Data Visualization in R

Dr. Alexander Mark Weber

Assistant Professor (Partner), Department of Pediatrics,
Division of Neurology, Faculty of Medicine,
Associate Member, Department of Neuroscience,
Associate Member, School of Biomedical Engineering,
Independent Investigator, BC Children's Hospital Research Institute,
University of British Columbia

November 19, 2020

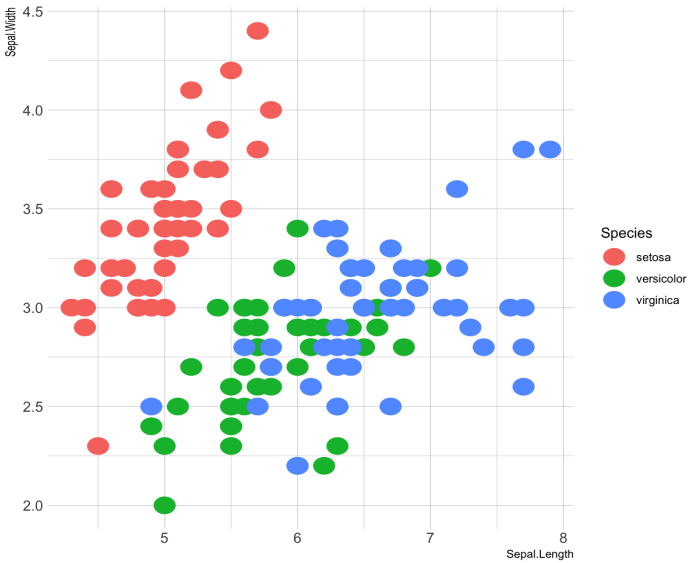
Outline

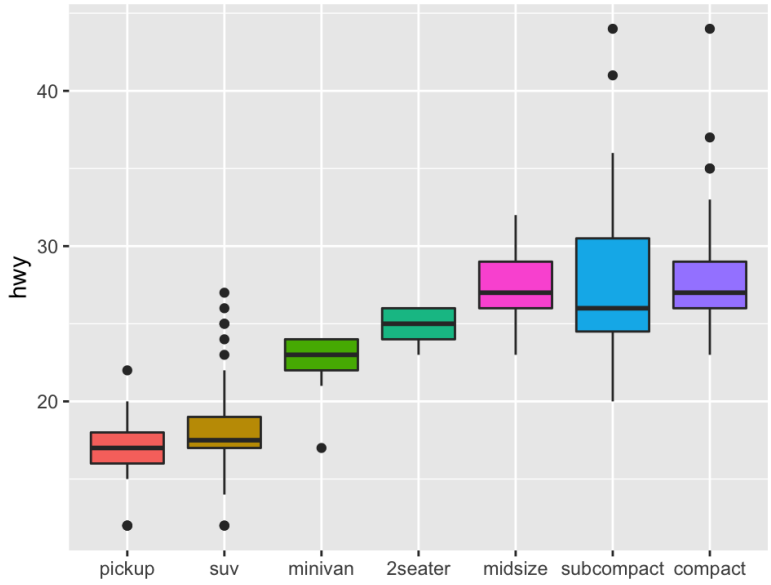
Learning Objectives

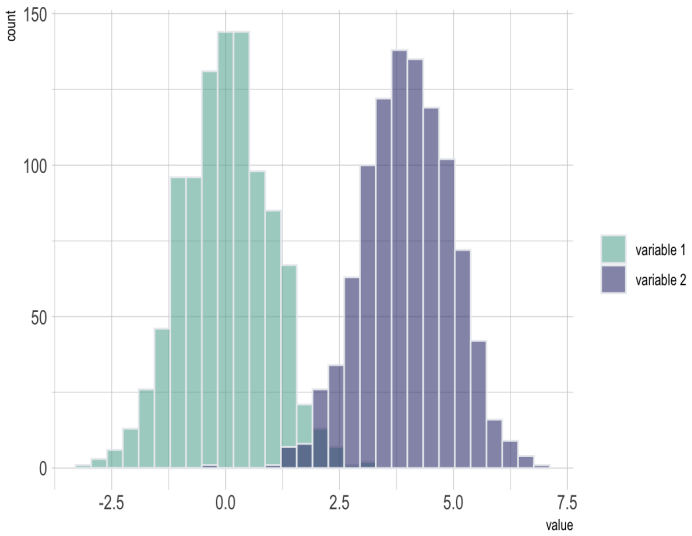
By the end of this lecture students will be able to...

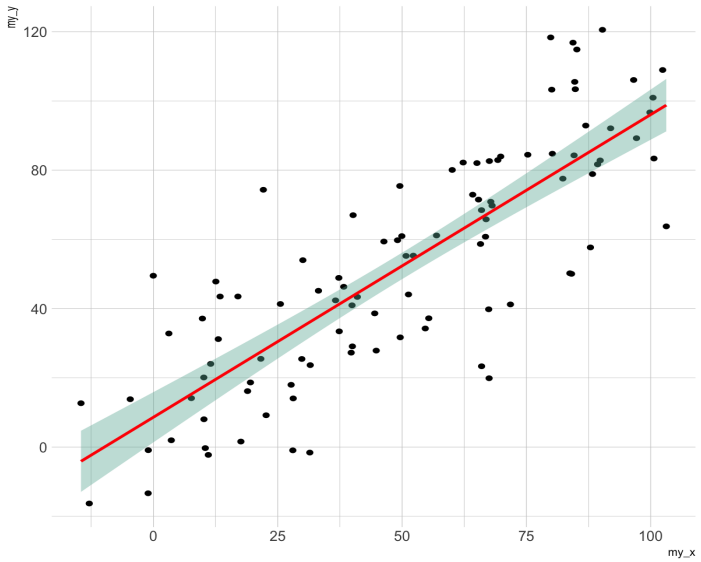
- explain to a colleague what R is
- install and run R and RStudio
- load data and packages in R and RStudio
- describe what a histogram is and when you would use it
- describe what a scatterplot is and when you would use it
- describe what a boxplot is, what each horizontal line refers to, and when you would use it
- describe and identify outliers
- explain concepts of correlation and simple linear regression
- perform correlation and regression analysis using R and RStudio
- interpret results from correlation and regression
- create and customize histograms, scatterplots, boxplots, and linear models in R

Basic Introduction to R









Basic Introduction to R

What is R?

Basic Introduction to R

What is R?

- A language and environment for **statistical computing and graphics**

Basic Introduction to R

What is R?

- A language and environment for **statistical computing and graphics**
- A GNU Project - 100% free software (other examples include Linux operating systems)

Basic Introduction to R

What is R?

- A language and environment for **statistical computing and graphics**
- A GNU Project - 100% free software (other examples include Linux operating systems)
- Widely used among statisticians, data scientists, and scientists alike

Basic Introduction to R

What is R?

- A language and environment for **statistical computing and graphics**
- A GNU Project - 100% free software (other examples include Linux operating systems)
- Widely used among statisticians, data scientists, and scientists alike

Why use R?

- Free, well documented, helpful community online, runs on everything

Basic Introduction to R

What is R?

- A language and environment for **statistical computing and graphics**
- A GNU Project - 100% free software (other examples include Linux operating systems)
- Widely used among statisticians, data scientists, and scientists alike

Why use R?

- Free, well documented, helpful community online, runs on everything
- Wide variety of statistical techniques: linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering

Basic Introduction to R

What is R?

- A language and environment for **statistical computing and graphics**
- A GNU Project - 100% free software (other examples include Linux operating systems)
- Widely used among statisticians, data scientists, and scientists alike

Why use R?

- Free, well documented, helpful community online, runs on everything
- Wide variety of statistical techniques: linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering
- Highly extensible via *packages*

Basic Introduction to R

What is R?

- A language and environment for **statistical computing and graphics**
- A GNU Project - 100% free software (other examples include Linux operating systems)
- Widely used among statisticians, data scientists, and scientists alike

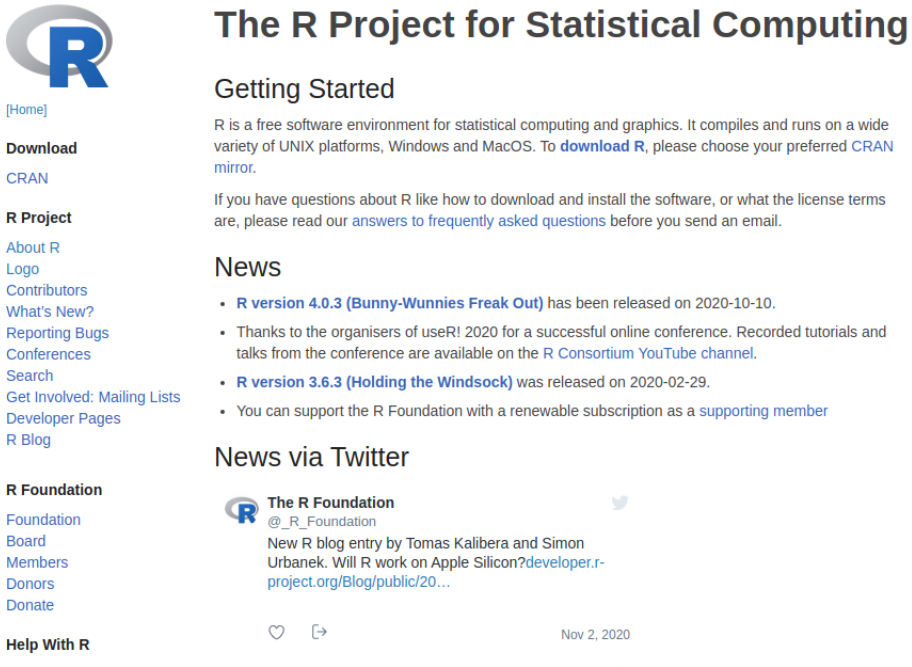
Why use R?

- Free, well documented, helpful community online, runs on everything
- Wide variety of statistical techniques: linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering
- Highly extensible via *packages*
- Well-designed publication-quality plots

Getting Started

Where to begin?

- Visit https://www.r-project.org/



[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Conferences
Search
Get Involved: Mailing Lists
Developer Pages
R Blog

R Foundation

Foundation
Board
Members
Donors
Donate

Help With R

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- **R version 4.0.3 (Bunny-Wunnies Freak Out)** has been released on 2020-10-10.
- Thanks to the organisers of useR! 2020 for a successful online conference. Recorded tutorials and talks from the conference are available on the R Consortium YouTube channel.
- **R version 3.6.3 (Holding the Windsock)** was released on 2020-02-29.
- You can support the R Foundation with a renewable subscription as a supporting member

News via Twitter

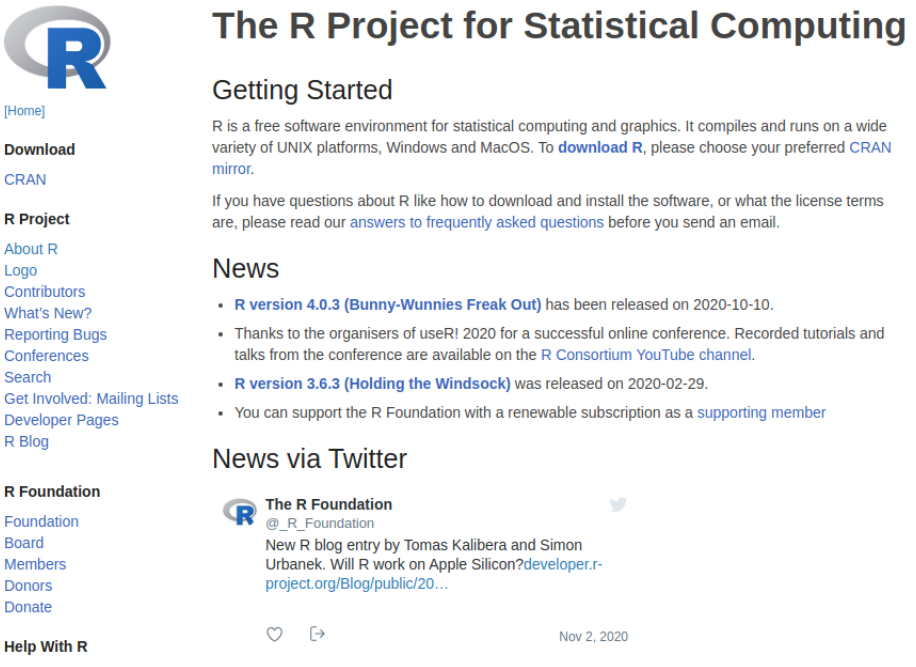
The R Foundation
@_R_Foundation
New R blog entry by Tomas Kalibera and Simon Urbanek. Will R work on Apple Silicon?developer.r-project.org/Blog/public/20…

Nov 2, 2020

Getting Started

Where to begin?

- Visit https://www.r-project.org/



[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Conferences
Search
Get Involved: Mailing Lists
Developer Pages
R Blog

R Foundation

Foundation
Board
Members
Donors
Donate

Help With R

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

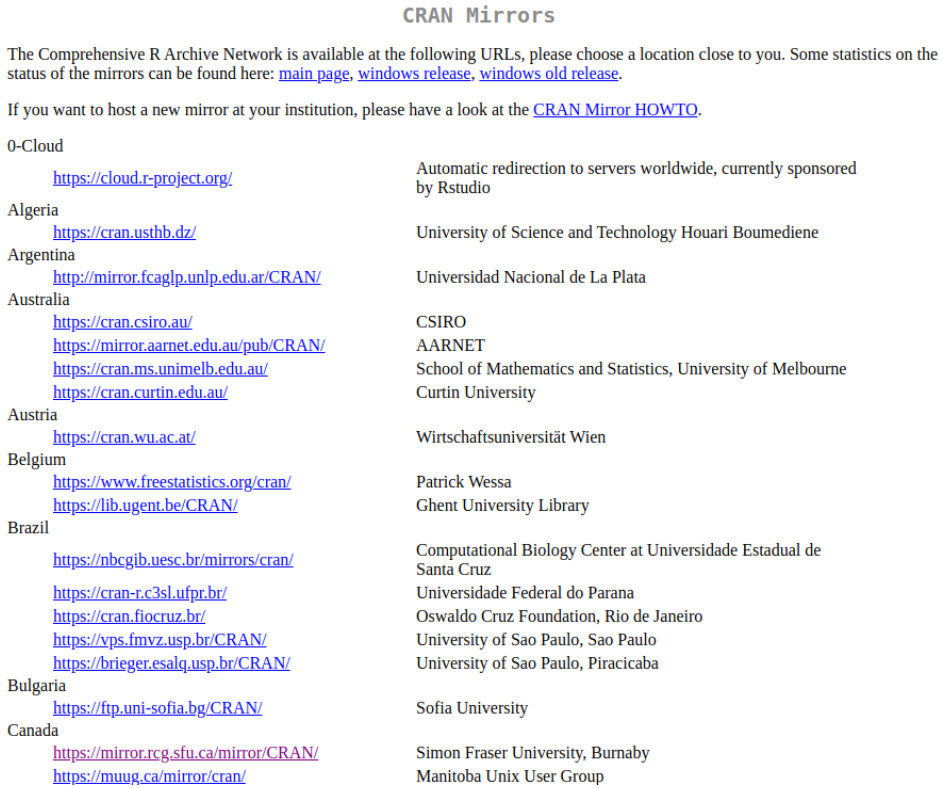
News

- **R version 4.0.3 (Bunny-Wunnies Freak Out)** has been released on 2020-10-10.
- Thanks to the organisers of useR! 2020 for a successful online conference. Recorded tutorials and talks from the conference are available on the R Consortium YouTube channel.
- **R version 3.6.3 (Holding the Windsock)** was released on 2020-02-29.
- You can support the R Foundation with a renewable subscription as a supporting member

News via Twitter

The R Foundation
@_R_Foundation
New R blog entry by Tomas Kalibera and Simon Urbanek. Will R work on Apple Silicon?developer.r-project.org/Blog/public/20…

Nov 2, 2020



CRAN Mirrors

The Comprehensive R Archive Network is available at the following URLs, please choose a location close to you. Some statistics on the status of the mirrors can be found here: main page, windows release, windows old release.

If you want to host a new mirror at your institution, please have a look at the CRAN Mirror HOWTO.

