COMP320: Research Practice

11: Visualising Data in R

Register Attendance

Figure 1: Attendance monitoring is in place. It is your responsability to ensure that you have signed yourself in.

Learning Outcomes

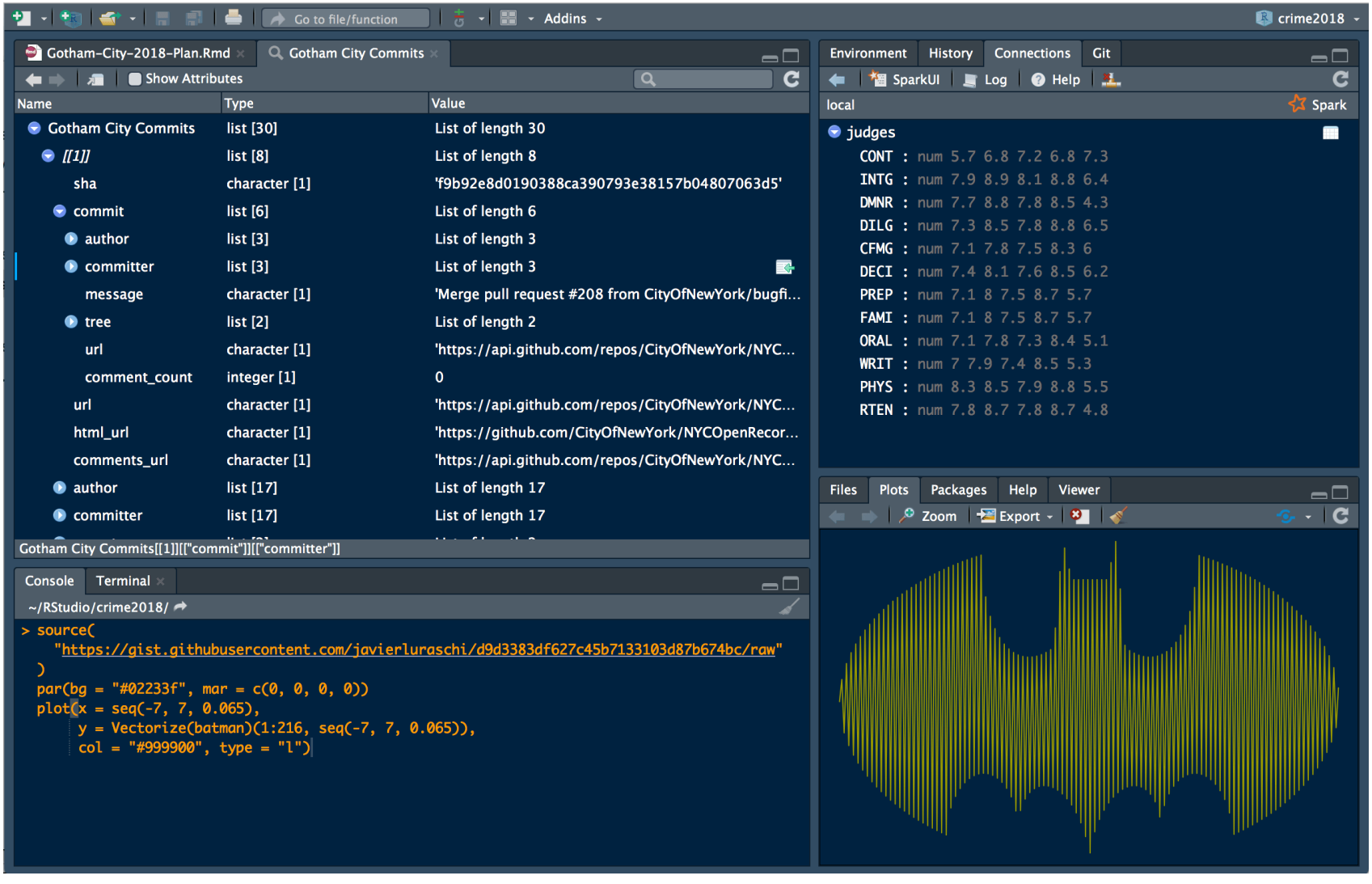
After this session you will be able to:

- **Import** data for analysis in R
- **Analyse** data in R
- **Visualise** data in R

What is R?

