Introduction to R (1)

Short course on modelling infectious disease dynamics in R

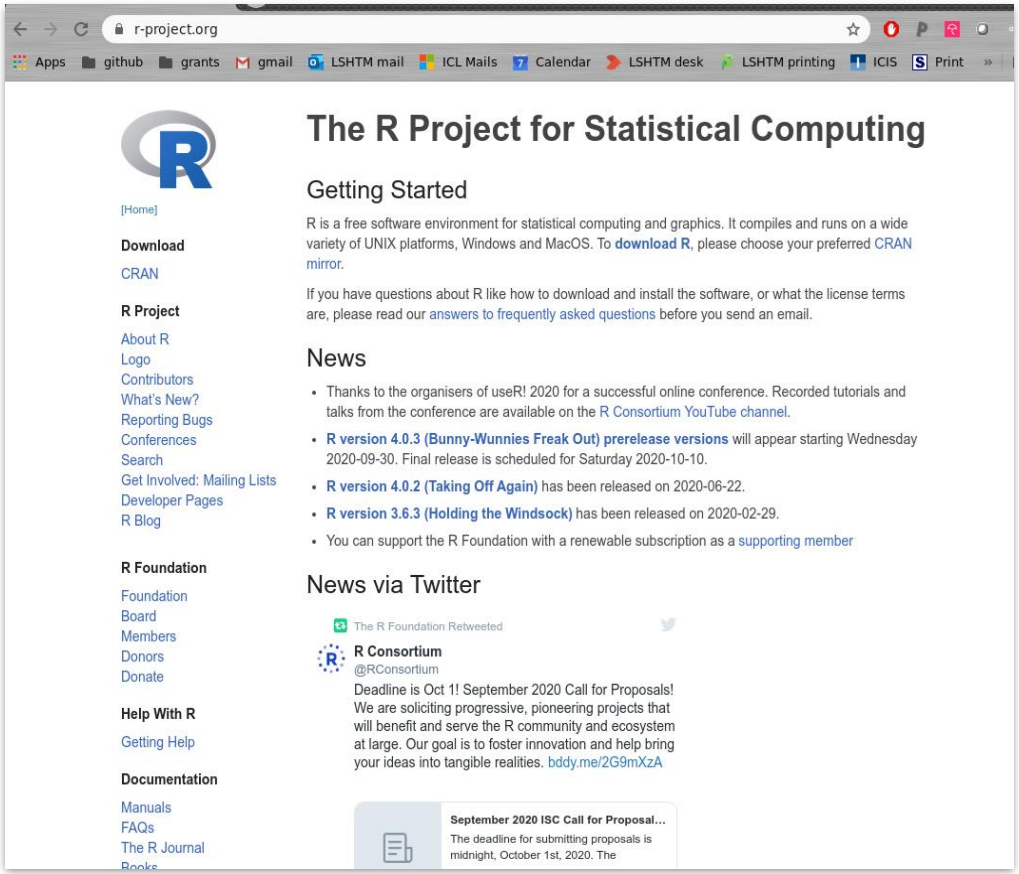
Ankara, Türkiye, September 2025

Dr Juan F Vesga

Aims of the session

- Understand why R is a great option for epidemiology
- Get started with R and Rstudio
- Understand the value of R and Rstudio compared to other programming languages
- Familiarize with the R environment

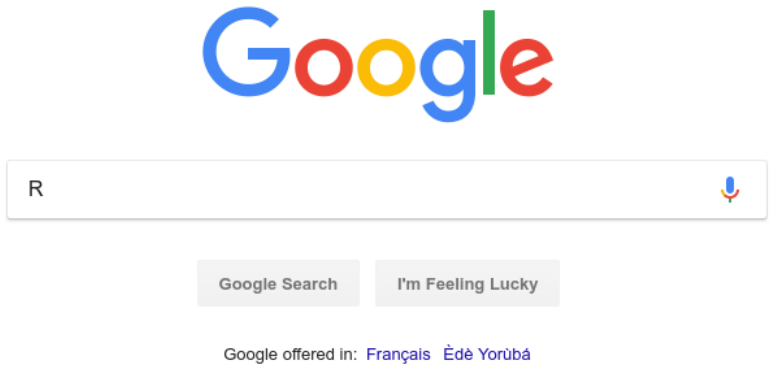
What is R ?