0-Cloud	
https://cloud.r-project.org/	Automatic redirection to servers worldwide, currently sponsored by Rstudio
Algeria	
https://cran.usthb.dz/	University of Science and Technology Houari Boumediene
Argentina	
http://mirror.fcaglp.unlp.edu.ar/CRAN/	Universidad Nacional de La Plata
Australia	
https://cran.csiro.au/	CSIRO
https://mirror.aarnet.edu.au/pub/CRAN/	AARNET
https://cran.ms.unimelb.edu.au/	School of Mathematics and Statistics, University of Melbourne
https://cran.curtin.edu.au/	Curtin University
Austria	
https://cran.wu.ac.at/	Wirtschaftsuniversität Wien
Belgium	
https://www.freestatistics.org/cran/	Patrick Wessa
https://lib.ugent.be/CRAN/	Ghent University Library
Brazil	
https://nbcgib.uesc.br/mirrors/cran/	Computational Biology Center at Universidade Estadual de Santa Cruz
https://cran-r.c3sl.ufpr.br/	Universidade Federal do Parana
https://cran.fiocruz.br/	Oswaldo Cruz Foundation, Rio de Janeiro
https://vps.fmvz.usp.br/CRAN/	University of Sao Paulo, Sao Paulo
https://brieger.esalq.usp.br/CRAN/	University of Sao Paulo, Piracicaba
Bulgaria	
https://ftp.uni-sofia.bg/CRAN/	Sofia University
Canada	
https://mirror.rcg.sfu.ca/mirror/CRAN/	Simon Fraser University, Burnaby
https://muug.ca/mirror/cran/	Manitoba Unix User Group

Getting Started

Where to begin?

- Download R for your operating system and install

The Comprehensive R Archive Network

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.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2020-10-10, Bunny-Wunnies Freak Out) R-4.0.3.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

- If you have questions about R like how to download and install the software,

Getting Started

Where to begin?

- Download R for your operating system and install

The Comprehensive R Archive Network

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.

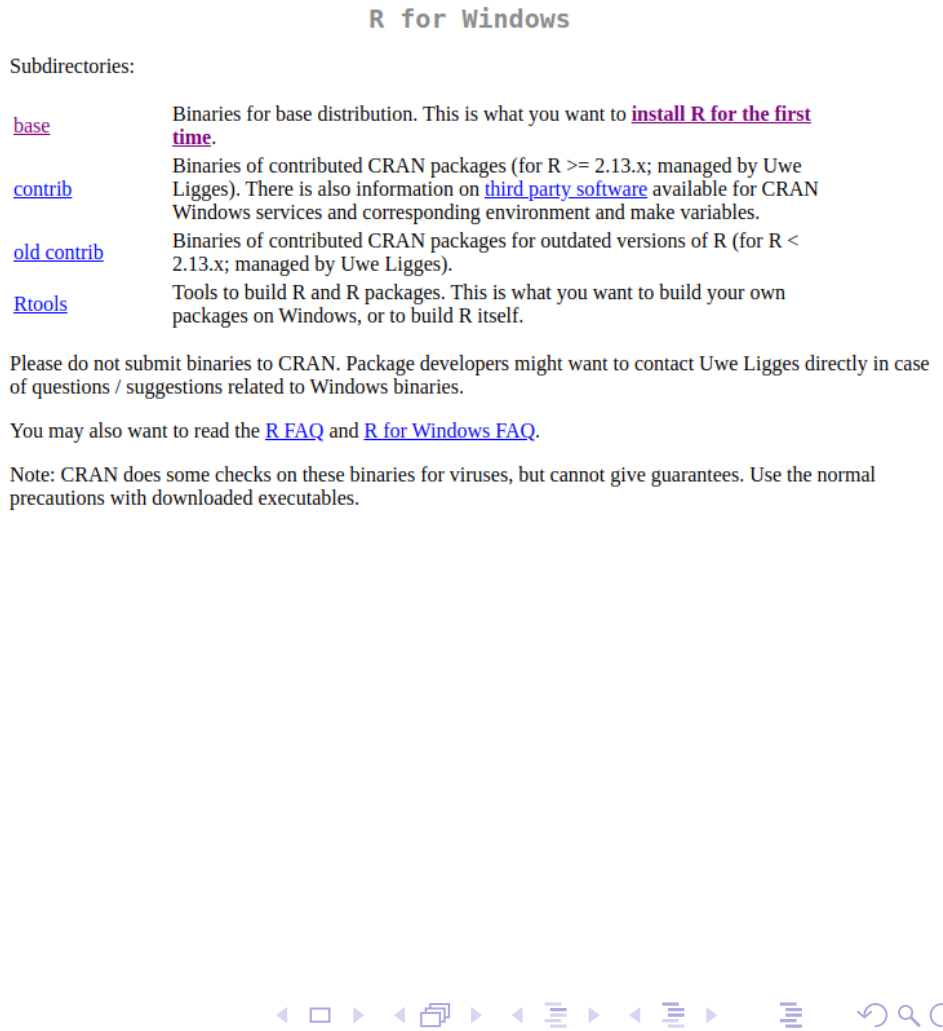
Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2020-10-10, Bunny-Wunnies Freak Out) R-4.0.3.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

- If you have questions about R like how to download and install the software,



R for Windows

Subdirectories:

base	Binaries for base distribution. This is what you want to **install R for the first time**.
contrib	Binaries of contributed CRAN packages (for R >= 2.13.x; managed by Uwe Ligges). There is also information on third party software available for CRAN Windows services and corresponding environment and make variables.
old contrib	Binaries of contributed CRAN packages for outdated versions of R (for R < 2.13.x; managed by Uwe Ligges).
Rtools	Tools to build R and R packages. This is what you want to build your own packages on Windows, or to build R itself.

Please do not submit binaries to CRAN. Package developers might want to contact Uwe Ligges directly in case of questions / suggestions related to Windows binaries.

You may also want to read the R FAQ and R for Windows FAQ.

Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded executables.

RStudio

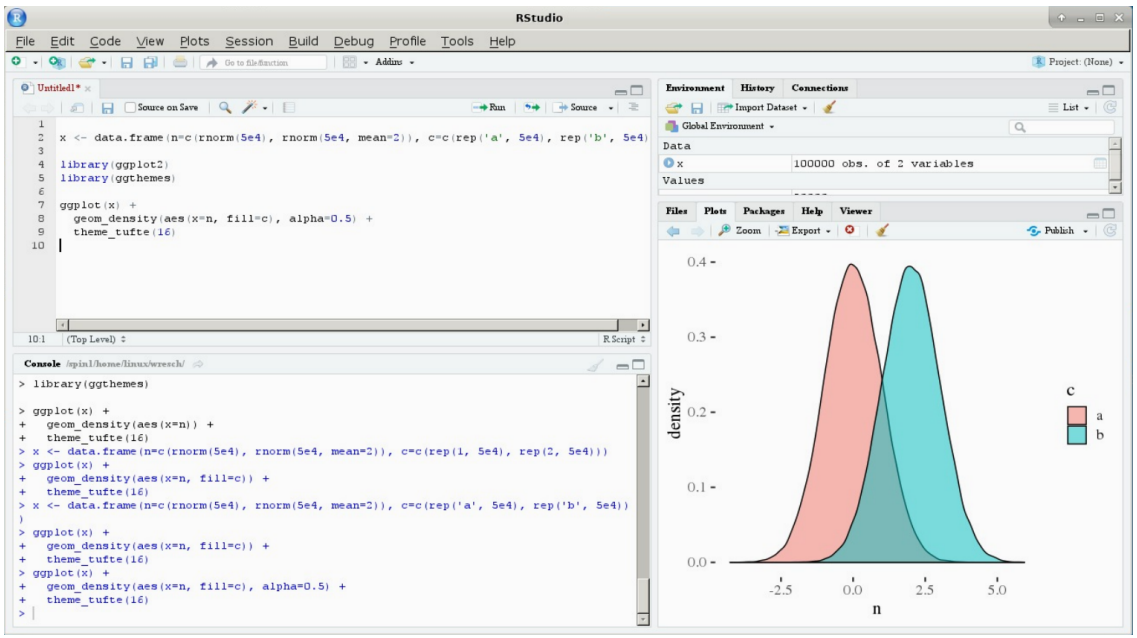
What is RStudio?

- RStudio is an integrated development environment (IDE) for R.

RStudio

What is RStudio?

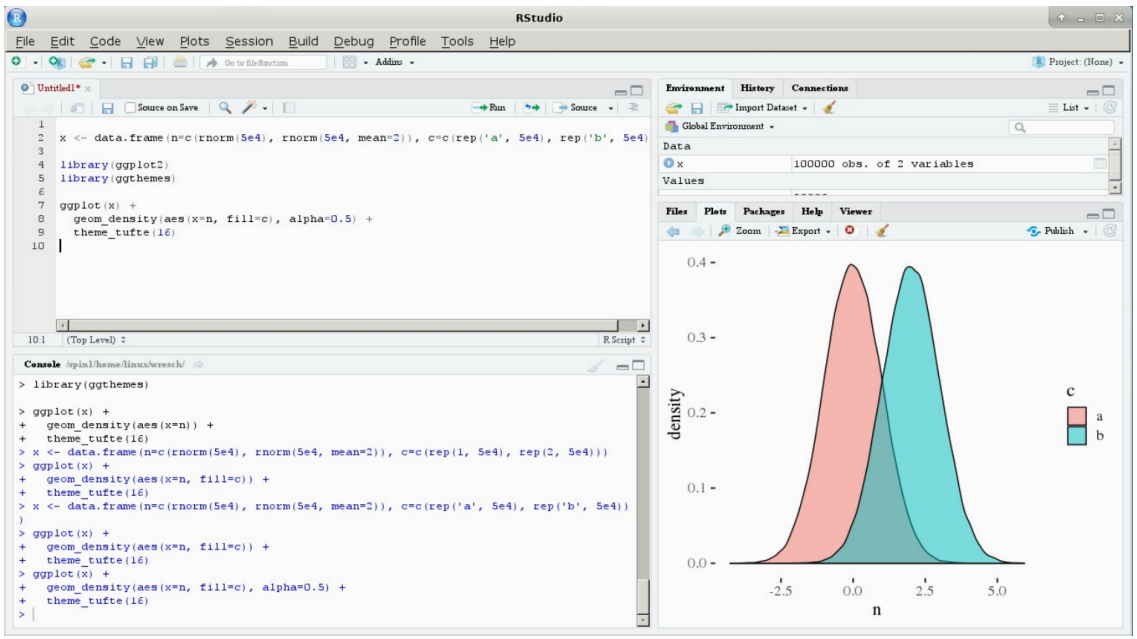
- RStudio is an integrated development environment (IDE) for R.
- Includes: console, syntax-highlighting editor, tools for plotting, history, debugging and workspace management

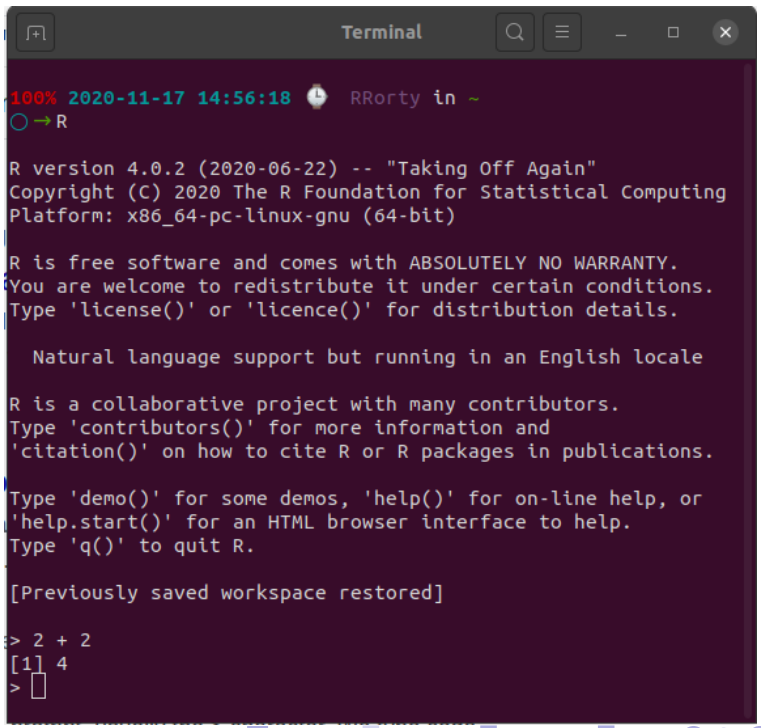


RStudio

What is RStudio?

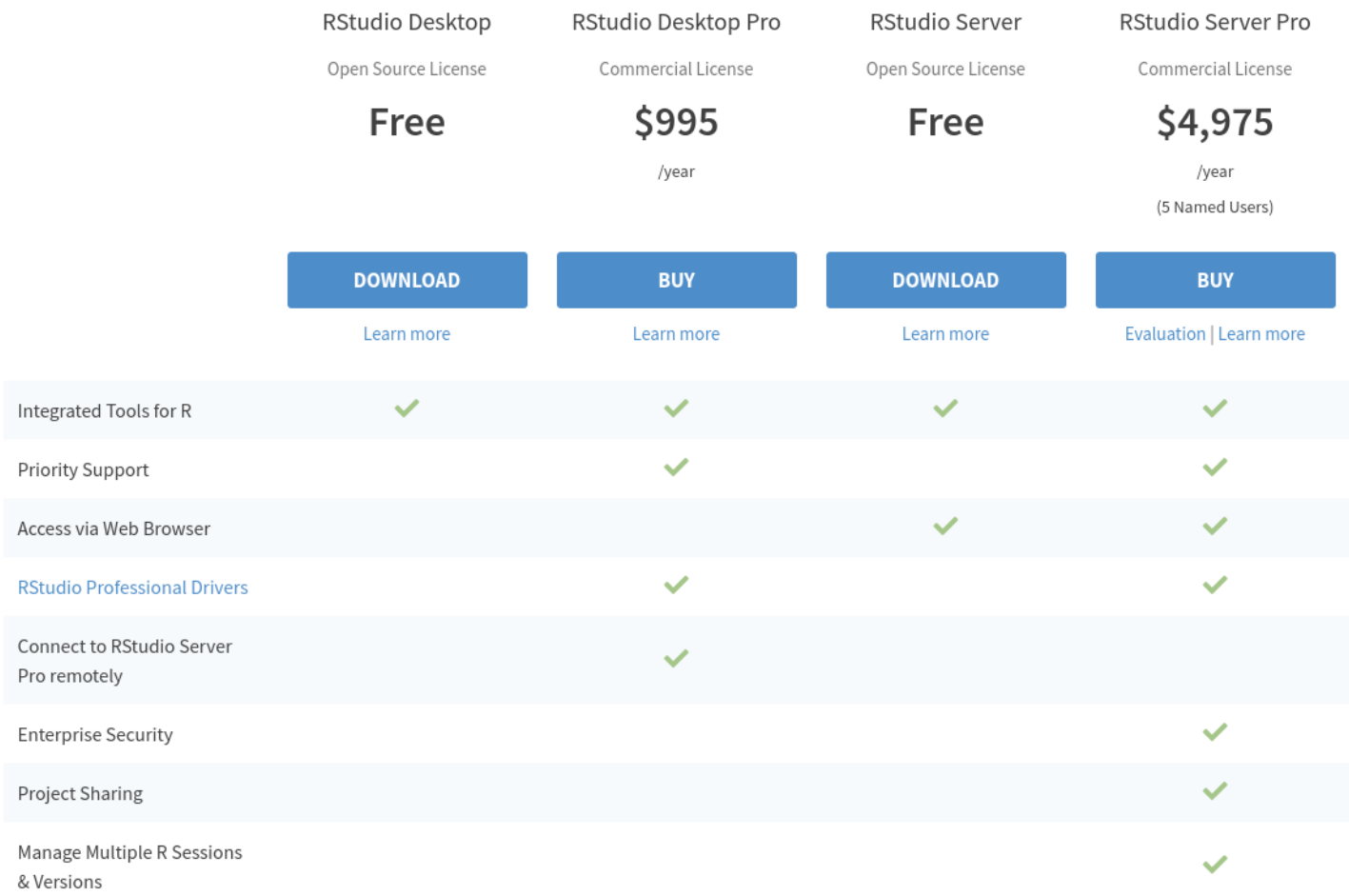
- RStudio is an integrated development environment (IDE) for R.
- Includes: console, syntax-highlighting editor, tools for plotting, history, debugging and workspace management
- Otherwise R is run through a command line interface (CLI)





