zeppelin universität

zwischen
Wirtschaft Kultur Politik

Data Analytics and Visualisation

0 An Introduction

David Zimmermann

2017-04-21

Zeppelin University

Idea of the Workshop

Today: until ~19:30

Tomorrow and the day after: 10-~16, this room

🍸Tomorrow social get together (Rathauscafe, FN at 19:00) 🍸

Course:

1. Theoretical input
2. Exercises
   - Standard
   - Advanced, if you are finished with the standard exercises (potentially in your free time)

Solutions will be posted to:

https://github.com/DavZim/RDataAnalytics

What we are going to do

- Friday: Base R and R Programming
- Saturday: Data Manipulation
- Sunday: Data Visualization

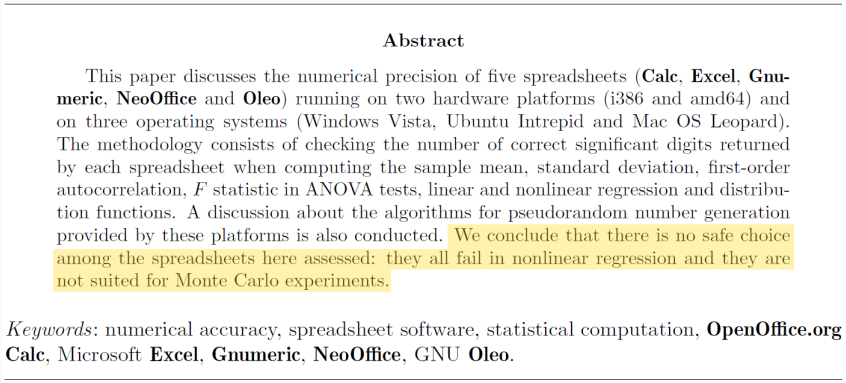
What we are going to leave out

- Statistics & Inference
- Machine Learning & Artificial Intelligence
- High-Performance Computing & “Big Data”

If you are interested maybe next time

R

Spreadsheets (such as Excel)?



Abstract

This paper discusses the numerical precision of five spreadsheets (**Calc**, **Excel**, **Gnumeric**, **NeoOffice** and **Oleo**) running on two hardware platforms (i386 and amd64) and on three operating systems (Windows Vista, Ubuntu Intrepid and Mac OS Leopard). The methodology consists of checking the number of correct significant digits returned by each spreadsheet when computing the sample mean, standard deviation, first-order autocorrelation, F statistic in ANOVA tests, linear and nonlinear regression and distribution functions. A discussion about the algorithms for pseudorandom number generation provided by these platforms is also conducted. We conclude that there is no safe choice among the spreadsheets here assessed: they all fail in nonlinear regression and they are not suited for Monte Carlo experiments.

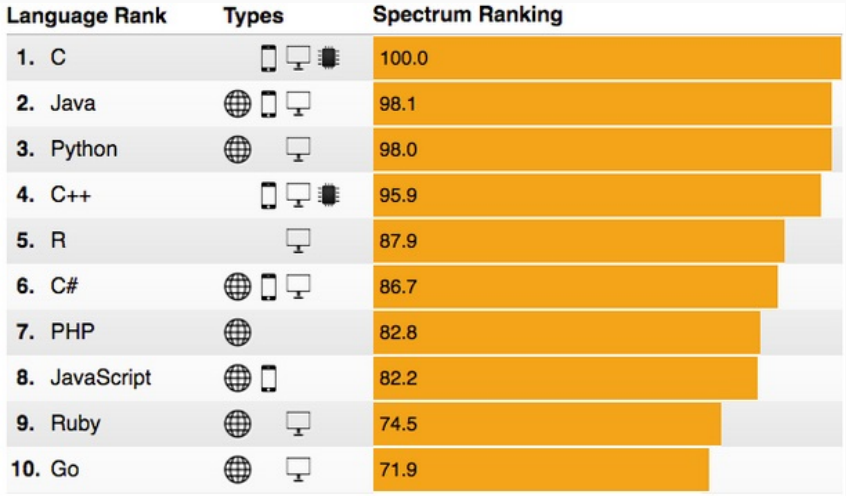
Keywords: numerical accuracy, spreadsheet software, statistical computation, **OpenOffice.org Calc**, Microsoft **Excel**, **Gnumeric**, **NeoOffice**, GNU **Oleo**.