- a free software for data analysis
- an interpreted programming language, derived from `S-plus'
- initially developed by R. Ihaka and R. Gentleman (1996)
- currently developed by the R Core Team (~20 people)

- **largest collection of tools for data analysis** (1,000s of contributors and specific packages)

Where can you get it?



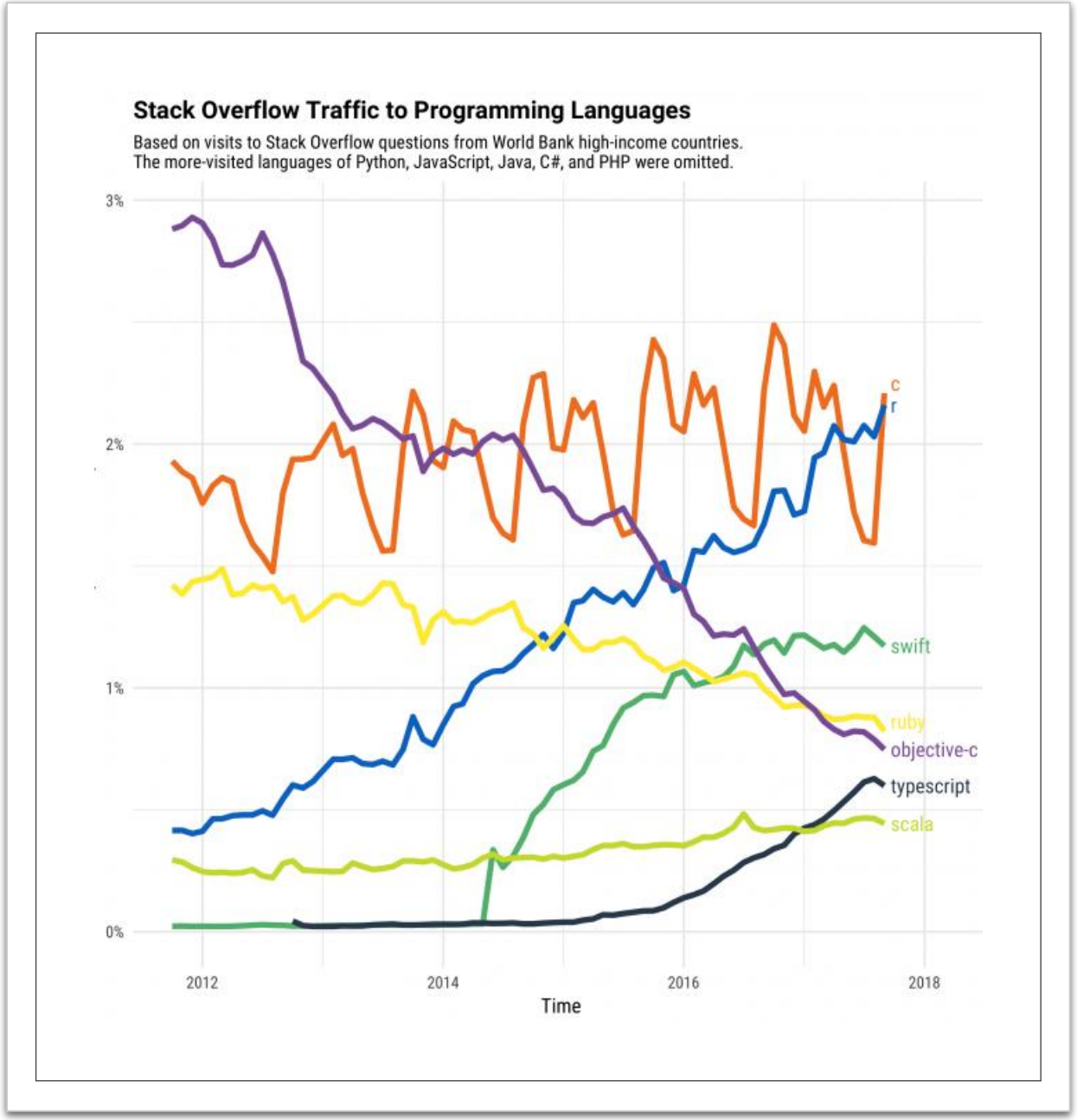
- The **R** project: www.r-project.org
- archiving / distribution network CRAN: cran.r-project.org/mirrors.html
- available on Windows, MacOSX, Linux

What can you do with it?

