Introduction to R

SIBB512: Biochemistry Laboratory
November 29, 2019

Bhoom Suktitipat, MD, PhD

Graduate Program in Medical Bioinformatics

Department of Biochemistry
Faculty of Medicine Siriraj Hospital
Integrative Computational BioScience Center (ICBS)
bhoom.suk@mahidol.edu

Data analysis using R

- Introduction to R
- Installation & Configuration
  - R & R-Studio
  - Install R library
- Reproducible research in R
  - R scripts, Markdown
- Simple R use case
  - Getting data into R
  - Filtering specific information
- Basic exploratory data analysis

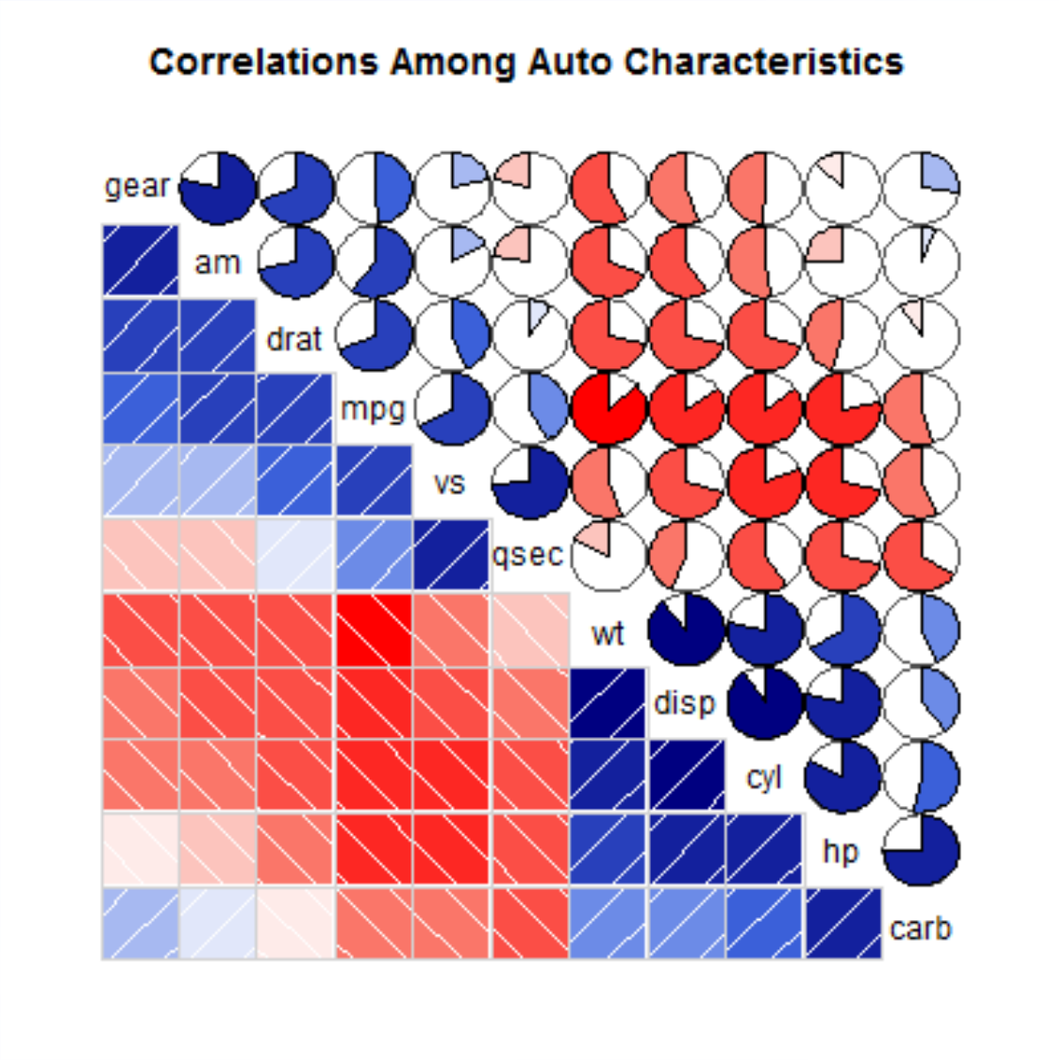
What is R?

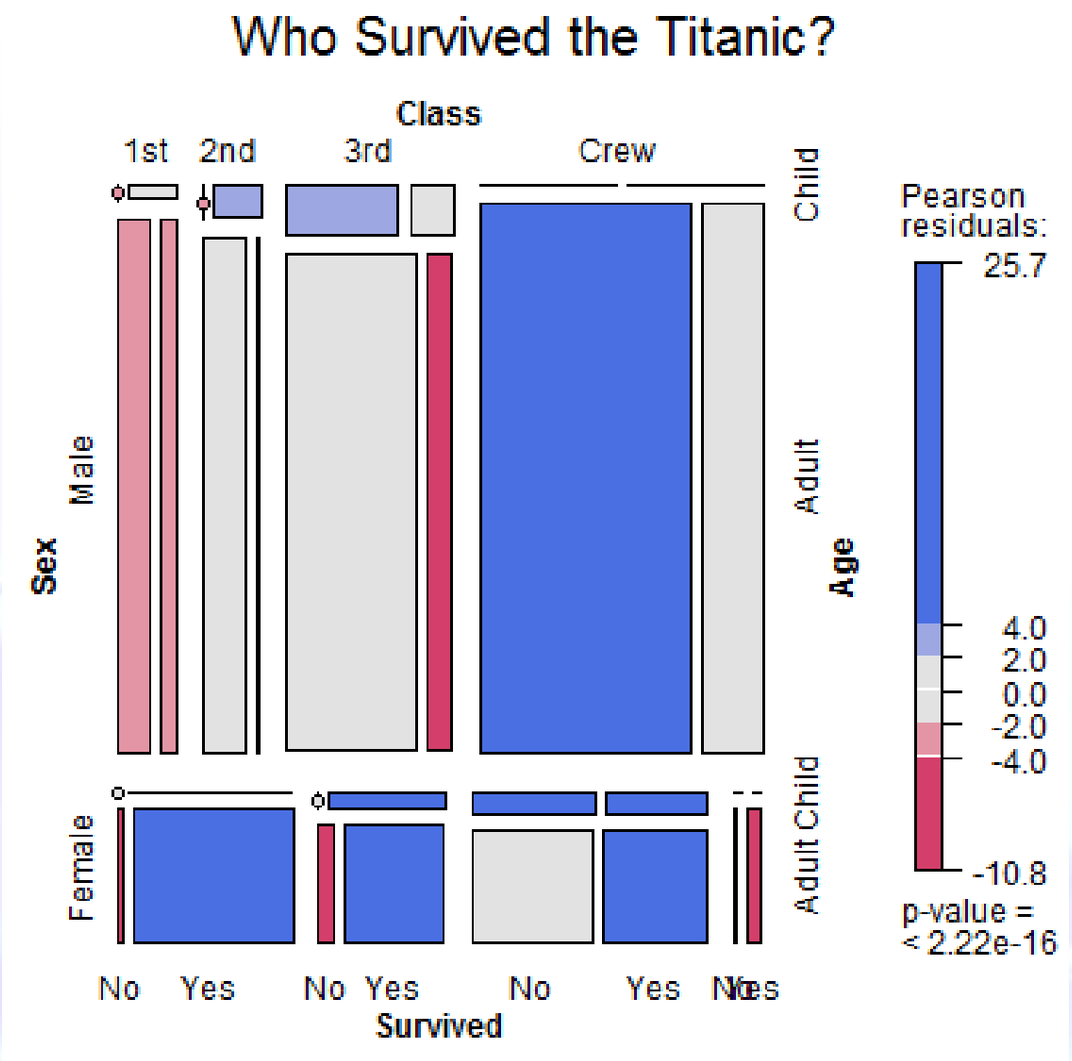
- Comprehensive ***Statistical*** & Graphical Programming Language
- Environment for statistical computing and graphics.
- A successor from “S” language
  - Developed by John M. Chambers, Rick Becker, & Allan Wilks
  - Bell Laboratories since 1976

Why R?

- It's *FREE*!
- Available on various platform
  - Windows, Mac, Linux, ...
- Easy to do **statistical** *programming*
- Advance statistical method libraries
  - High-throughput biological data: http://bioconductor.org
- Actively develop and maintain

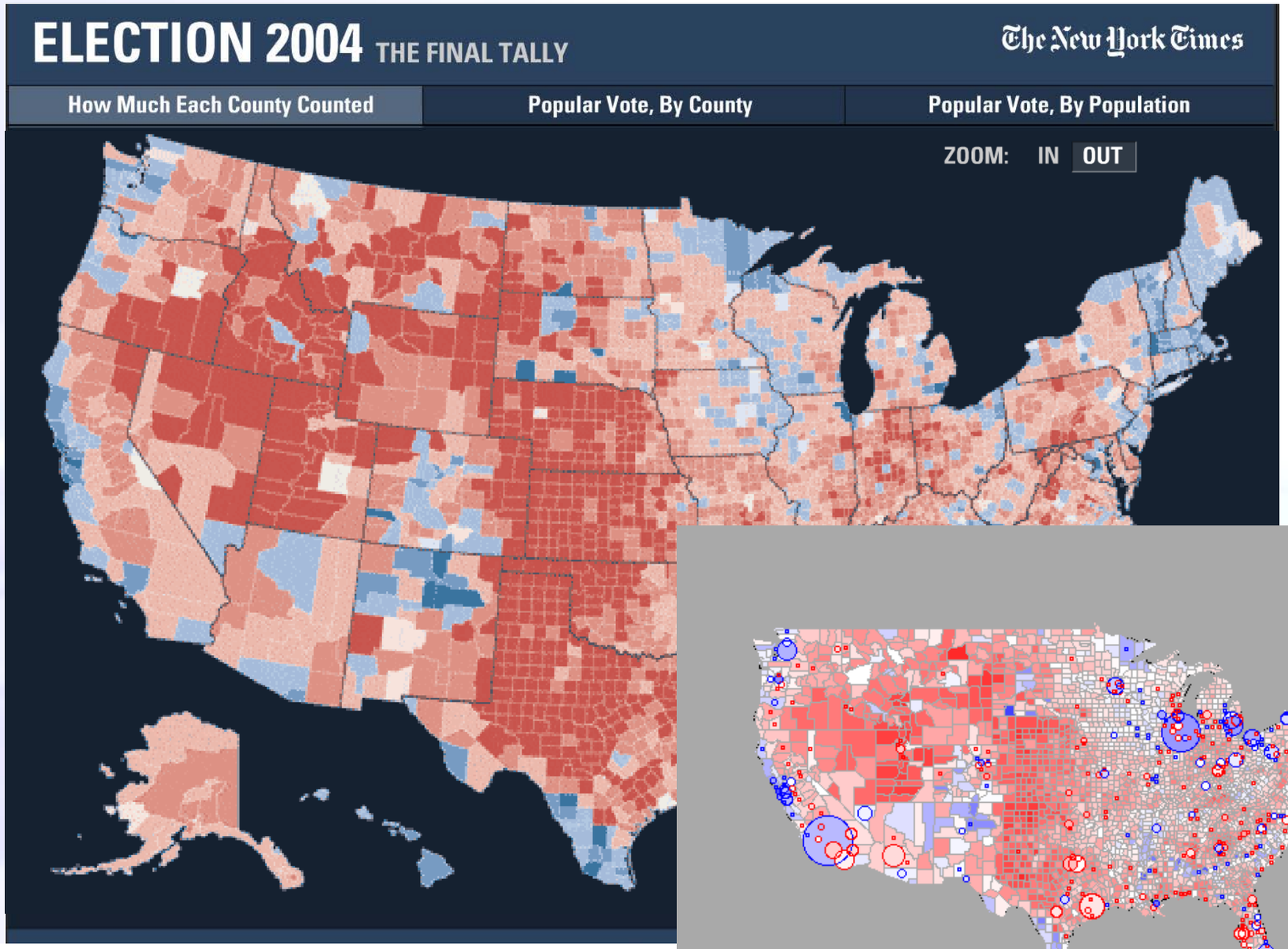
Graphs in R

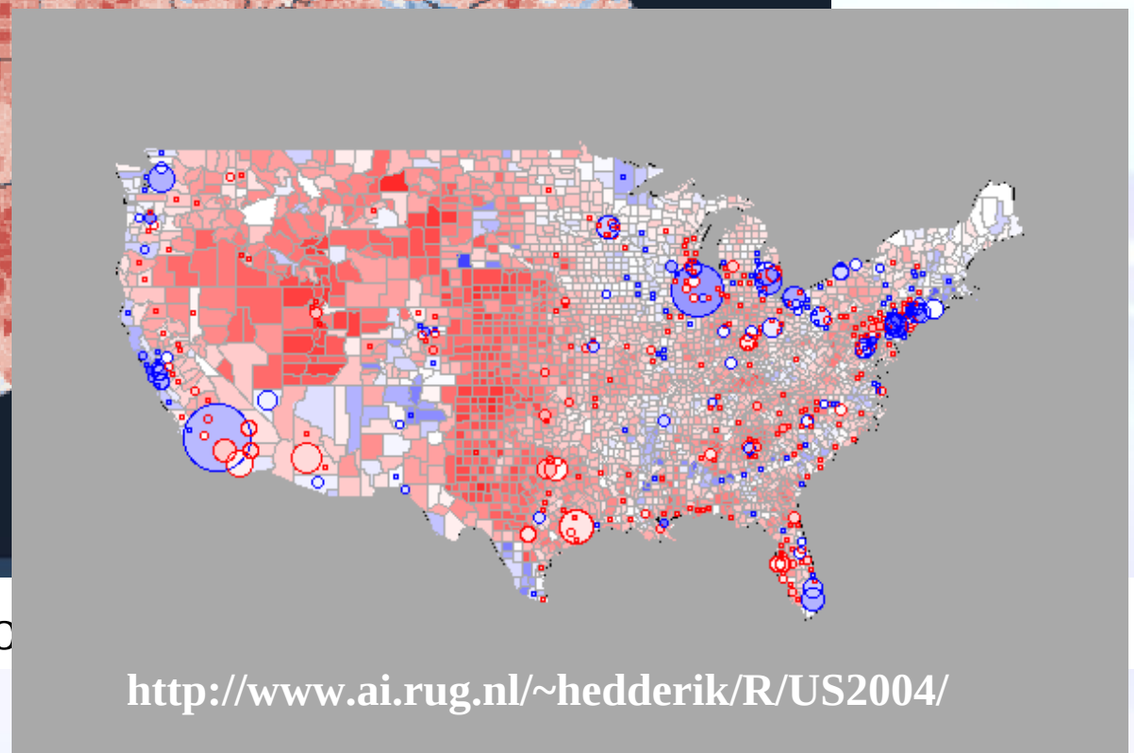




Source: Quick-R.com

State of the art graphics capability





http://www.nytimes.com/packages/khtml/2004/11/03/po

http://www.ai.rug.nl/~hedderik/R/US2004/

Installing R

1) R can be downloaded from http://cran.r-project.org/

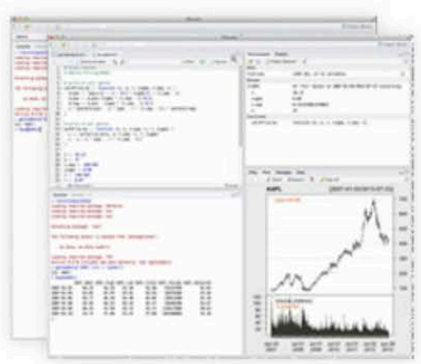
Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

