Organizing Your R Work

and staying sane

Paul Teetor
Chicago R User's Group
Jan 2017

What do I mean by a “project”?

- Doodling - for exploring
- **Project** - concrete result, or more than one day’s work; working alone
- Team project - multiple contributors

*Let’s assume we’re working on a **project**: homework, job assignment, on-going analysis, something like that.*

Today’s example: A request from my boss

How do we start? How can we stay organized?

First tip: Use RStudio!

- Makes R easier to use
- Easy to download & install (*www.rstudio.com*)
- *Tools for organizing your project*

RStudio

quantdevel.com:8787

Google | Remote | Finances | Pro | Markets | News | Social | Local | Audio | Reference | Other bookmarks

File Edit Code View Plots Session Build Debug Profile Tools Help

paul

Go to file/function | Addins | Project: (None)

Untitled1* | Untitled2*

Source on Save | Run | Source

```
#
# Playing with level vs momentum
#

library(tutils)
library(taudata)
library(ggplot2)

ast = as.Asset("^VIX")
px = getHistory(ast)
dpx = diff(px)

lev = (px - runMin(px, 252)) / (runMax(px,252) - runMin(px,252))
xlev = px / EMA(px, 252)
mom = EMA(dpx, 21) / px

```

4:1 (Top Level) R Script

Console ~/

```
(Intercept)  -0.16572    0.04260  -3.890  0.00010 ***
log(XLev)    -1.95829    0.16707 -11.721  < 2e-16 ***
Mom         -11.79708    3.94583  -2.990  0.00279 **
Signbull     -0.21305    0.06581  -3.237  0.00121 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9027.1  on 6545  degrees of freedom
Residual deviance: 8617.2  on 6542  degrees of freedom
AIC: 8625.2

Number of Fisher Scoring iterations: 4

Error in lapply(x, `[`, i) : object 'testset' not found
In addition: Warning messages:
1: Removed 251 rows containing missing values (geom_point).
2: Removed 251 rows containing missing values (geom_point).
> 
```

Environment | History

Import Dataset | List

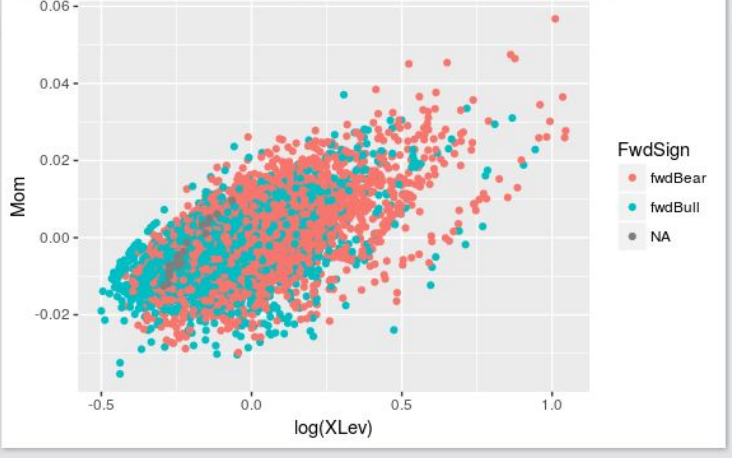
Global Environment

Data

clean	6546 obs. of 7 variables
dfrm	6818 obs. of 7 variables
dpx	An ‘xts’ object on 1990-01-02/2017-01-20 cont…
lev	An ‘xts’ object on 1990-01-02/2017-01-20 cont…
mom	An ‘xts’ object on 1990-01-02/2017-01-20 cont…
px	An ‘xts’ object on 1990-01-02/2017-01-20 cont…
xlev	An ‘xts’ object on 1990-01-02/2017-01-20 cont…

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



Start by creating a new RStudio project

- Creates an empty directory, dedicated to project
- Becomes a container & memory for your work
- Provides a “bookmark” to find your project later
- Project -> New Project … -> New Directory… -> Empty Project -> *give it a name*

Pro Tip: Consider having a directory of projects

RStudio

quantdevel.com:8787

Google | Remote | Finances | Pro | Markets | News | Social | Local | Audio | Reference | Other bookmarks

