

Quantitative Methods

Working Environment and data base

Ludovic Vigneron

University of Lille (IAE)

September 6th, 2022

Morning program

1. Intalling Softwares
2. R langage basis
3. WRDS data base
4. Extracting/importing data

Intalling Softwares

R language

R is a programming language dedicated to statistics and data sciences.

- **easy to use** (simple syntax especially using the tidyverse)
- **free** (that makes a huge difference with Stat, SAS, Eviews, SPSS...) and multi-platform (windows, OS, Linux)
- **evolutive** (a very important community develops applications -free packages- and makes 'R base' evolve on a regular basis)
- **a lot of information available** ([R journal](#), [R bigbook](#), [R-BLOGGERS](#), [Stack overflow](#), [freakonometrics](#), [MyEconTricks](#)...)

R language

You can download R at the following address <https://www.r-project.org/> .

R Studio IDE

- What is a *IDE* ?

It is just a *Interactif Developpement Environment*, a software that helps to deal with application creation using one or more programing language.

R Studio is dedicated to data analysis. It provides tools to get data, to shape data (data wrangling), to draw graphs, to do statistics (test, regression models, other modeling...) and to produce reports (presentations, text documents, interactive product...).

Actualy, R studio is becoming posit. It will give more space to other programing langage (like Pyhton, Julia,C++...).

R Studio IDE

You can download R Studio at the following address <https://www.rstudio.com/products/rstudio/download/>.

Video installation guide

To help you to install R and R Studio IDE, I invite you to watch the follow trought video (duration about 5 minutes) here after.

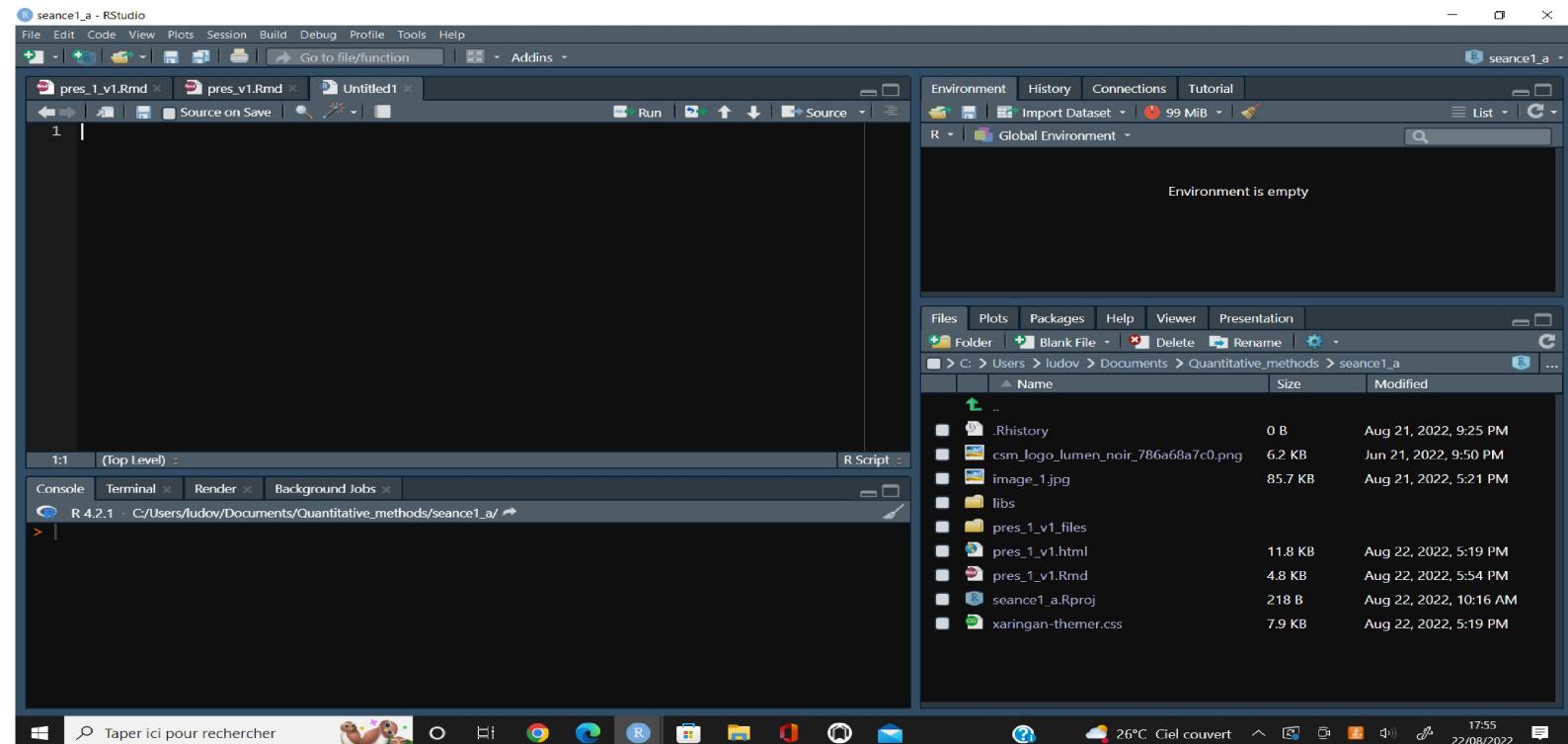
<https://youtu.be/hT2tSvj7d1A>

Presentation of the R Studio environment

You will find a short presentation of the R Studio in the following video (duration about 15 minutes).

<https://youtu.be/fB5QlyjWqK0>

Presentation of the R Studio environement



Presentation of the R Studio environment

Four panel layouts

- **Source** : where you write your program and save it to make your analysis reproducible.
- **Console** : where you interact with R.
- **Environment** : where you find the different objects that you have created during your session.
- **Files, Plots, Packages, Help, Viewer, Presentation** : you find the files in your working directory, you can view the graph made, the list of your installed packages, some descriptions of functions for which you have asked for help...

Presentation of the R Studio environment

It is important to create a new project when you start a data analysis.

file -> new project

Indicate where to put the directory associated with your project and name it.

If you have some doubt on your actual working directory type `getwd()` in the console space.

In order to make your work reproducible, you must at least use a script file.

file -> new file

Warning : it is not because you have saved your project (.Rproj) that your data and your programs (.R,.Rmd...) are also saved (.csv,.xlSx...). **You have to save them independently!!**

R language basis

The operators

- math : +, -, *, /, %/%,
%*%, t(), ^, log(), exp()...
- assignation : <- = ->
- logical: < > <= >= == != %in%

The data types

- Character
- numeric :
 - _ logical (0 FALSE 1 TRUE)
 - _ factor
 - _ integer
 - _ double
 - _ etc...

The data objects

- vector
- matrix
- array
- list
- data frame

The functions

There are commands that transform or produce new elements from a R object or a part of a R object.

They are shaped as follow : **name_of_the_function(main arguments, options)**

`sum(x,na.rm==TRUE)` make the sum of the element of vector named x. The option na.rm==TRUE excludes from the sum the NA values.

```
x<-c(1:6,NA,c(5:9)-2)
x

## [1] 1 2 3 4 5 6 NA 3 4 5 6 7

sum(x)

## [1] NA

sum(x,na.rm=TRUE)

## [1] 46
```

Find some help about a function

Use the help window or just put ? in front the function name.

```
?t.test()  
  
## démarrage du serveur d'aide httpd ... fini
```

The syntax

R base has a stacked syntax.

```
paste(mean(x[x<5],na.rm=TRUE), 'on average')  
## [1] "2.83333333333333 on average"
```

The syntax

R base has a stacked syntax.

```
x

## [1] 1 2 3 4 5 6 NA 3 4 5 6 7

x<5

## [1] TRUE TRUE TRUE TRUE FALSE FALSE NA TRUE TRUE FALSE FALSE FALSE

x[x<5]

## [1] 1 2 3 4 NA 3 4

mean(x[x<5],na.rm=TRUE)

## [1] 2.833333

paste(mean(x[x<5],na.rm=TRUE),'on average')

## [1] "2.83333333333333 on average"
```

Install and use packages

With R you have access to a wide panel of functions developped by the community. You can download packages of new functions from online repositories (like the CRAN, or github...).