2) R-Studio can be downloaded from https://www.rstudio.com

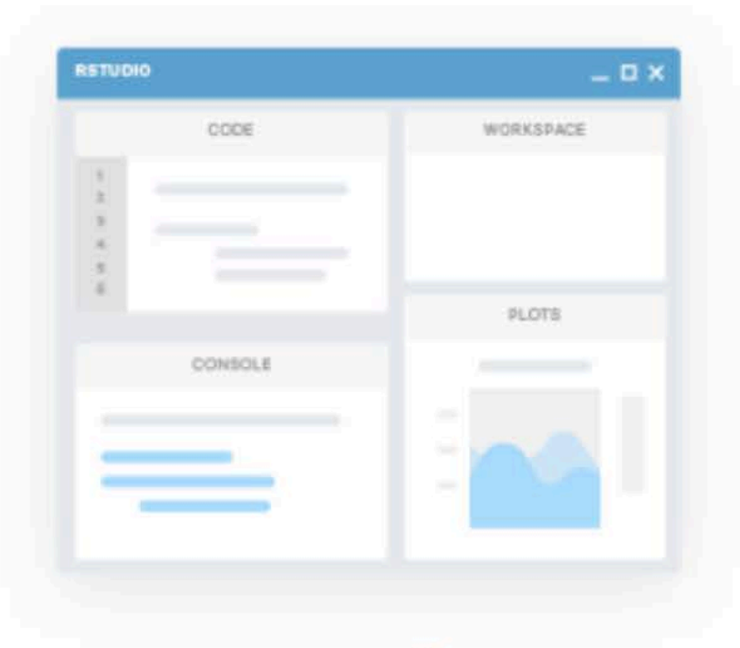


Choose Your Version of RStudio

RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging and managing your workspace. Learn More about RStudio features.

RStudio Desktop Open Source License	RStudio Desktop Commercial License	RStudio Server Open Source License	RStudio Server Pro Commercial License	RStudio Server Pro + RStudio Connect Commercial License
FREE	$995 per year	FREE	$9,995 per year	$29,995 per year

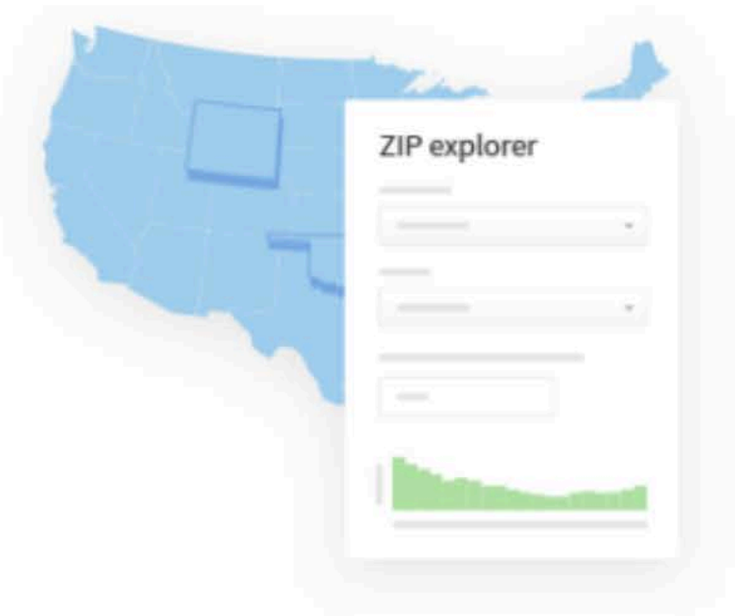
RStudio https://www.rstudio.com



RStudio

RStudio makes R easier to use. It includes a code editor, debugging & visualization tools.

Download Learn More



Shiny

Shiny helps you make interactive web applications for visualizing data. Bring R data analysis to life.

Learn More

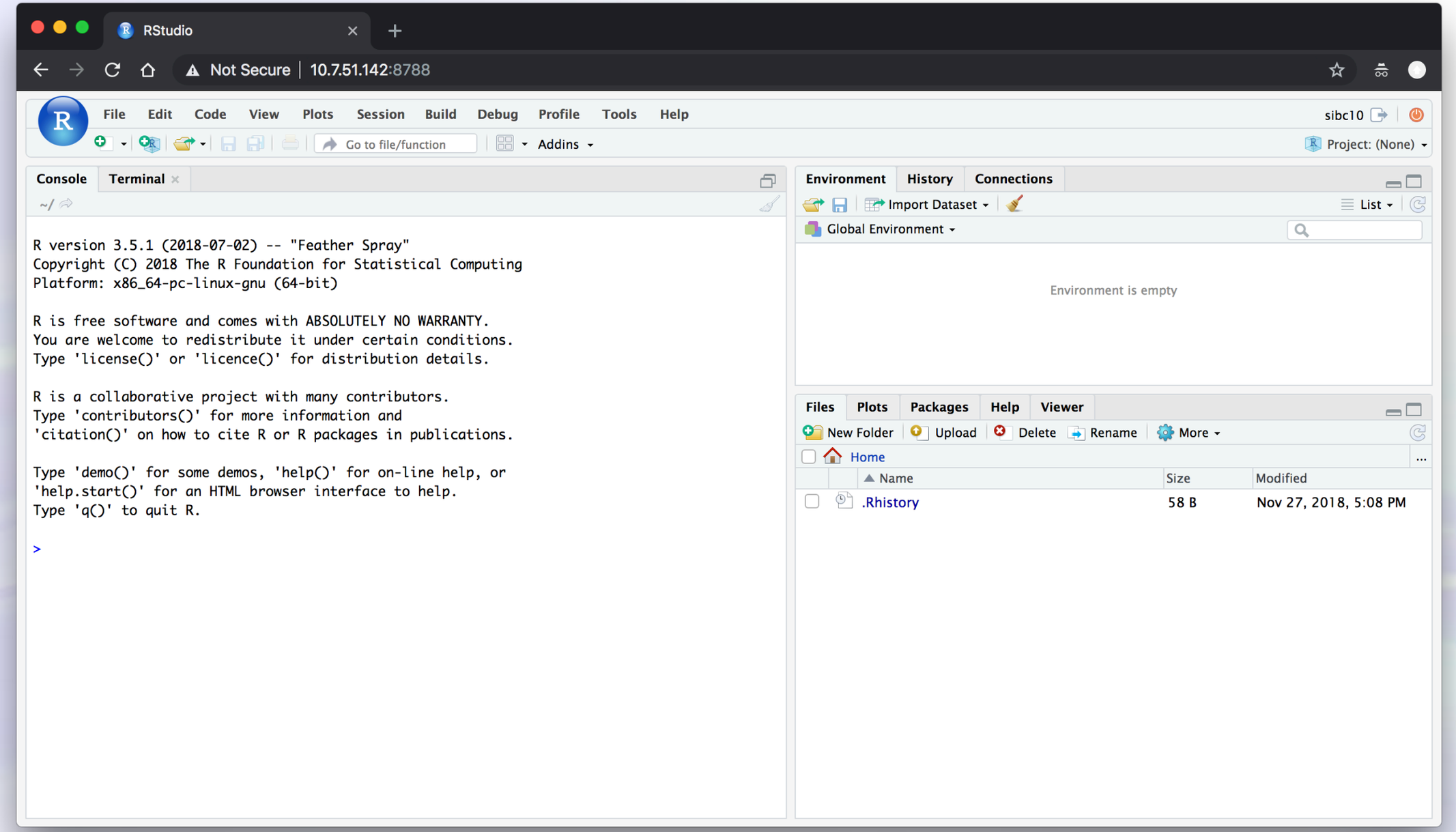
R Packages

Our developers create popular packages to expand the features of R. Includes ggplot2, dplyr, R Markdown & more.

