STEP- 5: Dimensionality Reduction

When using principal component analysis, one of the main goals is to reduce the dimensionality of the data — in effect, reducing the complexity of the problem. Dimensionality reduction comes at a cost: Fewer dimensions used implies less of the total variance in the data is being explained. Because of this, the *cumulative explained variance ratio* is extremely important for knowing how many dimensions are necessary for the problem.

In the code block below, you will need to implement the following:

- Assign the results of fitting PCA in two dimensions with `good_data` to `pca`.
- Apply a PCA transformation of `good_data` using `pca.transform`, and assign the reuslts to `reduced_data`.
- Apply a PCA transformation of the sample log-data `log_samples` using `pca.transform`, and assign the results to `pca_samples`.

```
# TODO: Apply PCA by fitting the good data with only two dimensions
# Instantiate
pca = PCA(n_components=2)
pca.fit(good_data)

# TODO: Transform the good data using the PCA fit above
reduced_data = pca.transform(good_data)

# TODO: Transform the sample log-data using the PCA fit above
pca_samples = pca.transform(log_samples)

# Create a DataFrame for the reduced data
reduced_data = pd.DataFrame(reduced_data, columns = ['Dimension 1', 'Dimension
2'])
```

STEP - 6: CLUSTERING

Depending on the problem, the number of clusters that you expect to be in the data may already be known. When the number of clusters is not known *a priority*, there is no guarantee that a given number of clusters best segments the data, since it is unclear what structure exists in the data — if any.

In the code block below, you will need to implement the following:

- Fit a clustering algorithm to the `reduced_data` and assign it to `clusterer`.
- Predict the cluster for each data point in `reduced_data` using `clusterer.predict` and assign them to `preds`.
- Find the cluster centers using the algorithm's respective attribute and assign them to `centers`.
- Predict the cluster for each sample data point in `pca_samples` and assign them `sample_preds`.
- Import sklearn.metrics.silhouette_score and calculate the silhouette score of `reduced_data` against `preds`.
    - Assign the silhouette score to `score` and print the result.

```
# Imports
from sklearn.mixture import GMM
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score

# Create range of clusters
range_n_clusters = list(range(2,11))
print(range_n_clusters)

[2, 3, 4, 5, 6, 7, 8, 9, 10]
```

GMM Implementation

```
# Loop through clusters
for n_clusters in range_n_clusters:
    # TODO: Apply your clustering algorithm of choice to the reduced data
    clusterer = GMM(n_components=n_clusters).fit(reduced_data)

    # TODO: Predict the cluster for each data point
    preds = clusterer.predict(reduced_data)

    # TODO: Find the cluster centers
    centers = clusterer.means_

    # TODO: Predict the cluster for each transformed sample data point
    sample_preds = clusterer.predict(pca_samples)

    # TODO: Calculate the mean silhouette coefficient for the number of clusters
chosen
    score = silhouette_score(reduced_data, preds, metric='mahalanobis')
    print "For n_clusters = {}. The average silhouette_score is :
{}".format(n_clusters, score)
```