Let's download the tidyverse package which includes a lot of usefull complementary functionality.

```
install.packages("tidyverse")
```

Install and use packages

Once a package downloaded, you have to call it in your environment in order to use the included functions.

```
library(tidyverse)

## — Attaching packages tidyverse 1.3.2 —
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.4
## ✓ tibble  3.1.8     ✓ dplyr   1.0.10
## ✓ tidyr   1.2.0     ✓ stringr 1.4.1
## ✓ readr   2.1.2     ✓ forcats 0.5.2
## — Conflicts tidyverse_conflicts() —
## ✗ dplyr::filter() masks stats::filter()
## ✗ dplyr::lag()   masks stats::lag()
```

Install and use packages

The tidyverse provides (among many other things) a usefull programing operator the simplify R syntax : the pipe operator `%>%` (keyboard shortcut 'ctrl' 'shift' 'm').

```
x<-1:5
y<-8:4
data.frame(x,y) %>% summarise(m_x=mean(x),m_y=mean(y))

##   m_x m_y
## 1   3   6
```

Install and use packages

It is simpler than :

```
lapply(data.frame(x,y), mean)
```

```
## $x  
## [1] 3  
##  
## $y  
## [1] 6
```

```
sapply(data.frame(x,y), mean)
```

```
## x y  
## 3 6
```

```
apply(data.frame(x,y), 2, mean)
```

```
## x y  
## 3 6
```

The home made functions

You can create your own function using the following syntax.

```
function_name<-function(arguements,...){  
  one program that do the job!!  
}
```

The home made functions

Let's build a function that compute the actual value of a given amount of money in n compounding periods when the opportunity interest rate is i .

The home made functions

Let's do the math.

$$Vf = VA \times (1 + i)^n$$

$$VA = \frac{VF}{(1 + i)^n}$$

The home made functions

Let's do the program.

```
val_act<-function(vf,i,d){  
  va<-vf*(1/(1+i)^d)  
  return(va)  
}
```

Let's test it.

```
val_act(1159.2741,0.03,5)  
  
## [1] 1000
```

The home made functions

Create a function that computes the annual interest rate allowing to obtain the final value from the initial one in n periods. (5 minutes)

The home made functions

Let's do the math.

$$Vf = VA \times (1 + i)^n$$

$$\frac{Vf}{VA} = (1 + i)^n$$

$$\left(\frac{Vf}{VA}\right)^{1/n} = 1 + i$$

$$i = \left(\frac{Vf}{VA}\right)^{1/n} - 1$$

The home made functions

Let's do the programing.

```
interet<-function(vf,va,n){  
  i<-(vf/va)^(1/n)-1  
  return(i)  
}
```

Let's test it.

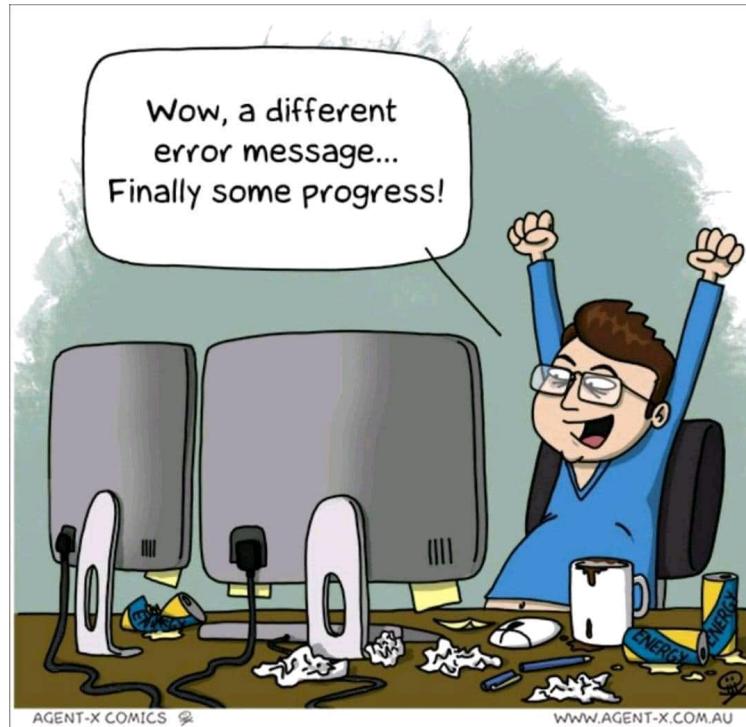
```
interet(1159.2741,1000,5)  
## [1] 0.03
```

More on R language

You can watch (later) a 40 minutes video that presents more extensively the points that we have just review.
You can find it at the following address:

<https://youtu.be/hGffaVR4Dtk>

short break



WRDS data base

Access to WRDS

You start from your Student ENT. You select "toutes les applications" and on the part dedicated to 'Bibliothèque' you click on 'Ressources électroniques'.

ent -> toutes les applications -> Bibliothèque -> ressources électroniques

Once the 'ressources électroniques' page opens go to 'Accueil' and section the 'gestion' discipline. You can find the link with WRDS at the bottom of the current page.

Accueil -> gestion -> wrds

Access to WRDS

The screenshot shows the 'Connect to WRDS' page. At the top, there's a navigation bar with links like 'FR', 'Chercheurs', 'Base de données', 'Cours', 'Conférences', 'Login', 'Méthodologies', 'Réviews', 'Réseaux sociaux', 'Ses cours', 'Opportunités', and 'Autre'. Below the bar, the Wharton logo and 'wrds' are displayed. The main content area has two columns: 'Connect to WRDS' on the left and 'Welcome to WRDS!' on the right. The 'Connect to WRDS' section asks for an email address and includes a CAPTCHA field. The 'Welcome to WRDS!' section provides a brief overview of the service. A footer at the bottom contains links for 'Analytics', 'Data', 'Classroom', 'About', 'Videos', and 'Demo'.

Access to WRDS

The screenshot shows the Wharton Research Data Services (WRDS) website. At the top, there is a navigation bar with links for Home, Analytics, Data, Classroom, About, Videos, and Demo. Below the navigation bar, there is a section titled "Email Sent" which contains a message about a password reset email. To the right of this, there is a "Welcome to WRDS!" section with a brief description of the platform's purpose and a "Connect with us on Facebook!" link.

Access to WRDS

The screenshot shows a webmail interface with the Université de Lille logo at the top. The menu includes Mail, Contacts, Calendrier, Tâches, and Préférences. The inbox contains several emails from 'WRDS'. One email from 'WRDS Day Pass' is highlighted, with the subject 'Your WRDS Day Pass - Dear WRDS User, Thank you for requesting a Day Pass!' and a preview of '29 new R posts from the past week.' Below the inbox, a calendar for August 2022 is visible, showing dates from 1 to 31.

Your WRDS Day Pass

Expéditeur : (WRDS)

A : Ludovic Vigneron

Les images externes ne seront pas affichées. Afficher les images
Requête afficher les images envoyées par wrds-support@wharton.upenn.edu

Clear WRDS User.

Thank you for requesting a Day Pass for WRDS access. Your connection string is

<https://ressources-electroniques.univ-lille.fr/> or <https://wrds.wharton.upenn.edu/>

The above link is valid for a day pass to WRDS, enabling you to access the data to which your institution subscribes via the WRDS platform. Clicking on the link will take you directly to the WRDS website.