File Edit Code View Plots Session Build Debug Profile Tools Help

paul

Go to file/function | Addins | Project: (None)

Untitled1* | Untitled2*

Source on Save | Run | Source

```
#
# Playing with level vs momentum
#

library(tutils)
library(taudata)
library(ggplot2)

ast = as.Asset("^VIX")
px = getHistory(ast)
dpx = diff(px)

lev = (px - runMin(px, 252)) / (runMax(px,252) - runMin(px,252))
xlev = px / EMA(px, 252)
mom = EMA(dpx, 21) / px

```

4:1 (Top Level) R Script

Environment | History

Import Dataset | Global Environment

Data

clean	6546 obs. of 7 varia
dfrm	6818 obs. of 7 varia
dpx	An 'xts' object on 1
lev	An 'xts' object on 1
mom	An 'xts' object on 1
px	An 'xts' object on 1
xlev	An 'xts' object on 1

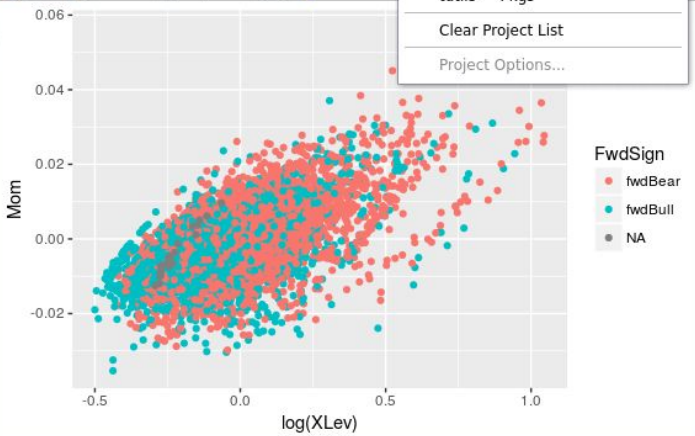
New Project...
Open Project...
Close Project
CRUG-sandbox
CRUG-2017
allotalot
asset
portfolio
portfolios
taufin
taudata
CSP-2017
tutils — Pkgs
Clear Project List
Project Options...