KNN Implementation

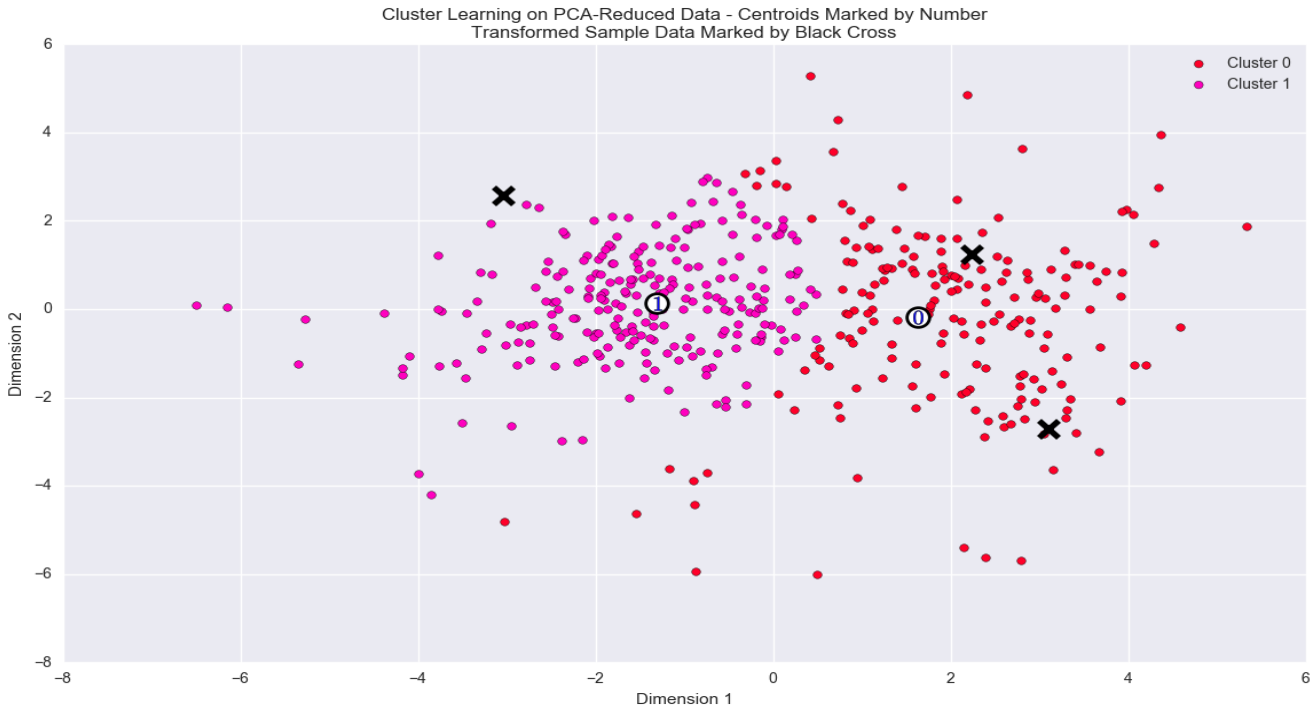
```
# Loop through clusters
for n_clusters in range_n_clusters:
    # TODO: Apply your clustering algorithm of choice to the reduced data
```

```
    clusterer = KMeans(n_clusters=n_clusters).fit(reduced_data)

    # TODO: Predict the cluster for each data point
    preds = clusterer.predict(reduced_data)

    # TODO: Find the cluster centers
    centers = clusterer.cluster_centers_

    # TODO: Predict the cluster for each transformed sample data point
    sample_preds = clusterer.predict(pca_samples)

    # TODO: Calculate the mean silhouette coefficient for the number of clusters
chosen
    score = silhouette_score(reduced_data, preds, metric='euclidean')
    print "For n_clusters = {}. The average silhouette_score is :
{}".format(n_clusters, score)
```

Cluster Visualization

Once you've chosen the optimal number of clusters for your clustering algorithm using the scoring metric above, you can now visualize the results by executing the code block below.The final visualization provided should, however, correspond with the optimal number of clusters.

```
# Extra code because we ran a loop on top and this resets to what we want
clusterer = GMM(n_components=2).fit(reduced_data)
preds = clusterer.predict(reduced_data)
centers = clusterer.means_
sample_preds = clusterer.predict(pca_samples)
```



