NOVA School of Business & Economics

Shaping powerful minds

Research Topics

PhD in Economics | Finance – 2019/2020

Lecture notes - Financial Database Management

Pedro Pires

Course contents:

➢ Section I

- Overview of WRDS
- Accessing WRDS - Web query in detail
- Data management in STATA
  - ✓ Aggregating/Summarizing data
  - ✓ Compounding returns
  - ✓ Merging datasets
  - ✓ Fuzzy matching

➢ Section II

- Brief introduction to SAS and SQL
- Data management in SAS Studio
  - ✓ I/O (Input-Output) files
  - ✓ Querying data
  - ✓ Aggregating/Summarizing data
  - ✓ Merging datasets
  - ✓ Fuzzy matching

Financial Information by Source

Company Financials

- Balance Sheet, Income Statement, and Cash Flows
- Annual and Quarterly Frequencies
- Industrial and Financial Services (Banks) Format

→ **US:** Compustat North America
Worldwide: WorldScope, Factset Fundamentals, Compustat Global
Banks: Compustat Bank, BvD BankFocus
Private: BvD Orbis, BvD Amadeus

Financial Markets

- Stock Prices, Returns, Shares Outstanding and Volume
- Daily and Monthly Frequencies
- Index, CDS, Bonds, Interest Rates , Futures, Options, Swaps

→ **US:** CRSP, Compustat North America
Worldwide: DataStream, Bloomberg

Ownership

- Institutional and Mutual Fund Holdings
- Ownership of securities at the institution level (e.g. Fidelity)
- Ownership by mutual funds (e.g. Fidelity Magellan)

→ **US:** CRSP Mutual Funds, Thomson Reuters
Worldwide: FactSet, Lipper

Other

- Bond Ratings
- Executive Compensation
- Analyst Forecasts (e.g. future EPS, recommendations)
- IPOs, SEOs, M&A
- Syndicated Loans

→ **US:** Compustat Execucomp, Boardex
Worldwide: Bloomberg, IBES, Thomson One, Eikon, SDC, DealScan

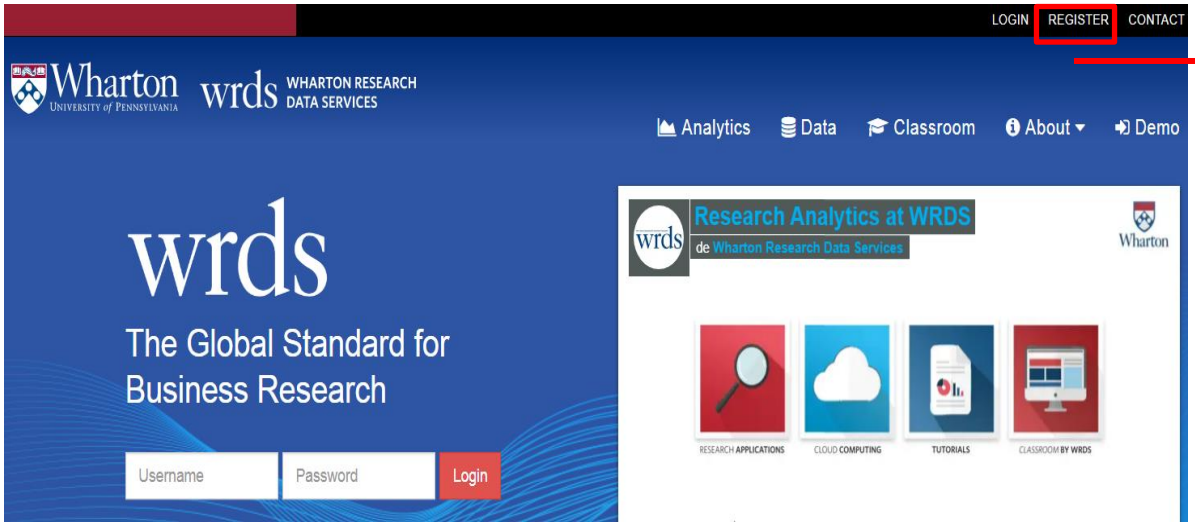
Contents coverage

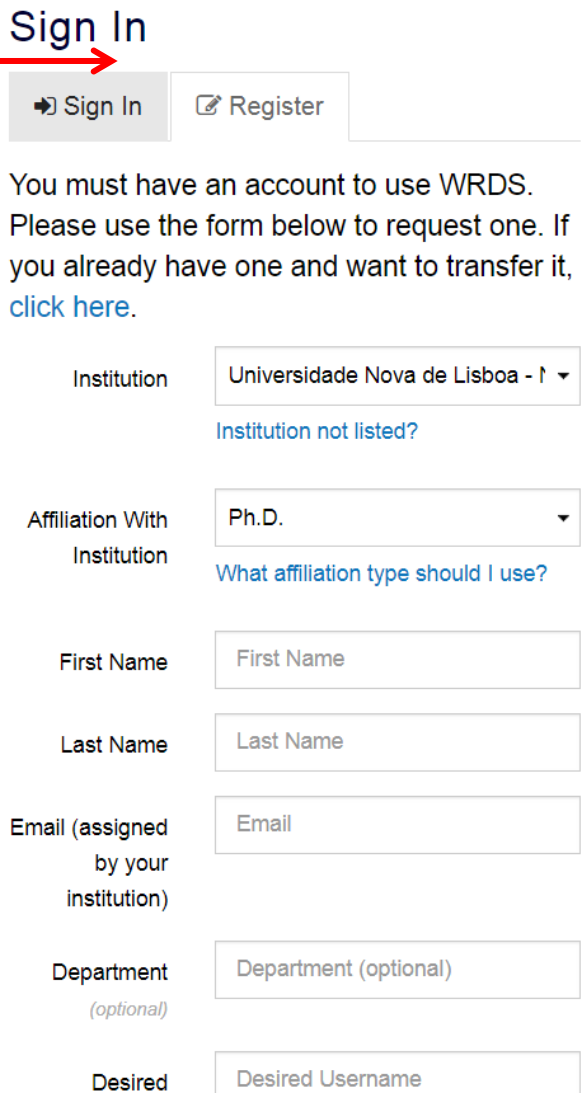
- **Company fundamentals**
  - Compustat North America
  - Compustat Bank
  - FactSet Fundamentals
  - Bureau Van Dijk
  - Thomson Reuters Eikon - DataStream/WorldScope
- **Segments: Geographical breakdown of sales**
  - FactSet Revere
  - Compustat Historical Segments
- **Issuer credit ratings**
  - Compustat S&P Ratings
- **Analyst estimates**
  - Thomson Reuters IBES
- **Executive compensation**
  - Compustat Execucomp – Executive Compensation
  - Boardex
- **M&A and Equity/Debt underwriting**
  - Thomson Reuters Eikon
  - Thomson Reuters SDC

Contents coverage

- **Stocks**
  - CRSP
  - Compustat Supplemental Short Interest File
- **Database management**
  - CRSP-Compustat Merged (CCM)
  - Merging CRSP and Compustat by CUSIP
  - Fuzzy Matching
- **Institutional ownership and mutual fund holdings**
  - CRSP Mutual Funds
  - Thomson Reuters - Mutual Fund Holdings
  - Lipper
  - Thomson Reuters Institutional Holdings
  - FactSet LionShares Ownership
- **Syndicated loans**
  - Thomson Reuters LPC DealScan
- **Others:**
  - Eventus
  - WorldBank WDI
  - IMF Macroeconomic and financial data

Register for a WRDS Account



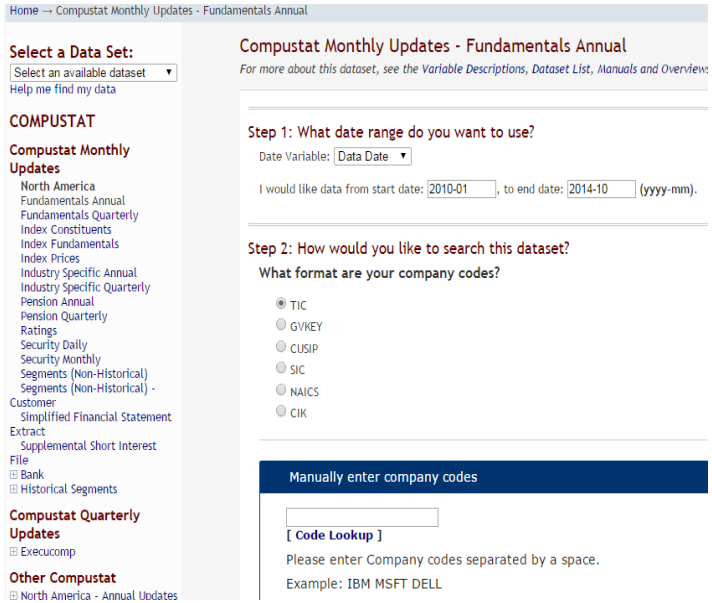


Three ways of using WRDS:

- 3 complementary ways of using WRDS
  - **Web Query**
  - UNIX
    - Secure SHell (SSH) Client (please install the **WinSCP** client)
  - **SAS Studio**
  - PCSAS (SAS)
- You can download the same exact data using any of the alternative methods. Different strengths.
- **In this course we will focus on (1) managing data obtained through Web Queries (using STATA); and (2) SAS Studio.**

Accessing WRDS

(1) Web Query



- Just need an Internet Browser
- Simple Point & Click query interface
- Access to search and help tools
- Choice of different output formats: txt, csv, xlsx, sas, stata, spss
- Instantaneously save output to your PC

(2.1) Install WinSCP for Windows or an alternative for Mac

winscp.net/eng/download.php

WinSCP 5.15 is a major application update. New features and enhancements include:

- Files can be optionally encrypted when storing them on SFTP server.
- Local UNC paths can be browsed.
- *Compare Files* extensions.
- Dark theme.
- Files can be copied to the clipboard.
- Coloring files in file panels based on a file mask.
- Improved incremental search in file panels.
- Support OpenSSH AES-256-CTR-encrypted keys.
- Improvements to directory synchronization.
- List of all changes.

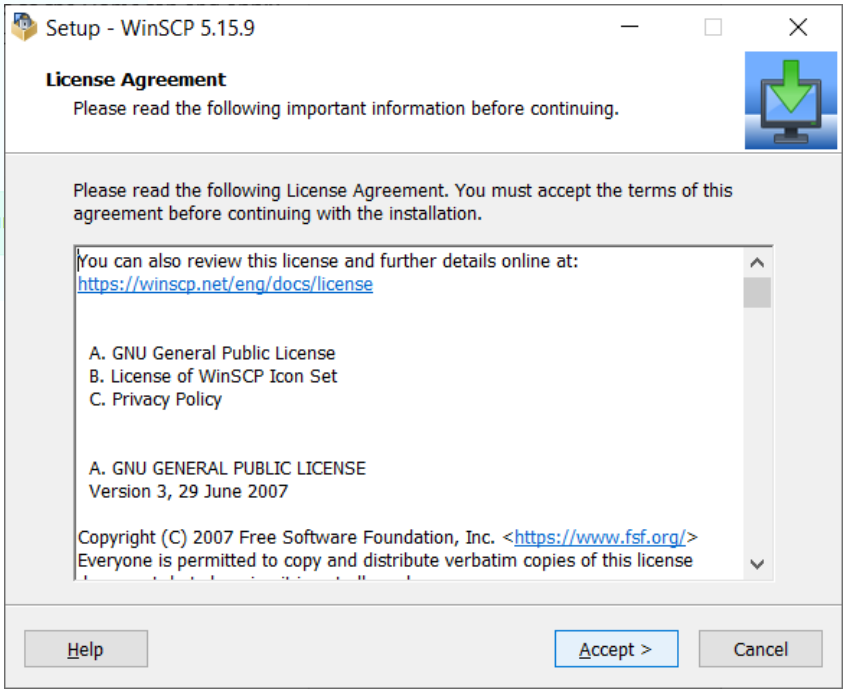
DOWNLOAD WINSCP 5.15.9 (9.4 MB)

Get it from Microsoft

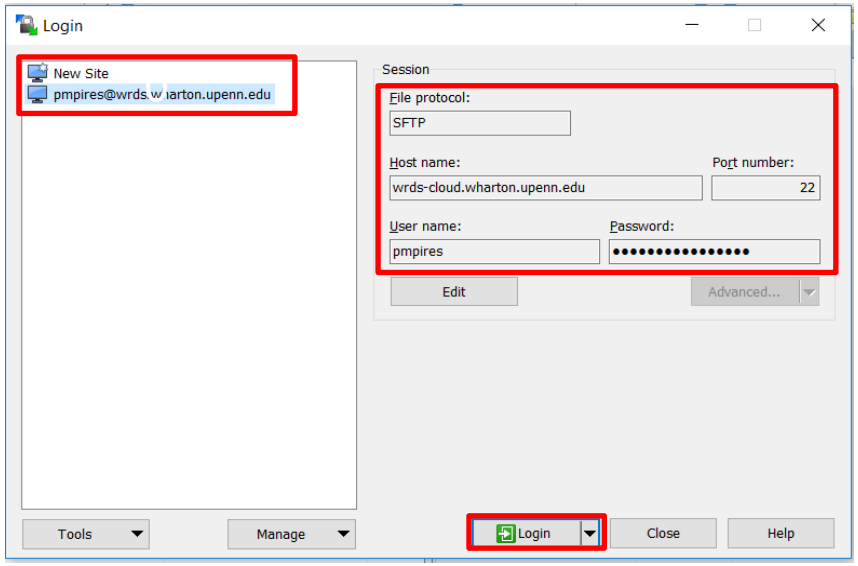
OTHER DOWN

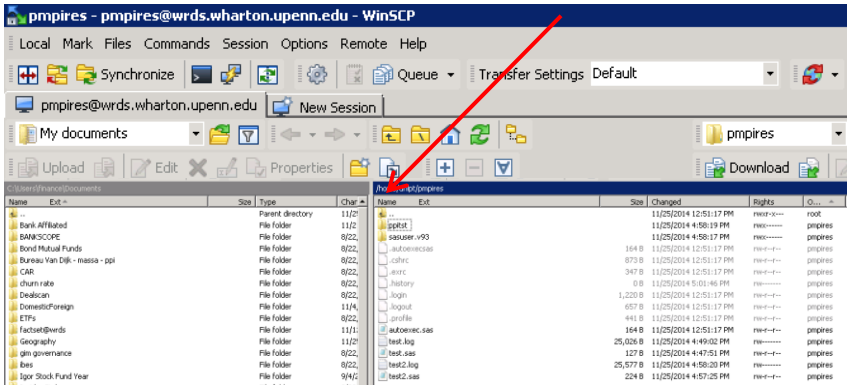
1,495,952 downloads since 2019-12-05

What is this?



(2.1) UNIX SAS - SSH Client



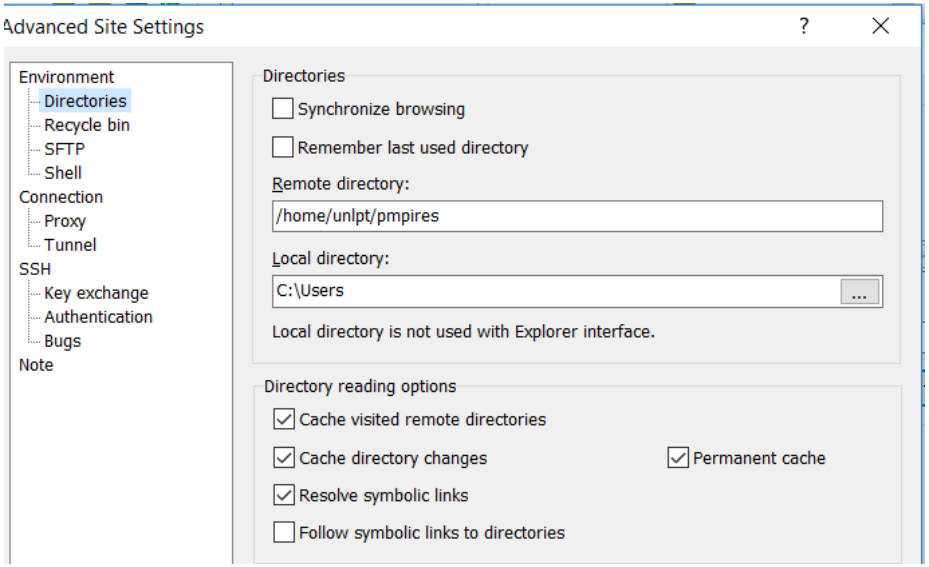


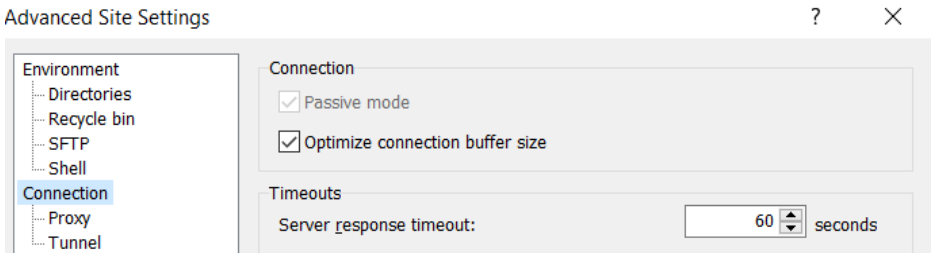
- Needs a SSH Client (*e.g.* WinSCP)

→ Host: *wrds-cloud.wharton.upenn.edu*

- Programming in SAS (No license needed) using a text editor (*e.g.* Notepad)
- Up to 16 GB RAM per process (Max 5 jobs)
- 10 GB individual storage space in directory /home/unlpt/**username**
- 4 TB temporary disk space located in /sastemp (deleted in 1 week or full)

(2.1) UNIX SAS - SSH Client

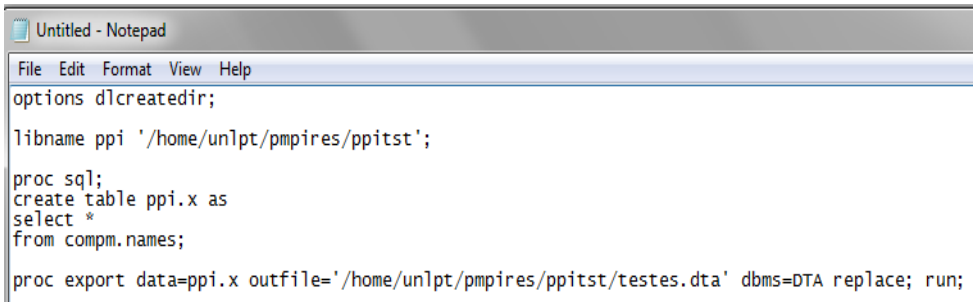


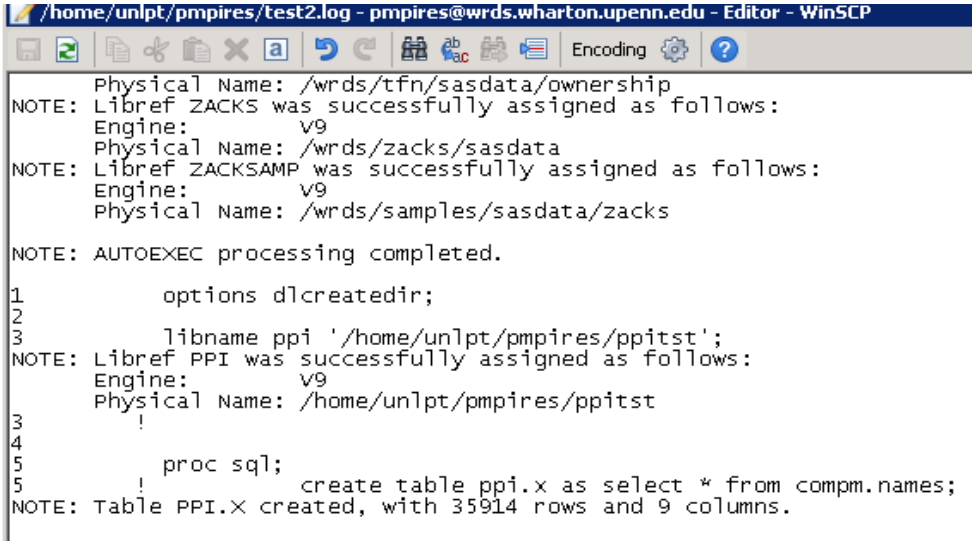


Advanced Settings

- Remote directory: /home/unlpt/username
- Uncheck “Remember last used directory”
- Connection: Timeouts – 60 seconds

(2.1) UNIX SAS - SSH Client (Programming in SAS)



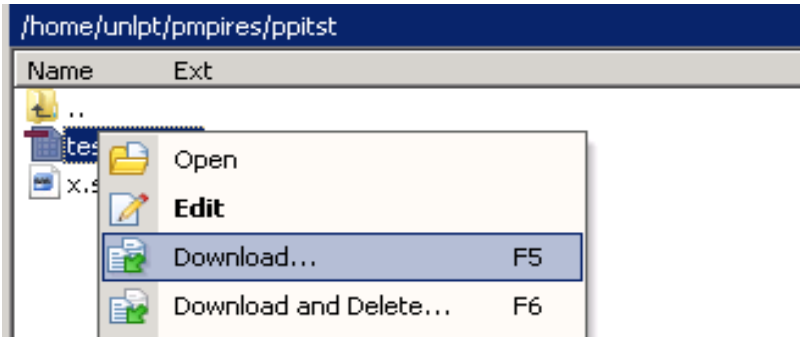


- Write a SAS progam using a text editor
- Upload the file to /home/unlpt/**username**

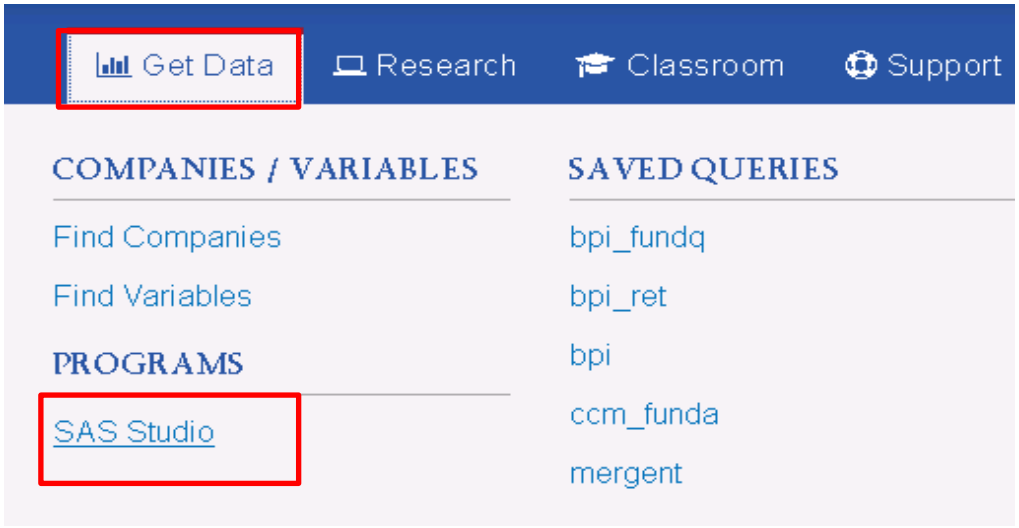
  (make sure the file extension is .sas)
- Run the program in the command window

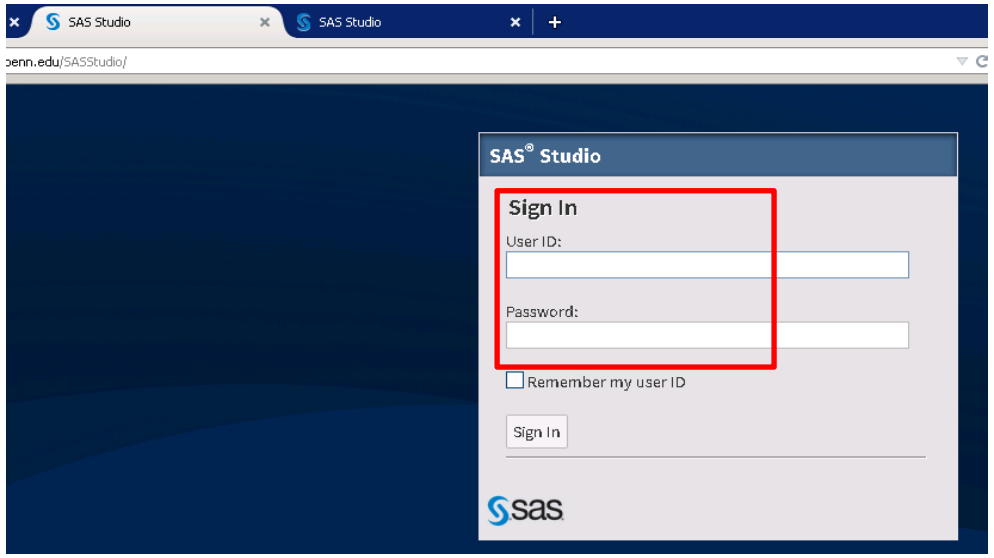
  (qsas **filename**)
- Download the file stored in the WRDS server

  to your pc



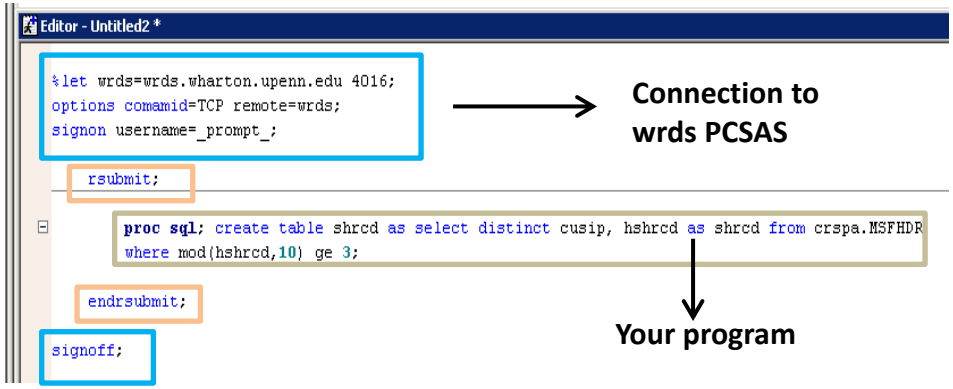
(2.2) SAS Studio

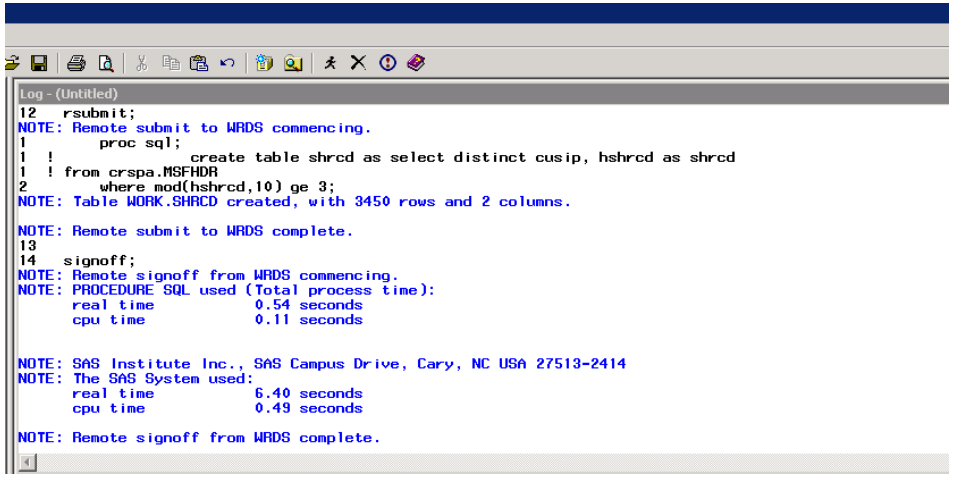




- SAS studio is a **friendly visual interface** to the SAS server running on WRDS
- All standard browsers are supported (Firefox, Google Chrome, Internet Explorer, Safari)

(3) PCSAS





- Needs a SAS License (software installed on your PC)
- Remote access to the WRDS server

→ *wrds-cloud.wharton.upenn.edu*

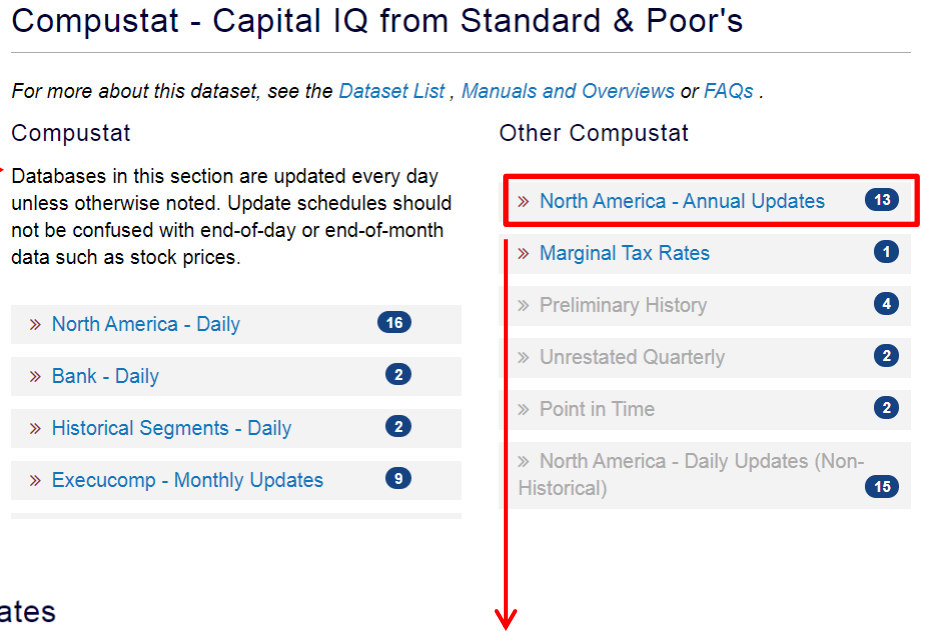
- Easier to edit and debug programs

Web Query Characteristics

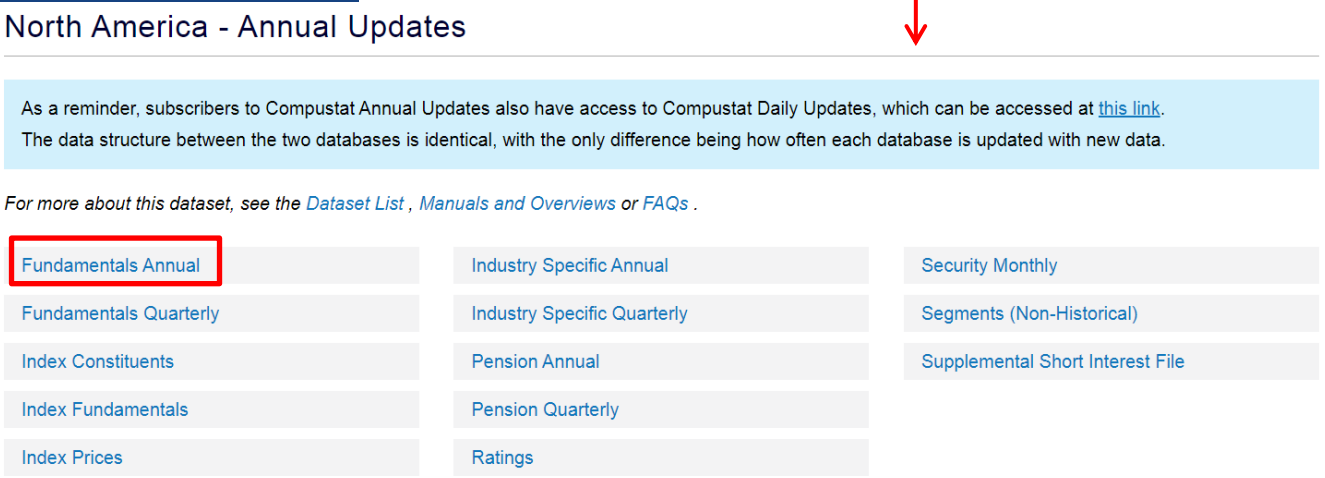
- Select Data Source
  - **Select Database** (e.g. COMPUSTAT)
  - **Select Dataset** (e.g. COMPUSTAT/North America/Fundamentals Annual)
- Uniform Query Interface (whatever the database you use in WRDS)
  - **Step 1: Select Date Range**
  - **Step 2: Select Companies/Entities** (e.g. IBM, MSFT, XOM)
  - **Step 3: Select Variables** (e.g. return, price, total assets, net income)
  - **Step 4: Select output format** (e.g. txt, xlsx, csv, stata, sas)
- Download file to your PC

COMPUSTAT (Web Query)

➢ Select Dataset

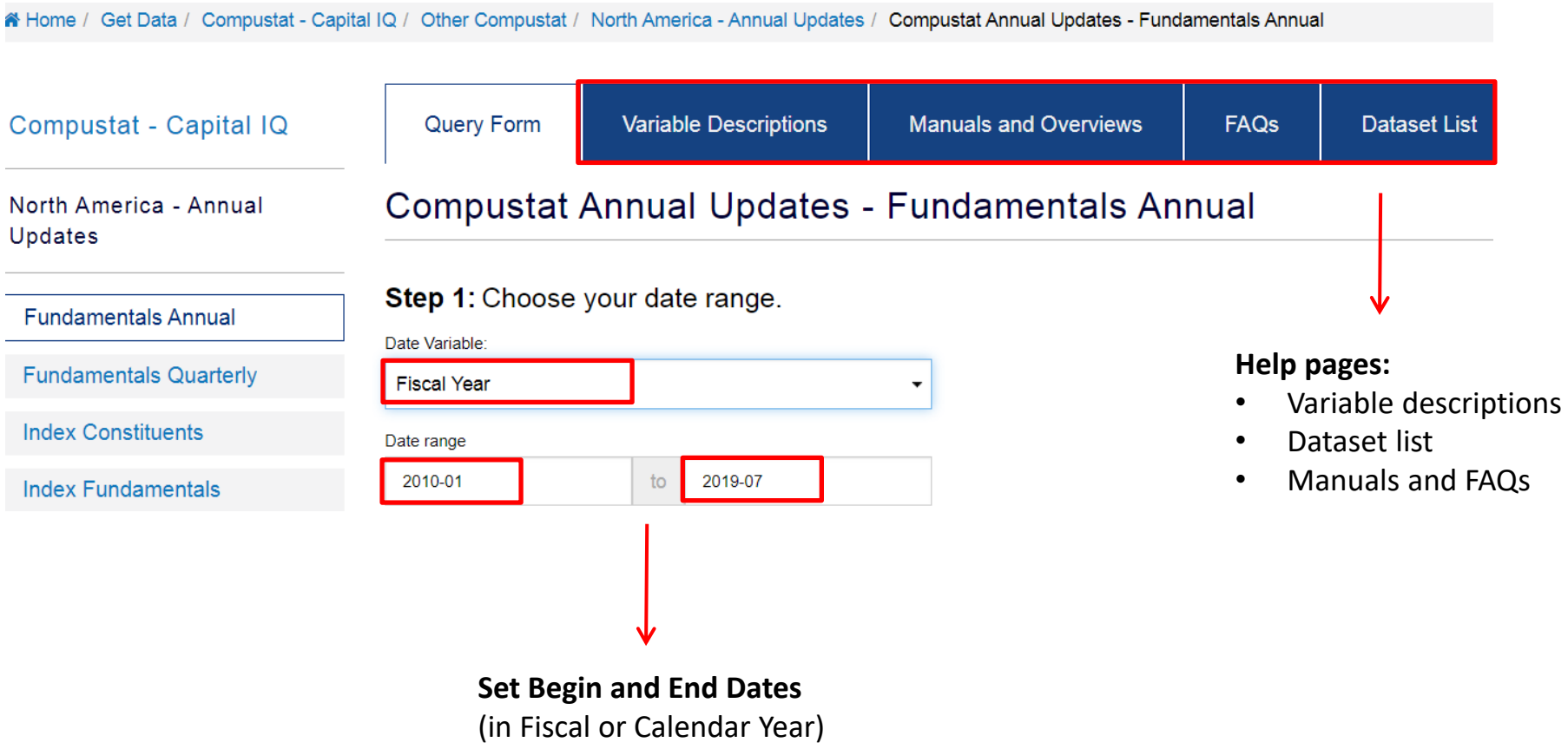


Point & Click



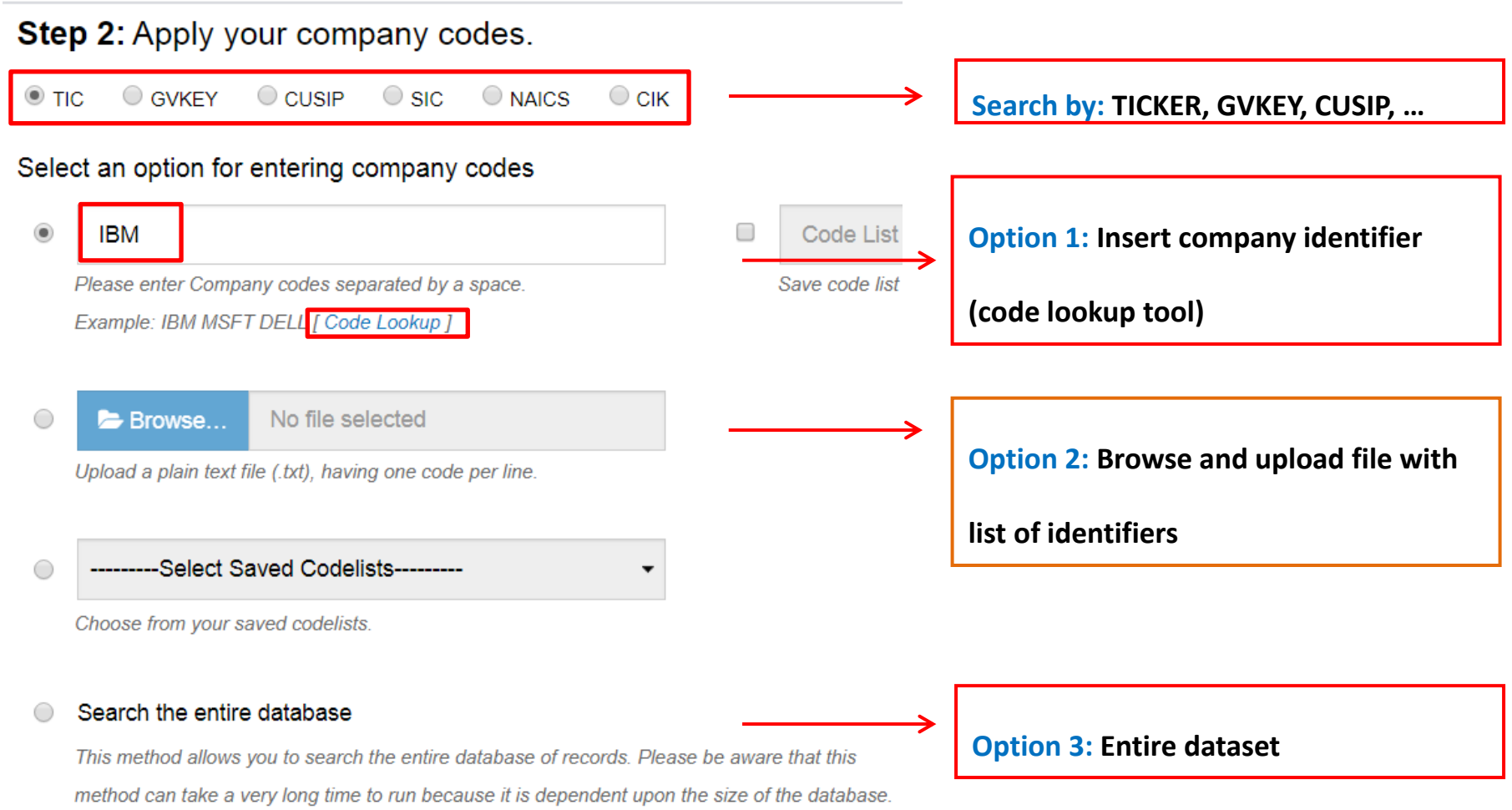
COMPUSTAT (Web Query)

➢ Step 1: Select Date Range



COMPUSTAT (Web Query)

➢ Step 2: Select Companies / Entities



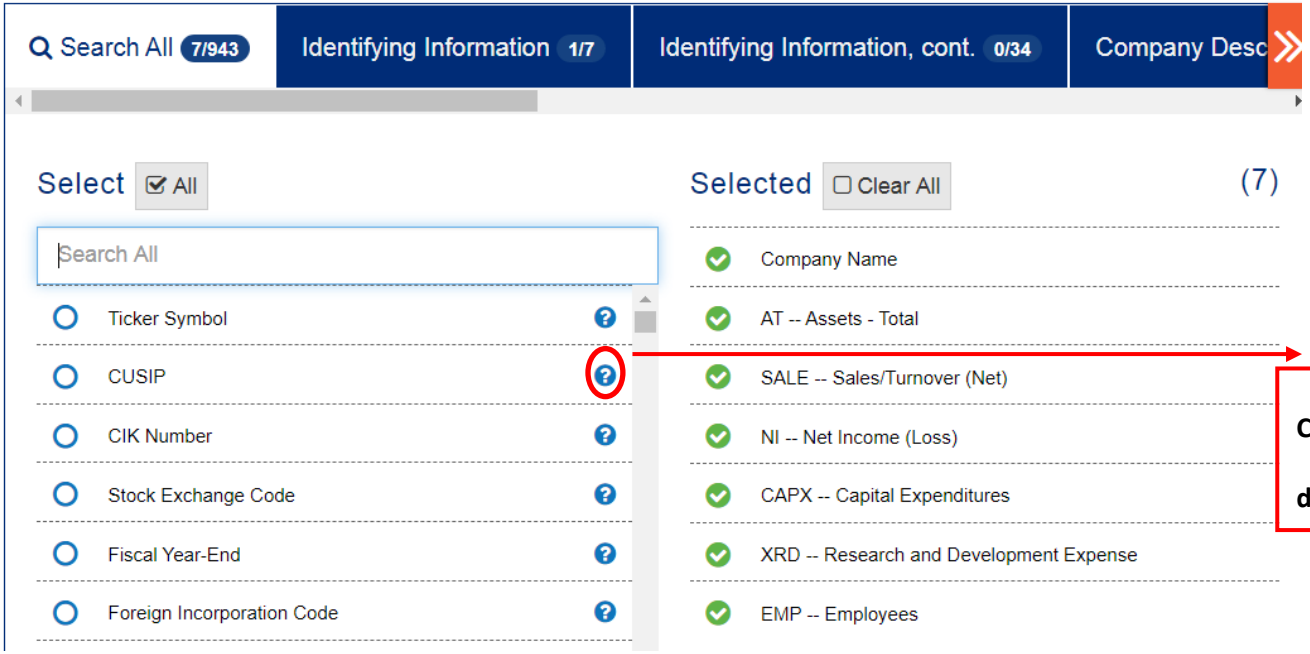
COMPUSTAT (Web Query)

➢ Step 3: Select Variables

Screening Variables

Several screening variables are pre-selected to produce one record per GVKEY-DATADATE pair, while keeping the vast majority of records. Examples of excluded rows include those with restated data, different views of the same data (pro forma, pre-FASB). Click on each variable for a more detailed explanation.

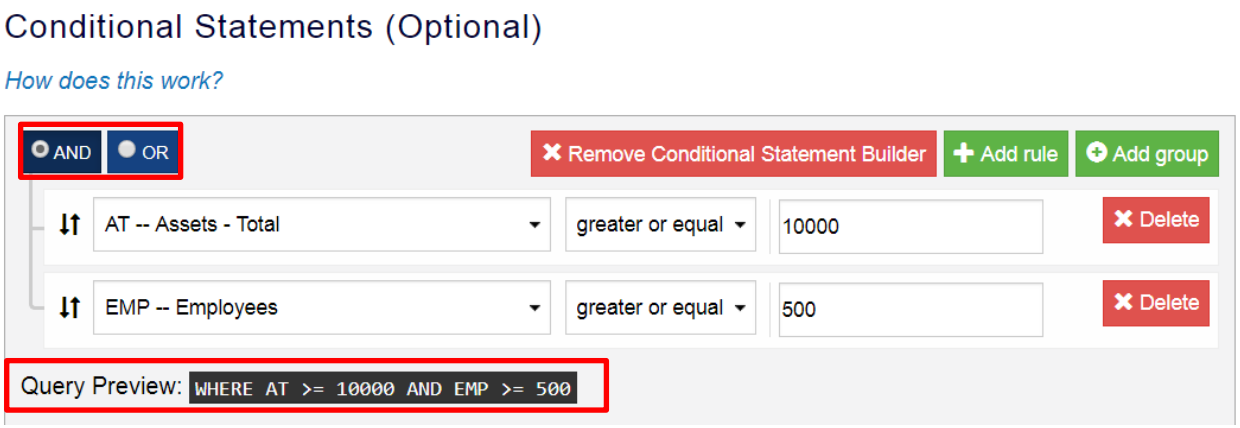
Consolidation Level	☑ C	☐ N	☐ R	☐ P	☐ D	☑ Output
Industry Format				☑ INDL	☐ FS	☑ Output
Data Format	☑ STD	☐ SUMM_STD		☐ PRE_AMENDS	☐ PRE_AMENDSS	☑ Output
Population Source				☑ D	☐ I	☑ Output
Currency				☑ USD	☐ CAD	☑ Output
Company Status				☑ Active	☑ Inactive	☑ Output



Click on the question mark button to get detailed information on each item

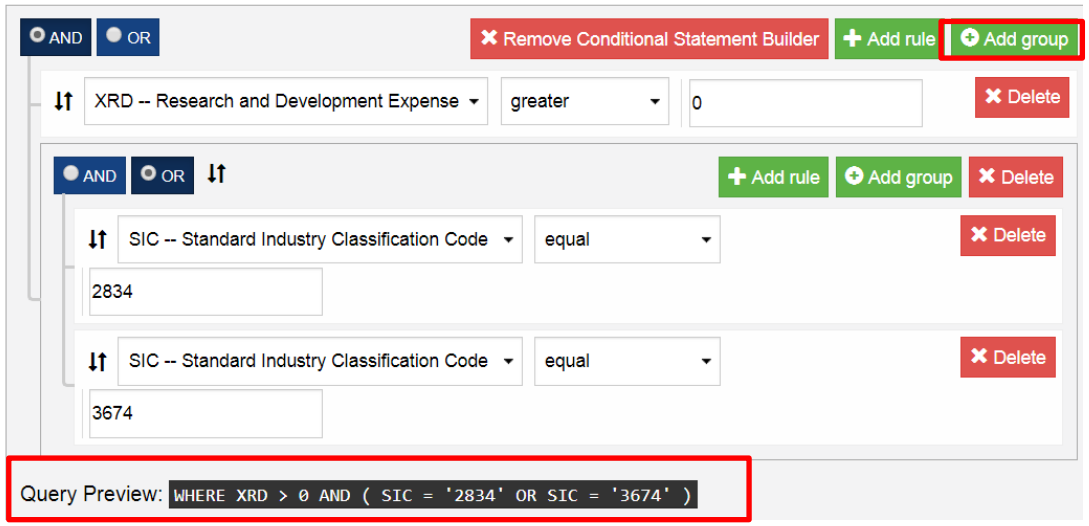
COMPUSTAT (Web Query)

➢ Step 2b: Conditional Statements



Example 1

- **Assets >= 10000 millions**
- **Employees >= 500**



Example 2

- **R&D expense > 0**
- **Standard Industry Classification Code (SIC) = 2834 OR 3674**

COMPUSTAT (Web Query)

➢ Step 4: Select Output Format

Step 4: Select query output.