Figure 2: 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

RStudio


Go to file/function
Addins
crime2018
Gotham-City-2018-Plan.Rmd
Gotham City Commits
Show Attributes
Name
Type
Value
Gotham City Commits
list [30]
List of length 30
[[1]]
list [8]
List of length 8
sha
character [1]
'f9b92e8d0190388ca390793e38157b04807063d5'
commit
list [6]
List of length 6
author
list [3]
List of length 3
committer
list [3]
List of length 3
message
character [1]
'Merge pull request #208 from CityOfNewYork/bugfi...
tree
list [2]
List of length 2
url
character [1]
'https://api.github.com/repos/CityOfNewYork/NYC...
comment_count
integer [1]
0
url
character [1]
'https://api.github.com/repos/CityOfNewYork/NYC...
html_url
character [1]
'https://github.com/CityOfNewYork/NYCOpenRecor...
comments_url
character [1]
'https://api.github.com/repos/CityOfNewYork/NYC...
author
list [17]
List of length 17
committer
list [17]
List of length 17
Gotham City Commits[[1]][["commit"]][["committer"]]
Console
Terminal
~/RStudio/crime2018/
> source(
"https://gist.githubusercontent.com/javierluraschi/d9d3383df627c45b7133103d87b674bc/raw"
)
par(bg = "#02233f", mar = c(0, 0, 0, 0))
plot(x = seq(-7, 7, 0.065),
y = Vectorize(batman)(1:216, seq(-7, 7, 0.065)),
col = "#999900", type = "l")
Environment
History
Connections
Git
SparkUI
Log
Help
local
Spark
judges
CONT : num 5.7 6.8 7.2 6.8 7.3
INTG : num 7.9 8.9 8.1 8.8 6.4
DMNR : num 7.7 8.8 7.8 8.5 4.3
DILG : num 7.3 8.5 7.8 8.8 6.5
CFMG : num 7.1 7.8 7.5 8.3 6
DECI : num 7.4 8.1 7.6 8.5 6.2
PREP : num 7.1 8 7.5 8.7 5.7
FAMI : num 7.1 8 7.5 8.7 5.7
ORAL : num 7.1 7.8 7.3 8.4 5.1
WRIT : num 7 7.9 7.4 8.5 5.3
PHYS : num 8.3 8.5 7.9 8.8 5.5
RTEN : num 7.8 8.7 7.8 8.7 4.8
Files
Plots
Packages
Help
Viewer
Zoom
Export