Source: Almiron et al. (2010): On the Numerical Accuracy of Spreadsheets
https://www.jstatsoft.org/article/view/v034i04

Further Motivation:

- Excel Errors and Science Papers:
  http://www.economist.com/blogs/graphicdetail/2016/09/daily-chart-3
- Herndon vs. Reinhart, Rogoff:
  http://www.bbc.com/news/magazine-22223190
- https://baselinescenario.com/2013/02/09/the-importance-of-excel/
- http://www.zerohedge.com/news/2013-02-12/how-rookie-excel-error-led-jpmorgan-misreport-its-var-years

Why use R? - Popularity



Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

Source: http://spectrum.ieee.org/computing/software/the-2016-top-programming-languages

Why use R? - Usage

Source: http://blog.revolutionanalytics.com/2014/05/companies-using-r-in-2014.html

Why use R? - Open Source

```
R version 3.3.3 (2017-03-06) -- "Another Canoe"
Copyright (C) 2017 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.
```

Source: https://www.r-project.org/Licenses/GPL-3

GPL-3:

- Free of charge
- Distribute as you like
- Open source code
- Contribute

$\sim$ peer-reviewed software

Why use R? - Community

CRAN Task Views

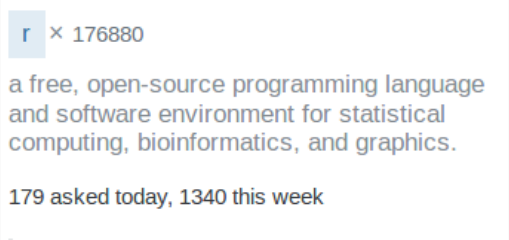
Special (curated) lists for topics such as:

- Finance
- Graphics
- Econometrics + Time Series
- Machine Learning
- Natural Language Processing (NLP)
- Spatial
- Genetics
- Psychometrics
- Environmetrics