Select the desired format of the output file. For large data requests, select a compression type to expedite downloads. If you enter your email address, you will receive an email that contains a URL to the output file when the data request is finished processing.

Output Format

- ○ fixed-width text (*.txt)
- ○ comma-delimited text (*.csv)
- ○ Excel spreadsheet (*.xlsx)
- ○ tab-delimited text (*.txt)
- ○ HTML table (*.htm)
- ○ SAS Windows_32 dataset (*.sas7bdat)
- ○ SAS Windows_64 dataset (*.sas7bdat)
- ○ SAS Solaris_64 dataset (*.sas7bdat)
- ○ dBase file (*.dbf)
- ◉ STATA file (*.dta)
- ○ SPSS file (*.sav)

Compression Type

- ◉ None
- ○ zip (*.zip)
- ○ gzip (*.gz)

Date Format

- ◉ YYMMDDn8. (e.g. 19840725)
- ○ DATE9. (e.g. 25JUL1984)
- ○ DDMMYY6. (e.g. 250784)
- ○ MMDDYY10. (e.g. 07/25/1984)
- ○ DDMMYY10. (e.g. 25/07/1984)
- ○ YYMMDDs10. (e.g. 1984/07/25)

E-Mail Address *(Optional)*

E-mail | Edit Preferences

Custom Field *(Optional)*

☐ Save this query to myWRDS

Query Name

Submit Query

COMPUSTAT (Web Query)

- Save File to your PC

Data Request Summary

Your output is complete. Click on the link below to open the output file.

ebde6dfa316fcf54.dta (4 KB, 9 observations 17 variables)

Download instructions

Internet Explorer and Firefox users... Right-click and select "Save Target As..."

Citation instructions

To cite this data use the following format:

Wharton Research Data Services. "Compustat Annual Updates " wrds.wharton.upenn.edu, accessed 01/23/2020.

Data Request ID	ebde6dfa316fcf54
Libraries/Data Sets	compa/funda /
Frequency/Date Range	ann / 01Jan2010 - 31Jul2019
Search Variable	TIC
Input Codes 1 item(s)	IBM
Conditional Statements	n/a
Output format/Compression	dta /
Variables Selected	CONM AT NI SALE XRD CAPX EMP
Extra Variables and Parameters Selected	C INDL STD

- **Save File when output is complete**
- **Right-Click and Select "Save Target (Link) As"**
- **Don't Left-Click the URL for large files**

COMPUSTAT (Web Query)

Example: IBM

gvkey	datadate	fyear	tic	at	capx	emp	ni	sale	xrd
006066	31dec2010	2010	IBM	113452	4185	426.751	14833	99871	5720
006066	31dec2011	2011	IBM	116433	4108	433.362	15855	106916	5990
006066	31dec2012	2012	IBM	119213	4081	434.246	16604	104507	6034
006066	31dec2013	2013	IBM	126223	3623	431.212	16483	99751	5959
006066	31dec2014	2014	IBM	117532	3740	379.592	12022	92793	5437
006066	31dec2015	2015	IBM	110495	3579	377.757	13190	81741	5247
006066	31dec2016	2016	IBM	117470	3567	380.3	11872	79920	5751
006066	31dec2017	2017	IBM	125356	3229	366.6	5753	79139	5787
006066	31dec2018	2018	IBM	123382	3396	350.6	8728	79591	5379

Consolidated Statement of Financial Position
International Business Machines Corporation and Subsidiary Companies

($ in millions except per share amounts)

At December 31:	Notes	2018	2017
Assets			
Total assets		$ 123,382	$ 125,356

Employees
Variable Name = EMP

This item represents the number of people employed by the company and its consolidated subsidiaries in thousands.

Company fundamentals

- **Compustat North America** [Coverage: North America | Public]
- **Compustat Global [Not Subscribed]** [Coverage: Global | Public]
- **Compustat Bank** [Coverage: US Banking Institutions]
- **FactSet Fundamentals** [Coverage: Global | Public + some Private]
- **Bureau Van Dijk – Orbis** [Coverage: Global | Public + Private]
- **Bureau Van Dijk – Other [Not Subscribed]**
  - **Amadeus** [Coverage: Europe | Public + Private]
  - **BankFocus** [Coverage: Global | Banks | Public + Private]
  - **Isis** [Coverage: Global | Insurance companies | Public + Private]
  - **Osiris** [Coverage: Global | Major companies | Public + Private]
- **Outside WRDS…**
  - **DataStream/WorldScope** [Coverage: Global | Public]
  - **Bloomberg** [Coverage: Global | Public] *

* Not covered in this course

➢ Financial reports for worldwide firms – Comparative Analysis

Rui Dai, 2012, "International Accounting Databases on WRDS: Comparative Analysis", Wharton Research Data Services, University of Pennsylvania.

https://wrds-www.wharton.upenn.edu/documents/692/March_16_2012_international_0HjpzsR.pdf

1) **Compustat Global** features greater coverage of large companies in more developed countries and provides a wider range of accounting data items than any other databases;

2) **BvD Osiris** offers lesser variety of accounting data items, but it also contains a higher number of small firms from developing countries;

3) **FactSet Fundamentals** Database provides a balance in the firm size and quantity of accounting items with a reasonable geographical coverage.

✓ Compustat North America Fundamentals

➢ Understanding the Data:

- Coverage: North America (US, Canada)
- Over 31,000 public companies (14,650 active)
- Includes ADR companies
- Company fundamental annual data beginning in 1950
- Monthly market data beginning in 1962
- Fiscal and Calendar presentations of the data

Compustat North America Fundamentals

- GVKEY: Global Company Key (GVKEY) is a unique identifier that represents each company in COMPUSTAT.
- DATADATE: Data Date represents the reporting date for a data record
- {GVKEY; DATADATE} DATADATE is often combined with GVKEY to make company records unique
- COMMON IDENTIFIERS: CUSIP, SEDOL, ISIN, TICKER
- CURRENCY: Data expressed in Local Currency.

Compustat North America Fundamentals

- Compustat has both active and inactive companies. Field DLRSN indicates the reason a company became inactive.

Research Company Reason for Deletion codes and descriptions.

Code	Reason for Deletion
01	Acquisition or merger
02	Bankruptcy
03	Liquidation
04	Reverse acquisition (1983 forward)
05	No longer fits original format (1978 forward)
06	Leveraged buyout (1982 forward)
07	Other (no longer files with SEC among other possible reasons), but pricing continues
09	Now a private company
10	Other (no longer files with SEC among other reasons)
11	Agency governing settlement of securities' trading inactivated the issue's Local Settlement Code because the issue matured, expired or was called. No successor settlement code was established.
12	Agency governing settlement of securities' trading inactivated the issue's Local Settlement Code. A successor settlement code was established; issue was changed for another, as in a par value change.
13	Price source for the SEDOL was no longer available. Issue now identified under different SEDOL.
14	Fully paid issue was replaced or partly paid issue was replaced by a subsequent installment. Successor settlement code was established.

Compustat North America Fundamentals

- Compustat is mainly used for its Fundamentals data…
- But Compustat also has other useful datasets:
  - Index Constituents
  - Ratings
  - Security Daily/Monthly
  - Bank Fundamentals
- Compustat is also the vendor of other databases: EXECUCOMP and CAPITAL IQ

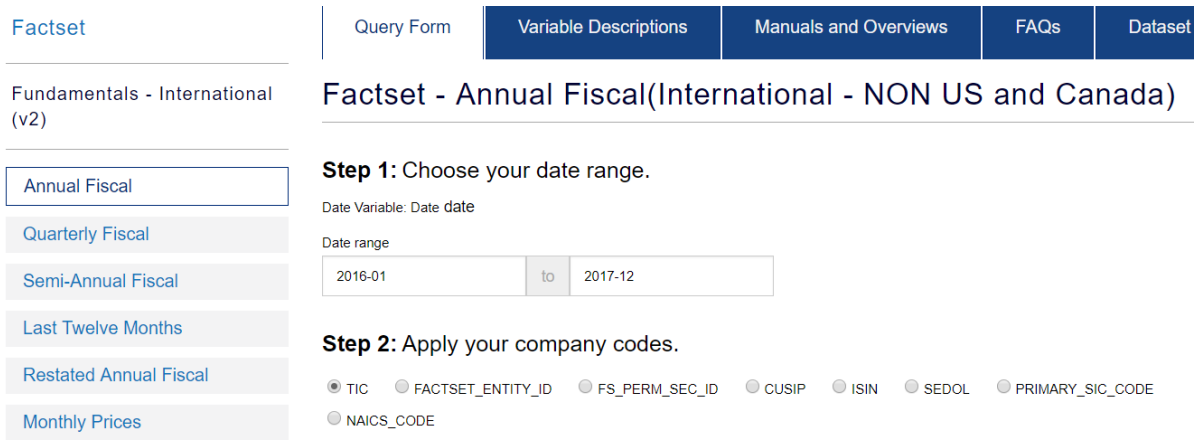
✓ Compustat Bank Fundamentals

- Compustat Bank Fundamentals contains financial data of the biggest and most important banks in the US since 1950.
- In accounting there is a difference between balance sheets and income statements for industrial companies and Banks.
- Compustat Industrial Annual includes both industrial companies and banks, with banks financial reports represented as Industrial statements.
- Compustat Bank annual features bank related items on the balance sheet and income statement.

Web Query in detail (FactSet Fundamentals)

✓ FactSet Fundamentals * (currently, not subscribed)

- FactSet purchased a copy of the Thomson Worldscope Database in April 2008, with historical information for over 43,000 companies dating back to 1980.
- FactSet covers more than 73,000 public and private companies in the Americas, Europe/Africa and Asia-Pacific regions.
- There are over 5,000 private companies within the Fundamentals database.



Web Query in detail (FactSet Fundamentals)

FactSet Fundamentals – Global coverage

Example: Find the sales and total assets for firms incorporated in Portugal with financial reports during 2017.

Dataset: Fundamentals - International (v2)

	factset_en~d	entity_proper_name	isin	iso_countr~p	date	ff_sales	ff_assets
1	003HW5-E	Banco Comercial Portugu□s SA	PTBCP0AM0015	PT	31dec2017	2972.6419999999998	72041.625
2	003J38-E	PHarol SGPS SA	PTPTC0AM0009	PT	31dec2017	0	269.10650399999997
3	0060F0-E	Litho Formas SA	PTLIT0AE0005	PT	31dec2017	.24729746	2.8972460999999998
4	006WFR-E	Sport Lisboa e Benfica-Futebol SA	PTSLB0AM0010	PT	30jun2017	128.23500000000001	507.71800000000002
5	006Y18-E	Altri SGPS SA	PTALT0AE0002	PT	31dec2017	656.05516	1210.0885069999999
6	05G6C7-E	Banco BPI SA	PTBPI0AM0004	PT	31dec2017	834.38099999999997	29640.208999999999
7	05G7DP-E	Caixa Geral de Dep□sitos SA		PT	31dec2017	3419.9884419999998	93247.913895999998
8	05HGD7-E	BRISA - Auto-estradas de Portugal SA	PTBRI0AM0000	PT	31dec2017	670.91800000000001	3338.703
9	05HH9S-E	EDP-Energias de Portugal SA	PTEDP0AM0009	PT	31dec2017	15745.986999999999	42075.048999999999
10	05HJCN-E	The Navigator Co. SA	PTPTI0AM0006	PT	31dec2017	1636.8344360000001	2439.1352959999999
11	05HZ0L-E	Jer□nimo Martins SGPS SA	PTJMT0AE0001	PT	31dec2017	16276.15	6441.8149999999996
12	05HZGB-E	Semapa Sociedade de Investimento e Gest□o SGPS SA	PTSEM0AM0004	PT	31dec2017	2164.6532779999998	3986.987071
13	05HZY4-E	Sonaecom SGPS SA	PTSNC0AM0006	PT	31dec2017	139.559562	1105.608849
14	05J0PJ-E	Mota-Engil SGPS SA	PTMEN0AE0005	PT	31dec2017	2597.2939999999999	4614.0889999999999
15	05J0WJ-E	SAG GEST - Solucoes Automovel Globais SGPS SA	PTSAG0AE0004	PT	31dec2017	619.70749999999998	550.12850000000003
16	05J1NL-E	Novabase SGPS SA	PTNBA0AM0006	PT	31dec2017	139.72200000000001	184.226
17	05J1WK-E	Impresa SGPS SA	PTIPR0AM0000	PT	31dec2017	200.67891299999999	390.03228300000001
18	05J32F-E	Inapa-Investimentos, Participa□□es e Gest□o SA	PTINA0AP0008	PT	31dec2017	894.69899999999996	670.71000000000004
19	05J3G3-E	COFINA SGPS SA	PTCFN0AE0003	PT	31dec2017	91.057924	113.389072
20	05J3PW-E	Corticeira Amorim SGPS SA	PTCOR0AE0006	PT	31dec2017	701.60900000000004	869.40700000000004
21	05J469-E	Toyota Caetano Portugal SA	PTSCT0AP0018	PT	31dec2017	390.03355299999998	298.48067099999997
22	05J5HG-E	Ibersol SGPS SA	PTIBS0AM0008	PT	31dec2017	448.329094	436.95330200000001
23	05J5XW-E	Estoril Sol SGPS SA	PTESO0AM0000	PT	31dec2017	107.657377	150.04283000000001
24	05J6WR-E	Parp□blica-Participa□□es P□blicas SGPS SA		PT	31dec2017	932.88999999999999	13902.038
25	05J6ZZ-E	Sonae SGPS SA	PTSON0AM0001	PT	31dec2017	5710.1519360000002	5604.6522359999999

Web Query in detail (FactSet Fundamentals)

FactSet Fundamentals – Global coverage

Example: Find the sales and total assets for firms incorporated in Portugal with financial reports during 2017.

Dataset: Fundamentals - International (v2)

	factset_en~d	entity_proper_name	isin	iso_countr~p	date	ff_sales	ff_assets
26	05J8C5-E	Sonae Capital, SGPS, SA	PTSNP0AE0008	PT	31dec2017	177.15676500000001	516.12660900000003
27	05JC5Z-E	Futebol Clube do Porto SAD	PTFCP0AM0008	PT	30jun2017	97.878227999999993	378.42498799999998
28	05JCQ0-E	Glintt-Global Intelligent Technologies SGPS SA	PTPAD0AM0007	PT	31dec2017	71.006651000000005	166.02282400000001
29	05JCZX-E	Cia Portuguesa de Amidos SA	PTCPA0AP0006	PT	31dec2017	33.038682600000001	19.051366389999998
30	05JFP7-E	Sporting Clube de Portugal Futebol SAD	PTSCP0AM0001	PT	30jun2017	61.402999999999999	316.49700000000001
31	05JHD4-E	Compta Equipamentos e Servicos de Inform□tica SA	PTCOM0AE0007	PT	31dec2017	29.187664000000002	25.761562000000001
32	05JHGP-E	Reditus SA	PTRED0AP0010	PT	31dec2017	41.139491	171.18457000000001
33	05JHP5-E	Imobili□ria Construtora Gr□o-Par□ SA	PTGPA0AP0007	PT	31dec2017	.029999999999999999	42.443810079999999
34	05JMCX-E	VAA Vista Alegre Atlantis SGPS SA	PTVAA0AE0001	PT	31dec2017	84.930000000000007	177.816
35	05JMR8-E	CIPAN Companhia Industrial Produtora de Antibi□ticos SA	PTCPN0AE0002	PT	31dec2017	19.262919279999998	27.820152570000001
36	05JMXS-E	Lisgr□fica - Impress□o e Artes Gr□ficas SA	PTLIG0AE0002	PT	31dec2017	16.969491999999999	16.603259000000001
37	05KB5L-E	Galp Energia SGPS SA	PTGAL0AM0009	PT	31dec2017	15193	12358
38	05KVVP-E	CTT - Correios de Portugal, SA	PTCTT0AM0001	PT	31dec2017	703.95816300000001	1608.765392
39	05KXTT-E	Luz Sa□de SA	PTEPT0AM0005	PT	31dec2017	481.924599	659.21839199999999
40	05L01N-E	Grupo Media Capital SGPS SA	PTGMC0AM0003	PT	31dec2017	126.91099199999999	301.24435699999998
41	05L61V-E	Conduril Engenharia SA	PTCDU0AE0003	PT	31dec2017	146.80751699999999	421.59400399999998
42	05MBL3-E	Empresa de Desenvolvimento e Infra-estruturas do Alqueva SA		PT	31dec2017	28.180444999999999	812.53116
43	05PHYY-E	Sonae Sierra SGPS SA	PTSOA0AE0008	PT	31dec2017	181.411	2322.0839999999998
44	05WS3M-E	Martifer SGPS	PTMFR0AM0003	PT	31dec2017	176.871771	376.10044599999998
45	06BF22-E	Sonagi SGPS SA	PTSNG0AM0007	PT	31dec2017	7.2181749999999996	117.505684
46	06BXQX-E	Sonae Ind□stria SGPS SA	PTS3P0AM0025	PT	31dec2017	230.97823099999999	401.03227099999998
47	06PHWQ-E	Redes Energeticas Nacionais SGPS SA	PTREL0AM0008	PT	31dec2017	716.14700000000005	5429.6030000000001
48	073TQR-E	NOS SGPS SA	PTZON0AM0006	PT	31dec2017	1548.9349999999999	2967.067
49	079MYN-E	Ramada Investimentos e Ind□stria SA	PTFRV0AE0004	PT	31dec2017	156.84656000000001	314.11801800000001
50	07KFYX-E	Sumol + Compal SA	PTSML0AM0009	PT	31dec2017	356.13121960000001	646.21593351000001
51	09ZPVB-E	Teixeira Duarte SA	PTTD10AM0000	PT	31dec2017	1035.6379999999999	2294.3589999999999
52	0BDJCW-E	Brisa Concess□o Rodovi□ria SA		PT	31dec2017	578.59822199999996	2974.026269

✓ BvD – Bureau Van Dijk

Orbis

- Flagship product, information on over 220 million public and private companies worldwide

Amadeus (not subscribed)

- European public and private company information (21 million companies)

BankFocus (not subscribed)

- Detailed information on 38,000 banks (28,000 US and 10,000 Non-US)

Isis (not subscribed)

- detailed information on public and private insurance companies around the world. ISIS contains information on 11,700 companies in 190 different countries.

Osiris (not subscribed)

- Listed and major unlisted/delisted companies globally (80,000 companies)

Web Query in detail (BvD Orbis)

✓ BvD Orbis

- **Company Financials**
  - Financials for Industrial Companies
  - Financials for Banks
  - Financials for Insurance Companies
- **Company size categories**
  - Very large companies (VL)
  - Large companies (L)
  - Medium companies (M)
  - Small companies (S)
- **Shareholders**
  - All current shareholders First Level

BvD Orbis

➢ Company size categories

- **Very large companies (VL)**

  Companies that match at least one of the following conditions:

  - ✓ Operating Revenue >= 100 million EUR (130 million USD)
  - ✓ Total assets >= 200 million EUR (260 million USD)
  - ✓ Employees >= 1,000
  - ✓ Listed

- **Large companies (L)**

  - ✓ Operating Revenue >= 10 million EUR (13 million USD)
  - ✓ Total assets >= 20 million EUR (26 million USD)
  - ✓ Employees >= 150
  - ✓ Not Very Large

- **Medium companies (M)**

  - ✓ Operating Revenue >= 1 million EUR (1.3 million USD)
  - ✓ Total assets >= 2 million EUR (2.6 million USD)
  - ✓ Employees >= 15
  - ✓ Not Very Large or Large

- **Small companies (S)**

  - ✓ Companies not included in another category

BvD Orbis

➢ Consolidation codes

- **For a given company, ORBIS provides financial statements with different consolidation codes:**
  - ✓ C1: consolidated account aggregating all companies belonging to the group (affiliates, subsidiaries, etc.), where the company has no unconsolidated account
  - ✓ C2: consolidated account aggregating all companies belonging to the group (affiliates, subsidiaries, etc.) where the company also presents an unconsolidated account
  - ✓ U1: unconsolidated account of a company with no consolidated account
  - ✓ U2: unconsolidated account of a company which also has a consolidated account

BvD Orbis

- **Survivorship bias (data availability up to 5/6 years)**

  NBER working paper: "How to construct nationally representative firm level data from the Orbis global database", Kalemli-Ozcan et al. (2015).

  - ✓ http://econweb.umd.edu/~kalemli/assets/workingpapers/constructorbis.pdf

  - Known issues:

    - ✓ Download speed (the BvD is in general slow)
    - ✓ Survivorship bias: Both Orbis (up to 5 years) and Amadeus (at most 10 years) contain a number of years of historic financial data
    - ✓ Reporting lag (about 2 years)

  - ✓ **EDP – Energias de Portugal:**

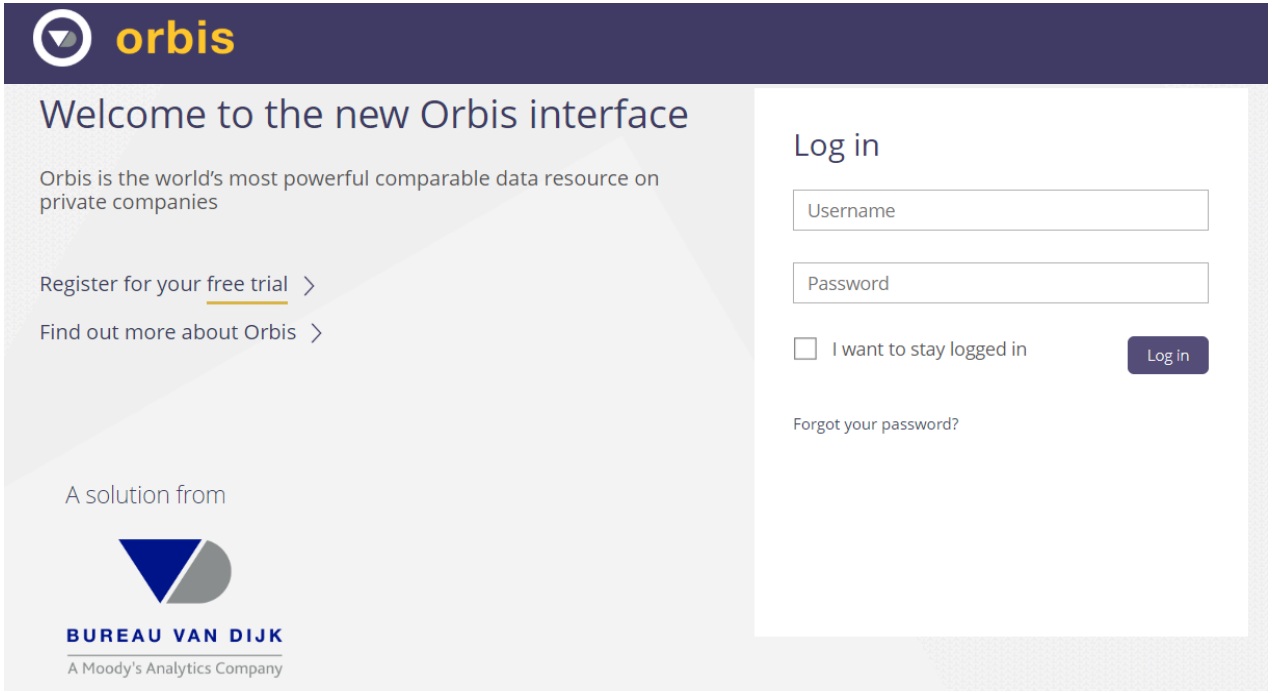
bvdid	name_internat	ctryiso	conscode	filing_type	closdate	closdate_y~r	exchrate	toas	empl
PT500697256	Edp - Energias De Portugal, S.A.	PT	C1	Annual report	31dec2009	2009	1.441	58000814926	12096
PT500697256	Edp - Energias De Portugal, S.A.	PT	C1	Annual report	31dec2010	2010	1.337	54129484659	12096
PT500697256	Edp - Energias De Portugal, S.A.	PT	C1	Annual report	31dec2011	2011	1.294	53412934195	12219
PT500697256	Edp - Energias De Portugal, S.A.	PT	C1	Annual report	31dec2012	2012	1.319	56243175365	12275
PT500697256	Edp - Energias De Portugal, S.A.	PT	C1	Annual report	31dec2013	2013	1.379	58818475601	12179
PT500697256	Edp - Energias De Portugal, S.A.	PT	C1	Annual report	31dec2014	2014	1.214	52052155633	11798
PT500697256	Edp - Energias De Portugal, S.A.	PT	C1	Annual report	31dec2015	2015	1.089	46309991110	12084
PT500697256	Edp - Energias De Portugal, S.A.	PT	C1	Annual report	31dec2016	2016	1.054	46468657252	11992
PT500697256	Edp - Energias De Portugal, S.A.	PT	C1	Annual report	31dec2017	2017	1.199	50460588337	11657
PT500697256	Edp - Energias De Portugal, S.A.	PT	C1	Annual report	31dec2018	2018	1.145	47662893218	11631

10 years

Web Query in detail (BvD Orbis)

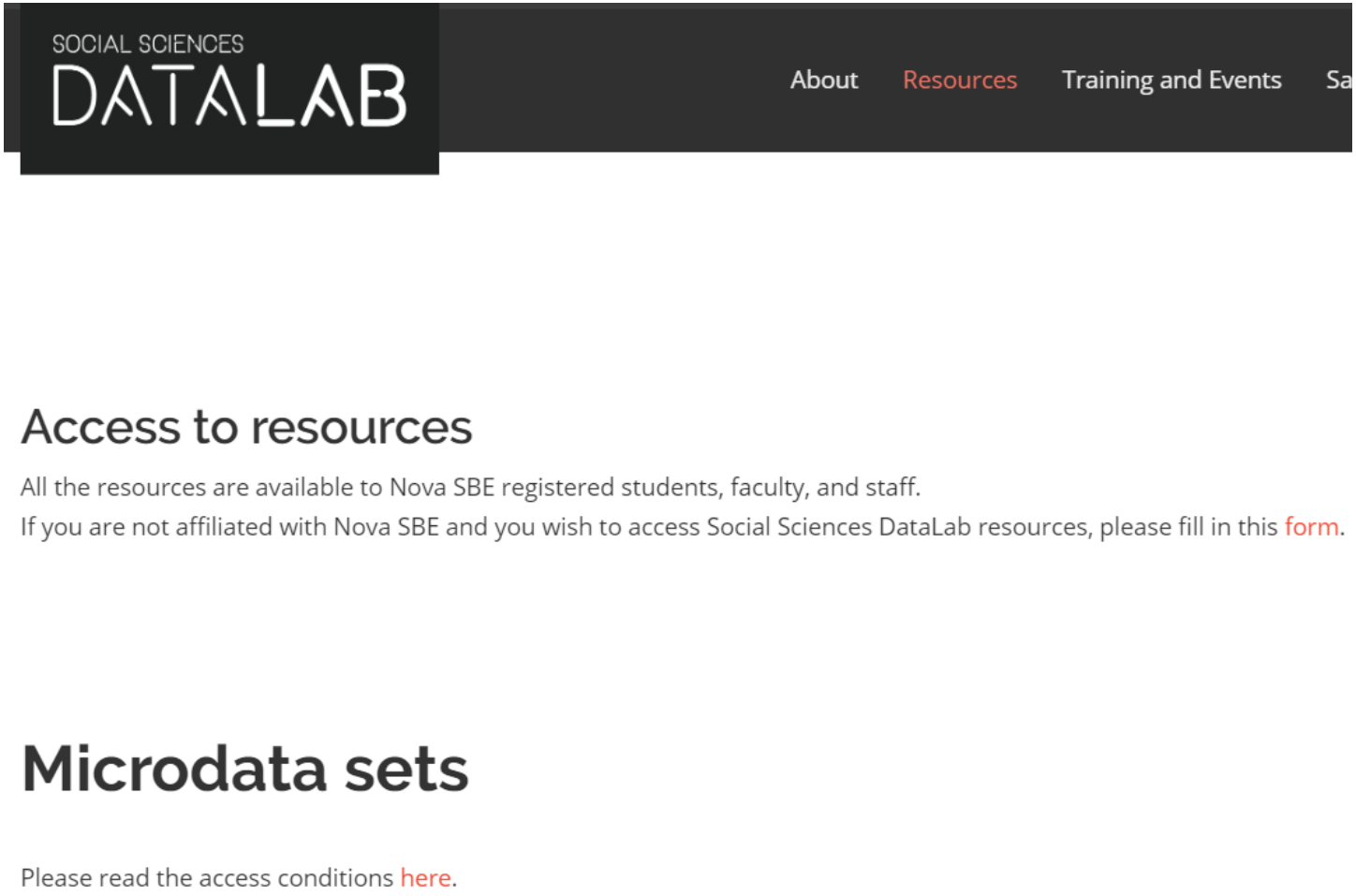
BvD Orbis – Web Interface

Orbis Web Interface @ Teresa and Alexandre Soares dos Santos Library



Web Query in detail (BvD Orbis)

DATALAB – Next class



Contents coverage

- **Segments: Geographical breakdown of sales**
  - **FactSet Revere**
  - **Compustat - Historical Segments**
- **Segments: breakdown of sales by customers**
  - **Compustat - Customer Segments**

✓ Factset Revere

FACTSET

FactSet Geographic Revenue Exposure (GeoRev™)

- Product designed to answer 3 questions: What do companies do? Who do they work with? Where do they sell?
- GeoRev focus on geographic revenue (Where do companies sell?)
- GeoRev currently covers more than 14,000 publicly traded companies around the world with historical data going back as far as 2005.

Figure 2: Intel's % Revenues by Geography as disclosed in FY 2011 10-K

Geography	Revenue (%)
Taiwan	31.6%
United States	15.6%
China	15.0%
Europe	12.9%
Asia/Pacific	10.2%
Japan	9.3%
Americas (ex. US)	5.4%

4 out of the 7 disclosed segments are country specific and require no further estimation.

Any inquiry regarding revenues exposure to countries beside Taiwan, U.S., China and Japan will require algorithmic estimations.

Factset Revere

FACTSET

FactSet Geographic Revenue Exposure (GeoRev™)

Features	FactSet GeoRev™	MSCI Economic Exposure Security Data Modules
Data Sources	**Public filings**	**Public filings**
Estimation Factor	**GDP weight**	**GDP weight**
History	**Up to 9 years**	**Up to 9 years (only for ≈ 2,500 large and mid companies in MSCI ACWI)**
Taxonomy for Data Mapping and Normalization	**Yes**	**Undisclosed**
Calculation Engine	**Estimates calculation 100% automated; no manual calculation to minimize errors**	**Undisclosed**
Confidence Factor	**Yes. Incorporates information from each company's unique disclosure**	**Yes, but rudimentary**
Data Granularity	**280+ super-regions, regions, areas, and countries. Ideally suited for user customization**	**244 regions and countries**
Coverage*	**14,000 large, mid and small cap companies in developed and emerging markets (Russell 3000 and MSCI ACWI included)**	**8,700 companies in MSCI ACWI FM IMI**
Customizable Estimation On Demand**	**Yes**	**Undisclosed**

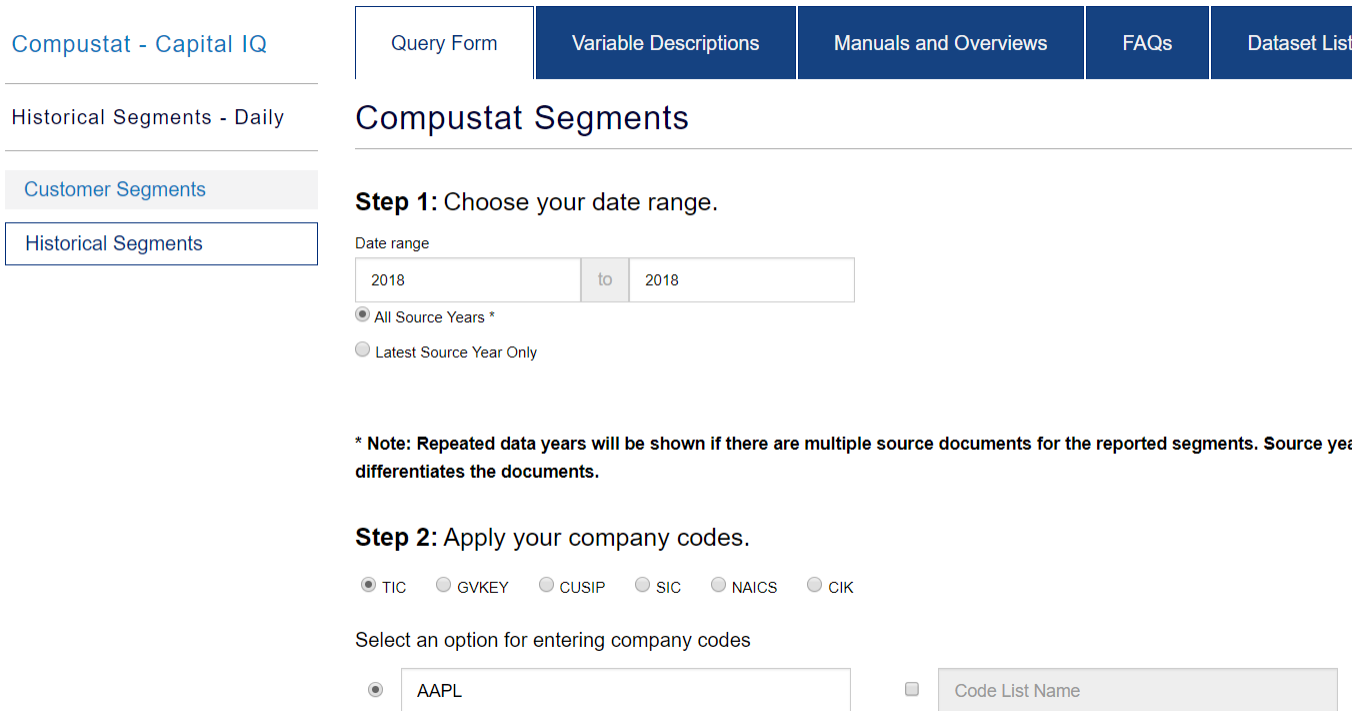
* Coverage as of Feb 2014. FactSet's coverage will expand to 15,000 companies by 2015 with a target goal of about 48,000 companies.
** Customers may have the option to request coverage of securities as we expand current coverage to entire global equity market.

✓ Compustat – Historical Segments

- Similar to FactSet GeoRev for US companies
- Find the geographical exposure of Apple sales (AAPL) in 2018

✓ Compustat – Historical Segments

➢ Find the geographical exposure of Apple sales (AAPL) in 2018



Web Query in detail (Compustat Historical Segments)

Exercise COMPSEG-1: Find the geographical exposure of Apple sales (AAPL) in 2018

5 - class1ex_COMPSEG1 × Untitled.do ×

```
1   use "D:\Documents\FDM - PhD 2020\data\compsegment_aapl.dta", clear
2
3   order gvkey conm tic datadate srcdate stype sales curcds srcs snms
4
5   keep if stype=="GEOSEG" & sales!=0
6
7   egen total_sales = sum(sales)
8
9   gen percent = sales/total_sales
10
11  gsort -percent
12
13  format conm %-16s
14  format snms %-24s
15
16  br gvkey conm tic datadate srcdate stype sales curcds srcs snms total_sales percent
17
```

gvkey	conm	tic	datadate	srcdate	stype	sales	curcds	srcs	snms	total_sales	percent
001690	APPLE INC	AAPL	30sep2018	30sep2018	GEOSEG	112093	USD	5	Americas	265595	.4220448
001690	APPLE INC	AAPL	30sep2018	30sep2018	GEOSEG	62420	USD	5	Europe	265595	.2350195
001690	APPLE INC	AAPL	30sep2018	30sep2018	GEOSEG	51942	USD	5	Greater China	265595	.1955684
001690	APPLE INC	AAPL	30sep2018	30sep2018	GEOSEG	21733	USD	5	Japan	265595	.0818276
001690	APPLE INC	AAPL	30sep2018	30sep2018	GEOSEG	17407	USD	5	Rest of Asia Pacific	265595	.0655396

Geog. Revenue Breakdown

- **Americas: 42.2%**
- **Asia/Pacific: 34.3%**
- **Europe: 23.5%**

✓ Compustat – Customer Segments

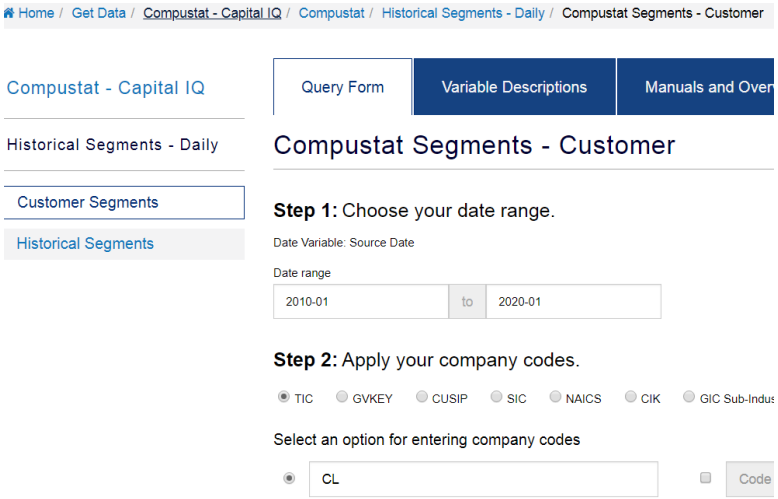
- Compustat Segments Data provides **business and geographic detail**, **product information** and **customer data** for over 70% of the companies in the North American (NA) database.
- Coverage
  - **Current Segments:** Past 8 years of information for over 8,000 North American companies.
  - **Historical Segments:** Since 1976, information for over 23,000 North American companies.

Web Query in detail (Compustat Customer Segments)

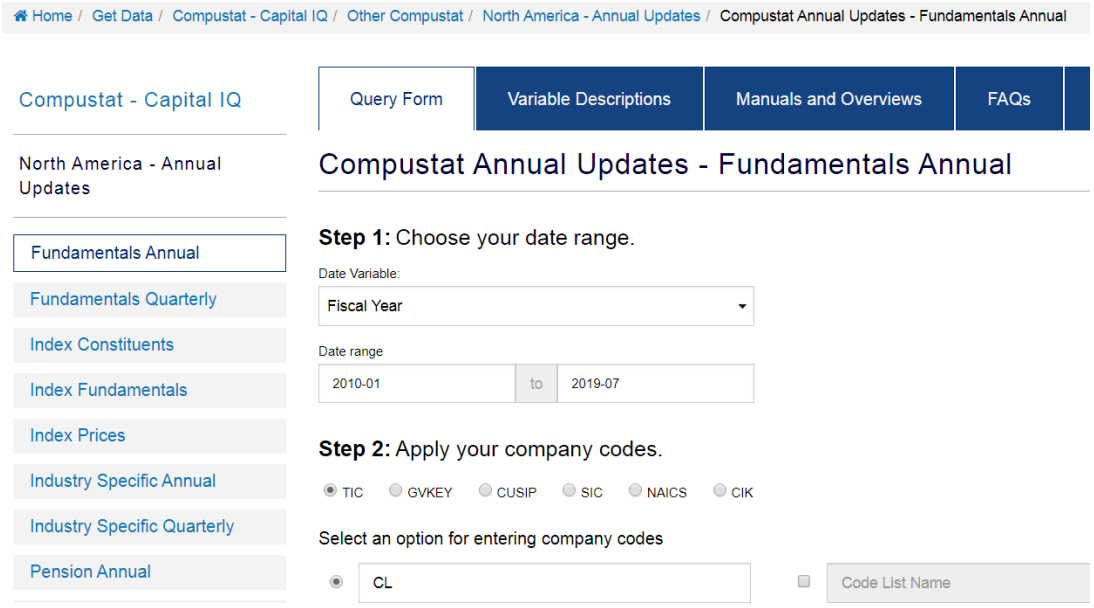
Compustat Segments – Customer

Example: Find the % sales of Colgate-Palmolive to reported customers in fiscal year 2017.

Dataset: Customer Segments



Dataset: Fundamentals Annual



Web Query in detail (Compustat Customer Segments)

Compustat Segments – Customer

Example: Find the % sales of Colgate-Palmolive to reported customers in fiscal year 2017.

```
use "C:\Users\Pedro Pires\Downloads\compsegment_customer.dta", clear

format conm %-24s
format cnms %-24s

merge m:1 gvkey using "C:\Users\Pedro Pires\Downloads\comp_colgatepalmolive.dta"

gen double salecs_pct = salecs/sale*100

format salecs_pct %4.2f

br gvkey cid cnms ctype salecs sid stype srcdate conm tic cusip sale salecs_pct if gvkey=="003170" & year(srcdate)==2017 & ctype=="COMPANY"
```

gvkey	cid	cnms	ctype	salecs	sid	stype	srcdate	conm	tic	cusip	sale	salecs_pct
003170	9	Wal-Mart Stores Inc	COMPANY	1699.94	0		31dec2017	COLGATE-PALMOLIVE CO	CL	194162103	15454	11.00

Distribution; Raw Materials; Competition; Trademarks and Patents

The Company's products are marketed by a direct sales force at individual operating subsidiaries or business units, and by distributors or brokers. The Oral, Personal and Home Care products are sold to a variety of retail and wholesale customers and distributors. Pet Nutrition products are sold by authorized pet supply retailers and veterinarians. Many of the Company's products are also sold online through various e-commerce platforms and retailers. The Company's sales to Wal-Mart Stores, Inc. and its affiliates represent approximately 11% of the Company's Net sales in 2017. No other customer represents more than 10% of the Company's Net sales.

Contents coverage

➢ Issuer credit ratings

- **Compustat S&P Ratings (North America)**
- **Alternative: Bloomberg terminal - RATC function (Global)**

July 29, 2016 Bloomberg Brief | Distress & Bankruptcy 6

DOWNGRADE WATCH BY JOHN E. MORRIS, BLOOMBERG BRIEFS

Tobacco, Graphite Makers See Ratings Cut

> **S&P** downgraded **Alliance One International** to CCC, citing an oversupply of tobacco and reduced demand, together with the possibility of a distressed exchange of its second-lien notes.
> **S&P** also cut **SGL Carbon SE** to CCC+, saying the company is likely to have an "unsustainable" debt-to-adjusted Ebitda ratio of 15 over the full year.

COMPANY	COUNTRY	CUT TO	FROM	AGENCY	DATE	RATING TYPE	INDUSTRY
Alliance One International Inc	US	CCC	CCC+	S&P	7/28	LT Local, Foreign Issuer Credit	Tobacco Leaf
Gibson Brands Inc	US	CCC	CCC+	S&P	7/28	LT Local, Foreign Issuer Credit	Music Instruments Manufacturing
Halcon Resources Corp	US	Ca	Caa2	Moody's	7/28	LT Corp Family Rating	Crude Oil Production
Olin Corp	US	BB	BB+	S&P	7/28	LT Local, Foreign Issuer Credit	Alkalis & Chlorine Manufacturing
SGL Carbon SE	DE	CCC+	B	S&P	7/27	LT Local, Foreign Issuer Credit	Graphite Products Manufacturing
Trader Corp	CA	B3	B2	Moody's	7/27	LT Corp Family Rating	Publishing
Turkiye Sise ve Cam Fabrikalari AS	TR	BB	BB+	S&P	7/26	LT Local, Foreign Issuer Credit	Glass Product Manufacturing - Building Materials
Basic Energy Services Inc	US	CCC-	CCC+	S&P	7/25	LT Local, Foreign Issuer Credit	Oilfield Services & Equipment Manufacturing
Holdikks SAS	FR	B-	B	S&P	7/25	LT Local, Foreign Issuer Credit	Apparel, Footwear & Accessories Design
Ottawa Holdings Pte Ltd	SG	Caa1	B3	Moody's	7/25	Senior Secured Debt	Publishing & Broadcasting
Outerwall Inc	US	B+	BB-	S&P	7/25	LT Local, Foreign Issuer Credit	Movie & Video Game Rental Stores

Source: Bloomberg

RATC <GO>

Note: Table shows downgrades for sub-investment grade companies in the July 25-28 period.