Learn More

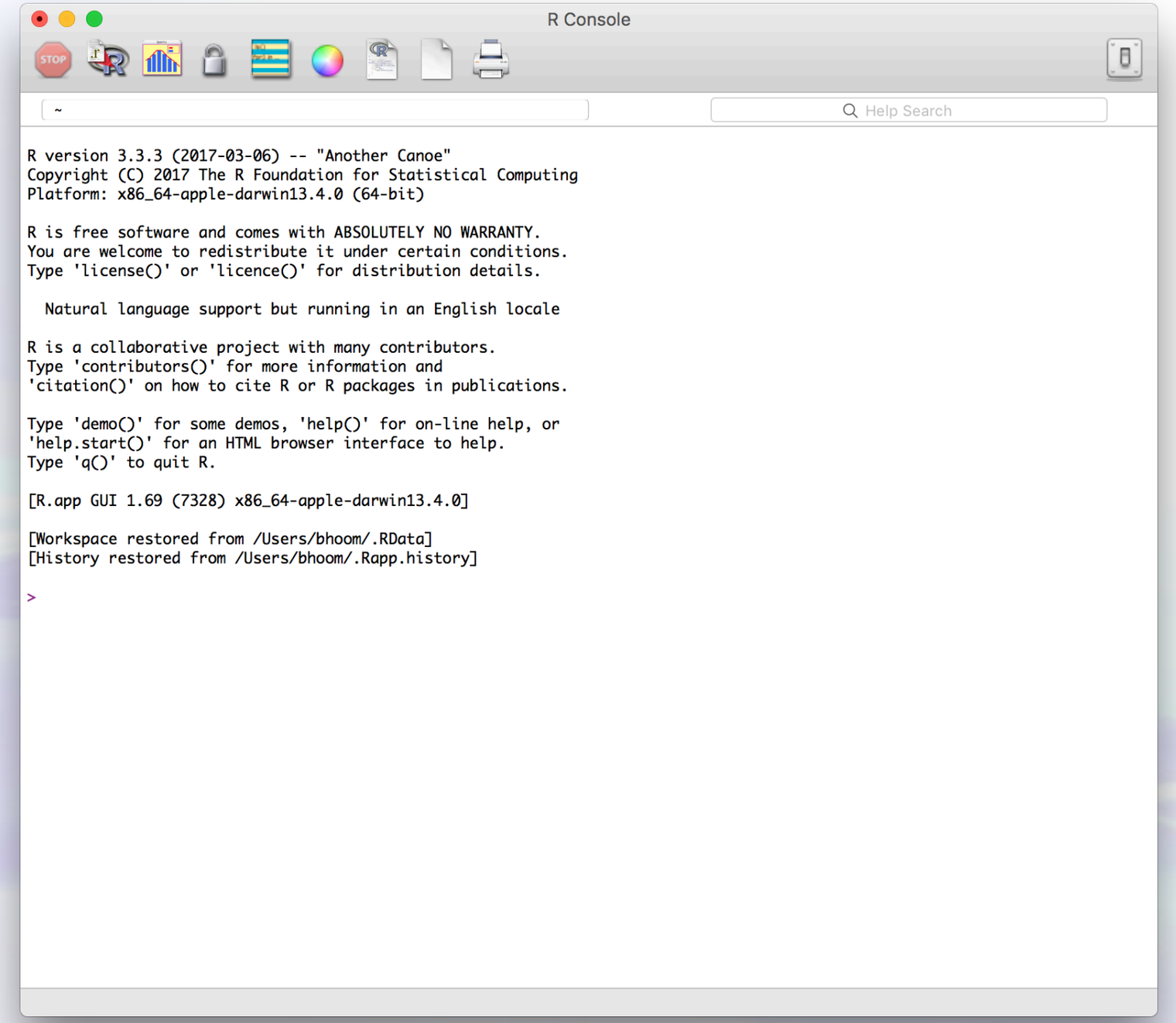
RStudio Cloud: https://rstudio.cloud

- Working with R-Studio through your browser



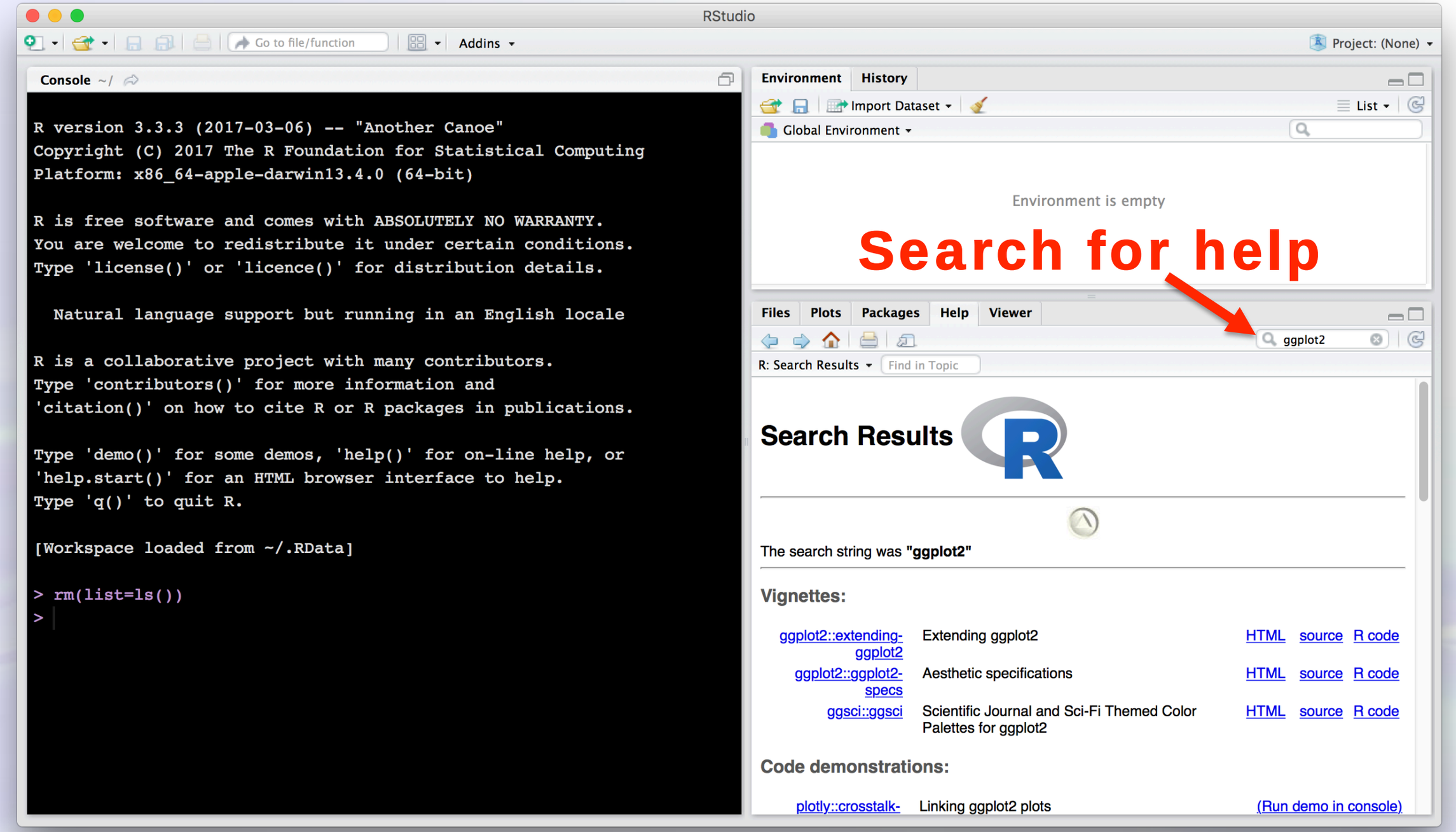
Starting R

R Console

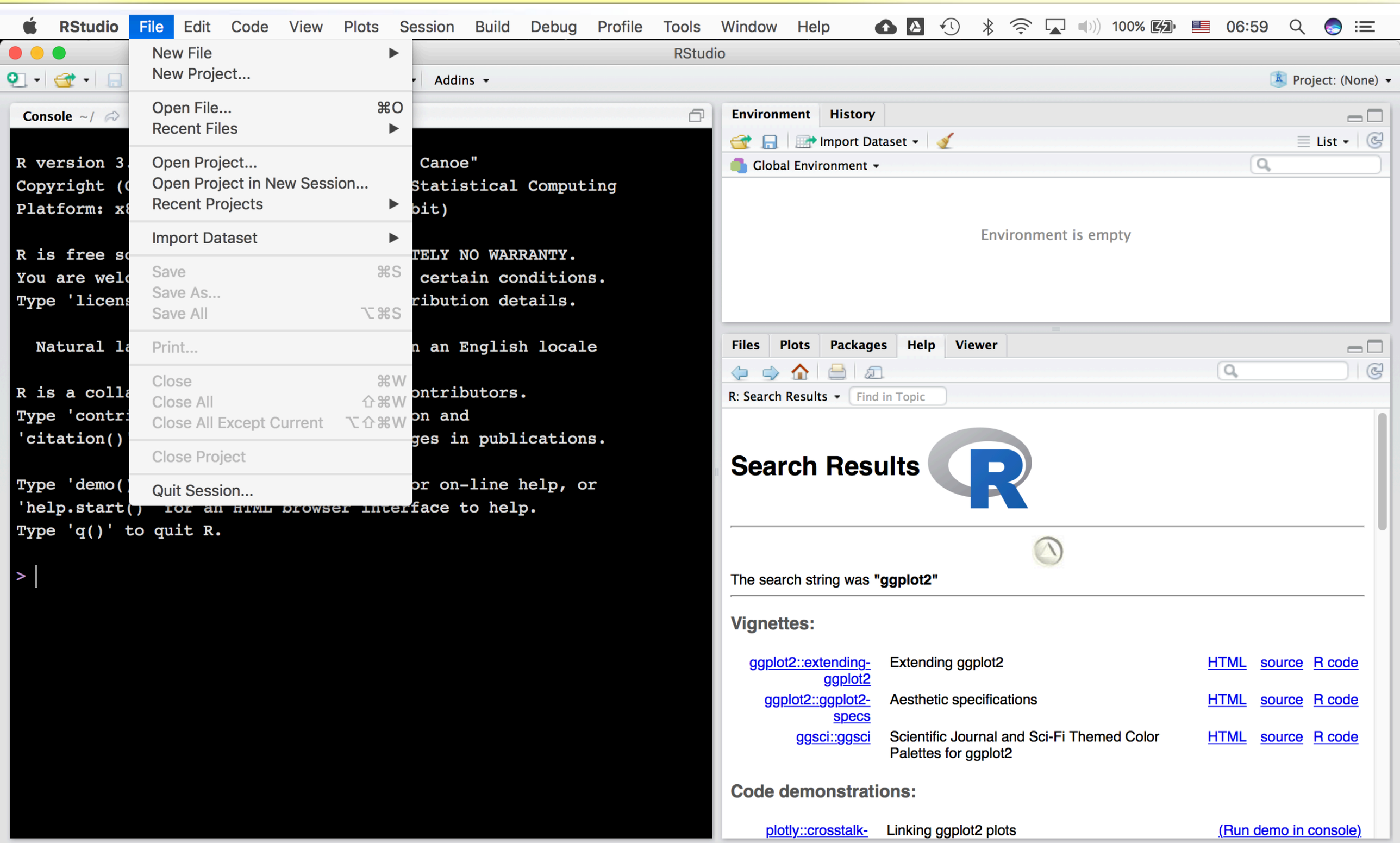


Starting R

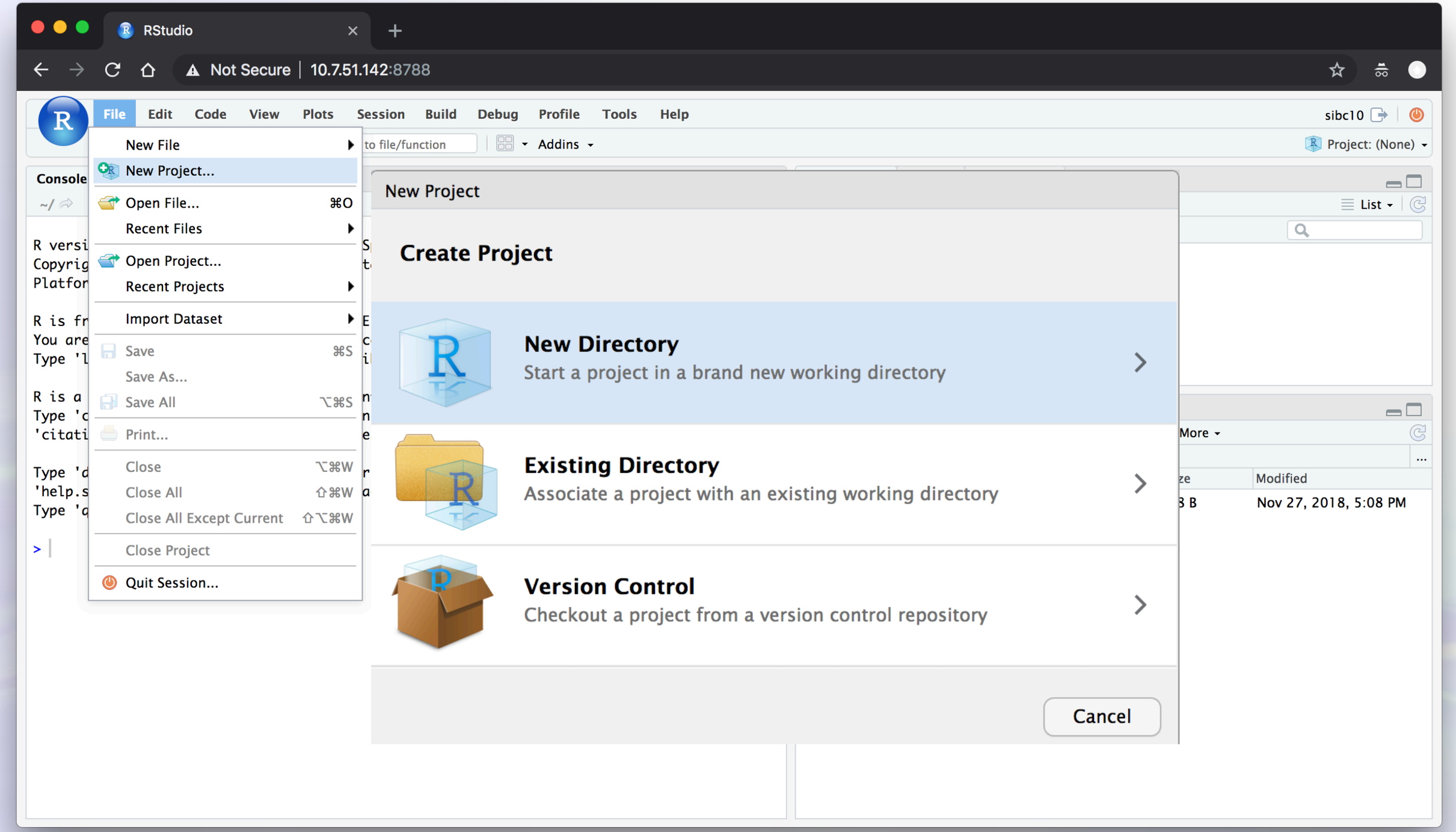
RStudio Console



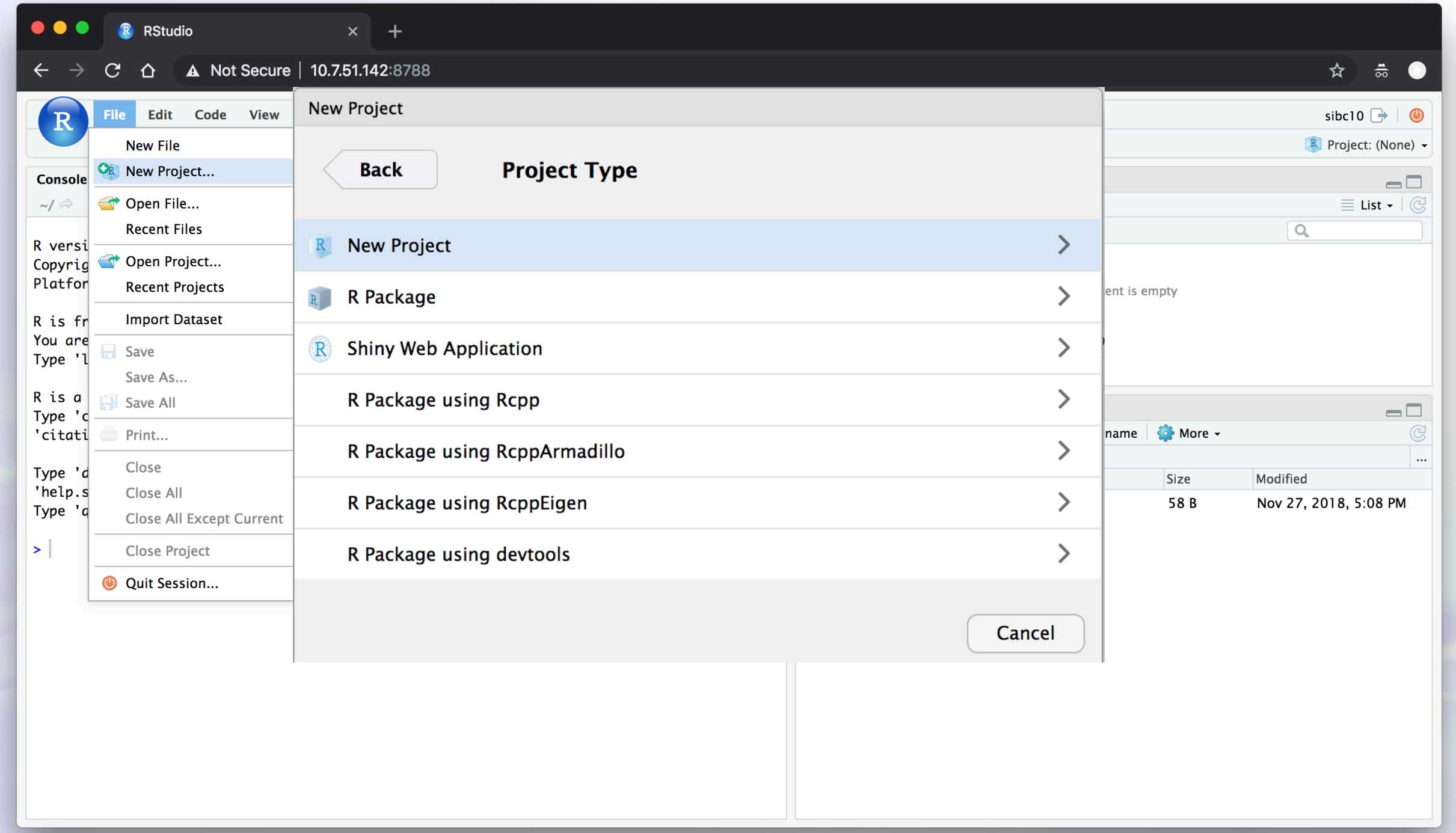
Create a Project



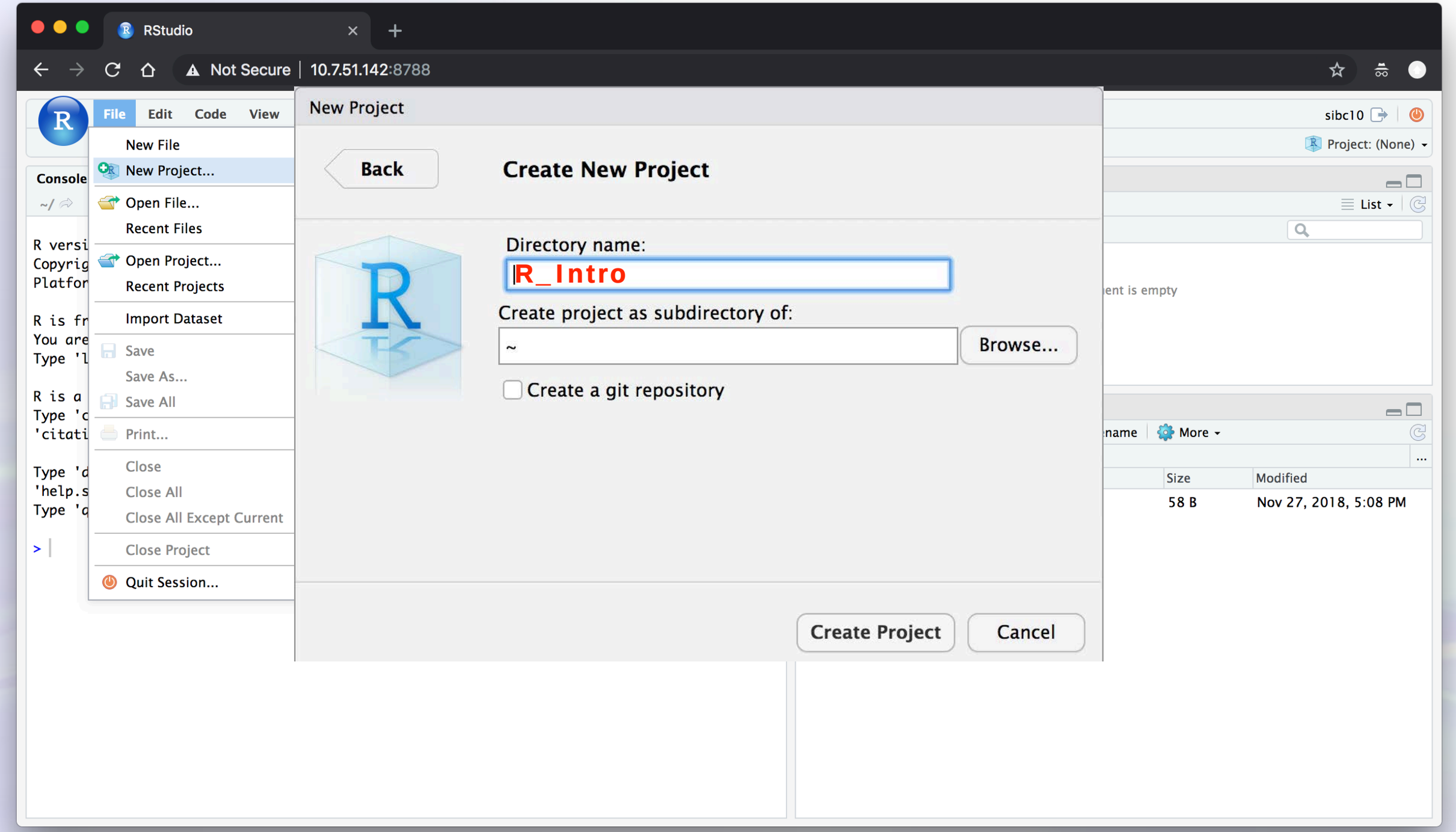
Create a Project



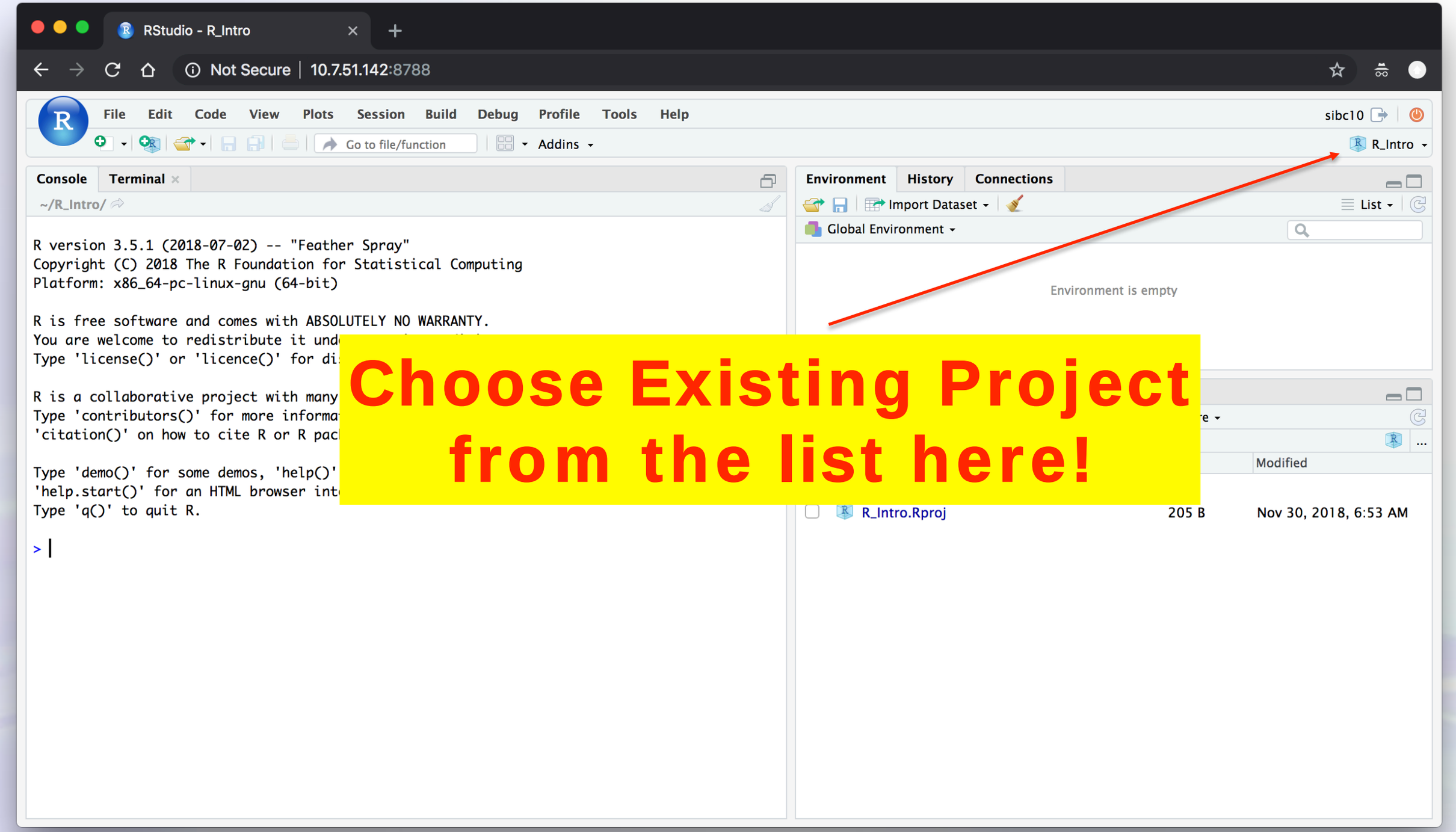
Create a Project



Create a Project



Create a Project



R & Proxy Server Authentication

.Renviron File

```
http_proxy=http://username:password@proxy-si.mahidol:8080/
```

Save .Renviron in your home directory
For Windows, this is probably C:/Users/[Username]/Documents

Command Line in R

```
> Sys.setenv(http_proxy="http://username:password@proxy-si.mahidol:8080")
## to check
> Sys.getenv("http_proxy")
```

Not permanent; have to issue this command every session.
Capital "S" at the beginning of "Sys.setenv"

*replace "username" and "password" with your own internet account username/password

Workspace

- Working directory
  - `getwd()`
  - `setwd()`
- Objects
  - `ls()`
  - `rm(list=ls())`
- Library

Session Info

```
> sessionInfo()
R version 3.5.1 (2018-07-02)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Debian GNU/Linux 9 (stretch)

Matrix products: default
BLAS: /usr/lib/openblas-base/libblas.so.3
LAPACK: /usr/lib/libopenblasp-r0.2.19.so

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C                   LC_TIME=en_US.UTF-8
 [4] LC_COLLATE=en_US.UTF-8     LC_MONETARY=en_US.UTF-8        LC_MESSAGES=C
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C                      LC_ADDRESS=C
[10] LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

loaded via a namespace (and not attached):
[1] compiler_3.5.1 tools_3.5.1    yaml_2.2.0
>
```

Simple use of R

- Calculator

Symbol	Operation
a + b	addition
a - b	subtraction
a * b	multiplication
a / b	division
a ^ b	exponentiation

Exercise: Calculate the percentage of your life spent at this university.
Hint: subtract the year you started at Mahidol University from 2017, divide by the difference between 2017 and the year you were born, multiply by 100.

Common R commands

- `ls()` # list all objects in workspace
- `rm(list=ls())` # remove all objects
- `q()` # quit
- `setwd("a:/my_dir")` # set working directory
- `install.packages("gee")` #install library
- `library("gee")` # load library
- `help("rnorm")` # Getting Help
- `help.search("keyword")`

Install Packages

- **Packages or Libraries:** collections of tools & functions in R
  - `library()` # show installed libraries
  - `install.packages(“ggplot2”)` # install a graphic library: ggplot2
- **Using bioconductor as package manager**
  - `source(“https://bioconductor.org/biocLite.R")`
  - `biocLite("ggplot2")`

About *Bioconductor*

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 1104 software packages, and an active user community. Bioconductor is also available as an AMI (Amazon Machine Image) and a series of Docker images.

News

- Bioconductor 3.2 is available.
- Bioconductor F1000 Research Channel launched.
- Orchestrating high-throughput genomic analysis with *Bioconductor* (abstract) and other recent literature.
- Read our latest newsletter and course material.
- Use the support site to get help installing, learning and using Bioconductor.

Install »

Get started with *Bioconductor*

- Install *Bioconductor*
- Explore packages
- Get support
- Latest newsletter
- Follow us on twitter
- Install R

Learn »

Master *Bioconductor* tools

- Courses
- Support site
- Package vignettes
- Literature citations
- Common work flows
- FAQ
- Community resources
- Videos

Use »

Create bioinformatic solutions with *Bioconductor*

- Software, Annotation, and Experiment packages
- Amazon Machine Image
- Latest release annoucement
- Support site

Develop »

Contribute to *Bioconductor*

- Use Bioc 'devel'