More: https://cran.r-project.org/web/views/

Why use R? - Community cont'd

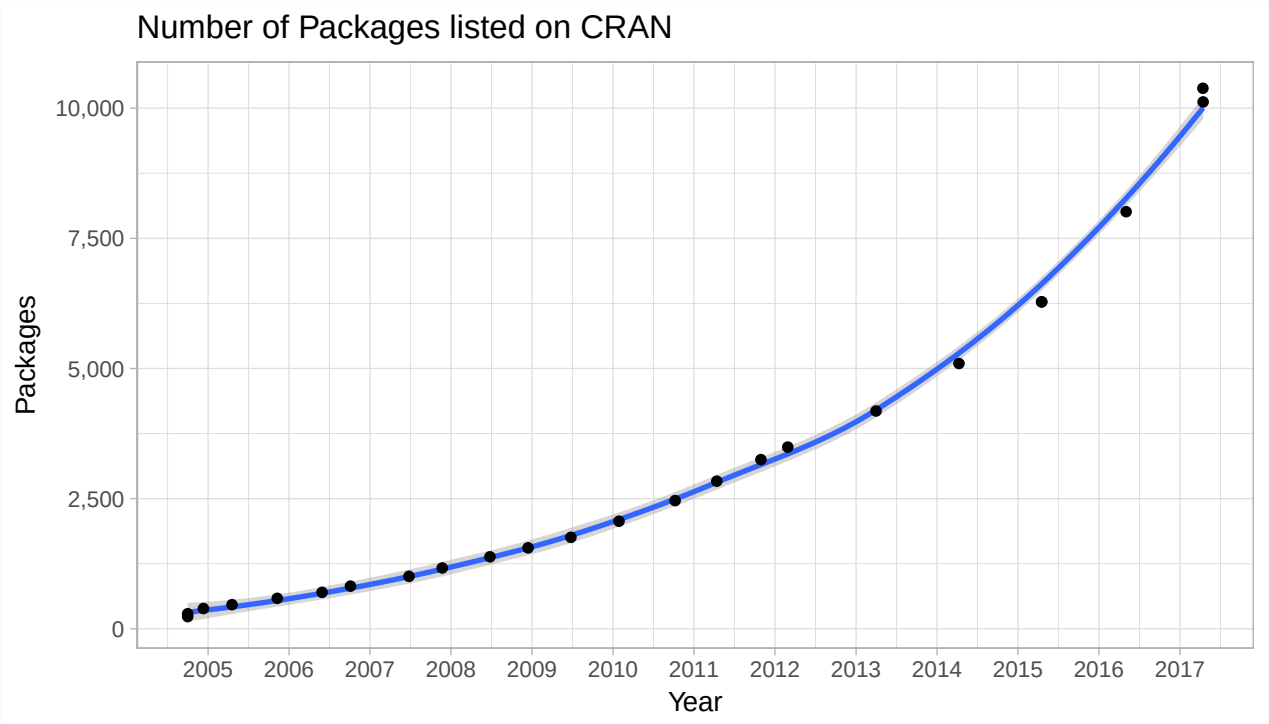
StackOverflow



Source: http://stackoverflow.com/tags

Why use R? - Community cont'd

Packages Libraries



Source: Author's own creation using
http://blog.revolutionanalytics.com/2016/04/cran-package-growth.html

Why use R? - Community cont'd

Blogs

Source: https://www.r-bloggers.com/

Why use R? - Documentation

Help Functions Asking yourself what function `plot` does, what arguments it takes, how to use/tweak it?

`?plot`: Inside of R, built-in help

If that doesn't help: Google and/or StackOverflow

Potential Goals

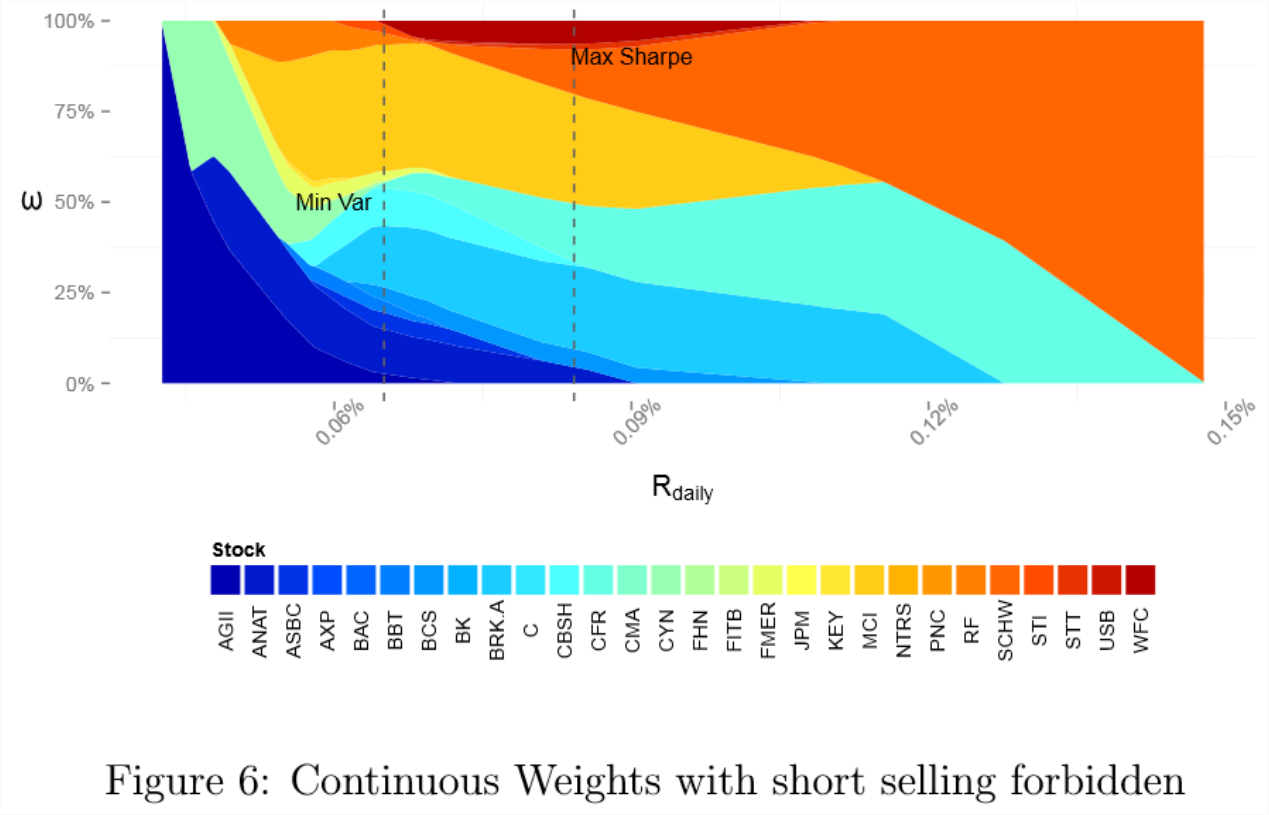
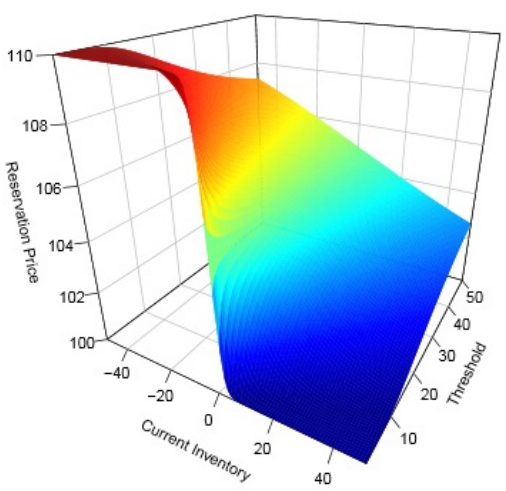