✓ S&P Ratings (data until FEB2017)

- Compustat S&P Ratings from 1973
  - **S&P Domestic Long-Term Issuer Credit Rating**
  - **S&P Domestic Short-Term Issuer Credit Rating**

Credit Rating Scales by Agency, Long-Term

Moody's	S&P	Fitch	
Aaa	AAA	AAA	Prime
Aa1	AA+	AA+	High grade
Aa2	AA	AA	
Aa3	AA-	AA-	
A1	A+	A+	Upper medium grade
A2	A	A	
A3	A-	A-	
Baa1	BBB+	BBB+	Lower medium grade
Baa2	BBB	BBB	
Baa3	BBB-	BBB-	
Ba1	BB+	BB+	Non-investment grade speculative
Ba2	BB	BB	
Ba3	BB-	BB-	
B1	B+	B+	Highly speculative
B2	B	B	
B3	B-	B-	
Caa1	CCC+	CCC	Substantial risk
Caa2	CCC		Extremely speculative
Caa3	CCC-		Default imminent with little prospect for recovery
Ca	CC		
	C		
C	D	DDD	In default
/		DD	
/		D	

"Junk"

WOLFSTREET.com

Contents coverage

- **Analyst estimates**
  - **Thomson Reuters IBES**

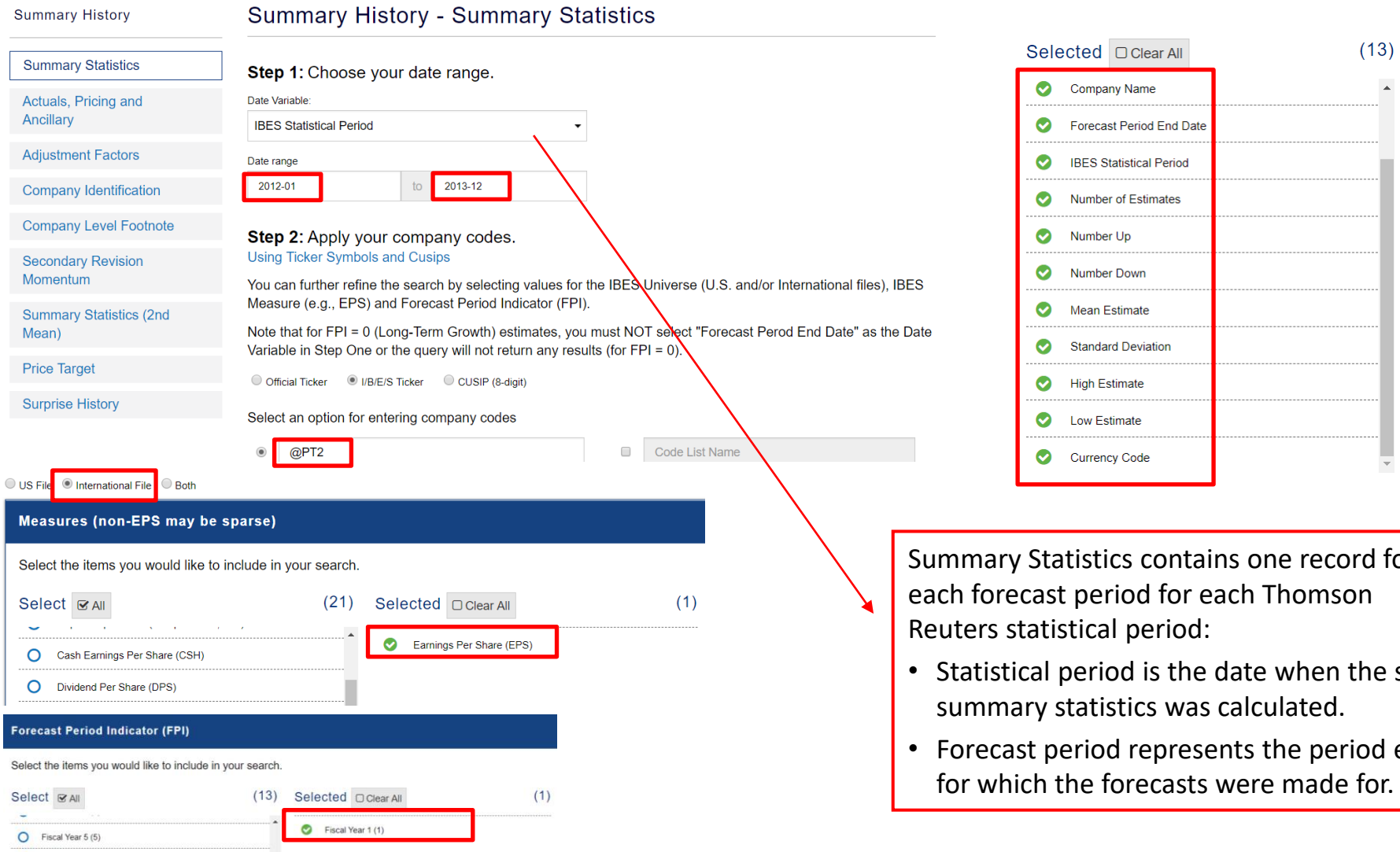
- ✓ **Thomson Reuters IBES (not subscribed – temporarily)**
- ➢ Institutional Brokers' Estimate System (I/B/E/S)
- ➢ Analysts forecast data (eps, sales, etc.), consensus estimates and trade recommendations
- ➢ Global coverage starting in 1980
- ➢ More than 22,000 thousand companies and 930 brokers

For additional insights about IBES watch e-learning video from WRDS:

→ http://wrds-web.wharton.upenn.edu/wrds/E-Learning/_000Video/IBES_at_WRDS/index.cfm

Web Query in detail (IBES)

I/B/E/S EPS Forecast for Portugal Telecom



Summary Statistics contains one record for each forecast period for each Thomson Reuters statistical period:

- Statistical period is the date when the set of summary statistics was calculated.
- Forecast period represents the period end for which the forecasts were made for.

Web Query in detail (IBES)

I/B/E/S EPS Forecast for Portugal Telecom

	ticker	oftic	cname	statpers	measure	fpi	curcode	numest	numup	numdown	meanest	stdev	highest	lowest	fpedats
1	@PT2	PTC	PORTUGAL TELECOM	19jan2012	EPS	1	EUR	28	2	3	.55	.1	.8	.41	31dec2011
2	@PT2	PTC	PORTUGAL TELECOM	16feb2012	EPS	1	EUR	28	0	4	.54	.1	.8	.41	31dec2011
3	@PT2	PTC	PORTUGAL TELECOM	15mar2012	EPS	1	EUR	28	2	5	.52	.09	.79	.4	31dec2011
4	@PT2	PTC	PORTUGAL TELECOM	19apr2012	EPS	1	EUR	29	1	13	.47	.11	.81	.31	31dec2012
5	@PT2	PTC	PORTUGAL TELECOM	17may2012	EPS	1	EUR	28	0	4	.45	.12	.81	.29	31dec2012
6	@PT2	PTC	PORTUGAL TELECOM	14jun2012	EPS	1	EUR	29	3	4	.45	.12	.81	.27	31dec2012
7	@PT2	PTC	PORTUGAL TELECOM	19jul2012	EPS	1	EUR	29	1	13	.4	.13	.81	.24	31dec2012
8	@PT2	PTC	PORTUGAL TELECOM	16aug2012	EPS	1	EUR	29	2	8	.38	.12	.81	.25	31dec2012
9	@PT2	PTC	PORTUGAL TELECOM	20sep2012	EPS	1	EUR	29	1	3	.38	.12	.81	.25	31dec2012
10	@PT2	PTC	PORTUGAL TELECOM	18oct2012	EPS	1	EUR	28	1	3	.36	.09	.6	.25	31dec2012
11	@PT2	PTC	PORTUGAL TELECOM	15nov2012	EPS	1	EUR	27	3	5	.34	.07	.52	.25	31dec2012
12	@PT2	PTC	PORTUGAL TELECOM	20dec2012	EPS	1	EUR	27	2	4	.34	.06	.52	.27	31dec2012
13	@PT2	PTC	PORTUGAL TELECOM	17jan2013	EPS	1	EUR	26	0	1	.33	.06	.52	.26	31dec2012
14	@PT2	PTC	PORTUGAL TELECOM	14feb2013	EPS	1	EUR	27	1	7	.32	.06	.5	.24	31dec2012
15	@PT2	PTC	PORTUGAL TELECOM	14mar2013	EPS	1	EUR	28	2	8	.34	.1	.62	.16	31dec2013
16	@PT2	PTC	PORTUGAL TELECOM	18apr2013	EPS	1	EUR	28	4	4	.34	.09	.55	.19	31dec2013
17	@PT2	PTC	PORTUGAL TELECOM	16may2013	EPS	1	EUR	28	2	2	.34	.09	.55	.19	31dec2013
18	@PT2	PTC	PORTUGAL TELECOM	20jun2013	EPS	1	EUR	27	1	10	.3	.09	.55	.18	31dec2013
19	@PT2	PTC	PORTUGAL TELECOM	18jul2013	EPS	1	EUR	28	4	5	.29	.1	.55	.15	31dec2013
20	@PT2	PTC	PORTUGAL TELECOM	15aug2013	EPS	1	EUR	28	2	5	.28	.1	.49	.01	31dec2013
21	@PT2	PTC	PORTUGAL TELECOM	19sep2013	EPS	1	EUR	28	8	4	.3	.12	.49	.01	31dec2013
22	@PT2	PTC	PORTUGAL TELECOM	17oct2013	EPS	1	EUR	23	4	6	.31	.14	.49	.02	31dec2013
23	@PT2	PTC	PORTUGAL TELECOM	14nov2013	EPS	1	EUR	23	2	4	.31	.14	.53	.05	31dec2013
24	@PT2	PTC	PORTUGAL TELECOM	19dec2013	EPS	1	EUR	22	2	2	.3	.14	.53	.05	31dec2013

Contents coverage

- **Executive compensation**
  - **Compustat Execucomp – Executive Compensation**
  - **Boardex [Not subscribed]**

✓ Compustat Execucomp – Executive Compensation

- Executive compensation data (DEF14A SEC form)
- Contains over 2872 companies, both active and inactive
- Covers the S&P 1500 since 1994, top 5 earners (e.g., CEO, CFO) in each company
- Date field is based on Fiscal Year
- Units: Thousands of Dollars

Compustat Execucomp – Executive Compensation

➢ Executive Compensation of Facebook in 2018

Compustat - Capital IQ

Execucomp - Monthly Updates

Annual Compensation

Company Financial and Director Compensation for 2005 and prior

Deferred Compensation

Director Compensation

Long Term Incentive Awards - 1992 Format

Query Form | Variable Descriptions | Manuals

Compustat Executive Compensation

Step 1: Choose your date range.

Date range: 2018 to 2018

Step 2: Apply your company codes.

TICKER | GVKEY | CUSIP | CO_PER_ROL | EXECID

Select an option for entering company codes

FB

Selected (5) Clear All

- Company Name
- EXEC_FULLNAME
- YEAR -- Fiscal Year
- TOTAL_SEC -- Total Compensation - As Reported in SEC Filin
- TDC2 -- Total Compensation (Salary + Bonus + Other An

Compustat Execucomp – Executive Compensation

- **Executive Compensation of Facebook in 2018**

https://www.sec.gov/Archives/edgar/data/1326801/000132680119000025/facebook2019definitiveprox.htm

exec_fullname	co_per_rol	coname	total_sec	tdc2	gvkey	year	ticker
Mark Elliot Zuckerberg	48275	FACEBOOK INC	22554.543	22554.543	170617	2018	FB
Michael T. Schroepfer	48276	FACEBOOK INC	19757.363	158767.482	170617	2018	FB
Sheryl Kara Sandberg	48278	FACEBOOK INC	23728.418	196746.341	170617	2018	FB
David M. Wehner	50979	FACEBOOK INC	19686.113	19686.113	170617	2018	FB
Christopher K. Cox	50981	FACEBOOK INC	19686.113	19686.113	170617	2018	FB

2018 Summary Compensation Table

The following table presents summary information regarding the total compensation awarded to, earned by, or paid to each of the named executive officers for services rendered to us for the years ended December 31, 2018, 2017, and 2016.

Name and Principal Position	Fiscal Year	Salary ($)[1]	Bonus ($)[2]	Stock Awards ($)[3]	All Other Compensation ($)	Total ($)
Mark Zuckerberg	2018	1	—	—	22,554,542[4]	22,554,543
CEO	2017	1	—	—	9,101,965[4]	9,101,966
	2016	1	—	—	6,015,431[4]	6,015,432
Sheryl K. Sandberg	2018	843,077	638,310	18,423,523	3,823,508[5]	23,728,418
COO	2017	795,769	640,378	21,072,431	2,687,643[5]	25,196,221
	2016	738,077	1,293,635	19,908,426	2,609,319[5]	24,549,457
David M. Wehner	2018	753,846	499,494	18,423,523	9,250	19,686,113
CFO	2017	711,539	633,317	21,072,431	9,000	22,426,287
	2016	662,692	940,421	14,931,596	9,566	16,544,275
Christopher K. Cox[6]	2018	753,846	499,494	18,423,523	9,250	19,686,113
Former CPO	2017	711,539	567,404	21,072,431	9,000	22,360,374
	2016	658,846	933,209	14,931,596	9,538	16,533,189
Mike Schroepfer	2018	753,846	570,744	18,423,523	9,250	19,757,363
CTO	2017	711,539	633,317	21,072,431	9,000	22,426,287
	2016	658,846	859,356	14,931,596	9,377	16,459,175

Outside WRDS… (BoardEx)

✓ Compensation Data (BoardEx)

Are US CEOs Paid More?
New International Evidence*

Nuno Fernandes, IMD International
Miguel A. Ferreira, Nova School of Business and Economics
Pedro Matos, University of Virginia, Darden School of Business
Kevin J. Murphy, University of Southern California, Marshall School of Business

(*forthcoming in The Review of Financial Studies*)

Abstract

This paper challenges the widely accepted stylized fact that CEOs in the United States are paid significantly more than their foreign counterparts. Using CEO pay data across 14 countries with mandated pay disclosures, we show that the US pay premium is economically modest and primarily reflects the performance-based pay demanded by institutional shareholders and independent boards. Indeed, we find no significant difference in either level of CEO pay or the use of equity-based pay between US and non-US firms exposed to international and US capital, product, and labor markets. We also show that US and non-US CEO pay has largely converged in the 2000s.

Appendix A: Variables definition and data sources

A. CEO PAY	
Total Pay	Total CEO compensation in US$ (US firms: ExecuComp; non-US firms: BoardEx, corporate filings)
Salary	Salary in US$ (ExecuComp: salary; BoardEx: salary)
Other pay	Other compensation in US$ (ExecuComp: other compensation; BoardEx: other pay and pensions)
Bonuses	Non-equity incentive-plan compensation in US$ (ExecuComp: bonus plus target value of non-equity incentive-plan compensation; BoardEx: bonus)
Equity pay (stock and options)	Stock and options awards in US$ (ExecuComp: grant-date fair value of stock awards plus grant-date fair value of option awards; BoardEx: market value of shares plus long-term incentive plans plus Black-Scholes option value)

The disclosure situation has improved over the past decade. Regulations mandating disclosure of executive pay were introduced in Ireland and South Africa in 2000 and in Australia in 2004. In May 2003, the European Union (EU) Commission issued an "Action Plan" recommending that all listed companies in the EU report details on individual compensation packages, and that EU member countries pass rules requiring such disclosure. By 2006, six EU members (in addition to the United Kingdom and Ireland) had mandated CEO-level disclosure: Belgium, France, Germany, Italy, Netherlands, and Sweden. In addition, although not in the EU, Norway also adopted EU-style disclosure rules, and Switzerland demanded similar disclosure for the "highest-paid" executive.

Our primary data source on compensation for US CEOs is Standard and Poor's (S&P's) ExecComp database, while our primary source for CEOs of firms based outside the United States is BoardEx, compiled by the UK-based firm Management Diagnostics Limited. Together, these two sources (identified as "BoardEx & Exec" in Table 1) account for 2,899 of the 3,263 firms in our sample. BoardEx is the leading database on board composition of publicly listed firms, and includes detailed biographic information on individual executives and board members in nearly 50 countries, including countries that do not have mandatory disclosure requirements for executive compensation. In addition to providing biographic information, BoardEx also includes detailed compensation data for top executives – including salaries, other pay, bonuses, payouts under long-term plans, option grants, and share grants.

Contents coverage

- **M&A and Equity/Debt underwriting**
  - **Thomson Reuters SDC**
  - **Thomson Reuters Eikon**

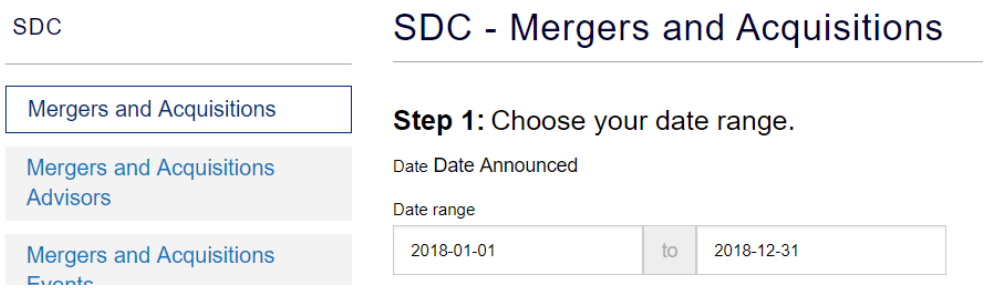
✓ SDC Global Issues – Thomson Reuters Eikon

- **Mergers and Acquisitions**
  - more than 500,000 transactions globally since 1980
- **Equity**
  - more than 100,000 transactions (IPO/SEO) globally since 1970
- **Bonds**
  - more than 300,000 transactions globally since 1970
- **Syndicated Loans**
  - more than 120,000 transactions globally since 1982

Outside WRDS… (SDC)

✓ Thomson Reuters SDC – WRDS

- **Mergers and Acquisitions**
- **New Issues**



- Duplicated records: One Deal_ID with 2 target companies

	master_dea~o	dateann	tmanames	amanames	dateeff	status	aupnames	tpublic	apublic	tnation	anation	acusip	master_cusip
1039	3312748040	19nov2018	Boulangerie Patisserie La	Mohamed A...	19nov2018	Completed	Mohamed A...	Priv.	Inv.	France	France	8H3787	8H4354
1040	3312748040	19nov2018	La Maison Mila	Mohamed A...	19nov2018	Completed	Mohamed A...	Priv.	Inv.	France	France	8H3787	8H3790

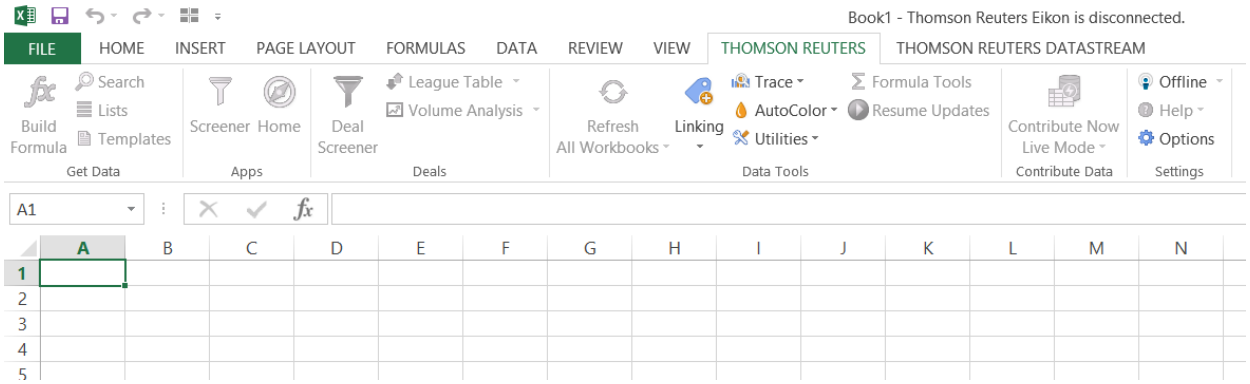
- Duplicated records: One Deal_ID with several records (*alegal* - legal advisor name, *datefin* – date of financial statement, …)

	master_dea~o	dateann	tmanames	amanames	dateeff	status	aupnames	tpublic	apublic	tnation	anation	acusip	master_cusip
71	3203226040	04jan2018	Societe Nouvelle Des Autocars	Keolis SA	04jan2018	Completed	SNCF	Priv.	Sub.	France	France	48979M	1H6286
72	3203226040	04jan2018	Societe Nouvelle Des Autocars	Keolis SA	04jan2018	Completed	SNCF	Priv.	Sub.	France	France	48979M	1H6286
73	3203226040	04jan2018	Societe Nouvelle Des Autocars	Keolis SA	04jan2018	Completed	SNCF	Priv.	Sub.	France	France	48979M	1H6286
74	3203226040	04jan2018	Societe Nouvelle Des Autocars	Keolis SA	04jan2018	Completed	SNCF	Priv.	Sub.	France	France	48979M	1H6286
75	3203226040	04jan2018	Societe Nouvelle Des Autocars	Keolis SA	04jan2018	Completed	SNCF	Priv.	Sub.	France	France	48979M	1H6286
76	3203226040	04jan2018	Societe Nouvelle Des Autocars	Keolis SA	04jan2018	Completed	SNCF	Priv.	Sub.	France	France	48979M	1H6286
77	3203226040	04jan2018	Societe Nouvelle Des Autocars	Keolis SA	04jan2018	Completed	SNCF	Priv.	Sub.	France	France	48979M	1H6286
78	3203226040	04jan2018	Societe Nouvelle Des Autocars	Keolis SA	04jan2018	Completed	SNCF	Priv.	Sub.	France	France	48979M	1H6286
79	3203226040	04jan2018	Societe Nouvelle Des Autocars	Keolis SA	04jan2018	Completed	SNCF	Priv.	Sub.	France	France	48979M	1H6286
80	3203226040	04jan2018	Societe Nouvelle Des Autocars	Keolis SA	04jan2018	Completed	SNCF	Priv.	Sub.	France	France	48979M	1H6286
81	3203226040	04jan2018	Societe Nouvelle Des Autocars	Keolis SA	04jan2018	Completed	SNCF	Priv.	Sub.	France	France	48979M	1H6286
82	3203226040	04jan2018	Societe Nouvelle Des Autocars	Keolis SA	04jan2018	Completed	SNCF	Priv.	Sub.	France	France	48979M	1H6286

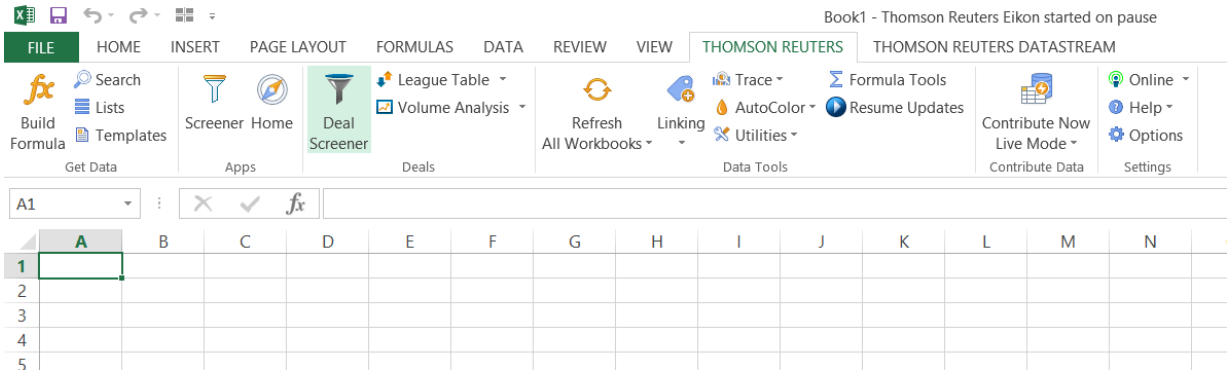
SDC Global Issues – Thomson Reuters Eikon

How to use TR Eikon's Deal Analytics (User Guide)

1 – Open Excel and Select the "Thomson Reuters" Add-in Tab. Click on offline/login and insert login credentials



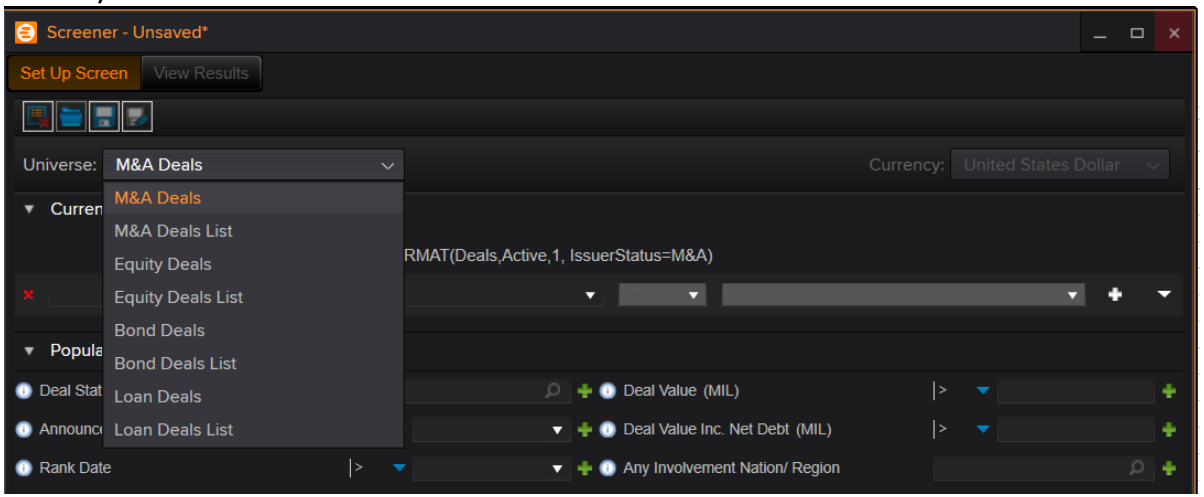
2 – Click on "Deal Screener"



SDC Global Issues – Thomson Reuters Eikon

How to use TR Eikon's Deal Analytics

3 – The "Set Up Screen" window will pop up. Select the deal type in the "Universe" ListBox (M&A Deals, Equity Deals, Bond Deals, Loan Deals).

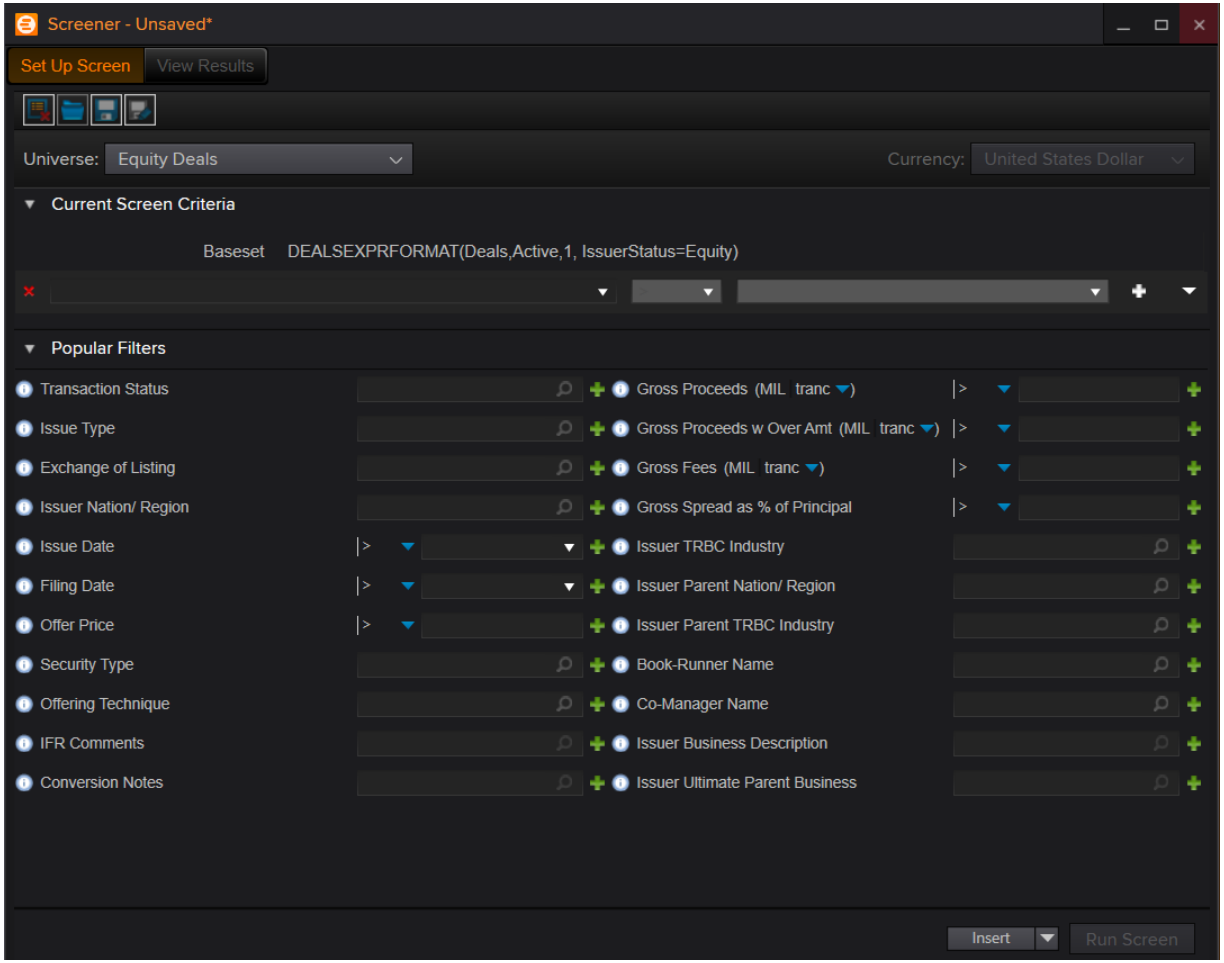


4 – Pick "Equity Deals" as an example. The extraction process of data from TR Eikon Deal Analytics is a bit cumbersome as there is a download limit. So the data needs to be subdivided before exporting to excel. Subset the data either by setting up "Popular Filters" which are fields already available in the screener or by inserting other available fields in the screener.

SDC Global Issues – Thomson Reuters Eikon

How to use TR Eikon's Deal Analytics

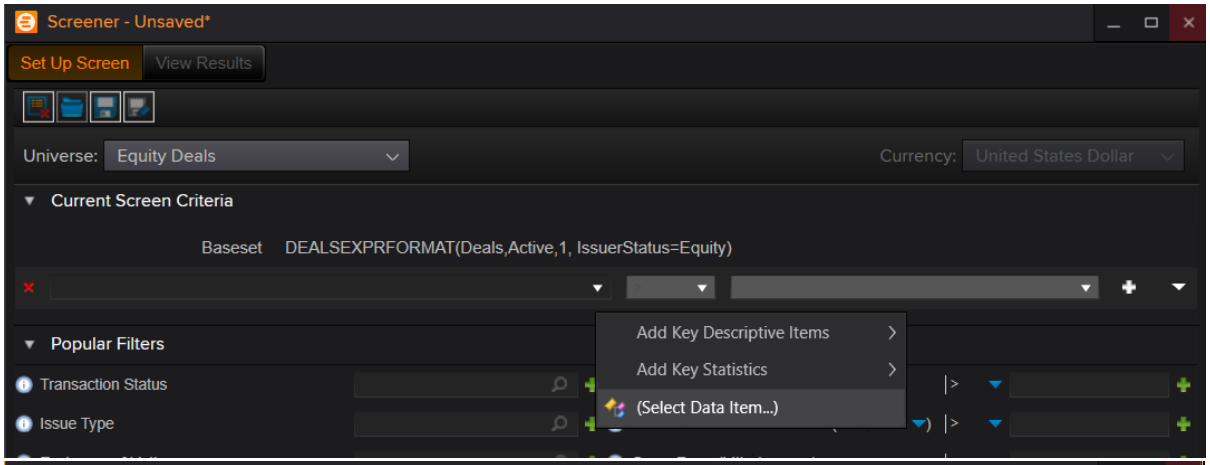
4.1 – "Popular filters"

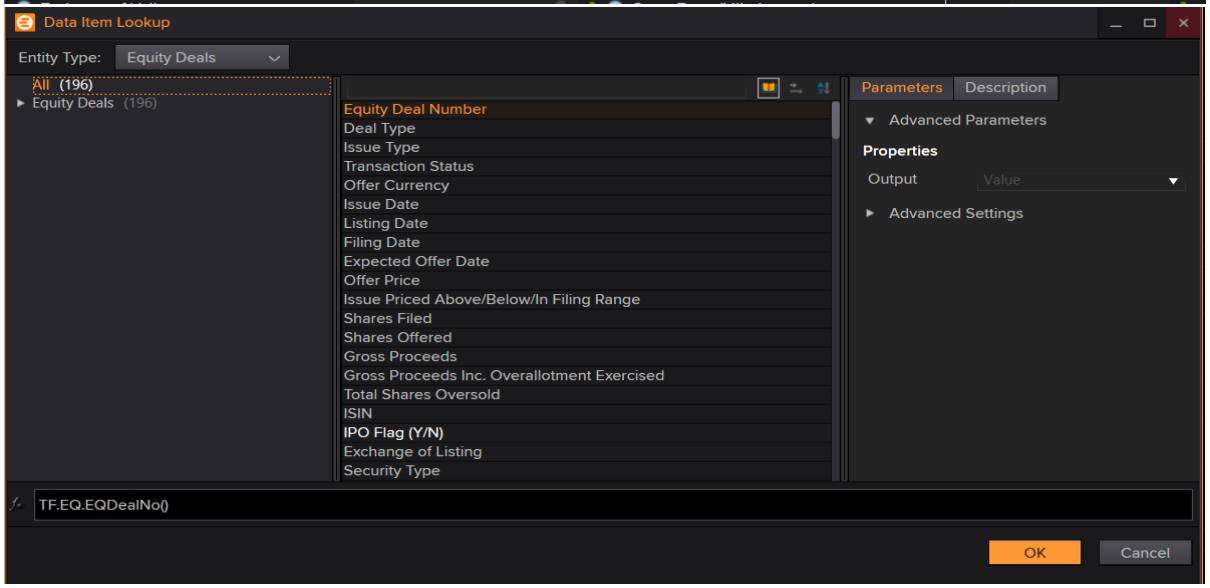


SDC Global Issues – Thomson Reuters Eikon

How to use TR Eikon's Deal Analytics

4.2 – Insert other available Fields



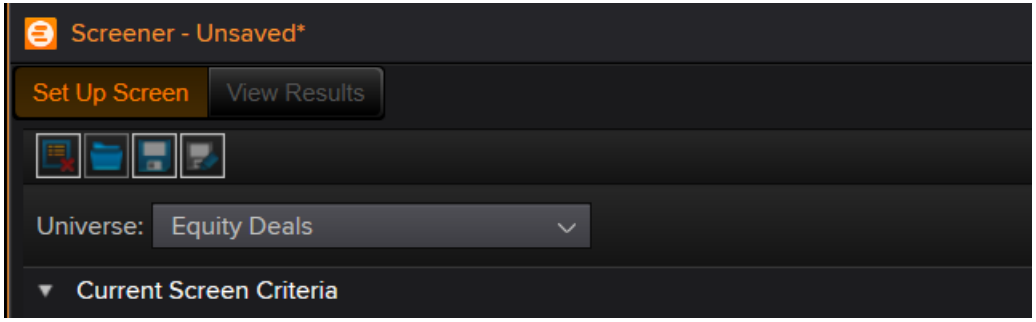


SDC Global Issues – Thomson Reuters Eikon

How to use TR Eikon's Deal Analytics

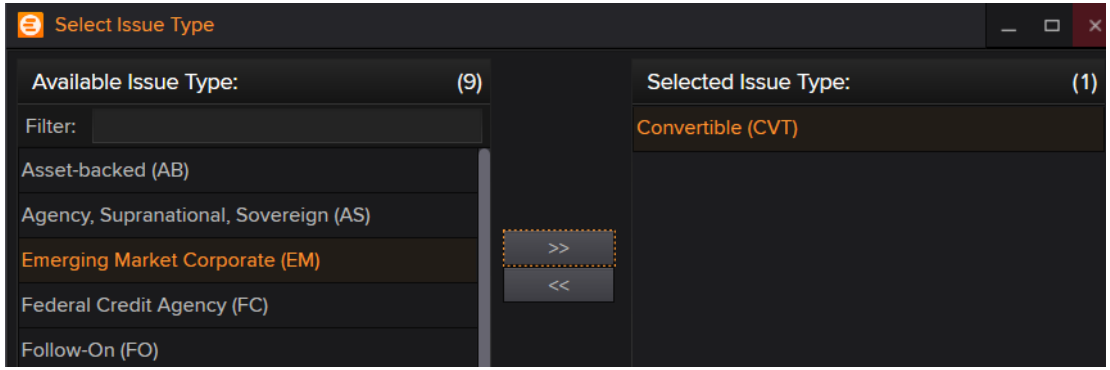
5 – Example: Set-Up Screen for Convertibles.

5.1 – Select "Equity Deals" in the "Universe" ListBox



5.2 – Click on the magnifier next to "Issue Type"

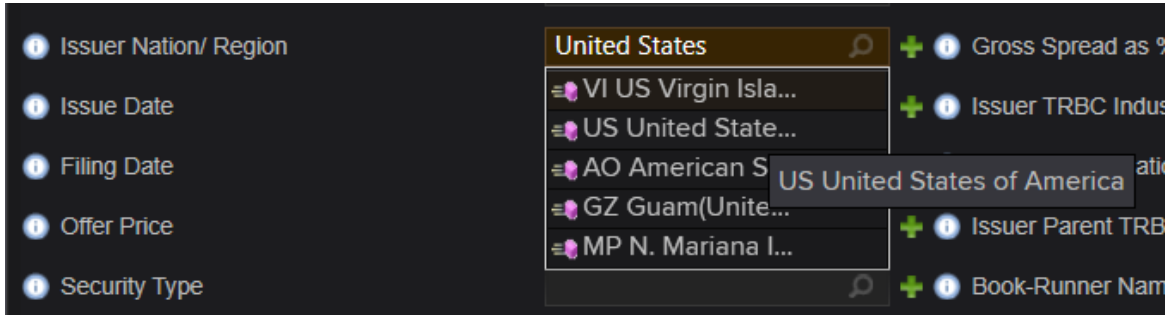
5.3 – Select "Convertible (CVT)" from the "Issue Type" ListBox. Click OK.



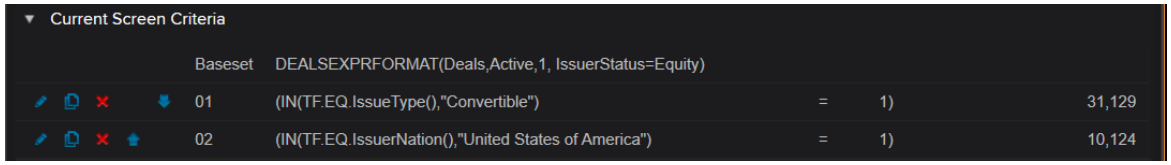
SDC Global Issues – Thomson Reuters Eikon

How to use TR Eikon's Deal Analytics

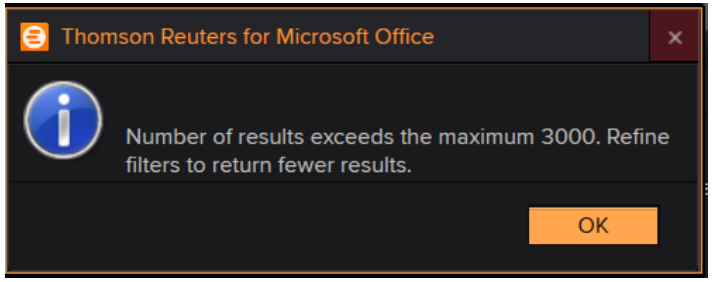
5.4 – Select "United States" from the "Issuer Nation/Region" TextBox



5.5 – Check the record count (number of records in the query to the Deal Analytics dataset)



5.6 – Make sure the number of records doesn't exceed the 3'000 download limit

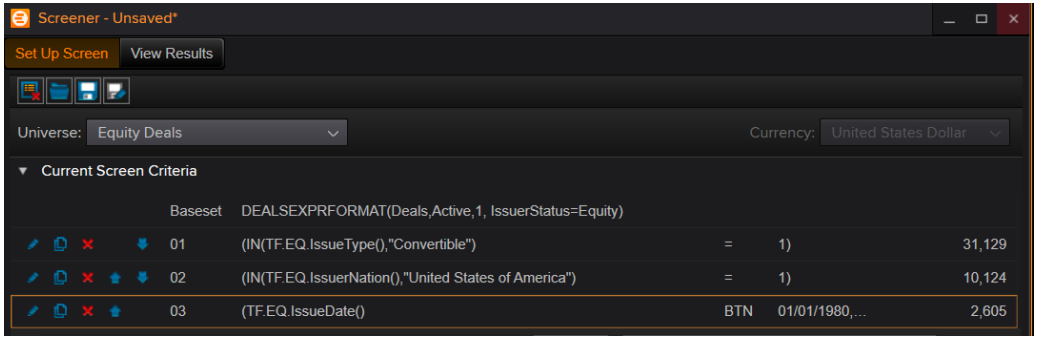


SDC Global Issues – Thomson Reuters Eikon

How to use TR Eikon's Deal Analytics

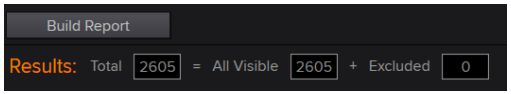
5.7 – A way to comply with the download limit is to subset data using date intervals using the field "Issue Date"

5.8 – Check the record count until it's below 3'000



5.9 – Run Screen

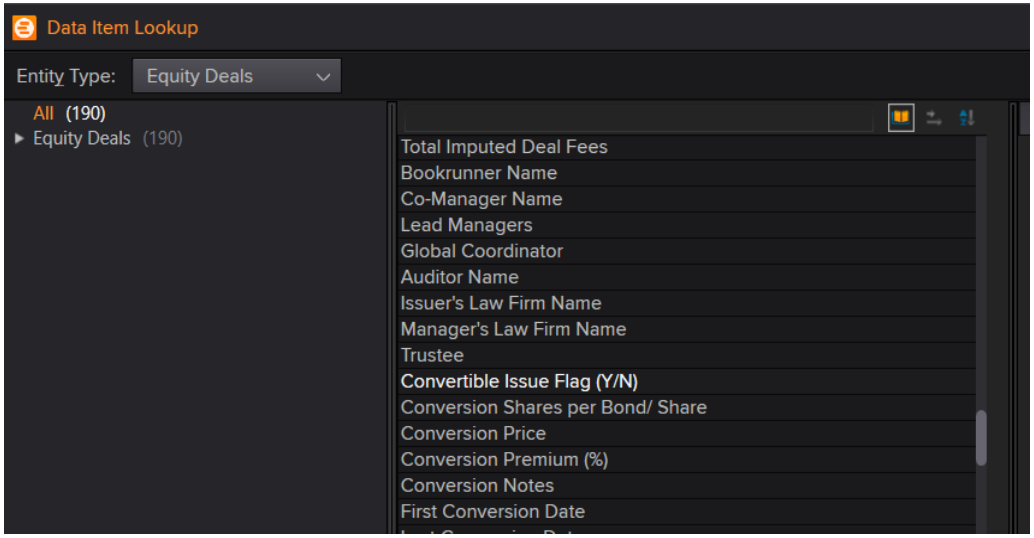
5.10 – Click "Build Report" and "Add report items" to add more fields to the screener



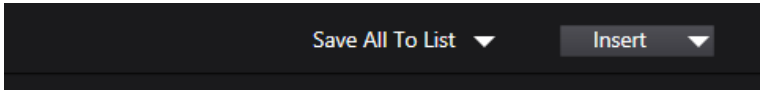
SDC Global Issues – Thomson Reuters Eikon

How to use TR Eikon's Deal Analytics

5.11 – Pick the fields you wish to include in the output. Select e.g. "Offer Currency"; "ISIN"; "Security Type"; "Use of Proceeds"; "Gross Spread"; "Principal Amount"; "Maturity Date"; "Equity Deal Type"; "Issuer SDC CUSIP"; "Issuer Name"; "Convertible Issue Flag"; "Conversion Shares per Bond/Share"; "Conversion Price"; "Conversion Premium (%)"; "Conversion Notes"; "First Conversion Date"; "Last Conversion Date".