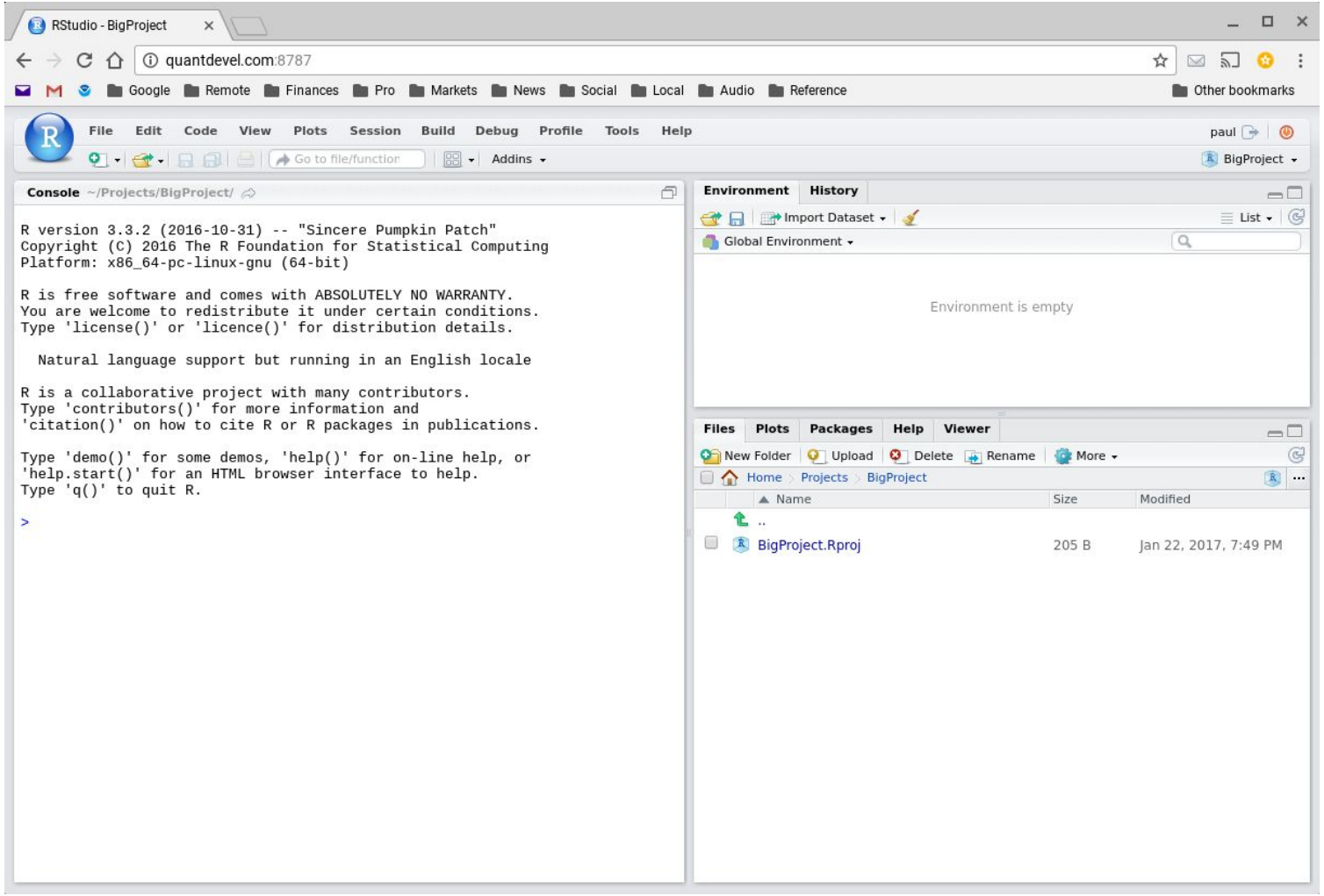
Console ~/

```
(Intercept)  -0.16572    0.04260  -3.890  0.00010 ***
log(XLev)    -1.95829    0.16707 -11.721  < 2e-16 ***
Mom         -11.79708    3.94583  -2.990  0.00279 **
Signbull     -0.21305    0.06581  -3.237  0.00121 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9027.1  on 6545  degrees of freedom
Residual deviance: 8617.2  on 6542  degrees of freedom
AIC: 8625.2

Number of Fisher Scoring iterations: 4

Error in lapply(x, `[`, i) : object 'testset' not found
In addition: Warning messages:
1: Removed 251 rows containing missing values (geom_point).
2: Removed 251 rows containing missing values (geom_point).
> 
```