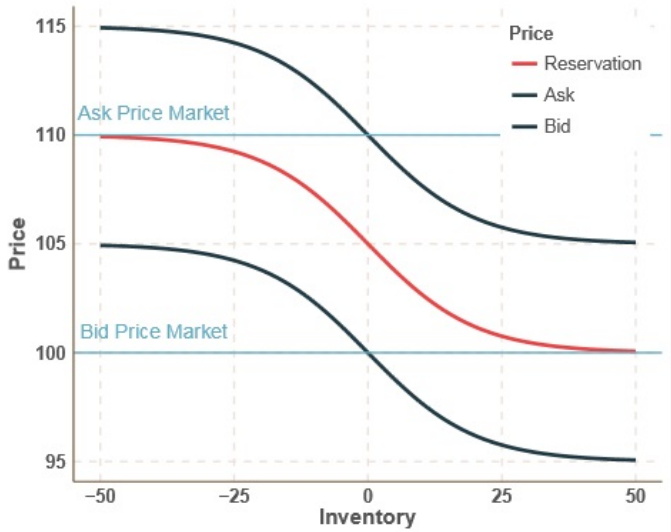
Figure 6: Continuous Weights with short selling forbidden

Source: Author's own creation

Potential Goals cont'd

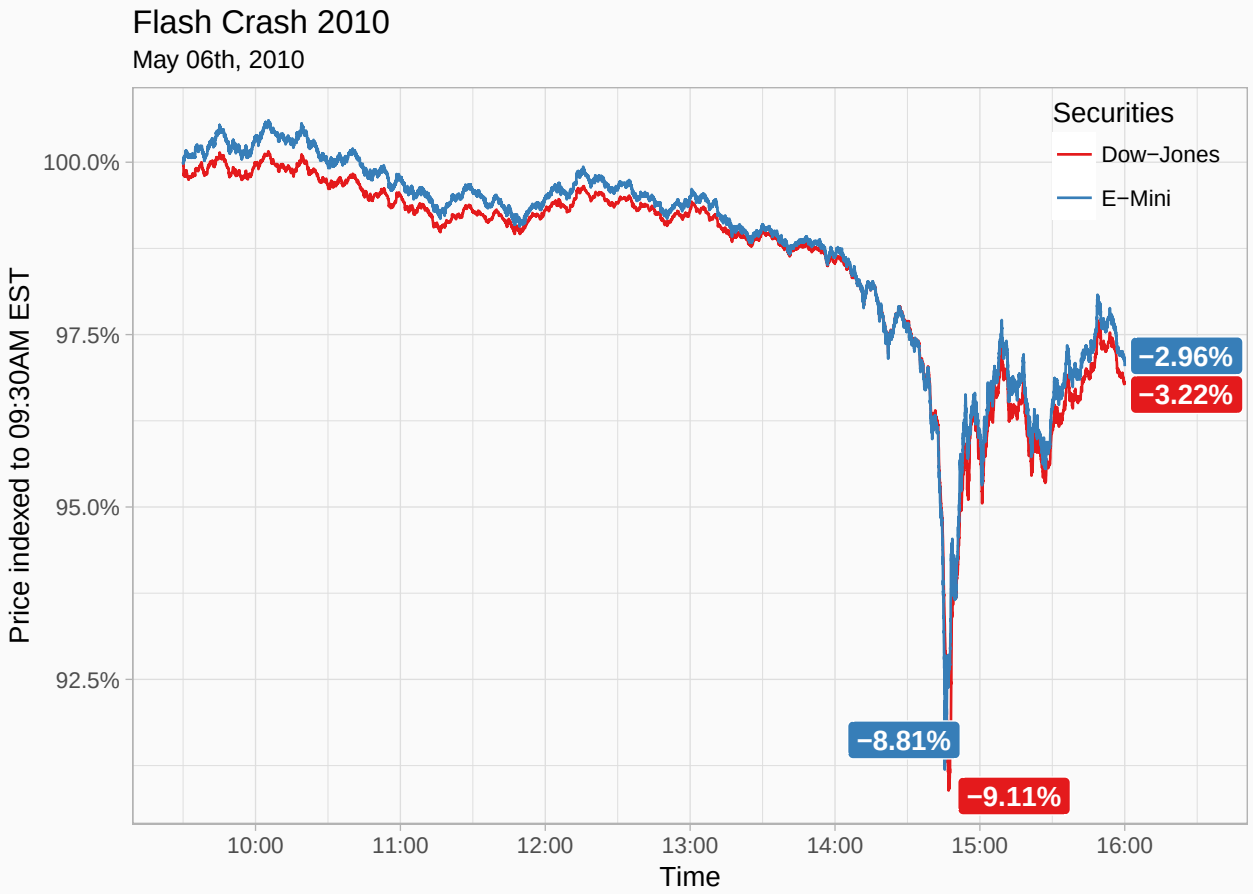
Figure 6: Reservation Price of High-Frequency Traders