- 'Devel' Software, Annotation and Experiment packages
- Package guidelines
- New package submission
- Developer resources
- Build reports

Getting Help

```
> help.start()
```

The R Language

Statistical Data Analysis

Manuals

An Introduction to R
Writing R Extensions
R Data Import/Export
The R Language Definition
R Installation and Administration
R Internals

Reference

Packages
Search Engine & Keywords

Miscellaneous Material

About R
License
NEWS
Authors
Frequently Asked Questions
User Manuals
Resources
Thanks
Technical papers

An Introduction to R

Table of Contents

- An Introduction to R
- Preface
- 1 Introduction and preliminaries
  - 1.1 The R environment
  - 1.2 Related software and documentation
  - 1.3 R and statistics
  - 1.4 R and the window system
  - 1.5 Using R interactively
  - 1.6 An introductory session
  - 1.7 Getting help with functions and features
  - 1.8 R commands, case sensitivity, etc.
  - 1.9 Recall and correction of previous commands
  - 1.10 Executing commands from or diverting output to a file
  - 1.11 Data permanency and removing objects
- 2 Simple manipulations; numbers and vectors
  - 2.1 Vectors and assignment
  - 2.2 Vector arithmetic
  - 2.3 Generating regular sequences
  - 2.4 Logical vectors
  - 2.5 Missing values
  - 2.6 Character vectors
  - 2.7 Index vectors; selecting and modifying subsets of a data set
  - 2.8 Other types of objects
- 3 Objects, their modes and attributes
  - 3.1 Intrinsic attributes: mode and length
  - 3.2 Changing the length of an object
  - 3.3 Getting and setting attributes
  - 3.4 The class of an object

Getting Help

```
> help("t.test")
```

t.test {stats} **R Documentation**

Student's t-Test

Description

Performs one and two sample t-tests on vectors of data.

Usage

```
t.test(x, ...)

## Default S3 method:
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)

## S3 method for class 'formula'
t.test(formula, data, subset, na.action, ...)
```

Arguments

`x`	a (non-empty) numeric vector of data values.
`y`	an optional (non-empty) numeric vector of data values.
`alternative`	a character string specifying the alternative hypothesis, must be one of `"two.sided"` (default), `"greater"` or `"less"`. You can specify just the initial letter.
`mu`	a number indicating the true value of the mean (or difference in means if you are performing a two sample test).
`paired`	a logical indicating whether you want a paired t-test.
`var.equal`	a logical variable indicating whether to treat the two variances as being equal. If `TRUE` then the pooled variance is used to estimate the variance otherwise the Welch (or Satterthwaite) approximation to the degrees of freedom is used.
`conf.level`	confidence level of the interval.
`formula`	a formula of the form `lhs ~ rhs` where `lhs` is a numeric variable giving the data values and `rhs` a factor with two levels giving the corresponding groups.

Examples

```
require(graphics)

t.test(1:10, y = c(7:20))      # P = .00001855
t.test(1:10, y = c(7:20, 200)) # P = .1245    -- NOT significant anymore

## Classical example: Student's sleep data
plot(extra ~ group, data = sleep)
## Traditional interface
with(sleep, t.test(extra[group == 1], extra[group == 2]))
## Formula interface
t.test(extra ~ group, data = sleep)
```

More help to learn R

- Quick-R: http://www.statmethods.net/
- A short reference card: http://cran.r-project.org/doc/contrib/ Short-refcard.pdf
- very rich source of examples: http://zoonek2.free.fr/UNIX/48_R/all.html
- An R-wiki
  http://rwiki.sciviews.org/doku.php
- R for MatLab users:
  http://mathesaurus.sourceforge.net/
- Google “R” and something…

how sidebar

Home | Interface | Input | Manage | Stats | Adv Stats | Graphs | Adv Graphs | Blog

Quick-R

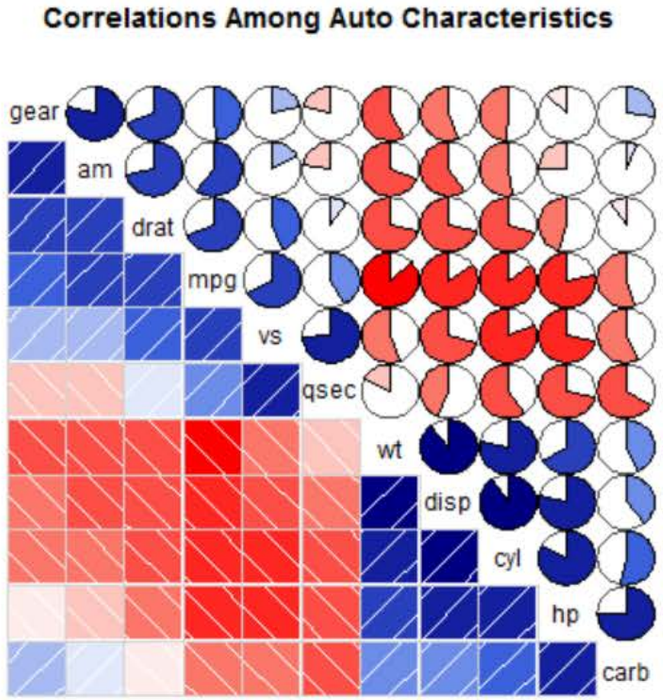
accessing the power of R

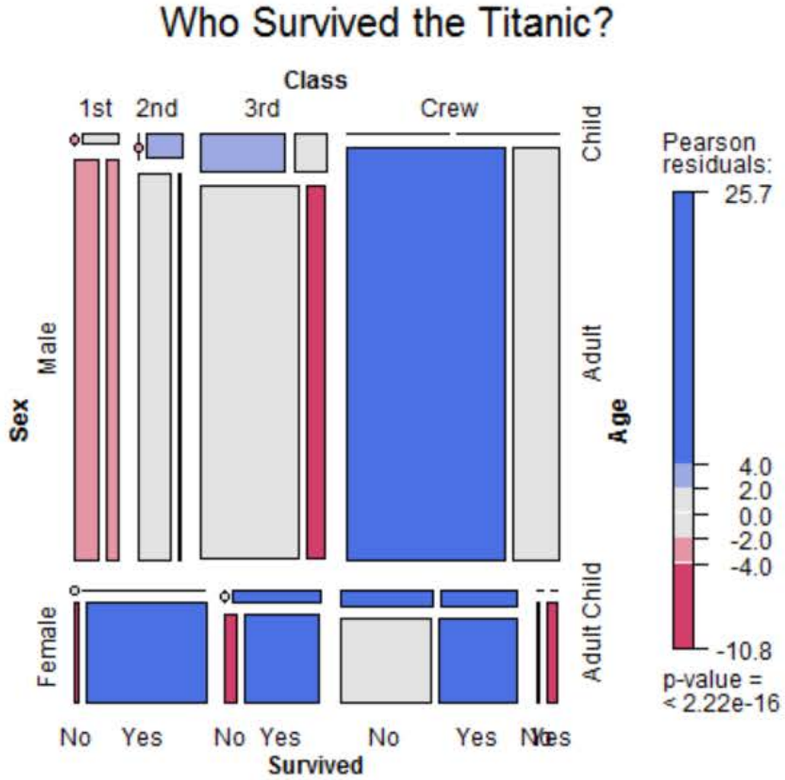
Search

Top Menu

- Home
- The R Interface
- Data Input
- Data Management
- Basic Statistics
- Advanced Statistics
- Basic Graphs
- Advanced Graphs
- Blog