5.12 – Click "Insert" when the query is ready to download



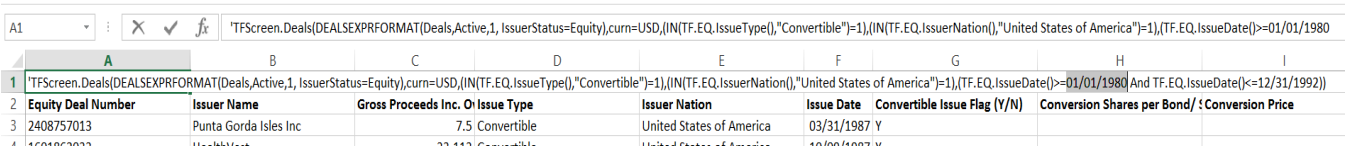
SDC Global Issues – Thomson Reuters Eikon

How to use TR Eikon's Deal Analytics

5.13 – Data exported to Excel:

	A	B	C	D	E	F	G
1	TFScreen.Deals(DEALSEXPRFORMAT(Deals,Active,1, IssuerStatus=Equity),curn=USD,(IN(TF.EQ.IssueType(),"Convertible")=1),(IN(TF.EQ.IssuerNation(),"United States of America")=1),(TF.EQ.IssueDate						
2	**Equity Deal Number**	**Issuer Name**	**Gross Proceeds Inc. O**	**Issue Type**	**Issuer Nation**	**Issue Date**	**Convertible Issue Flag (Y/N)**
3	2408757013	Punta Gorda Isles Inc	7.5	Convertible	United States of America	03/31/1987	Y
4	1601862033	HealthVest	33.113	Convertible	United States of America	10/09/1987	Y
5	1241482007	Dennison Manufacturing Co	37.099	Convertible	United States of America	02/01/1985	Y
6	1196434097	Club Med Inc	0.13	Convertible	United States of America	06/17/1992	Y
7	1103498007	Community Psychiatric Centers	15	Convertible	United States of America	02/01/1980	Y
8	1103473007	Oak Industries Inc	35	Convertible	United States of America	09/01/1980	Y
9	1103465007	Digicon Finance NV	18	Convertible	United States of America	09/01/1980	Y
10	1103419007	Community Psychiatric Centers	15	Convertible	United States of America	04/01/1981	Y
11	1103401007	Massmutual Mortgage Realty	21	Convertible	United States of America	07/01/1981	Y
12	1103397007	Blue Ridge Petroleum Corp	25	Convertible	United States of America	07/01/1981	Y
13	1103344007	AMI American Medical Intl Inc	25	Convertible	United States of America	04/30/1982	Y
14	1103339007	Southern California Edison Co	50	Convertible	United States of America	07/01/1982	Y
15	1103314007	Duncan Lawrie Security Holding	15	Convertible	United States of America	04/29/1983	Y
16	1103252007	Viacom International Inc	50	Convertible	United States of America	10/01/1984	Y
17	1103235007	National Patent Development Corp		Convertible	United States of America	01/01/1985	Y
18	1103225007	Wang Laboratories Inc	74.198	Convertible	United States of America	02/01/1985	Y
19	1103195007	St Paul Cos Inc	100	Convertible	United States of America	03/01/1985	Y
20	1103191007	Newmont Mining Corp	60	Convertible	United States of America	03/01/1985	Y
21	1103190007	National Patent Development C	17.495	Convertible	United States of America	03/01/1985	Y
22	1103188007	Louisiana Land & Exploration C	50	Convertible	United States of America	03/01/1985	Y
23	1103187007	Louisiana Land & Exploration C	50	Convertible	United States of America	03/01/1985	Y
24	1103154007	American General Corp	300	Convertible	United States of America	05/01/1985	Y
25	1103153007	Viacom International Inc	50	Convertible	United States of America	05/31/1985	Y
26	1103149007	Pan American Corporation	0.767	Convertible	United States of America	05/31/1985	Y
27	1103141007	ICN Pharmaceuticals Inc	19.227	Convertible	United States of America	05/31/1985	Y
28	1103136007	Coleco Inc	0.614	Convertible	United States of America	05/31/1985	Y
29	1103135007	Continental Health Affiliates Inc	13.425	Convertible	United States of America	05/31/1985	Y
30	1103131007	Limited Inc	50	Convertible	United States of America	07/01/1985	Y

5.14 – Use the Excel Formula to change date ranges thus avoiding using the set up screener window



'TFScreen.Deals(DEALSEXPRFORMAT(Deals,Active,1, IssuerStatus=Equity),curn=USD,(IN(TF.EQ.IssueType(),"Convertible")=1),(IN(TF.EQ.IssuerNation(),"United States of America")=1),(TF.EQ.IssueDate()>=01/01/1980 And TF.EQ.IssueDate()<=12/31/1992))

Contents coverage

- **Stocks**
  - **CRSP**
  - **Compustat Supplemental Short Interest File**
- **Database Management (Merging Datasets)**
  - **CRSP-Compustat Merged (CCM)**
  - **Merging CRSP and Compustat by CUSIP**
  - **Fuzzy Matching**

✓ CRSP – The Center for Research in Security Prices

- Comprehensive collection of daily and monthly security price, return, and volume data for the NYSE, AMEX and NASDAQ stock markets.
- CRSP U.S. Stock databases do not include securities for international companies unless they are ADRs, cross-listed, or traded on the major stock exchanges mentioned above.
- Daily and Monthly data for roughly 28,000 securities traded on major exchanges from 1925-present.
- Complete historical information (bias-free).
- Accurate total returns calculation.

CRSP

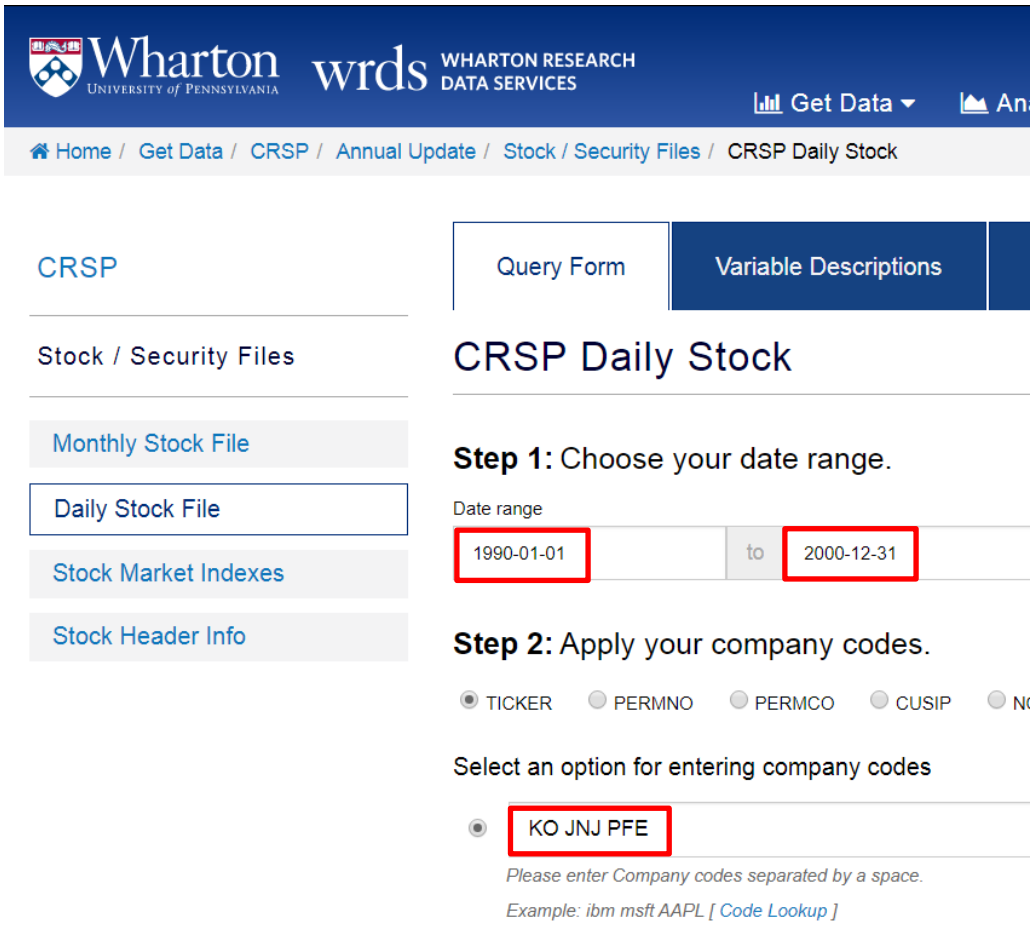
- Information in CRSP:
  - Price and quote data.
  - Holding period returns with and without dividends.
  - Shares outstanding.
  - Trading volume.
  - Security delisting information.

CRSP - Identifiers

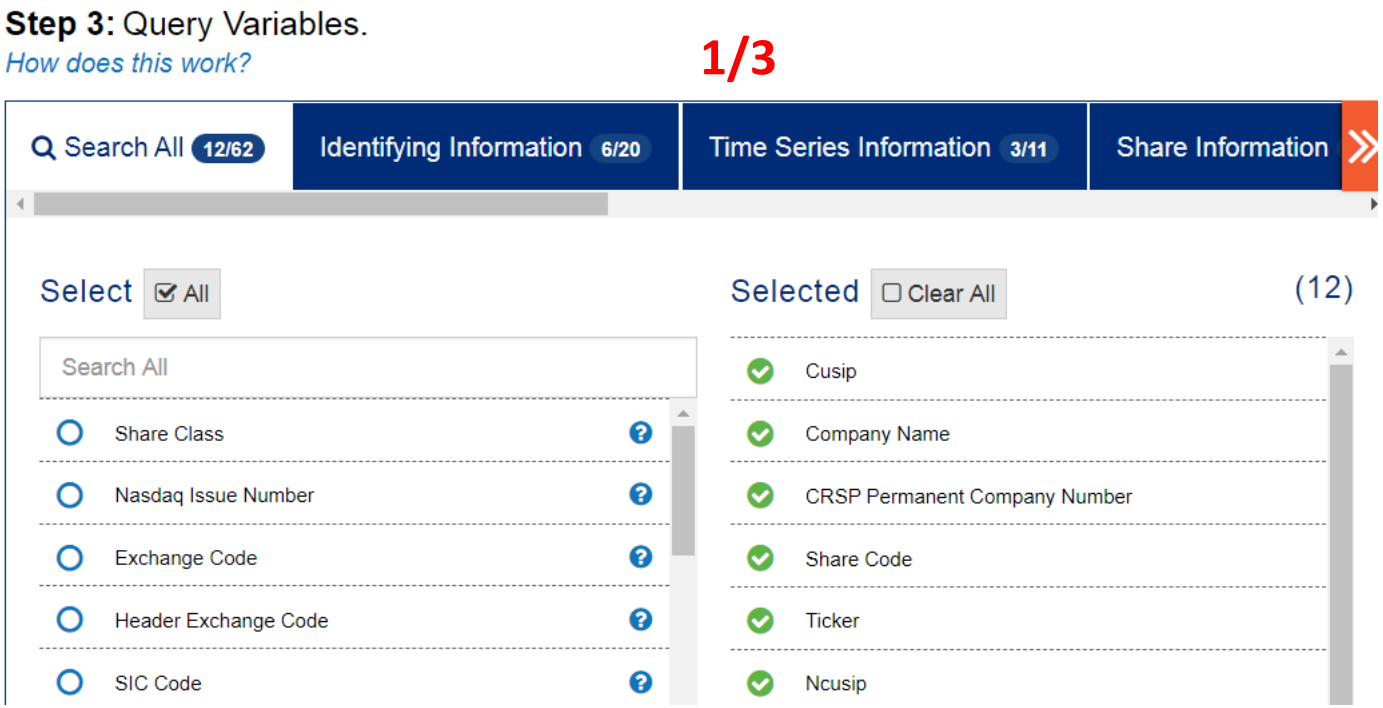
- PERMNO: Unique and permanent issue identification number
- PERMCO: Permanent company identification number
- While PERMCO applies to only one company more than one PERMNO (securities) can be associated with one PERMCO (company)
- Common Identifiers:
  - CUSIP (latest eight-character CUSIP identifier)
  - Ticker

CRSP (Web Query)

- Get <u>daily</u> Price, Volume, Holding Return and Shares Outstanding for Coca-Cola, Johnson & Johnson and Pfizer from 1990-01-01 to 2000-12-31.



CRSP (Web Query)

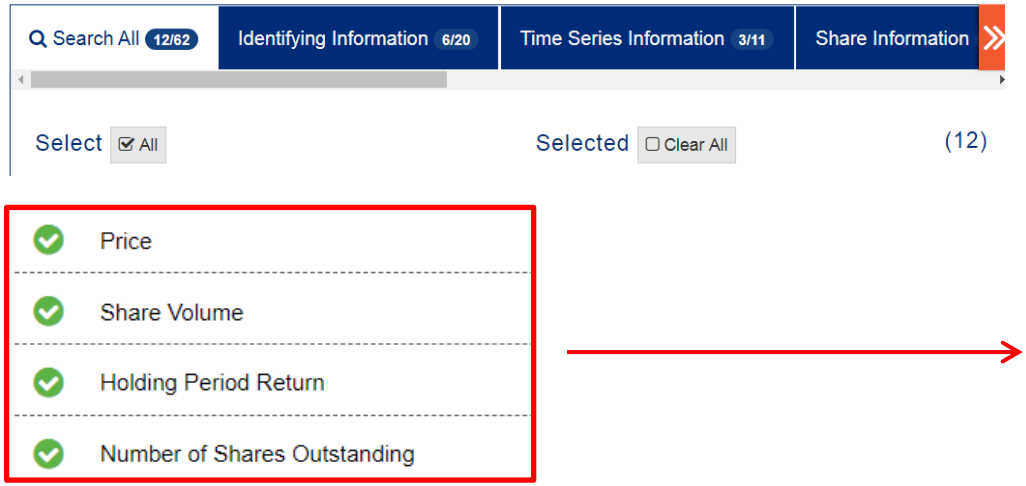


Web Query in detail (CRSP)

CRSP (Web Query)

Step 3: Query Variables. 2/3

How does this work?



Price (PRC) is the closing price or the negative bid/ask average for a trading day. If the number in the price field has a negative sign it is a bid/ask average and not an actual closing price.

Share Volume (VOL) is the sum of the trading volumes during that period. It is expressed in units of one share, for daily data, and on hundred shares for monthly data. Volume is set to -99 if the value is missing.

Holding Period Return (RET) is the change in the total value of an investment in a common stock over some period of time per dollar of initial investment. It's already adjusted to split events and includes dividends.

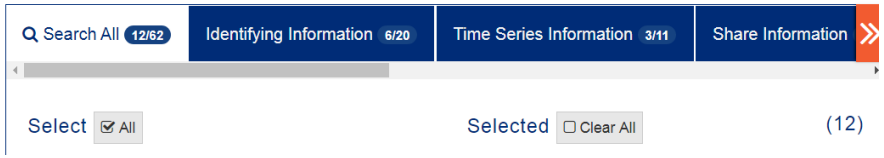
Nbr. of Shares Outstanding (SHROUT) is the number of publicly held shares recorded in thousands

Web Query in detail (CRSP)

CRSP (Web Query)

3/3

Step 3: Query Variables.
How does this work?



- Cumulative Factor to Adjust Price
- Cumulative Factor to Adjust Shares

- Factors to adjust price and shares outstanding for split events (stock splits, stock dividends, spin-offs, stock distributions, and rights)
- Adjusted Price = PRC / CFACPR
- Adjusted Shares Outstanding = SHROUT * CFACSHR
- Holding period returns are already adjusted!

For most distribution cases (excl. e.g. DISTCD 6225):

$$r(t) = \left(\frac{|PRC(t)| + d(t)}{CFACPR(t)}\right) \Big/ \left(\frac{|PRC(t')|}{CFACPR(t')}\right) - 1$$

RET(t)	Reason For Missing Return
-66.0	more than 10 periods between time t and the time of the preceding price t?
-77.0	not trading on the current exchange at time t
-88.0	no return, array index t not within range of BEGRET and ENDRET
-99.0	missing return due to missing price at time t

	A	B	E	F	J	M	O	Q
1	PERMNO	Names Date	Ticker Symbol	Company Name	Dividend Cash Amount	Price or Bid/Ask Average	Returns	Factor to Adjust Prices
48	11308	1990/03/08	KO	COCA COLA CO		72.375	0.014011	16
49	11308	1990/03/09	KO	COCA COLA CO	0.4	72	0.000345	16
50	11308	1990/03/12	KO	COCA COLA CO		72.125	0.001736	16
51	11308	1990/03/13	KO	COCA COLA CO		71.875	-0.003466	16
52	11308	1990/03/14	KO	COCA COLA CO		72.375	0.006957	16

r(KO, 1990-03-09) = [(72+0.4)/16]/[72.375/16]-1 = 0.000345

CRSP – Share Code

First Digit	
Code	Definition
1	Ordinary Common Shares
2	Certificates
3	ADRs (American Depository Receipts)
4	SBIs (Shares of Beneficial Interest)
7	Units (Depository Units, Units of Beneficial Interest, Units of Limited Partnership Interest, Depository Receipts, etc.)

Second Digit	
Code	Definition
0	Securities which have not been further defined.
1	Securities which need not be further defined.
2	Companies incorporated outside the US
3	Americus Trust Components (Primes and Scores).
4	Closed-end funds.
5	Closed-end fund companies incorporated outside the US
8	REIT's (Real Estate Investment Trusts).

Usually you want to include share codes 10 and 11 (Ordinary common shares incorporated in the US)

✓ Merging datasets in STATA

excel:

	A	B	C	D	E	F	G	H	I	J
1	Ticker	Year	Industry		Ticker	Year	Assets		Industry	IndustryName
2	AAPL	2015	A		AAPL	2015	80		A	Technology
3	AAPL	2016	A		AAPL	2016	90		B	Banking
4	JPM	2015	B		JPM	2015	200		C	Beverage
5	JPM	2016	B		JPM	2016	210			
6	KO	2015	C							
7	KO	2016	C							

Stata:

- **edit**
- **copy paste**
- **save the 3 files (e.g. ds1, ds2, ds3)**

```
*#################################
*# STATA MERGE - EXAMPLE
*#################################
cd "D:\FDM - PhD 2020\data"

* example - merge 1:1
use ds1, clear
merge 1:1 ticker year using ds2
br

* example - merge m:1
use ds1, clear
merge m:1 industry using ds3
br

* example - merge 1:m
use ds3, clear
merge 1:m industry using ds1
sort ticker year
br
*#################################
```

✓ Merging datasets in STATA

merge 1:1 ticker year using "path\filename.dta"

ds1

	ticker	year	industry
1	AAPL	2015	A
2	AAPL	2016	A
3	JPM	2015	B
4	JPM	2016	B
5	KO	2015	C
6	KO	2016	C

ds2

	ticker	year	assets
1	AAPL	2015	80
2	AAPL	2016	90
3	JPM	2015	200
4	JPM	2016	210

	ticker	year	industry	assets	_merge
1	AAPL	2015	A	80	matched (3)
2	AAPL	2016	A	90	matched (3)
3	JPM	2015	B	200	matched (3)
4	JPM	2016	B	210	matched (3)
5	KO	2015	C	.	master only (1)
6	KO	2016	C	.	master only (1)

✓ Merging datasets in STATA

merge m:1 industry using "path\filename.dta"

ds1

	ticker	year	industry
1	AAPL	2015	A
2	AAPL	2016	A
3	JPM	2015	B
4	JPM	2016	B
5	KO	2015	C
6	KO	2016	C

ds3

	industry	industryname
1	A	Technology
2	B	Banking
3	C	Beverage

	ticker	year	industry	industryname	_merge
1	AAPL	2015	A	Technology	matched (3)
2	AAPL	2016	A	Technology	matched (3)
3	JPM	2015	B	Banking	matched (3)
4	JPM	2016	B	Banking	matched (3)
5	KO	2015	C	Beverage	matched (3)
6	KO	2016	C	Beverage	matched (3)

✓ Merging datasets in STATA

merge 1:m industry using "path\filename.dta"

ds3

	industry	industryname
1	A	Technology
2	B	Banking
3	C	Beverage

ds1

	ticker	year	industry
1	AAPL	2015	A
2	AAPL	2016	A
3	JPM	2015	B
4	JPM	2016	B
5	KO	2015	C
6	KO	2016	C

	industry	industryname	ticker	year	_merge
1	A	Technology	AAPL	2015	matched (3)
2	A	Technology	AAPL	2016	matched (3)
3	B	Banking	JPM	2015	matched (3)
4	B	Banking	JPM	2016	matched (3)
5	C	Beverage	KO	2015	matched (3)
6	C	Beverage	KO	2016	matched (3)

✓ Merging datasets in STATA

merge — Merge datasets

Syntax

One-to-one merge on specified key variables

`merge 1:1` *varlist* `using` *filename* [, *options*]

Many-to-one merge on specified key variables

`merge m:1` *varlist* `using` *filename* [, *options*]

One-to-many merge on specified key variables

`merge 1:m` *varlist* `using` *filename* [, *options*]

Many-to-many merge on specified key variables

`merge m:m` *varlist* `using` *filename* [, *options*]

Description

`merge` joins corresponding observations from the dataset currently in memory (called the master dataset) with those from *filename*`.dta` (called the using dataset), matching on one or more key variables. `merge` can perform match merges (one-to-one, one-to-many, many-to-one, and many-to-many), which are often called *joins* by database people. `merge` can also perform sequential merges, which have no equivalent in the relational database world.

`merge` is for adding new variables from a second dataset to existing observations. You use `merge`, for instance, when combining hospital patient and discharge datasets. If you wish to add new observations to existing variables, then see [D] **append**. You use `append`, for instance, when adding current discharges to past discharges.

By default, `merge` creates a new variable, `_merge`, containing numeric codes concerning the source and the contents of each observation in the merged dataset. These codes are explained below in the match results table.

Key variables cannot be `strL`s.

If *filename* is specified without an extension, then `.dta` is assumed.

Web Query in detail (CRSP)

```
. merge 1:1 id using filename
```

in memory *master* + in *filename*.dta *using* = *merged result*

id	age
1	22
2	56
5	17

id	wgt
1	130
2	180
4	110

id	age	wgt	
1	22	130	(matched)
2	56	180	(matched)
5	17	.	(master only)
4	.	110	(using only)

The original data in memory are called the master data. The data in *filename*`.dta` are called the using data. After `merge`, the merged result is left in memory. The `id` variable is called the key variable. Stata jargon is that the datasets were merged on `id`.

Observations for `id==1` existed in both the master and using datasets and so were combined in the merged result. The same occurred for `id==2`. For `id==5` and `id==4`, however, no matches were found and thus each became a separate observation in the merged result. Thus each observation in the merged result came from one of three possible sources:

m:1 merges

In an `m:1` merge, the key variable or variables uniquely identify the observations in the using data, but not necessarily in the master data. Suppose you had person-level data within regions and you wished to bring in regional data. Here is an example:

```
. merge m:1 region using filename
```

master + *using* = *merged result*

id	region	a
1	2	26
2	1	29
3	2	22
4	3	21
5	1	24
6	5	20

region	x
1	15
2	13
3	12
4	11

id	region	a	x	_merge
1	2	26	13	3
2	1	29	15	3
3	2	22	13	3
4	3	21	12	3
5	1	24	15	3
6	5	20	.	1
.	4	.	11	2

To bring in the regional information, we need to merge on `region`. The values of `region` identify individual observations in the using data, but it is not an identifier in the master data.

We show the merged dataset sorted by `id` because this makes it easier to see how the merged dataset was constructed. For each observation in the master data, `merge` finds the corresponding observation in the using data. `merge` combines the values of the variables in the using dataset to the observations in the master dataset.

✓ CRSP-Compustat Merged (CCM)

- Easiest way to merge CRSP and Compustat
- Use either CRSP or Compustat identifiers, (i.e. PERMNO|PERMCO; GVKEY)
- The match between CRSP and COMPUSTAT data is highly accurate, but it is not complete or entirely unambiguous
  - CRSP database covers stock prices on public stock exchanges while COMPUSTAT does not require a company to have a traded stock. There are cases where a company that is covered by COMPUSTAT has valid stock price data for only part of its history.
  - Mergers: sometimes there is disagreement between CRSP and COMPUSTAT over which is the surviving company.
  - The match between COMPUSTAT's GVKEY and CRSP's PERMNO is not one-to-one. For example, a company might have multiple equity issues.
  - Different calendar/fiscal regime.
  - Different frequencies: Compustat data is either annual or quarterly while CRSP data is either monthly or daily.

CRSP-Compustat Merged (CCM)

- Primary link types - 41%
  - **LC: Link research complete.**
  - **LU: Link is unresearched by CRSP and is established by comparing CUSIPs**
  - LS: these links mainly relate to ETFs
- Secondary link types – 2%
  - LX, LD, LN
- Non-matching link types – 57%
  - NR: No link available; confirmed by research
  - NU: No link available; not yet confirmed

CRSP-Compustat Merged (CCM)

- CCM Link structure

MNEMONIC	FIELD NAME
LPERMNO	CRSP PERMNO link during link period. It is set to zero if there is no CRSP link during the range.
LPERMCO	CRSP PERMCO link during link period. It is set to zero if there is no CRSP link during the range.
LIID	Security identifier
LNKTYPE	Link type code. Each link is given a code describing the connection between the CRSP and Compustat data. Values are: • LC – Link research complete. Standard connection between databases. • LU – Unresearched link to issue by CUSIP • LX – Link to a security that trades on another exchange system not included in CRSP data. • LD – Duplicate link to a security. Another GVKEY/IID is a better link to that CRSP record. • LS – Link valid for this security only. Other CRSP PERMNOs with the same PERMCO will link to other GVKEYs. • LN – Primary link exists but Compustat does not have prices. • NR – No link available, confirmed by research • NU – No link available, not yet confirmed

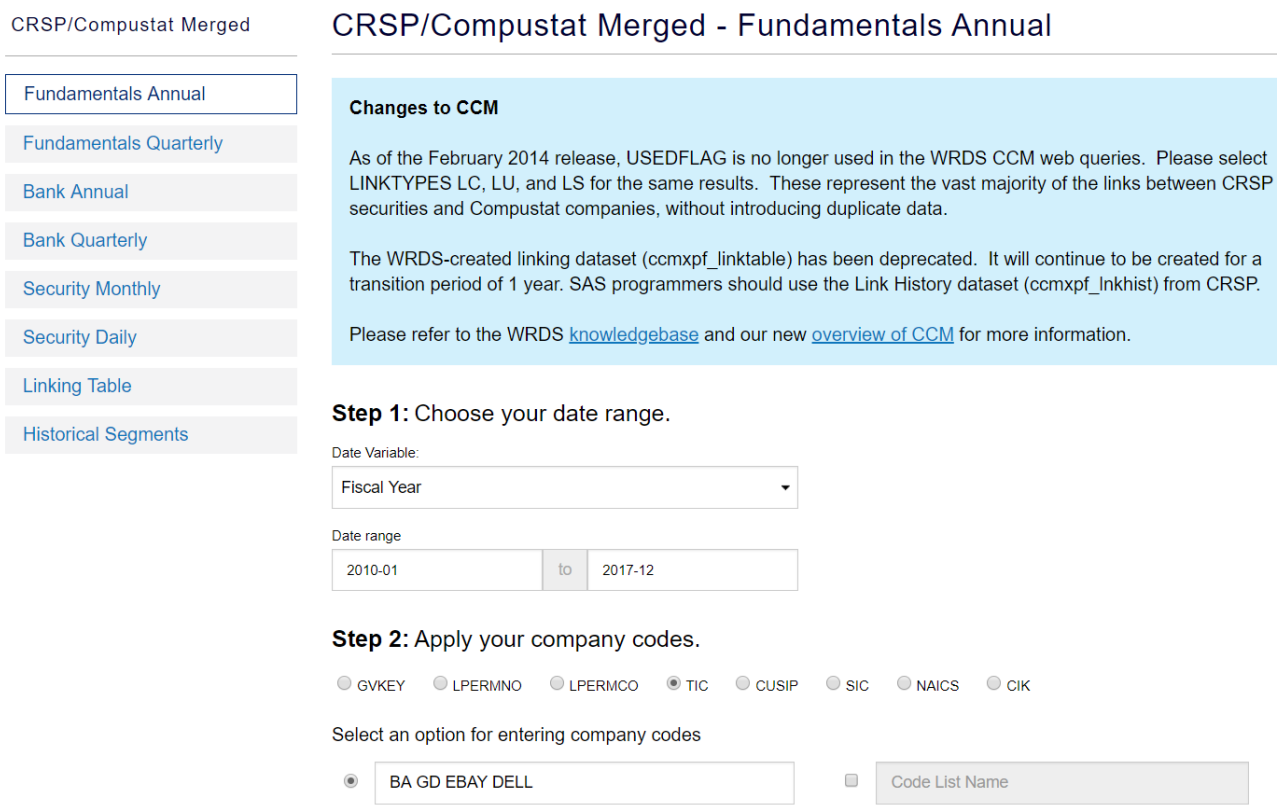
LINKPRIM	Primary issue marker for the link. Based on Compustat Primary/Joiner flag (PRIMISS), indicating whether this link is to Compustat's marked primary security during this range. P = Primary, identified by Compustat in monthly security data. J = Joiner secondary issue of a company, identified by Compustat in monthly security data. C = Primary, assigned by CRSP to resolve ranges of overlapping or missing primary markers from Compustat in order to produce one primary security throughout the company history. N = Secondary, assigned by CRSP to override Compustat. Compustat allows a US and Canadian security to both be marked as Primary at the same time. For Purposes of the link, CRSP allows only one primary at a time and marks the others as N.

Basic data management in STATA (Merge CRSP - Compustat)

CCM – CRSP/Compustat Merged

Example: Get the annual sales and assets for Boeing (BA), Dell (DELL), Ebay (EBAY), and General Dynamics (GD) for the fiscal years 2010:2017.

Dataset: CCM

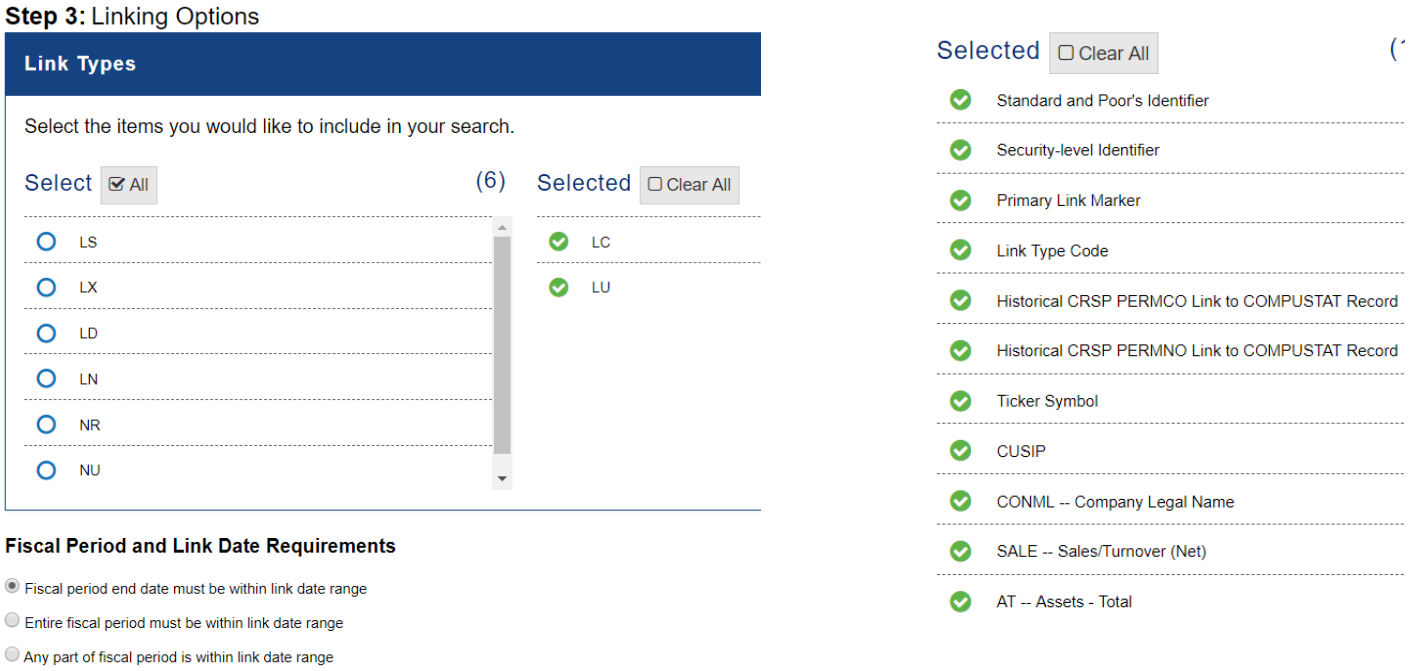


Basic data management in STATA (Merge CRSP - Compustat)

CCM – CRSP/Compustat Merged

Example: Get the annual sales and assets for Boeing (BA), Dell (DELL), Ebay (EBAY), and General Dynamics (GD) for the fiscal years 2010:2017.

Dataset: CCM




```
use "C:\Users\Pedro Pires\Downloads\ccm_crsp_comp_example1.dta", clear

format conm %-24s

br gvkey linkprim liid linktype lpermno lpermco datadate fyear tic conm at sale
```

Basic data management in STATA (Merge CRSP - Compustat)

CCM – CRSP/Compustat Merged

Example: Get the annual sales and assets for Boeing (BA), Dell (DELL), Ebay (EBAY), and General Dynamics (GD) for the fiscal years 2010:2017.

Dataset: CCM

	gvkey	linkprim	liid	linktype	lpermno	lpermco	datadate	fyear	tic	conm	at	sale
1	002285	P	01	LU	19561	20315	31dec2010	2010	BA	BOEING CO	68565	64306
2	002285	P	01	LU	19561	20315	31dec2011	2011	BA	BOEING CO	79986	68735
3	002285	P	01	LU	19561	20315	31dec2012	2012	BA	BOEING CO	88896	81698
4	002285	P	01	LU	19561	20315	31dec2013	2013	BA	BOEING CO	92663	86623
5	002285	P	01	LU	19561	20315	31dec2014	2014	BA	BOEING CO	99198	90762
6	002285	P	01	LU	19561	20315	31dec2015	2015	BA	BOEING CO	94408	96114
7	002285	P	01	LU	19561	20315	31dec2016	2016	BA	BOEING CO	89997	94571
8	002285	P	01	LU	19561	20315	31dec2017	2017	BA	BOEING CO	92333	93392
9	005046	P	01	LC	12052	20791	31dec2010	2010	GD	GENERAL DYNAMICS CORP	32545	32466
10	005046	P	01	LC	12052	20791	31dec2011	2011	GD	GENERAL DYNAMICS CORP	34883	32677
11	005046	P	01	LC	12052	20791	31dec2012	2012	GD	GENERAL DYNAMICS CORP	34309	31682
12	005046	P	01	LC	12052	20791	31dec2013	2013	GD	GENERAL DYNAMICS CORP	35448	31218
13	005046	P	01	LC	12052	20791	31dec2014	2014	GD	GENERAL DYNAMICS CORP	35355	30852
14	005046	P	01	LC	12052	20791	31dec2015	2015	GD	GENERAL DYNAMICS CORP	31997	31469
15	005046	P	01	LC	12052	20791	31dec2016	2016	GD	GENERAL DYNAMICS CORP	32872	31353
16	005046	P	01	LC	12052	20791	31dec2017	2017	GD	GENERAL DYNAMICS CORP	35046	30973
17	014489	P	01	LC	11081	9833	31jan2011	2010	DELL	DELL TECHNOLOGIES INC	38599	61494
18	014489	P	01	LC	11081	9833	31jan2012	2011	DELL	DELL TECHNOLOGIES INC	44533	62071
19	014489	P	01	LC	11081	9833	31jan2013	2012	DELL	DELL TECHNOLOGIES INC	47540	56940
20	114524	P	01	LU	86356	16285	31dec2010	2010	EBAY	EBAY INC	22003.762	9156.273999999999
21	114524	P	01	LU	86356	16285	31dec2011	2011	EBAY	EBAY INC	27320.218	11651.654
22	114524	P	01	LU	86356	16285	31dec2012	2012	EBAY	EBAY INC	37074	14072
23	114524	P	01	LU	86356	16285	31dec2013	2013	EBAY	EBAY INC	41488	16047
24	114524	P	01	LU	86356	16285	31dec2014	2014	EBAY	EBAY INC	45132	17902
25	114524	P	01	LU	86356	16285	31dec2015	2015	EBAY	EBAY INC	17785	8592
26	114524	P	01	LU	86356	16285	31dec2016	2016	EBAY	EBAY INC	23847	8979
27	114524	P	01	LU	86356	16285	31dec2017	2017	EBAY	EBAY INC	25981	9567

✓ Compustat Supplemental Short Interest File

- Data on shorted stocks for listed companies at the New York Stock Exchange, American Stock Exchange, and NASDAQ.
- short interest positions are compiled by the exchanges twice each month, at mid-month and end-month.

ELSEVIER

Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Financial Economics 78 (2005) 243–276

JOURNAL OF Financial ECONOMICS

www.elsevier.com/locate/jfec

Short interest, institutional ownership, and stock returns☆

Paul Asquith[a,b,*], Parag A. Pathak[c], Jay R. Ritter[d]

[a]M.I.T Sloan School of Management, Cambridge, MA 02142, USA
[b]NBER, Cambridge, MA, USA
[c]Harvard University, Cambridge, MA 02138, USA
[d]University of Florida, Gainesville, FL 32611, USA

Received 4 April 2004; received in revised form 21 December 2004; accepted 5 January 2005
Available online 2 August 2005

Abstract

Stocks are short-sale constrained when there is a strong demand to sell short and a limited supply of shares to borrow. Using data on both short interest (a proxy for demand) and institutional ownership (a proxy for supply) we find that constrained stocks underperform during the period 1988–2002 by a significant 215 basis points per month on an equally weighted basis, although by only an insignificant 39 basis points per month on a value-weighted basis. For the overwhelming majority of stocks, short interest and institutional ownership levels make short selling constraints unlikely.

1. Introduction

Shares sold short, as a percentage of shares outstanding, have more than doubled in the last 20 years. In dollar terms, the increase is more than twentyfold. While our understanding of short sales has not increased nearly as much, it is now widely accepted that if short selling is costly and there are heterogeneous investor beliefs, a stock can be overvalued and generate low subsequent returns. This hypothesis originates with Miller (1977), and its cross-sectional predictions have motivated many recent empirical studies.

There are now three distinct empirical literatures on short selling. The oldest contends that high short interest ratios (shares sold short/shares outstanding) forecast low future returns. The second literature investigates the actual cost of short selling by looking at the rebate rate on borrowed stock. The third and most recent empirical short-sale literature assumes that short sales depend on stock ownership by mutual funds and institutions. This literature assumes either that the number of institutions owning a stock is a proxy for heterogeneous expectations or that most lendable shares are from institutional owners.

✓ Merging without CCM

- Compustat and CRSP have different universes with slightly different coverage

	CRSP	COMPUSTAT
Coverage	Stock market data from major stock exchanges (NYSE, AMEX, NASDAQ)	Accounting data for public, OTC, and private companies
Primary Identifiers	PERMNO, PERMCO	GVKEY
Secondary Identifiers	Header & Historical 8-digit CUSIP	Header 9-digit CUSIP
Additional Identifiers	Ticker Symbol, Company Name (Header & Historical)	Ticker Symbol, Company Name (Header only)
Inclusion Policy	All data after IPO	A firm must have some number of years of history as public company before inclusion
Historical Coverage	1925–present	1950–present

- WRDS recommends starting with the SAS sample program merge_funda_crsp_bycusip.sas to merge CRSP and Compustat Xpressfeed databases by CUSIP.

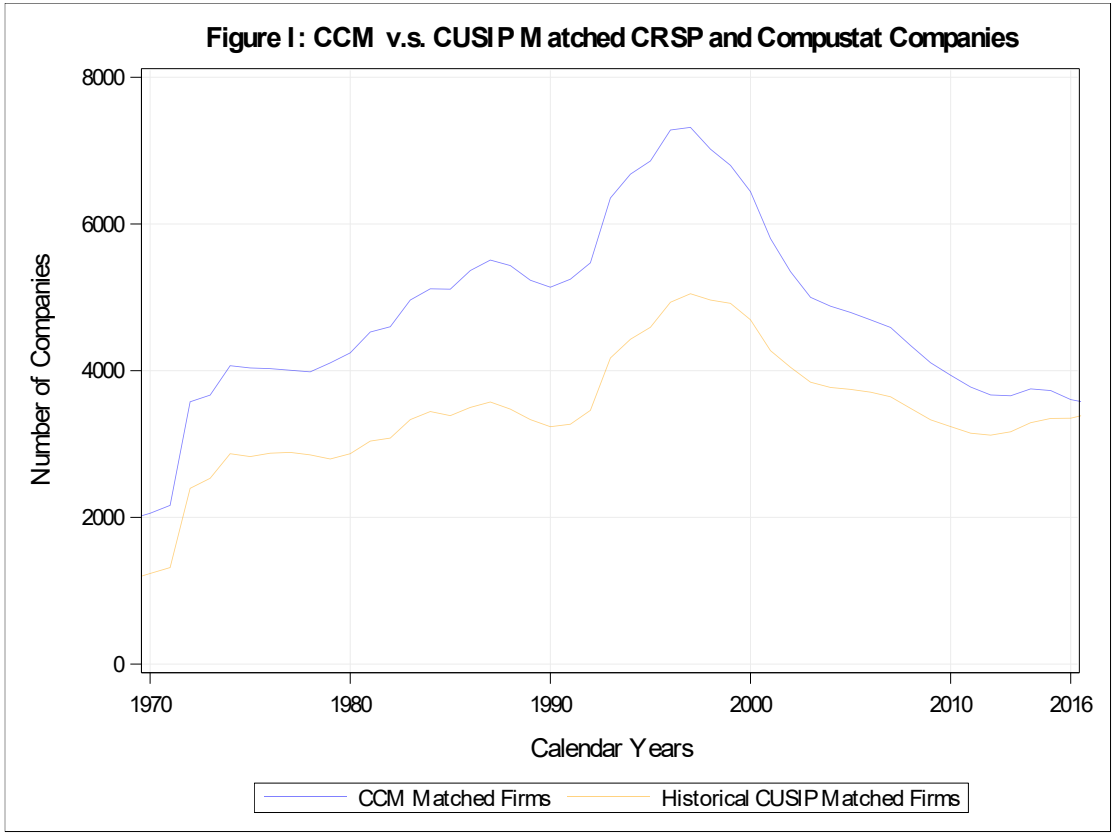
https://wrds-www.wharton.upenn.edu/pages/support/research-wrds/macros/wrds-macro-merge_funda_crsp_bycusipsas/

Merging without CCM

- CCM vs. CUSIP-Matched CRSP and Compustat companies

SAS code provided in:

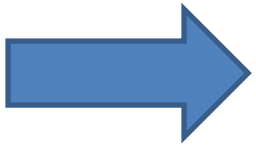
https://wrds-www.wharton.upenn.edu/pages/support/applications/linking-databases/linking-crsp-and-compustat/



Source: SAS code provided by WRDS

Merging without CCM

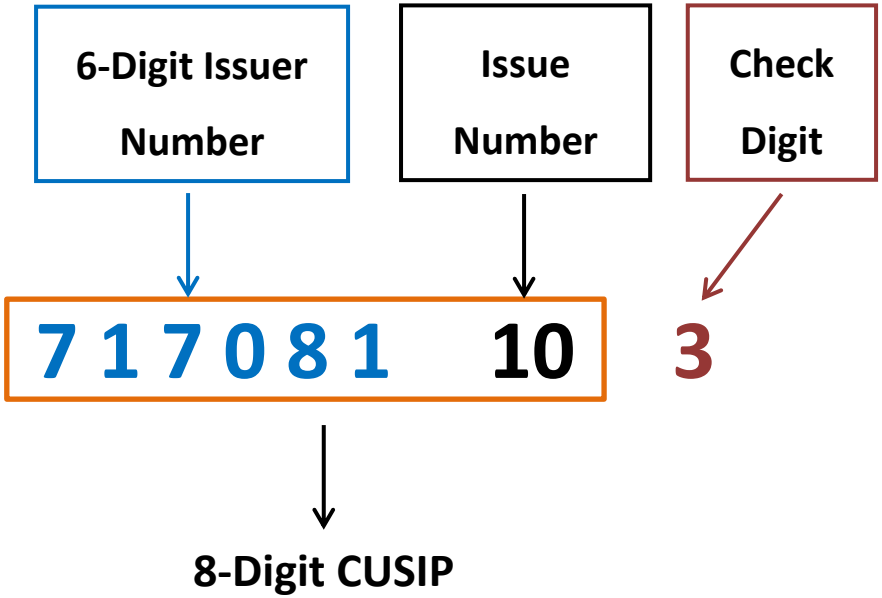
- Compustat data is annual or quarterly, while CRSP data is monthly or daily.
- Different calendar/fiscal regime: Compustat data needs to be converted from fiscal to calendar regime.
- When merging both sets, it is important to ask the question, "When did the market know about the accounting information?".



Merging CRSP and Compustat by CUSIP

Understanding CUSIP's

Example using CUSIP for Pfizer (PFE)

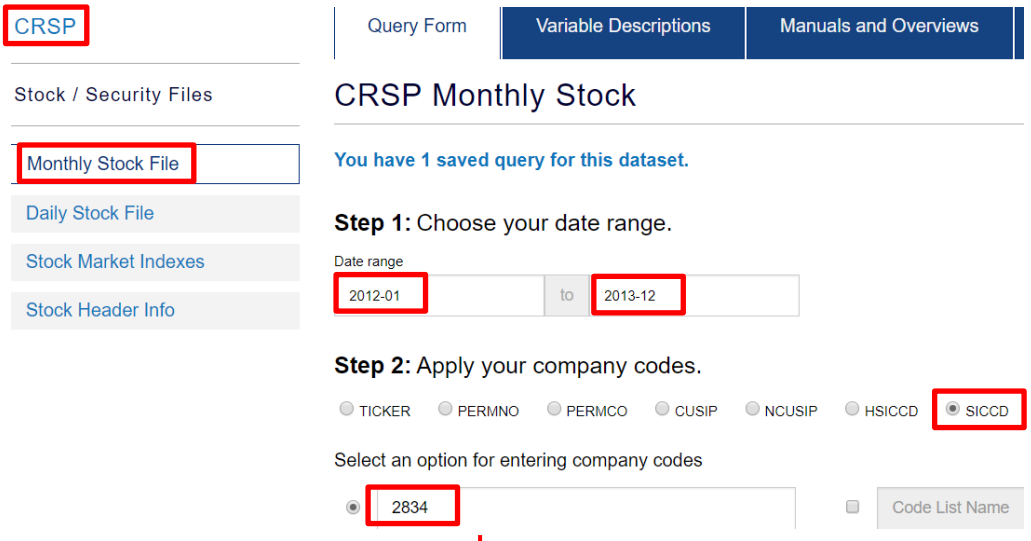


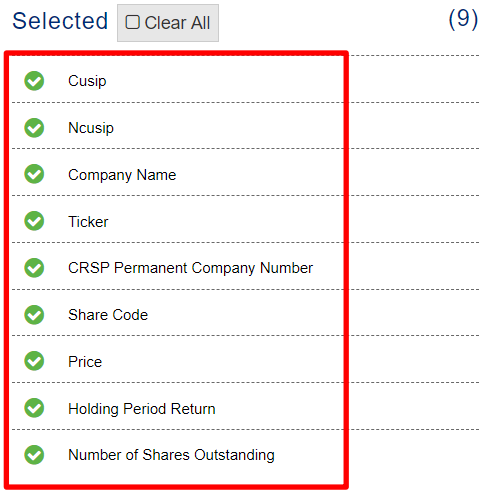
Basic data management in STATA (Merge CRSP - Compustat)

Merge CRSP/Compustat by CUSIP

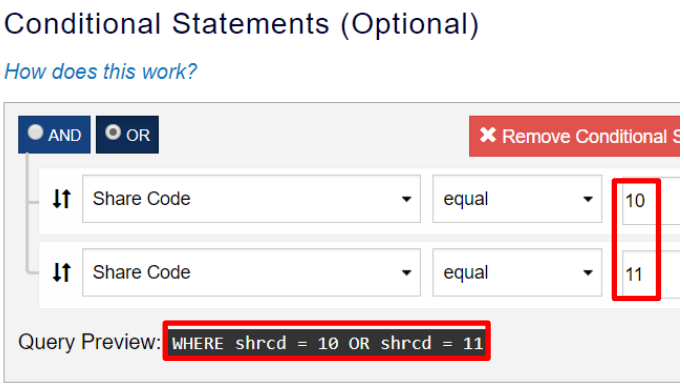
Example: Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

STEP1 - Dataset: CRSP





Extract CRSP Monthly data for pharmaceutical preparation stocks (sic 2834, share code 10/11) for 2 yrs. ending in 2013.