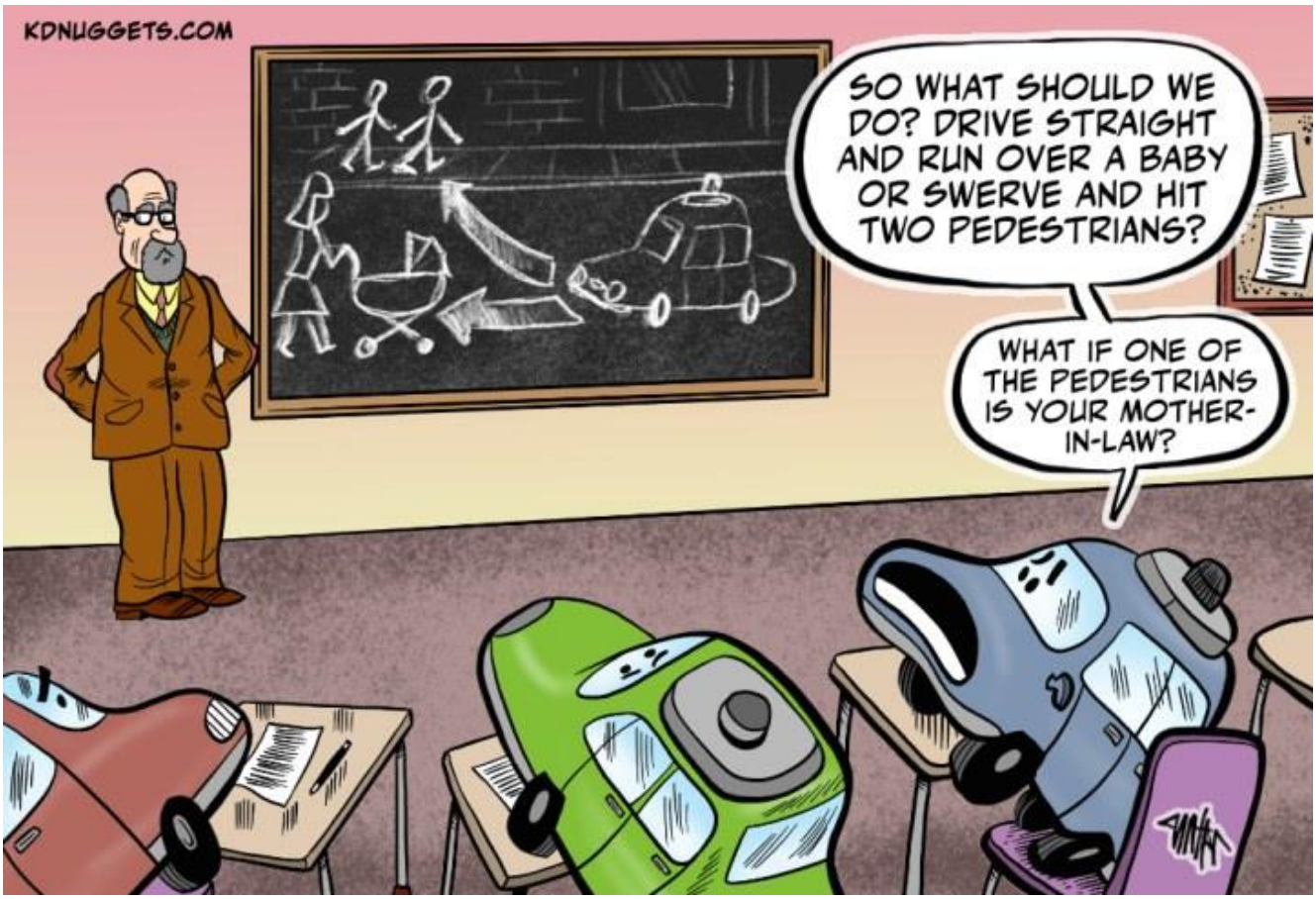

```
Display the results of the clustering from implementation
rs.cluster_results(reduced_data, preds, centers, pca_samples)
```

Unit 10

AI Values

Title: AI Values	**Approach: Interactive/ Discussion, Team Activity, Case studies**
Summary: AI is progressing towards a stage where, eventually it will replicate human's general intelligence. The possibility of making a thinking machine raises a host of ethical issues. These ethical questions ensure that such machines do not harm humans and the society at large. To harness the potential of AI in the right way, guidelines and ethical standards are therefore required. As a result, ethical guidelines have been developed in recent years and developers are expected to be adhere to these principles. Inspite of the standards, collectively as a society we have to face the challenges arising from current AI techniques and implementations, in the form of systematic decrease in privacy; increasing reliance on AI for our safety, and the ongoing job losses due to mechanization and automatic control of work processes.	
Objectives: **1.** Understand and debate on Ethics of AI 2. Understand biases and its types 3. Scope of biases in data and how it impacts AI	
Key Concepts: Data, Bias, Data Bias, Types of Bias	

1. AI Values



Before we begin this chapter, let us watch the few essential videos. (Total watch time may be 30-35 minutes).