About Quick-R





R Objects

- Scalar
- Vector
- Matrix
- Function

```
plus <- function(a,b) {
  c = a+b
  return(c)
}

plus(10,1)
```

```
[1] 11
```

Reproducible Analyses & Scripts

- **Script**: a set of instruction for the **R**
- Allow us to keep track of what we do
- Allow us to replicate our analyses

```
# R Script Example
# Author: Bhoom Suktitipat
# Date: 2014 Nov 24

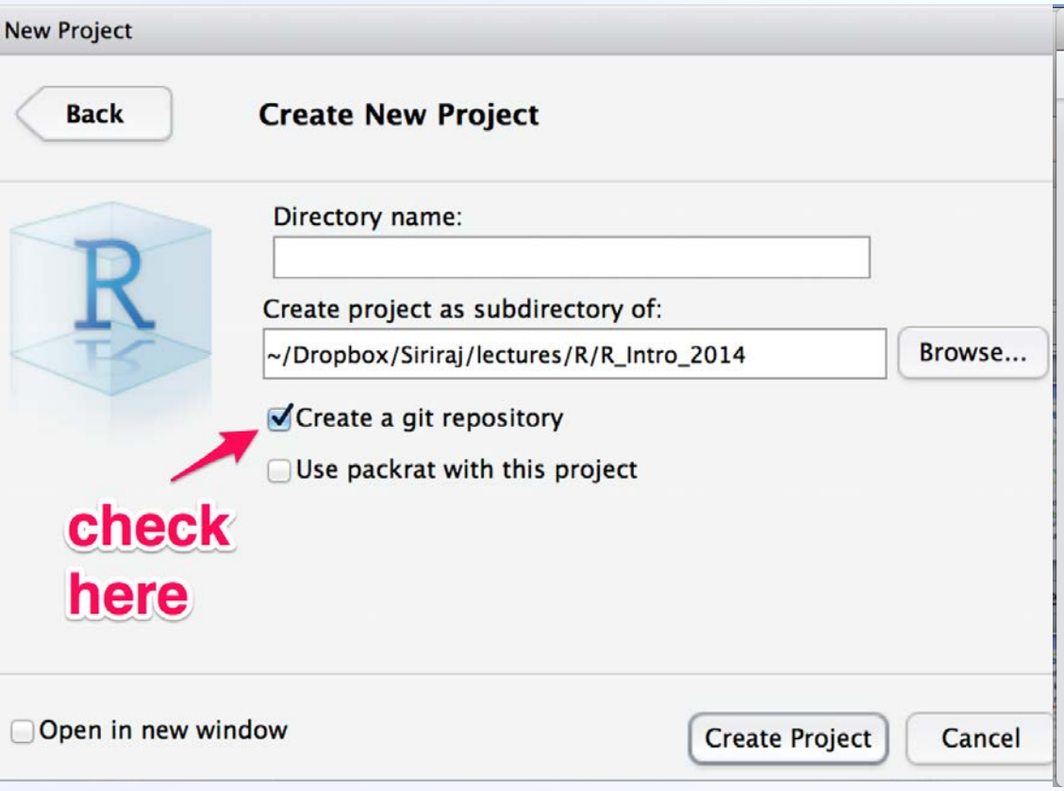
set.seed(12345)
# 1) Simulate 1000 random numbers
x <- runif(1000)
y <- runif(1000)

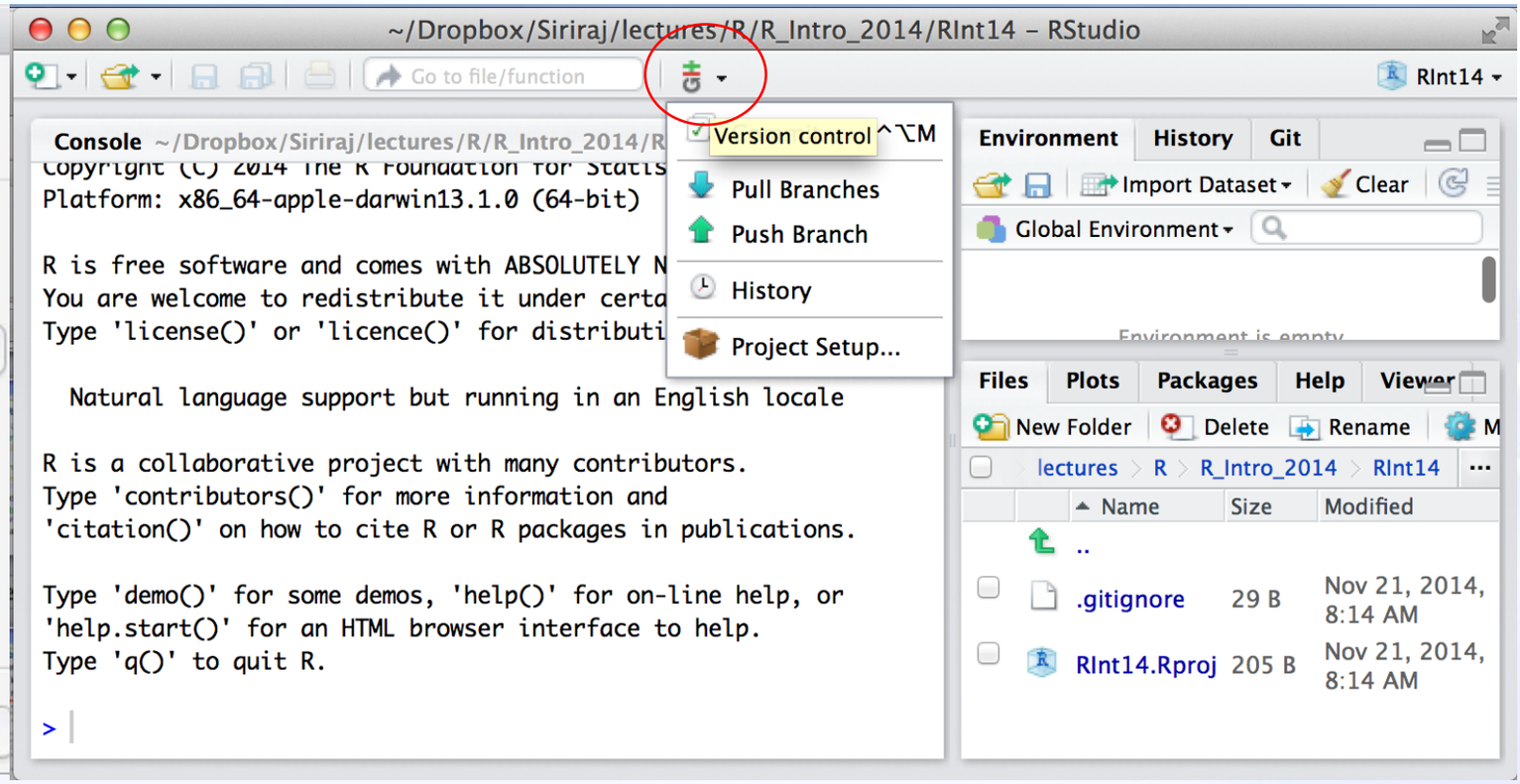
# 2) Calculate a summary statistics of “x”
summary(x)

# 3) Calculate Pearson’s correlation
cor(x,y,method=“pearson”)
```

Version Control

- The Rstudio-way
  - File > New Project > New Directory > Empty Project “Click Create git repository”
  - File > New File > R Script “Edit”
  - ‘Git’ icon on menu bar > “Commit”

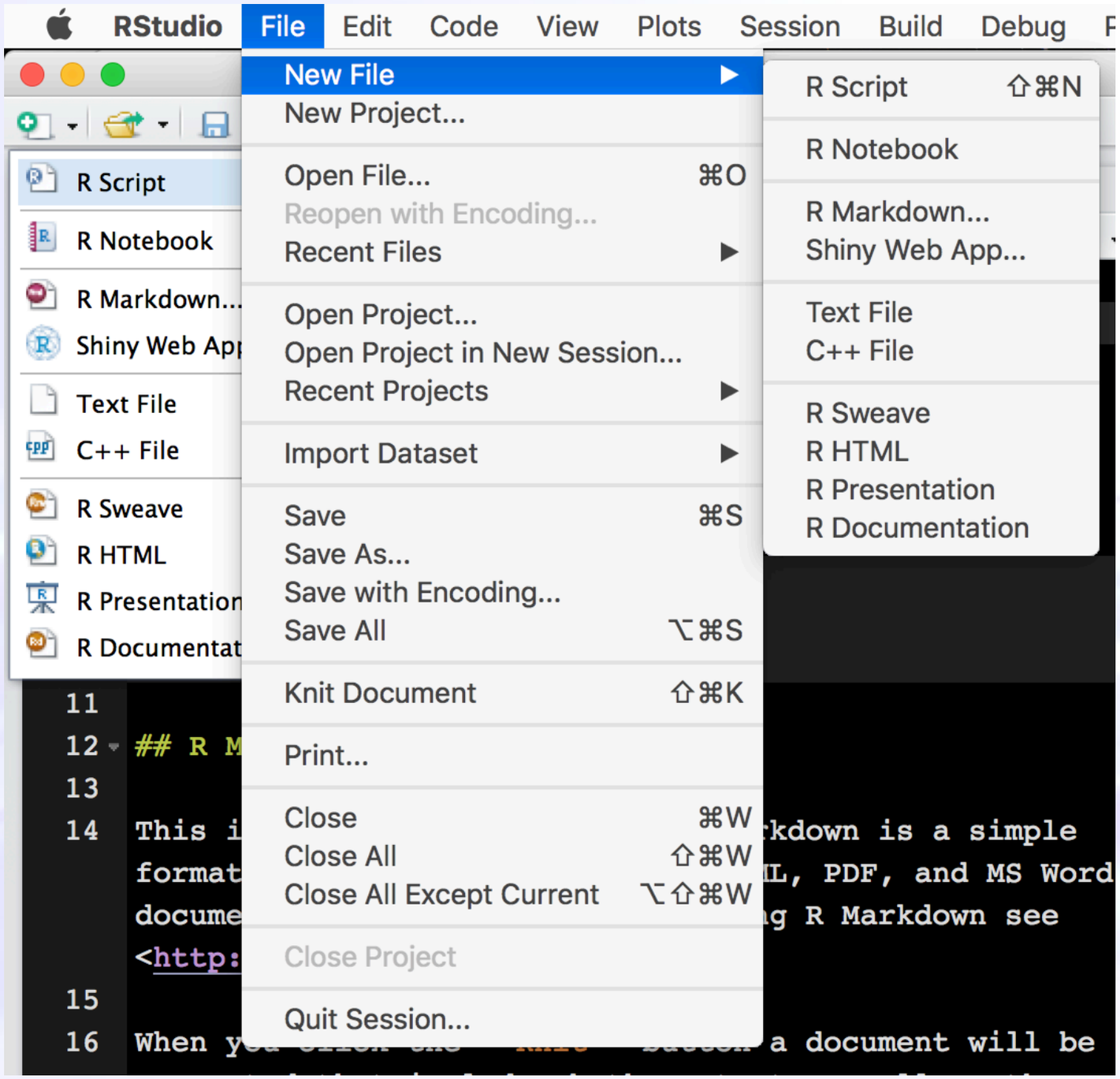




R Script

- Modern ways to keep a record (html, pdf, md, etc)
  - knitr: http://yihui.name/knitr/
  - R markdown: http://www.rstudio.com/ide/docs/authoring/using_markdown
  - *LaTeX*
  - *html5 presentation* *http://slides.html5rocks.com/*

Reproducible Analyses & Scripts



Option 1 — Menu File "New File"

Option 2 — Menu Icon "New File"

R Markdown from RStudio

Introduction

How It Works

Code Chunks

Inline Code

Code Languages

Parameters

Tables

Markdown Basics

Output Formats

Notebooks

Slide Presentations

Dashboards

Websites

Interactive Documents

R Markdown and R Notebook

http://rmarkdown.rstudio.com/lesson-1.html

- Allows us to create document with integrated R code
- render() or *knit* into various document formats
  - pdf, html, word, books
  - http://rmarkdown.rstudio.com/lesson-9.html

Figures

Margin Figures

Images and graphics play an integral role in Tufte's work. To place figures or tables in the margin you can use the `fig.margin` knitr chunk option. For example:

```
library(ggplot2)
qplot(Sepal.Length, Petal.Length, data = iris,
    color = Species)
```

Note the use of the `fig.cap` chunk option to provide a figure caption. You can adjust the proportions of figures using the `fig.width` and `fig.height` chunk options. These are specified in inches, and will be automatically scaled down to fit within the handout margin.

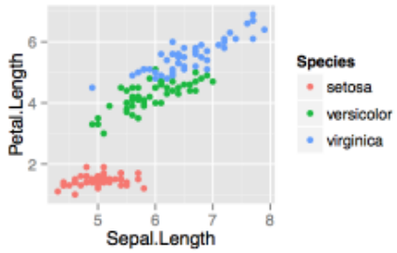


Figure 1: Sepal length vs. petal length, colored by species

Equations

You can also include LaTeX equations in the margin by explicitly invoking the `marginfigure` environment.