Installation

- Visit https://rstudio.com/products/rstudio/download/

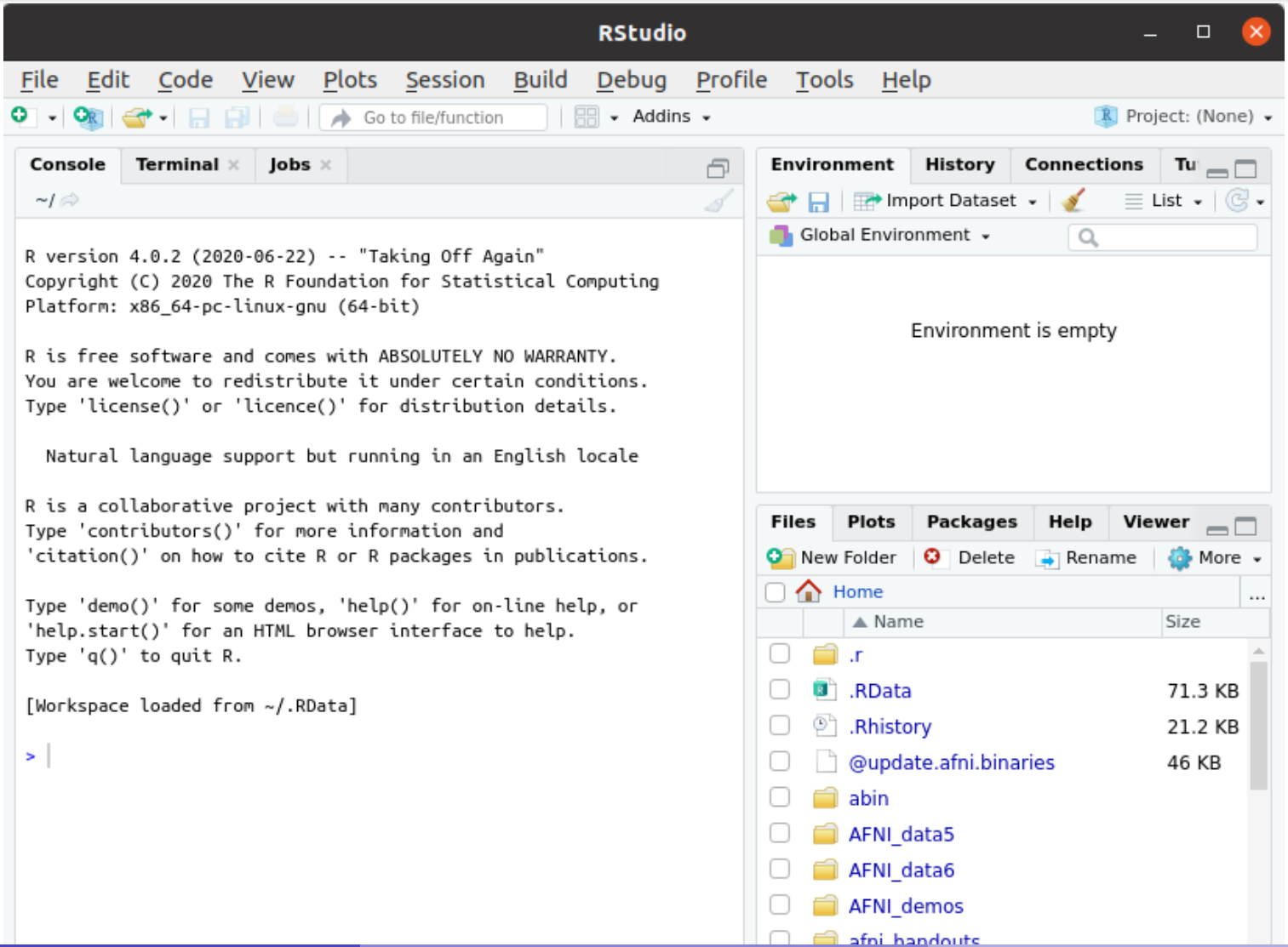


	RStudio Desktop	RStudio Desktop Pro	RStudio Server	RStudio Server Pro
	Open Source License	Commercial License	Open Source License	Commercial License
	Free	$995 /year	Free	$4,975 /year (5 Named Users)
	DOWNLOAD	BUY	DOWNLOAD	BUY
	Learn more	Learn more	Learn more	Evaluation | Learn more
Integrated Tools for R	✓	✓	✓	✓
Priority Support		✓		✓
Access via Web Browser			✓	✓
RStudio Professional Drivers		✓		✓
Connect to RStudio Server Pro remotely		✓		
Enterprise Security				✓
Project Sharing				✓
Manage Multiple R Sessions & Versions				✓

RStudio

Getting Started

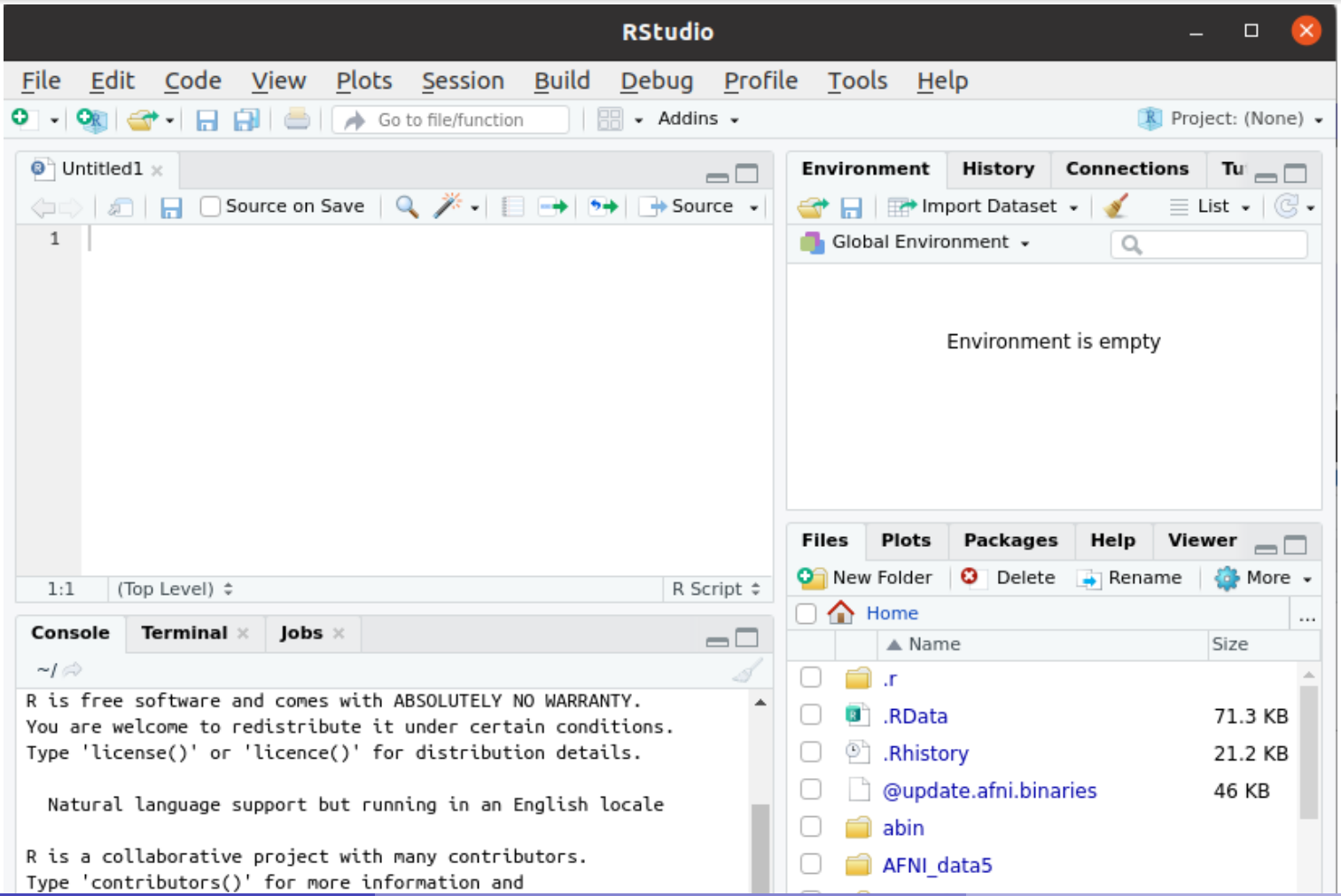
- Let's run Rstudio and take a look



RStudio

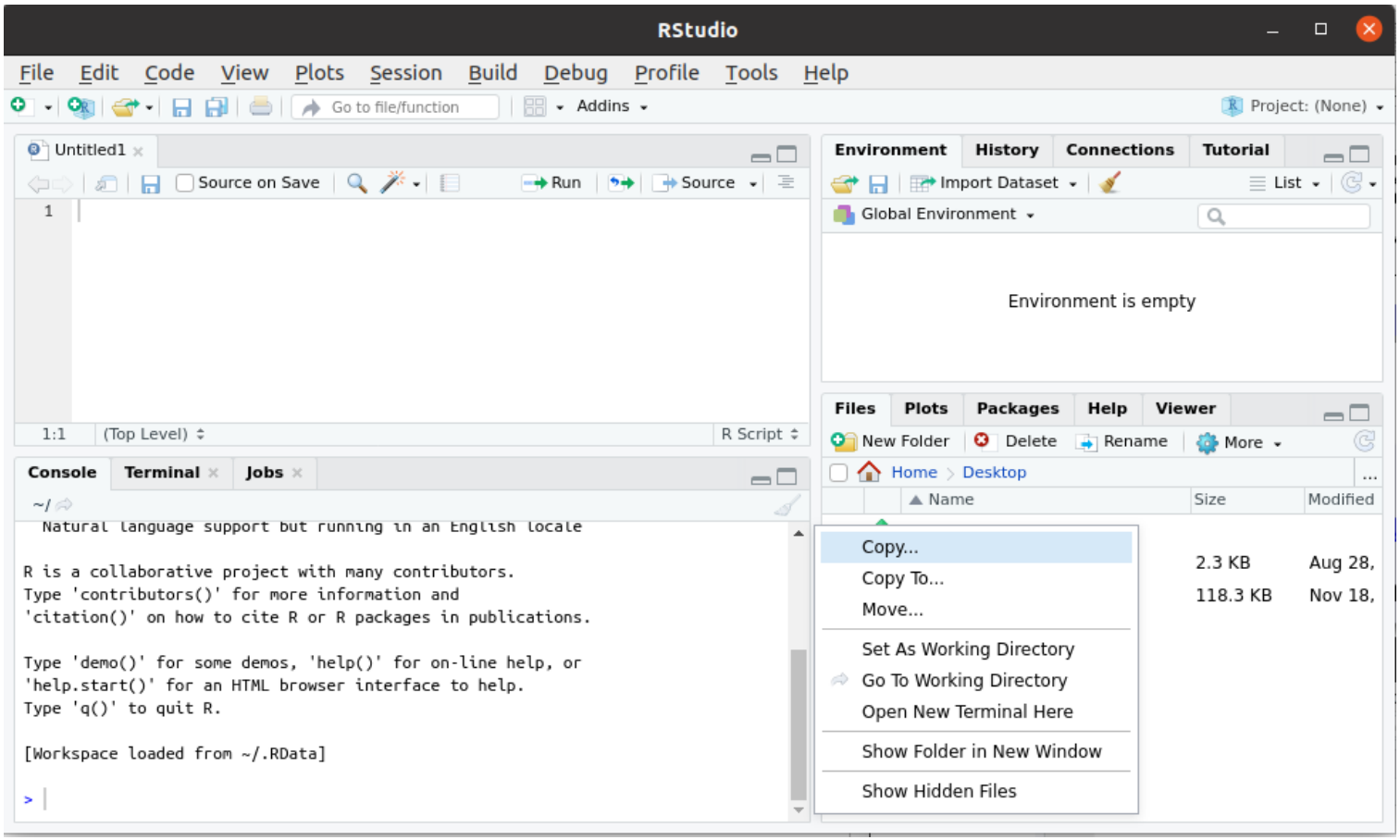
Getting Started

- Let's open up a scripting window
- File > New File > RScript

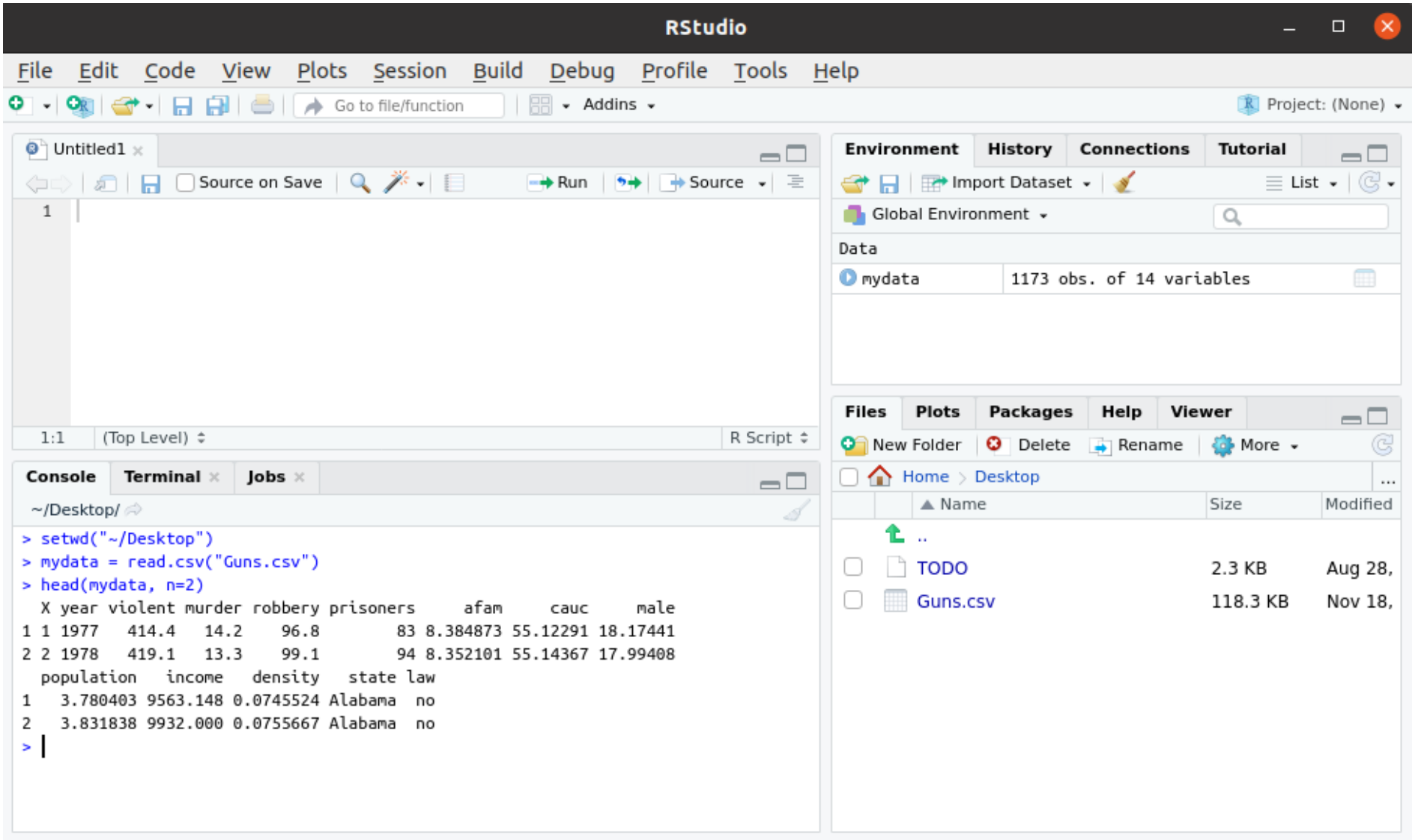


Loading Data

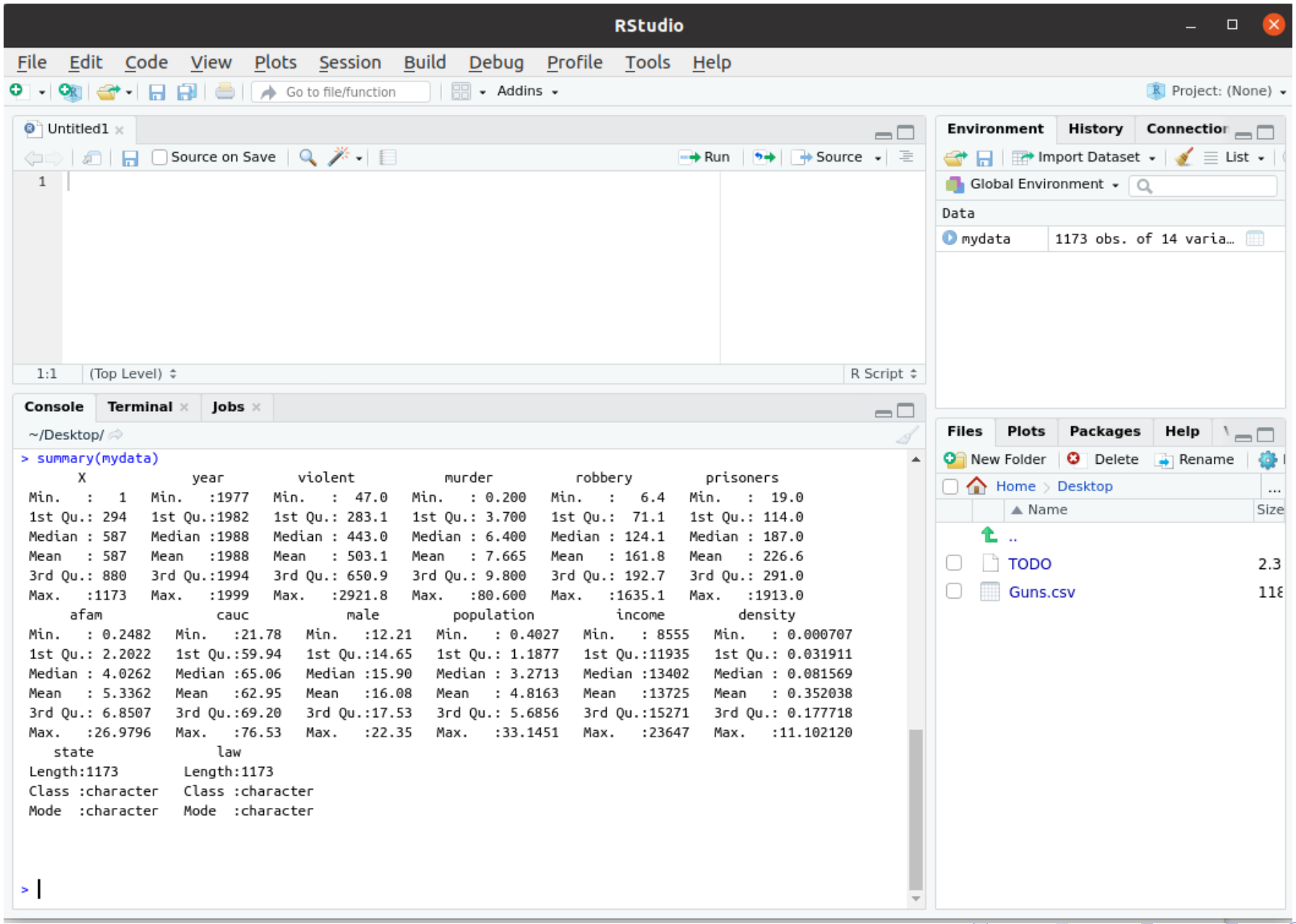
- I'm going to grab some public data from https://vincentarelbundock.github.io/Rdatasets/datasets.html
- Place file in Desktop, and change the working directory