Files | Plots | Packages | Help | Viewer

Zoom | Export



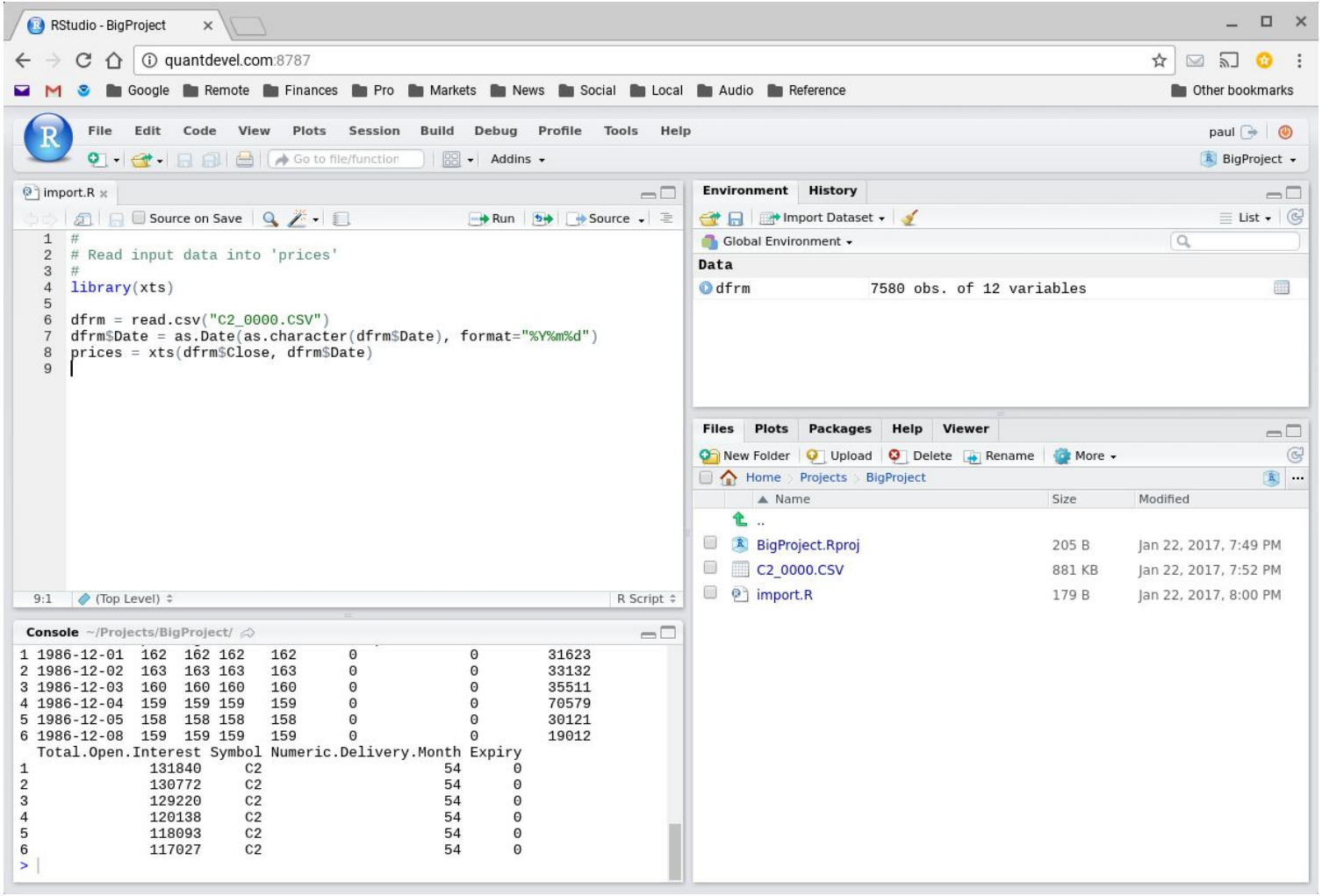

RStudio - BigProject
quantdevel.com:8787
Google
Remote
Finances
Pro
Markets
News
Social
Local
Audio
Reference
Other bookmarks
File
Edit
Code
View
Plots
Session
Build
Debug
Profile
Tools
Help
paul
Go to file/function
Addins
BigProject
Console ~/Projects/BigProject/
R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
Environment
History
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
New Folder
Upload
Delete
Rename
More
Home
Projects
BigProject
Name
Size
Modified
..
BigProject.Rproj
205 B
Jan 22, 2017, 7:49 PM

Next, capture and load your data

- Save data file in project directory
- Look at the data file
- Try loading data file into R
- Transform into something useful for this project
- Capture the load-and-transforms steps into a script (or function)

Let's call that script "import.R"

The script's job is to reliably load your data.