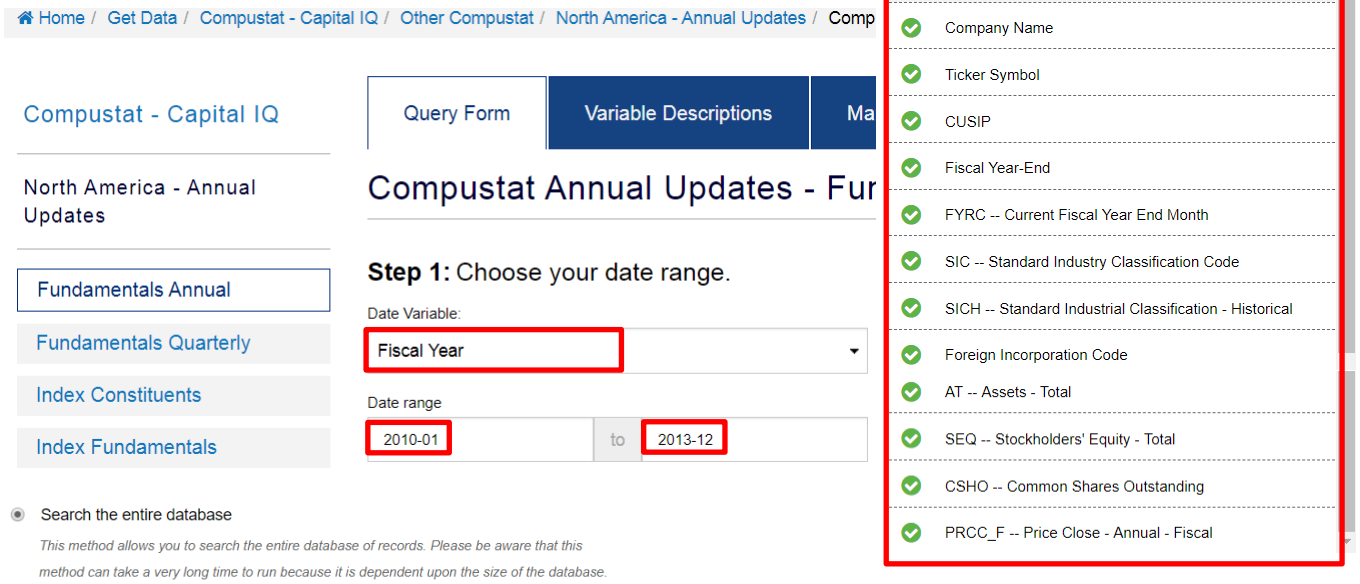


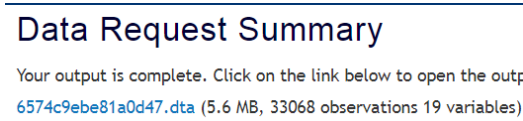
Basic data management in STATA (Merge CRSP - Compustat)

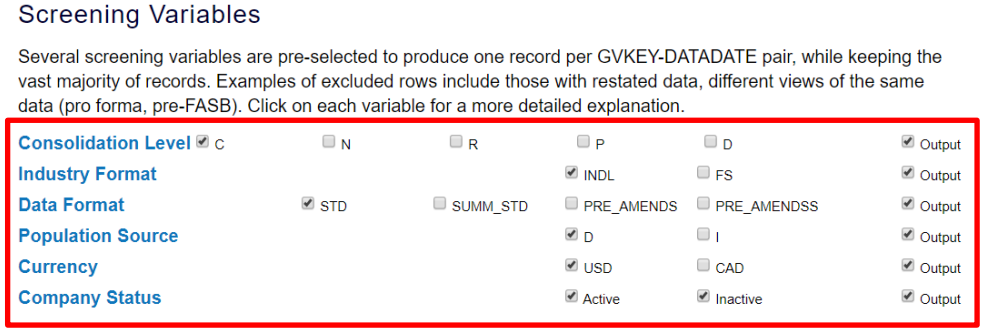
Merge CRSP/Compustat by CUSIP

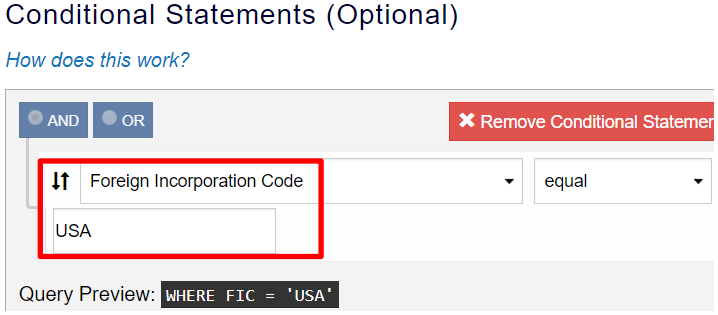
Example: Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

STEP2 - Dataset: Compustat Fundamentals Annual









Merge CRSP/Compustat by CUSIP

Example: Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP

```
clear all
set more off

cd "D:\Documents\FDM - PhD 2020\data\"

* We are going to begin with the Compustat file
use ex1_crspcomp_merge_cusip_comp.dta, clear

* drop records with missing cusip - common identifier in CRSP & COMPUSTAT
drop if cusip==""

* CRSP CUSIP only has 8 digits - Remove last digit from Compustat CUSIP
gen cusip8 = substr(cusip,1,8)

* Check for duplicates on the merging keys (cusip8-fyear)
duplicates report cusip8 fyear

* create new variable comp_fyear with a copy of fyear from Compustat
gen comp_fyear = fyear

* Save file
sa ex1_crspcomp_aux1.dta, replace
```

Basic data management in STATA (Merge CRSP - Compustat)

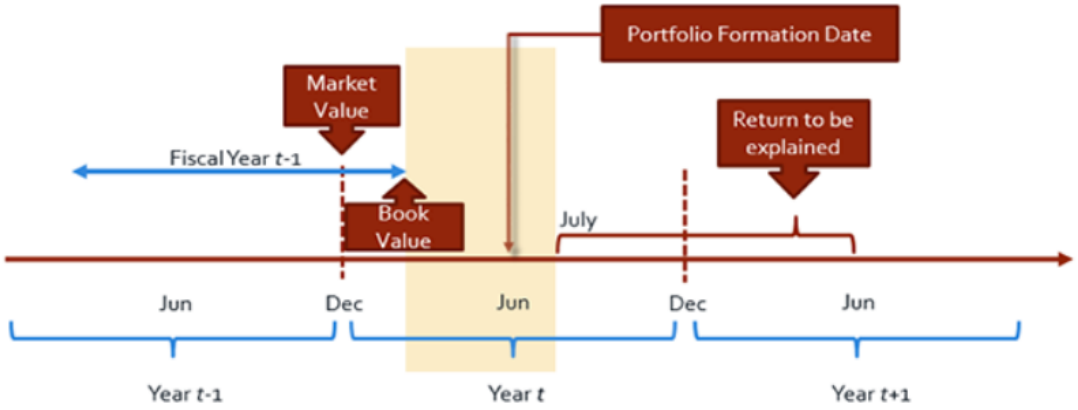
Merge CRSP/Compustat by CUSIP

Example: Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP

To make sure accounting information is publicly available when a new portfolio is formatted, Fama and French (1993) chose the last trading date in June as the portfolio construction date, while using the Market Capitalization for the previous year end, to ensure both data availability and consistency of their empirical analysis.

The following figure demonstrates the logic of time dimension alignment in Fama and French (1993). A more detailed description of this strategy, as well as a reproduction of the methodology, is documented in Fama-French Factors.



Merge CRSP/Compustat by CUSIP

Example: Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP

Compustat fiscal years vs. calendar years

Compustat annual files include a fiscal year-end month variable (fyr) that takes values form 1 (January) to 12 (December) depending on the month in which a company's fiscal year ends.

- For firms with a fiscal year end between January and May, the fiscal year lags one year behind the calendar year
- For firms with a fiscal year end between June and December, the fiscal year is the same as the calendar year

FYR	Fiscal year-end	Calendar year	Fiscal year
1	January 31	t	t-1
2	February 28	t	t-1
3	March 31	t	t-1
4	April 30	t	t-1
5	May 31	t	t-1
6	June 30	t	t
7	July 31	t	t
8	August 31	t	t
9	September 30	t	t
10	October 31	t	t
11	November 30	t	t
12	December 31	t	t

Basic data management in STATA (Merge CRSP - Compustat)

Example (1): Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP

```
* Now, We are going to open the CRSP file
use ex1_crspcomp_merge_cusip_crsp.dta, clear

* drop recors with missing cusip - common identifier in CRSP & COMPUSTAT
drop if cusip==""

* create a new variable cusip8 (copy of cusip) to facilitate the merge command
gen cusip8 = cusip

* Check for duplicates
duplicates report cusip8 date

* You need to think on how you want to merge the CRSP data with the Compustat data
* In CRSP you'll get daily or monthly data in a calendar regime
* In Compustat you'll get annual or quarterly data in a fiscal regime
* In compustat datadate identifies the calendar date of the financial report
* For illustrative purposes only, in this example we will merge CRSP market data
* on calendar months from July of year(t) to June of year(t+1)
* with Compustat accounting data in fiscal year (t-1)
* Accordingly, create a key variable with the time identifier needed to merge with Compustat
gen fyear = year(date)-1
replace fyear = year(date)-2 if month(date) <= 6
```

1

Basic data management in STATA (Merge CRSP - Compustat)

Example (1): Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP

```
* Now, we can merge CRSP with Compustat on cusip8-fyear

* issue the merge command
merge m:1 cusip8 fyear using ex1_crspcomp_aux1.dta, keep(match master)

* drop observations without a match
drop if _merge!=3

format comnam %32s
format conm %32s

sort permno date

gen yearfrac = (date - datadate)/365

su yearfrac, d

br permno date cusip ticker comnam ret gvkey comp fyear datadate conm at
```

1

Syntax

One-to-one merge on specified key variables

`merge 1:1` *varlist* `using` *filename* [, *options*]

Many-to-one merge on specified key variables

`merge m:1` *varlist* `using` *filename* [, *options*]

Basic data management in STATA (Merge CRSP - Compustat)

Example (1): Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

1

STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP

permno	date	cusip	comnam	ret	gvkey	comp_fyear	datadate	at
10116	31jan2012	00509L80	ACURA PHARMACEUTICALS INC	-.051575951	011929	2010	31dec2010	25.493
10116	29feb2012	00509L80	ACURA PHARMACEUTICALS INC	-.021148017	011929	2010	31dec2010	25.493
10116	30mar2012	00509L80	ACURA PHARMACEUTICALS INC	.067901246	011929	2010	31dec2010	25.493
10116	30apr2012	00509L80	ACURA PHARMACEUTICALS INC	-.066473991	011929	2010	31dec2010	25.493
10116	31may2012	00509L80	ACURA PHARMACEUTICALS INC	-.126935005	011929	2010	31dec2010	25.493
10116	29jun2012	00509L80	ACURA PHARMACEUTICALS INC	.113475241	011929	2010	31dec2010	25.493
10116	31jul2012	00509L80	ACURA PHARMACEUTICALS INC	-.407643318	011929	2011	31dec2011	37.173
10116	31aug2012	00509L80	ACURA PHARMACEUTICALS INC	-.123655923	011929	2011	31dec2011	37.173
10116	28sep2012	00509L80	ACURA PHARMACEUTICALS INC	.067484669	011929	2011	31dec2011	37.173
10116	31oct2012	00509L80	ACURA PHARMACEUTICALS INC	-.132183924	011929	2011	31dec2011	37.173
10116	30nov2012	00509L80	ACURA PHARMACEUTICALS INC	.211920559	011929	2011	31dec2011	37.173
10116	31dec2012	00509L80	ACURA PHARMACEUTICALS INC	.213114738	011929	2011	31dec2011	37.173
10116	31jan2013	00509L80	ACURA PHARMACEUTICALS INC	-.157657668	011929	2011	31dec2011	37.173
10116	28feb2013	00509L80	ACURA PHARMACEUTICALS INC	.101604246	011929	2011	31dec2011	37.173
10116	28mar2013	00509L80	ACURA PHARMACEUTICALS INC	.033980668	011929	2011	31dec2011	37.173
10116	30apr2013	00509L80	ACURA PHARMACEUTICALS INC	.150234699	011929	2011	31dec2011	37.173
10116	31may2013	00509L80	ACURA PHARMACEUTICALS INC	-.077551045	011929	2011	31dec2011	37.173
10116	28jun2013	00509L80	ACURA PHARMACEUTICALS INC	-.168141589	011929	2011	31dec2011	37.173
10116	31jul2013	00509L80	ACURA PHARMACEUTICALS INC	.085106365	011929	2012	31dec2012	29.054

t = 2012 / Jul2012 – Jun2013 / fyear t-1 = 2011

Basic data management in STATA (Merge CRSP - Compustat)

Example (1): Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP

1

```
. su yearfrac, d

                          yearfrac
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .5808219       .1671233
 5%     .5835617       .2520548
10%     .6630137       .3287671       Obs               2,851
25%     .8328767       .3342466       Sum of Wgt.       2,851

50%     1.084931                      Mean           1.079154
                       Largest        Std. Dev.      .3248256
75%     1.331507              2
90%     1.493151              2       Variance       .1055117
95%     1.506849              2       Skewness       .2090855
99%     1.920548              2       Kurtosis       2.500372

.
. br permno date cusip ticker comnam ret gvkey comp_fyear datadate conm at
```

Basic data management in STATA (Merge CRSP - Compustat)

Example (2): Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013. **Most recent report as long as it's at least 3 months old. – JOINBY --**

STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP

```
clear all
set more off

cd "D:\Documents\FDM - PhD 2020\data"

* open CRSP file
use ex1_crspcomp_merge_cusip_crsp.dta, clear

* drop recors with missing cusip - common identifier in CRSP & COMPUSTAT
drop if cusip==""

* create a new variable cusip8 (copy of cusip) to facilitate the merge command
gen cusip8 = cusip

* Check for duplicates
duplicates report cusip8 date

* check distinct # obs. before joinby
distinct cusip8 date, joint

* For illustrative purposes only, in this example we will merge CRSP market data,
* for each CRSP date, with the most recent financial report provided that
* the report is at least 3 months old (consider a year fraction of 0.25)

* issue the joinby command by cusip8 to create a cartesian product
* that includes every compustat observations with the same cusip8 for each cusip8-date of CRSP
joinby cusip8 using ex1_crspcomp_aux1.dta, unmatched(master)

* check distinct # obs. after joinby
distinct cusip8 date, joint
```

Basic data management in STATA (Merge CRSP - Compustat)

Example (2): Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP

2

```
* apx. year fraction difference in days / 365 days
gen yearfrac = (date - datadate)/365

* drop if compustat datadate is missing
drop if datadate == .

* drop if financial data is less than 3 months old
* Prior to 2003, most public companies were allowed to file the annual meeting report (10K) within 90 days
* of fiscal year-end. After 2005 certain large companies were required to file their 10K within 60 days.
drop if yearfrac < 0.25

* sort by permno date and descendind compustat report date
gsort permno date -datadate

* count distinct permno-date before dropping duplicates
distinct permno date, joint

duplicates drop permno date, force

* count distinct permno-date after dropping duplicates
distinct permno date, joint

* Check the nr. ob observations following the merge command
su

format comnam %32s
format conm %32s

sort permno date
br permno date cusip ticker comnam ret gvkey comp_fyear datadate conm at
su yearfrac, d
```

Instead of year fraction (yearfrac) you can use something more precise:

gen ddays = date - datadate

drop if ddays < 90

Basic data management in STATA (Merge CRSP - Compustat)

Example (2): Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

2

STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP

permno	date	cusip	ticker	comnam	ret	gvkey	comp_fyear	datadate	at
10116	31jan2012	00509L80	ACUR	ACURA PHARMACEUTICALS INC	-.051575951	011929	2010	31dec2010	25.493
10116	29feb2012	00509L80	ACUR	ACURA PHARMACEUTICALS INC	-.021148017	011929	2010	31dec2010	25.493
10116	30mar2012	00509L80	ACUR	ACURA PHARMACEUTICALS INC	.067901246	011929	2010	31dec2010	25.493
10116	30apr2012	00509L80	ACUR	ACURA PHARMACEUTICALS INC	-.066473991	011929	2011	31dec2011	37.173
10116	31may2012	00509L80	ACUR	ACURA PHARMACEUTICALS INC	-.126935005	011929	2011	31dec2011	37.173
10116	29jun2012	00509L80	ACUR	ACURA PHARMACEUTICALS INC	.113475241	011929	2011	31dec2011	37.173
10116	31jul2012	00509L80	ACUR	ACURA PHARMACEUTICALS INC	-.407643318	011929	2011	31dec2011	37.173
10116	31aug2012	00509L80	ACUR	ACURA PHARMACEUTICALS INC	-.123655923	011929	2011	31dec2011	37.173
10116	28sep2012	00509L80	ACUR	ACURA PHARMACEUTICALS INC	.067484669	011929	2011	31dec2011	37.173
10116	31oct2012	00509L80	ACUR	ACURA PHARMACEUTICALS INC	-.132183924	011929	2011	31dec2011	37.173
10116	30nov2012	00509L80	ACUR	ACURA PHARMACEUTICALS INC	.211920559	011929	2011	31dec2011	37.173
10116	31dec2012	00509L80	ACUR	ACURA PHARMACEUTICALS INC	.213114738	011929	2011	31dec2011	37.173
10116	31jan2013	00509L80	ACUR	ACURA PHARMACEUTICALS INC	-.157657668	011929	2011	31dec2011	37.173
10116	28feb2013	00509L80	ACUR	ACURA PHARMACEUTICALS INC	.101604246	011929	2011	31dec2011	37.173
10116	28mar2013	00509L80	ACUR	ACURA PHARMACEUTICALS INC	.033980668	011929	2011	31dec2011	37.173
10116	30apr2013	00509L80	ACUR	ACURA PHARMACEUTICALS INC	.150234699	011929	2012	31dec2012	29.054
10116	31may2013	00509L80	ACUR	ACURA PHARMACEUTICALS INC	-.077551045	011929	2012	31dec2012	29.054

31jan2012 / most recent report date 31dec2011 / 1 month difference only

30apr2012 / most recent report date 31dec2011 / 4 months difference

Basic data management in STATA (Merge CRSP - Compustat)

Example (2): Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP

2

```
. su yearfrac, d

                          yearfrac
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .3287671       .2520548
 5%     .3315068       .2520548
10%     .4136986       .2520548       Obs                2,863
25%     .4986301       .2520548       Sum of Wgt.        2,863

50%     .7506849                      Mean           .7851856
                       Largest        Std. Dev.      .2908822
75%      1.00274       1.334247
90%     1.164384       1.589041       Variance       .0846124
95%     1.241096       1.668493       Skewness       .0142461
99%     1.246575       1.750685       Kurtosis       1.835334
```

Fuzzy matching in STATA – Database management functions

CROSS

Cartesian Product → Generate all possible combinations → CROSS

[D] cross — Form every pairwise combination of two datasets

Syntax

cross using *filename*

Menu

Data > Combine datasets > Form every pairwise combination of two datasets

Description

cross forms every pairwise combination of the data in memory with the data in *filename*.

Crossing refers to merging two datasets in every way possible. That is, the first observation of the data in memory is merged with every observation of *filename*, followed by the second, and so on. Thus the result will have $N_1 N_2$ observations, where N_1 and N_2 are the number of observations in memory and in *filename*, respectively.

Fuzzy matching in STATA – Database management functions

JOINBY

Cartesian Product → Generate all possible combinations within groups → JOINBY

Title

[D] **joinby** — Form all pairwise combinations within groups

Syntax

joinby [*varlist*] **using** *filename* [, *options*]

options	Description
Options	
When observations match:	
update	replace missing data in memory with values from *filename*
replace	replace all data in memory with values from *filename*
When observations do not match:	
unmatched(none)	ignore all; the default
unmatched(both)	include from both datasets
unmatched(master)	include from data in memory
unmatched(using)	include from data in *filename*
_merge(*varname***)**	*varname* marks source of resulting observation; default is **_merge**
nolabel	do not copy value-label definitions from *filename*

varlist may not contain **strL**s.

Basic data management in STATA (Merging datasets)

STATA – edit – copy/paste

	A	B	C	D	E	F	G	H
1		ds1				ds2		
2	ticker	conm	state		tic	name	state	cstate
3	NVDA	NVIDIA	CA		NVDA	NVIDIA	CA	CA
4	WFC	WELLS FARGO	CA		WFC	WELLS FARGO	CA	CA
5	INTC	INTEL	CA		INTC	INTEL	CA	CA
6	AAPL	APPLE	CA		AAPL	APPLE	CA	CA
7	BA	BOEING	IL		BA	BOEING	IL	IL
8	CAT	CATERPILLAR	IL		CAT	CATERPILLAR	IL	IL
9	HON	HONEYWELL	NJ		HON	HONEYWELL	NJ	NJ
10	JNJ	J&J	NJ		JNJ	J&J	NJ	NJ
11	JPM	JPMORGAN	NY		JPM	JPMORGAN	NY	NY
12	XOM	EXXON MOBIL	TX		KO	COCA-COLA	GA	GA
13					MSFT	MICROSOFT	WA	WA

	ticker	conm	state
1	NVDA	NVIDIA	CA
2	WFC	WELLS FARGO	CA
3	INTC	INTEL	CA
4	AAPL	APPLE	CA
5	BA	BOEING	IL
6	CAT	CATERPILLAR	IL
7	HON	HONEYWELL	NJ
8	JNJ	J&J	NJ
9	JPM	JPMORGAN	NY
10	XOM	EXXON MOBIL	TX

	tic	name	state	cstate
1	NVDA	NVIDIA	CA	CA
2	WFC	WELLS FARGO	CA	CA
3	INTC	INTEL	CA	CA
4	AAPL	APPLE	CA	CA
5	BA	BOEING	IL	IL
6	CAT	CATERPILLAR	IL	IL
7	HON	HONEYWELL	NJ	NJ
8	JNJ	J&J	NJ	NJ
9	JPM	JPMORGAN	NY	NY
10	KO	COCA-COLA	GA	GA
11	MSFT	MICROSOFT	WA	WA

Basic data management in STATA (Merging datasets)

```
* ####################################################
* CROSS - example
* ####################################################

use "C:\Users\Pedro Pires\Desktop\ds1.dta", clear

cross using "C:\Users\Pedro Pires\Desktop\ds2.dta"

sort ticker name

br
```

	ticker	conm	state	tic	name	cstate
1	AAPL	APPLE	CA	AAPL	APPLE	CA
2	AAPL	APPLE	CA	BA	BOEING	IL
3	AAPL	APPLE	CA	CAT	CATERPILLAR	IL
4	AAPL	APPLE	CA	KO	COCA-COLA	GA
5	AAPL	APPLE	CA	HON	HONEYWELL	NJ
6	AAPL	APPLE	CA	INTC	INTEL	CA
7	AAPL	APPLE	CA	JNJ	J&J	NJ
8	AAPL	APPLE	CA	JPM	JPMORGAN	NY
9	AAPL	APPLE	CA	MSFT	MICROSOFT	WA
10	AAPL	APPLE	CA	NVDA	NVIDIA	CA
11	AAPL	APPLE	CA	WFC	WELLS FARGO	CA
12	BA	BOEING	IL	AAPL	APPLE	CA
13	BA	BOEING	IL	BA	BOEING	IL
14	BA	BOEING	IL	CAT	CATERPILLAR	IL
15	BA	BOEING	IL	KO	COCA-COLA	GA
16	BA	BOEING	IL	HON	HONEYWELL	NJ
17	BA	BOEING	IL	INTC	INTEL	CA
18	BA	BOEING	IL	JNJ	J&J	NJ
19	BA	BOEING	IL	JPM	JPMORGAN	NY
20	BA	BOEING	IL	MSFT	MICROSOFT	WA
21	BA	BOEING	IL	NVDA	NVIDIA	CA
22	BA	BOEING	IL	WFC	WELLS FARGO	CA

Basic data management in STATA (Merging datasets)

```
* ####################################################
* JOINBY - example
* ####################################################

use "C:\Users\Pedro Pires\Desktop\ds1.dta", clear

joinby state using "C:\Users\Pedro Pires\Desktop\ds2.dta", unmatched(master)

sort ticker name

br
```

	ticker	conm	state	_merge	tic	name	cstate
1	AAPL	APPLE	CA	both in master and using data	AAPL	APPLE	CA
2	AAPL	APPLE	CA	both in master and using data	INTC	INTEL	CA
3	AAPL	APPLE	CA	both in master and using data	NVDA	NVIDIA	CA
4	AAPL	APPLE	CA	both in master and using data	WFC	WELLS FARGO	CA
5	BA	BOEING	IL	both in master and using data	BA	BOEING	IL
6	BA	BOEING	IL	both in master and using data	CAT	CATERPILLAR	IL
7	CAT	CATERPILLAR	IL	both in master and using data	BA	BOEING	IL
8	CAT	CATERPILLAR	IL	both in master and using data	CAT	CATERPILLAR	IL
9	HON	HONEYWELL	NJ	both in master and using data	HON	HONEYWELL	NJ
10	HON	HONEYWELL	NJ	both in master and using data	JNJ	J&J	NJ
11	INTC	INTEL	CA	both in master and using data	AAPL	APPLE	CA
12	INTC	INTEL	CA	both in master and using data	INTC	INTEL	CA
13	INTC	INTEL	CA	both in master and using data	NVDA	NVIDIA	CA
14	INTC	INTEL	CA	both in master and using data	WFC	WELLS FARGO	CA
15	JNJ	J&J	NJ	both in master and using data	HON	HONEYWELL	NJ
16	JNJ	J&J	NJ	both in master and using data	JNJ	J&J	NJ
17	JPM	JPMORGAN	NY	both in master and using data	JPM	JPMORGAN	NY
18	NVDA	NVIDIA	CA	both in master and using data	AAPL	APPLE	CA
19	NVDA	NVIDIA	CA	both in master and using data	INTC	INTEL	CA
20	NVDA	NVIDIA	CA	both in master and using data	NVDA	NVIDIA	CA
21	NVDA	NVIDIA	CA	both in master and using data	WFC	WELLS FARGO	CA
22	WFC	WELLS FARGO	CA	both in master and using data	AAPL	APPLE	CA
23	WFC	WELLS FARGO	CA	both in master and using data	INTC	INTEL	CA
24	WFC	WELLS FARGO	CA	both in master and using data	NVDA	NVIDIA	CA
25	WFC	WELLS FARGO	CA	both in master and using data	WFC	WELLS FARGO	CA
26	XOM	EXXON MOBIL	TX	only in master data			

✓ Fuzzy matching in STATA

- Using common identifiers (such as ISIN, CUSIP, and SEDOL) is the best way to merge datasets from different providers.
- In practical applications, however, sometimes one has to merge datasets which only contain a string (e.g. company name, fund name). In these cases, fuzzy matching algorithms provide similarity scores that help identifying correct matches.

Matchit — Matches two columns or two datasets based on similar text patterns

Syntax

Data in two columns in the same dataset

matchit *varname1 varname2* [*, options*]

Data in two different datasets (with indexation)

matchit *idmaster txtmaster* **using** *filename.dta* , **idusing**(*varname*) **txtusing**(*varname*) [*options*]

options	Description
similmethod(*simfcn*)	String matching method. Default is *bigram*. Other built-in *simfcn* are: *ngram*, *ngram_circ*, *token*, *soundex* and *token_soundex*.
score(*scrfcn*)	Specifies similarity score. Default is *jaccard*. Other built-in options are *simple* and *minsimple*.
weights(*wgtfcn*)	Weighting transformation. Default is *noweights*. Built-in options are *simple*, *log* and *root*.
generate(*varname*)	Specifies the name for the similarity score variable. Default is *similscore*.

Fuzzy matching in STATA

Decomposition of text into *grams* of 2 moving chars (*bigrams*)

Weights: based on *grams* frequency (default: no weights – all *grams* eq. 1)

Description

matchit provides a similarity score between two different text strings by performing many different string-based matching techniques. It returns a new numeric variable (*similscore*) containing the similarity score, which ranges from 0 to 1. A *similscore* of 1 implies a perfect similarity according to the string matching technique chosen and decreases when the match is less similar. *similscore* is a relative measure which can (and often do) change depending on the technique chosen. For more information on these techniques refer to Raffo & Lhuillery (2009).

Examples for "John Smith":

# grams	bigram	token	soundex	token_ soundex	ngram,1	ngram,3
1	Jo	John	J525	J500	J	Joh
2	oh	Smith		S530	o	ohn
3	hn				h	hn_
4	n_				n	n_S
5	_S				_	_Sm
6	Sm				S	Smi
7	mi				m	mit
8	it				i	ith
9	th				t	
10					h	
11						

Notes: "_" = a blank space.
ngram, 2 is equivalent to bigram.

Fuzzy matching in STATA - GED (Generalized Edit Distance)

Obs	String1	String2	Generalized Edit Distance	Operation
1	baboon	baboon	0	match
2	baXboon	baboon	100	insert
3	baoon	baboon	100	delete
4	baXoon	baboon	100	replace
5	baboonX	baboon	50	append
6	baboo—	baboon	10	truncate
7	babboon	baboon	20	double
8	babon	baboon	20	single
9	baobon	baboon	20	swap
10	bab oon	baboon	10	blank
11	bab,oon	baboon	30	punctuation
12	bXaoon	baboon	200	insert+delete
13	bXaYoon	baboon	200	insert+replace
14	bXoon	baboon	200	delete+replace
15	Xbaboon	baboon	200	finsert
16	aboon	baboon	200	trick question: swap+delete
17	Xaboon	baboon	200	freplace
18	axoon	baboon	300	fdelete+replace
19	axoo	baboon	310	fdelete+replace+truncate
20	axon	baboon	320	fdelete+replace+single
21	baby	baboon	120	replace+truncate*2
22	balloon	baboon	200	replace+insert

Fuzzy matching in STATA - Levenshtein distance

Calculates the distance between strings using the Levenshtein distance metric. Levenshtein distance, or edit distance, is the smallest number of edits required to make one string match a second string. An edit may be an insertion, deletion, or substitution of any single letter.

```
help strdist
```

Title

strdist — Calculate the Levenshtein distance, or edit distance, between strings.

Syntax

strdist {*varname1*|"*string1*"} {*varname2*|"*string2*"} [*if*] [*in*] [, **generate**(*newvar*) **maxdist**(*integer*)]

Fuzzy matching in STATA

COMPGED Function

The COMPGED function returns the generalized edit distance between two strings. Specifically, the COMPGED function returns a generalization of the Levenshtein edit distance, which is a measure of dissimilarity between two strings. The Levenshtein edit distance is the number of operations (deletions, insertions, or replacement) of a single characters that are required to transform string-1 into string-2.

Each operation basically 'costs' a certain value. For example, if *string-1* is the result of inserting a space into *string-2*, this has a cost of 10. The more dramatic the operation, the greater the cost. The COMPGED will return the total cost for all operations that occur. The costs returned by COMPGED can be altered by using CALL COMPCOST so that the cost are specific to your needs. A common use I have seen for using the COMPGED function is using it to compare email addresses.

```
email1='JohnDoe@abc.com';
email2='John_Doe@abc.com';
cost=compged(email1,email2);
```

The value of COST will be 30 which is the cost of adding punctuation to a string.

COMPLEV Function

The COMPLEV function is very similar to the COMPGED function. The difference is that the Levenshtein edit distance that is computed by COMPLEV is a special case of the generalized edit distance that is computed by COMPGED. The result is the COMPLEV executes much more quickly than COMPGED. However, the COMPLEV function is not as powerful or versatile as the COMPGED function. The COMPLEV function is generally most useful when comparing simple strings and when speed of comparison is important.

Contents coverage

- **Institutional ownership and mutual fund holdings**
  - **CRSP Mutual Funds**
  - **Thomson Reuters - Mutual Fund Holdings**
  - **Lipper**
  - **Thomson Reuters Institutional Holdings**
  - **FactSet LionShares Ownership**

CRSP Mutual Funds

- **Mutual Fund data (1962-2018)**
  - Survivorship-bias-free data
  - Historical performance of open-end mutual funds
  - History of each mutual fund's name, investment style, fee structure, and asset allocation.
  - Monthly total returns, monthly total net assets, monthly/daily net asset values, and dividends.
  - https://wrds-www.wharton.upenn.edu/documents/410/CRSP_MFDB_Guide.pdf

CRSP Mutual Funds

NOTE ABOUT RETURNS:

Daily and monthly returns values are calculated as a change in NAV including reinvested dividends from one period to the next. Navs are net of all management expenses and 12b-fees. Front and rear load fees are excluded.

Returns are calculated as follows:

$$R_t = \left[\frac{Nav_t * cumfact}{Nav_{t-1}} \right] - 1$$

DATABASE TABLES OVERVIEW

TABLE	NAME	DEFINITION
contact_info	Contact Information	Current and historical contact information
daily_nav	Daily Net Asset Value	Net Asset Value for each trading day
daily_returns	Daily Returns	Returns for each trading day
dividends	Dividends	Fund dividends
front_load_det	Front Load Detail	Details of front load fees
front_load_grp	Front Load Group	Effective dates for front load fees
fund_fees	Fund Fees	Fees associated with each fund
fund_hdr	Fund Header	Most recent identification information for each fund
fund_hdr_hist	Historical Fund Header	Historical identification information for each fund
fund_style	Fund Style	Style attributes for each fund
fund_summary	Fund Summary	Summary data for each fund
holdings	Holdings	Portfolio holding information
holdings_co_info	Holdings Company Information	Information about companies held in portfolios
crsp_portno_map	CRSP PORTNO Map	Map to portfolio for security holdings info
monthly_nav	Monthly Net Asset Value	Net Asset Values as of the last trading day of each month
monthly_returns	Monthly Returns	Monthly holding period returns
monthly_tna	Monthly Total Net Assets	Total Net Assets as of the last trading day of each month
rear_load_det	Rear Load Detail	Details of rear load fees
rear_load_grp	Rear Load Group	Effective dates for rear load fees

CRSP Mutual Funds

FUND FEES "FUND_FEES"

NAME	DATATYPE	DEFINITION
*crsp_fundno	INTEGER	Unique identifier for fund
*begdt	DATE	Beginning of date range for fee information. For annual data, information is valid for the year portion of the begin date.
enddt	DATE	End of date range for fee information
actual_12b1	FLOAT	Reported as the ratio of the total assets attributed to marketing and distribution costs. Represents the actual fee paid in the most recently completed fiscal year as reported in the Annual Report Statement of Operations. Represented in decimal format. • -99 & 0 (prior to 1998) are used when no 12b-1 fee is reported.
max_12b1	FLOAT	Maximum contractual 12b-1 fee. Represented in decimal format. • A fund with an actual 12b-1 value of -99 may have a max_12b1 fee indicating the maximum even though none is currently being reported. • Null values are used when no maximum value is reported.
exp_ratio	FLOAT	Expense Ratio as of the most recently completed fiscal year. Represented in decimal format. Ratio of total investment that shareholders pay for the fund's operating expenses, which include 12b-1 fees exp_ratio may include waivers and reimbursements, causing it to appear to be less than the fund management fee.
mgmt_fee	FLOAT	Management fee ($)/ Average Net Assets ($) The fee is calculated using ratios based on the line items reported in the Statement of Operations. The management fee can be offset by fee waivers and/or reimbursements which will make this value differ from the contractual fees found in the prospectus. Reimbursements can lead to negative Management Fees.

NAME	DATATYPE	DEFINITION
turn_ratio	FLOAT	Fund Turnover Ratio. Minimum (of aggregated sales or aggregated purchases of securities), divided by the average 12-month Total Net Assets of the fund. If fiscal_yearend is present, turn_ratio is for the twelve months ending on fiscal_yearend. If fiscal_yearend is missing, then turn_ratio is for the twelve months ending on the begdt.

CRSP Mutual Funds

NAME	DATATYPE	DEFINITION	DATA AVAILABILITY
fund_name	VARCHAR(140)	The full name of the fund	
nasdaq	VARCHAR (5)	NASDAQ ticker symbol. Available for NASDAQ listed funds only.	
ncusip	VARCHAR(9)	Fund CUSIP. Available for internal use by client only. Go to www.crsp.ChicagoBooth.edu/crsp/policies/index.html for CUSIP licensing information.	
first_offer_dt	DATE	Date when the fund was first offered	
mgmt_name	VARCHAR(80)	Management Company Name	
mgmt_cd	VARCHAR(4)	Management Company Code. An identifier for the Management Company.	December 1999
mgr_name	VARCHAR(30)	Portfolio Manager Name	
mgr_dt	DATE	Date that the current portfolio manager assumed responsibility for the portfolio	
adv_name	VARCHAR(40)	Fund Advisor Name	December 1999
open_to_inv	VARCHAR(1)	Identifies if the fund is open to investors: Y = Yes N = No	December 1999
retail_fund	VARCHAR(1)	Identifies if a fund is a retail fund or not: Y = Yes N = No	December 1999
Inst_fund	VARCHAR(1)	Identifies if a fund is an institutional fund: Y = Yes N = No	December 1999

NOVA School of Business & Economics — Shaping powerful minds

CRSP Mutual Funds

NAME	DATATYPE	DEFINITION	DATA AVAILABILITY
m_fund	VARCHAR(1)	Identifies if a fund was originally a "M" fund: Y = Yes N = No Note: "M" funds are "dead" funds that were introduced to the database when originally built by Mark Carhart in order to eliminate survivor bias.	
index_fund_flag	VARCHAR(3)	Identifies if a fund is an index fund: B = Index-based fund - utilizes indexes as its primary filter for the purchase and sale of securities. This is accomplished by investing in the components of one or more indexes, or by investing in a small percentage of securities within the index in an attempt to capture the best performers. In each case, the option to invest a portion of assets outside the securities held by the index is left open. D = Pure Index fund - objective is to match the total investment performance of a publicly recognized securities market index. The fund will hold virtually all securities in the noted index with weightings equal to those in the index. E = Index fund enhanced - objective is to exceed the total investment performance of a publicly recognized securities market. This is accomplished by investing primarily in derivatives based on the index itself and/or the securities within the index, or by utilizing different weightings for the securities held by the index	June 2008
vau_fund		Y = Fund is a variable annuity underlying fund N = Fund is not a variable annuity underlying fund	
et_flag	VARCHAR(1)	Identifies if a fund is an ETF or ETN: F = ETF N = ETN	
end_dt	DATE	Date of latest NAV data	
dead_flag	VARCHAR(1)	Identifies if a fund is dead or still active: Y = Yes N = No	

CRSP Mutual Funds

CRSP Style Code

< Wiesenberger Objective Codes Lipper Objective and Classification Codes >

The CRSP US Survivor-Bias-Free Mutual Funds database includes style and objective codes from three different sources over the life of the database. No single source exists for its full-time range.

- Wiesenberger Objective codes are populated between 1962 – 1993.
- Strategic Insight Objective codes are populated between 1993 – 1998.
- Lipper Objective codes begin 1998.

The CRSP Style Code builds continuity within the database by using the three afore mentioned codes as its base and provides consistency with those codes provided by our different sources.
The CRSP Style Code consists of up to four characters, with each position defined. Reading Left to Right, the four codes represent an increasing level of granularity. For example, a code for a particular mutual fund is EDYG, where:
E = Equity, D = Domestic, Y = Style, G = Growth
Codes with less than four characters exist, and it simply means that they are defined to a less granular level.

Web Query in detail (CRSP Mutual Funds)

Level 1	Level 2	Level 3	Level 4	Lipper	Strategic Insights	Wiesenberger
Equity (E)	Domestic (D)	Sector (S)	Gold (O)	AU Gold Oriented Funds	GLD Equity Gold	GPM Gold and Precious Metals
			Health (H)	H Health/Biotechnology Funds	HLT Equity USA Health	HLT Health Sector
			Financial (F)	FS Financial Services Funds	FIN Equity USA Financial Sector	FIN Financial Sector
			Natural Resources (N)	NR Natural Resources Funds	NTR Equity Natural Resources & Energy	ENR Energy/Natural Resources
			Real Estate (R)	RE Real Estate Funds	RLE Equity USA Real Estate	
			Technology (T)	TK Science & Technology Funds	TEC Equity USA Technology	TCH Technology Sector
			Utilities (U)	UT Utility Funds	UTI Equity USA Utilities	UTL Utilities
			Consumer Goods (G)	CG CONSUMER GOODS FUNDS		
			Commodities (C)	CMD COMMODITIES FUNDS		
			Consumer Services (S)	CS CONSUMER SERVICES FUNDS		
			Industrials (I)	ID INDUSTRIALS FUNDS		
			Materials (M)	BM BASIC MATERIALS FUNDS		
			Telecom (A)	TL Telecommunication Funds		
					SEC Equity USA Misc Sectors	
		Cap-based (C)	Large Cap (L)	SP S&P 500 Index Objective Funds		
			Mid Cap (M)	MC Mid-Cap Funds	GMC Equity USA Midcaps	
			Small Cap (S)	SG Small-Cap Funds	SCG Equity USA Small Companies	SCG Small Capitalization Growth
			Micro Cap (I)	MR Micro-Cap Funds		
		Style (Y)	Growth (G)	CA Capital Appreciation Funds G Growth Funds	AGG Equity USA Aggressive Growth GRO Equity USA Growth	G Growth LTG Long-Term Growth MCG Maximum Capital Gains
			Growth & Income (B)	GI Growth and Income Funds	GRI Equity USA Growth & Income ING Equity USA Income & Growth	GCI Growth and Current Income
			Hedged (H)	LSE Long/Short Equity Funds EMN Equity Market Neutral Funds ABR ABSOLUTE RETURN FUNDS DL EQUITY LEVERAGE FUNDS		
			Short (S)	DSB Dedicated Short Bias Funds		
			Income (I)	EI Equity Income Funds	OPI Option Income	IEQ Equity Income
					ENV Equity USA Environmental	

Thomson Reuters - Mutual Funds Holdings

- Holding information for mutual funds that report with the SEC
  - **Type 1: Fund characteristics** (fund name, country, …)
  - **Type 2: Stock characteristics** (cusip, stock name, shares outstanding, price)
  - **Type 3: Stock holdings** (cusip, # of shares held)
  - **Type 4: Change in holdings** (net change in shares held)
  - **s12 master file (Types 1/2/3/4)**
  - Link to the CRSP Mutual Funds database using WRDS MFLinks (Not subscribed)

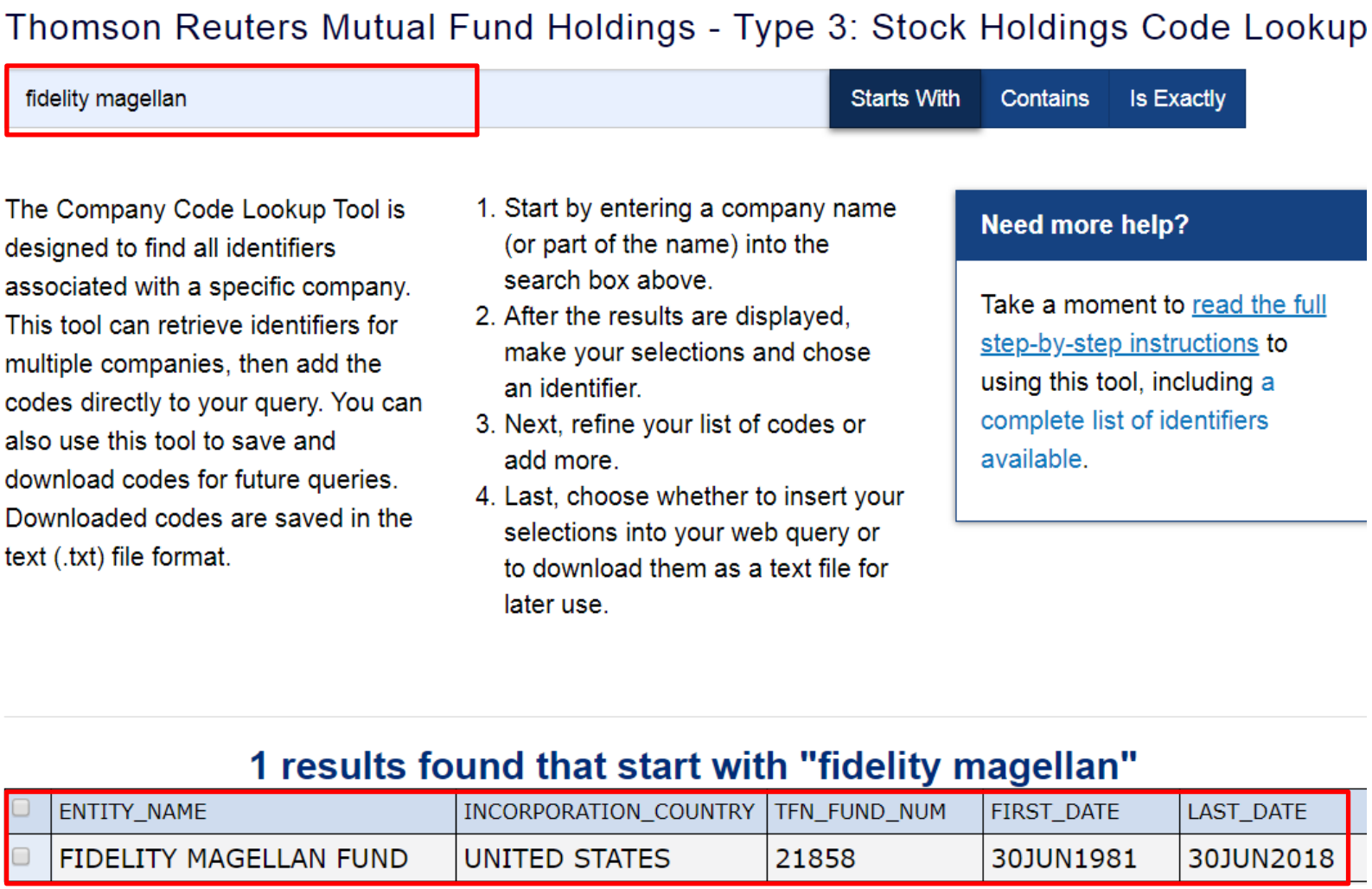
  - Details on the holdings of Fidelity Magellan and other Fidelity funds are available in the S12 sets, while the aggregate for all Fidelity funds is in the S34 set.