Loading Data



Loading Data



Histogram

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1 | Source on Save | Source

1

1:1 (Top Level) R Script

Environment History Connections Tutorial

Import Dataset | List

Global Environment

Data

mydata 1173 obs. of 14 variables

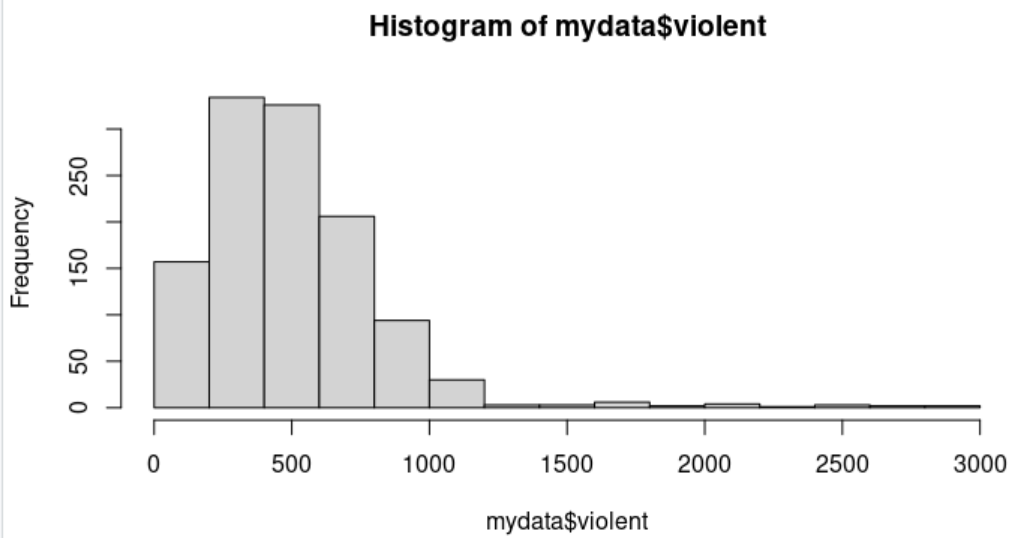
Console Terminal Jobs

~/Desktop/

```
 3rd Qu.: 192.7   3rd Qu.: 291.0
 Max.   :1173   Max.   :1999   Max.   :2921.8   Max.   :80.600
 Max.   :1635.1   Max.   :1913.0
      afam              cauc             male          population
       income          density
 Min.   : 0.2482   Min.   :21.78   Min.   :12.21   Min.   : 0.402
7   Min.   : 8555   Min.   : 0.000707
 1st Qu.: 2.2022   1st Qu.:59.94   1st Qu.:14.65   1st Qu.: 1.187
7   1st Qu.:11935   1st Qu.: 0.031911
 Median : 4.0262   Median :65.06   Median :15.90   Median : 3.271
3   Median :13402   Median : 0.081569
 Mean   : 5.3362   Mean   :62.95   Mean   :16.08   Mean   : 4.816
3   Mean   :13725   Mean   : 0.352038
 3rd Qu.: 6.8507   3rd Qu.:69.20   3rd Qu.:17.53   3rd Qu.: 5.685
6   3rd Qu.:15271   3rd Qu.: 0.177718
 Max.   :26.9796   Max.   :76.53   Max.   :22.35   Max.   :33.145
1   Max.   :23647   Max.   :11.102120
    state               law
 Length:1173        Length:1173
 Class :character   Class :character
 Mode  :character   Mode  :character


> hist(mydata$violent)
>
```

Files Plots Packages Help Viewer

Zoom | Export | Publish



Histogram

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1 | Source on Save | Source

1

1:1 (Top Level) R Script

Environment History Connections Tutorial

Import Dataset | List

Global Environment

Data

mydata 1173 obs. of 14 variables

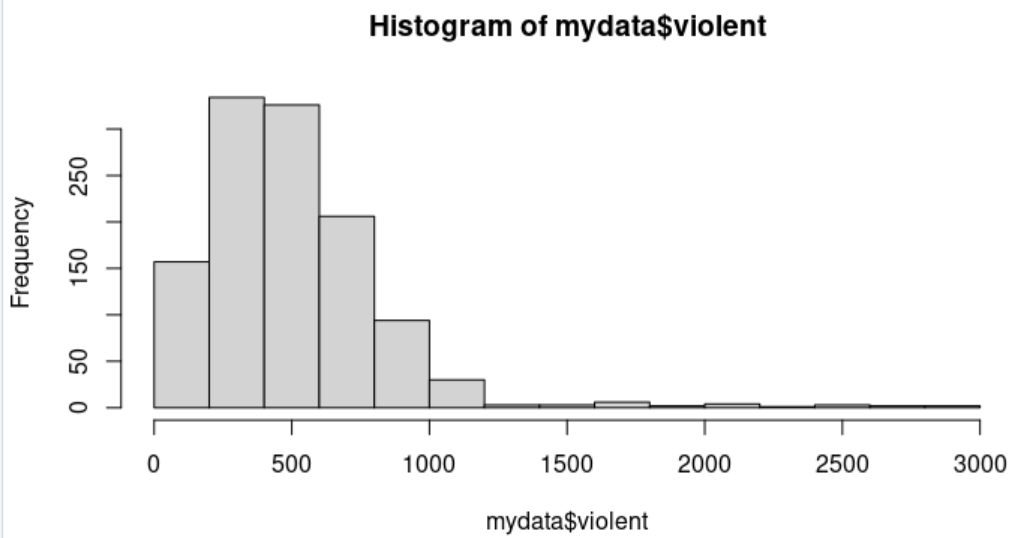
Console Terminal Jobs

~/Desktop/

```
 3rd Qu.: 192.7   3rd Qu.: 291.0
 Max.   :1173   Max.   :1999   Max.   :2921.8   Max.   :80.600
 Max.   :1635.1   Max.   :1913.0
      afam              cauc             male          population
       income          density
 Min.   : 0.2482   Min.   :21.78   Min.   :12.21   Min.   : 0.402
7   Min.   : 8555   Min.   : 0.000707
 1st Qu.: 2.2022   1st Qu.:59.94   1st Qu.:14.65   1st Qu.: 1.187
7   1st Qu.:11935   1st Qu.: 0.031911
 Median : 4.0262   Median :65.06   Median :15.90   Median : 3.271
3   Median :13402   Median : 0.081569
 Mean   : 5.3362   Mean   :62.95   Mean   :16.08   Mean   : 4.816
3   Mean   :13725   Mean   : 0.352038
 3rd Qu.: 6.8507   3rd Qu.:69.20   3rd Qu.:17.53   3rd Qu.: 5.685
6   3rd Qu.:15271   3rd Qu.: 0.177718
 Max.   :26.9796   Max.   :76.53   Max.   :22.35   Max.   :33.145
1   Max.   :23647   Max.   :11.102120
    state               law
 Length:1173        Length:1173
 Class :character   Class :character
 Mode  :character   Mode  :character


> hist(mydata$violent)
>
```

Files Plots Packages Help Viewer

Zoom | Export | Publish



Histogram

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1*

Source on Save | Source

```
1  setwd("~/Desktop")
2  mydata = read.csv("Guns.csv")
3  head(mydata, n=2)
4  summary(mydata)
5  hist(mydata$violent)
6  Violence <- mydata$violent
7  png(file="ViolenceHist.png",
8      units="in", width=5, height=5, res=300)
9  hist(Violence, col="gold")
10 dev.off()
11 hist(Violence, col="gold")
```

7:1 (Top Level) R Script

Console Terminal Jobs

~/Desktop/

```
> Violence <- mydata$violent
> png(file="ViolenceHist.png",
+     units="in", width=5, height=5, res=300)
> hist(Violence, col="gold")
> dev.off()
RStudioGD
        2
> hist(Violence, col="gold")
>
```

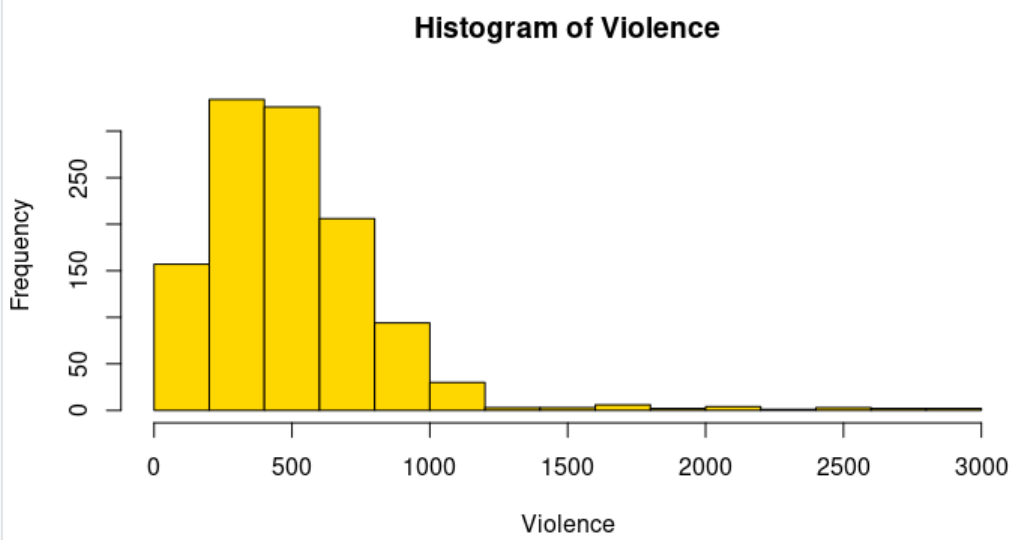
Environment History Connections Tutorial

To Console | To Source

```
summary(mydata)
summary(mydata)
hist(mydata$violent)
violent <- mydata$violent
clear(violent)
rm(violent)
Violence <- mydata$violent
png(file="ViolenceHist.png",
units="in", width=5, height=5, res=300)
hist(Violence, col="gold")
dev.off()
hist(Violence, col="gold")
```

Files Plots Packages Help Viewer

Zoom | Export | Publish



Scatterplot

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1*

Source on Save | Source

```
1  setwd("~/Desktop")
2  mydata = read.csv("Guns.csv")
3  head(mydata, n=2)
4  summary(mydata)
5  hist(mydata$violent)
6  Violence <- mydata$violent
7  png(file="ViolenceHist.png",
8      units="in", width=5, height=5, res=300)
9  hist(Violence, col="gold")
10 dev.off()
11 hist(Violence, col="gold")
```

7:1 (Top Level) R Script

Environment History Connections Tutorial

To Console | To Source

```
summary(mydata)
hist(mydata$violent)
violent <- mydata$violent
clear(violent)
rm(violent)
Violence <- mydata$violent
png(file="ViolenceHist.png",
units="in", width=5, height=5, res=300)
hist(Violence, col="gold")
dev.off()
hist(Violence, col="gold")
plot(Violence)
```

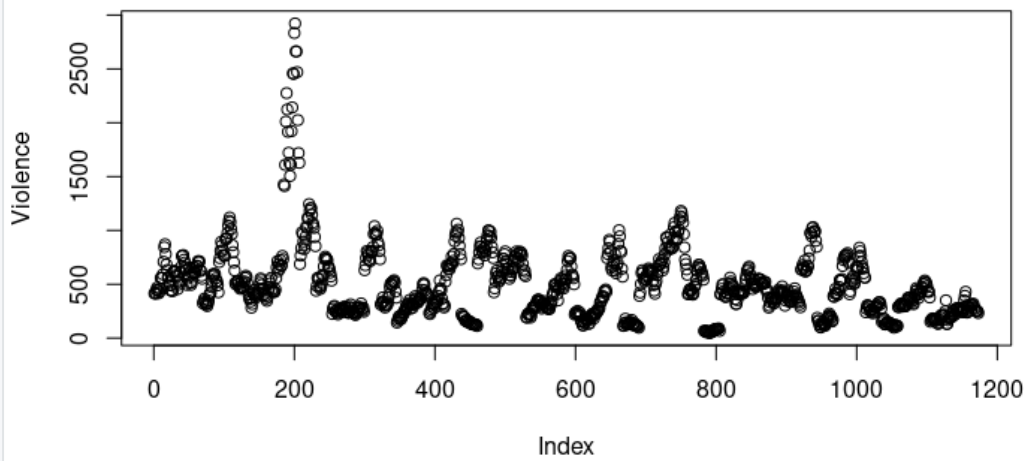
Console Terminal Jobs

~/Desktop/

```
> Violence <- mydata$violent
> png(file="ViolenceHist.png",
+     units="in", width=5, height=5, res=300)
> hist(Violence, col="gold")
> dev.off()
RStudioGD
        2
> hist(Violence, col="gold")
> plot(Violence)
>
```

Files Plots Packages Help Viewer

Zoom | Export | Publish



Scatterplot

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1* | mydata

Filter

	X	year	violent	murder	robbery	prisoners	afam	ca
1	1	1977	414.4	14.2	96.8	83	8.384873	
2	2	1978	419.1	13.3	99.1	94	8.352101	
3	3	1979	413.3	13.2	109.5	144	8.329575	
4	4	1980	448.5	13.2	132.1	141	8.408386	
5	5	1981	470.5	11.9	126.5	149	8.483435	
6	6	1982	447.7	10.6	112.0	183	8.514000	

Showing 1 to 6 of 1,173 entries, 14 total columns

Environment History Connections Tutorial

Import Dataset | List

Global Environment

Data

mydata 1173 obs. of 14 variables

Values

Violence num [1:1173] 414 419 413 448 470 ...