The Ethical Robot

How To Build A Moral Robot

Humans Need Not Apply

Activity 1: After watching the video "The Ethical Robot" what are the two ethical questions that strike you? Write them down.

Activity 2: With the video "How to build a moral robot" as your baseline, please write down the moral and ethical values you would like incorporate in your robot? The video is only a guide, let it not limit your imagination and creativity.

Activity 3: Form a group of 5 students and watch the video "Humans need not apply" as a group. Please watch the video more than once. At the end, submit a paper as a group on your learnings from the vide.

1. AI Working for Good

At the World Economic Forum 2019 in Davos, Paul Daugherty, Accenture's Chief Technology and Innovation Officer floated the idea of Human + Machine = Superpowers. What if people could predict natural disasters before they happen? Better protect endangered species? Track disease as it spreads, to eliminate it sooner? These are the actual AI projects on which companies like Google, IBM etc. are working to serve humanity.

Find below a compilation of a few AI projects in progress:

1. **IBM** (https://www.research.ibm.com/artificial-intelligence/#quicklinks)
    - Applying AI to accelerate COVID-19 Research. As the COVID-19 pandemic unfolds, we continue to ask how these technologies and our scientific knowledge can help in the global battle against the corona virus.
    - The potential benefits of AI for breast cancer detection
    - AI Enables Foreign Language Study Abroad, No Travel Required
2. **Google** (https://ai.google/social-good/)
    - Keeping people safe with AI-enabled flood forecasting
3. Assessing Cardiovascular Risk Factors with Computer Vision Agricultural productivity can be increased through digitization and analysis of images from automated drones and satellites
4. AI can be instrumental in providing personalized learning experience to students
5. AI can help the people in special needs in numerous ways. AI is getting better at doing text-to-voice translation as well as voice-to-text translation, and could thus help visually impaired people, or people with hearing impairments, to use information and communication technologies (ICTs)
6. Pattern recognition can track marine life migration, concentrations of life undersea and fishing activities to enhance sustainable marine ecosystems and combat illegal fishing
7. With global warming, climate change, and water pollution on the rise, we could be dealing with a harsh future. Food shortages are not something we want to add to the list. Thankfully, one startup is already working hard on using AI for good in this regard.

8. Imago AI an India-based agri-tech startup that aims to use AI to increase crop yields and reduce food waste. The company's vision is to use technology to feed the world's growing population by optimizing agricultural methods.