- basic statistics: statistical tests, linear modelling, multivariate analysis
- spatial statistics: GIS, mapping, clustering
- graph theory: social sciences, networ analysis, graph algorithm
- genetics: phylogenetic trees, genetic markers, genomics
- epidemiology!

What does “free” mean?

- Freedom = *ability to make informed decisions*
- you don’t pay for it
- the code is accessible by anyone
- anyone can use, modify and share the code


Stack Overflow Traffic to Programming Languages
Based on visits to Stack Overflow questions from World Bank high-income countries.
The more-visited languages of Python, JavaScript, Java, C#, and PHP were omitted.
3%
2%
1%
0%
2012
2014
2016
2018
Time
c
r
swift
ruby
objective-c
typescript
scala

Growing use

and the 'epi' community

- R Epidemics Consortium (RECON)
  - NGO for the development of resources for health emergencies, with strong focus on outbreak response
  - 20+ packages for outbreak analytics
  - Share needs on COVID-19 challenge
  - www.repidemicsconsortium.org
- R4epis
  - Partnership between MSF + RECON
  - Analysis templates for field epidemiologists
  - R4epis.netlify.app
- RECON learn
  - Free / open training material by RECON
  - www.reconlearn.org

Alternatives to R

Python

- Both are frequently used in data science, and both are free
- Python is better for analyses that need scaling up
- R is better for investigative “one-time” analyses
- R more commonly used in medical research
- Python more commonly used in industry

Stata

- R and Stata most commonly used languages in medical research
- Commonly taught in medical statistics & epidemiology courses
- Requires a licence, but has official help
- Strong statistical analyses capabilities
- Less developed data visualization

Alternatives to R

SPSS

- Statistical Package for the Social Sciences
- Many researchers' first exposure to a statistical program
- Usually used as a "point and click", although can write code
- Use is declining, and also requires a licence

Excel

- Very widely used
- Can be helpful for "having a look" at the data
- Generally not suitable for research-level inferential statistical analyses

Getting started

- get **R** for your system (download from CRAN)
- get a Graphical User Interface (GUI): **RStudio**, emacs + ESS, Tinn-R
- (or at least) get a text editor to write code: notepad++, emacs, vi, Tinn-R, ...

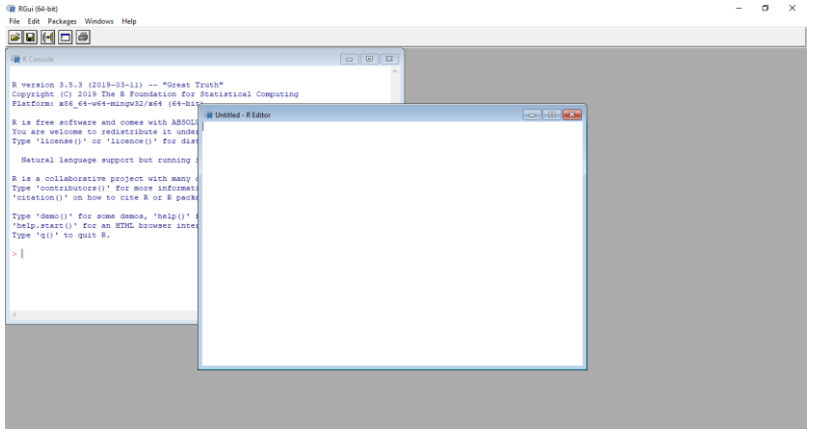
R Studio

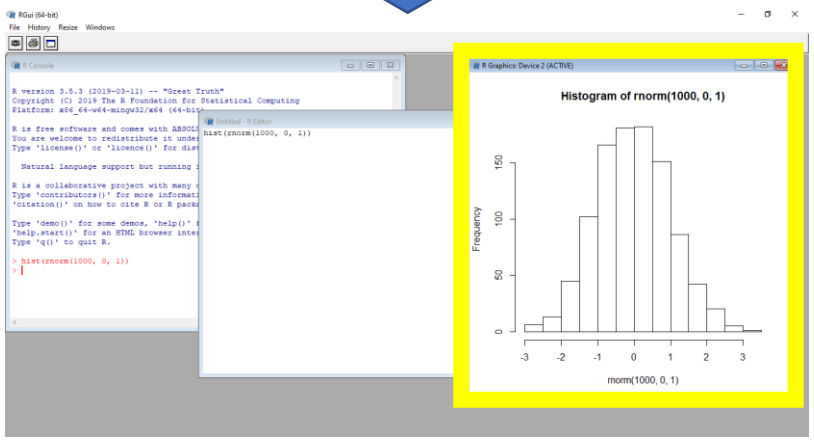
- R Studio is an Integrated Development Environment (IDE) for R
  - i.e. it makes doing things in R easier and more structured