Tidyverse

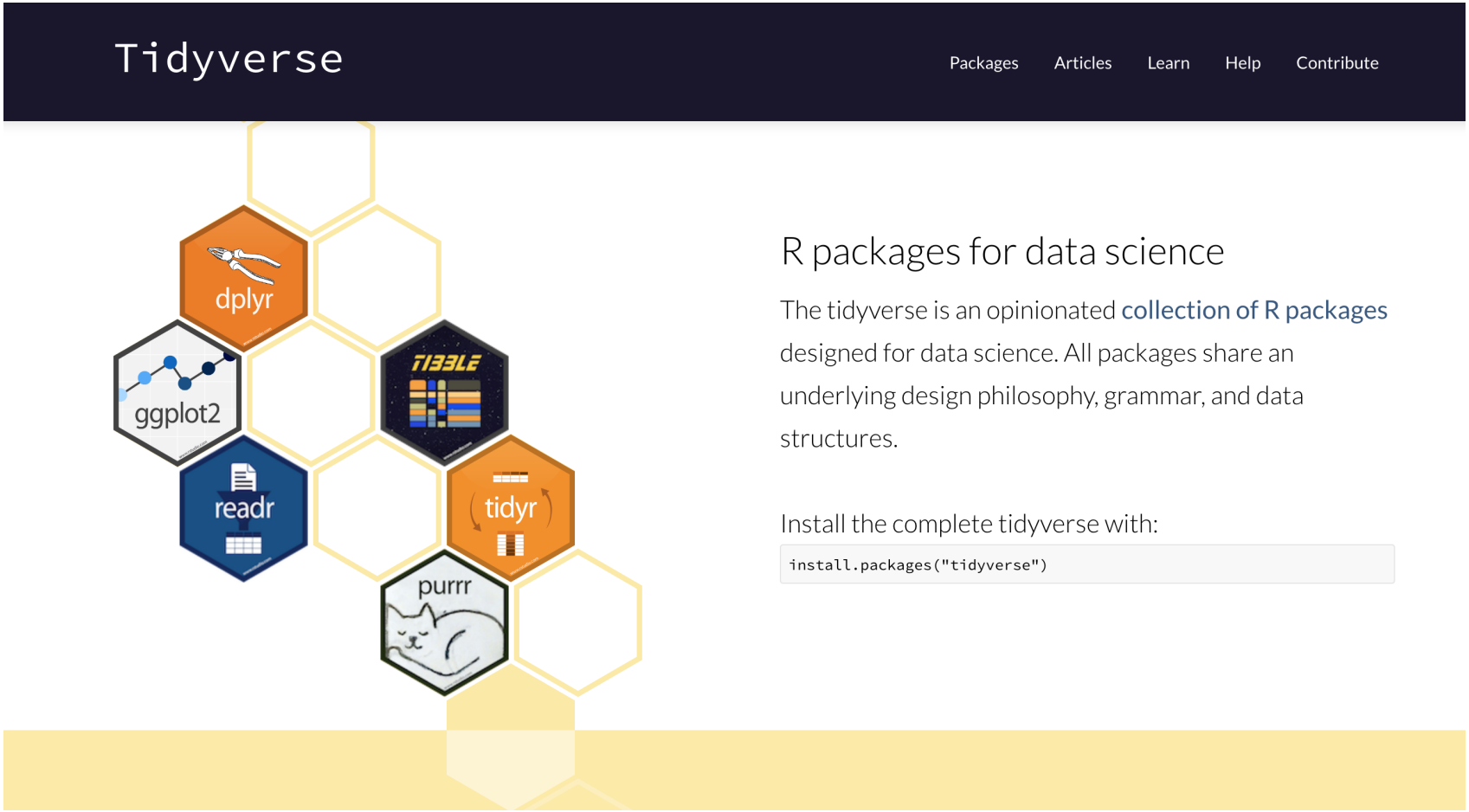


Figure 3: An opinionated collection of R packages designed for data science

Psych

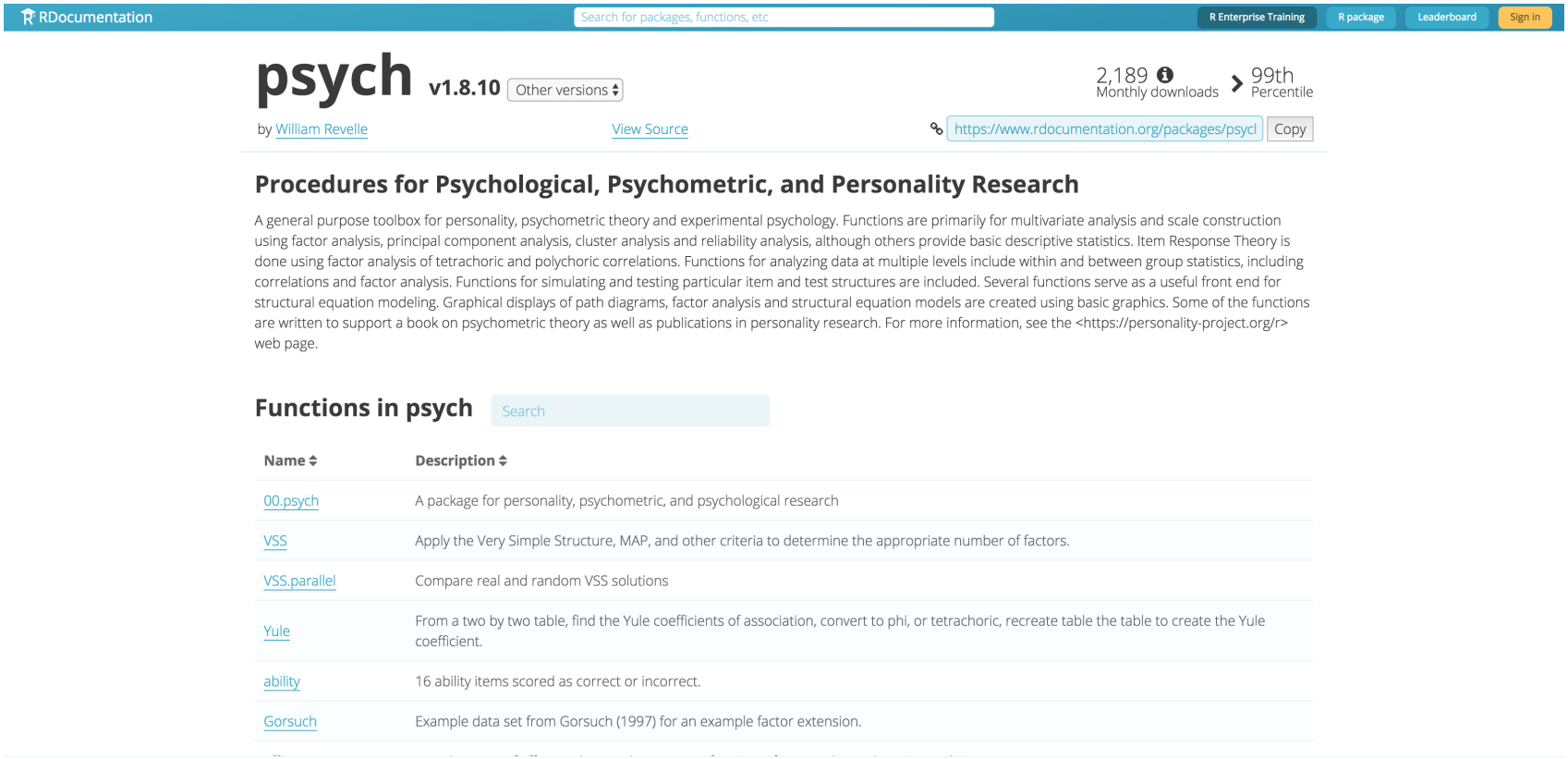


Figure 4: A general purpose toolbox for personality, psychometric theory and experimental psychology. Functions are primarily for multivariate analysis and scale construction using factor analysis, principal component analysis, cluster analysis and reliability analysis, although others provide basic descriptive statistics.