a) Original

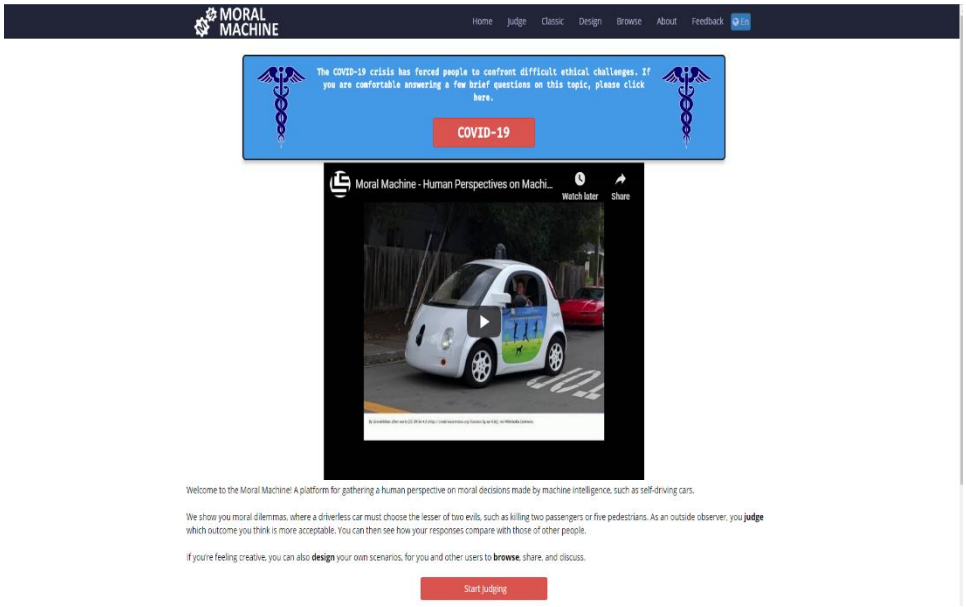
a) Processed

The company combines machine learning and computer vision to automate tedious tasks like measuring crop quality and weighing yields. This won't just speed up the process, but it will also help farmers to identify plants that have diseases. TechCrunch reports that 40% of the world's crops are lost to disease, so the work from Imago AI could be a major breakthrough for agriculture, especially in poorer countries.

(https://www.springboard.com/blog/ai-for-good/)

Activity – Moral Machines – For student to explore!

From self-driving cars on public roads to self-piloting reusable rockets landing on self-sailing ships, machine intelligence is supporting or entirely taking over ever more complex human activities at an ever-increasing pace. The greater autonomy given machine intelligence in these roles can result in situations where they have to make autonomous choices involving human life and limb. This calls for not just a clearer understanding of how humans make such choices, but also a clearer understanding of how humans perceive machine intelligence making such choices. URL - https://www.moralmachine.net/



2. Principles for Ethical AI

The 1980s classic film "The Terminator" stimulated our imaginations, but it also triggered fears about autonomous, intelligent robots eliminating the human race, though this scenario is far-fetched. Artificial Intelligence (AI) is a technology to design a machine which can perform tasks normally requiring human intelligence. In light of its powerful transformative force and profound impact across various societal domains, AI has sparked ample debate about the principles and values that should guide its development and use.

UNI Global Union (http://www.thefutureworldofwork.org/) has identified **10 key principles for Ethical AI**

1. AI systems must be transparent

Consumers should have the right to demand transparency in the decisions and outcomes of AI systems as well as their underlying algorithms. They must also be consulted on AI systems' implementation, development and deployment.

2. AI systems must be equipped with an "ethical black box"

The ethical "black box" should not only contain relevant data to ensure system transparency and accountability, but also include clear data and information on the ethical considerations built into the system.

3. AI must serve people and planet

Codes of ethics for the development, application and use of AI are needed so that throughout their entire operational process, AI systems remain compatible and increase the principles of human dignity, integrity, freedom, privacy and cultural and gender diversity, as well as fundamental human rights.

4. Adopt a human-in-command approach

The development of AI must be responsible, safe and useful where machines maintain the legal status of tools, and legal persons retain control over, and responsibility for these machines at all times.

5. Ensure a gender less, unbiased AI

In the design and maintenance of AI and artificial systems, it is vital that the system is controlled for negative or harmful human-bias, and that any bias–be it gender, race, sexual orientation, age–is identified and is not propagated by the system.

6. Share the benefits of AI systems

The economic prosperity created by AI should be distributed broadly and equally, to benefit all of humanity. Global as well as national policies aimed at bridging the economic, technological and social digital divide are therefore necessary.

7. Secure a just transition and ensure support for fundamental freedoms and rights

As AI systems develop and augmented realities are formed, workers and work tasks will be displaced. It is vital that policies are put in place that ensure a just transition to the digital reality, including specific governmental measures to help displaced workers find new employment.

8. Establish global governance mechanism

Establish multi-stakeholder Decent Work and Ethical AI governance bodies on global and regional levels. The bodies should include AI designers, manufacturers, owners, developers, researchers, employers, lawyers, CSOs and trade unions.