- Some advantages include:
  - An organised environment for your projects
  - Colour-coding, auto-completed brackets, auto-structuring
  - Keeping track of which packages have been installed, and which variables have been defined
  - Fixed quadrants for code and graphical outputs

R (without R Studio)

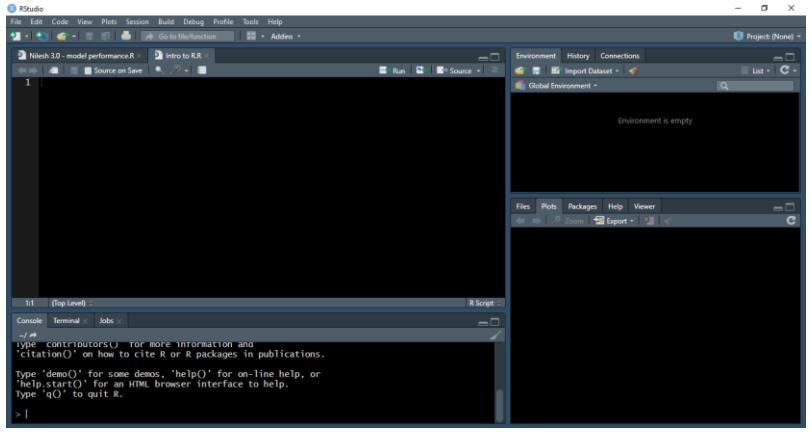
graphs appear in a new window

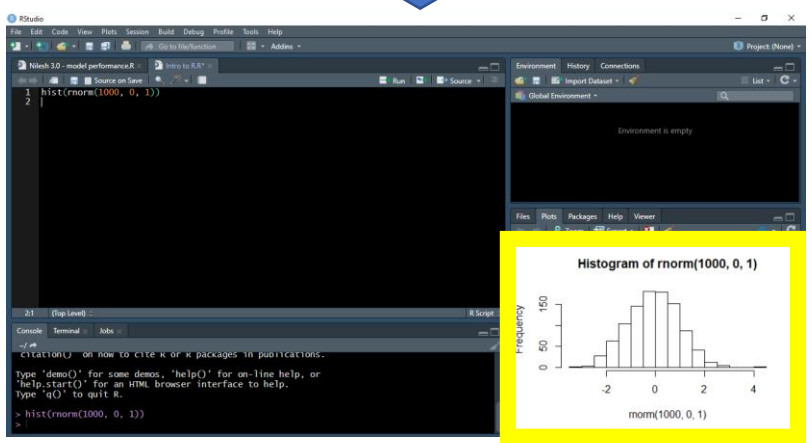




R Studio

graphs appear in bottom-right quadrant

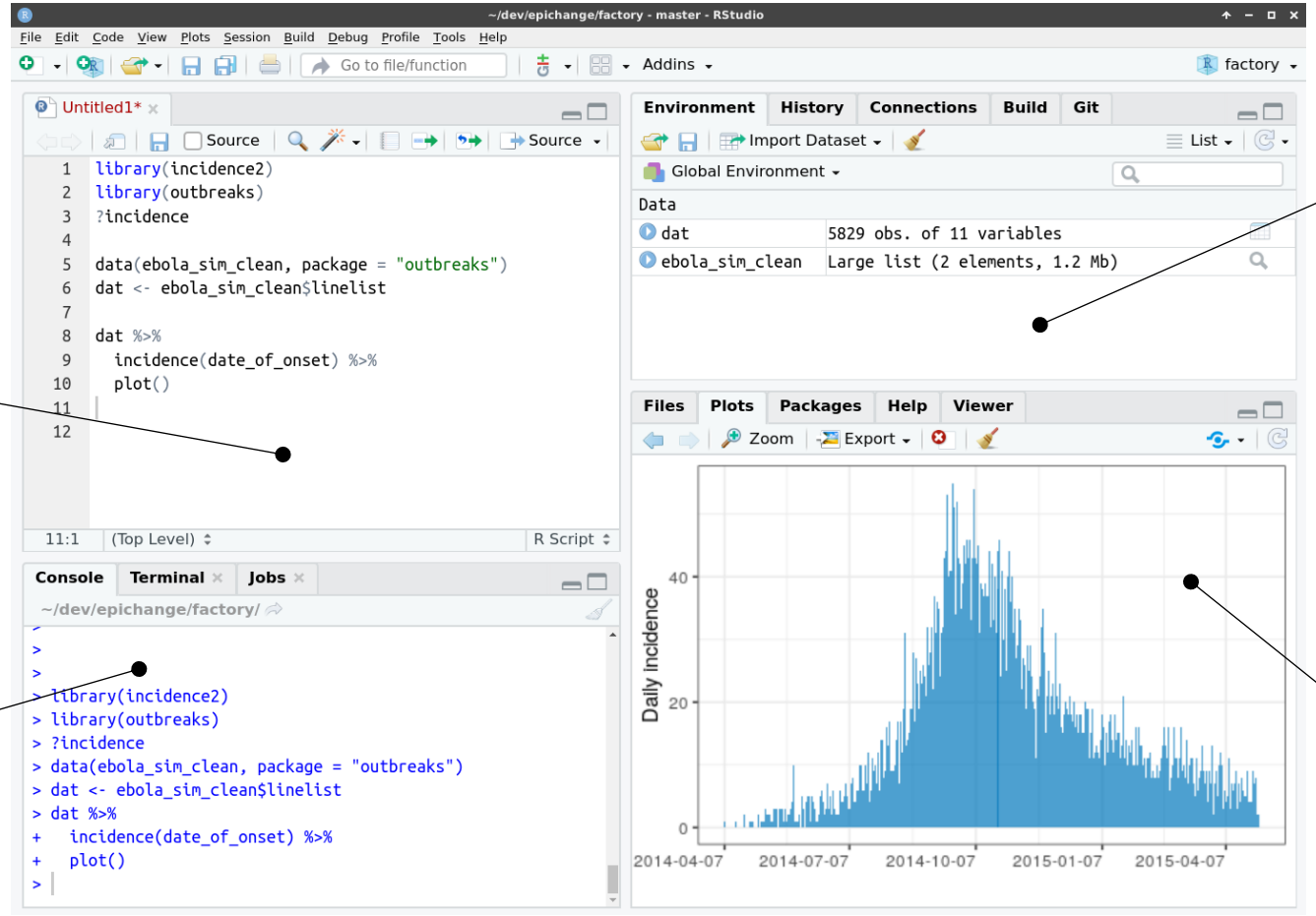




Getting familiar with RStudio

Script file:
write code here

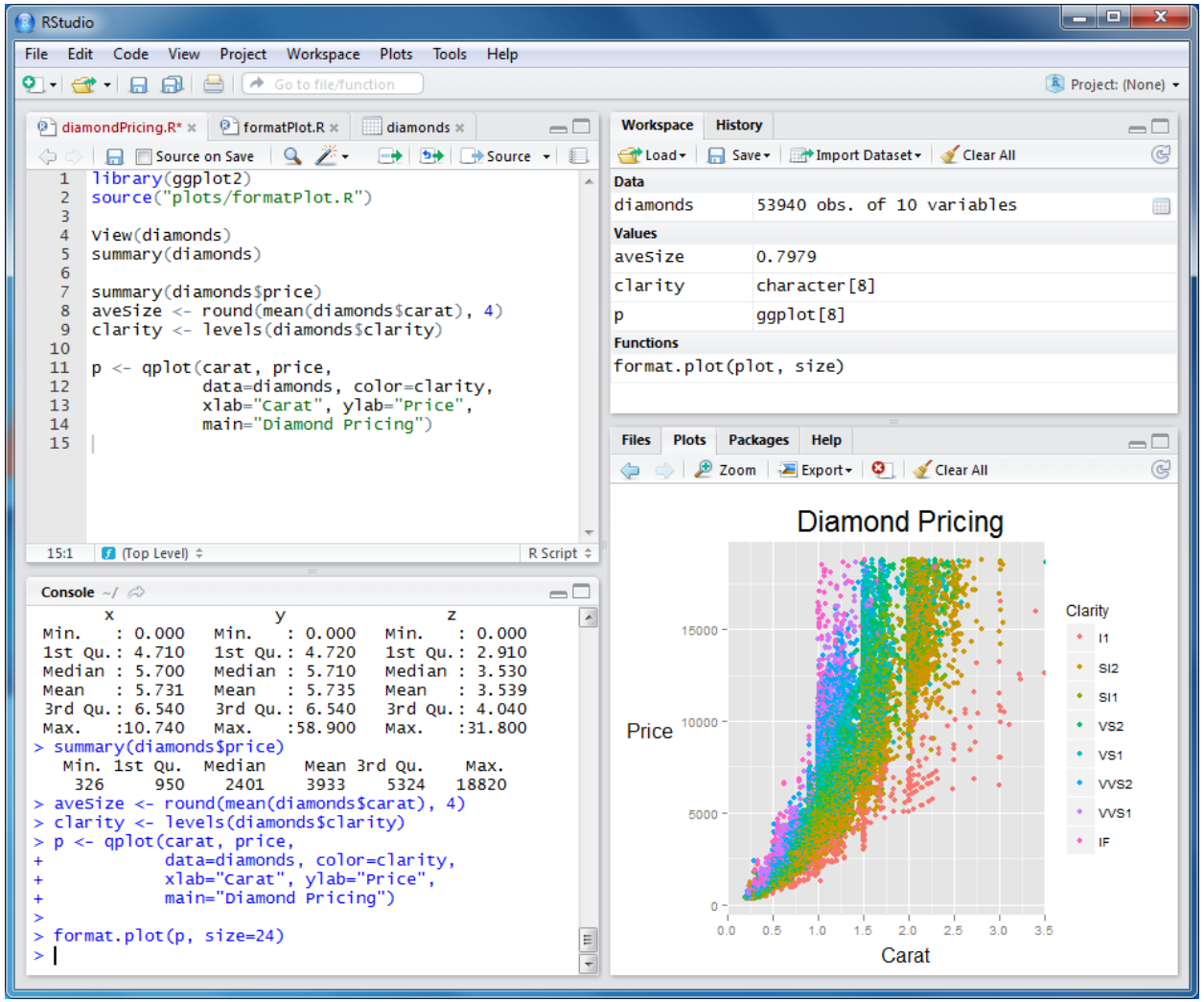
R console:
Evaluate code here



Overview of objects in the session

File browser, Graphics, help, packages loaded

And then…


RStudio
File Edit Code View Project Workspace Plots Tools Help
diamondPricing.R* formatPlot.R diamonds