RStudio - BigProject
quantdevel.com:8787
Google
Remote
Finances
Pro
Markets
News
Social
Local
Audio
Reference
Other bookmarks
File
Edit
Code
View
Plots
Session
Build
Debug
Profile
Tools
Help
paul
Go to file/function
Addins
BigProject
import.R
Source on Save
Run
Source
# Read input data into 'prices'
library(xts)
dfrm = read.csv("C2_0000.CSV")
dfrm$Date = as.Date(as.character(dfrm$Date), format="%Y%m%d")
prices = xts(dfrm$Close, dfrm$Date)
9:1
(Top Level)
R Script
Environment
History
Import Dataset
List
Global Environment
Data
dfrm
7580 obs. of 12 variables
Files
Plots
Packages
Help
Viewer
New Folder
Upload
Delete
Rename
More
Home
Projects
BigProject
Name
Size
Modified
..
BigProject.Rproj
205 B
Jan 22, 2017, 7:49 PM
C2_0000.CSV
881 KB
Jan 22, 2017, 7:52 PM
import.R
179 B
Jan 22, 2017, 8:00 PM
Console ~/Projects/BigProject/
1 1986-12-01 162 162 162 162 0 0 31623
2 1986-12-02 163 163 163 163 0 0 33132
3 1986-12-03 160 160 160 160 0 0 35511
4 1986-12-04 159 159 159 159 0 0 70579
5 1986-12-05 158 158 158 158 0 0 30121
6 1986-12-08 159 159 159 159 0 0 19012
Total.Open.Interest Symbol Numeric.Delivery.Month Expiry
1 131840 C2 54 0
2 130772 C2 54 0
3 129220 C2 54 0
4 120138 C2 54 0
5 118093 C2 54 0
6 117027 C2 54 0
>

Then build your analysis incrementally

- Doodle! Play with the data, play with your model, make plots
- Select the useful doodles
- Capture the useful ones in a file
- That becomes your analysis

Let's call that file "analysis.R"

Pro tip: Use the RStudio History pane to grab useful doodles

RStudio - BigProject

quantdevel.com:8787

Google | Remote | Finances | Pro | Markets | News | Social | Local | Audio | Reference | Other bookmarks

File Edit Code View Plots Session Build Debug Profile Tools Help

paul

Go to file/function | Addins | BigProject

analysis.R | import.R

Source on Save | Run | Source

```
1  #
2  # Analysis of price changes
3  #
4
5  # Compute relative changes
6  changes = diff(prices) / head(prices,-1)
7
8  # Show some descriptive stats
9  print(summary(changes))
10 cat("SD =", sd(changes, na.rm=TRUE), "\n")
11 cat("Date range is", format(range(dates)), "\n")
12
13 # Produce quantile-quantile plot of changes
14 qqnorm(changes)
15 qqline(changes)
16
```

16:1 (Top Level) R Script

Environment | History

Import Dataset | List

Global Environment

Data

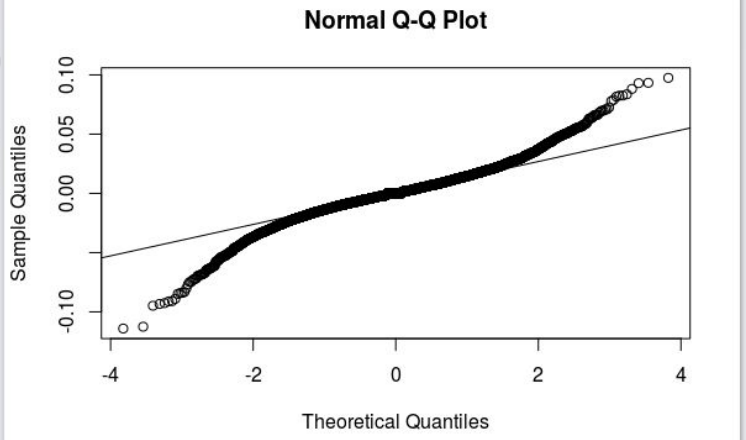
dfrm	7580 obs. of 12 variables

Values

changes	num [1:7579] 0.00617 -0.0184 -0.00625 -0.00629 0.…
dates	Date[1:7580], format: "1986-12-01" "1986-12-02" "…
prices	num [1:7580] 162 163 160 159 158 ...