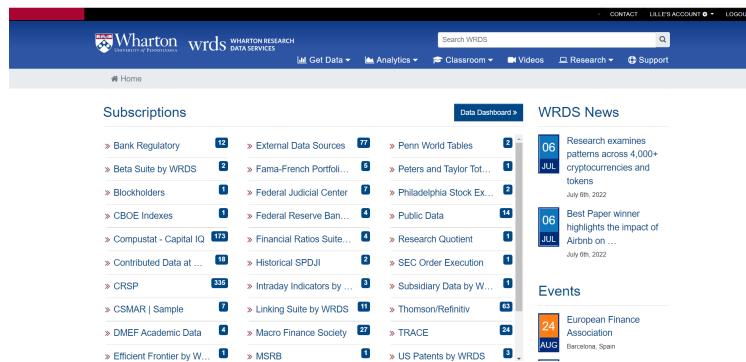
Please make sure to enter your email address on the query page of any job you run which may finish after this pass has expired. Obtaining any results which become available after this pass has expired will then simply require a new day pass.

Please note that your use of the WRDS website and other services offered by WRDS is contingent upon your acceptance of and agreement to the WRDS Terms of Use, available [here](#).

<https://wrds.wharton.upenn.edu/en/>

If you have any issues or questions, please reply directly to this email (wrds-support@wharton.upenn.edu).

Access to WRDS



Subscriptions

- > Bank Regulatory 12
- > Beta Suite by WRDS 2
- > Blockholders 1
- > CBOE Indexes 1
- > Compustat - Capital IQ 173
- > Contributed Data at ... 18
- > CRSP 336
- > CSMAR | Sample 7
- > DMEF Academic Data 4
- > Efficient Frontier by W... 1
- > External Data Sources 77
- > Fama-French Portfoli... 3
- > Federal Judicial Center 7
- > Financial Ratios Suite... 4
- > Historical SPDI 2
- > Intraday Indicators by ... 3
- > Linking Suite by WRDS 11
- > Macro Finance Society 27
- > MSRB 1
- > Peters and Taylor Tot... 1
- > Philadelphia Stock Ex... 2
- > Public Data 14
- > Research Quotient 1
- > SEC Order Execution 1
- > Subsidiary Data by W... 1
- > Thomson/Refinitiv 63
- > TRACE 24
- > US Patents by WRDS 3

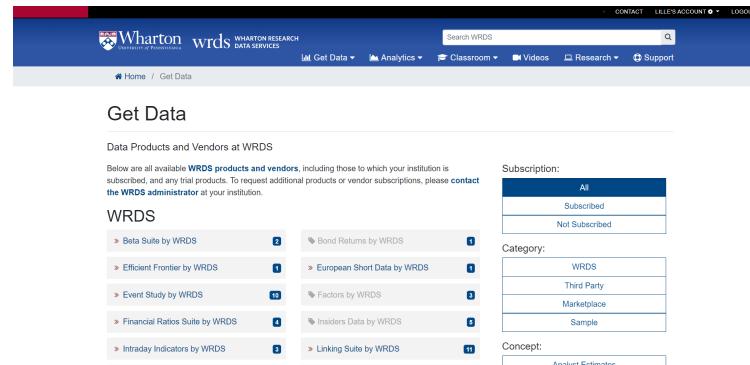
WRDS News

- 06 JUL** Research examines patterns across 4,000+ cryptocurrencies and tokens
July 6th, 2022
- 06 JUL** Best Paper winner highlights the impact of Airbnb on ...
July 6th, 2022

Events

- 24 AUG** European Finance Association
Barcelona, Spain

Access to WRDS



The screenshot displays the 'Get Data' section of the WRDS website. At the top, there's a navigation bar with links for 'Home', 'Get Data', 'Analytics', 'Classroom', 'Videos', 'Research', and 'Support'. Below the navigation is a search bar labeled 'Search WRDS'.

The main content area is titled 'Get Data' and features a heading 'Data Products and Vendors at WRDS'. It includes a note about available products and vendors, a contact link for additional subscriptions, and a 'WRDS' section listing various products like 'Beta Suite by WRDS', 'Efficient Frontier by WRDS', etc.

On the right side, there are three filter sections: 'Subscription' (with options All, Subscribed, Not Subscribed), 'Category' (with options WRDS, Third Party, Marketplace, Sample), and 'Concept' (with one option shown: Analyst Estimates).

Access to WRDS

You only have access to ressources marked by the plain blue links.

If you spend too much time without doing anything, you will be disconnected.

Your access is can be activated using the email link during a day long.

What WRDS is?

It is a data base aggregator provided by the Wharton school of business of the university of Pennsylvania.

It allows to access multiple resources in one place using a central data extraction interface.

It also provides some data treatment services (compute event studies, efficient portfolio, Beta...).

What the main ressources available are ?

- Compustat : accounting data about listed firms
- CRPS : share prices
- Thomson/Refinitiv (WRDS-Reuter Dealscan) : syndicated loan deals
- Trace : bonds

etc...

What the main ressources available are ?

Each database has its own individuals' (company or security or CEO...) identifier.

One of the first difficulties that you will have to deal with using those data base will be to manage those identifiers in order to merge the data accurately.

WRDS helps providing tables with merging keys (most of times).

Extracting/importing data

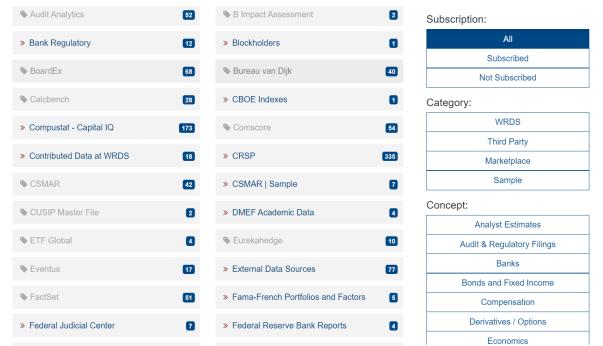
Extracting data from CRSP

The Center for Research in Security Prices, LLC (CRSP) maintains the most comprehensive collection of security price, return, and volume data for the NYSE, AMEX and NASDAQ stock markets.

It is the gold standard data base on stocks in empirical finance (in US).

Most of the more important research papers published in top scientific journals (JF,JFE,JFQA...) have been produced using CRSP.

Extracting data from CRSP



The screenshot shows a search interface for extracting data from CRSP. On the left, there is a list of data sources with their counts:

- Audit Analytics: 82
- Bank Regulatory: 12
- BoardEx: 68
- Calcbench: 28
- Compustat - Capital IQ: 172
- Contributed Data at WRDS: 19
- CSMAR: 42
- CUSIP Master File: 2
- ETF Global: 6
- Eventus: 17
- FadSet: 11
- Federal Judicial Center: 7
- B Impact Assessment: 3
- Blockholders: 3
- Bureau van Dijk: 43
- CBOE Indexes: 3
- Comscore: 24
- CRSP: 333
- CSMAR | Sample: 3
- DMEF Academic Data: 3
- Eurekahedge: 10
- External Data Sources: 7
- Fama-French Portfolios and Factors: 3
- Federal Reserve Bank Reports: 3

On the right, there are three filter panels:

- Subscription:** All, Subscribed, Not Subscribed
- Category:** WRDS, Third Party, Marketplace, Sample
- Concept:** Analyst Estimates, Audit & Regulatory Filings, Banks, Bonds and Fixed Income, Compensation, Derivatives / Options, Economics

Extracting data from CRSP

The screenshot shows the Wharton WRDS homepage. A modal window titled "New CRSP Format: Details available" is open, providing information about the enhanced Flat File Format 2.0 (CIZ). The sidebar on the right contains links to "CRSP Basics", "Knowledge Base", and categories like "Stock" and "Index and Deciles".