9. Ban the attribution of responsibility to robots

Robots should be designed and operated as far as is practicable to comply with existing laws, and fundamental rights and freedoms, including privacy.

10. Ban AI arms race

Lethal autonomous weapons, including cyber warfare, should be banned.

Now please watch the two videos below:

1. https://www.youtube.com/watch?v=cplucNW70II&ab_channel=TEDxTalks
2. https://www.youtube.com/watch?v=vgUWKXVvO9Q

Question 1: How do you decide if something deserves to be called intelligent? Does it have to pass exams to earn this certificate? Apply your imagination and creativity to answer this question.

Question 2: A village needs your help to prevent the spread of a nearby forest fire. Design, develop and train the Agent to identify what causes fires, remove materials that help fires spread, and then bring life back to a forest destroyed by fire — all with Flowchart / pseudo code.

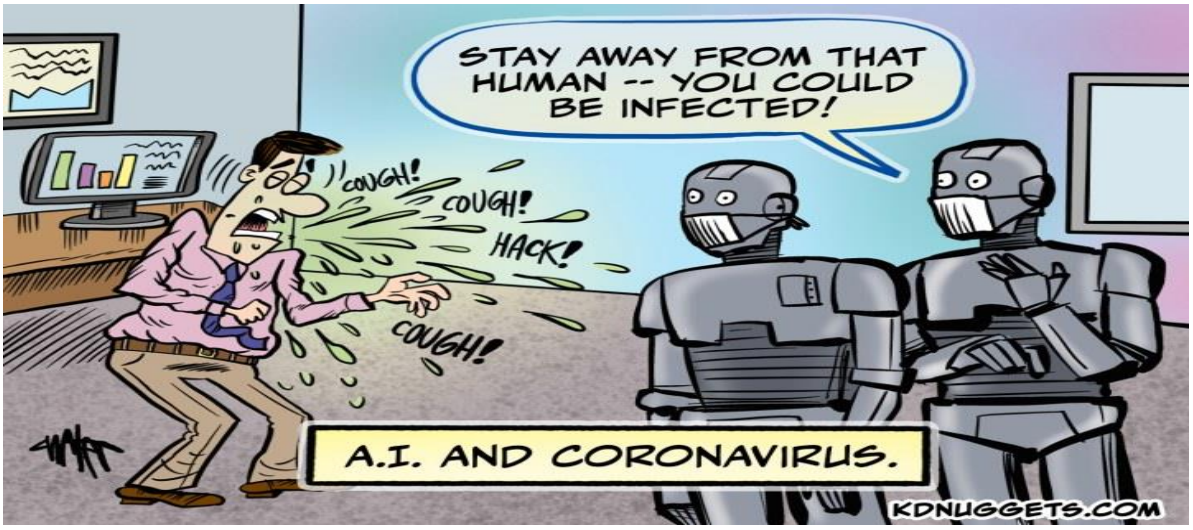
Be informed that while designing the AI agent, most likely your own biases will get inside the algorithm.

Stage 1: Design the whole solution alone first

Stage 2: Include more students in the group and take their perspective about your solution. You will come to know there were biases in your solution

Stage 3: Increase your group size by including students from different classes and different age groups. Let them give their point of view about the solution you've arrived at in Stage 2 and you would be surprised to know that the solution still has a lot of biases.

3. Types of Bias (personal /cultural /societal)



Example 1

Suppose a CCTV camera were to spot your face in a crowd outside a sports stadium. In the police data center somewhere in the city/ country, an artificial neural network analyzes images from the CCTV footage frame-by-frame. A floating cloud in the sky causes a shadow on your face and neural network (by mistake) finds your face similar to the face of a wanted criminal.

If the police were to call you aside for questioning and tell you they had reason to detain you, how would you defend yourself? Was it your fault that your shadowed face has resemblance by few degrees with a person in the police record?

Example 2: This happened in the USA in 2018. An AI system was being used to allocate care to nearly 200 million patients in the US. It was discovered later that AI system was offering a lower standard of care to the black patients. Across the board, black people were assigned lower risk scores than white people. This in turn meant that black patients were less likely to be able to access the necessary standard of care.