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



Console ~/Projects/BigProject/

```
 num [1:7579] 1 -3 -1 -1 1 -0.5 0 -4 -2 0 ...
> str(head(prices,-1))
 num [1:7579] 162 163 160 159 158 ...
> changes = diff(prices) / head(prices,-1)
> print(summary(changes))
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-0.1141000 -0.0085590  0.0000000  0.0002539  0.0093460  0.0975200
> cat("SD =", sd(changes, na.rm=TRUE), "\n")
SD = 0.01739884
> cat("Date range is", format(range(dates)), "\n")
Date range is 1986-12-01 2017-01-20
> qqnorm(changes)
> qqline(changes)
>
```

Use debugging tools to pinpoint your errors

- To debug a simple script, use Ctrl-Enter to “single step” through.
- Set breakpoints in editor before running script
- *Or*, insert breakpoints with `browser()` function

Pro tip: When you write functions, set On Error to “Break in Code” to easily pinpoint errors.

RStudio - BigProject

quantdevel.com:8787

Google | Remote | Finances | Pro | Markets | News | Social | Local | Audio | Reference | Other bookmarks

File Edit Code View Plots Session Build Debug Profile Tools Help

paul

BigProject

Debug menu:
- Toggle Breakpoint Shift+F9
- Clear All Breakpoints...
- Execute Next Line F10
- Step Into Function Shift+F4
- Finish Function/Loop Shift+F6
- Continue Shift+F5
- Stop Debugging Shift+F8
- On Error
  - Message Only
  - Error Inspector
  - ✓ Break in Code
- Debugging Help

analysis.R | import.R

Source on Save

```
1  #
2  # Analysis of price changes
3  #
4
5  # Compute relative changes
6  changes = diff(prices) / head(prices,-1)
7
8  # Show some descriptive stats
9  print(summary(changes))
10 cat("SD =", sd(changes), "\n")
11 cat("Date range is", format(range(date)), "\n")
12
13 # Produce quantile-quantile plot of changes
14 qqnorm(changes)
15 qqline(changes)
16
```

11:39 (Top Level) R Script

Environment | History

Import Dataset List

Global Environment

Data

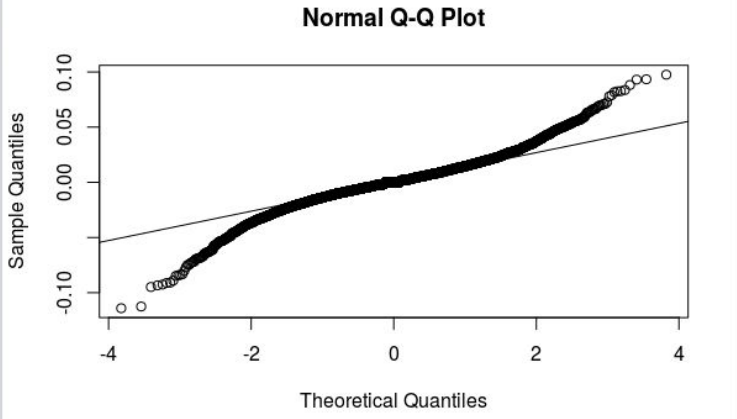
dfrm	7580 obs. of 12 variables

Values

changes	num [1:7579] 0.00617 -0.0184 -0.00625 -0.00629 0.…
date	Date[1:7580], format: "1986-12-01" "1986-12-02" "…
prices	num [1:7580] 162 163 160 159 158 ...