Web Query in detail (TR s12)

Thomson Reuters – Mutual Fund Holdings

Example: Get the security holdings for Fidelity Magellan and show the top 20 holdings for June2018.

Dataset: s12 master file



Thomson Reuters Mutual Fund Holdings - Type 3: Stock Holdings Code Lookup

fidelity magellan

Starts With | Contains | Is Exactly

The Company Code Lookup Tool is designed to find all identifiers associated with a specific company. This tool can retrieve identifiers for multiple companies, then add the codes directly to your query. You can also use this tool to save and download codes for future queries. Downloaded codes are saved in the text (.txt) file format.

1. Start by entering a company name (or part of the name) into the search box above.
2. After the results are displayed, make your selections and chose an identifier.
3. Next, refine your list of codes or add more.
4. Last, choose whether to insert your selections into your web query or to download them as a text file for later use.

Need more help?

Take a moment to read the full step-by-step instructions to using this tool, including a complete list of identifiers available.

1 results found that start with "fidelity magellan"

ENTITY_NAME	INCORPORATION_COUNTRY	TFN_FUND_NUM	FIRST_DATE	LAST_DATE
FIDELITY MAGELLAN FUND	UNITED STATES	21858	30JUN1981	30JUN2018

Web Query in detail (TR s12)

Thomson Reuters – Mutual Fund Holdings

Example: Get the security holdings for Fidelity Magellan and show the top 20 holdings for June2018.

Dataset: sec filing form N-Q https://www.sec.gov/Archives/edgar/data/61397/000137949118004145/filing924.htm

	SEC filing JUN2018				TR S12 file			
	stock	shares	value	Dif	stkname	shares	prc	holding
3	Microsoft Corp.	9,704,700	956,989	9	MICROSOFT CORP	9,704,700	98.61 $	956,980
4	Amazon.com, Inc. (a)	362,200	615,668	591	AMAZON.COM INC	362,200	1698.17 $	615,077
5	UnitedHealth Group, Inc.	1,947,584	477,820	0	UNITEDHEALTH GROUP INC	1,947,584	245.34 $	477,820
6	Alphabet Class A (a)	393,316	444,128	0	ALPHABET INC	393,316	1129.19 $	444,128
7	Alphabet Class C (a)	396,554	442,415	0	ALPHABET INC	396,554	1115.65 $	442,415
8	Apple, Inc.	2,355,513	436,029	5,182	APPLE INC	2,355,513	182.91 $	430,847
9	Berkshire Hathaway, Inc. Class B (a)	1,929,673	360,173	0	BERKSHIRE HATHAWAY INC	1,929,673	186.65 $	360,173
10	JPMorgan Chase & Co.	3,456,533	360,171	0	JPMORGAN CHASE & CO	3,456,533	104.2 $	360,171
11	Facebook, Inc. Class A (a)	1,716,200	333,492	0	FACEBOOK INC	1,716,200	194.32 $	333,492
12	Home Depot, Inc.	1,643,163	320,581	0	HOME DEPOT INC	1,643,163	195.1 $	320,581
13	Northrop Grumman Corp.	1,025,342	315,498	0	NORTHROP GRUMMAN CORP	1,025,342	307.7 $	315,498
14	Bank of America Corp.	10,378,136	292,560	0	BANK OF AMERICA CORP	10,378,136	28.19 $	292,560
15	Boston Scientific Corp. (a)	7,789,172	254,706	0	BOSTON SCIENTIFIC CORP	7,789,172	32.7 $	254,706
16	DowDuPont, Inc.	3,829,470	252,439	0	DOWDUPONT INC	3,829,470	65.92 $	252,439
17	United Technologies Corp.	2,017,321	252,226	0	UNITED TECHNOLOGIES CORP	2,017,321	125.03 $	252,226
18	ConocoPhillips Co.	3,601,083	250,707	0	CONOCOPHILLIPS	3,601,083	69.62 $	250,707
19	American Tower Corp.	1,730,050	249,421	0	AMERICAN TOWER CORP	1,730,050	144.17 $	249,421
20	Monster Beverage Corp. (a)	4,310,100	246,969	0	MONSTER BEVERAGE CORP	4,310,100	57.3 $	246,969
21	Intuit, Inc.	1,188,200	242,755	-30	INTUIT INC	1,188,200	204.33 $	242,785
22	EOG Resources, Inc.	1,946,292	242,177	0	EOG RESOURCES INC	1,946,292	124.43 $	242,177

Lipper for Investment Management **Not subscribed**

- Provides mutual fund data (returns, TNA, benchmark, fund family, total expense ratio (TER), loads, fund domicile, countries notified for sale, etc.)
- Coverage of over 213,000 share classes of more than 117,000 funds located in over 60 countries
- Ferreira, Keswani, Miguel and Ramos (2012) find that Lipper covers 87% of the total TNA of worldwide equity mutual funds (as of Dec/2007)

Lipper for Investment Management

	Lipper_ID	Name	CUSIP	ISIN_Code	Asset_Universe	Asset_Type	Asset_Status	Geographical_Focus	Lipper_Global
1	40000006	Lipper Growth Fund Index			Mutual Funds	Equity	Liquidated	United States of America	
2	40000013	Reynolds Opportunity Fund	761724202	US7617242026	Mutual Funds	Equity	Liquidated	United States of America	Equity North America
3	40000020	PNC Large Cap Growth;A	69351J249	US69351J2490	Mutual Funds	Equity	Active	United States of America	Equity North America
4	40000027	T Rowe Price Japan Fund	77956H708	US77956H7089	Mutual Funds	Equity	Active	Japan	Equity Japan
5	40000028	Dreyfus International Growth Fund;A	261986103	US2619861030	Mutual Funds	Equity	Liquidated	Global Ex US	Equity Global ex US
6	40000029	Touchstone HLAM Large Cap Quality Stock Fund;II	89155H702	US89155H7026	Mutual Funds	Equity	Liquidated	United States of America	Equity North America
7	40000031	Scudder Dynamic Growth Fund (Class A)	81114R103	US81114R1032	Mutual Funds	Equity	Merged	United States of America	Equity Nth America Sm&Mid Cap
8	40000033	RiverSource New Dimensions Fund;A	00245V106	US00245V1061	Mutual Funds	Equity	Merged	United States of America	Equity North America
9	40000035	Lord Abbett Affiliated Fund;A	544001100	US5440011006	Mutual Funds	Equity	Active	United States of America	Equity North America
10	40000038	Seligman Capital Fund;A	816326102	US8163261022	Mutual Funds	Equity	Merged	United States of America	Equity Nth America Sm&Mid Cap
11	40000039	American Funds AMCAP Fund;A	023375108	US0233751082	Mutual Funds	Equity	Active	United States of America	Equity North America
12	40000043	Midas Investors Ltd.	595632100	US5956321005	Mutual Funds	Equity	Merged	United States of America	Equity Sector Gold&Prec Metals
13	40000044	American Growth Fund;D	026393108	US0263931083	Mutual Funds	Equity	Active	United States of America	Equity North America
14	40000045	Invesco Charter Fund;A	001413103	US0014131033	Mutual Funds	Equity	Active	United States of America	Equity North America
15	40000046	Smith Barney Equity Funds, Inc.; Class A Shares			Mutual Funds	Equity	Liquidated	United States of America	

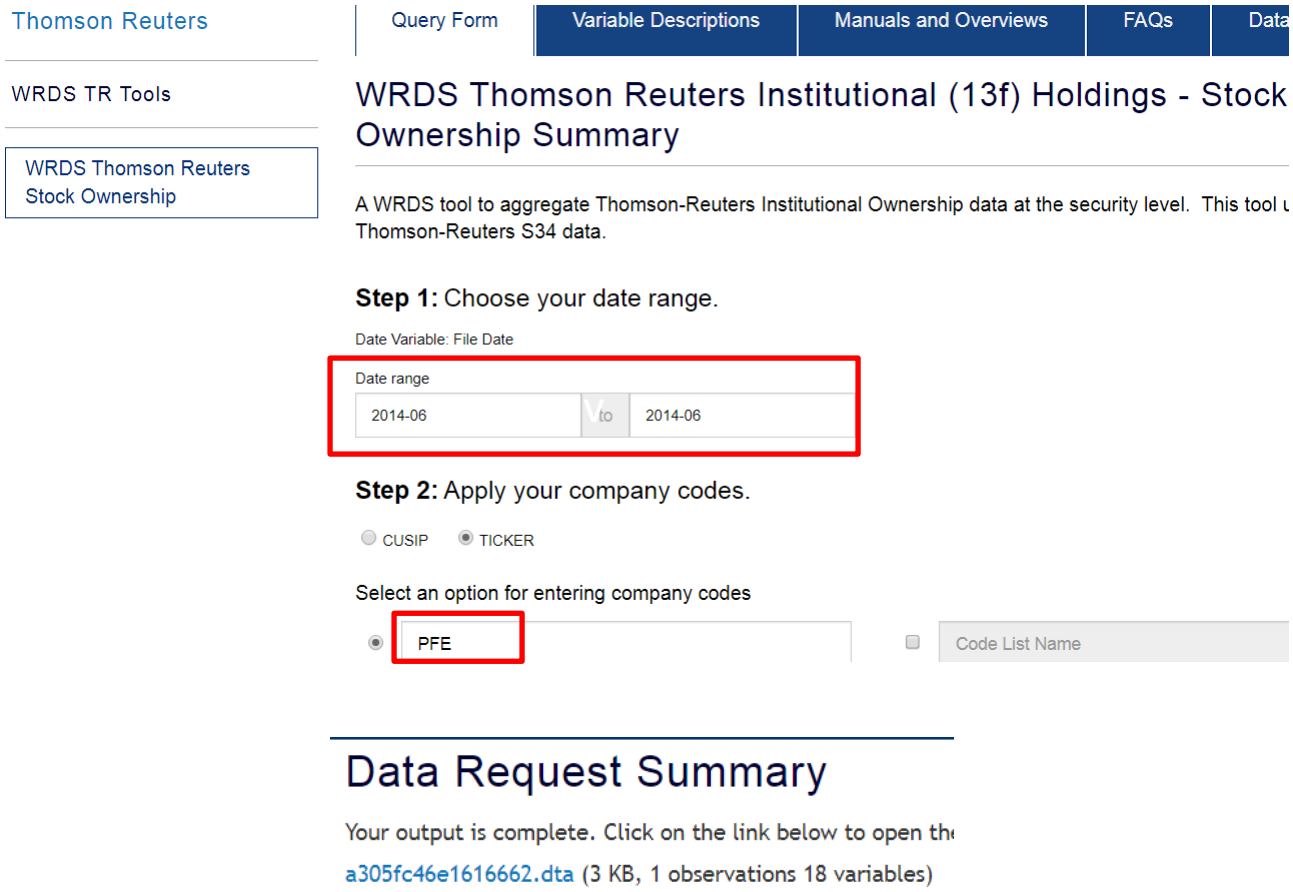
	Fund_Manager_Benchmark	Technical_Indicator_Benchmark_Co	Technical_Indicator_Benchmark	Fund_Management_Company_Code	Fund_Management_Company_Name	Fund_Management_Company_Web_Site	Fund_Ma
1		.					
2	S&P 500 TR	11020361	Russell 2000 Growth TR	REY$	Reynolds Capital Management	www.reynoldsfunds.com	
3	Russell 1000 Growth TR	11000689	Russell 1000 Growth TR	NCC$	PNC Capital Advisors LLC	www.allegiantfunds.com	
4	MSCI Japan NR USD	11004832	MSCI Japan NR USD	PRI$	T Rowe Price Associates Inc	www.troweprice.com	
5	MSCI World NR USD	11004203	MSCI EU Growth TR USD	DRY$	Dreyfus Corporation	www.dreyfus.com	
6	Russell 1000 Growth TR	11000696	S&P 500 TR	TOU$	Touchstone Advisors Inc	www.touchstoneinvestments.com	
7		11021725	Russell MidCap Growth TR	SCD$	DWS Investments	www.dws-investments.com	
8		11021728	Russell 3000 TR	RPUH	Ameriprise Financial Inc	www.ameriprise.com	
9	Russell 1000 Value TR	11028480	S&P 500 Value Total Return Index	LRD$	Lord Abbett & Co LLC	www.lordabbett.com	
10	Russell MidCap Growth TR	11021725	Russell MidCap Growth TR	AXP$	Columbia Management Inv Advisers LLC	www.columbiafunds.com	
11	S&P 500 TR	11000689	Russell 1000 Growth TR	CRE$	Capital Research & Management Company		
12		.		BNB$	Midas Management Corporation	www.midasfunds.com	
13	S&P 500 TR	11021728	Russell 3000 TR	IVM$	Investment Research Corporation	www.americangrowthfund.com	
14	S&P 500 TR	11021727	Russell 1000 TR	AMA$	Invesco Advisers Inc	www.invesco.com	

Thomson Reuters Institutional Holdings

- Formerly known as CDA/Spectrum
- Section 13(f) of the Securities Exchange Act (1975)
- Institutional Common Stock Holdings as reported on Form 13F filed with the SEC
- Ownership information by institutional managers with $100 million or more in Assets Under Management

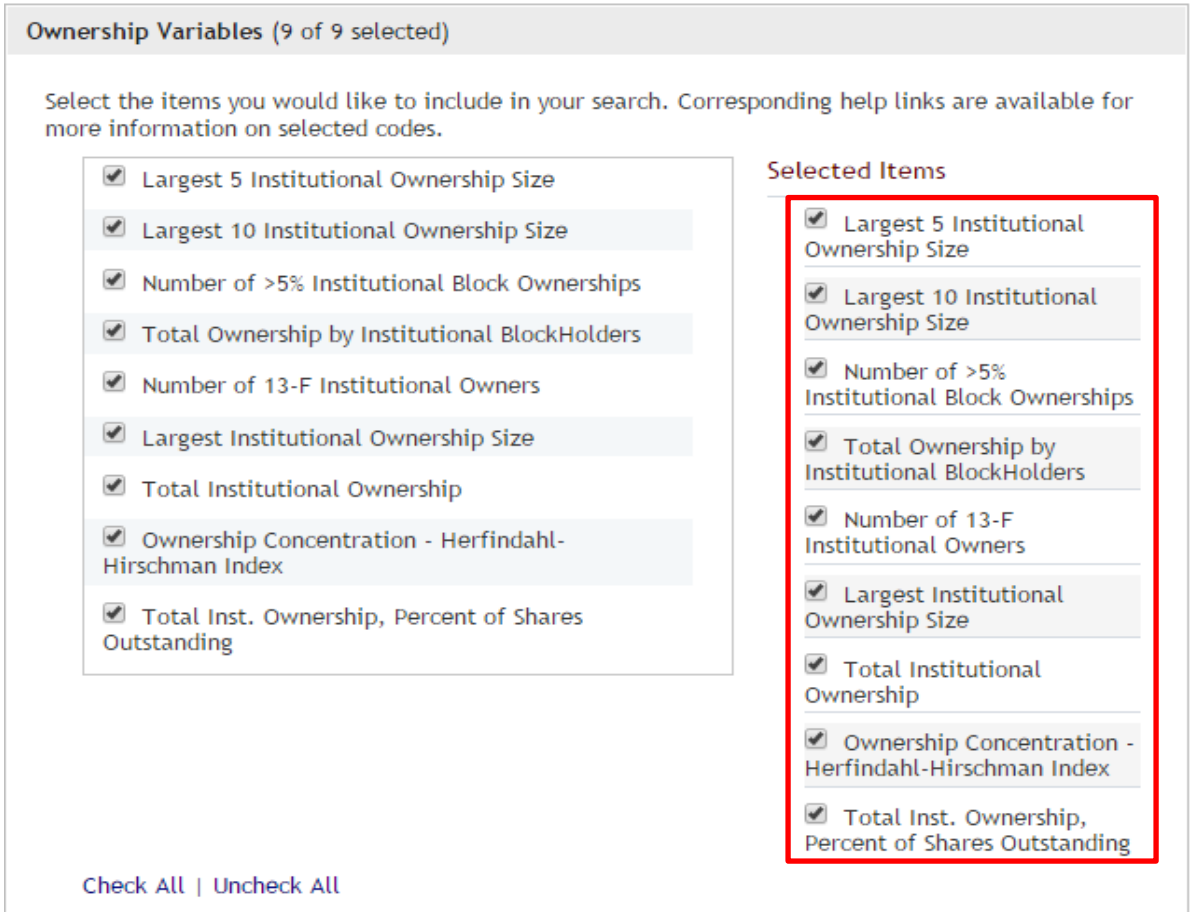
Thomson Reuters 13F Ownership Tool

➢ **Ownership of Pfizer in 2014-06-30**



Thomson Reuters 13F Ownership Tool

➢ **Ownership of Pfizer in 2014-06-30**



Thomson Reuters 13F Ownership Tool

- **Ownership of Pfizer in 2014-06-30**

	rdate	stkname	ticker	instown	instown_hhi	instown_perc
1	30jun2014	PFIZER INC	PFE	4475315092	.024950791	.6309282948

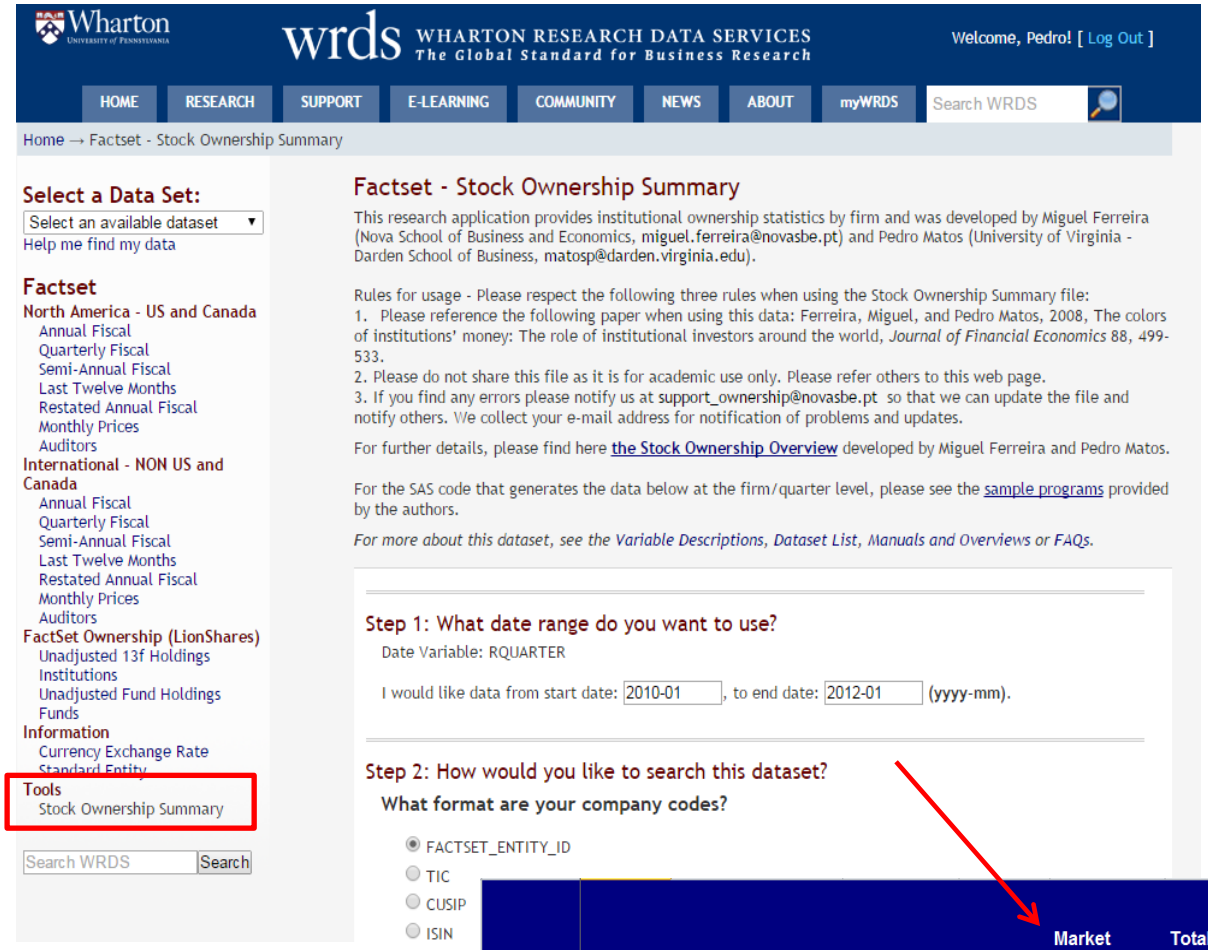
You can access more detailed data (e.g. by institution or fund) using the SAS code used in the Tool provided by WRDS (Moussawi and Palacios, 2009) or by using the Thomson Reuters raw files (Web Query)

http://wrds-web.wharton.upenn.edu/wrds/research/applications/index.cfm

FactSet/LionShares Fundamentals and Ownership Not Subscribed

- Contains global equity ownership data for approximately 13,000 institutions and 33,000 mutual funds
- History dating back to 1999
- Unadjusted Package (survivorship bias free)
- Fundamentals Data (same as Worldscope)
- Stock Ownership Summary Tool (Ferreira and Matos, JFE 2008) programmed by Pedro Pires

Web Query in detail (FactSet/LionShares)



Firms Factset Entity ID	Quarter (Report Date)	Firms name	ISIN (primary)	TICKER (primary	Market capitalization in US$100000	Total institutional ownership in US$1000000	Number of institutional owner	Total institutional ownership ratio in percentage of market capitalization	Domestic institutional ownership ratio in percentage of market	Foreign institutional ownership ratio in percentage of market
05HH9S-E	201001	Energias de Portugal SA	PTEDP0AM0009	EDP	14427.14	1850.49	367	12.8%	0.3%	12.5%
05HH9S-E	201002	Energias de Portugal SA	PTEDP0AM0009	EDP	10828.32	1320.96	354	12.2%	0.2%	12.0%
05HH9S-E	201003	Energias de Portugal SA	PTEDP0AM0009	EDP	12434.37	1572.67	341	12.6%	0.2%	12.4%
05HH9S-E	201004	Energias de Portugal SA	PTEDP0AM0009	EDP	12108.05	1783.98	330	14.7%	0.2%	14.5%
05HH9S-E	201101	Energias de Portugal SA	PTEDP0AM0009	EDP	13635.41	2079.03	317	15.2%	0.2%	15.1%
05HH9S-E	201102	Energias de Portugal SA	PTEDP0AM0009	EDP	12870.02	1894.01	320	14.7%	0.2%	14.6%
05HH9S-E	201103	Energias de Portugal SA	PTEDP0AM0009	EDP	11257.14	1604.33	304	14.3%	0.2%	14.0%
05HH9S-E	201104	Energias de Portugal SA	PTEDP0AM0009	EDP	11249.01	1647.42	307	14.6%	0.2%	14.4%
05HH9S-E	201201	Energias de Portugal SA	PTEDP0AM0009	EDP	10524.96	1536.23	305	14.6%	0.1%	14.5%

Contents coverage

- **Syndicated loans**
  - **Thomson Reuters LPC DealScan**

✓ Thomson Reuters LPC DealScan

- Also known as *Loan Pricing Corporation Deal Scan*
- Global syndicated bank loan market
- Detailed historical (origination) information on loan pricing, contracts details, terms and conditions on over 240,000 transactions since 1988
- Information includes borrower, lender, purpose, amount, fees, and covenants.
- Carey and Hrycray (1999) estimated that DealScan loans covered between half and three-quarters of the volume for outstanding commercial and industrial loans in the US

For additional insights about DealScan watch e-learning video from WRDS:

→ http://wrds-web.wharton.upenn.edu/wrds/E-Learning/_000Video/Overview_of_Dealscan/index.cfm

Thomson Reuters LPC DealScan

- **Company**
  - Company information on all borrowers & lenders
  - Main identifier: companyID
  - No historical information
  - Contains the ultimate parent of each company
- **Package**
  - Details on each loan contract or deal (named package)
  - Uniquely identified by packageID
  - One package may include several facilities (groups facilities into each deal)
  - Information on deal status: (completed 95%, noinfo 1.5%, in process, closed, cancelled, suspended, rumour, etc.)
- **Facility**
  - Details the information provided in the package file
  - Uniquely identified by facilityID
  - Contains packageID, BorrowerCompanyID

Thomson Reuters LPC DealScan

- **Lenders**
  - Company information on the lenders & participants in each facilityID
  - Contains the LenderRole
- **Current Facility Pricing**
  - Details on loan pricing: fees, base rates and spreads
- **DealScan-Compustat linking table** (mostly for US firms)
  - Prof. Michael Roberts
  - gvkey

Contents coverage

- **Others:**
  - **Eventus**
  - **Datastream**
  - **WorldBank WDI**
  - **IMF Macroeconomic and financial data**

Web Query in detail (Eventus)

✓ Eventus

➢ **Performs event studies using CRSP market data**

Eventus

Basic Event Study

- Daily
- Monthly
- Daily - with Fama French
- Monthly - with Fama French

Query Form | Variable Descriptions | Manuals and Overviews

Eventus Basic Event Study - Daily

Basic Event Study (Daily) executes an event study with daily data centered on a si
most common and simple event study.

Directions for running an Eventus query can be found in the Overview of Eventus.
on WRDS requires a request file (i.e. input file with event dates and firm identifiers)
Sample Request Files for more information.

You can use the Eventus File Validator to test your input text file to make sure they

Step 1: Choose identifier and request file.

The request file must have a specific format to be read correctly by Eventus: a simple text file \
and the date. MS Word or MS Excel files will not work.

Select identifier used in Request File

- PERMNO
- CUSIP

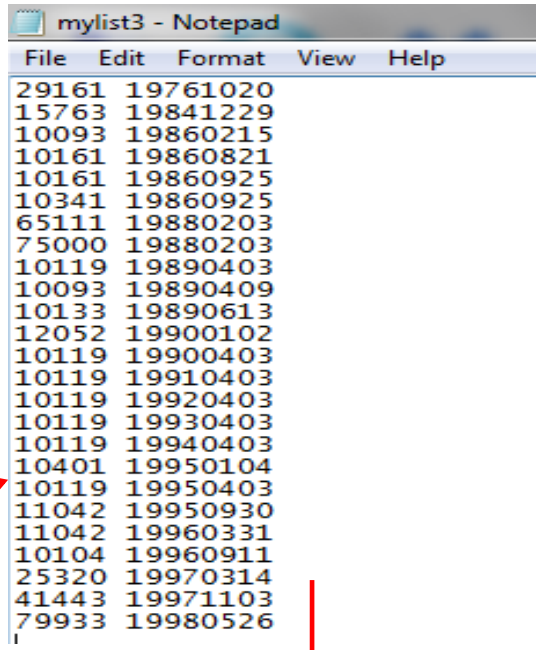
Request File

Escolher ficheiro | mylist3.txt

[Code Search]

Group data file options:

- none



Text file with 2 columns:

- **Col (1) PERMNO or CUSIP**
- **Col (2) Date**

Web Query in detail (Eventus)

Step 2: Choose Market Indices and Benchmark Options.

Market Index

- ○ CRSP Equally Weighted
- ◉ CRSP Value Weighted
- ○ CRSP Equally Weighted + Value Weighted
- ○ CRSP Equally Weighted + SP500
- ☐ Exclude Dividends (NODIVIDX)

Additional Benchmark Options

(MM - market model is selected by default)

- ☐ No Market Model (NOMM)
- ☑ Market-Adjusted Returns (MAR)
- ☐ No Market Adjustment (RAW)
- ☐ Comparison Period Mean-Adjusted Returns (CP)

Market Model:

- **Using CRSP Value Weighted**
- **Parameter Estimation (α , β) using pre-defined window (-246,-46)**
- **Cumulative abnormal return window (-30,+30)**

Alternative Benchmark: Market-Adjusted Returns (MAR, β=1)

Step 3: Choose Estimation Options.

Estimation Period:

End Before Event Date (EST)	-46	days
Minimum Estimation Length (MINESTN)	15	days
Maximum Estimation Length (ESTLEN)	200	days

☐ POOL

Autodate

- ◉ None

Estimate Method

- ◉ OLS
- ○ EGARCH
- ○ GARCH
- ○ ScholesWilliams

Step 4: Choose windows to search.

Event Period:

PRE 30 POST 30

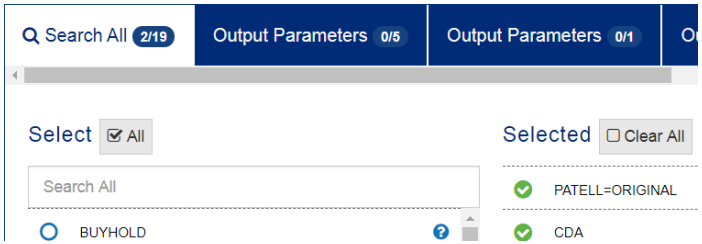
☐ Allow OVERLAP with estimation period

Alternative Windows

	Begin		End	
1.	-30	days	-2	days
2.	-1	days	1	days
3.	1	days	10	days
4.	2	days	30	days
5.		days		days
6.		days		days

Step 5: Choose tests to search.

How does this work?



Web Query in detail (Eventus)

Market Adjusted Returns, Value Weighted Index

Days	N	Mean Cumulative Abnormal Return	Precision Weighted CAAR	Positive: Negative	Uncorrected Patell Z	Portfolio Time-Series (CDA) t	Generalized Sign Z
(-30,-2)	18	-1.09%	-1.17%	6:12	-0.488	-0.169	-1.133
(-1,+1)	18	2.00%	1.13%	10:8	1.437$	0.966	0.757
(+1,+10)	18	5.93%	2.98%	10:8	2.100*	1.567$	0.757
(+2,+30)	18	5.18%	2.33%	11:7	0.954	0.804	1.229

The symbols $,*,**, and *** denote statistical significance at the 0.10, 0.05, 0.01 and 0.001 levels, respectively, using a generic one-tail test. The symbols (,< or),> etc. correspond to $,* and show the direction and significance of a generic one-tail generalized sign test.

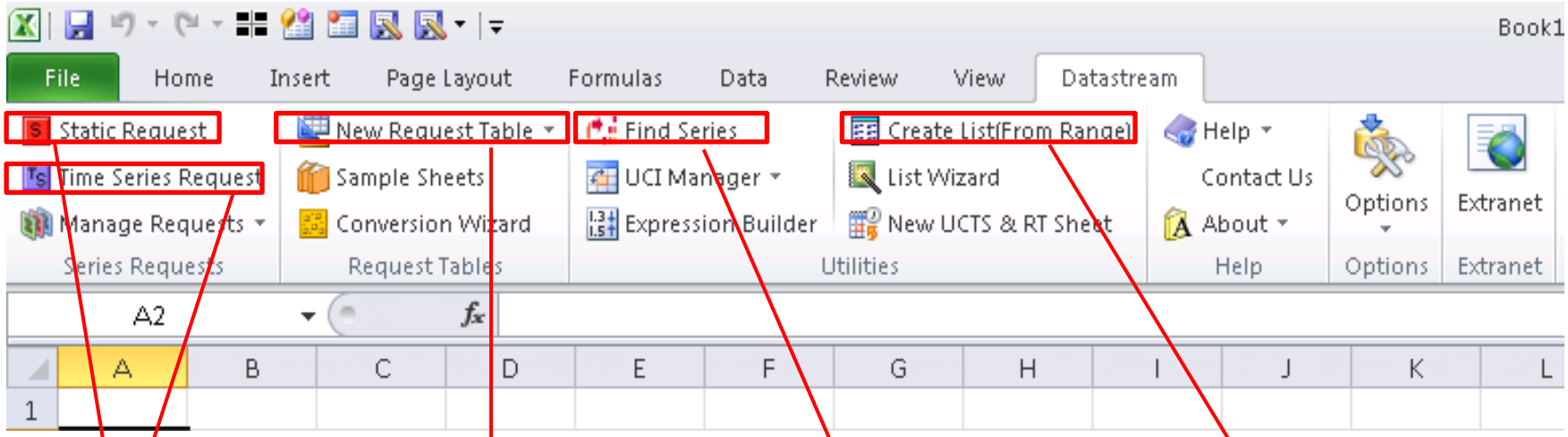
Market Model Abnormal Returns, Value Weighted Index

Days	N	Mean Cumulative Abnormal Return	Precision Weighted CAAR	Positive: Negative	Uncorrected Patell Z	Portfolio Time-Series (CDA) t	Generalized Sign Z
(-30,-2)	18	-2.26%	-1.64%	6:12	-0.677	-0.352	-1.052
(-1,+1)	18	1.61%	1.04%	9:9	1.322$	0.778	0.368
(+1,+10)	18	5.39%	2.69%	10:8	1.889*	1.427$	0.841
(+2,+30)	18	4.65%	2.57%	10:8	1.046	0.723	0.841

The symbols $,*,**, and *** denote statistical significance at the 0.10, 0.05, 0.01 and 0.001 levels, respectively, using a generic one-tail test. The symbols (,< or),> etc. correspond to $,* and show the direction and significance of a generic one-tail generalized sign test.

Outside WRDS… (Eikon)

✓ Thomson Reuters Eikon - DataStream/WorldScope



Tools to request Static Data and Time-Series Data

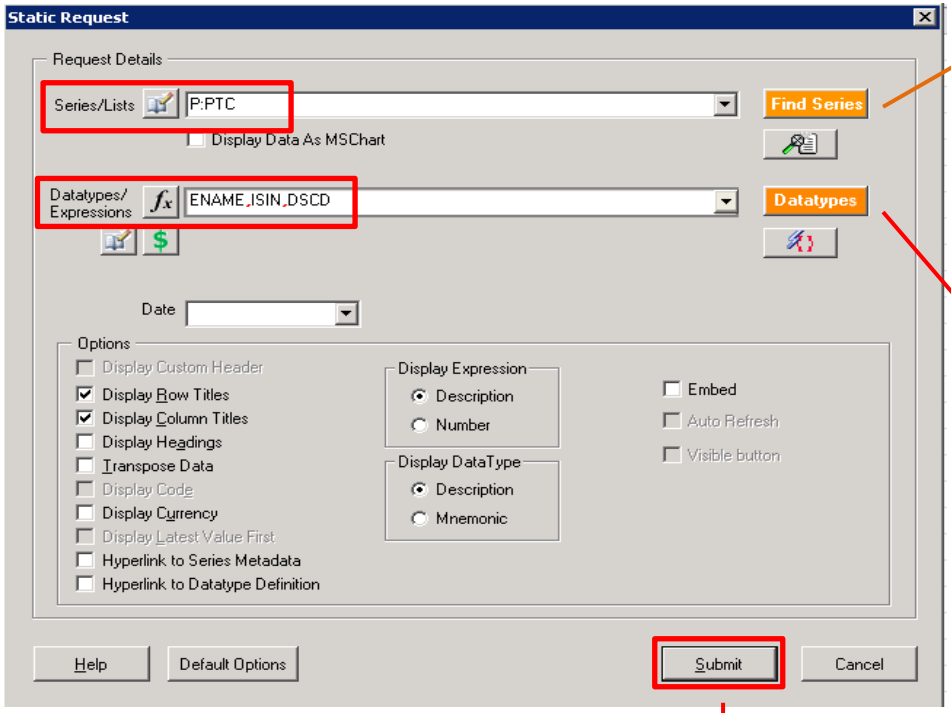
Opens excel lay-out useful to download huge datasets (several lists and data fields)

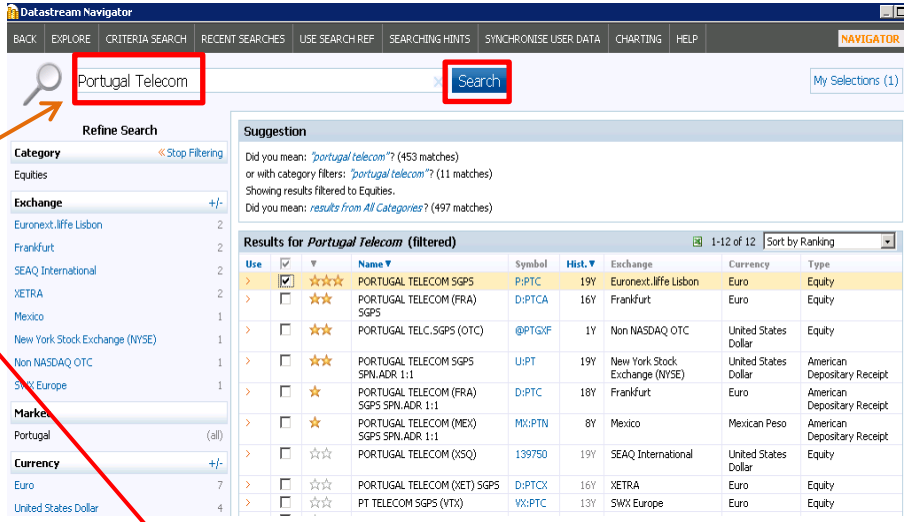
Datastream Navigator to explore data

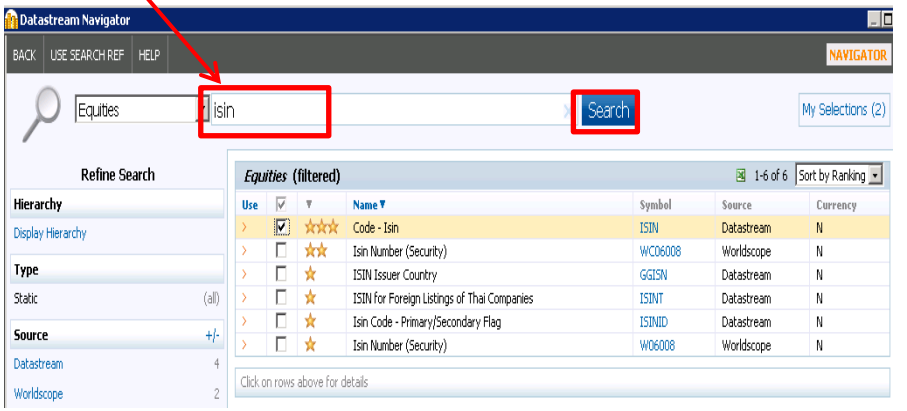
Tool to create a single list of securities using identifiers (preferred key: datastream codes)

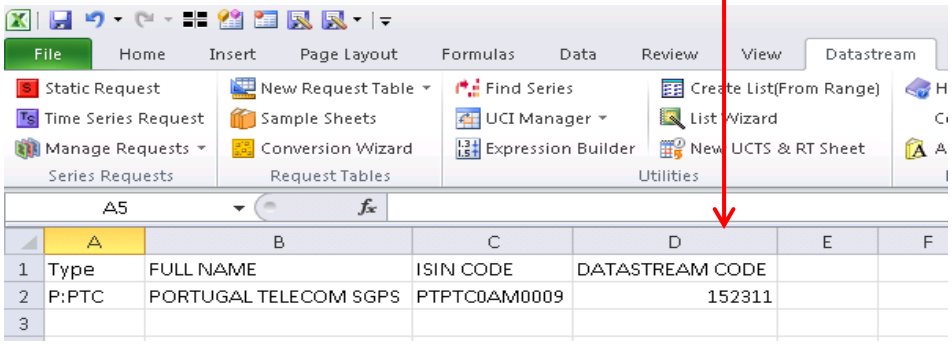
Outside WRDS… (Eikon)

Static Request



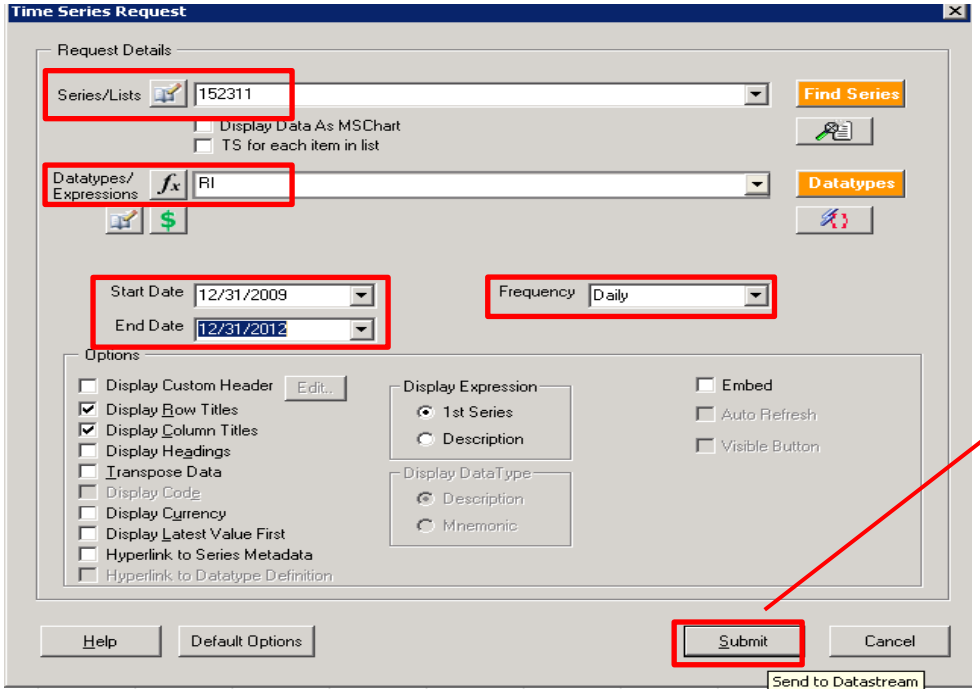


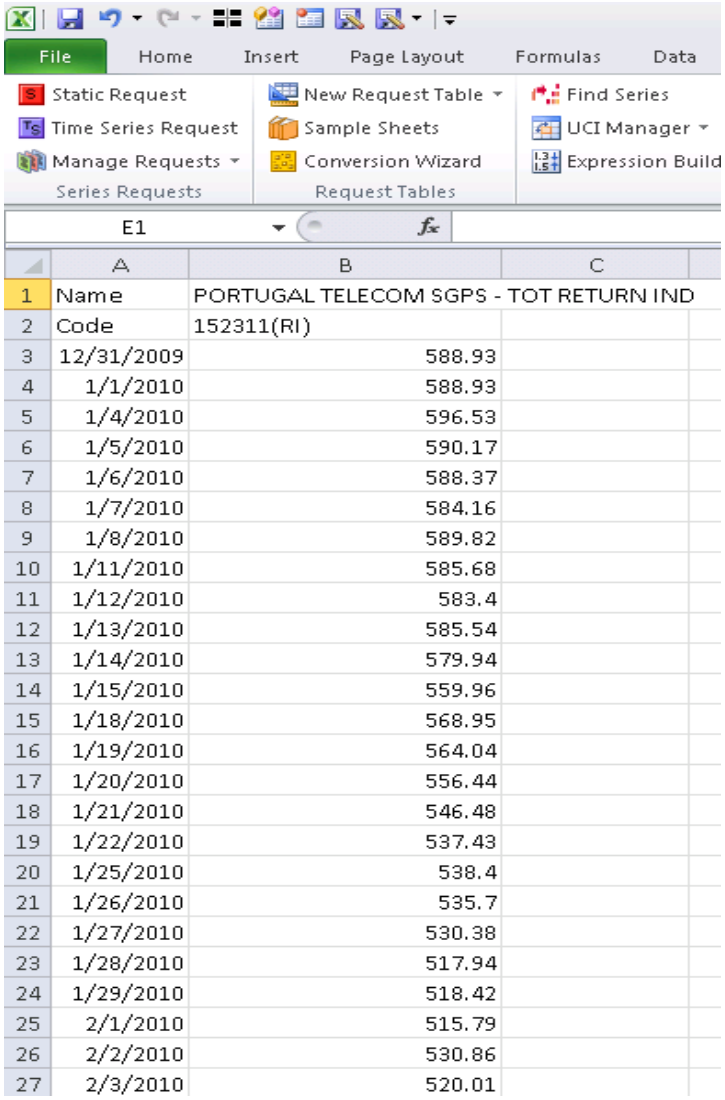




Outside WRDS… (Eikon)