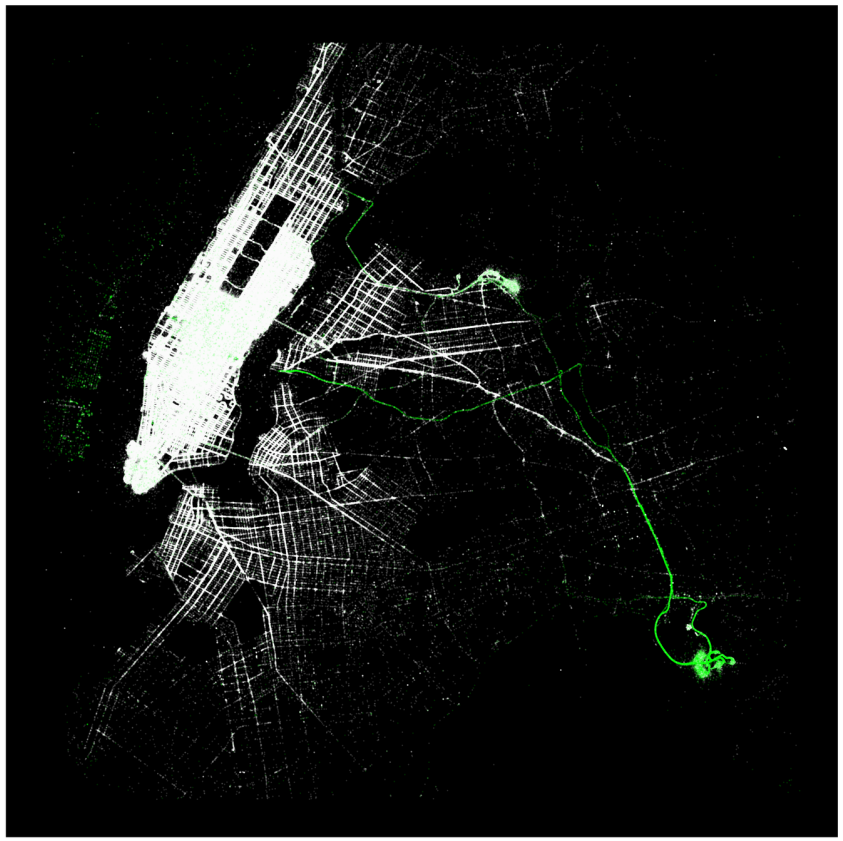
Source: Author's own creation

Potential Goals cont'd



Source: Author's own creation

Potential Goals cont'd



Source: Author's own creation using 10 mil. 2016 yellow-cab entries from http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

Potential Goals cont'd

Source: http://blog.revolutionanalytics.com/2010/12/facebooks-social-network-graph.html and https://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919