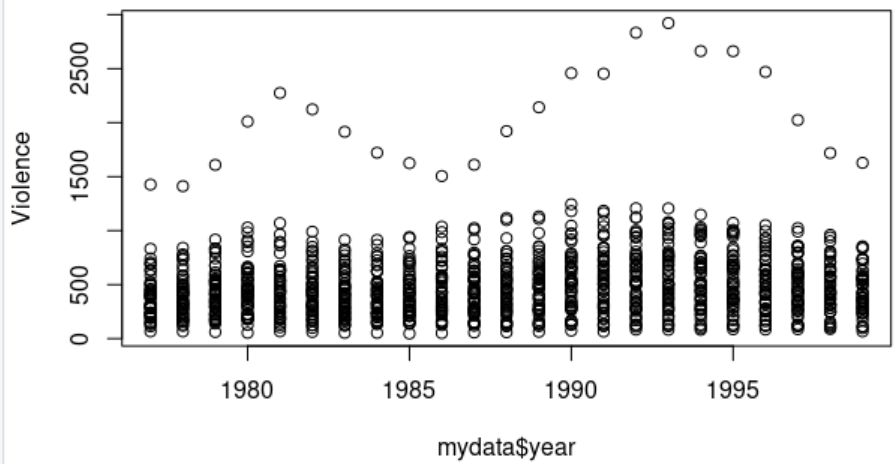
Console Terminal Jobs

~/Desktop/

```
> Violence <- mydata$violent
> png(file="ViolenceHist.png",
+     units="in", width=5, height=5, res=300)
> hist(Violence, col="gold")
> dev.off()
RStudioGD
        2
> hist(Violence, col="gold")
> plot(Violence)
> plot(mydata$year,Violence)
> View(mydata)
>
```

Files Plots Packages Help Viewer

Zoom | Export | Publish



Scatterplot

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1* | mydata

Filter

	cauc	male	population	income	density	state	law
873	55.12291	18.17441	3.780403	9563.148	0.0745524	Alabama	no
101	55.14367	17.99408	3.831838	9932.000	0.0755667	Alabama	no
575	55.13586	17.83934	3.866248	9877.028	0.0762453	Alabama	no
386	54.91259	17.73420	3.900368	9541.428	0.0768288	Alabama	no
435	54.92513	17.67372	3.918531	9548.351	0.0771866	Alabama	no
000	54.89621	17.51052	3.925229	9478.919	0.0773185	Alabama	no

Showing 1 to 6 of 1,173 entries, 14 total columns

Environment | History | Connections | Tutorial

Import Dataset | List

Global Environment

Data

mydata — 1173 obs. of 14 variables

Values

Violence — num [1:1173] 414 419 413 448 470 ...

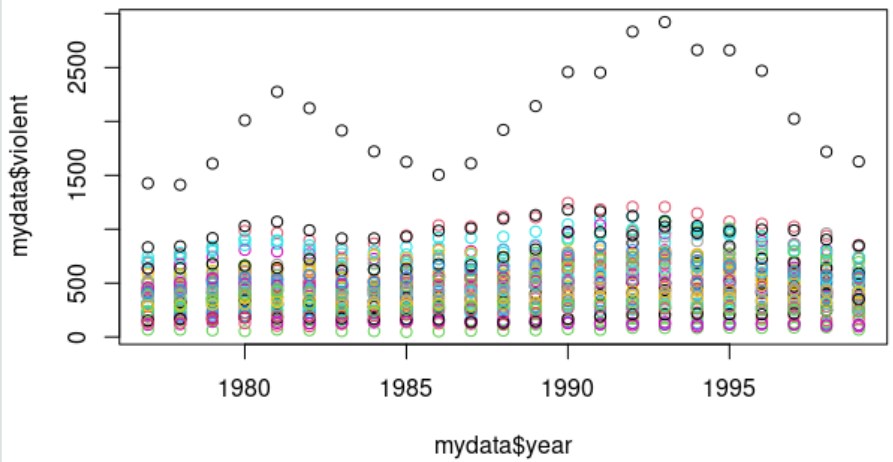
Console | Terminal | Jobs

~/Desktop/

```
> plot(mydata$year, mydata$violent, col=mydata$state)
Error in plot.xy(xy, type, ...) : invalid color name 'Alabama'
> str(mydata$state)
 chr [1:1173] "Alabama" "Alabama" "Alabama" "Alabama" "Alabama" ...
> mydata$state <- as.factor(mydata$state)
> str(mydata$state)
 Factor w/ 51 levels "Alabama","Alaska",..: 1 1 1 1 1 1 1 1 1 1 ...
> plot(mydata$year, mydata$violent, col=mydata$state)
>
```

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



Scatterplot

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1* | mydata

Filter

	cauc	male	population	income	density	state	law
873	55.12291	18.17441	3.780403	9563.148	0.0745524	Alabama	no
101	55.14367	17.99408	3.831838	9932.000	0.0755667	Alabama	no
575	55.13586	17.83934	3.866248	9877.028	0.0762453	Alabama	no
386	54.91259	17.73420	3.900368	9541.428	0.0768288	Alabama	no
435	54.92513	17.67372	3.918531	9548.351	0.0771866	Alabama	no
000	54.89621	17.51052	3.925229	9478.919	0.0773185	Alabama	no

Showing 1 to 6 of 1,173 entries, 14 total columns

Environment | History | Connections | Tutorial

To Console | To Source

```
legend("topright",legend=unique(mydata$state), fill=unique(mydata$state), cex=0.3)
legend("outside",legend=unique(mydata$state), fill=unique(mydata$state), cex=0.3)
legend("right",legend=unique(mydata$state), fill=unique(mydata$state), cex=0.3)
plot(mydata$year, mydata$violent, col=mydata$state)
legend("right",legend=unique(mydata$state), fill=unique(mydata$state), cex=0.3)
plot(mydata$year, mydata$violent, col=mydata$state, main="Violence per Year by State", xlab…
legend("right",legend=unique(mydata$state), fill=unique(mydata$state), cex=0.3)
plot(mydata$year, mydata$violent, col=mydata$state, main="Violence per Year by State", xlab…
legend("right",legend=unique(mydata$state), fill=unique(mydata$state), cex=0.3)
palette(rainbow(51))
plot(mydata$year, mydata$violent, col=mydata$state, main="Violence per Year by State", xlab…
legend("right",legend=unique(mydata$state), fill=unique(mydata$state), cex=0.3)
```

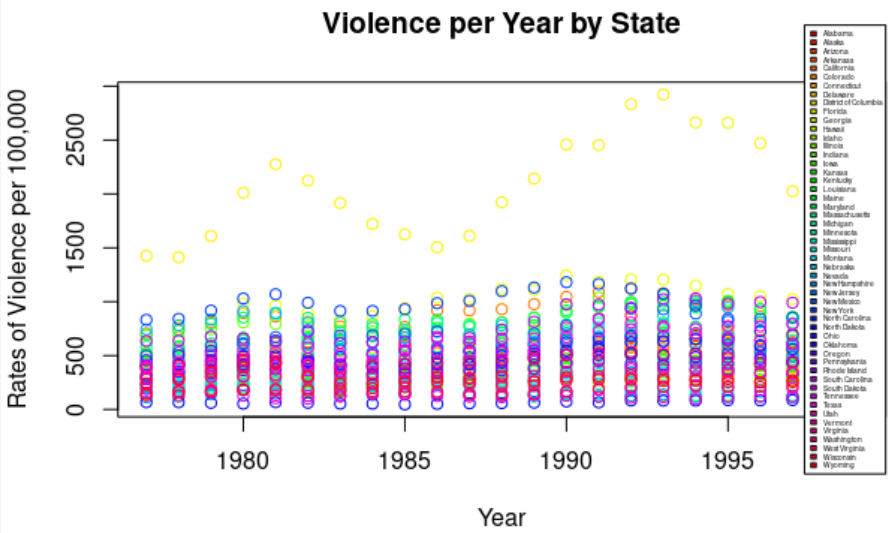
Console | Terminal | Jobs

~/Desktop/

```
> palette(rainbow(51))
> plot(mydata$year, mydata$violent, col=mydata$state, main="Violence per Year b
y State", xlab="Year", ylab="Rates of Violence per 100,000")
> legend("right",legend=unique(mydata$state), fill=unique(mydata$state), cex=0.
3)
>
```

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



Boxplot

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1* | mydata

Filter

	X	year	violent	murder	robbery	prisoners	afam	cau
1	1	1977	414.4	14.2	96.8	83	8.384873	5
2	2	1978	419.1	13.3	99.1	94	8.352101	5
3	3	1979	413.3	13.2	109.5	144	8.329575	5
4	4	1980	448.5	13.2	132.1	141	8.408386	5
5	5	1981	470.5	11.9	126.5	149	8.483435	5
6	6	1982	447.7	10.6	112.0	183	8.514000	5

Showing 1 to 6 of 1,173 entries, 14 total columns

Environment | History | Connections | Tutorial

Import Dataset | List

Global Environment

Data

mydata — 1173 obs. of 14 variables

Values

Violence — num [1:1173] 414 419 413 448 470 ...

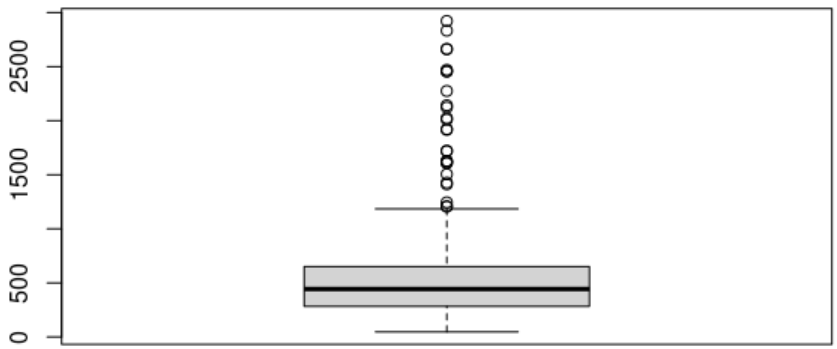
Console | Terminal | Jobs

~/Desktop/

```
> boxplot(Violence)
>
```

Files | Plots | Packages | Help | Viewer

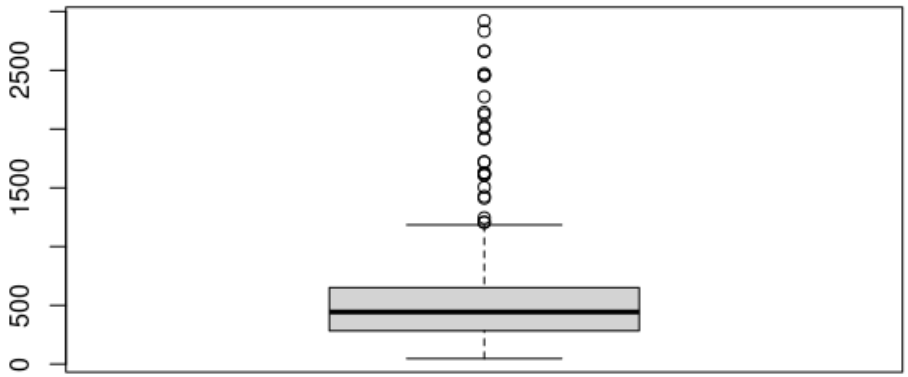
Zoom | Export | Publish



Boxplot

Boxplot Elements

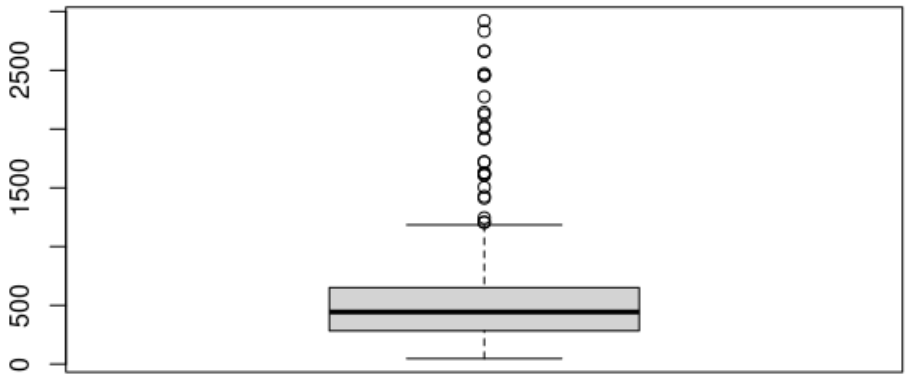
- Median (Q_2): middle value of the dataset



Boxplot

Boxplot Elements

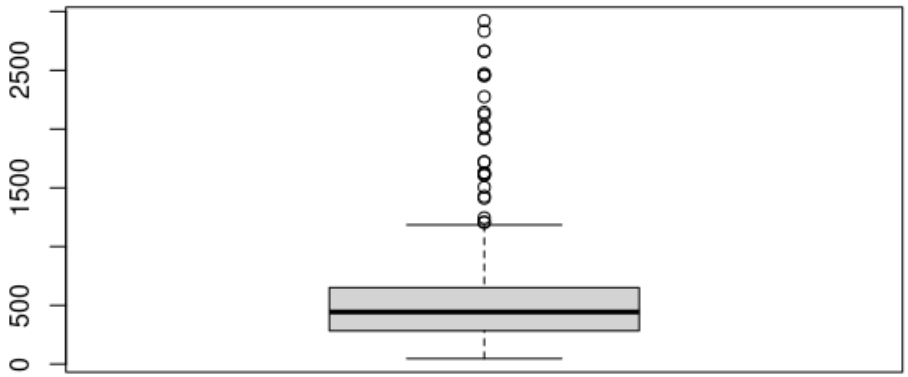
- Median (Q_2): middle value of the dataset
- First/lower quartile (Q_1): median of the lower half of the values



Boxplot

Boxplot Elements

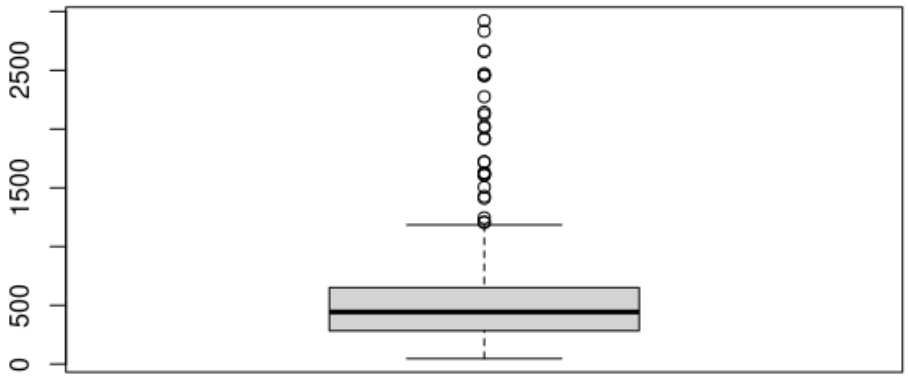
- Median (Q_2): middle value of the dataset
- First/lower quartile (Q_1): median of the lower half of the values
- Third / upper quartile (Q_3): median of the upper half of the values



Boxplot

Boxplot Elements

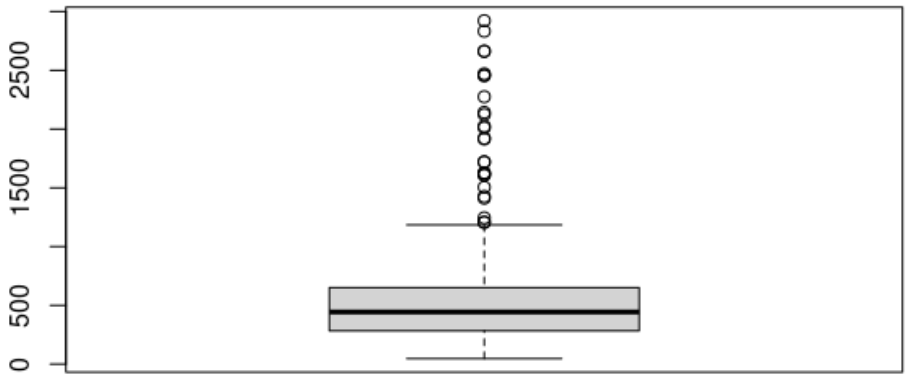
- Median (Q_2): middle value of the dataset
- First/lower quartile (Q_1): median of the lower half of the values
- Third / upper quartile (Q_3): median of the upper half of the values
- The interquartile range, or IQR, is Q_3 - Q_1



Boxplot

Boxplot Elements

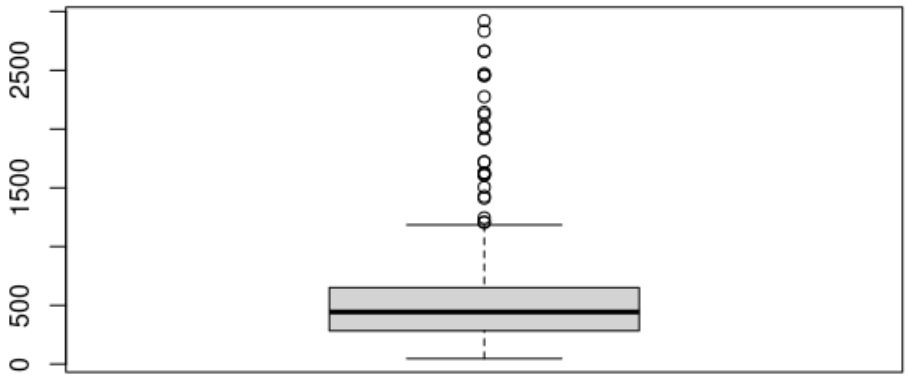
- Median (Q_2): middle value of the dataset
- First/lower quartile (Q_1): median of the lower half of the values
- Third / upper quartile (Q_3): median of the upper half of the values
- The interquartile range, or IQR, is Q_3 - Q_1
- Whiskers are the max value that lies within 1.5xIQR of first and third quartiles



Boxplot

Boxplot Elements

- Median (Q_2): middle value of the dataset
- First/lower quartile (Q_1): median of the lower half of the values
- Third / upper quartile (Q_3): median of the upper half of the values
- The interquartile range, or IQR, is Q_3 - Q_1
- Whiskers are the max value that lies within 1.5xIQR of first and third quartiles
- Outliers are circles that don't lie within the box or whiskers



Boxplot

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1* | mydata

Filter

	X	year	violent	murder	robbery	prisoners	afam	cau
1	1	1977	414.4	14.2	96.8	83	8.384873	5
2	2	1978	419.1	13.3	99.1	94	8.352101	5
3	3	1979	413.3	13.2	109.5	144	8.329575	5
4	4	1980	448.5	13.2	132.1	141	8.408386	5
5	5	1981	470.5	11.9	126.5	149	8.483435	5
6	6	1982	447.7	10.6	112.0	183	8.514000	5

Showing 1 to 6 of 1,173 entries, 14 total columns

Environment | History | Connections | Tutorial

Import Dataset | List

Global Environment

Data

mydata — 1173 obs. of 14 variables

Values

Violence — num [1:1173] 414 419 413 448 470 ...