Files | Plots | Packages | Help | Viewer

Zoom Export Publish



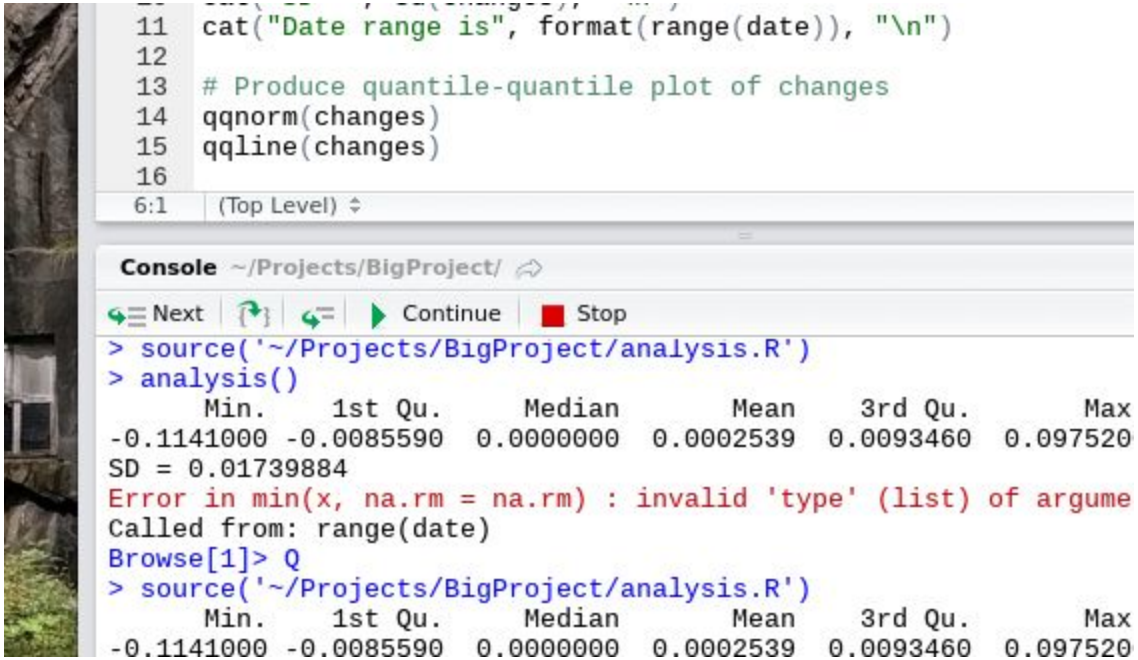
Console ~/Projects/BigProject/

```
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-0.1141000 -0.0085590  0.0000000  0.0002539  0.0093460  0.0975200
SD = 0.01739884
Called from: eval(expr, envir, enclos)
Browse[1]> n
debug at ~/Projects/BigProject/analysis.R#11: cat("Date range is", format(rang
e(dates)), "\n")
Browse[2]> head(changes)
[1]  0.006172840 -0.018404908 -0.006250000 -0.006289308  0.006329114
[6] -0.003144654
Browse[2]> summary(prices)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  130.0   209.5   257.0   310.0   357.0   849.0
Browse[2]> n
Date range is 1986-12-01 2017-01-20
debug at ~/Projects/BigProject/analysis.R#14: qqnorm(changes)
Browse[2]> n
debug at ~/Projects/BigProject/analysis.R#15: qqline(changes)
Browse[2]> n
>
```

Know the debugger commands

The common ones are

- Evaluate and print an expression
- *n* = Next
- *c* = Continue
- *Q* = Quit the debugger and return to the command line



```
11  cat("Date range is", format(range(date)), "\n")
12
13  # Produce quantile-quantile plot of changes
14  qqnorm(changes)
15  qqline(changes)
16
```

6:1 (Top Level)

Console ~/Projects/BigProject/

Next | Continue | Stop

```
> source('~/Projects/BigProject/analysis.R')
> analysis()
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max
-0.1141000 -0.0085590  0.0000000  0.0002539  0.0093460  0.097520
SD = 0.01739884
Error in min(x, na.rm = na.rm) : invalid 'type' (list) of argume
Called from: range(date)
Browse[1]> Q
> source('~/Projects/BigProject/analysis.R')
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max
-0.1141000 -0.0085590  0.0000000  0.0002539  0.0093460  0.097520
```

The next level: Split into load, analyse, report

- import.R - load data
- analysis.R - analyze data, *save results to intermediate file (e.g., models) rather than print them*
- report.R - *read results from file and format nicely for printing*

This works much better than intermixing everything in “one big script”.

Why split into parts?

- “One big script” will become a headache over time
- Easier to debug analysis when it’s separate
- Easier to run and view printing & formatting when it’s separate
- Easier to reuse parts

Some final ideas

- Create a *Sandbox* project for yourself - *not really a project, a place to keep all the little, miscellaneous stuff*
- Learn and use RMarkdown - *for beautifully formatted output*
- Consider keeping a README file of notes, ideas, lessons, conclusions for each project - *because you will forget*
- For long-term projects, consider learning about *git* for managing your files