The problem stemmed from the fact that the AI algorithm was allocating risk values using the predicted cost of healthcare. Because black patients were often less able to pay or were *perceived* as less able to pay for the higher standard of care, the AI essentially learned that they were not entitled to such a standard of treatment.

Though the system was fixed / improved after being discovered but the big question is – whose problem was this? The AI system developers or the US black people data (which was true to an extent)?

3.1 What is Bias?

Bias is a tendency to lean and act in a certain direction, either in favor of or against a particular thing. Bias lacks the neutral viewpoint. If you're biased *toward* something, then you lean favorably toward it; you tend to think positively of it. Meanwhile, if you're biased *against* something, then you lean negatively against it; you tend to think poorly of it.

The sources of 'Bias in AI' usually are our own cultural, societal or personal biases regarding race, gender, nationality, age or personal habits.

Did you answer "bananas"? Why didn't you mention the plastic bag roll? Or the color of the banana? Or the plastic stand holding the bananas?

Although all answers are technically correct, for some reason we have a bias to prefer one of them. Not all people would share that bias; what we perceive and how we respond is influenced by our norms, culture and habits. If you live on a planet where all bananas are blue, you might answer "yellow bananas" here. If you've never seen a banana before, you might say "shelves with yellow stuff on them."

Question:

Make a list of 10 biases which you observe in your home, classroom or in your society. You don't need to get all 10 biases in one go. You can start with one and keep adding as you observe more.

5. How data driven decisions can be de-biased

"AI bias doesn't come from AI algorithm; it comes from people" - some may disagree to this saying bias comes NOT from people but from dataset. But people make and collect the data. Textbooks reflect the biases of their authors. Like textbooks, datasets have authors. They're collected according to instructions made by people.

While ML and AI are technologies often dissociated from human thinking, they are always based on algorithms created by humans. And like anything created by humans, these algorithms are prone to incorporating the biases of their creators.

Because AI algorithms learn from data, any historical data can quickly create biased AI that bases decisions on unfair datasets.

But there are tangible things we can do to manage bias in AI. Here are some of them:

Educate and check yourself
The first step to removing bias is to proactively look out for it and keep checking your own behavior, as a lot of bias is unconscious.

Build a diverse team
Another way to reduce the risk of bias and to create more inclusive experiences is to ensure the team building the AI system is diverse (for example, with regard to gender, race, education, thinking process, disability status, skill set and problem framing approach). This should include the engineer teams, as well as project and middle management, and design teams.

Be able to explain automated decisions

Explainable AI is new normal. With AI in system, ability to explain the algorithm under the hood is critical. This involves ensuring transparency at both the macro level as well as at the individual level.

It's all about the data – make sure you choose a representative dataset

Choosing data that is diverse and includes different groups to prevent your model from having trouble identifying unlabeled examples that are outside the norm. Make sure you have properly grouped and managed the data so you aren't forced to face similar situations as Google and its facial recognition system.

Activities to teach AI ethics in the classrooms

Activity 1: What do You think

Most western countries in the world have a regulation that says that fire retardants must be added to the foams and fabrics in furniture. New Zealand does not. Do you think New Zealand should have such a regulation? YES or NO?

Give two reasons:

1. _____

2. _____

Activity 2: AI Bingo

Step 1: First take a look at this PPT as a class: **Introduction to AI and AI Bingo_AI-Ethics.pptx**

Step 2: After reviewing the introductory slides, pass out bingo cards. Bingo cards are available here: https://www.media.mit.edu/projects/ai-ethics-for-middle-school/overview/

Step 3: Club students into teams of 2. Teams must identify the prediction the AI system is trying to make and the dataset it might use to make that prediction. The first team to get five squares filled out in a row, diagonal, or column wins (or, for longer play, the first student to get two rows/diagonals/columns).

Step 4: After playing, have students discuss the squares they have filled out