Time-Series Request

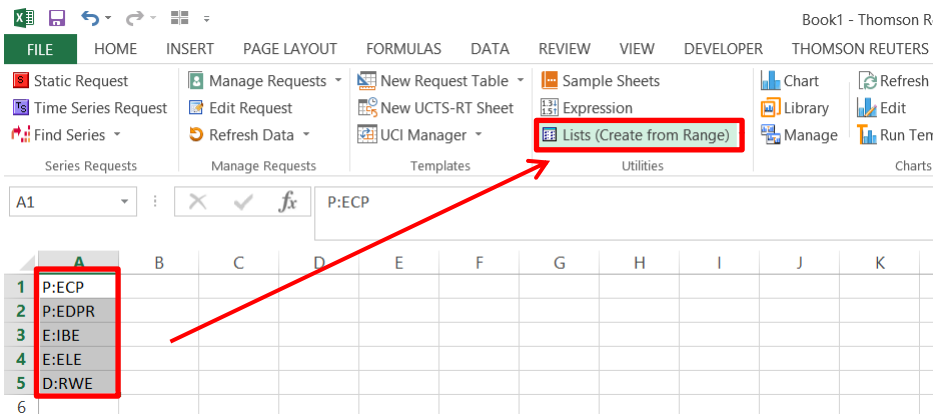


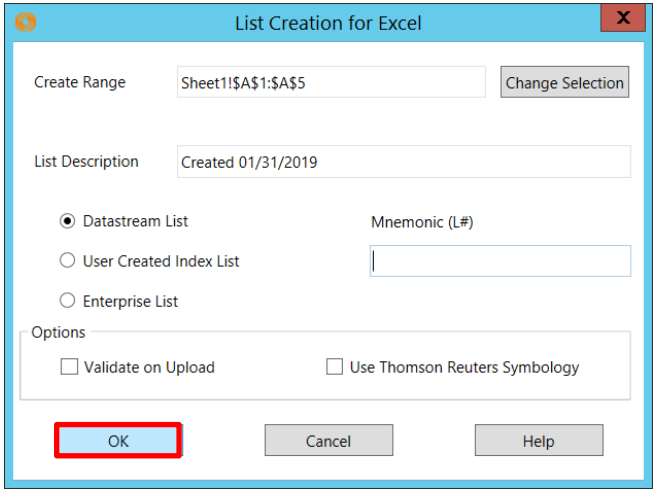


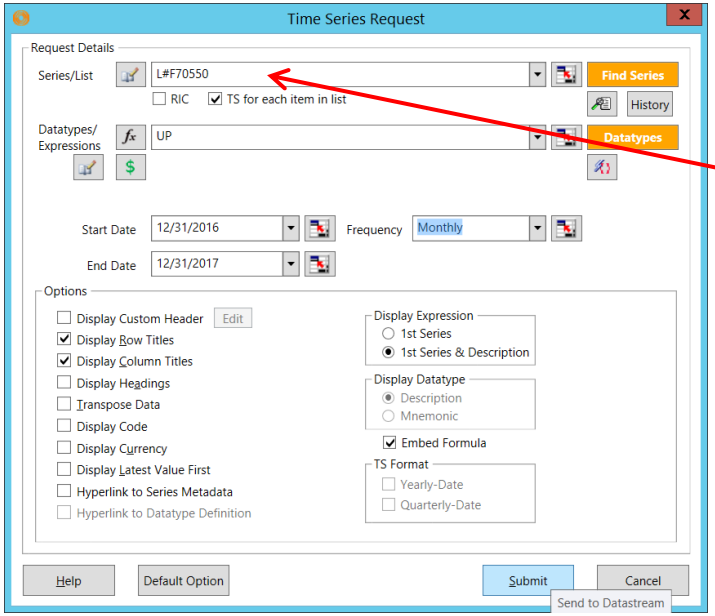
	A	B	C
1	Name	PORTUGAL TELECOM SGPS - TOT RETURN IND	
2	Code	152311(RI)	
3	12/31/2009	588.93	
4	1/1/2010	588.93	
5	1/4/2010	596.53	
6	1/5/2010	590.17	
7	1/6/2010	588.37	
8	1/7/2010	584.16	
9	1/8/2010	589.82	
10	1/11/2010	585.68	
11	1/12/2010	583.4	
12	1/13/2010	585.54	
13	1/14/2010	579.94	
14	1/15/2010	559.96	
15	1/18/2010	568.95	
16	1/19/2010	564.04	
17	1/20/2010	556.44	
18	1/21/2010	546.48	
19	1/22/2010	537.43	
20	1/25/2010	538.4	
21	1/26/2010	535.7	
22	1/27/2010	530.38	
23	1/28/2010	517.94	
24	1/29/2010	518.42	
25	2/1/2010	515.79	
26	2/2/2010	530.86	
27	2/3/2010	520.01	

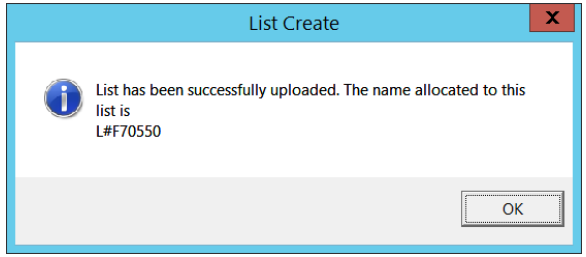
Outside WRDS… (Eikon)

Create List of Securities



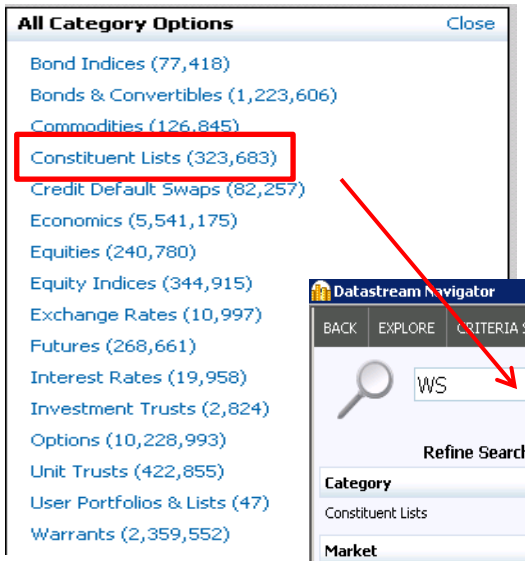




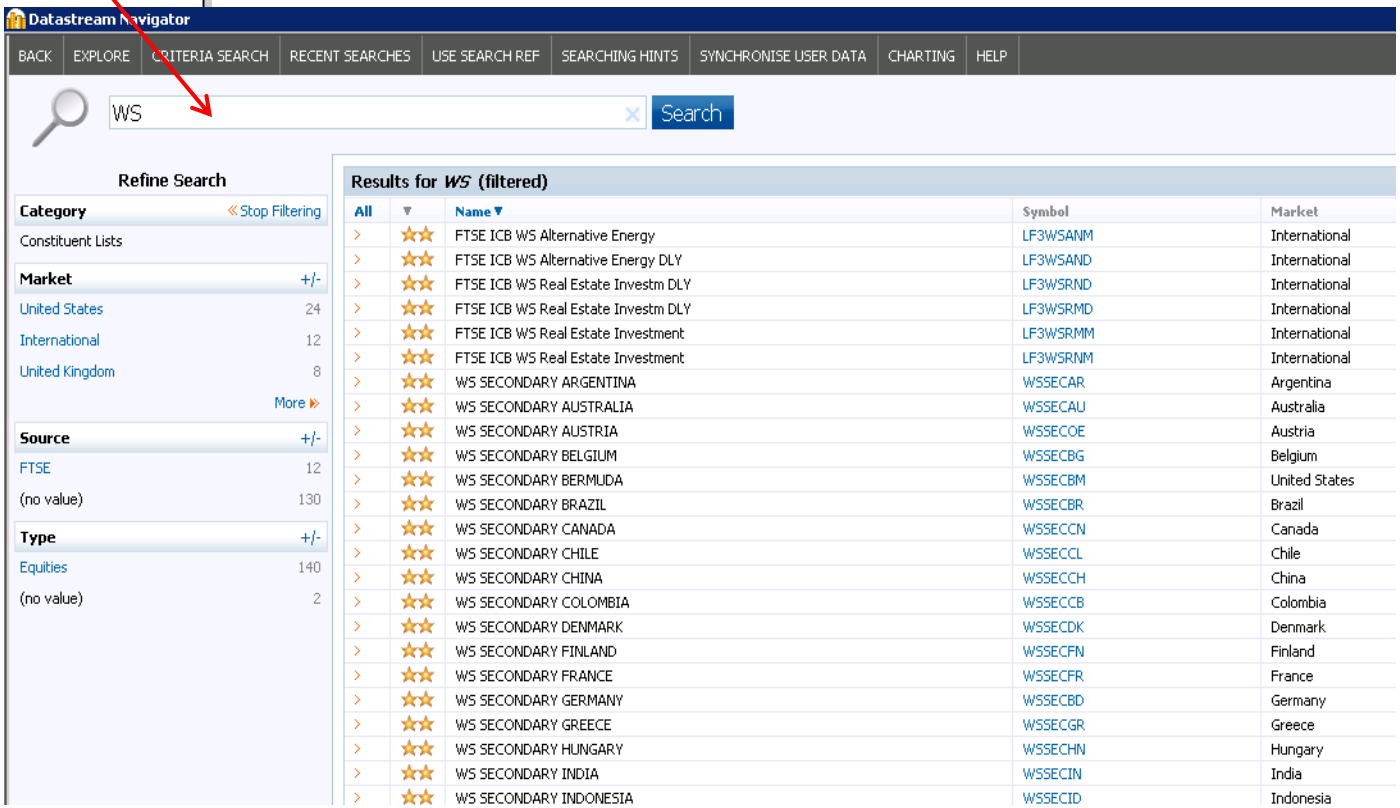


Outside WRDS… (Eikon)

Constituent Lists of Securities

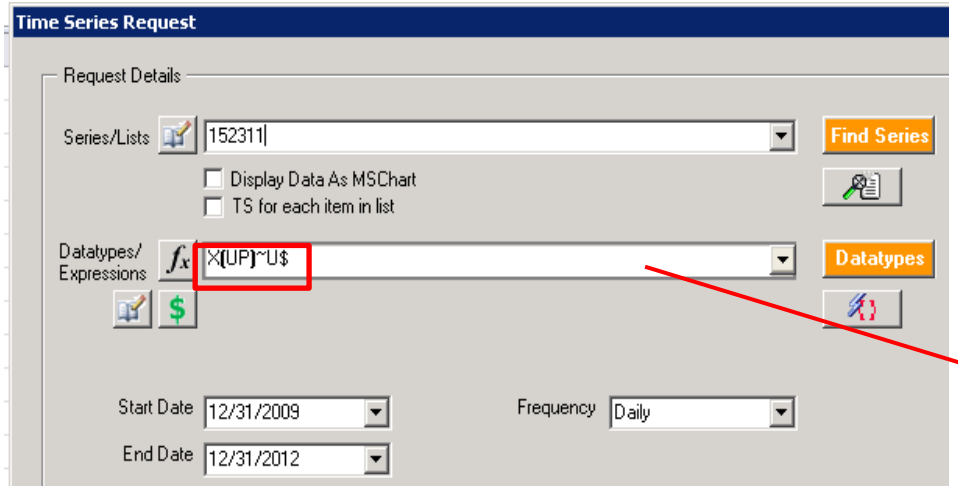


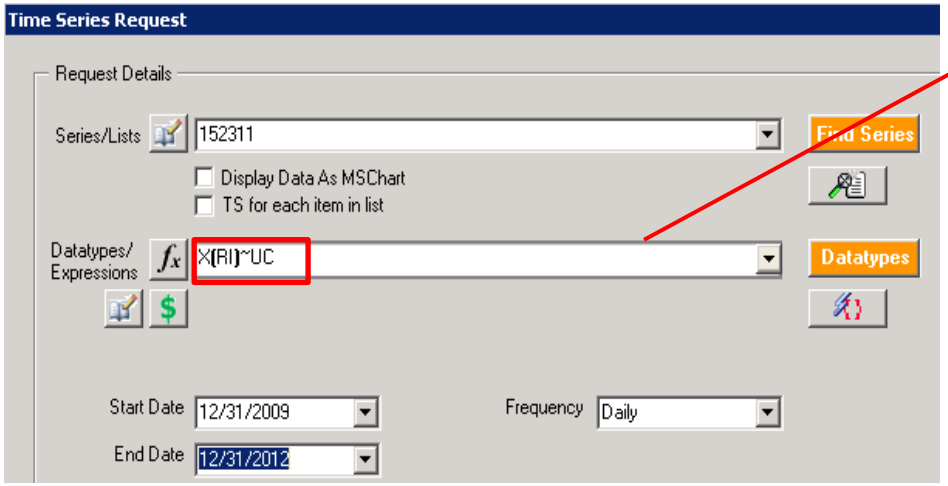
Market Indices and Research Lists

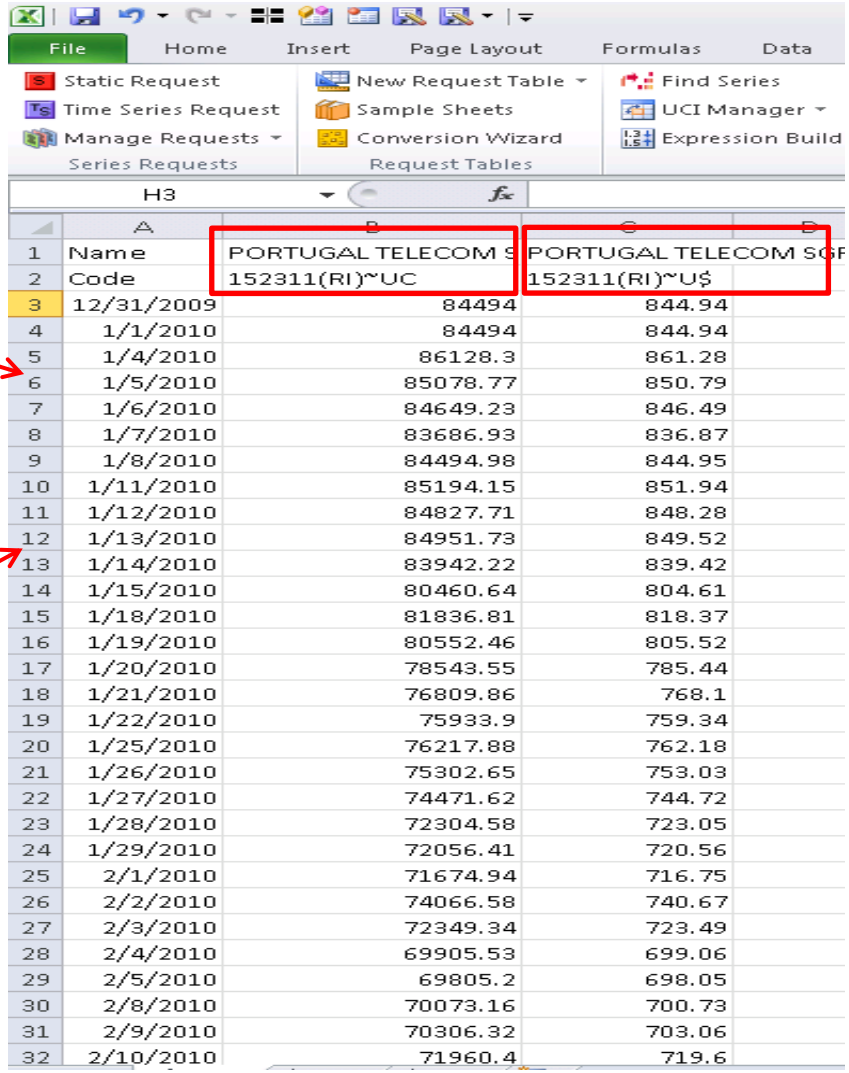


Outside WRDS… (Eikon)

Converting data to US$ or US$ Cents.



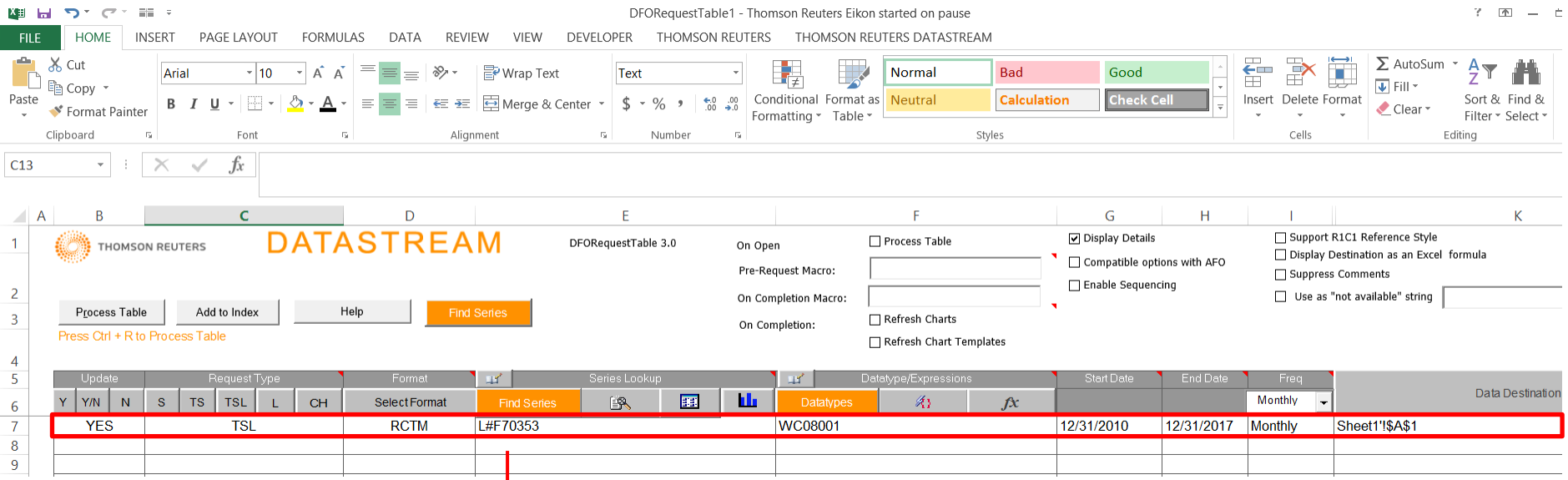


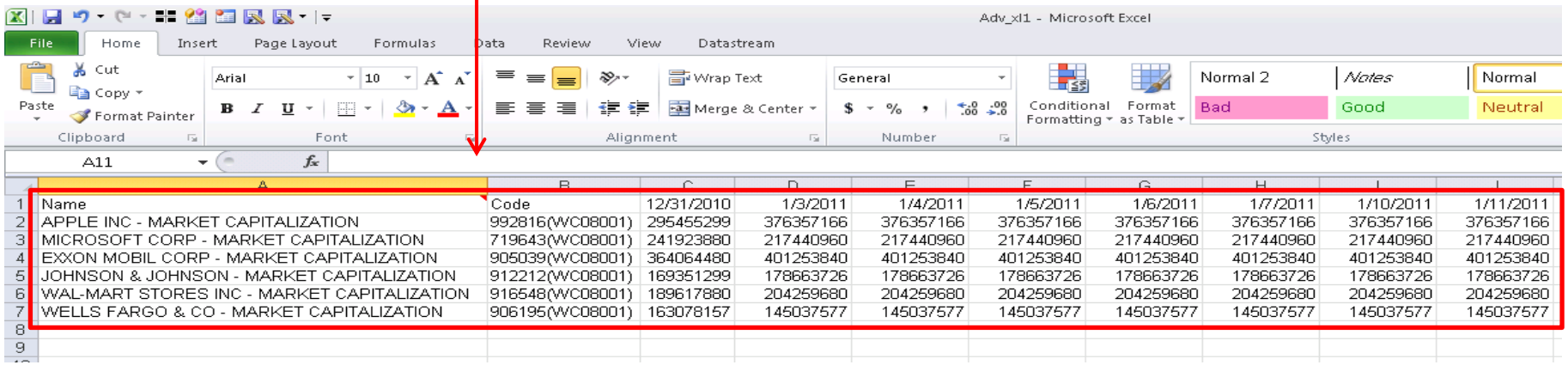


	A	B	C	D
1	Name	PORTUGAL TELECOM S	PORTUGAL TELECOM SGF	
2	Code	152311(RI)~UC	152311(RI)~U$	
3	12/31/2009	84494	844.94	
4	1/1/2010	84494	844.94	
5	1/4/2010	86128.3	861.28	
6	1/5/2010	85078.77	850.79	
7	1/6/2010	84649.23	846.49	
8	1/7/2010	83686.93	836.87	
9	1/8/2010	84494.98	844.95	
10	1/11/2010	85194.15	851.94	
11	1/12/2010	84827.71	848.28	
12	1/13/2010	84951.73	849.52	
13	1/14/2010	83942.22	839.42	
14	1/15/2010	80460.64	804.61	
15	1/18/2010	81836.81	818.37	
16	1/19/2010	80552.46	805.52	
17	1/20/2010	78543.55	785.44	
18	1/21/2010	76809.86	768.1	
19	1/22/2010	75933.9	759.34	
20	1/25/2010	76217.88	762.18	
21	1/26/2010	75302.65	753.03	
22	1/27/2010	74471.62	744.72	
23	1/28/2010	72304.58	723.05	
24	1/29/2010	72056.41	720.56	
25	2/1/2010	71674.94	716.75	
26	2/2/2010	74066.58	740.67	
27	2/3/2010	72349.34	723.49	
28	2/4/2010	69905.53	699.06	
29	2/5/2010	69805.2	698.05	
30	2/8/2010	70073.16	700.73	
31	2/9/2010	70306.32	703.06	
32	2/10/2010	71960.4	719.6	

Outside WRDS… (Eikon)

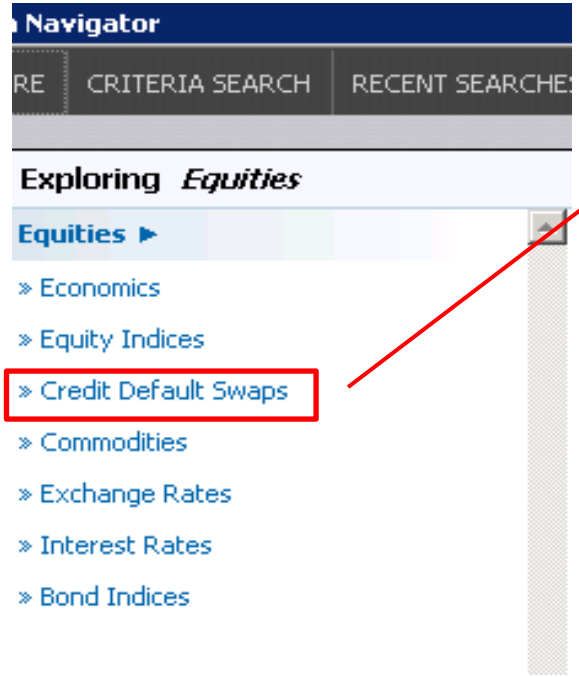
Request Table Tool

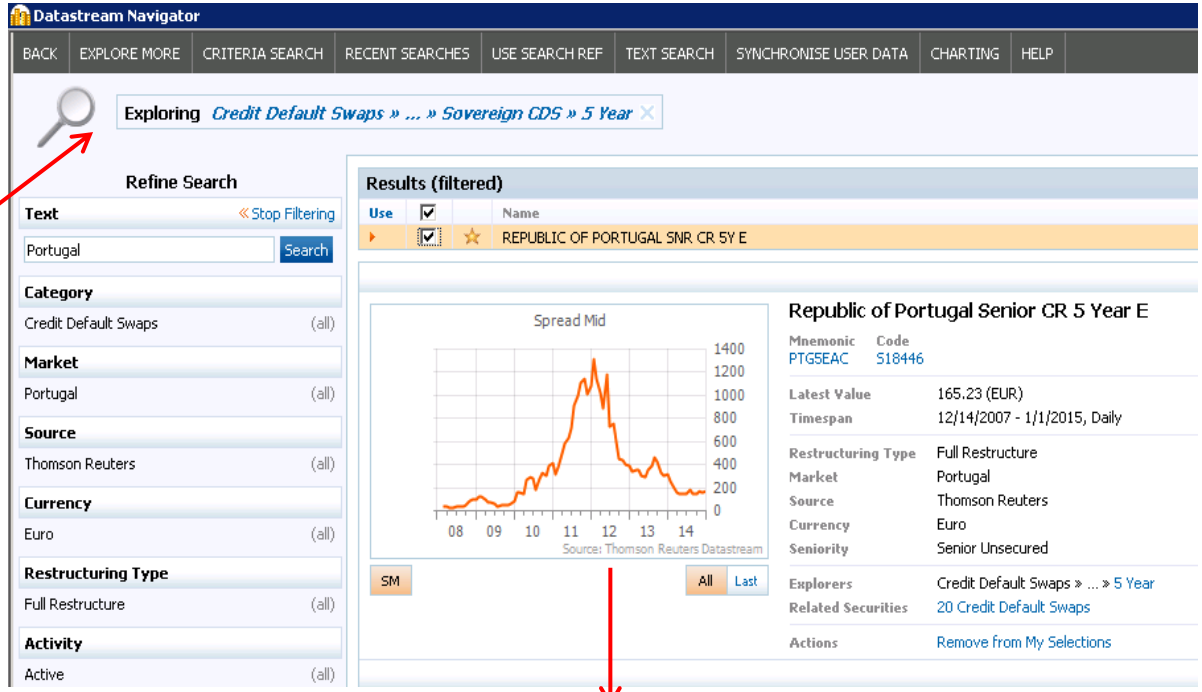




Outside WRDS… (Eikon)

Other Data Categories…





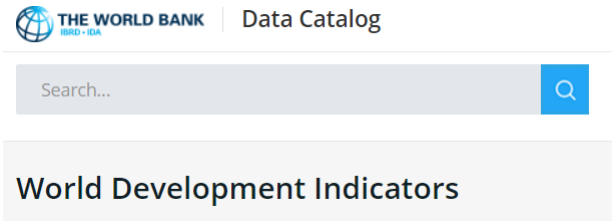
1	Name	REPUBLIC OF POR
2	Code	PTG5EAC(SM)
3	CURRENCY	E
4	3/31/2008	40.5
5	4/30/2008	25.5
6	5/30/2008	25
7	6/30/2008	36.5
8	7/31/2008	36
9	8/29/2008	37
10	9/30/2008	50
11	10/31/2008	83.5
12	11/28/2008	92
13	12/31/2008	92

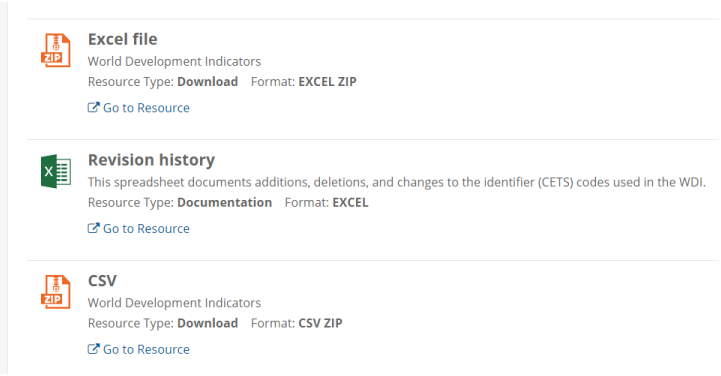
Example: Republic of Portugal 5 Yrs. CDS Spread

Outside WRDS… (WDI)

WDI

- **World Bank collection of development indicators**
  - Download excel file https://datacatalog.worldbank.org/dataset/world-development-indicators

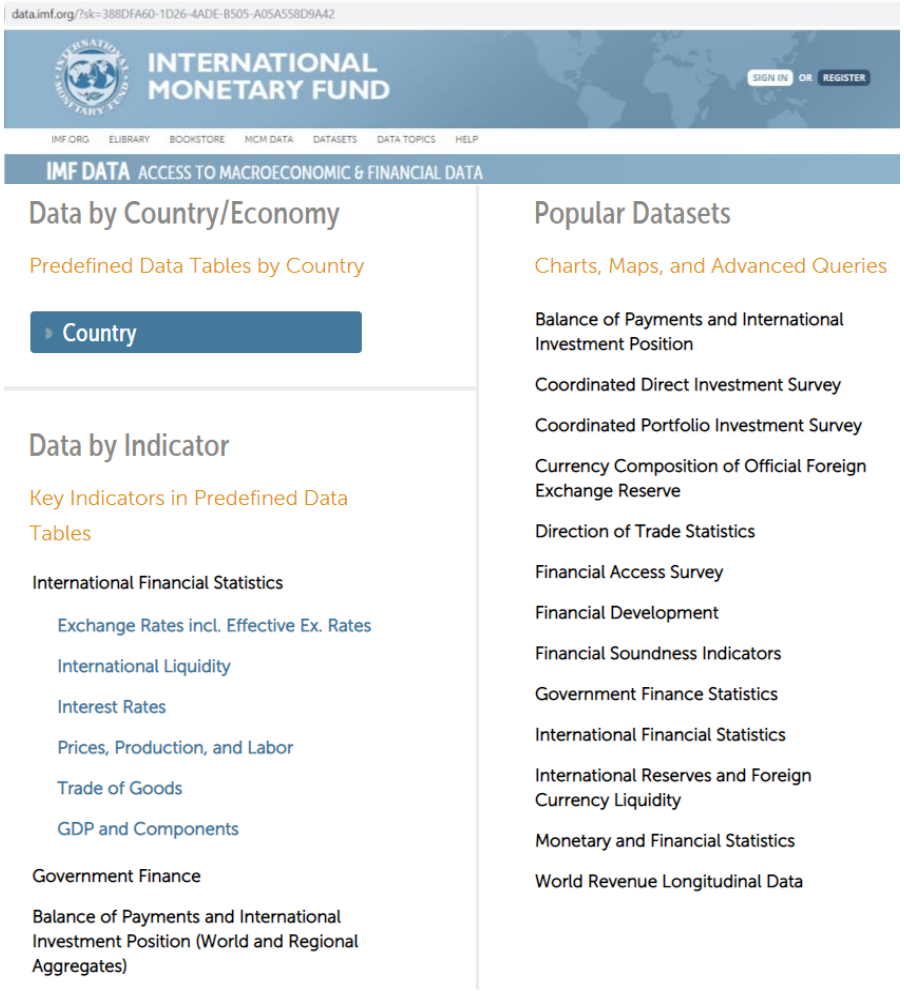




Country Name	Country Co	Indicator Name	Indicator Code
Portugal	PRT	CO2 emissions (kt)	EN.ATM.CO2E.KT
Portugal	PRT	CO2 emissions (metric tons per capita)	EN.ATM.CO2E.PC
Portugal	PRT	GDP (current US$)	NY.GDP.MKTP.CD
Portugal	PRT	GDP growth (annual %)	NY.GDP.MKTP.KD.ZG
Portugal	PRT	Inflation, consumer prices (annual %)	FP.CPI.TOTL.ZG
Portugal	PRT	Market capitalization of listed domestic companies (% of GDP)	CM.MKT.LCAP.GD.ZS
Portugal	PRT	Net migration	SM.POP.NETM
Portugal	PRT	Patent applications, nonresidents	IP.PAT.NRES
Portugal	PRT	Patent applications, residents	IP.PAT.RESD
Portugal	PRT	PM2.5 air pollution, mean annual exposure (micrograms per cubic meter)	EN.ATM.PM25.MC.M3
Portugal	PRT	Population, total	SP.POP.TOTL
Portugal	PRT	Portfolio investment, net (BoP, current US$)	BN.KLT.PTXL.CD
Portugal	PRT	Profit tax (% of commercial profits)	IC.TAX.PRFT.CP.ZS
Portugal	PRT	Research and development expenditure (% of GDP)	GB.XPD.RSDV.GD.ZS
Portugal	PRT	Researchers in R&D (per million people)	SP.POP.SCIE.RD.P6
Portugal	PRT	Technicians in R&D (per million people)	SP.POP.TECH.RD.P6
Portugal	PRT	Unemployment, total (% of total labor force) (modeled ILO estimate)	SL.UEM.TOTL.ZS
Portugal	PRT	Unemployment, total (% of total labor force) (national estimate)	SL.UEM.TOTL.NE.ZS

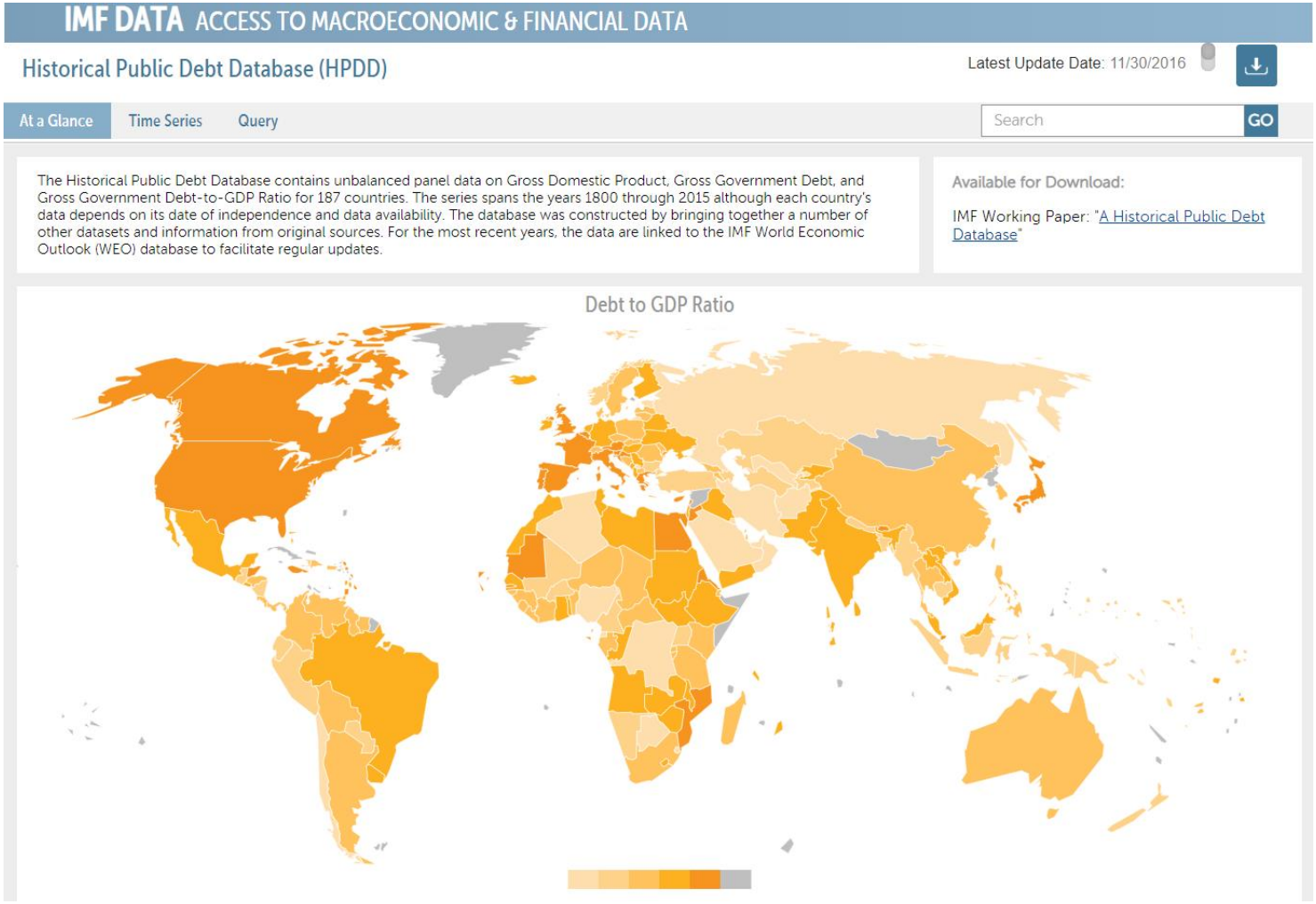
IMF

- **International Monetary Fund (Macroeconomic and financial data)**
  - www.data.imf.org



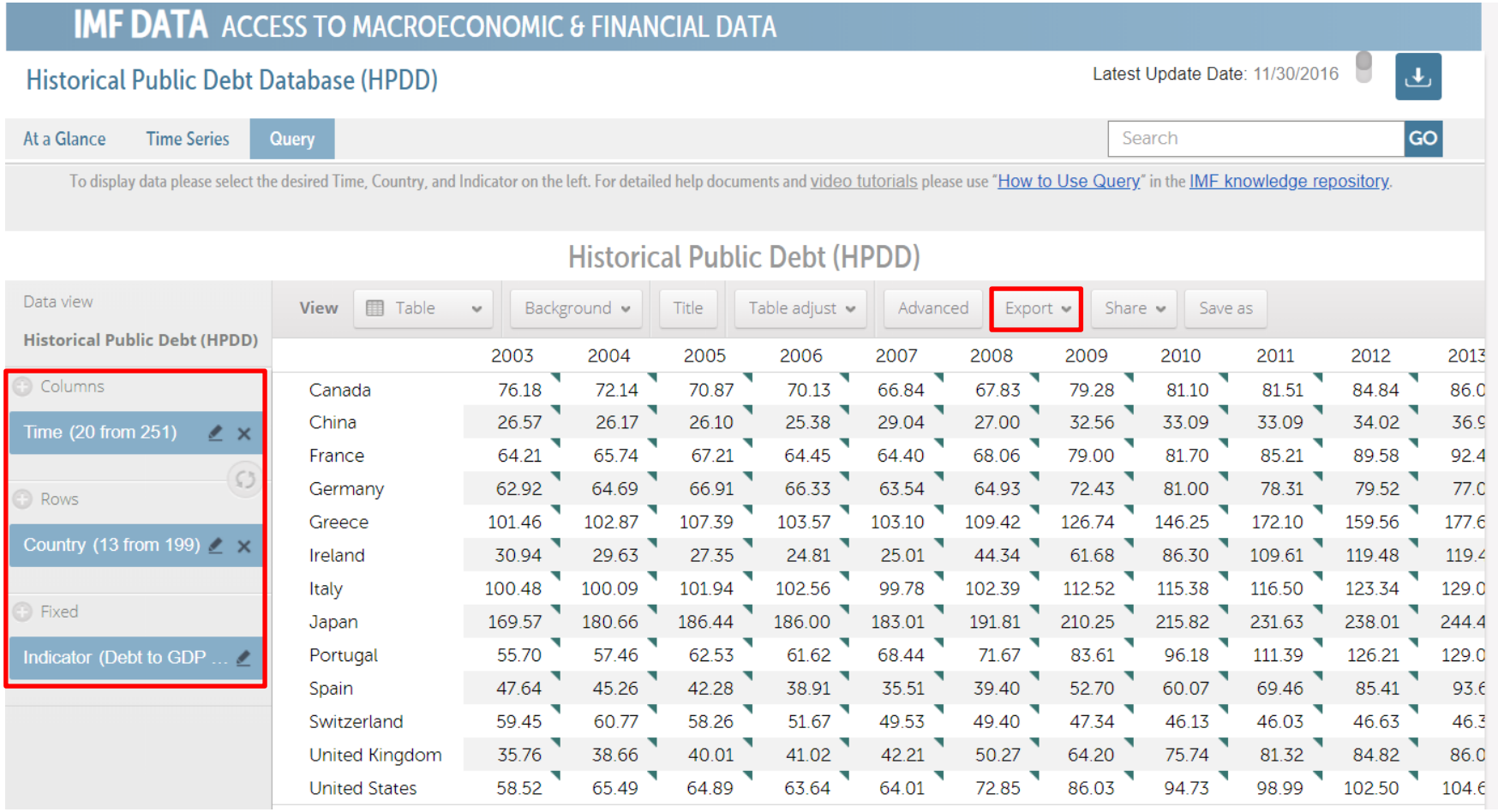
Outside WRDS… (IMF)

IMF



Outside WRDS… (IMF)

IMF

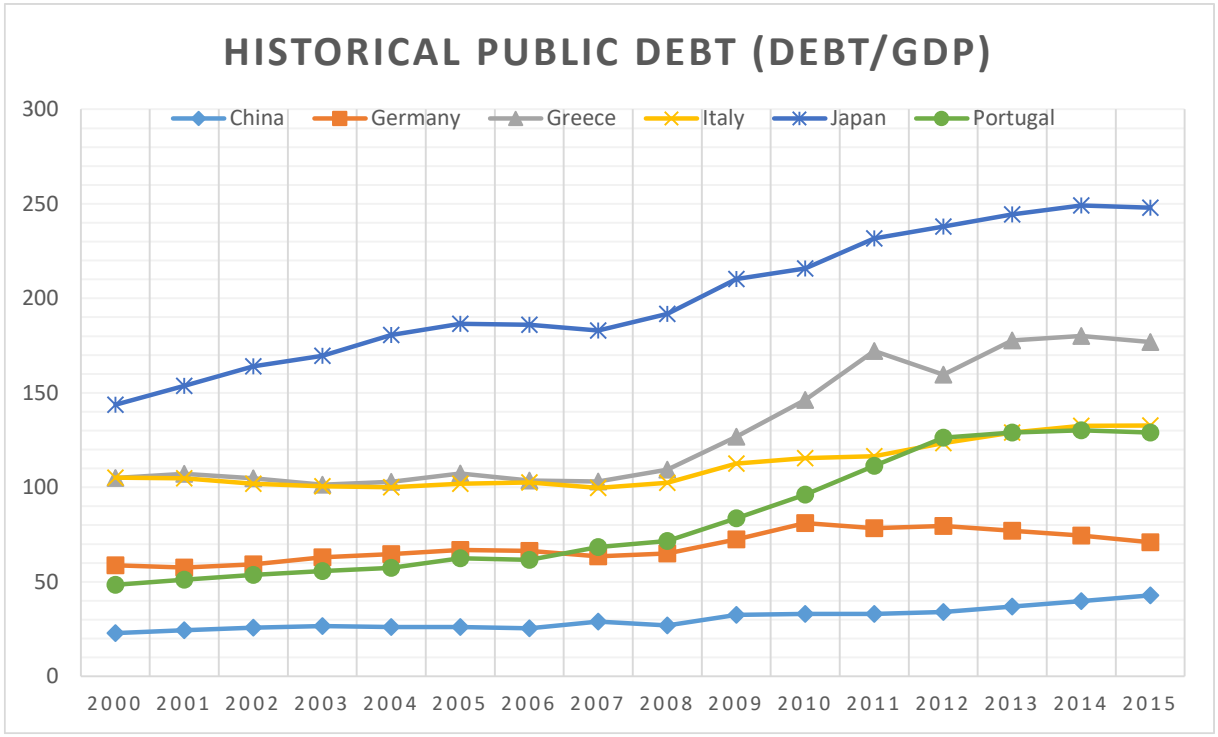


	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Canada	76.18	72.14	70.87	70.13	66.84	67.83	79.28	81.10	81.51	84.84	86.0
China	26.57	26.17	26.10	25.38	29.04	27.00	32.56	33.09	33.09	34.02	36.9
France	64.21	65.74	67.21	64.45	64.40	68.06	79.00	81.70	85.21	89.58	92.4
Germany	62.92	64.69	66.91	66.33	63.54	64.93	72.43	81.00	78.31	79.52	77.0
Greece	101.46	102.87	107.39	103.57	103.10	109.42	126.74	146.25	172.10	159.56	177.6
Ireland	30.94	29.63	27.35	24.81	25.01	44.34	61.68	86.30	109.61	119.48	119.4
Italy	100.48	100.09	101.94	102.56	99.78	102.39	112.52	115.38	116.50	123.34	129.0
Japan	169.57	180.66	186.44	186.00	183.01	191.81	210.25	215.82	231.63	238.01	244.4
Portugal	55.70	57.46	62.53	61.62	68.44	71.67	83.61	96.18	111.39	126.21	129.0
Spain	47.64	45.26	42.28	38.91	35.51	39.40	52.70	60.07	69.46	85.41	93.6
Switzerland	59.45	60.77	58.26	51.67	49.53	49.40	47.34	46.13	46.03	46.63	46.3
United Kingdom	35.76	38.66	40.01	41.02	42.21	50.27	64.20	75.74	81.32	84.82	86.0
United States	58.52	65.49	64.89	63.64	64.01	72.85	86.03	94.73	98.99	102.50	104.6

Outside WRDS… (IMF)

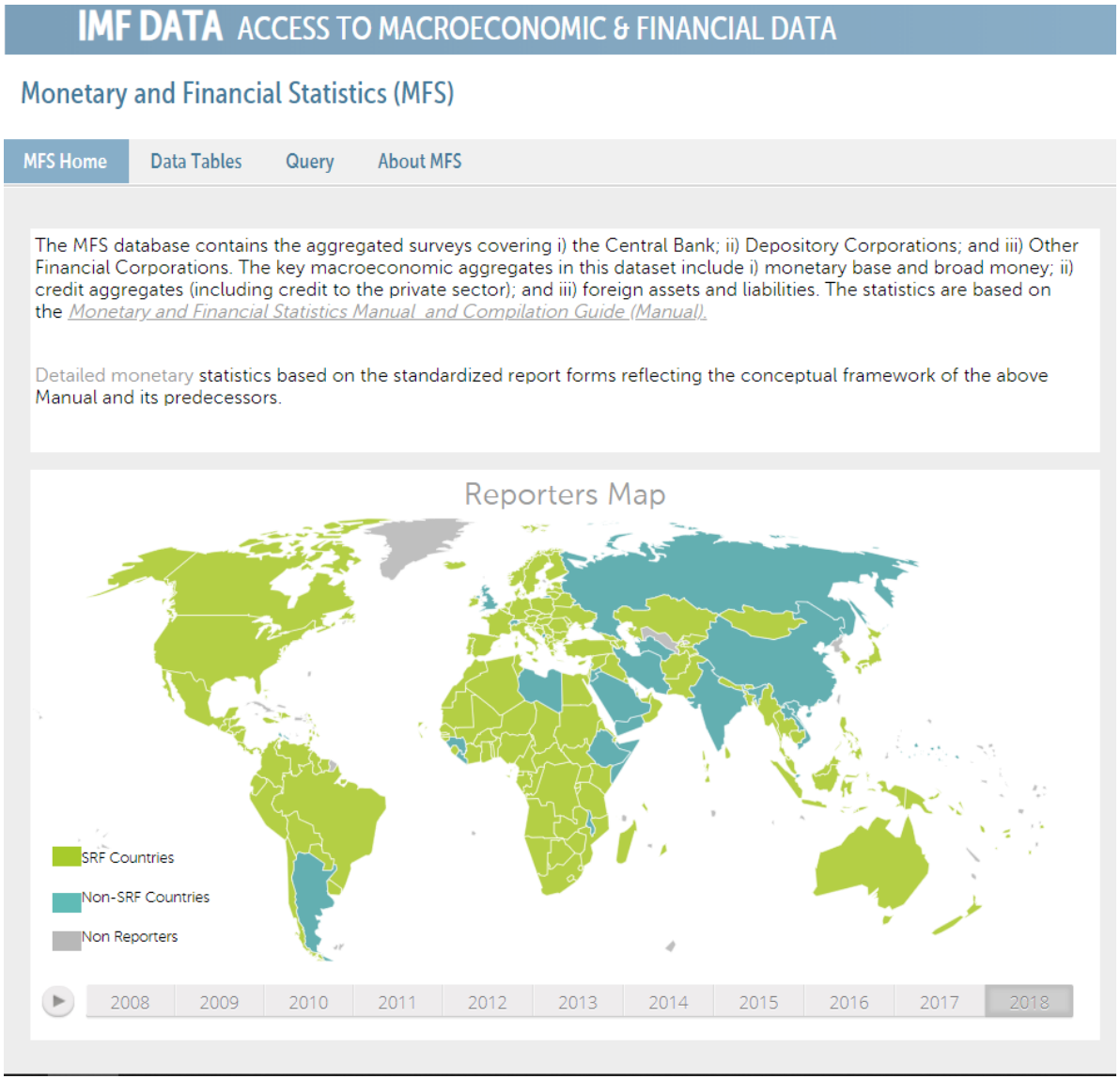
IMF

	Historical Public Debt (HPDD)															
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
China	22.81	24.38	25.71	26.57	26.17	26.10	25.38	29.04	27.00	32.56	33.09	33.09	34.02	36.93	39.83	42.92
Germany	58.76	57.58	59.22	62.92	64.69	66.91	66.33	63.54	64.93	72.43	81.00	78.31	79.52	77.06	74.48	70.99
Greece	104.93	107.08	104.86	101.46	102.87	107.39	103.57	103.10	109.42	126.74	146.25	172.10	159.56	177.68	180.06	176.94
Italy	105.11	104.73	101.92	100.48	100.09	101.94	102.56	99.78	102.39	112.52	115.38	116.50	123.34	129.00	132.53	132.71
Japan	143.78	153.63	163.99	169.57	180.66	186.44	186.00	183.01	191.81	210.25	215.82	231.63	238.01	244.48	249.11	247.98
Portugal	48.36	51.07	53.68	55.70	57.46	62.53	61.62	68.44	71.67	83.61	96.18	111.39	126.21	129.00	130.17	128.98



Outside WRDS… (IMF)

IMF

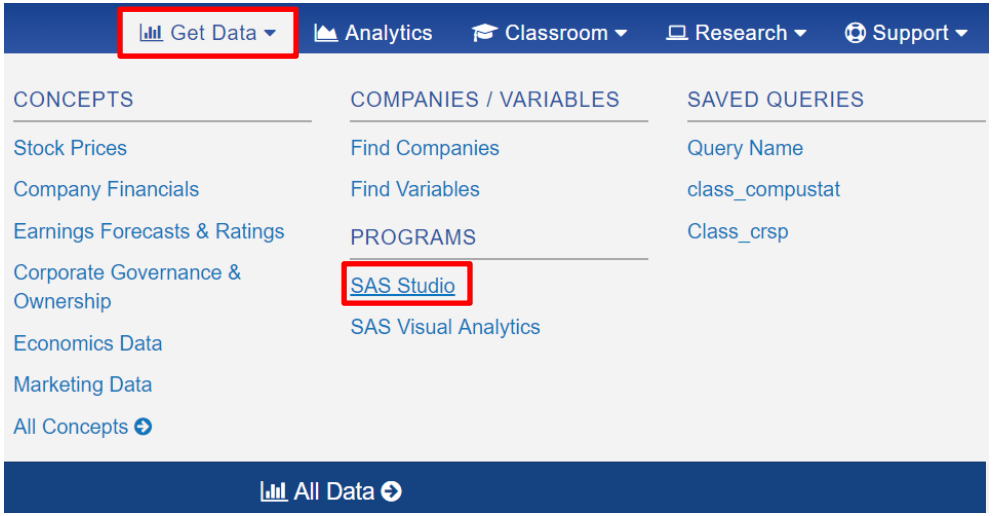


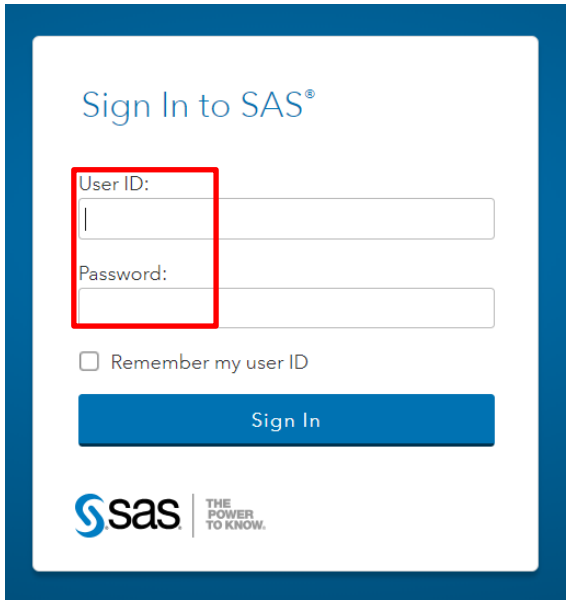
Course contents:

➢ Section II

- **Brief introduction to SAS and SQL**
- **Data management in SAS Studio**
  - ✓ **I/O (Input-Output) files**
  - ✓ **Querying data**
  - ✓ **Aggregating/Summarizing data**
  - ✓ **Merging datasets**
  - ✓ **Fuzzy matching**

UNIX – SAS Studio





- SAS studio is a **friendly visual interface** to the SAS server running on WRDS
- It looks like the SAS interface on the PC but has remote access to WRDS data like on UNIX
- All standard browsers are supported (Firefox, Google Chrome, Internet Explorer, Safari)

SAS Studio on WRDS

SAS Studio is a web application that lets you access SAS through your browser.