New CRSP Format: Details available

Details are now available about the enhanced Flat File Format 2.0 (CIZ)

CRSP has introduced a new Flat File Format 2.0 (CIZ) for the CRSP US 1925 and 1962 Stock and Stock with Index Databases. All subscribers – monthly, quarterly, and annual – will receive this premier data cut when it becomes available in early August. WRDS will maintain both current and new formats for two years. Please see the [full announcement here](#).

Posted June 10th, 2022

Don't show this again ■

CRSP

Filter by Concept / Region (show)

CRSP Basics

Wharton Research Data Services

WTDS

The Online Facet for Business Research

Knowledge Base

Stock

Index and Deciles

Extracting data from CRSP

The screenshot shows the CRSP website's homepage. At the top, there is a navigation bar with links for 'Get Data', 'Analytics', 'Classroom', 'Videos', 'Research', and 'Support'. Below the navigation bar, the main content area has a title 'Stock / Security Files'. A prominent blue modal window is open, titled 'New CRSP Format: Details available'. It contains text about the introduction of the enhanced Flat File Format 2.0 (CIZ) for CRSP US 1925 and 1962 Stock and Stock with Index Databases. It mentions that all subscribers – monthly, quarterly, and annual – will receive the premiere data cut when it becomes available in early August. WRDS will maintain both current and new formats for two years. There is a link to an announcement. A 'Don't show this again' button is at the bottom right of the modal. To the right of the modal, there is a video player showing a presentation slide with the text 'CRSP - Basics' and 'WRDS'. Below the video player, there is a 'Knowledge Base' section with several links: 'Stock', 'Index and Deciles', 'Identifying ETFs', 'Stock Header Info', 'U.S. Daily Event Study: Upload your own ...', 'Daily Stock File', 'Stock Market Indexes', and 'Monthly Stock File'. At the bottom, there is a 'Manuals and Overviews' section.

Extracting data from CRSP

[Home](#) / [Get Data](#) / [CRSP](#) / [Annual Update](#) / [Stock / Security Files](#) / [Daily Stock File](#)

CRSP Daily Stock

Query Form Variable Descriptions Manuals and Overviews Knowledge Base Data Preview

CRSP Daily Stock

You have 10 saved queries for this Query Form.

Step 1: Choose your date range.
 Date Variable:
 to

Step 2: Apply your company codes.
 What format are your company codes?
 TICKER PERMNO PERMCO CUSIP NOCUSIP HSICCD SICCD

Select an option for entering your company codes:
 Company Codes Code List Name
 Please enter company codes separated by a space.

Save this code list to Saved Codes

CRSP CENTER FOR RESEARCH IN SECURITY PRICES, LLC
 An Affiliate of the University of Chicago Booth School of Business

More About This Vendor »

CRSP »
 Annual Update »
 Stock / Security Files (hide)

- » Daily Stock File
- » Stock Market Indexes
- » Monthly Stock File
- » Stock Header Info
- » U.S. Daily Event Study: Upload your own ...


 CRSP - Basics
 Wharton Research Data Services

Extracting data from CRSP

Step 3: Choose query variables.
How does this work?

Q Search All **CRSP** Identifying Information Time Series Information Share Info

Select All Selected Clear All (0)

Search All

- Cusp (cusp)
- Nousip (nousip)
- Company Name (comnam)
- Ticker (ticker)
- CRSP Permanent Company Number (permco)
- Share Code (shrcd)
- Share Class (shrccls)
- Nasdaq Issue Number (issuno)
- Exchange Code (exchcd)
- Header Exchange Code (hewxcn)

CRSP CENTER FOR RESEARCH IN SECURITY PRICES, LLC
An Affiliat of the University of Chicago Booth School of Business

More About This Vendor >

CRSP >
Annual Update >
Stock / Security Files (hide)

- > Daily Stock File
- > Stock Market Indexes
- > Monthly Stock File
- > Stock Header Info
- > U.S. Daily Event Study: Upload your ow...

CRSP - Basics Wharton Research Data Services WINDS The Global Standard for Financial Research

Extracting data from CRSP

Step 4: Select query output.

How does this work?

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

- Uncompressed
- zip (*.zip)
- gzip (*.gz)

Date Format

- YYMMDDn8 (e.g. 19840725)
- DATE9 (e.g. 25JUL1984)
- DDMMYY6 (e.g. 250784)
- MMDDYY10 (e.g. 07251984)
- DDMMYY10 (e.g. 25/07/1984)
- YYMMDDs10 (e.g. 1984/07/25)

Manuals and Overviews

- Stock
- Index and Deciles
- Identifying ETFs
- Overview of CRSP U.S. Stock Database
- WRDS Overview of CRSP/COMPUSTAT Merged (CCM)
- Stocks and Indices

Additional Tools



E-Mail Address (Optional)

Custom Field (Optional)

Save This Query (Optional)

Notes on this Query (Optional)

Extracting data from CRSP

Mission 1:

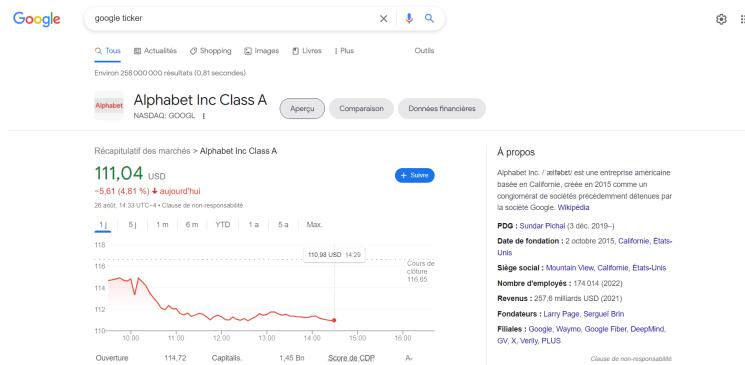
Find the ticker of a stock of your choice and put it in the chat window (company name followed by the corresponding actual ticker).

Extracting data from CRSP

Let's take Google.

A quick research on it provides us its ticker.

Extracting data from CRSP



Extracting data from CRSP

The screenshot shows the CRSP website interface. At the top, there is a search bar with placeholder text "Type here to search queries for this query form...". Below it, a section titled "Step 1: Choose your date range." contains a date range selector with "Date Variable:" dropdowns set to "2000-01-01" and "2022-03-31". A note "Maximum allowed date: 2022-03-31" is displayed above the second dropdown. To the right, a sidebar titled "CRSP CENTER FOR RESEARCH IN SECURITY PRICES, LLC" includes links like "More About This Vendor >", "CRSP >", "Annual Update >", and "Stock / Security Files (HTML)" with sub-links for "Daily Stock File", "Stock Market Indexes", "Monthly Stock File", "Stock Header Info", and "U.S. Daily Event Study: Upload your ow...". A video player at the bottom left shows a thumbnail for "CRSP - Basics" by Wharton Research Data Services.

Extracting data from CRSP

Step 3: Choose query variables.
How does this work?

The screenshot shows the CRSP query builder interface. At the top, there are tabs for "Search All" (selected), "Identifying Information", "Time Series Information", and "Share". Below the tabs, there's a search bar and a "Select" button with a checked "All" checkbox. A "Selected" section lists 61 variables, including Cusip, Nicusp, Company Name, Ticker, CRSP Permanent Company Number, Share Code, Share Class, Nasdaq Issue Number, Exchange Code, and Header Exchange Code. There are also sections for "Stock / Security Files" (Daily Stock File, Stock Market Indexes, Monthly Stock File, Stock Header Info, U.S. Daily Event Study) and a video player for "CRSP - Basics: Wharton Research Data Services WIDS".

Extracting data from CRSP

Step 4: Select query output.

How does this work?