Note the use of the `\caption` command to add additional text below the equation.

$$\frac{d}{dx}\left(\int_0^x f(u)\,du\right) = f(x).$$

Figure 2: An equation

Markdown Quick Reference

- Emphasize
  - *italic*
  - **bold**
- Headers
  - # Header 1
  - ## Header 2
  - ### Header 3

- Lists

  * Item 1
  * Item 2
    + Item 2a
    + Item 2b

- Ordered Lists

  1. Item 1
  2. Item 2
     + Item 2a
     + Item 2b

- Links: **plain html link**
- Line breaks: **two or more spaces at line ending**
- Images: **![alt text](figures/img.png)**
- Blockquotes: **> Blockquote text**
- Superscript (x^y): **x^y^**
- ~~strikethrough~~: **~~strikethrough~~**
- Assigned bookmark: **[Bookmark name][id]**
- Refer to bookmark: **[id] description later on**

Markdown Quick Reference

- R Blockquotes

```
```{r}
summary(cars$dist)
summary(cars$speed)
```
```

- Inline Code

```
`r nrow(cars)` cars studied
```

- Plain code block

```
```
This text is displayed verbatim
```
```

- Tables

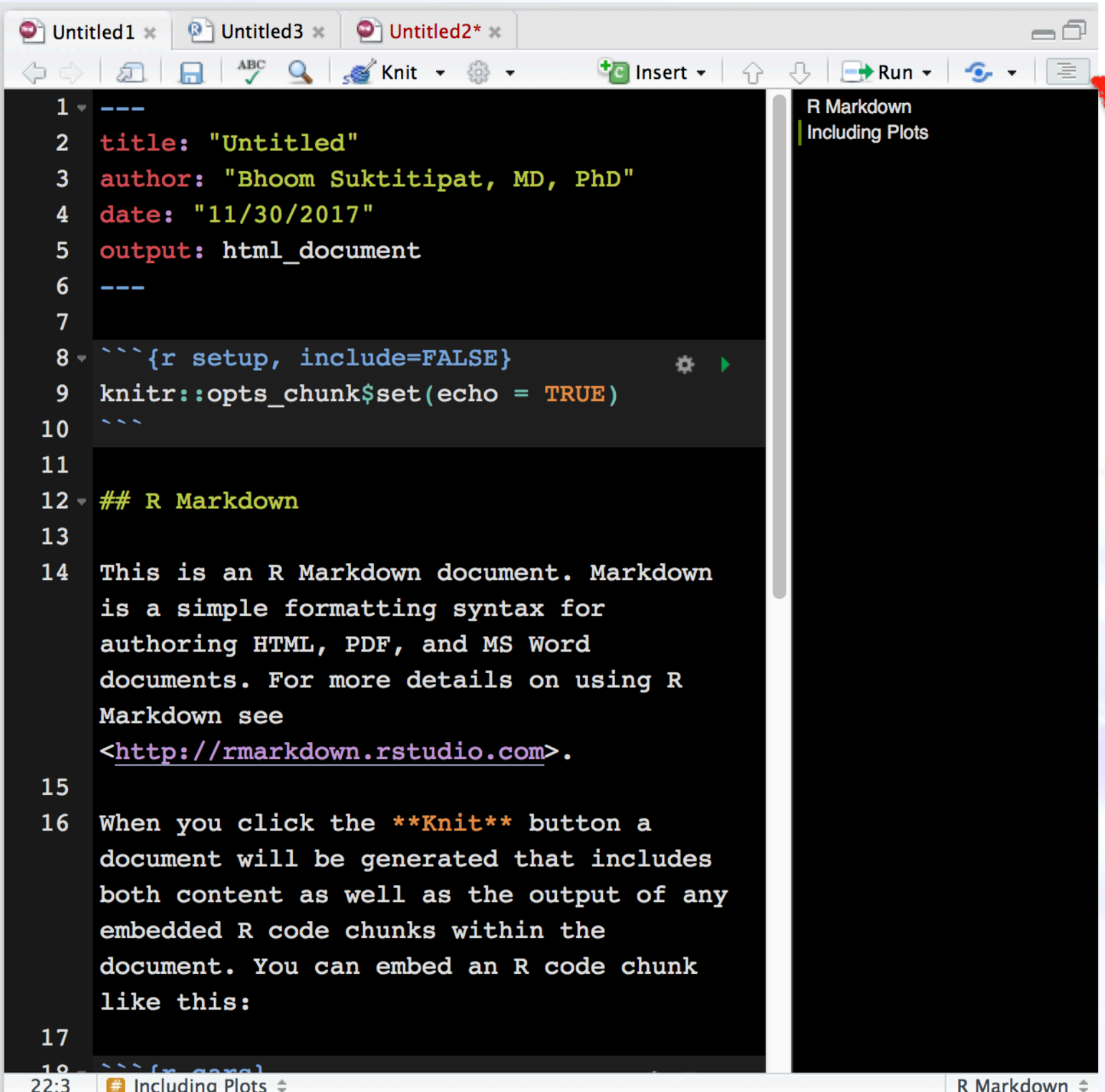
```
First Header  | Second Header
------------- | -------------
Content Cell  | Content Cell
Content Cell  | Content Cell
```

- Lines/page break
  - `*****`
  - `-------`

- LaTeX
  - `$inline equation$`
  - `$$display Equation$$`

http://rmarkdown.rstudio.com/?version=0.98.1028&mode=desktop

R Markdown — Navigation menu



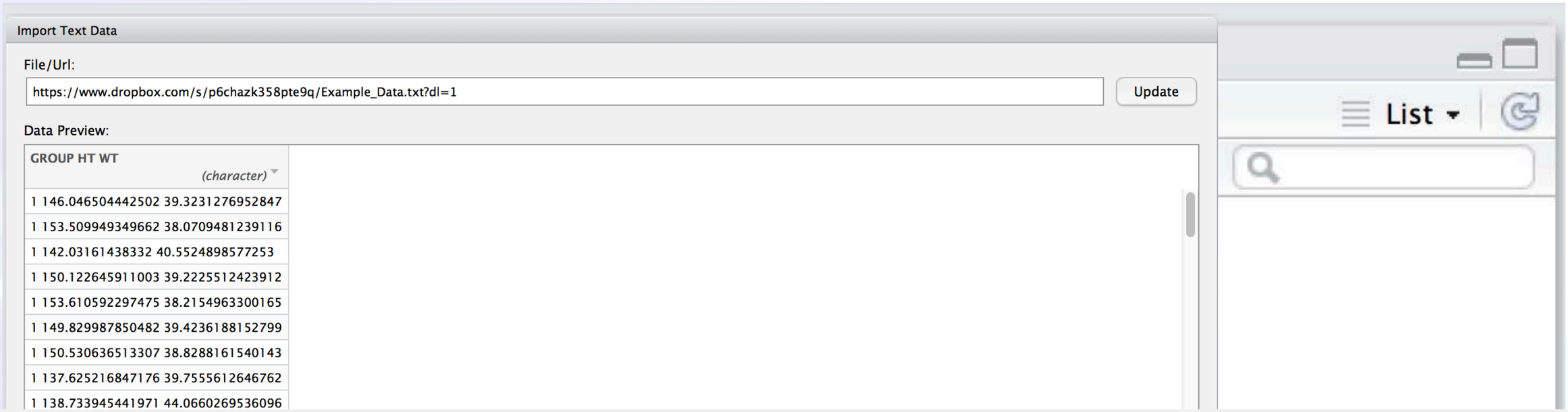
Heading menu
— H1
— H2
— H3

Data

- Spreadsheet
  - Data input
  - Data management
- Save data file plain texts
  - “.csv” [comma separated]
  - “.txt” [tab delimited]
- Can also import files from other statistical software: SPSS, SAS, STATA

Import “Example_Data.txt”

- `?read.csv`
- `?read.table`



Import Options:

Name: Example_Data

Skip: 0

First Row as Names

Trim Spaces

Open Data Viewer

Delimiter: Comma

Quotes: Default

Locale: Configure...

Escape: None

Comment: Default

NA: Default

Code Preview:

```
library(readr)
Example_Data <- read_csv("https://www
.dropbox.com/s/p6chazk358pte9q
/Example_Data.txt?dl=1")
View(Example_Data)
```

Import

Cancel

Data Structure

- Vector
  - `x1 <- c(1, 3, 5, 7, 9)`
  - `x2 <- seq(from=1, to=9, by=2)`
  - `x3 <- rep(0.25,times=5)`
  - `x1 + x3`

```
> x1+x3
[1] 1.25 3.25 5.25 7.25 9.25
```

- Matrix
  - `mat1 <- matrix(data=c(1,2,3,4,5,6),ncol=3)`

```
> mat1
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

```
> rbind(x1,x3,x2)
   [,1] [,2] [,3] [,4] [,5]
x1 1.00 3.00 5.00 7.00 9.00
x3 0.25 0.25 0.25 0.25 0.25
x2 1.00 3.00 5.00 7.00 9.00
```

Data Structure

- Data Frame
  - `df = data.frame(a=x1, b=x2, c=x3)`
- Refer to variable “a”

```
> df$a
[1] 1 3 5 7 9
```

```
> df
  a b    c
1 1 1 0.25
2 3 3 0.25
3 5 5 0.25
4 7 7 0.25
5 9 9 0.25
```

- List
  - `L <- list(one=1,two=c(1,2), five=seq(0,1,length=5))`
  - `L[[“one”]]`
  - `L[[2]]`

