R VISUALIZING DATA

Spring 2017

Goals for this lecture:

- Review constructing Data Frame, Categorizing variables
- Construct basic graph, learn how to recode variables

- Load built-in mtcars example Data Frame
- Recoding Variables
- Learn how to present data graphically in a bar graph, pie chart, histogram, or box plot.

- Learn how to load data from a file into a Data Frame

Review: Problem

For the given CS130 class information, create a data frame, **cs130DataFrame.R** that contains the following data

ID	Year	Age
0001	FR	18
0002	FR	18
0003	SR	22
0004	JR	22
0005	SO	19
0006	FR	19
0007	SR	23
0008	SO	19
0009	SR	22

Review: Continued

- Using the command **str(cs130DataFrame),** classify each variable ID, Year, Age as:
  - quantitative or qualitative
  - discrete, continuous, neither
  - nominal, ordinal, neither

Variable	Quantitative or Qualitative?	Discrete, continuous, neither?	Nominal, ordinal, neither?
ID			
Year			
Age			

Table function, barplots

- The **`table`** function will return a vector of table counts
- For instance

```
className = table(cs130DataFrame$Year)
```

will return a count of the number of FR, SO, JR, and SR's:

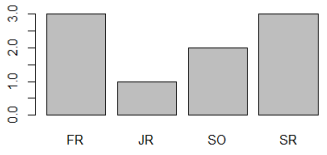
```
> className=table(cs130DataFrame$Year)
> className

FR JR SO SR
 3  1  2  3
>
```

- To graph this table, enter **`barplot(className)`**
- **Problems?**

Barplot: Categorical Data

- Need to create categorical (factor) data so desired order is maintained:




```
> classNameCategorical= factor(cs130DataFrame$Year,
                               levels=c("FR","SO","JR","SR"),
                               labels=c("Fr","So","Jr","Sr"))

Note: "levels" parameter is our desired order
      "labels" parameter is optional (defaults to levels)

> Barplot(table(classNameCategorical))  //Order now correct
```

mtcars Data Frame

R has a built-in data frame called **mtcars:** Data was extracted from the 1974 *Motor Trend* US magazine, and 11 aspects of automobile design and performance for 32 automobiles(1973–74 models).

[1]	mpg	Miles/(US) gallon
[2]	cyl	Number of cylinders
[3]	disp	Displacement (cu.in.)
[4]	hp	Gross horsepower
[5]	drat	Rear axle ratio
[6]	wt	Weight (1000 lbs)
[7]	qsec	1/4 mile time
[8]	vs	V/S (vshape or straight line engine)
[9]	am	Transmission (0 = automatic, 1 = manual)
[10]	gear	Number of forward gears
[11]	carb	Number of carburetors

```
                  mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4        21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag    21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710       22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive   21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
                                    ...
```


R: Useful functions

Copy **mtcars** to **tempMtcars** to protect mtcars data

```
> tempMtcars = mtcars
```

- Useful R functions
  - length(object) # number of variables
  - str(object) # structure of an object
  - class(object) # class or type of an object
  - names(object) # names
  - dim(object) # number of observations and variables
- In the console, call each function using **tempMtcars** as the object

Recoding Variables

Recode am variable as categorical variable amCategorical:

1. Default method with no parameters:

```
>tempMtcars$amCategorical = as.factor (tempMtcars$am)
```

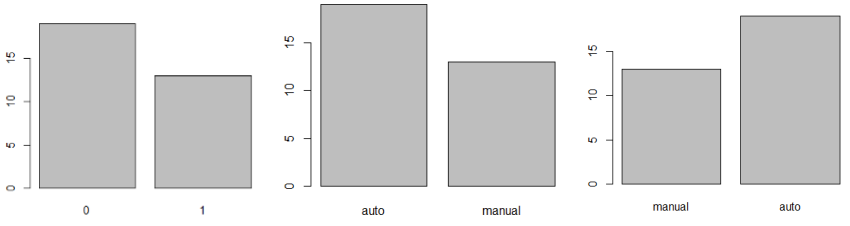
2. Method that allows levels and labels

```
> tempMtcars$amLabels = factor(tempMtcars$am,
levels=c('0','1'),labels=c("auto","manual"))

> tempMtcars$amOrdered = factor(tempMtcars$am,
levels=c('1','0'), labels=c("manual","auto"),ordered=TRUE)
```


Graphing Recoded Variables

```
> barplot(table(tempMtcars$amCategorical)
> barplot(table(tempMtcars$amOrdered))
> barplot(table(tempMtcars$amLabels))
```

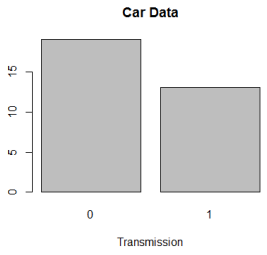


Bar Chart

http://statmethods.net/graphs/bar.html

- A **bar chart** or **bar graph** is a chart that presents grouped data with rectangular bars with lengths proportional to the values that they represent.
- function table returns a vector of frequency data

```
barplot(table(tempMtcars$amCategorical),
  main = "Car Data",
  xlab = "Transmission")
```