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 <input type="radio"/> fixed-width text (*.txt) <input checked="" type="radio"/> comma-delimited text (*.csv) <input type="radio"/> Excel spreadsheet (*.xlsx) <input type="radio"/> tab-delimited text (*.txt) <input type="radio"/> HTML table (*.htm) <input type="radio"/> SAS Windows_32 dataset (*.sas7bdat) <input type="radio"/> SAS Windows_64 dataset (*.sas7bdat) <input type="radio"/> SAS Solaris_64 dataset (*.sas7bdat) <input type="radio"/> dBase file (*.dbf) <input type="radio"/> STATA file (*.dta) <input type="radio"/> SPSS file (*.sav)	Compression Type <input checked="" type="radio"/> Uncompressed <input type="radio"/> zip (*.zip) <input type="radio"/> gzip (*.gz)	Date Format <input checked="" type="radio"/> YYMMDDhh (e.g. 19840725) <input type="radio"/> DATES (e.g. 25JUL1984) <input type="radio"/> DDMMYY6 (e.g. 250784) <input type="radio"/> MMDDYY10 (e.g. 07251984) <input type="radio"/> YYMMDDs10 (e.g. 1984/07/25)
E-Mail Address (Optional) <input type="text" value="E-mail"/> <input type="button" value="Edit Preferences"/>		
Custom Field (Optional) <input type="text" value="Custom Field"/> <input type="button" value="Edit"/>		
Save This Query (Optional) <input type="checkbox"/> Saved Query Name <input type="button" value="Edit"/>		
Notes on This Query (Optional) <input type="text" value="Saved Query Notes"/>		
<input type="button" value="Submit Form"/>		

CRSP CENTER FOR RESEARCH
IN SECURITY PRICES, LLC
An Offshoot of the University of Chicago Booth School of Business

[More About This Vendor >](#)

CRSP [»](#)

Annual Update [»](#)

Stock / Security Files [\(hide\)](#)

[» Daily Stock File](#)

[» Stock Market Indexes](#)

[» Monthly Stock File](#)

[» Stock Header Info](#)

[» U.S. Daily Event Study: Upload your ow...](#)



Extracting data from CRSP

The screenshot shows a web page for the CRSP Daily Stock service. At the top, there is a navigation bar with links to Home, Get Data, CRSP, Annual Update, Stock / Security Files, and Daily Stock File. Below this, a green success message says "Query 6004167 successfully submitted." The main content area has a title "CRSP Daily Stock". Below it, a blue box contains the heading "New CRSP Format: Details available". The text explains that CRSP has introduced a new Flat File Format 2.0 (CIZ) for the CRSP US 1925 and 1962 Stock and Stock with Index Databases. It mentions that all subscribers – monthly, quarterly, and annual – will receive this premiere data cut when it becomes available in early August. WRS will maintain both current and new formats for two years. A link to the announcement is provided. The date "Posted June 10th, 2022" is at the bottom left of this box. On the right side, there is a sidebar titled "More About This Vendor" which includes sections for CRSP, Annual Update, and Stock / Security Files. Under Stock / Security Files, "Daily Stock File" is highlighted with a blue border. Other options like "Stock Market Indexes", "Monthly Stock File", "Stock Header Info", and "U.S. Daily Event Study: Upload your own ..." are also listed.

Extracting data from CRSP

Query 6064167, for crsp_a_stock

Status:	Success	Run		
Product:	crsp_a_stock			
Results:	Result Size: 1.2 MB	Result Count: 4,436 Rows		
	Output Files: • ~wrds_query_output/danvxe@cruvhcppb.csv			
Timing:	Elapsed Time: 10 seconds	Submitted: 2022-08-26 14:41		
	Work Begun: 2022-08-26 14:41	Work Finished: 2022-08-26 14:41		
Input Parameters:	Toggle Input Parameters			
Data Retention				
Results data will be accessible for 2 days, until 2022-08-28, 14:41				
Notice				
Your use of WRDS and this data extract must comply with the WRDS Terms of Use.				
There may be additional usage restrictions that are governed by your institution's licensing of specific databases. If you have any questions about data licensing and appropriate usage, please contact WRDS using the Support form.				

Extracting data from CRSP

Home / Your Account / Queries / Query 6064167, for crsp_a_stock

Query 6064167, for crsp_a_stock

Status:	Success	Run
Product:	crsp_a_stock	
Results:	Result Size: 1.2 MiB	Result Count: 4,436 Rows
	Output Files: <ul style="list-style-type: none"> • ~/wrds_query_output/ydanvxe@cruvhgb.csv 	
Timing:	Elapsed Time: 10 seconds	Submitted: 2022-08-26 14:41
	Work Begun: 2022-08-26 14:41	Work Finished: 2022-08-26 14:41

Data Retention

Results data will be accessible for 2 days, until 2022-08-28, 14:41

Notice

Your use of WRDS and this data extract must comply with the WRDS Terms of Use.

There may be additional usage restrictions that are governed by your institution's licensing of specific databases. If you have any questions about data licensing and appropriate usage, please contact WRDS using the Support form.

[ydanvxe@cruvhgb.csv](#)

Extracting data from CRSP

Copy the downloaded file in your working directory.

Once done import your data in R.

```
google <- read_csv("google.csv")  
  
## Warning: One or more parsing issues, see `problems()` for details  
  
## Rows: 4436 Columns: 63  
## — Column specification ——————  
## Delimiter: ","  
## chr (13): NCUSIP, TICKER, COMNAM, SHRCLS, TSYMBOL, PRIMEXCH, TRDSTAT, SECSTA...  
## dbl (39): PERMNO, date, NAMEENDT, SHRCD, EXCHCD, SICCD, NAICS, PERMCO, ISSUN...  
## lgl (11): DCLRDT, DLPDT, NEXTDT, PAYDT, RCRDDT, DISTCD, DIVAMT, FACPR, FACSH...  
##  
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Extracting data from CRSP

let's see what do we have. head(google)

Show entries Search:

	PERMNO	date	NAMEENDT	SHRCD	EXCHCD	SICCD	NCUSIP	TICKER	COMM
1	90319	20040819	20050818	11	3	7375	38259P50	GOOG	GOOG INC
2	90319	20040820		11	3	7375	38259P50	GOOG	GOOG INC
3	90319	20040823		11	3	7375	38259P50	GOOG	GOOG INC
4	90319	20040824		11	3	7375	38259P50	GOOG	GOOG INC
5	90319	20040825		11	3	7375	38259P50	GOOG	GOOG INC
6	90319	20040826		11	3	7375	38259P50	GOOG	GOOG INC

Showing 1 to 6 of 6 entries Previous Next

Extracting data from CRSP

```
unique(google$PERMNO)
## [1] 90319

unique(google$PERMCO)
## [1] 45483

unique(google$TICKER)
## [1] "GOOG"   "GOOGL"

unique(google$CUSIP)
## [1] "02079K30"

unique(google$COMNAM)
## [1] "GOOGLE INC"    "ALPHABET INC"
```

Extracting data from CRSP

PERMNO (security) and PERMCO (company) are unique identifier given once by crsp respectivily to a security and a company they never change and are never reaffected.

There is no survivor bias in CRSP.

Extracting data from CRSP

Put the date variable on a date format.

```
str(google$date)  
## num [1:4436] 2e+07 2e+07 2e+07 2e+07 2e+07 ...
```

Extracting data from CRSP

You have to install the lubridate package to manage easily date format variables.

```
install.packages(lubridate)
```

Once the new package installed, call it in your working environment.

```
library(lubridate)

## 
## Attachement du package : 'lubridate'

## Les objets suivants sont masqués depuis 'package:base':
## 
##     date, intersect, setdiff, union
```

Extracting data from CRSP

So you can use ymd() function to format the date variable.