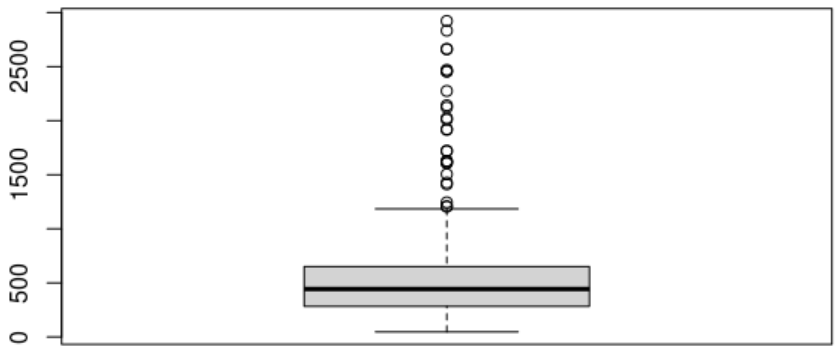
Console | Terminal | Jobs

~/Desktop/

```
> boxplot(Violence)
>
```

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



Packages

What are packages?

- Packages are sets of *functions* developed by the community

Packages

What are packages?

- Packages are sets of *functions* developed by the community
- They increase the power of R by improving existing base R functionalities, or by adding new ones

Packages

What are packages?

- Packages are sets of *functions* developed by the community
- They increase the power of R by improving existing base R functionalities, or by adding new ones
- Can be loaded with `library(packagename)`

Packages

What are packages?

- Packages are sets of *functions* developed by the community
- They increase the power of R by improving existing base R functionalities, or by adding new ones
- Can be loaded with `library(packagename)`
- If you don't have the package yet, you can install it with `install.packages("packagename")`

Boxplot

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1* | mydata | stats | stats

Filter

	item	group1	vars	n	mean	sd	median
X16	6	Colorado	1	23	482.78696	63.53352	49
X17	7	Connecticut	1	23	420.67391	65.36586	41
X18	8	Delaware	1	23	564.35652	122.20878	55
X19	9	District of Columbia	1	23	2048.97826	466.43853	201
X110	10	Florida	1	23	999.23478	151.54791	102
X111	11	Georgia	1	23	595.57826	99.63111	57

Showing 6 to 11 of 51 entries, 15 total columns

Environment | History | Connections | Tutorial

Import Dataset | List

Global Environment

Data

mydata	1173 obs. of 14 variables
stats	51 obs. of 15 variables

Values

outliers	2048.97826086957
Violence	num [1:1173] 414 419 413 448 470 ...

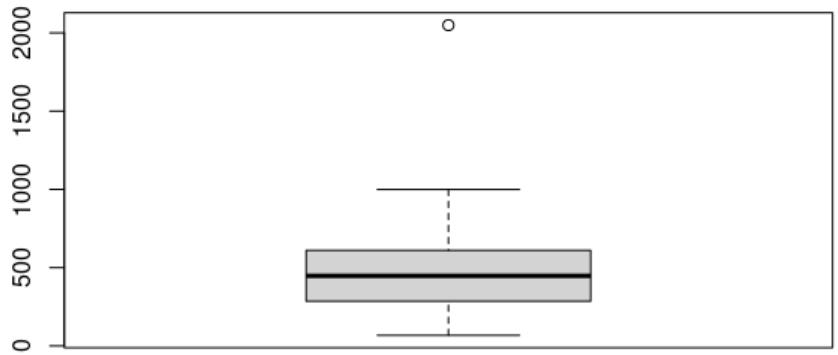
Console | Terminal | Jobs

~/Desktop/

```
> library(psych)
> stats <- describeBy(mydata$violent, mydata$state, mat=TRUE)
> stats$mean
 [1]  558.17391  596.79565  608.09565  438.40000  877.48261  482.78696
 [7]  420.67391  564.35652 2048.97826  999.23478  595.57826  259.43478
[13]  262.64348  828.16522  399.86957  247.60000  391.70870  334.69565
[19]  778.43913  154.04783  853.63478  616.71739  695.21304  277.26522
[25]  355.52609  582.65217  183.47826  300.73478  755.03478  130.84783
[31]  552.82174  743.86957  941.31739  526.39565   68.00435  445.45217
[37]  491.58696  496.33913  383.45652  369.40000  800.26522  157.36087
[43]  580.68261  612.65652  283.50870  133.20870  322.95217  447.72174
[49]  188.76087  225.28261  287.50000
> boxplot(stats$mean)
> outliers = boxplot(stats$mean, plot=FALSE)$out
> stats[stats$mean %in% outliers,]
    item              group1 vars  n     mean       sd median  trimmed
X19    9 District of Columbia    1 23 2048.978 466.4385 2010.6 2028.089
         mad    min    max  range      skew  kurtosis       se
X19 593.4848 1411.7 2921.8 1510.1 0.3306733 -1.260257 97.25916
> View(stats)
>
```

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



Boxplot

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1* | mydata | stats | stats

Filter

	item	group1	vars	n	mean	sd	median
X16	6	Colorado	1	23	482.78696	63.53352	49
X17	7	Connecticut	1	23	420.67391	65.36586	41
X18	8	Delaware	1	23	564.35652	122.20878	55
X19	9	District of Columbia	1	23	2048.97826	466.43853	201
X110	10	Florida	1	23	999.23478	151.54791	102
X111	11	Georgia	1	23	595.57826	99.63111	57

Showing 6 to 11 of 51 entries, 15 total columns

Console | Terminal | Jobs

~/Desktop/

```
> boxplot(Violence ~ mydata$state)
>
```

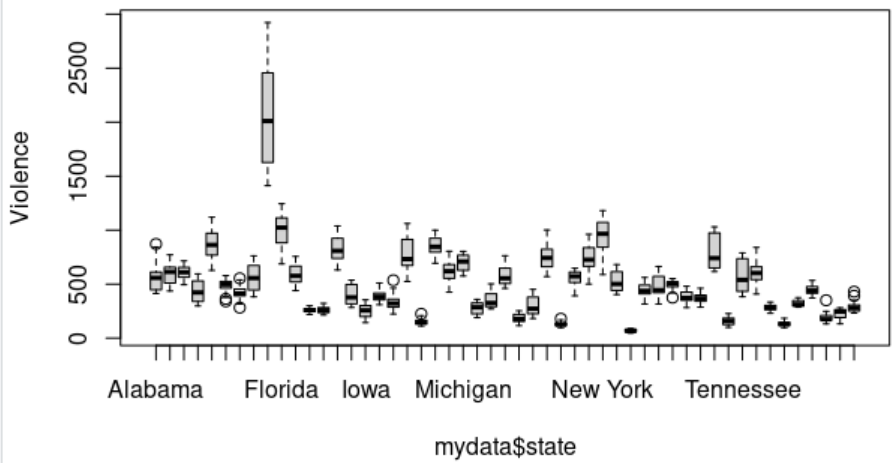
Environment | History | Connections | Tutorial

To Console | To Source

```
rm(outliers)
rm (stats)
boxplot(Violence)
boxplot(Violence ~ mydata$state)
boxplot(Violence ~ mydata$state, las=2)
par(cex.axis=1.5)
boxplot(Violence ~ mydata$state, las=2)
par(cex.axis=.5)
boxplot(Violence ~ mydata$state, las=2)
boxplot(Violence ~ mydata$state, las=2, cex.axis=.5)
boxplot(Violence ~ mydata$state, las=2, cex.axis=1)
boxplot(Violence ~ mydata$state, las=2, cex.axis=.5)
```

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



Boxplot

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1* | mydata | stats | stats

Filter

	X	year	violent	murder	robbery	prisoners	afam	cau
7	7	1983	416.0	9.2	98.4	215	8.545608	5
8	8	1984	431.2	9.4	96.1	243	8.559511	5
9	9	1985	457.5	9.8	105.4	256	8.562801	5
10	10	1986	558.0	10.1	111.6	267	8.566521	5
11	11	1987	559.2	9.3	112.2	283	8.592103	5
12	12	1988	558.6	9.9	117.8	307	8.618144	5

Showing 6 to 12 of 1,173 entries, 14 total columns

Environment | History | Connections | Tutorial

Import Dataset | List

Global Environment

Data	
mydata	1173 obs. of 14 variables
stats	51 obs. of 15 variables
Values	
outliers	2048.97826086957
Violence	num [1:1173] 414 419 413 448 470 ...

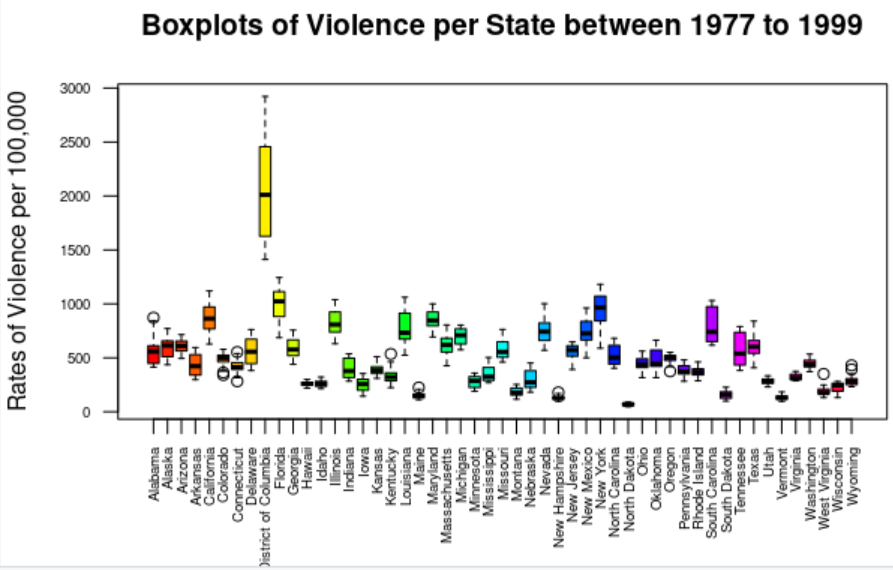
Console | Terminal | Jobs

~/Desktop/

```
> boxplot(Violence ~ mydata$state, las=2, cex.axis=.6, col=unique(mydata$stat
e), xlab = "", main = "Boxplots of Violence per State between 1977 to 1999", yl
ab = "Rates of Violence per 100,000")
>
```

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



tidyverse and ggplot2

tidyverse

- tidyverse is a collection of packages for tidy data

tidyverse and ggplot2

tidyverse

- tidyverse is a collection of packages for tidy data
- The core packages are ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr, and forcats

tidyverse and ggplot2

tidyverse

- tidyverse is a collection of packages for tidy data
- The core packages are ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr, and forcats
- As of November 2018, the packages make up 5 out of the top 10 most downloaded R packages

tidyverse and ggplot2

tidyverse

- tidyverse is a collection of packages for tidy data
- The core packages are ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr, and forcats
- As of November 2018, the packages make up 5 out of the top 10 most downloaded R packages
- `install.packages("tidyverse")`

tidyverse and ggplot2

tidyverse

- tidyverse is a collection of packages for tidy data
- The core packages are ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr, and forcats
- As of November 2018, the packages make up 5 out of the top 10 most downloaded R packages
- `install.packages("tidyverse")`

ggplot2

- ggplot2 is a data visualization package

tidyverse and ggplot2

tidyverse

- tidyverse is a collection of packages for tidy data
- The core packages are ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr, and forcats
- As of November 2018, the packages make up 5 out of the top 10 most downloaded R packages
- `install.packages("tidyverse")`

ggplot2

- ggplot2 is a data visualization package
- follows 'Grammar of Graphics': a general scheme for data visualization which breaks up graphs into semantic components such as scales and layers

tidyverse and ggplot2

tidyverse

- tidyverse is a collection of packages for tidy data
- The core packages are ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr, and forcats
- As of November 2018, the packages make up 5 out of the top 10 most downloaded R packages
- `install.packages("tidyverse")`

ggplot2

- ggplot2 is a data visualization package
- follows 'Grammar of Graphics': a general scheme for data visualization which breaks up graphs into semantic components such as scales and layers
- `library(tidyverse)`

Correlation

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1* | meddata

Filter

	X	Accept	Acceptance	Sex	BCPM	GPA	VR	PS	W
1	1	D	0	F	3.59	3.62	11	9	
2	2	A	1	M	3.75	3.84	12	13	
3	3	A	1	F	3.24	3.23	9	10	
4	4	A	1	F	3.74	3.69	12	11	
5	5	A	1	F	3.53	3.38	9	11	
6	6	A	1	M	3.59	3.72	10	9	

Showing 1 to 6 of 55 entries, 12 total columns

Environment | History | Connections | Tutorial

Import Dataset | List

Global Environment

Data

meddata | 55 obs. of 12 variables

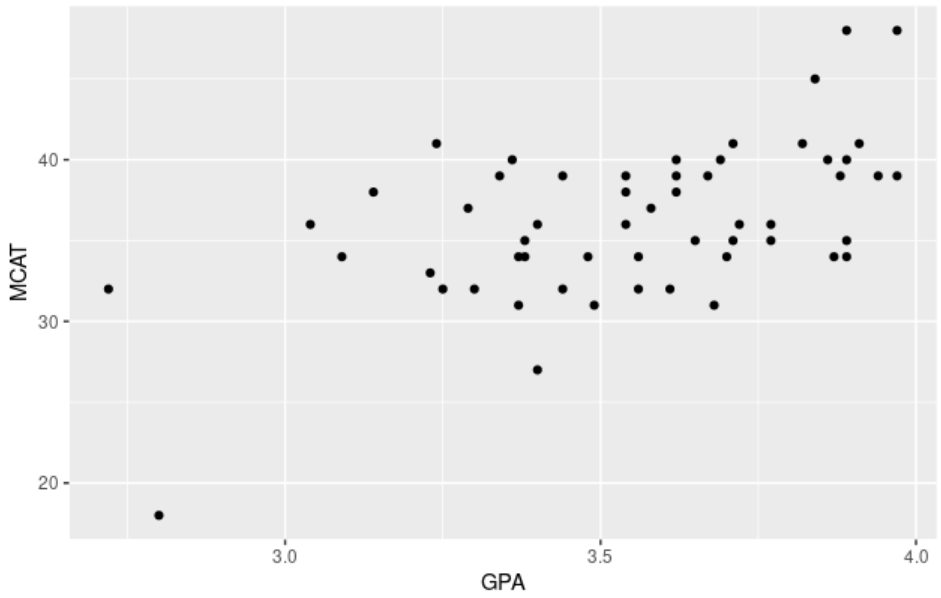
Console | Terminal | Jobs

~/Desktop/

```
*** copying figures
** building package indices
** installing vignettes
** testing if installed package can be loaded from temporary location
** testing if installed package can be loaded from final location
** testing if installed package keeps a record of temporary installation path
* DONE (tidyverse)

The downloaded source packages are in
        '/tmp/RtmpcXe70R/downloaded_packages'
> library(tidyverse)
── Attaching packages ─────────────────────────────────── tidyverse 1.3.0 ──
✓ ggplot2 3.3.2     ✓ purrr   0.3.4
✓ tibble  3.0.1     ✓ dplyr   1.0.0
✓ tidyr   1.1.0     ✓ stringr 1.4.0
✓ readr   1.3.1     ✓ forcats 0.5.0
── Conflicts ────────────────────────────────────── tidyverse_conflicts() ──
x ggplot2::%+%()   masks psych::%+%()
x ggplot2::alpha() masks psych::alpha()
x dplyr::filter()  masks stats::filter()
x dplyr::lag()     masks stats::lag()
> meddata <- read.csv("MedGPA.csv")
> rm(mydata,stats,outliers,Violence)
> View(meddata)
> ggplot(data = meddata, aes(x = GPA, y = MCAT)) + geom_point()
>
```

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



Correlation

What is correlation?

- Correlation is a statistical measure that suggests the level of linear dependence between two variables, that occur in pair

Correlation

What is correlation?