How It Works

SAS Studio is an interface to the SAS server running on the WRDS environment. It complements SAS on Unix, SAS Connect on a PC, and the SAS Share interface to R, Matlab, and other languages. Since a new connection is established, you will be prompted to **enter your password**.

It looks like the SAS interface found on the PC, but has direct access to WRDS data like on Unix.

All standard browsers are supported, including Firefox, Chrome, Internet Explorer, and Safari.

Who should use it?

At this time, SAS Share is available to Faculty, staff, PhD students, and research assistants.

People using a **Mac** and those who happen to be at a computer that does not have PC SAS installed will find SAS Share very convenient. Unix users looking for a more **friendly interface** may also find it useful.

» Get Started

SAS Studio is currently being beta tested. If you have feedback or experience any problems, please let us know!

Wharton University of Pennsylvania | wrds The Global Standard for Business Research

Home / Support / Programming With WRDS

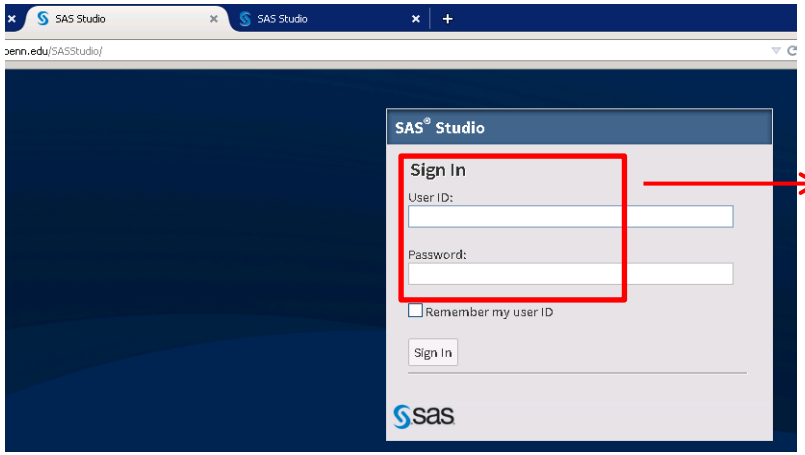
Programming With WRDS

- SAS Programming
- MATLAB Programming
- R Programming
- Python Programming
- Other Languages: C, Fortran

Why SAS?

- WRDS platform is built under a SAS/SQL backend.
- SAS doesn't load the entire dataset to your PC RAM. It works sequentially in the dataset which is stored in the disk. Good at handling large datasets.
- SAS allows you to work remotely in the WRDS Server.
  - You don't need to have a super laptop even when working with big datasets.
  - You don't need to download all datasets to your PC. You can work in the server and then just download the final output file.
  - WRDS datasets and sample programs/research applications are in SAS.
- Other Options: Stata, R, Matlab, Python, etc.. **R** is the Open source counterpart of Matlab or SAS. The number of researchers using it is expanding fast.

User-friendly interface



Create new folder “bootcamp” in your home directory

Hands-on example:

Download the monthly stock prices from CRSP for Chevron, Exxon Mobil and Conoco Phillips (permnos: 14541, 11850, 13928) in 2014.

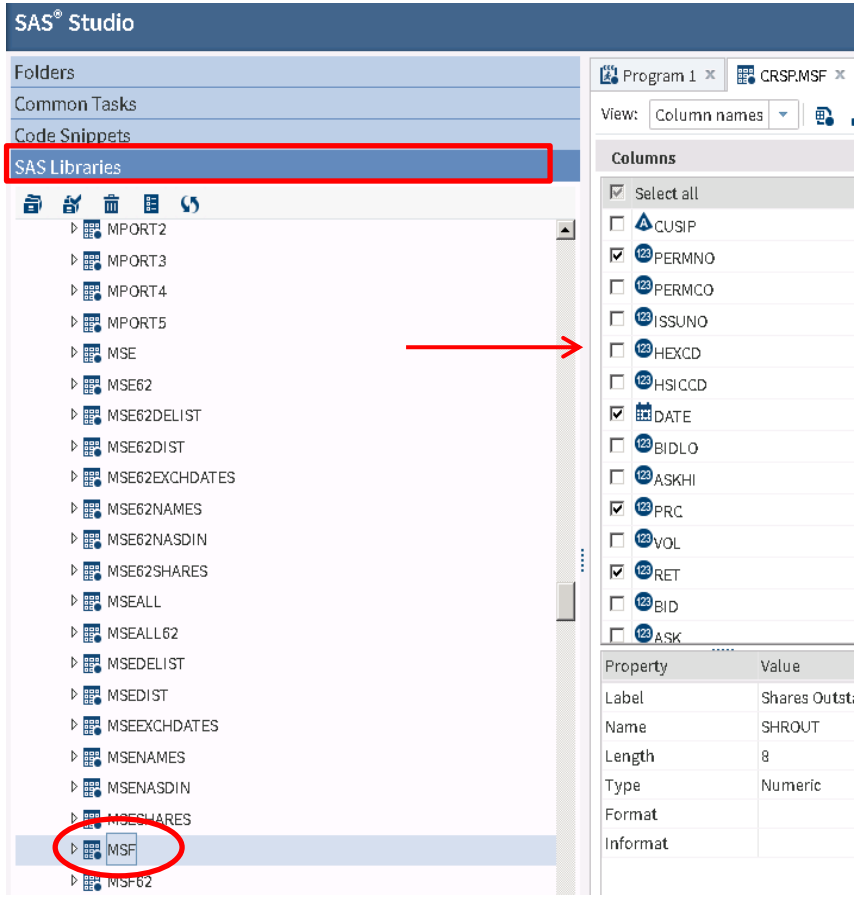
Hands-on example:

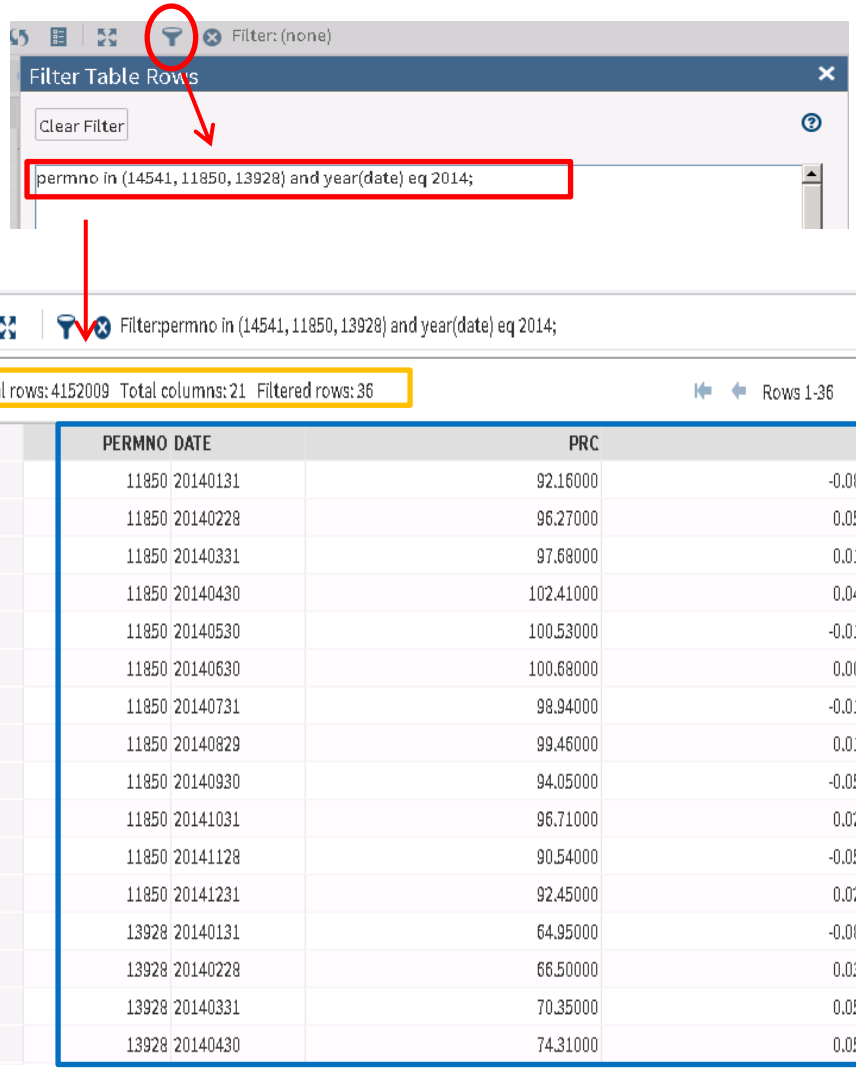
Download the monthly stock prices from CRSP for Chevron, Exxon Mobil and Conoco Phillips (permnos: 14541, 11850, 13928) in 2014.

Filtering in the Table View

- Select **CRSP** in the left navigation menu under **SAS libraries**
- Inside **CRSP** click on **MSF**
- Select the columns **permno, date, prc, ret**
- **Filter** the table to get the results needed

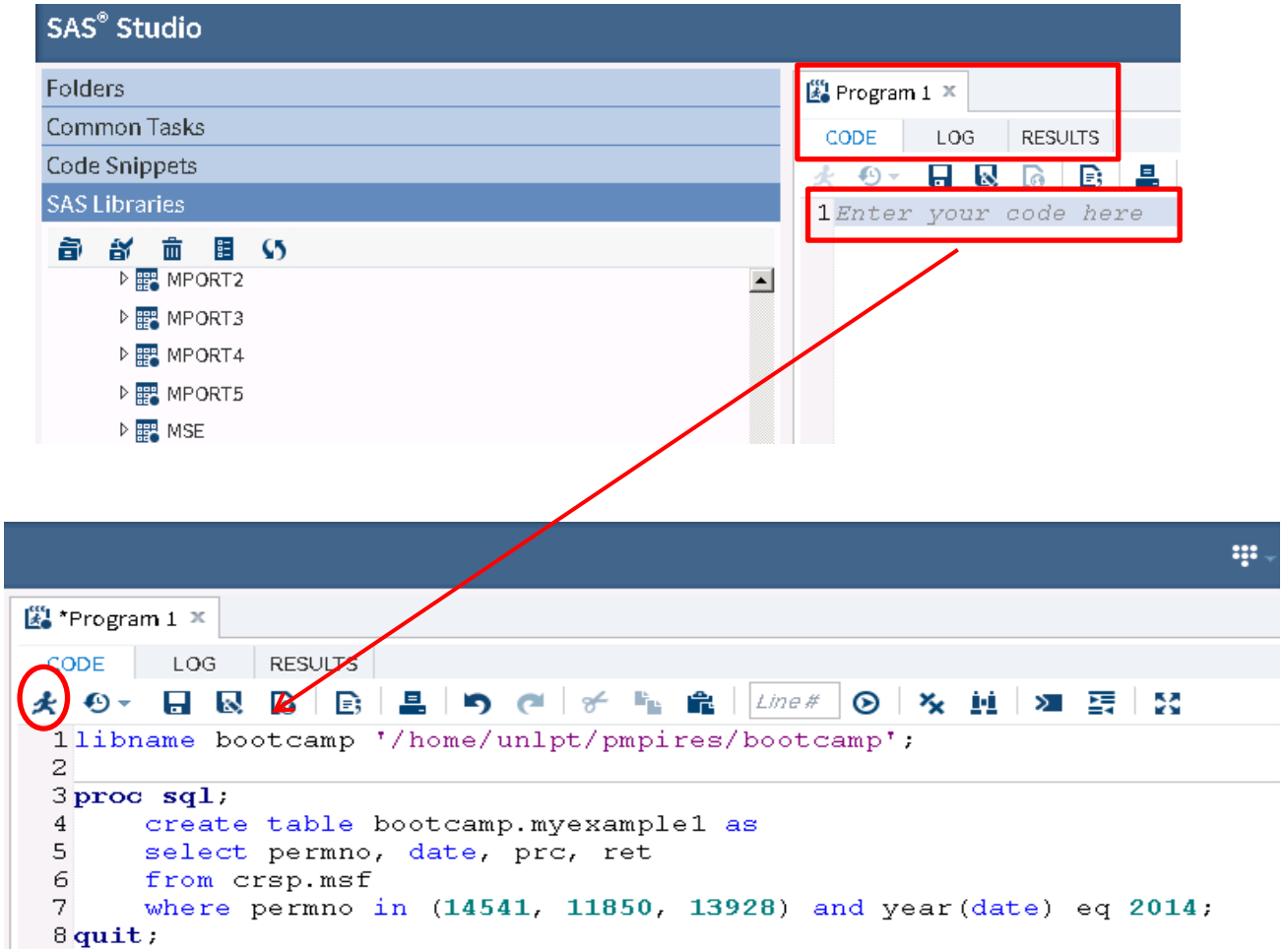
(1) Filtering in the Table View



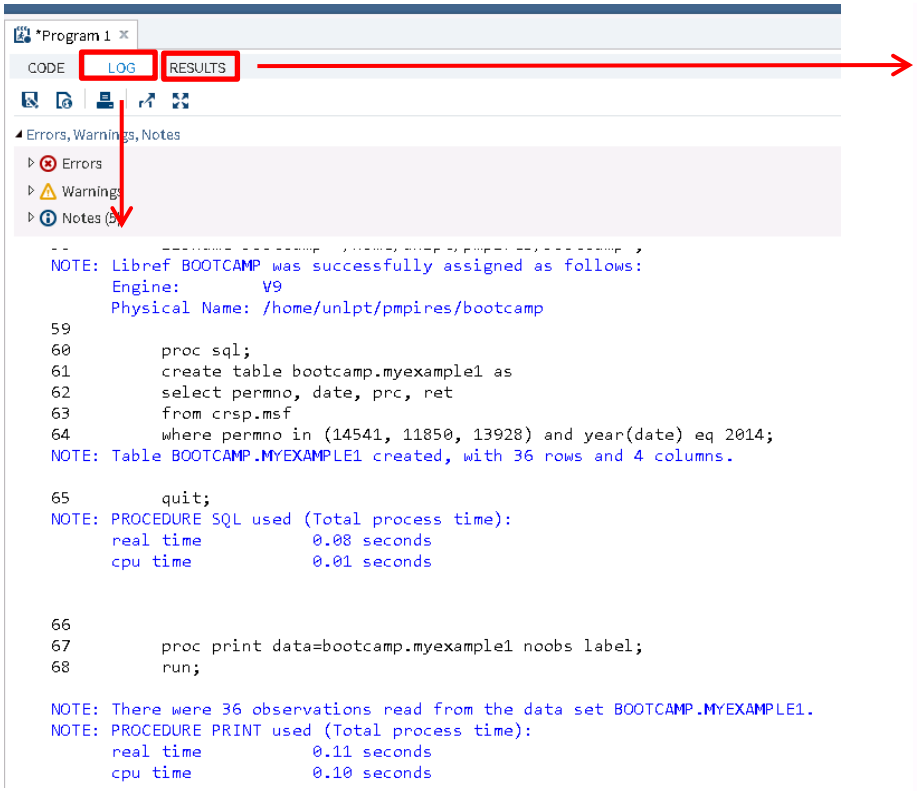


	PERMNO	DATE	PRC	RET
1	11850	20140131	92.16000	-0.089328
2	11850	20140228	96.27000	0.051432
3	11850	20140331	97.68000	0.014646
4	11850	20140430	102.41000	0.048423
5	11850	20140530	100.53000	-0.011620
6	11850	20140630	100.68000	0.001492
7	11850	20140731	98.94000	-0.017282
8	11850	20140829	99.46000	0.012230
9	11850	20140930	94.05000	-0.054394
10	11850	20141031	96.71000	0.028283
11	11850	20141128	90.54000	-0.056664
12	11850	20141231	92.45000	0.021096
13	13928	20140131	64.95000	-0.080679
14	13928	20140228	66.50000	0.034488
15	13928	20140331	70.35000	0.057895
16	13928	20140430	74.31000	0.056290

(2) Program a Query



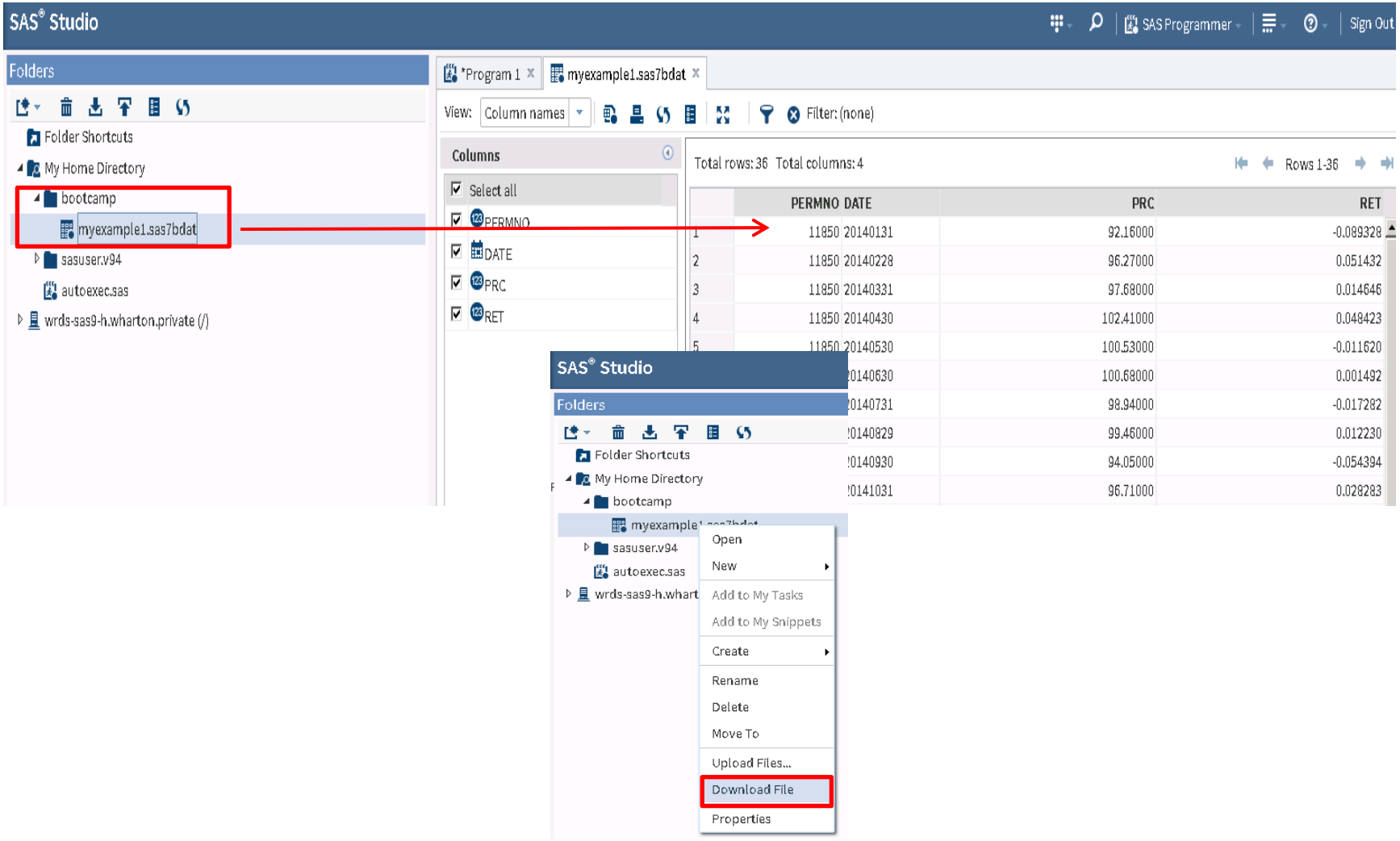
(2) Program a Query (cont.)




```
NOTE: Libref BOOTCAMP was successfully assigned as follows:
      Engine:        V9
      Physical Name: /home/unlpt/pmpires/bootcamp
59
60         proc sql;
61         create table bootcamp.myexample1 as
62         select permno, date, prc, ret
63         from crsp.msf
64         where permno in (14541, 11850, 13928) and year(date) eq 2014;
NOTE: Table BOOTCAMP.MYEXAMPLE1 created, with 36 rows and 4 columns.

65         quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.08 seconds
      cpu time            0.01 seconds


66
67         proc print data=bootcamp.myexample1 noobs label;
68         run;

NOTE: There were 36 observations read from the data set BOOTCAMP.MYEXAMPLE1.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.11 seconds
      cpu time            0.10 seconds
```

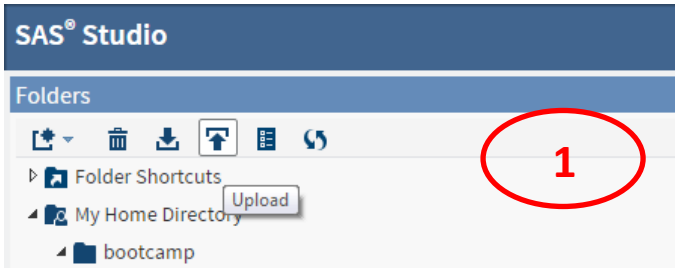
PERMNO	Date of Observation	Price or Bid/Ask Average	Returns
11850	20140131	92.16000	-0.089328
11850	20140228	96.27000	0.051432
11850	20140331	97.68000	0.014646
11850	20140430	102.41000	0.048423
11850	20140530	100.53000	-0.011620
11850	20140630	100.68000	0.001492
11850	20140731	98.94000	-0.017282
11850	20140829	99.46000	0.012230
11850	20140930	94.05000	-0.054394
11850	20141031	96.71000	0.028283
11850	20141128	90.54000	-0.056664
11850	20141231	92.45000	0.021096
13928	20140131	64.95000	-0.080679
13928	20140228	66.50000	0.034488
13928	20140331	70.35000	0.057895
13928	20140430	74.31000	0.056290
13928	20140530	79.94000	0.085049
13928	20140630	85.73000	0.072429
13928	20140731	82.50000	-0.029161
13928	20140829	81.22000	-0.015515
13928	20140930	76.52000	-0.057868
13928	20141031	72.15000	-0.047569
13928	20141128	66.07000	-0.084269
13928	20141231	69.06000	0.045255
14541	20140131	111.63000	-0.106317
14541	20140228	115.33000	0.042103
14541	20140331	118.91000	0.031041
14541	20140430	125.52000	0.055588
14541	20140530	122.79000	-0.013225
14541	20140630	130.55000	0.063197

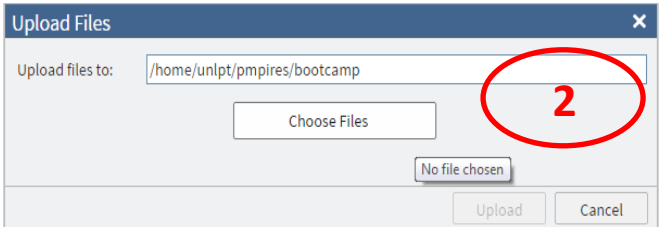
(2) Program a Query (cont.) - Download a file

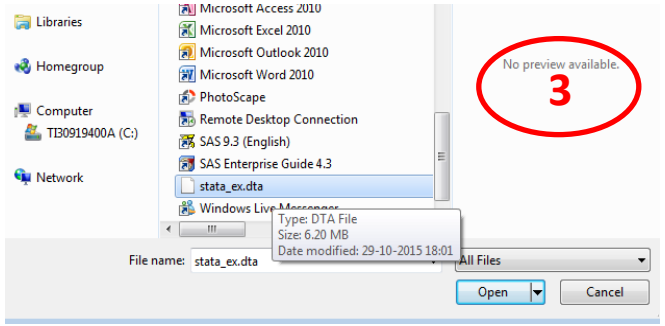


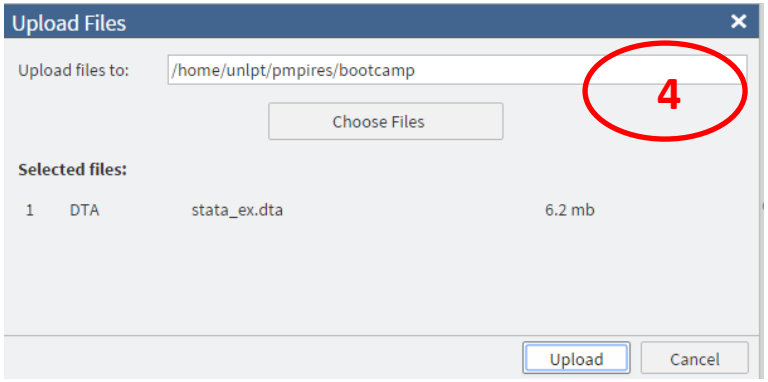
Uploa a file

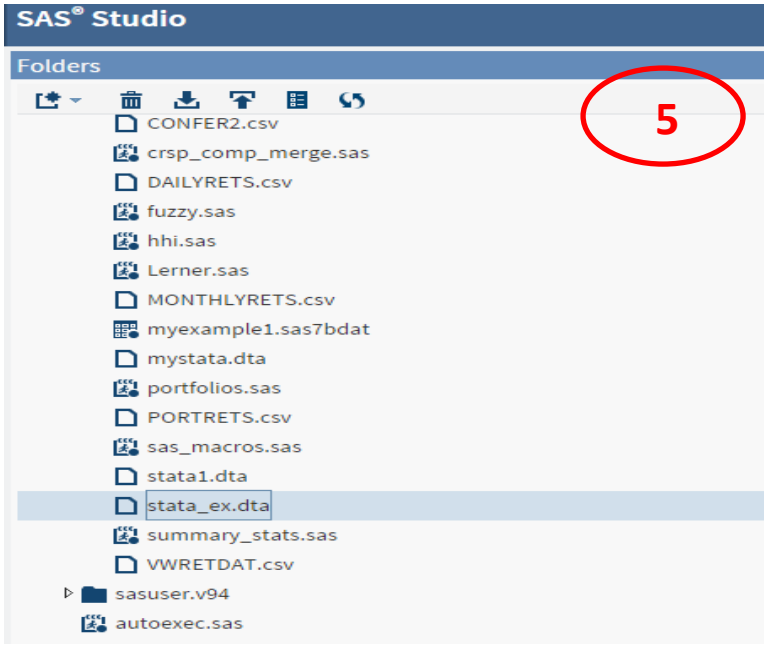
➢ Upload a file from your PC to WRDS





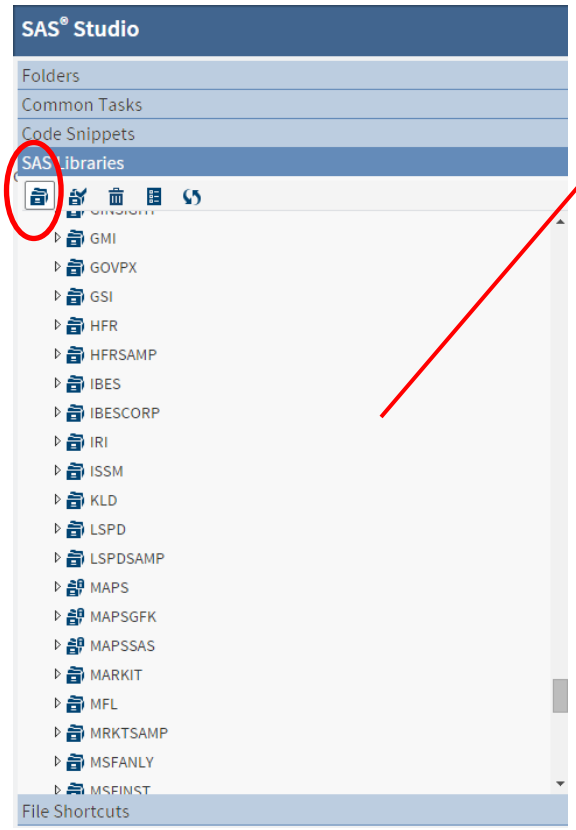


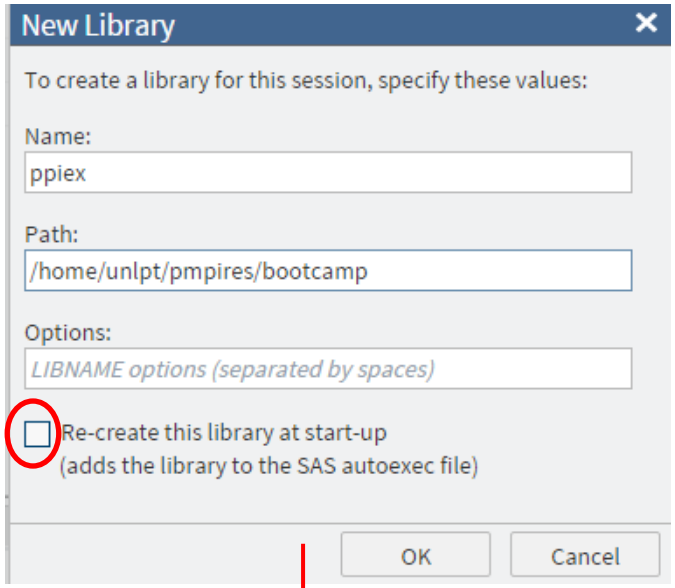


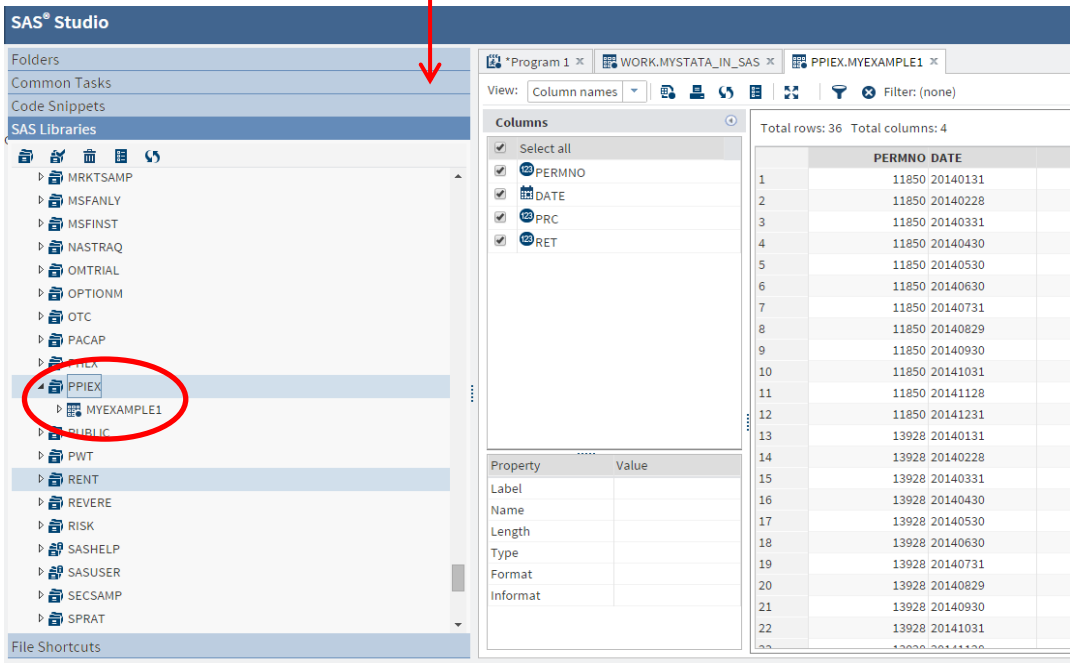


More on using SAS Studio…

- **Defining a SAS Library**







Introduction

- A Table is a two-dimensional representation of data consisting of **columns** and **rows**

	name	year	score
1	John	2012	55
2	John	2013	65
3	John	2014	70
4	Pedro	2012	75
5	Pedro	2013	80
6	Pedro	2014	90
7	Marta	2012	75
8	Marta	2013	65
9	Marta	2014	80

Terminology

Data Processing	SAS	SQL
File	Dataset	Table
Record	Observation	Row
Field	Variable	Column

PROC SQL

- Use the “proc sql” instruction to turn SAS into an efficient SQL environment

```
proc sql;
    [SQL Instruction]
quit;
```

Structure of a SQL Query

- SELECT columns (variables)
- FROM tables (datasets)
- WHERE row (observation) conditions that must be met
- GROUP BY summarize by these variables
- HAVING summary conditions that must be met
- ORDER BY sort by these columns

Querying data and storing results

- By default the results of a query are displayed in the SAS output window

```
proc sql;
    select *
    from saslibrary.table;
quit;
```

Select ALL variables in the dataset using *

- You can create a dataset (table) to store the output of your query

```
proc sql;
    create table table as
    select *
    from saslibrary.table;
quit;
```

Sorting

```
select name, year, score
from work.scores
order by year, score desc;
```

Creating new variables

```
select name, year, score, 90-score as distance
from work.scores
```

Duplicates

```
select distinct name
from work.scores
```

Conditional logic (select statement)

```
select name, year, score,
      case
         when score < 50 then 'Fail'
         else 'Pass'
      end as mark
from work.scores
```

Subsetting data

```
select name, year, score
from work.scores
where year = 2014

(…)

where year = 2014 and name = 'Pedro'
```

BETWEEN operator

```
select name, year, score
from work.scores
where score between 80 and 90
```

LIKE operator

```
select name, year, score
from work.scores
where name like '%Ped%'
```

Contains: '%Ped%'

Starts with: 'Ped%'

Ends with: '%dro'

IN operator

```
select name, year, score
from work.scores
where name in ('Pedro', 'Marta')
```

- Strings in SQL are <u>case sensitive</u>
- Use lower and upper functions to convert strings to lower- or upper-case

Standard comparison operators

EQ	=	Equal to
NE	^= or <>	Not equal to
GE	>=	Greater than or equal to
GT	>	Greater than
LE	<=	Less than or equal to
LT	<	Less than

Summarizing or aggregating data

```
select year, avg(score) as avg_score
from work.scores
group by year
```

- SUM
- AVG
- MIN
- MAX
- COUNT
- MEDIAN
- (...)

Subsetting with group by clauses

```
select name, avg(score) as avg_score
from work.scores
group by name
having avg(score) > 70
```

Subqueries

- A query expression nested as part of another query expression

```
select *
from address
where name in
              (select distinct name from scores
                       where score ge 75 and year eq 2014);
```

Creating a table and sorting with SAS Data Step

```
data table1;
set scores (keep=name score);
where year=2014;
run;

proc sort data=table1;
by descending score;
run;
```

Relational database

- Data in tables can be logically related according to common keys (variables)

Table: Scores

	name	year	score
1	John	2012	55
2	John	2013	65
3	John	2014	70
4	Pedro	2012	75
5	Pedro	2013	80
6	Pedro	2014	90
7	Marta	2012	75
8	Marta	2013	65
9	Marta	2014	80

9 records

Table: Address

	name	city	zip
1	John	Almada	5220
2	Pedro	Lisbon	1500
3	Marta	Cascais	2850

3 records

Relational database

- Data in tables can be logically related according to common keys (variables)

```
data scores;
input name $ year score;
datalines;
John 2012 55
John 2013 65
John 2014 70
Pedro 2012 75
Pedro 2013 80
Pedro 2014 90
Marta 2012 75
Marta 2013 65
Marta 2014 80
;
run;

data address;
input name $ city $ zip;
datalines;
John Almada 5220
Pedro Lisbon 1500
Marta Cascais 2850
;
run;
```

Brief introduction to SAS

CODE LOG RESULTS OUTPUT DATA

Table: WORK.SCORES | View: Column names | Filter: (none)

Columns

Select all

- name
- year
- score

Total rows: 9 Total columns: 3

	name	year	score
1	John	2012	55
2	John	2013	65
3	John	2014	70
4	Pedro	2012	75
5	Pedro	2013	80
6	Pedro	2014	90
7	Marta	2012	75
8	Marta	2013	65
9	Marta	2014	80

CODE LOG RESULTS OUTPUT DATA

Table: WORK.ADDRESS | View: Column names | Filter:

Columns

Select all

- name
- city
- zip

Total rows: 3 Total columns: 3

	name	city	zip
1	John	Almada	5220
2	Pedro	Lisbon	1500
3	Marta	Cascais	2850

Joining Tables

- Cartesian Product: Each row of the first table is combined with every row from the second table (number of rows table 1 * number of rows table 2; can be huge!)

```
create table cartesian as
select *
from work.scores, work.address
order by name, year;
```

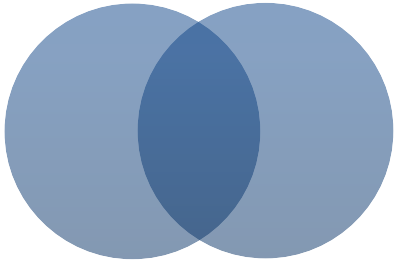
- **27 records (=9*3)**
- **Now, we have three duplicate recors for each pair (name, year)**

NOTE: The execution of this query involves performing one or more Cartesian product joins that can not be optimized.

	name	year	score	city	zip
1	John	2012	55	Almada	5220
2	John	2012	55	Cascais	2850
3	John	2012	55	Lisbon	1500
4	John	2013	65	Almada	5220
5	John	2013	65	Cascais	2850
6	John	2013	65	Lisbon	1500
7	John	2014	70	Almada	5220
8	John	2014	70	Cascais	2850
9	John	2014	70	Lisbon	1500
10	Marta	2012	75	Almada	5220
11	Marta	2012	75	Cascais	2850
12	Marta	2012	75	Lisbon	1500
13	Marta	2013	65	Cascais	2850
14	Marta	2013	65	Almada	5220
15	Marta	2013	65	Lisbon	1500
16	Marta	2014	80	Cascais	2850
17	Marta	2014	80	Almada	5220
18	Marta	2014	80	Lisbon	1500
19	Pedro	2012	75	Lisbon	1500
20	Pedro	2012	75	Cascais	2850
21	Pedro	2012	75	Almada	5220
22	Pedro	2013	80	Lisbon	1500
23	Pedro	2013	80	Almada	5220
24	Pedro	2013	80	Cascais	2850
25	Pedro	2014	90	Lisbon	1500
26	Pedro	2014	90	Almada	5220
27	Pedro	2014	90	Cascais	2850

Inner Join

- Returns the subset of rows from the first table that matches rows from the second table



```
create table inner as
select *
from scores as t1, address as t2
where t1.name = t2.name
order by t1.name, t1.year;

(...)
from scores as t1
inner join address as t2
   on t1.name = t2.name
```

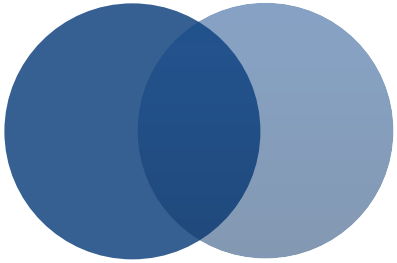
Use table alias to abbreviate table names

	name	year	score	city	zip
1	John	2012	55	Almada	5220
2	John	2013	65	Almada	5220
3	John	2014	70	Almada	5220
4	Marta	2012	75	Cascais	2850
5	Marta	2013	65	Cascais	2850
6	Marta	2014	80	Cascais	2850
7	Pedro	2012	75	Lisbon	1500
8	Pedro	2013	80	Lisbon	1500
9	Pedro	2014	90	Lisbon	1500

Left Outer Join

- Returns all the rows from the first table and the fields matched with the second table

```
proc sql;
delete from address
where name eq 'John';
```



```
create table left as
select *
from scores as t1
Left join address as t2
   on t1.name = t2.name
order by t1.name, t1.year;
```

Table: Address *

	name	city	zip
1	Pedro	Lisbon	1500
2	Marta	Cascais	2850

	name	year	score	city	zip
1	John	2012	55		.
2	John	2013	65		.
3	John	2014	70		.
4	Marta	2012	75	Cascais	2850
5	Marta	2013	65	Cascais	2850
6	Marta	2014	80	Cascais	2850
7	Pedro	2012	75	Lisbon	1500
8	Pedro	2013	80	Lisbon	1500
9	Pedro	2014	90	Lisbon	1500

Full Outer Join

- Returns all matching and non-matching rows

```
create table full as
select coalesce (t1.name, t2.name) as name,
t1.year, t1.score, t2.city, t2.zip
from scores as t1
full join address as t2
   on t1.name = t2.name
order by calculated name,
t1.year;
```

Table: Address *

	name	city	zip
1	John	Almada	5220
2	Pedro	Lisbon	1500
3	Marta	Cascais	2850
4	Joana	Porto	3225

	name	year	score	city	zip
1	Joana	.	.	Porto	3225
2	John	2012	55	Almada	5220
3	John	2013	65	Almada	5220
4	John	2014	70	Almada	5220
5	Marta	2012	75	Cascais	2850
6	Marta	2013	65	Cascais	2850
7	Marta	2014	80	Cascais	2850
8	Pedro	2012	75	Lisbon	1500
9	Pedro	2013	80	Lisbon	1500
10	Pedro	2014	90	Lisbon	1500

Brief introduction to SAS

```
data scores;
input name $ year score;
datalines;
John 2012 55
John 2013 65
John 2014 70
Pedro 2012 75
Pedro 2013 80
Pedro 2014 90
Marta 2012 75
Marta 2013 65
Marta 2014 80
;
run;
```

1

```
data address;
input name $ city $ zip;
datalines;
John Almada 5220
Pedro Lisbon 1500
Marta Cascais 2850
;
run;
```

```
proc sql;
create table cartesian as
select *
from work.scores, work.address
order by name, year;
```

```
proc sql;
create table inner as
select *
from scores as t1, address as t2
where t1.name = t2.name
order by t1.name, t1.year;
```

2

```
proc sql; delete from address where name eq 'John';
```

```
proc sql;
create table left as
select *
from scores as t1
Left join address as t2
    on t1.name = t2.name
order by t1.name, t1.year;
```

```
data address;
input name $ city $ zip;
datalines;
John Almada 5220
Pedro Lisbon 1500
Marta Cascais 2850
Joana Porto 3225
;
run;
```

3

```
proc sql;
create table full as
select coalesce (t1.name, t2.name) as name,
t1.year, t1.score, t2.city, t2.zip
from scores as t1
full join address as t2
    on t1.name = t2.name
order by calculated name, t1.year;
```