```
google$date<-ymd(google$date)
```

```
str(google$date)
```

```
##  Date[1:4436], format: "2004-08-19" "2004-08-20" "2004-08-23" "2004-08-24" "2004-08-25" ...
```

Extracting data from CRSP

let's see what do we have now. head(google)

Show entries Search:

	PERMNO	date	NAMEENDT	SHRCD	EXCHCD	SICCD	NCUSIP	TICKER	COMNA
1	90319	2004-08-19	20050818	11	3	7375	38259P50	GOOG	GOOGLE INC
2	90319	2004-08-20		11	3	7375	38259P50	GOOG	GOOGLE INC
3	90319	2004-08-23		11	3	7375	38259P50	GOOG	GOOGLE INC
4	90319	2004-08-24		11	3	7375	38259P50	GOOG	GOOGLE INC
5	90319	2004-08-25		11	3	7375	38259P50	GOOG	GOOGLE INC
6	90319	2004-08-26		11	3	7375	38259P50	GOOG	GOOGLE INC

Showing 1 to 6 of 6 entries ◀ ▶

Previous 1 Next

Extracting data from CRSP

Let's check if we have duplicated observations.

```
google<-google %>% arrange(PERMNO,date) %>%
  group_by(PERMNO,date) %>%
  mutate(dup=n()) %>% ungroup()
table(google$dup)

## 
##      1
## 4436
```

Extracting data from CRSP

There is no duplicated observations. We can drop the dup variable.

```
google <- google %>% select(-dup)
```

Extracting data from CRSP

There is too much information in the actual data base. Let reduce it to the ones we need.

```
google<-google %>% select(PERMNO, COMNAM, date, PRC, RET, RETX, DIVAMT, DISTCD,  
DCLRDT, RCRDDT, PAYDT)
```

Extracting data from CRSP

When did google start to be listed?

```
min(google$date)
```

```
## [1] "2004-08-19"
```

Created in september 1998.

Did Google (Alphabet) pay some dividends since its listing?

```
unique(google$DIVAMT)
```

```
## [1] NA
```

Extracting data from CRSP

No dividence, we can also drop the related information.

```
google<-google %>% select(PERMNO, COMNAM, date, PRC, RET, RETX)
```

Extracting data from CRSP

Let's check if RET and RETX are always equal.

```
sum(google$RET==google$RETX)
```

```
## [1] 4436
```

We can drop RETX.

```
google<-google %>% select(-RETX)
```

Extracting data from CRSP

How are returns established in CRSP?

Simple return

$$r_t = \frac{P_1 - P_0}{P_0} = \frac{P_1}{P_0} - 1$$

```
google$PRC[2]/google$PRC[1]-1
```

```
## [1] 0.07948373
```

```
google$RET[2]
```

```
## [1] "0.079484"
```

CRSP provides simple returns.

Extracting data from CRSP

Returns (RET) are in character format. Lets transform the variable to make it numeric.

```
google$RET[1:10]
```

```
## [1] "C"          "0.079484"   "0.010064"   "-0.041408"  "0.010775"   "0.018019"  
## [7] "-0.016310" "-0.039001" "0.003529"  "-0.020709"
```

```
google$RET<-as.numeric(google$RET)
```

```
## Warning: NAs introduits lors de la conversion automatique
```

```
google$RET[1:10]
```

```
## [1] NA  0.079484  0.010064 -0.041408  0.010775  0.018019 -0.016310  
## [8] -0.039001  0.003529 -0.020709
```

Extracting data from CRSP

Let's compute our own simple returns for the entire data base and compare with the precomputed returns

```
google <- google %>% arrange(date) %>%
  mutate(ret_s=round(PRC/lag(PRC)-1,digits=6),
        test=RET==ret_s)
google

## # A tibble: 4,436 × 7
##   PERMNO COMNAM     date      PRC      RET    ret_s test
##   <dbl> <chr>     <date>    <dbl>    <dbl>    <dbl> <lgl>
## 1 90319 GOOGLE INC 2004-08-19 100.    NA       NA     NA
## 2 90319 GOOGLE INC 2004-08-20 108.   0.0795   0.0795 TRUE
## 3 90319 GOOGLE INC 2004-08-23 109.   0.0101   0.0101 TRUE
## 4 90319 GOOGLE INC 2004-08-24 105.  -0.0414  -0.0414 TRUE
## 5 90319 GOOGLE INC 2004-08-25 106.   0.0108   0.0108 TRUE
## 6 90319 GOOGLE INC 2004-08-26 108.   0.0180   0.0180 TRUE
## 7 90319 GOOGLE INC 2004-08-27 106.  -0.0163  -0.0163 TRUE
## 8 90319 GOOGLE INC 2004-08-30 102.  -0.0390  -0.0390 TRUE
## 9 90319 GOOGLE INC 2004-08-31 102.  0.00353  0.00353 TRUE
## 10 90319 GOOGLE INC 2004-09-01 100.  -0.0207  -0.0207 TRUE
## # ... with 4,426 more rows
```

Extracting data from CRSP

Let see if the two variables always match.

```
sum(google$test,na.rm=TRUE)
```

```
## [1] 4408
```

It is not the case.

```
google %>% filter(test==FALSE)
```

```
## # A tibble: 27 × 7
##   PERMNO COMNAM     date      PRC      RET    ret_s test
##   <dbl> <chr>     <date>    <dbl>    <dbl>    <dbl> <lgl>
## 1 90319 GOOGLE INC 2004-10-21 149.  0.0633  0.0633 FALSE
## 2 90319 GOOGLE INC 2005-01-11 194. -0.00779 -0.00779 FALSE
## 3 90319 GOOGLE INC 2005-02-01 192. -0.0190  -0.0190 FALSE
## 4 90319 GOOGLE INC 2005-02-16 198.  0.0163  0.0163 FALSE
## 5 90319 GOOGLE INC 2005-10-28 358.  0.0145  0.0145 FALSE
## 6 90319 GOOGLE INC 2006-12-04 485.  0.00842 0.00842 FALSE
## 7 90319 GOOGLE INC 2007-04-13 466. -0.00235 -0.00235 FALSE
## 8 90319 GOOGLE INC 2007-11-19 626. -0.0123  -0.0123 FALSE
## 9 90319 GOOGLE INC 2008-07-28 477. -0.0302  -0.0302 FALSE
## 10 90319 GOOGLE INC 2008-10-31 359. -0.000918 -0.000917 FALSE
## # ... with 17 more rows
```

It is just marginal rounding problems.

Extracting data from CRSP

Let consider the continuous returns starting with some math.

$$vf = va \cdot e^{r \cdot n}$$

$$\frac{vf}{va} = e^{r \cdot n}$$

$$\ln\left(\frac{vf}{va}\right) = \ln(e^{r \cdot n})$$

$$r \cdot n = \ln\left(\frac{vf}{va}\right)$$

$$r = \frac{1}{n} \cdot \ln\left(\frac{vf}{va}\right)$$

here we just consider one period (so n=1).

$$r_t = \ln\left(\frac{P_1}{P_0}\right)$$

Extracting data from CRSP

Let compute the continuous daily returns for Google.

```
google<-google %>% mutate(ret_c=log(PRC/lag(PRC)))
```

Extracting data from CRSP

Let compute the final value after one period using continuous returns.

```
google$PRC[1]*exp(google$ret_c[2])  
  
## [1] 108.31
```

Extracting data from CRSP

Let consider now multiperiods returns.

Compute the return between date 1 [2004-08-19] and date 4 [2004-08-24].

```
google$PRC[1]
```

```
## [1] 100.335
```

```
google$PRC[4]
```

```
## [1] 104.87
```

Extracting data from CRSP

Let consider now multiperiods returns.

simple return

```
google$PRC[4]/google$PRC[1]-1  
  
## [1] 0.04519858  
  
google$PRC[1]*(1+(google$PRC[4]/google$PRC[1]-1))  
  
## [1] 104.87
```

continuous returns

```
log(google$PRC[4]/google$PRC[1])  
  
## [1] 0.0442069  
  
google$PRC[1]*exp(log(google$PRC[4]/google$PRC[1]))  
  
## [1] 104.87
```

Extracting data from CRSP

simple returns intra period

```
google$PRC[1]*(1+google$ret_s[2])*(1+google$ret_s[3])*(1+google$ret_s[4])  
## [1] 104.87
```

using cumprod()

```
google$PRC[1]*cumprod(1+google$ret_s[2:4])[3]  
## [1] 104.87
```

Extracting data from CRSP

simple returns intra period

using geometric average

```
google$PRC[1]*(prod(1+google$ret_s[2:4])^(1/3))^3  
## [1] 104.87
```

using a home made function for geometric average.