- Correlation is a statistical measure that suggests the level of linear dependence between two variables, that occur in pair
- Correlation can take values between -1 to $+1$

Correlation

What is correlation?

- Correlation is a statistical measure that suggests the level of linear dependence between two variables, that occur in pair
- Correlation can take values between -1 to $+1$
- If for every time an independent variable increases and the dependent variable increases, we say this is a positive correlation, and will have a value closer to 1

Correlation

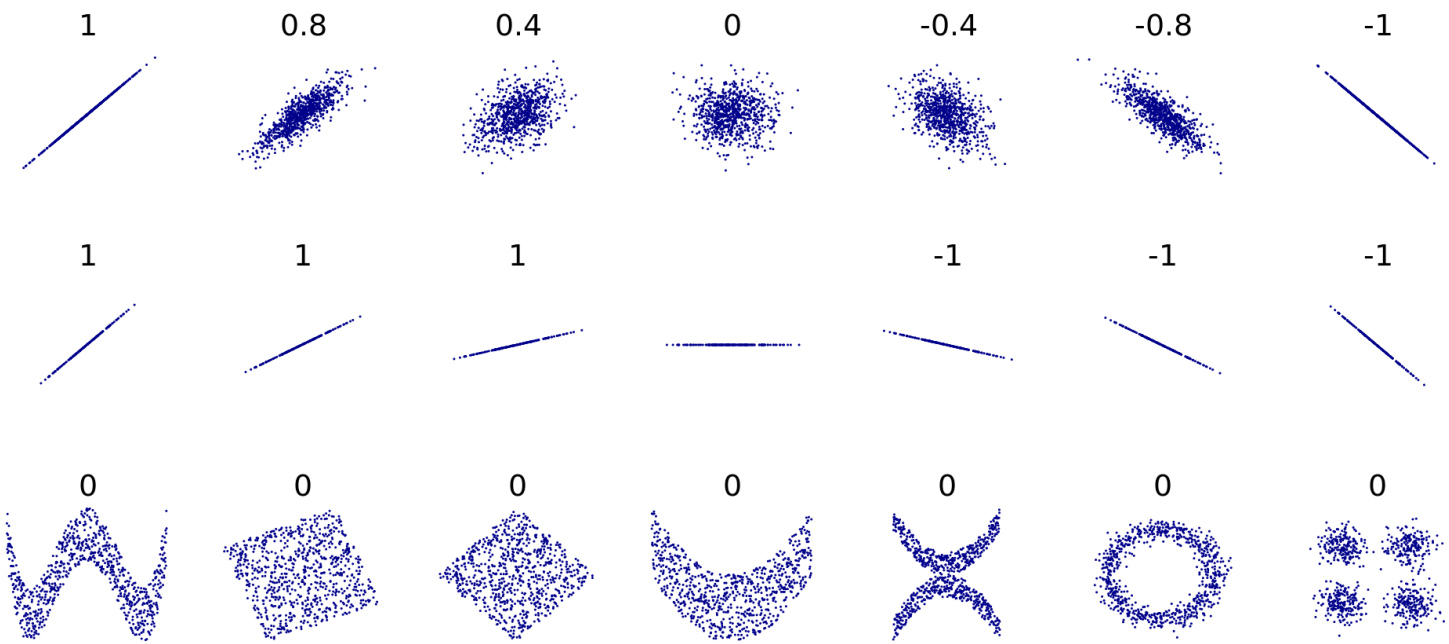
What is correlation?

- Correlation is a statistical measure that suggests the level of linear dependence between two variables, that occur in pair
- Correlation can take values between -1 to $+1$
- If for every time an independent variable increases and the dependent variable increases, we say this is a positive correlation, and will have a value closer to 1
- If for every time an independent variable increases and the dependent variable decreases, we say this is a negative correlation, and will have a value closer to -1

Correlation

Pearson Correlation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}}$$



Correlation

Two Assumptions

- The two variables are normally distributed. We can test this assumption using a histogram

Correlation

Two Assumptions

- The two variables are normally distributed. We can test this assumption using a histogram
- The relationship between the two variables is linear. If this relationship is found to be curved, etc. we need to use another correlation test. We can test this assumption by examining the scatterplot between the two variables.

Correlation

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1* | meddata

Filter

	X	Accept	Acceptance	Sex	BCPM	GPA	VR	PS	W
1	1	D	0	F	3.59	3.62	11	9	
2	2	A	1	M	3.75	3.84	12	13	
3	3	A	1	F	3.24	3.23	9	10	
4	4	A	1	F	3.74	3.69	12	11	
5	5	A	1	F	3.53	3.38	9	11	
6	6	A	1	M	3.59	3.72	10	9	

Showing 1 to 6 of 55 entries, 12 total columns

Console | Terminal | Jobs

~/Desktop/

```
> ggplot(data = meddata, aes(x = GPA, y = MCAT)) + geom_point()
>
```

Environment | History | Connections | Tutorial

Import Dataset | List

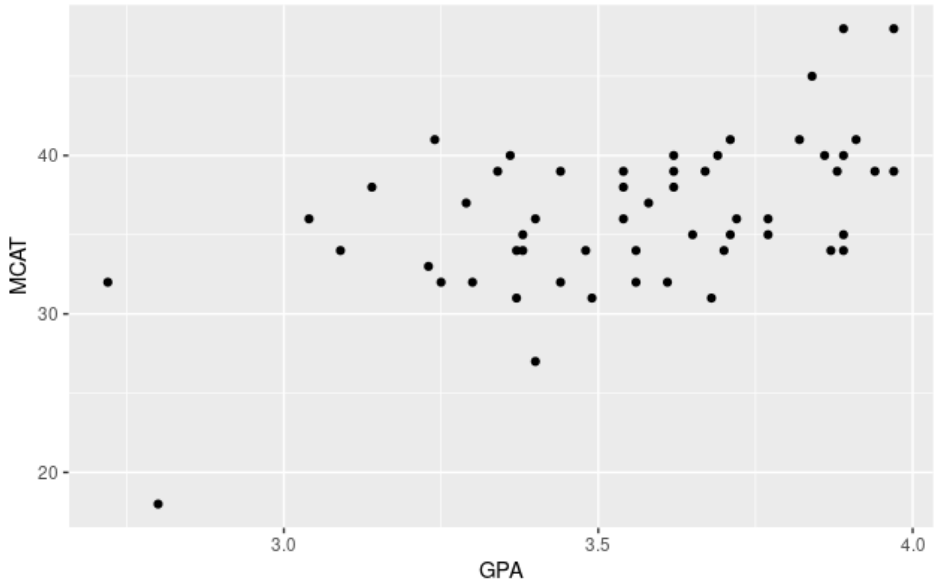
Global Environment

Data

meddata	55 obs. of 12 variables

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



Correlation

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1* | meddata

Filter

	X	Accept	Acceptance	Sex	BCPM	GPA	VR	PS	W:
1	1	D	0	F	3.59	3.62	11	9	
2	2	A	1	M	3.75	3.84	12	13	
3	3	A	1	F	3.24	3.23	9	10	
4	4	A	1	F	3.74	3.69	12	11	
5	5	A	1	F	3.53	3.38	9	11	
6	6	A	1	M	3.59	3.72	10	9	

Showing 1 to 6 of 55 entries, 12 total columns

Environment | History | Connections | Tutorial

Import Dataset | List

Global Environment

Data

meddata	55 obs. of 12 variables
plot.data	num [1:2, 1:55] 3.62 38 3.84 45 3.23 33 3.69 40 3.38 35 ...
plot1	List of 9
plot2	List of 9

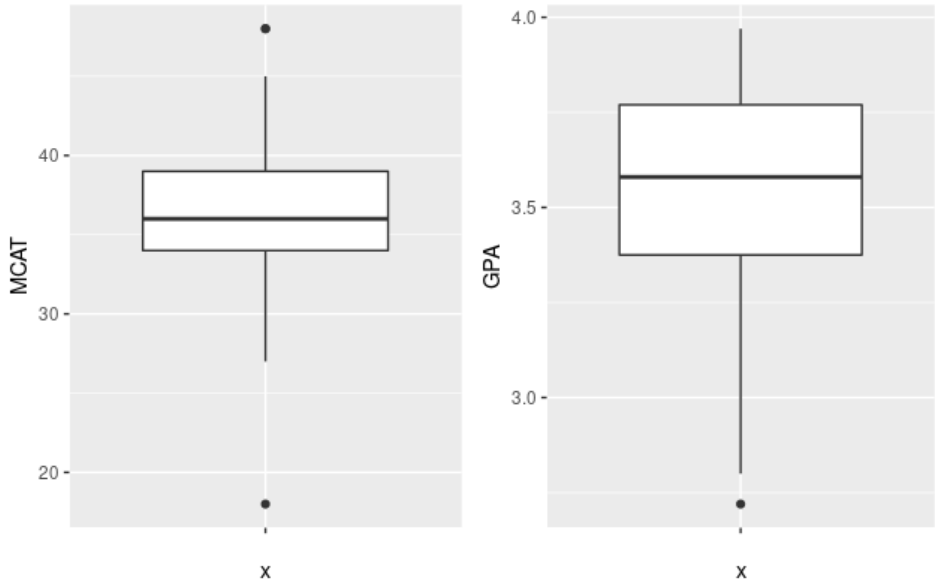
Console | Terminal | Jobs

~/Desktop/

```
** R
** inst
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
** building package indices
** installing vignettes
** testing if installed package can be loaded from temporary location
** testing if installed package can be loaded from final location
** testing if installed package keeps a record of temporary installation path
* DONE (gridExtra)

The downloaded source packages are in
        '/tmp/RtmpcXe70R/downloaded_packages'
> library(gridExtra)

Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':

    combine

> plot1 <- ggplot(data = meddata, aes(x = "", y = MCAT)) + geom_boxplot()
> plot2 <- ggplot(data = meddata, aes(x = "", y = GPA)) + geom_boxplot()
> grid.arrange(plot1,plot2,ncol=2)
>
```

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



A Word of Warning

To Remove or Keep Outliers?

There are generally three reasons why we may have an outlier:

1. There was a recording error
2. Someone made a fundamental mistake collecting the data
3. The data point is perfectly valid, in which case the model we are trying to use cannot account for the behavior.

"Sometimes we throw out perfectly good data when we should be throwing out questionable models."

Correlation

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1* | meddata

Filter

	X	Accept	Acceptance	Sex	BCPM	GPA	VR	PS	WS
1	1	D	0	F	3.59	3.62	11	9	
2	2	A	1	M	3.75	3.84	12	13	
3	3	A	1	F	3.24	3.23	9	10	
4	4	A	1	F	3.74	3.69	12	11	
5	5	A	1	F	3.53	3.38	9	11	
6	6	A	1	M	3.59	3.72	10	9	

Showing 1 to 6 of 55 entries, 12 total columns

Environment | History | Connections | Tutorial

Import Dataset | List

Global Environment

Data

meddata	55 obs. of 12 variables
meddata_nooutliers	51 obs. of 12 variables
plot1	List of 9
plot2	List of 9

Values

outliers_GPA	2.72
outliers_MCAT	num [1:3] 48 48 18

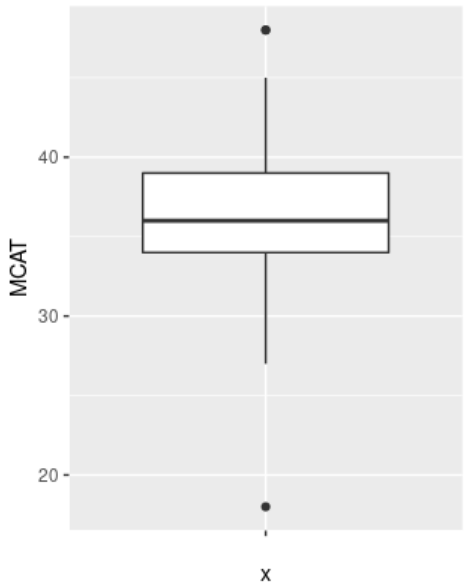
Console | Terminal | Jobs

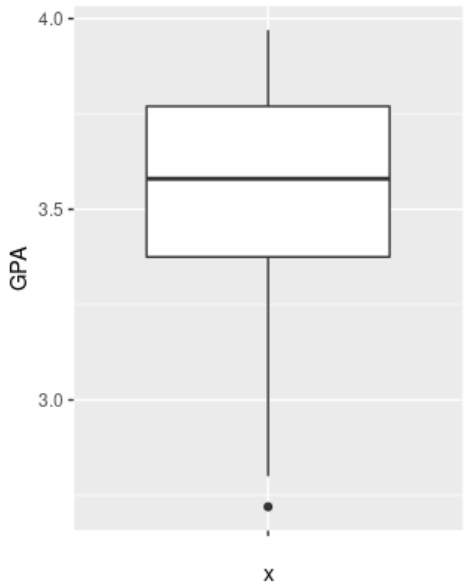
~/Desktop/

```
> outliers_MCAT = boxplot(meddata$MCAT, plot=FALSE)$out
> outliers_GPA = boxplot(meddata$GPA, plot=FALSE)$out
> meddata[meddata$MCAT %in% outliers_MCAT,]
    X Accept Acceptance Sex BCPM  GPA VR PS WS BS MCAT Apps
20 20      A          1   M 3.95 3.89 13 14  8 13   48    5
47 47      A          1   F 3.98 3.97 11 13 10 14   48    6
54 54      D          0   M 2.61 2.80  7  5 NA  6   18    6
> meddata[meddata$GPA %in% outliers_GPA,]
    X Accept Acceptance Sex BCPM  GPA VR PS WS BS MCAT Apps
51 51      D          0   M 2.41 2.72  8  8  8  8   32    7
> meddata_nooutliers <- meddata[-c(20,47,51,54),]
>
```

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish





Correlation

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1* | meddata

Filter

	X	Accept	Acceptance	Sex	BCPM	GPA	VR	PS	W:
1	1	D	0	F	3.59	3.62	11	9	
2	2	A	1	M	3.75	3.84	12	13	
3	3	A	1	F	3.24	3.23	9	10	
4	4	A	1	F	3.74	3.69	12	11	
5	5	A	1	F	3.53	3.38	9	11	
6	6	A	1	M	3.59	3.72	10	9	

Showing 1 to 6 of 55 entries, 12 total columns

Console | Terminal | Jobs

~/Desktop/

```
> ggplot(data = meddata_nooutliers, aes(x = GPA, y = MCAT)) + geom_point()
>
```

Environment | History | Connections | Tutorial

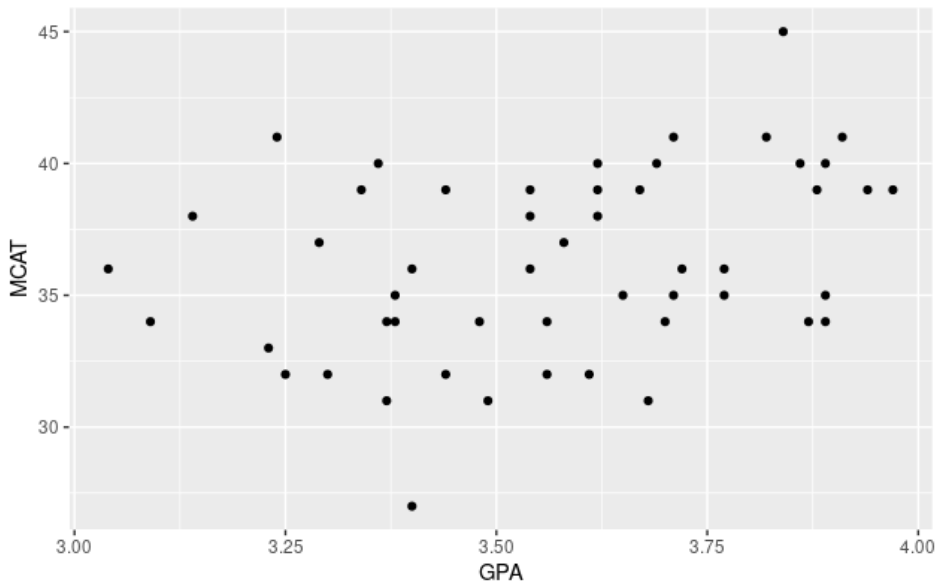
Import Dataset | List

Global Environment

Data	
meddata	55 obs. of 12 variables
meddata_nooutliers	51 obs. of 12 variables
plot1	List of 9
plot2	List of 9
Values	
outliers_GPA	2.72
outliers_MCAT	num [1:3] 48 48 18

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



Correlation

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1* | meddata

Filter

	X	Accept	Acceptance	Sex	BCPM	GPA	VR	PS
1	1	D	0	F	3.59	3.62	11	9
2	2	A	1	M	3.75	3.84	12	13
3	3	A	1	F	3.24	3.23	9	10
4	4	A	1	F	3.74	3.69	12	11
5	5	A	1	F	3.53	3.38	9	11
6	6	A	1	M	3.59	3.72	10	9