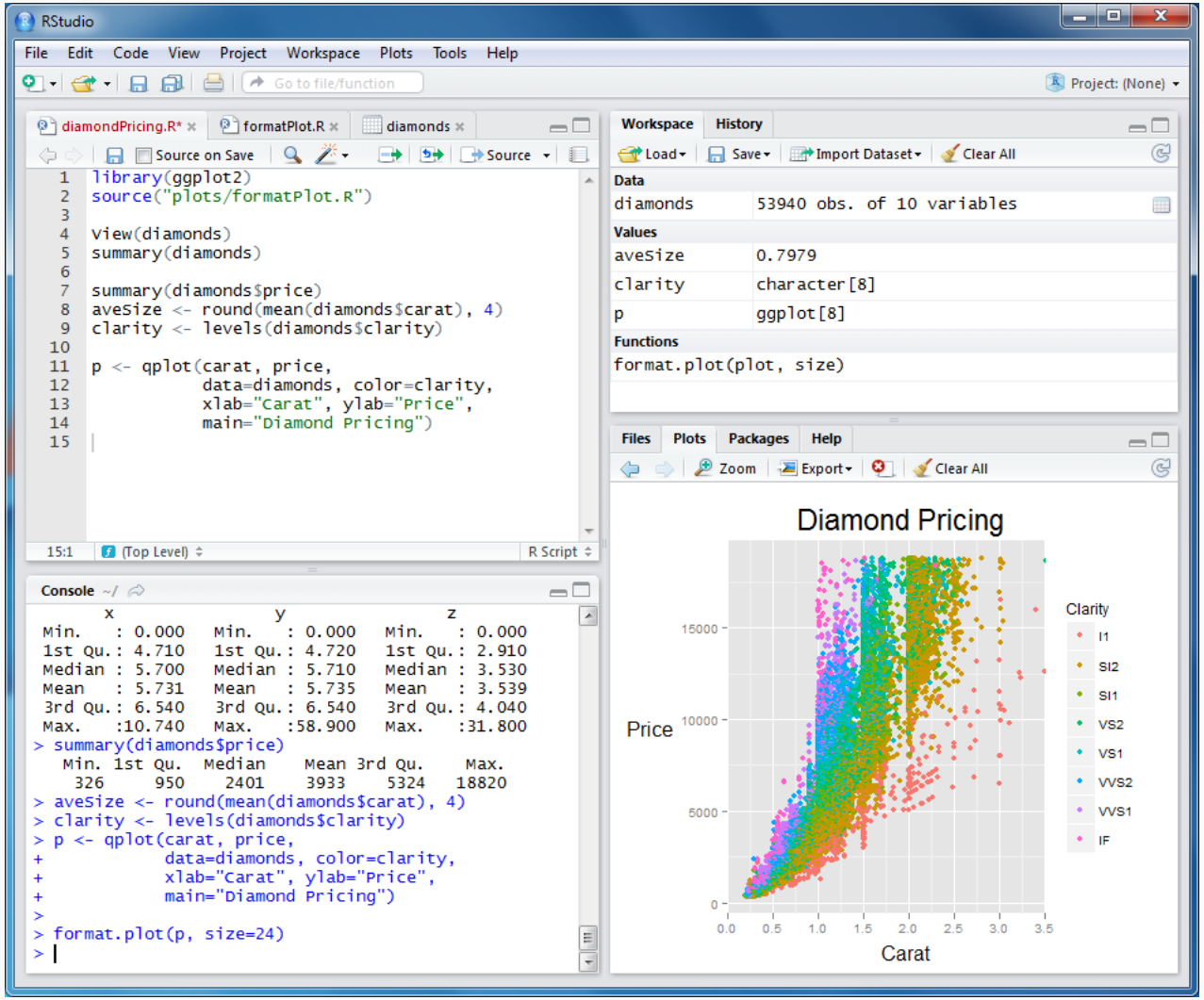
Getting down to R

Installing R and RStudio

R: https://cran.r-project.org/

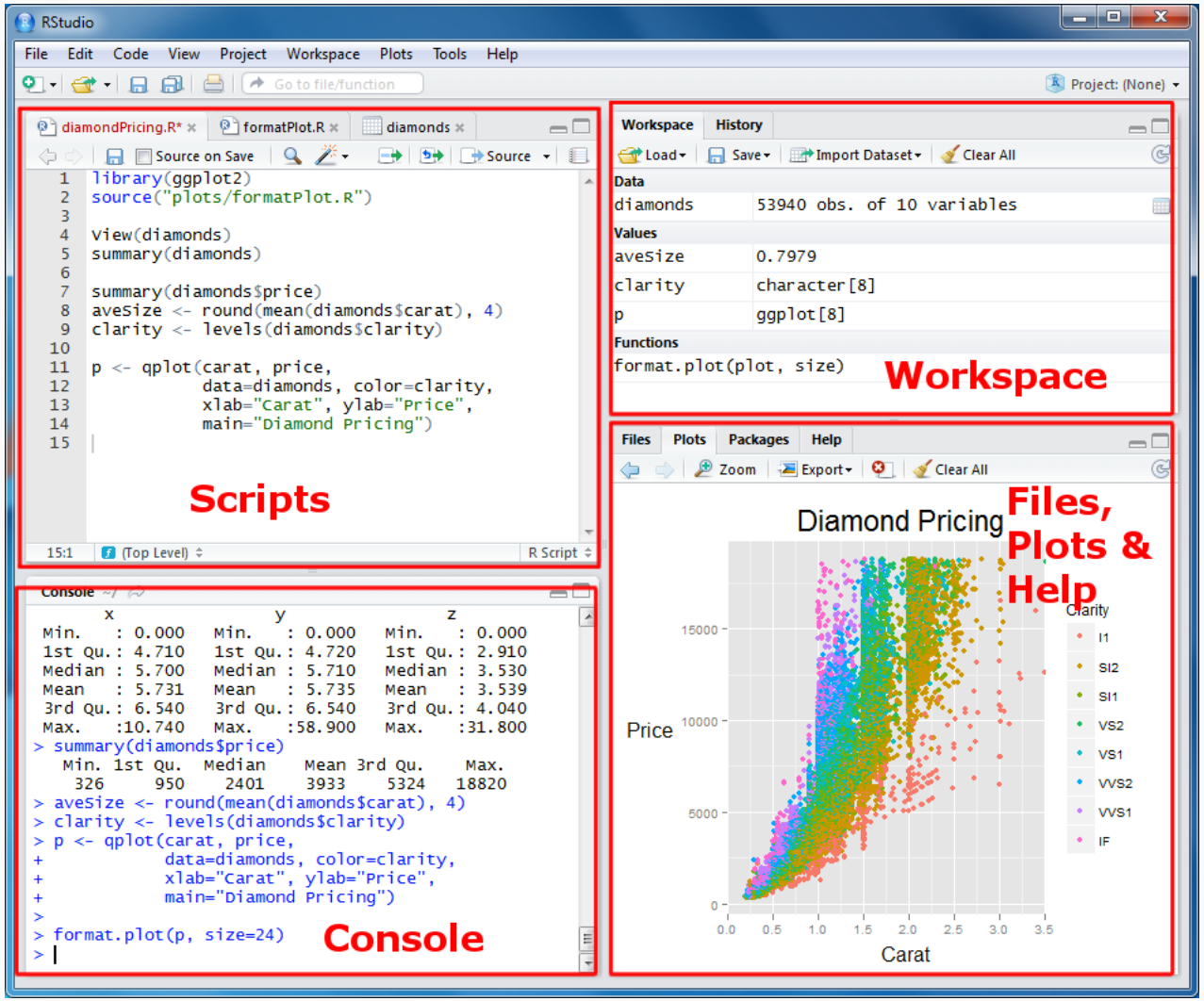
RStudio: https://www.rstudio.com/products/RStudio/

RStudio



Source: http://rprogramming.net/download-and-install-rstudio/

RStudio cont'd



Demonstration

Additional Information

Learning R elsewhere

- www.swirlstats.com/students
- http://tryr.codeschool.com
- https://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf
- www.datacamp.com/courses/free-introduction-to-r/
- www.rstudio.com/online-learning/
- www.ats.ucla.edu/stat/r/
- https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
- http://r4ds.had.co.nz/
- Reproducible Resarch: www.kbroman.org/pages/tutorials.html

RMarkdown

Another option to scripting. Output is by default a html-page or a pdf.

Use markdown for writing, insert r-code chunks with (`cmd/strg` + `alt` + `i` or click "insert new code chunk") and hit `knit` to create the output

Next up: base R and R Programming