```
geomet_moy<-function(x){  
  y<-prod(1+x)^(1/length(x))  
  return(y)  
}
```

```
google$PRC[1]*geomet_moy(google$ret_s[2:4])^3  
## [1] 104.87
```

Extracting data from CRSP

continuous returns intra period

```
google$PRC[1]*exp(sum(google$ret_c[2:4]))  
## [1] 104.87  
  
google$PRC[1]*exp(mean(google$ret_c[2:4])*3)  
## [1] 104.87
```

Extracting data from CRSP

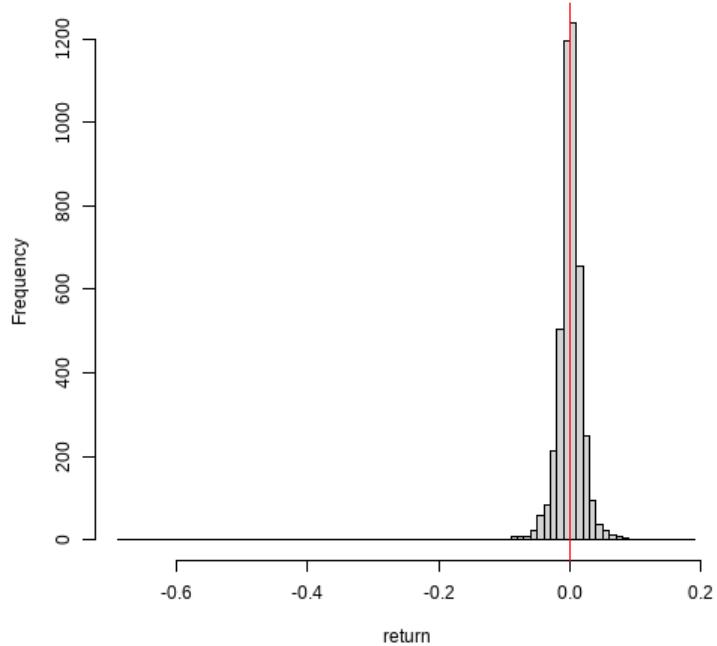
Are Google return normally distributed?

Extracting data from CRSP

Let see it drawing a histogram.

```
hist(google$ret_c,nclass = 100,  
main="histogram of google stock return",xlab='return')  
abline(v=mean(google$ret_c,na.rm=TRUE),col='red')
```

histogram of google stock return



Extracting data from CRSP

To be sure, let perform a Shapiro-Wilk test.

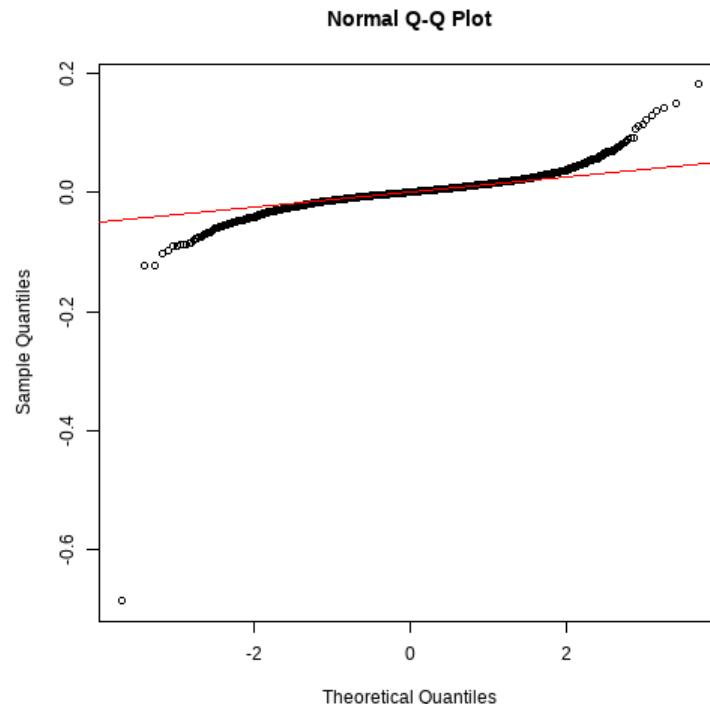
```
shapiro.test(google$ret_c)

##
##      Shapiro-Wilk normality test
##
## data: google$ret_c
## W = 0.74424, p-value < 2.2e-16
```

Extracting data from CRSP

We also can examine the normality with a quantile-quantile plot (QQplot).

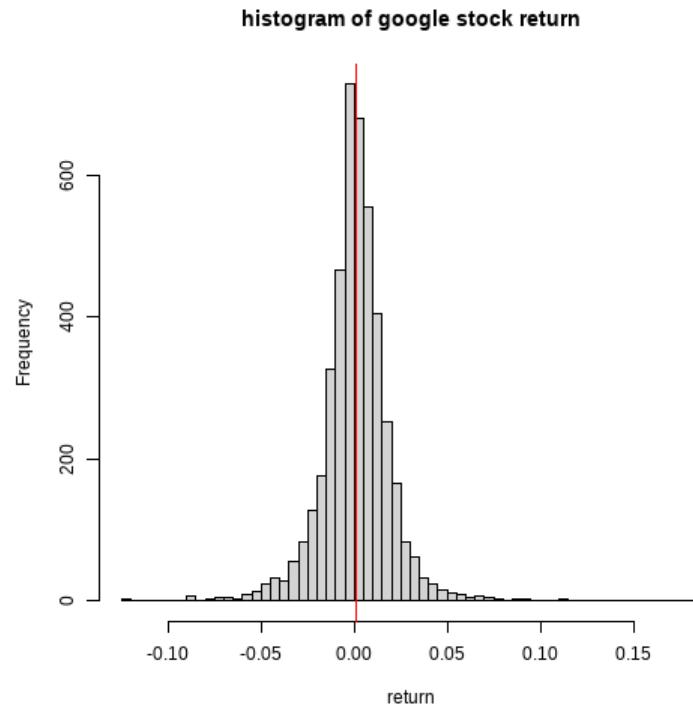
```
qqnorm(google$ret_c)
qqline(google$ret_c,col='red')
```



Extracting data from CRSP

Let drop the more extrem observation.

```
hist(google$ret_c[which(google$ret_c>-0.68)],nclass = 100,  
      main="histogram of google stock return",xlab='return')  
abline(v=mean(google$ret_c[which(google$ret_c>-0.68)]),  
       na.rm=TRUE),col='red')
```



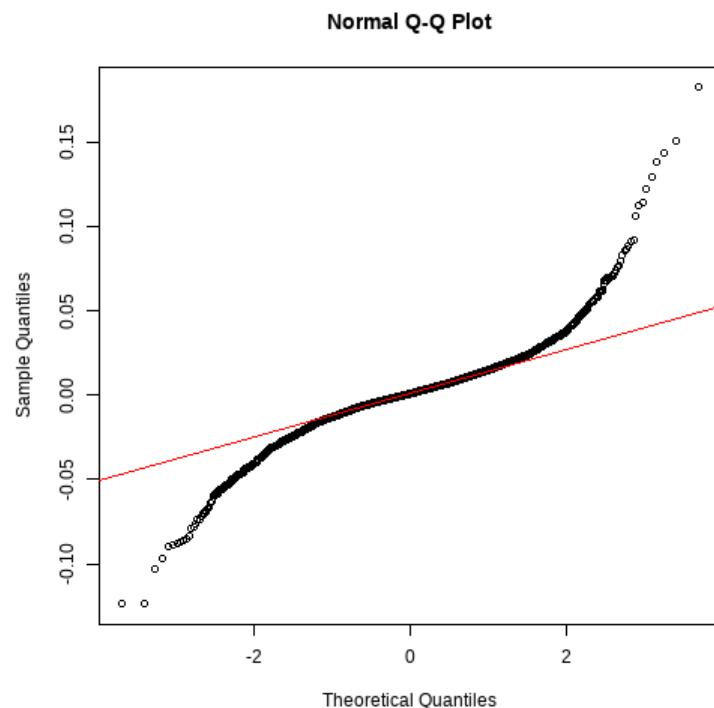
Extracting data from CRSP