Showing 1 to 6 of 55 entries, 12 total columns

Environment | History | Connections | Tutorial

Import Dataset | List

Global Environment

Data	
meddata	55 obs. of 12 variables
meddata_nooutliers	51 obs. of 12 variables
plot1	List of 9
plot2	List of 9
Values	
outliers_GPA	2.72
outliers_MCAT	num [1:3] 48 48 18

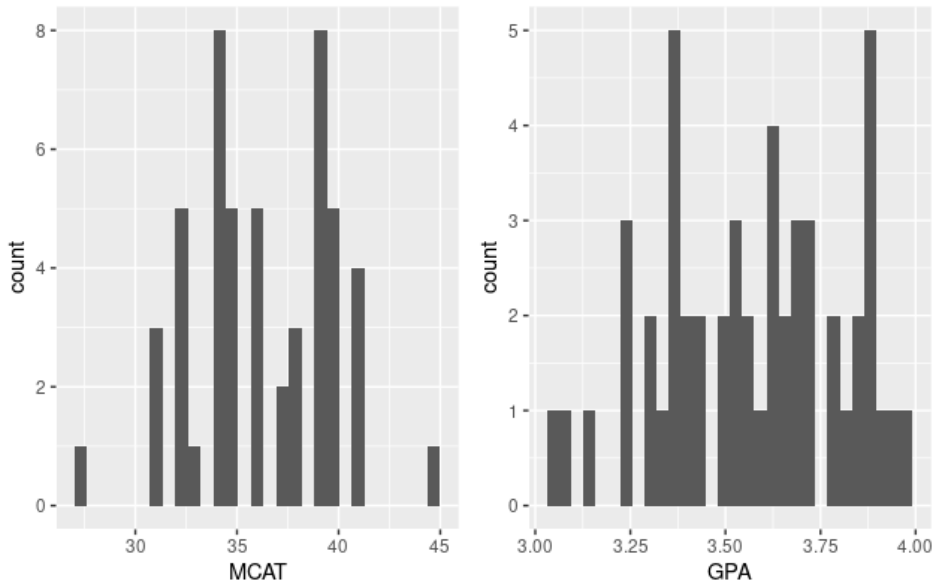
Console | Terminal | Jobs

~/Desktop/

```
> plot1 <- ggplot(meddata_nooutliers, aes(x=MCAT)) + geom_histogram()
> plot2 <- ggplot(meddata_nooutliers, aes(x=GPA)) + geom_histogram()
> grid.arrange(plot1,plot2,ncol=2)
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
>
```

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



Correlation

```
> cor(meddata_nooutliers$GPA,meddata_nooutliers$MCAT, method="pearson")
[1] 0.3350624
> cor.test(meddata_nooutliers$GPA,meddata_nooutliers$MCAT, method="pearson")

        Pearson's product-moment correlation

data:  meddata_nooutliers$GPA and meddata_nooutliers$MCAT
t = 2.4893, df = 49, p-value = 0.01624
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.06552957 0.55902680
sample estimates:
      cor
0.3350624

> 
```

Linear Regression

What is linear regression?

- linear regression is a linear approach to modelling the relationship between a dependent variable and an explanatory variable

Linear Regression

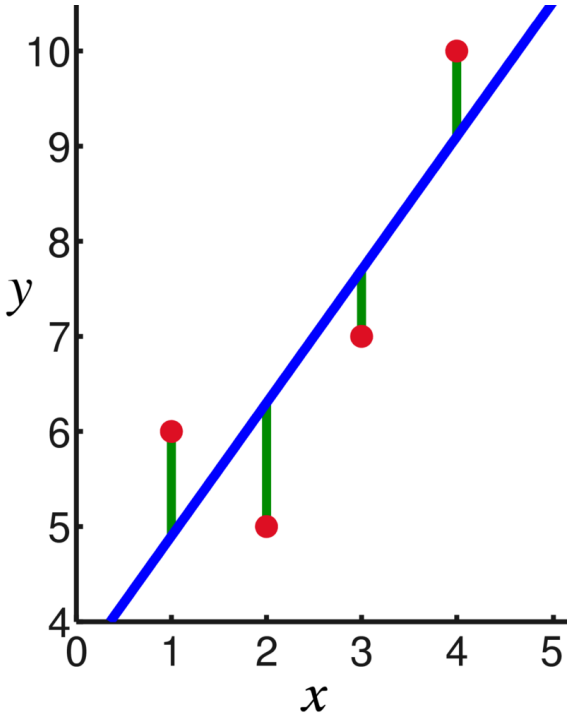
What is linear regression?

- linear regression is a linear approach to modelling the relationship between a dependent variable and an explanatory variable
- can be used to predict values of one variable based on the other

Linear Regression

What is linear regression?

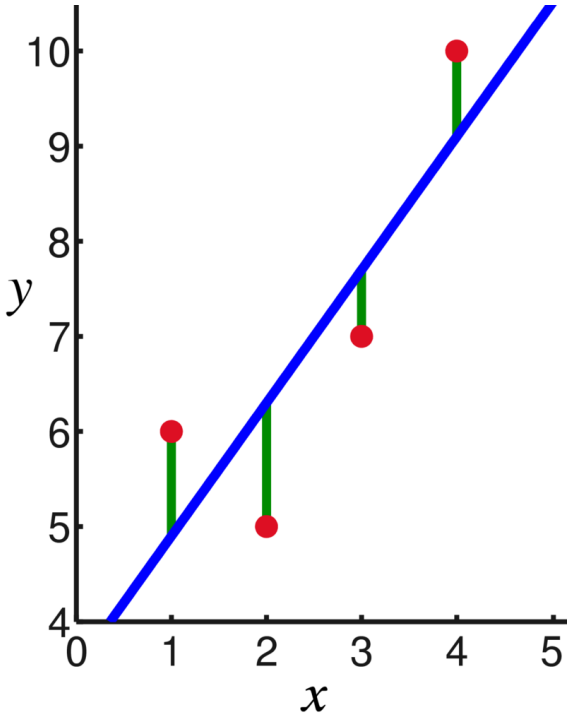
- linear regression is a linear approach to modelling the relationship between a dependent variable and an explanatory variable
- can be used to predict values of one variable based on the other
- the observations (red) are assumed to be the result of random deviations (green) from an underlying relationship (blue)



Linear Regression

What is linear regression?

- linear regression is a linear approach to modelling the relationship between a dependent variable and an explanatory variable
- can be used to predict values of one variable based on the other
- the observations (red) are assumed to be the result of random deviations (green) from an underlying relationship (blue)



Ordinary Least Squares

- minimize the sum of the squares of the differences (green) between the observed dependent variable (red) in the given dataset and those predicted by the linear function (blue).

$$y_i = \alpha + \beta x_i + \epsilon$$

Linear Regression

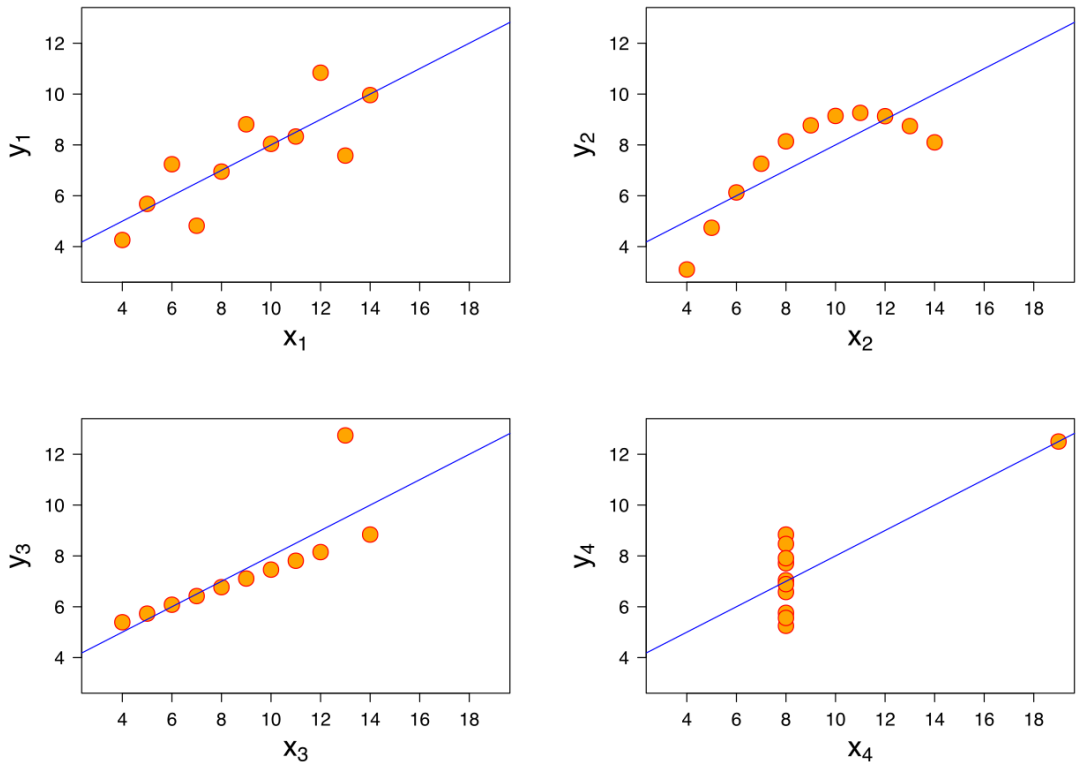
Warning

- just because you can fit a data, doesn't mean you should

Linear Regression

Warning

- just because you can fit a data, doesn't mean you should
- the data below have approximately the same linear regression line (as well as nearly identical means, standard deviations, and correlations)



Linear Regression

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Untitled1* | meddata

Filter

	X	Accept	Acceptance	Sex	BCPM	GPA	VR	PS	W
1	1	D	0	F	3.59	3.62	11	9	
2	2	A	1	M	3.75	3.84	12	13	
3	3	A	1	F	3.24	3.23	9	10	
4	4	A	1	F	3.74	3.69	12	11	
5	5	A	1	F	3.53	3.38	9	11	
6	6	A	1	M	3.59	3.72	10	9	

Showing 1 to 6 of 55 entries, 12 total columns

Environment | History | Connections | Tutorial

Import Dataset | List | Global Environment

Data	
lm1	List of 12
meddata	55 obs. of 12 variables
meddata_nooutliers	51 obs. of 12 variables
plot1	List of 9
plot2	List of 9
Values	
outliers_GPA	2.72
outliers_MCAT	num [1:3] 48 48 18

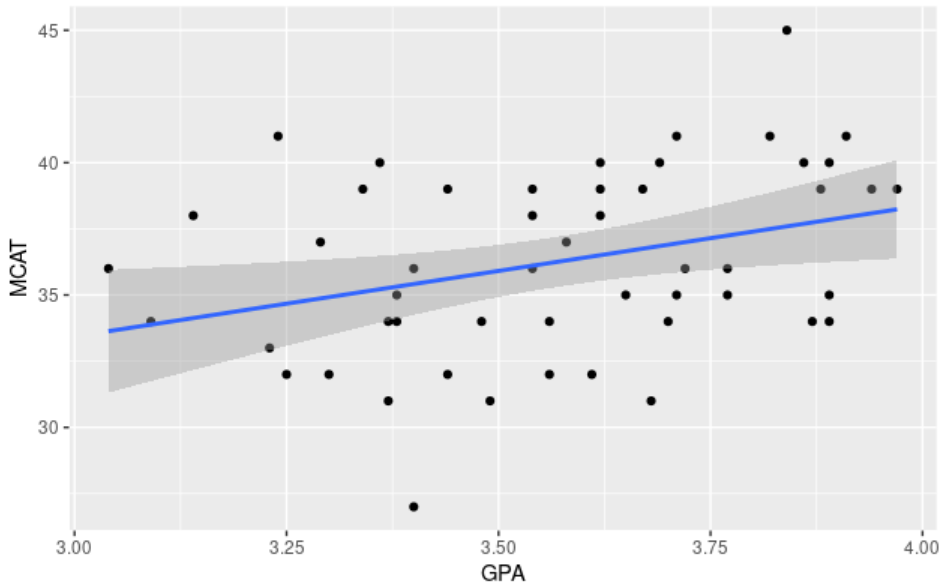
Console | Terminal | Jobs

~/Desktop/

```
> ggplot(data = meddata_nooutliers, aes(x=GPA, y=MCAT)) + geom_point() + geom_s
mooth(method = "lm")
`geom_smooth()` using formula 'y ~ x'
> lm1 <- lm(MCAT ~ GPA, data = meddata_nooutliers)
> summary(lm1)

Call:
lm(formula = MCAT ~ GPA, data = meddata_nooutliers)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4156 -2.4599  0.1184  2.4308  7.4071

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   18.591      7.112   2.614   0.0118 *
GPA            4.948      1.988   2.489   0.0162 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.367 on 49 degrees of freedom
Multiple R-squared:  0.1123,    Adjusted R-squared:  0.09415
F-statistic: 6.197 on 1 and 49 DF,  p-value: 0.01624

> |
```

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



Correlation and Linear Regression

Summary

- Pearson's correlation gives a value r, which tells us the level of linear dependence between two variables (-1 to +1)

Correlation and Linear Regression

Summary

- Pearson's correlation gives a value r, which tells us the level of linear dependence between two variables (-1 to +1)
- Based on two assumptions: data are normally distributed without any outliers; the relationship between the variables is linear

Correlation and Linear Regression

Summary

- Pearson's correlation gives a value r, which tells us the level of linear dependence between two variables (-1 to +1)
- Based on two assumptions: data are normally distributed without any outliers; the relationship between the variables is linear
- Linear Regression models the linear correlation between an independent variable and a dependent one

Correlation and Linear Regression

Summary

- Pearson's correlation gives a value r, which tells us the level of linear dependence between two variables (-1 to +1)
- Based on two assumptions: data are normally distributed without any outliers; the relationship between the variables is linear
- Linear Regression models the linear correlation between an independent variable and a dependent one
- Ordinary Least Squares is one method of determining the linear regression, where it minimizes the sum of the squares of the differences between the observed dependent variable and those predicted by the linear function.

Correlation and Linear Regression

Summary

- Ultimately we end up with a linear model:

$$y_i = \alpha + \beta x_i + \epsilon$$

Correlation and Linear Regression

Summary

- Ultimately we end up with a linear model:

$$y_i = \alpha + \beta x_i + \epsilon$$

- And an r-squared, which tells us how much of the variation of the response variable is explained by our predictors, and not by error

Summary

Today we learned...

- what R and RStudio are, how to download and install them, and how to load data and packages

Summary

Today we learned...

- what R and RStudio are, how to download and install them, and how to load data and packages
- what a histogram is and how to use it to get a visual idea of how a single set of our data values are distributed

Summary

Today we learned...

- what R and RStudio are, how to download and install them, and how to load data and packages
- what a histogram is and how to use it to get a visual idea of how a single set of our data values are distributed
- what a scatterplot is and how to use it to visualize two sets of our data and visually see how they are potentially linked

Summary

Today we learned...

- what R and RStudio are, how to download and install them, and how to load data and packages
- what a histogram is and how to use it to get a visual idea of how a single set of our data values are distributed
- what a scatterplot is and how to use it to visualize two sets of our data and visually see how they are potentially linked
- what a boxplot is and how to use it to visualize our data

Summary

Today we learned...

- what R and RStudio are, how to download and install them, and how to load data and packages
- what a histogram is and how to use it to get a visual idea of how a single set of our data values are distributed
- what a scatterplot is and how to use it to visualize two sets of our data and visually see how they are potentially linked
- what a boxplot is and how to use it to visualize our data
- what an outlier is and how to identify and remove it in R
- about Pearson correlation and r

Summary

Today we learned...

- what R and RStudio are, how to download and install them, and how to load data and packages
- what a histogram is and how to use it to get a visual idea of how a single set of our data values are distributed
- what a scatterplot is and how to use it to visualize two sets of our data and visually see how they are potentially linked
- what a boxplot is and how to use it to visualize our data
- what an outlier is and how to identify and remove it in R
- about Pearson correlation and r
- about linear regression, ordinary least squares, a linear regression model, and r^2
- how to plot all of these using base R packages, and with ggplot2

Going Further

Additional Resources

- Getting Started with R and RStudio
  https://www.youtube.com/watch?v=lVKMsaWju8w
- Programming with R - Tutorial
  http://swcarpentry.github.io/r-novice-inflammation/
- R for Reproducible Scientific Analysis - Tutorial
  http://swcarpentry.github.io/r-novice-gapminder/
- Scatterplots, Boxplots, Histograms in R
  http://becomingvisual.com/rfundamentals/statistical-graphs.html
- Linear Regression in R
  https://www.datacamp.com/community/tutorials/linear-regression-R