```
> L[["one"]]
[1] 1
```

```
> L[[2]]
[1] 1 2
```

```
> L
$one
[1] 1

$two
[1] 1 2

$five
[1] 0.00 0.25 0.50 0.75
1.00
```

Selecting Data

```
> df
  a b    c
1 1 1 0.25
2 3 3 0.25
3 5 5 0.25
4 7 7 0.25
5 9 9 0.25
```

```
> df[3,1]
[1] 5
```

```
> df$a
[1] 1 3 5 7 9
```

```
> df$a[3] #3rd element
[1] 5
```

```
> df[["a"]][1:3]
[1] 1 3 5
```

```
> df$a[6] #6th element?!
[1] NA
```

Clopidogrel Resistance Data

- A study of platelet function and genotype (contain 5 variables)
- IID: 211 patients with CAD
- ADP: platelet function assay (AU)
- resistance: clopidogrel resistance
- 3 SNPs genotype data
  - 0, 1, 2 based on the number of minor alleles
  - rs4244285: *CYP2C19*2*
  - rs4986893: *CYP2C19*3*
  - rs662: variant in *PON1*

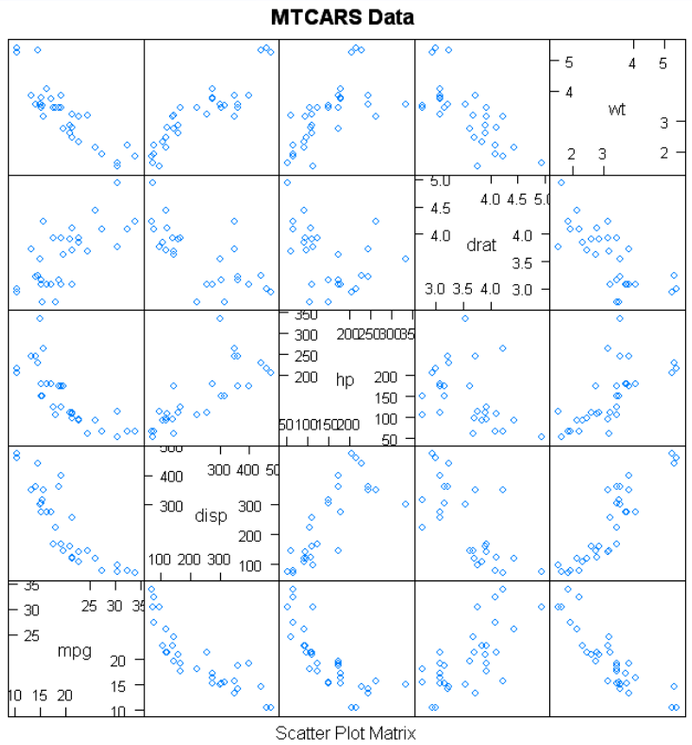
http://dx.plos.org/10.1371/journal.pone.0110188

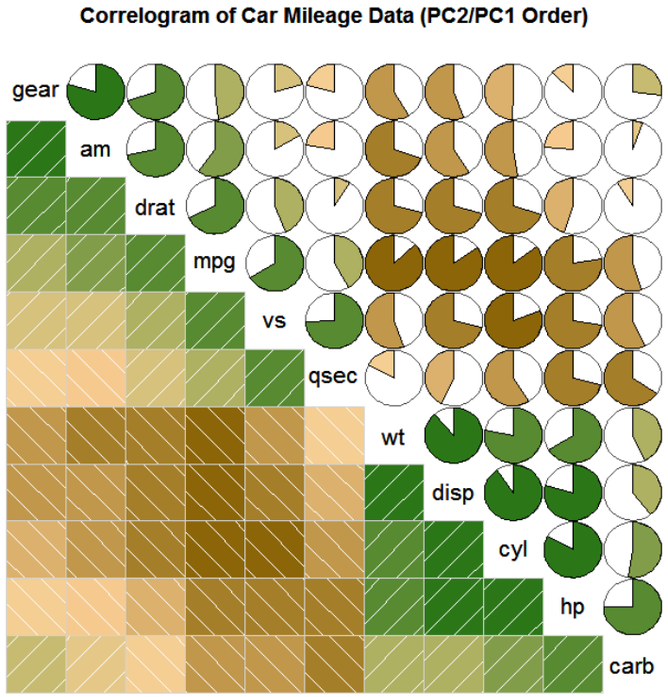
	A	B	C	D	E	F
1	IID	ADP	Resistance	rs4244285	rs4986893	rs662
2	1	1.59957315	0	0	0	1
3	2	9.52714594	0	0	0	2
4	3	12.8094598	0	0	0	1
5	4	14.5233349	0	0	0	1
6	5	18.2618036	0	0	0	2
7	6	23.2778581	0	0	0	0
8	7	32.8712885	0	0	0	2
9	8	13.2975173	0	0	0	1
10	9	33.6278596	0	1	0	1
11	10	51.5061123	0	2	0	2
12	11	76.0005883	1	0	0	1
13	12	11.513361	0	0	0	2
14	13	15.3118085	0	0	0	2
15	14	5.18451304	0	1	0	2
16	15	22.8071414	0	0	0	2
17	16	27.2004589	0	0	0	1
18	17	15.1600453	0	1	0	2
19	18	12.0639912	0	1	0	2
20	19	13.3450161	0	0	0	2
21	20	10.8272303	0	1	0	1

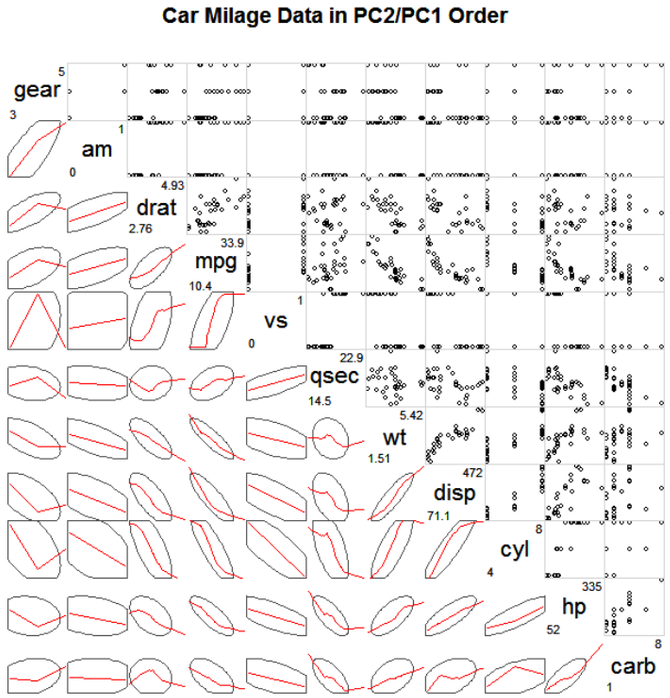
 Tresukosol, D. et al. PLoS ONE 9, e110188 (2014).

What questions can we ask?

- Descriptive statistics
  - distribution (mean, SD)
  - Any outliers?
- Do ADP-induced platelet aggregation differ among different genotype groups?
  - Try graphs and test
- Genetic association
  - Tests if SNP is associated with ADP-induced platelet aggregation or clopidogrel resistance



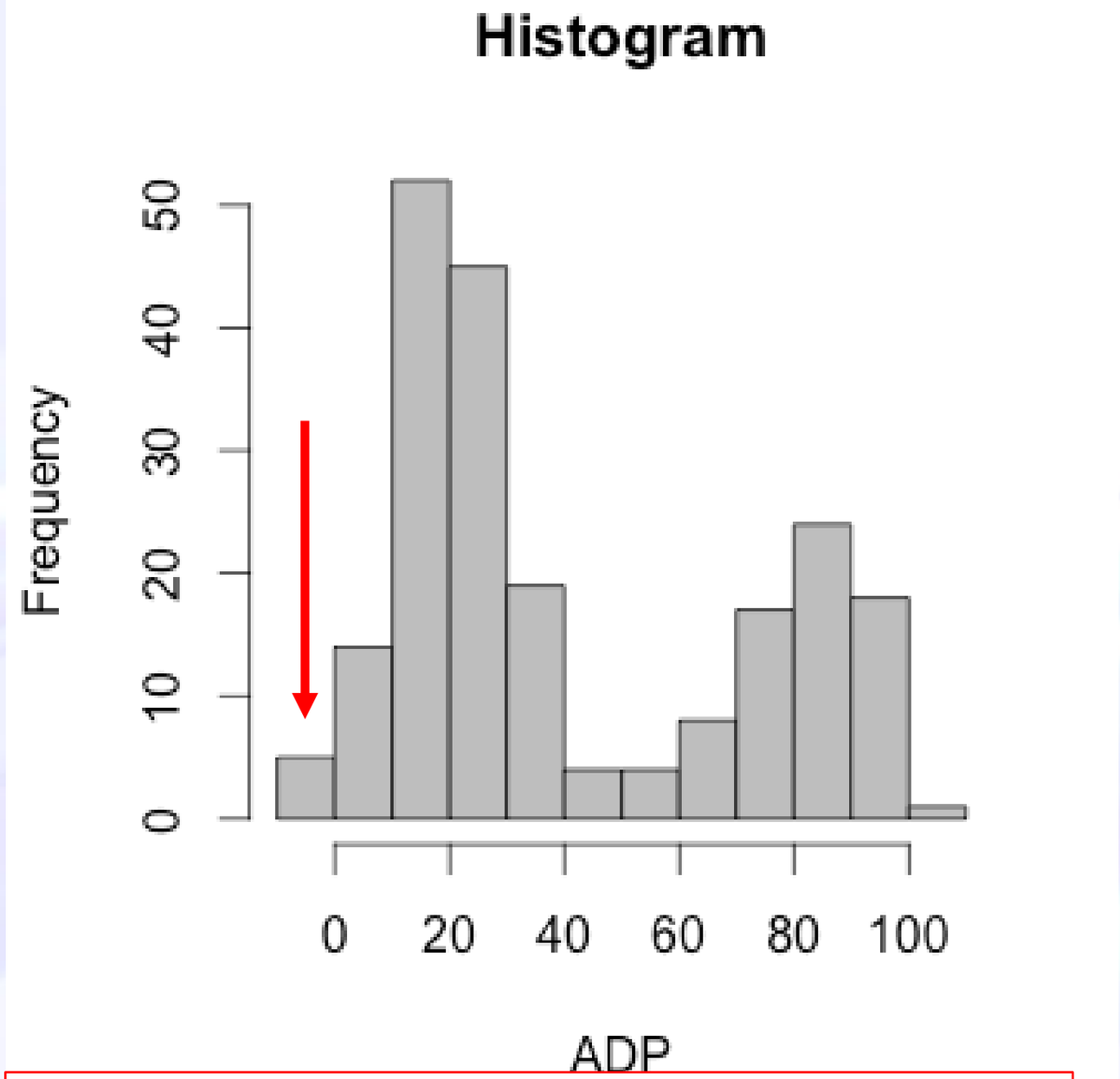




EXPLORATORY DATA ANALYSIS

Distribution of ADP-induced platelet aggregation

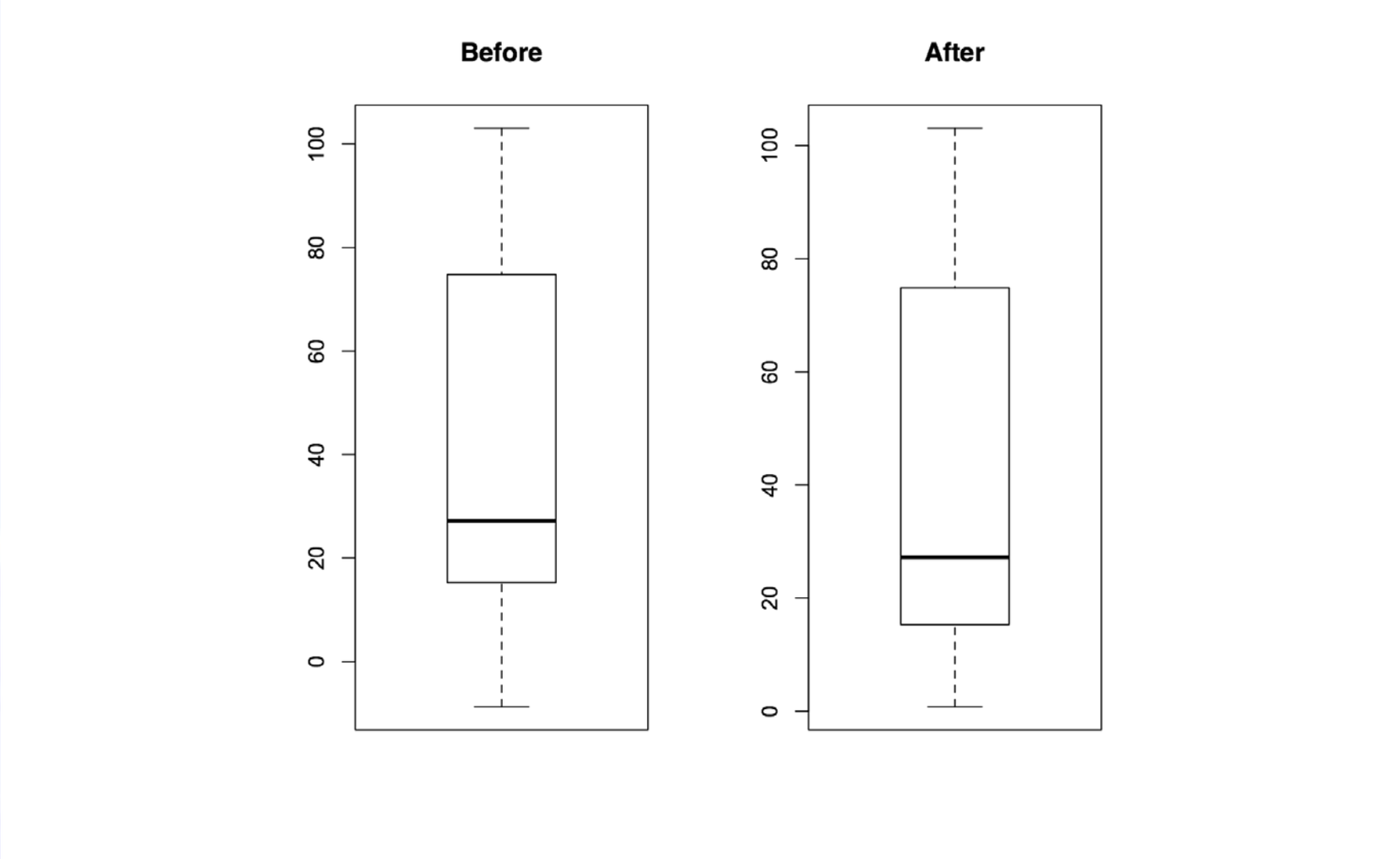
```
> hist(data$ADP,main="Histogram",xlab="ADP",col="grey")
```




```
> summary(data$ADP)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -8.721  15.280  27.200  40.850  74.810 103.100
```

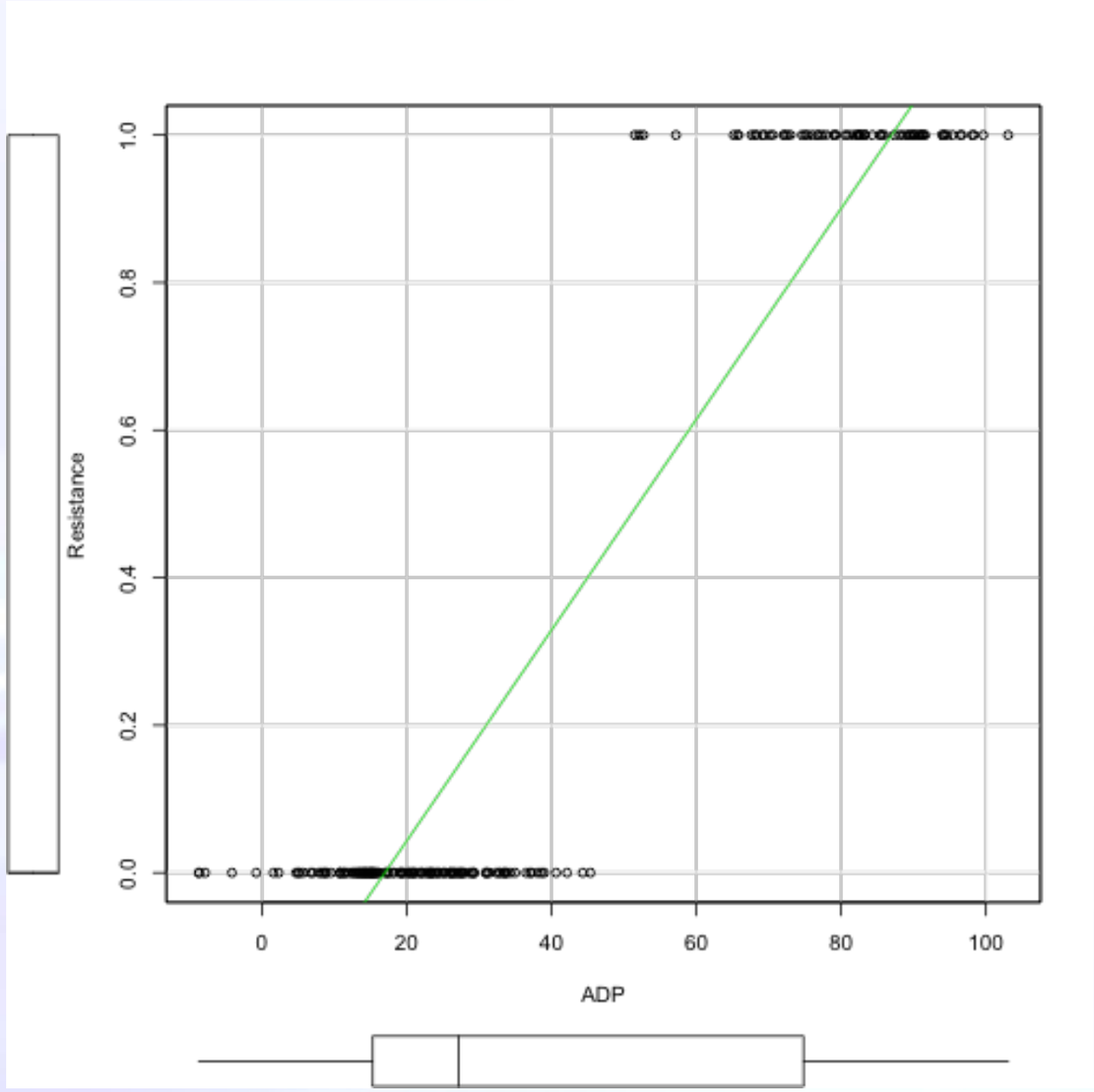
Boxplot

Correcting incorrect data entry with "-" values.



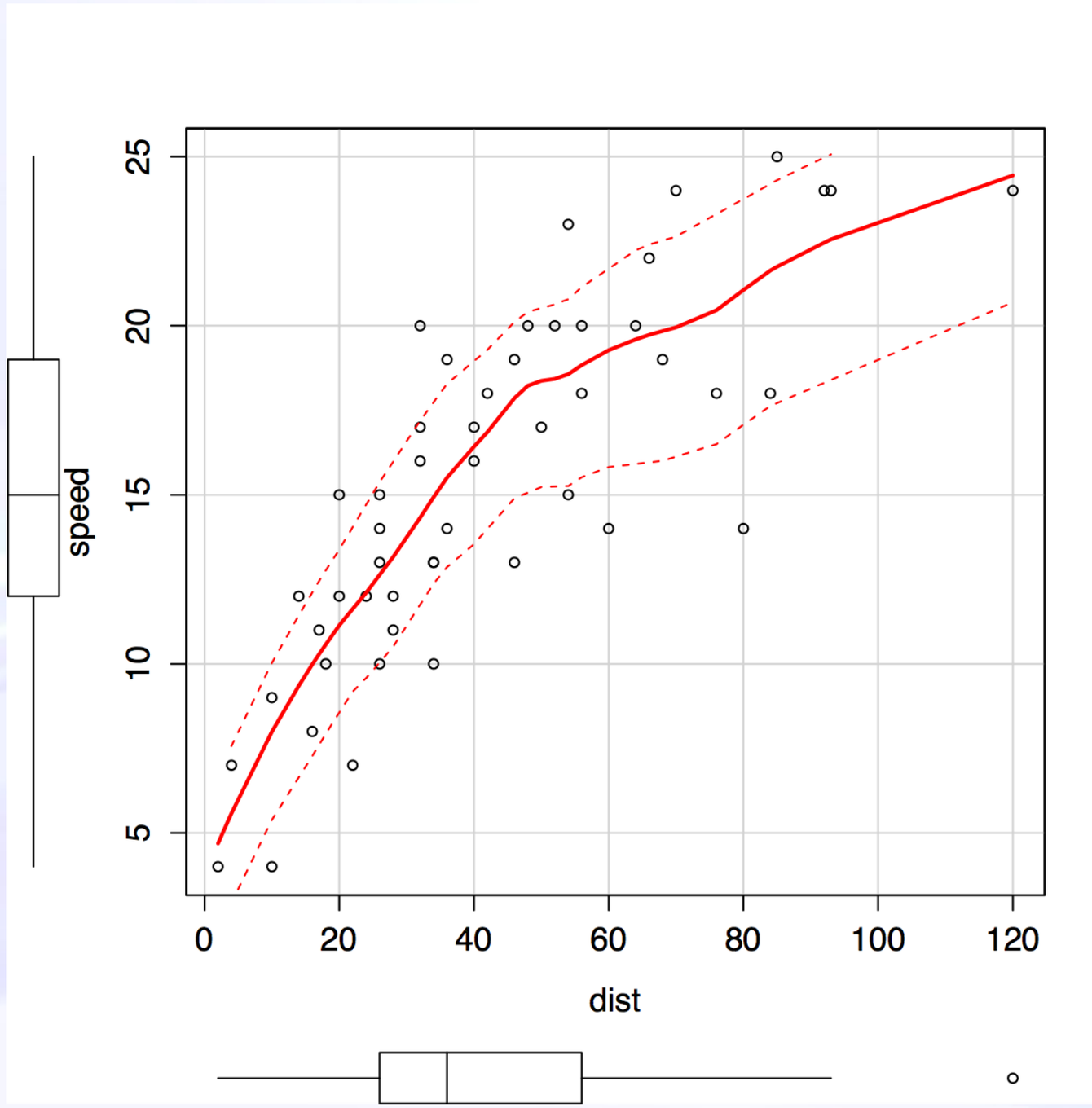