```
shapiro.test(google$ret_c[which(google$ret_c>-0.68)])  
  
##  
##      Shapiro-Wilk normality test  
##  
## data: google$ret_c[which(google$ret_c > -0.68)]  
## W = 0.90833, p-value < 2.2e-16
```

Extracting data from CRSP

```
qqnorm(google$ret_c[which(google$ret_c>-0.68)])
qqline(google$ret_c[which(google$ret_c>-0.68)], col='red')
```

Extracting data from CRSP



Extracting data from CRSP

Let estimate the CAPM Beta of Google.

What do we need?

- stock return
- market return
- risk free rate

$$E(r_{it}) = \alpha_i + \beta_i(rm_t - rf_t) + \epsilon_{i,t}$$

Extracting data from CRSP (market index)

Extracting data from CRSP (market index)

```
index_CRPS <- read_csv("index_CRPS.csv")  
  
## Rows: 4594 Columns: 11  
## — Column specification ——————  
## Delimiter: ","  
## dbl (11): DATE, vwretd, vwretx, ewretd, ewretx, sprtrn, spindx, totval, totc...  
##  
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Extracting data from CRSP (market index)

Import data in R

```
index_CRPS$DATE<-ymd(index_CRPS$DATE)
```

Extracting data from CRSP (market index)

format data.

```
index_CRPS_<-index_CRPS %>% select(DATE,vwretd) %>%  
  filter(DATE>="2004-08-19") %>%  
  rename(date=DATE)
```

Extracting data from CRSP (market index)

merge the data with google returns.

```
mer_base<-left_join(google,index_CRPS_)

## Joining, by = "date"

drop omit values (NA)

mer_base<-mer_base %>% na.omit()
```

Extracting data from Fama-French Portofolio and factors (risk free rate)

Fama-French Portfolios and Factors

Eugene Fama and Kenneth French showed that their factors capture a statistically significant fraction of the variation in stock returns (see "Common Risk Factors in the Returns on Stocks and Bonds", Journal of Financial Economics 33, 1993). The Fama-French data source is Kenneth French's web site at Dartmouth.

Eugene Fama and Kenneth French showed that their factors capture a statistically significant fraction of the variation in stock returns (see "Common Risk Factors in the Returns on Stocks and Bonds", Journal of Financial Economics 33, 1993). The Fama-French data source is Kenneth French's web site at Dartmouth.

Fama-French Portfolios

Fama-French Portfolios

[Indices and Factors](#) [North America](#)

[» 2x3 Research Portfolios](#)

[» 5x5 Research Portfolios](#)

[» Factors - Daily Frequency](#)

[» Factors - Monthly Frequency](#)



[More About This Vendor >](#)

Knowledge Base >

[» Fama-French Portfolios and Factors](#)

[» Daily risk-free rate](#)

Manuals and Overviews >

[» Fama French Research Portfolios and Factors](#)

[» Size and Value in China](#)

Extracting data from Fama-French Portofolio and factors (risk free rate)

This is the query form for the Fama French Factors - Daily Frequency

Step 1: Choose your date range.

Date range

 to


[More About This Vendor >](#)

Step 2: Choose factors for query.

How does this work?

Factors

Select All

Selected Clear All (5)

Search All

- Excess Return on the Market (mktf)
- Small-Minus-Big Return (smb)
- High-Minus-Low Return (hml)
- Risk-Free Return Rate (One Month Treasury Bill Rate) (rf)
- Momentum (umdt)

Fama-French Portfolios and Factors

Fama-French Portfolios

- [» 2x3 Research Portfolios](#)
- [» 5x5 Research Portfolios](#)
- [**» Factors - Daily Frequency**](#)
- [» Factors - Monthly Frequency](#)

Knowledge Base

- [» Fama-French Portfolios and Factors](#)
- [» Daily risk-free rate](#)

Manuals and Overviews

- [» Fama French Research Portfolios and Factors](#)
- [» Size and Value in China](#)

Step 3: Select query output.

How does this work?

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	Compression Type	Date Format
<input type="radio"/> fixed-width text (*.txt) <input checked="" type="radio"/> comma-delimited text (*.csv) <input type="radio"/> Excel spreadsheet (*.xlsx) <input type="radio"/> tab-delimited text (*.txt) <input type="radio"/> HTML table (*.htm) <input type="radio"/> SAS Windows_32 dataset (*.sas7bdat) <input type="radio"/> SAS Windows_64 dataset (*.sas7bdat) <input type="radio"/> SAS Solaris_64 dataset (*.sas7bdat) <input type="radio"/> dBase file (*.dbf) <input type="radio"/> STATA file (*.dta) <input type="radio"/> SPSS file (*.sav)	<input checked="" type="radio"/> Uncompressed <input type="radio"/> zip (*.zip) <input type="radio"/> gzip (*.gz)	<input checked="" type="radio"/> YYMMDD08 (e.g. 19840725) <input type="radio"/> DATE09 (e.g. 25JUL1984) <input type="radio"/> DMMDYY08 (e.g. 25071984) <input type="radio"/> MMDDYY10 (e.g. 07251984) <input type="radio"/> DDMMYY10 (e.g. 25071984) <input type="radio"/> YYMMDD08 (e.g. 1984/07/25)

E-Mail Address (Optional)

Custom Field (Optional)

Save This Query (Optional)

Notes on this Query (Optional)

Extracting data from Fama-French Portofolio and factors (risk free rate)

Now, we can import the new data in R

```
risk_free_rate <- read_csv("risk_free_rate.csv")  
  
## Rows: 4656 Columns: 6  
## — Column specification ——————  
## Delimiter: ","  
## dbl (6): date, mktrf, smb, hml, rf, umd  
##  
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Extracting data from Fama-French Portofolio and factors (risk free rate)

and prepare them to be merged with our estimation sample data.

```
risk_free_rate <- risk_free_rate %>% select(date,rf) %>%
  mutate(date=ymd(date)) %>%
  filter(date>="2004-08-19"&date<='2022-03-31')
```

Extracting data from Fama-French Portofolio and factors (risk free rate)

Merge the data

```
mer_base<-left_join(mer_base,risk_free_rate)  
## Joining, by = "date"
```

Estimating MEDAF BETA

Variables computation

- risk premium
- market risk premium

Estimating MEDAF BETA

```
mer_base<-mer_base %>% mutate(risp=RET-rf,  
                                markrisp=vwretd-rf)
```

Estimating MEDAF BETA

Let limit us to a estimation over 130 cotation days.

```
rega<-lm(risp~markrisp,data=filter(mer_base,date<=date[130]))
summary(rega)

##
## Call:
## lm(formula = risp ~ markrisp, data = filter(mer_base, date <=
##     date[130]))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.090092 -0.021006  0.000732  0.013994  0.154748
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.004790  0.002821   1.698   0.0919 .
## markrisp    0.567737  0.451241   1.258   0.2106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03182 on 128 degrees of freedom
## Multiple R-squared:  0.01222,    Adjusted R-squared:  0.004499
## F-statistic: 1.583 on 1 and 128 DF,  p-value: 0.2106
```

If you want more about importing and managing data with R, you can watch the following video
<https://youtu.be/1CSE0mQVwAQ> (31 minutes).