Source on Save Source
1 library(ggplot2)
2 source("plots/formatPlot.R")
4 view(diamonds)
5 summary(diamonds)
7 summary(diamonds$price)
8 aveSize <- round(mean(diamonds$carat), 4)
9 clarity <- levels(diamonds$clarity)
11 p <- qplot(carat, price,
12 data=diamonds, color=clarity,
13 xlab="Carat", ylab="Price",
14 main="Diamond Pricing")
15:1 (Top Level) R Script
Console ~/
x y z
Min. : 0.000 Min. : 0.000 Min. : 0.000
1st Qu.: 4.710 1st Qu.: 4.720 1st Qu.: 2.910
Median : 5.700 Median : 5.710 Median : 3.530
Mean : 5.731 Mean : 5.735 Mean : 3.539
3rd Qu.: 6.540 3rd Qu.: 6.540 3rd Qu.: 4.040
Max. :10.740 Max. :58.900 Max. :31.800
> summary(diamonds$price)
Min. 1st Qu. Median Mean 3rd Qu. Max.
326 950 2401 3933 5324 18820
> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+ data=diamonds, color=clarity,
+ xlab="Carat", ylab="Price",
+ main="Diamond Pricing")
> format.plot(p, size=24)
Workspace History
Load Save Import Dataset Clear All
Data
diamonds 53940 obs. of 10 variables
Values
aveSize 0.7979
clarity character[8]
p ggplot[8]
Functions
format.plot(plot, size)
Files Plots Packages Help
Zoom Export Clear All
Diamond Pricing
Price
Carat
Clarity
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF

Getting help

- `?foo` or `help("foo")`: access the help page of `foo`
- `??bar` or `help.search("foo")`: look for `foo` in help pages
- dedicated mailing lists: stat.ethz.ch/mailman/listinfo
- the RECON forum: http://www.repidemicsconsortium.org/forum/
- **google**

Someone's done it before!

- Stack **overflow** https://stackoverflow.com/
- Stack **exchange** (methods) https://stats.stackexchange.com/
- R-bloggers (chat, templates, examples) https://www.r-bloggers.com/