```
data$ADP2 <- abs(data$ADP) # replace ADP < 0 with abs(ADP)
par(mfcol=c(1,2))
boxplot(data$ADP, main="Before")
boxplot(data$ADP2, main="After")
```

Relationship between clopidogrel resistance and ADP-induced platelet aggregation




```
library(car)
scatterplot(Resistance ~ ADP, data=data, reg.line = lm, smooth=FALSE)
```

Scatterplot of continuous variables



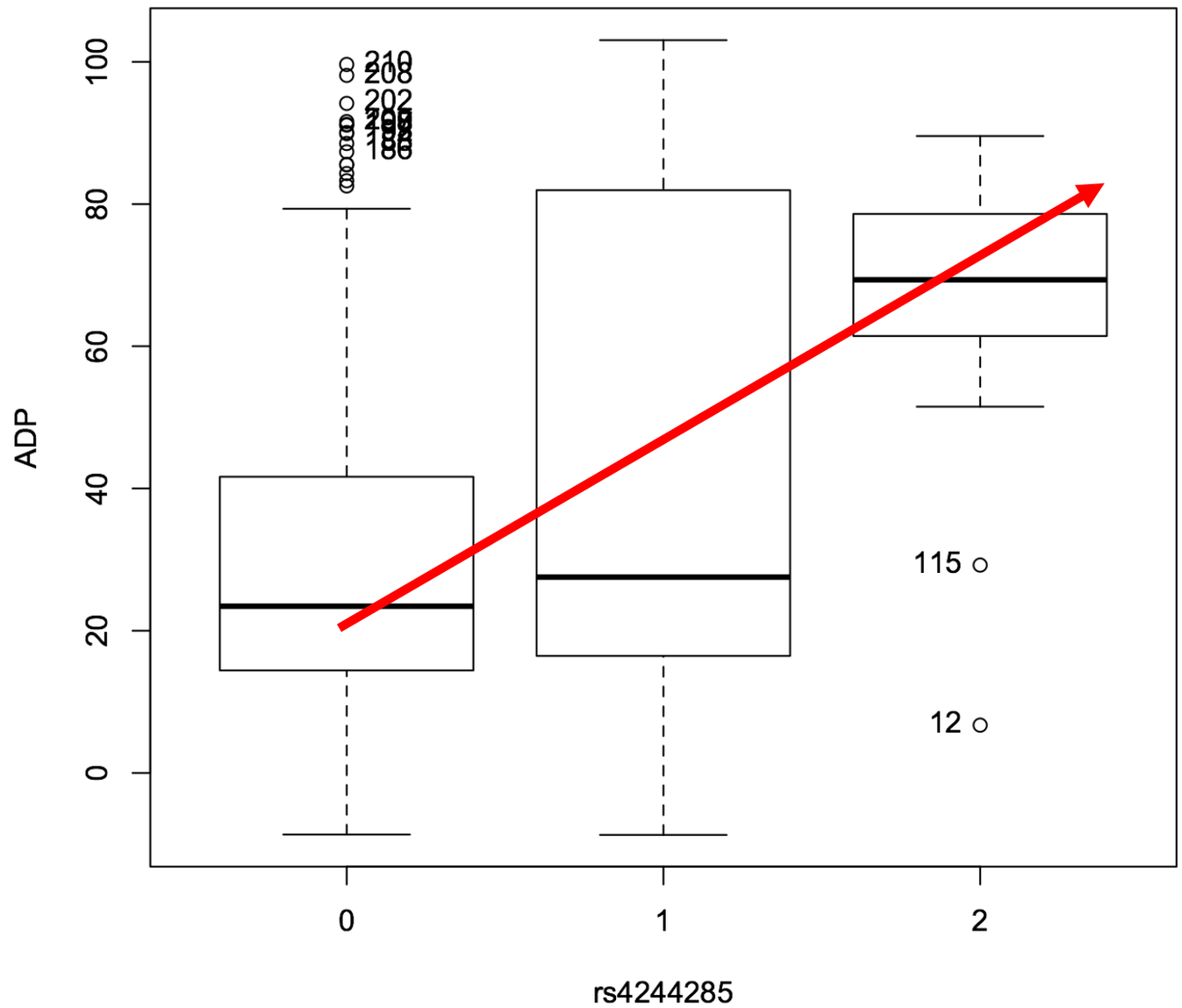

```
scatterplot(speed~dist,data=cars,smooth=TRUE,reg.line='lm')
```

What questions can we ask?

- Descriptive statistics
  - distribution (mean, SD)
  - Any outliers?
- Do ADP-induced platelet aggregation differ among different genotype groups?
  - *Try graphs and test*
- Genetic association
  - Tests if SNP is associated with ADP-induced platelet aggregation or clopidogrel resistance

Distribution of ADP-induced platelet aggregation by genotype group rs4244285

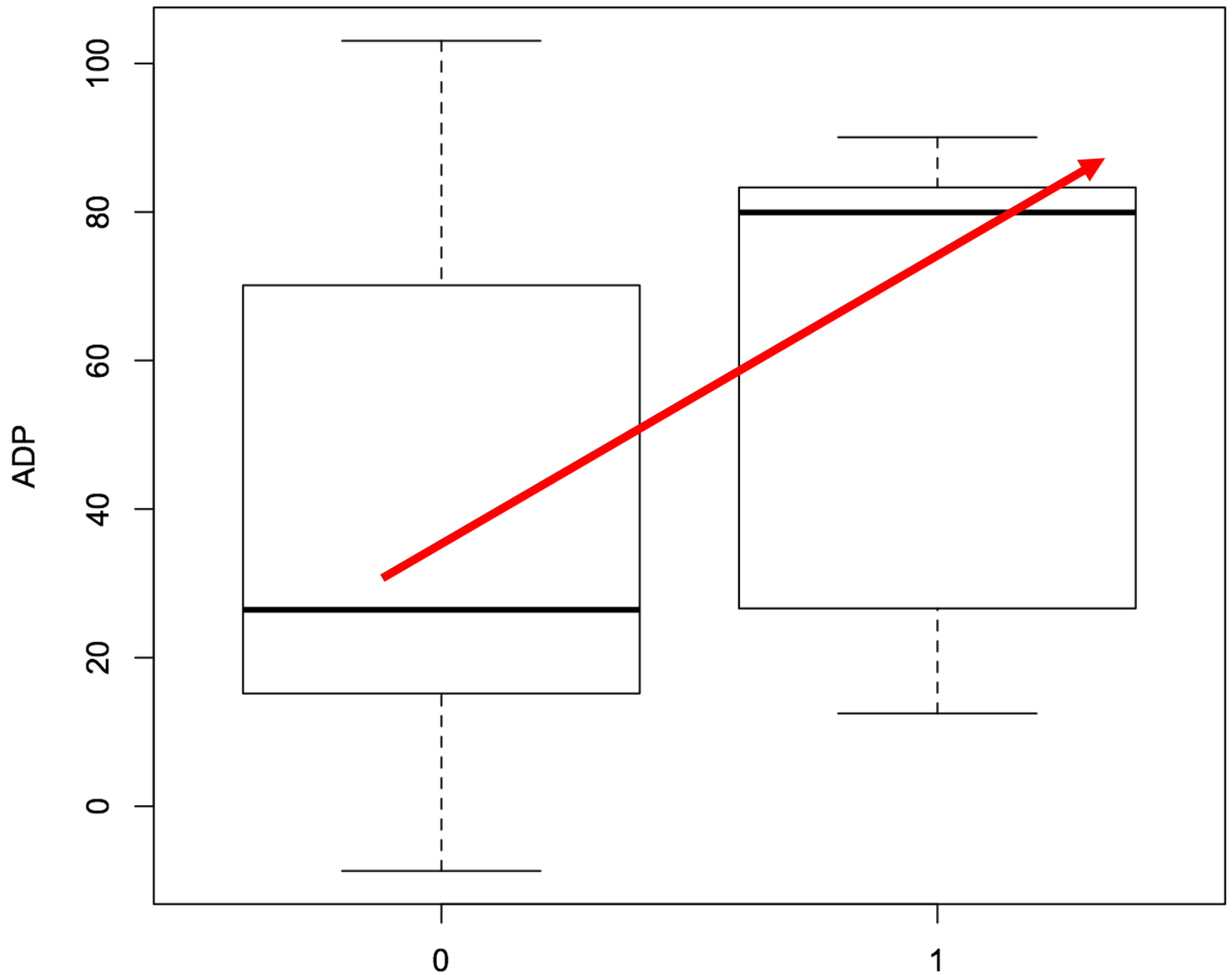
Any relationship between genotype and platelet aggregation?




```
Boxplot(data$ADP~data$rs4244285,xlab="rs4244285",ylab="ADP")
```

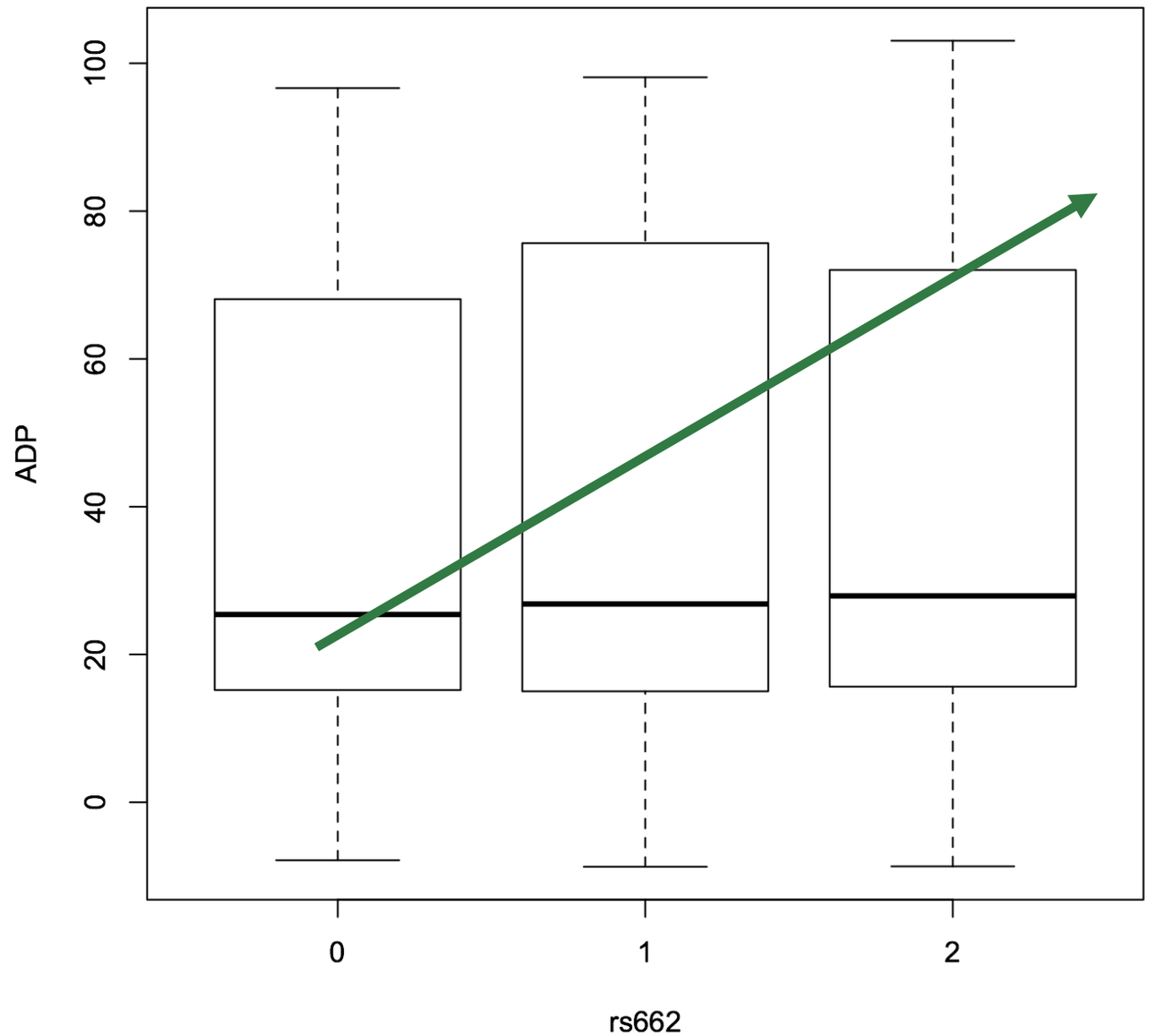
Relationship between platelet aggregation and genotypes

What relationship can you see?



Distribution by genotype group rs662

Still see the trend?



Exercise & Assignment

- Heading to R-Studio Cloud
  - https://rstudio.cloud
- Sign-up for an account
- Create a new Project from Github repository
  - https://github.com/hypotheses/SIBB512-R

Mail - Bhoom Suktitipat - Outloo | RStudio Cloud | hypotheses/SIBB512-R

rstudio.cloud/projects

Studio Cloud

Your Workspace | Projects | Info

Blast Oppenheimer

Spaces

- Your Workspace
- SIBB512
- New Space

Learn

- Guide
- What's New
- Primers
- Cheat Sheets
- Feedback and Questions

Info

- Terms and Conditions
- System Status

Your Projects

New Project

New Project from Git Repo

URL of your Git repository

https://github.com/hypotheses/SIBB512-R

OK

Options

Search Projects

Sort Projects

- By name
- By date created

Capacity

This your personal workspace, where you can create a virtually unlimited number of projects.

Learn more about Your Workspace in the Guide.

Studio Cloud alpha

© 2019 RStudio Inc.

Assignment Submission

- **Check 00-SIBB512-Assignment.Rmd**
- **knit** your **Rmd file** to create an html file.
- Send the file by email to bhoom.suk@mahidol.edu
- Name the file using this template **SIBB512-FirstName-Lastname.html**
- **Due Dec 5, 2019**