Recoding Variables

Create a new variable mpgClass where mpg<=25 is “low”, mpg>25 is “high”

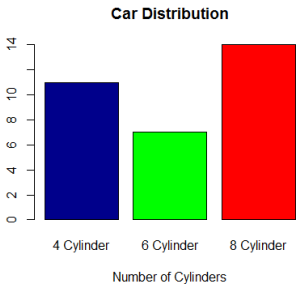
```
> tempMtcars$mpgClass[tempMtcars$mpg <= 25] = "low"
> tempMtcars$mpgClass[tempMtcars$mpg > 25] = "high"
> tempMtcars$mpgClass
[1] "low" "low" "low" "low" "low" "low" "low" "low"
[9] "low" "low" "low" "low" "low" "low" "low" "low"
[17] "low" "high" "high" "high" "low" "low" "low" "low"
[25] "low" "high" "high" "high" "low" "low" "low" "low"

> typeof(tempMtcars$mpgClass)
[1] "character"

> barplot(table(tempMtcars$mpgClass), main = "Car
Data", xlab="MPG")
```

Bar Chart

```
> barplot (table(tempMtcars$cyl),
main = "Car Distribution",
xlab = "Number of Cylinders",
col = c("darkblue", "green", "red"),
names.arg = c("4 Cylinder", "6 Cylinder", "8 Cylinder"))
```

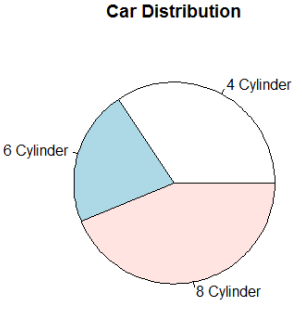


Pie Chart

http://statmethods.net/graphs/pie.html

- A pie chart is a circular graphical representation of data that illustrates a numerical proportion
- A pie chart gives a better visualization of the frequency of occurrence as a percent

```
> pie(table (tempMtcars$cyl),
labels = c("4 Cylinder", "6 Cylinder", "8 Cylinder"),
main="Car Distribution")
```

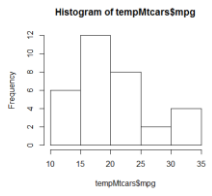


Histogram

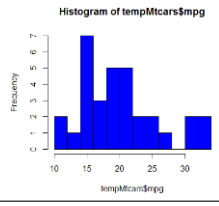
http://statmethods.net/graphs/density.html

- A histogram is a graphical representation of the distribution of numerical data
- Bin – are adjacent intervals usually of equal size
- Notice: breaks <> number of bins and breaks is just a suggestion and not guaranteed

```
>hist (tempMtcars$mpg)
```




```
> hist (tempMtcars$mpg, breaks=10,
col="blue")
```



Can also use main, xlab, ylab attributes as with barchar

Boxplots

http://statmethods.net/graphs/boxplot.html

- A boxplot is a way of graphically showing numerical data through quartiles
- A box-and-whisker plot is a boxplot that shows variability outside the upper and lower quartiles
- Quartile – the three points that divide the ranked data values into 4 equal sized groups

Quartile Definitions

- **first quartile/lower quartile/25th percentile/** Q_1
  - splits off the lowest 25% of data from the highest 75%
- **second quartile /median/50th percentile /** Q_2
  - cuts data set in half
- **third quartile/upper quartile/75th percentile /** Q_3
  - splits off the highest 25% of data from the lowest 75%
- **interquartile range /** IQR
  - IQR = Q_3 - Q_1

Problem Continued

- Using R, show the box-and-whisker plot and quantiles for
  - 6, 7, 19, 20, 42, 100, 200
  - 6, 7, 20, 100, 200

Paint Problem

- Let's put everything together
- A paint manufacturer tested two experimental brands of paint over a period of months to determine how long they would last without fading. Here are the results:

BrandA	**BrandB**	Report on the following
10	25	-Mean
20	35	-Median
60	40	
40	45	-Std Deviation
50	35	-Minimum
30	30	-Maximum

Paint Problem

1. Using Rstudio, create an R script on your desktop called paintDataFrame.R that creates a data frame paintData for the paint data.
    1. Name the variables brandAPaint and brandBPaint
2. Enter the data
3. Output the data frame
4. Save and run the script. Show me.

Paint Problem Continued

5. Compute and output the mean, median, std deviation, minimum, and maximum for each brand of paint

[1] "Brand A Mean = 35"
[1] "Brand A Median = 35"
[1] "Brand A Std Dev = 18.7082869338697"
[1] "Brand A Minimum = 10"
[1] "Brand A Maximum = 60"
[1] ""
[1] "Brand B Mean = 35"
[1] "Brand B Median = 35"
[1] "Brand B Std Dev = 7.07106781186548"
[1] "Brand B Minimum = 25"
[1] "Brand B Maximum = 45"

Paint Problem Continued

6. Output a Box-and-Whisker Plot for each brand of paint as follows. Get as close as possible. This isn't easy but give it a try.
7. What do the descriptive statistics tell us?
8. Which paint would you buy? Justify your answer.