Installing Packages

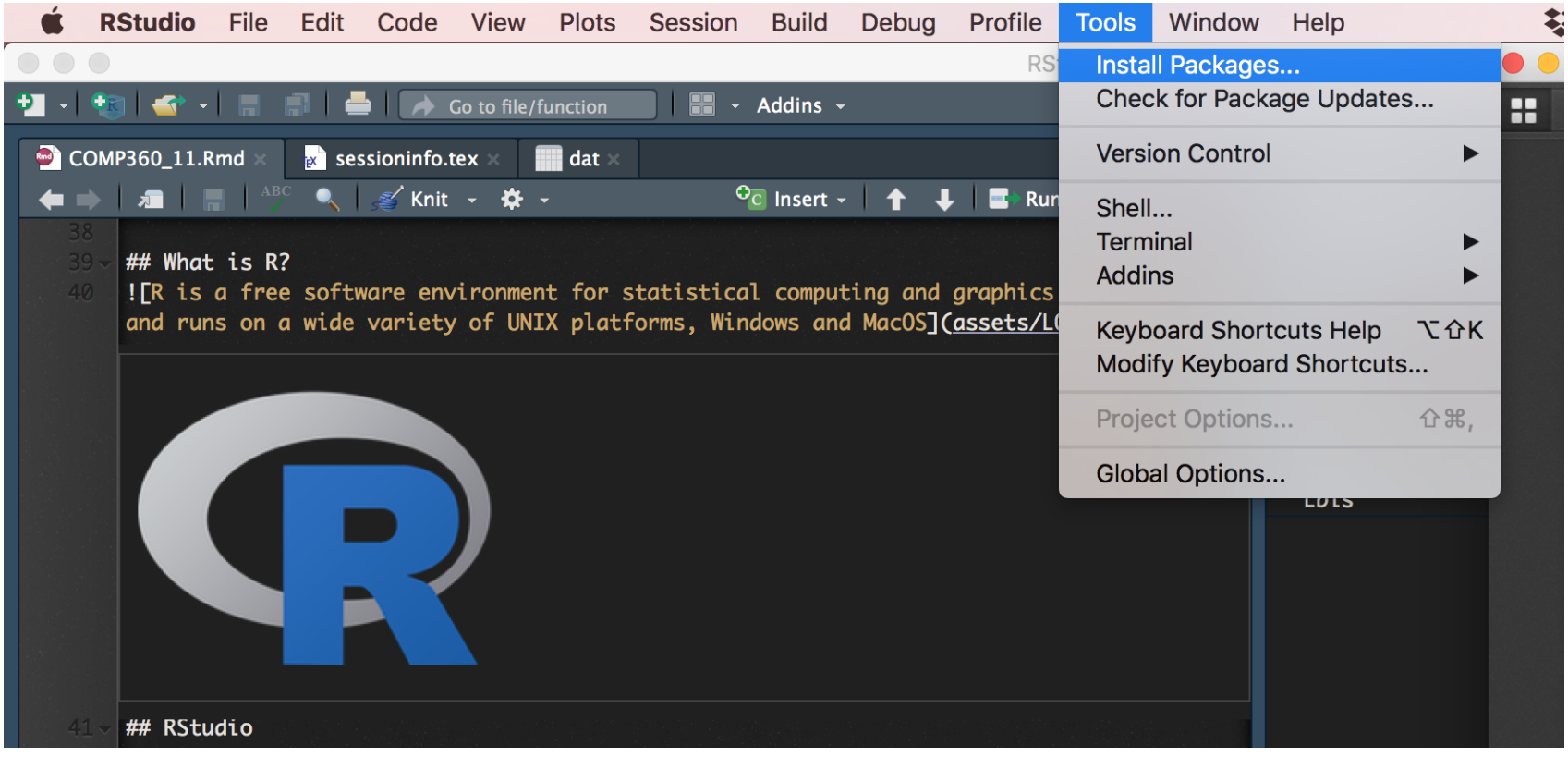


Figure 5: Location of the menu for importing packages

Importing Packages

```
> library(psych)
> library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------ tidy
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.7
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts --------------------------------------------------------------- tidyverse_
## x ggplot2::%+%()   masks psych::%+%()
## x ggplot2::alpha() masks psych::alpha()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

Data

R and Tidyverse come packaged with some interesting datasets. Try:

```
> data()
```

Also try viewing a `dataframe’:

```
> mpg
```

Download Data

Download and inspect this comma-seperated values (CSV) file:

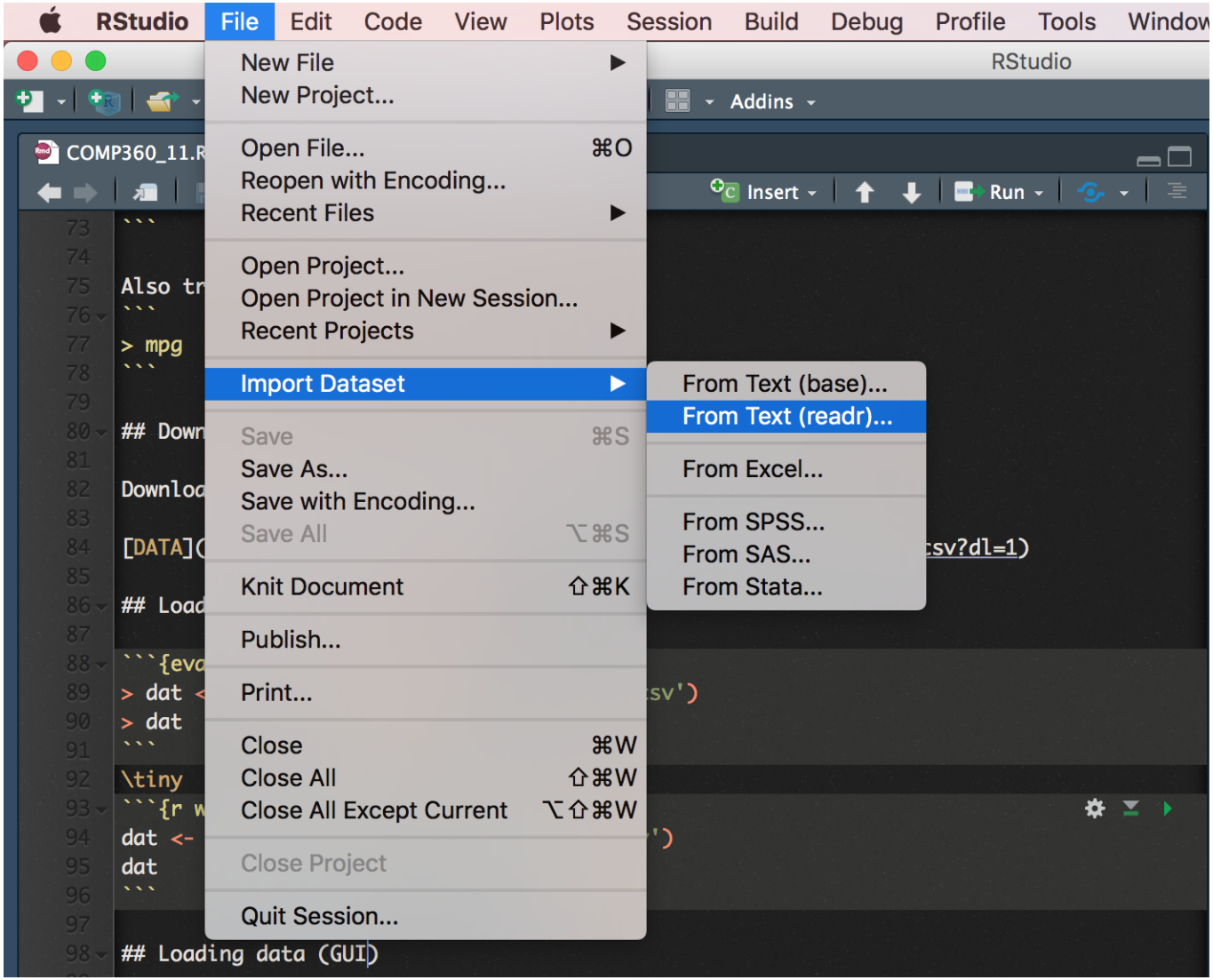
DATA

Loading data

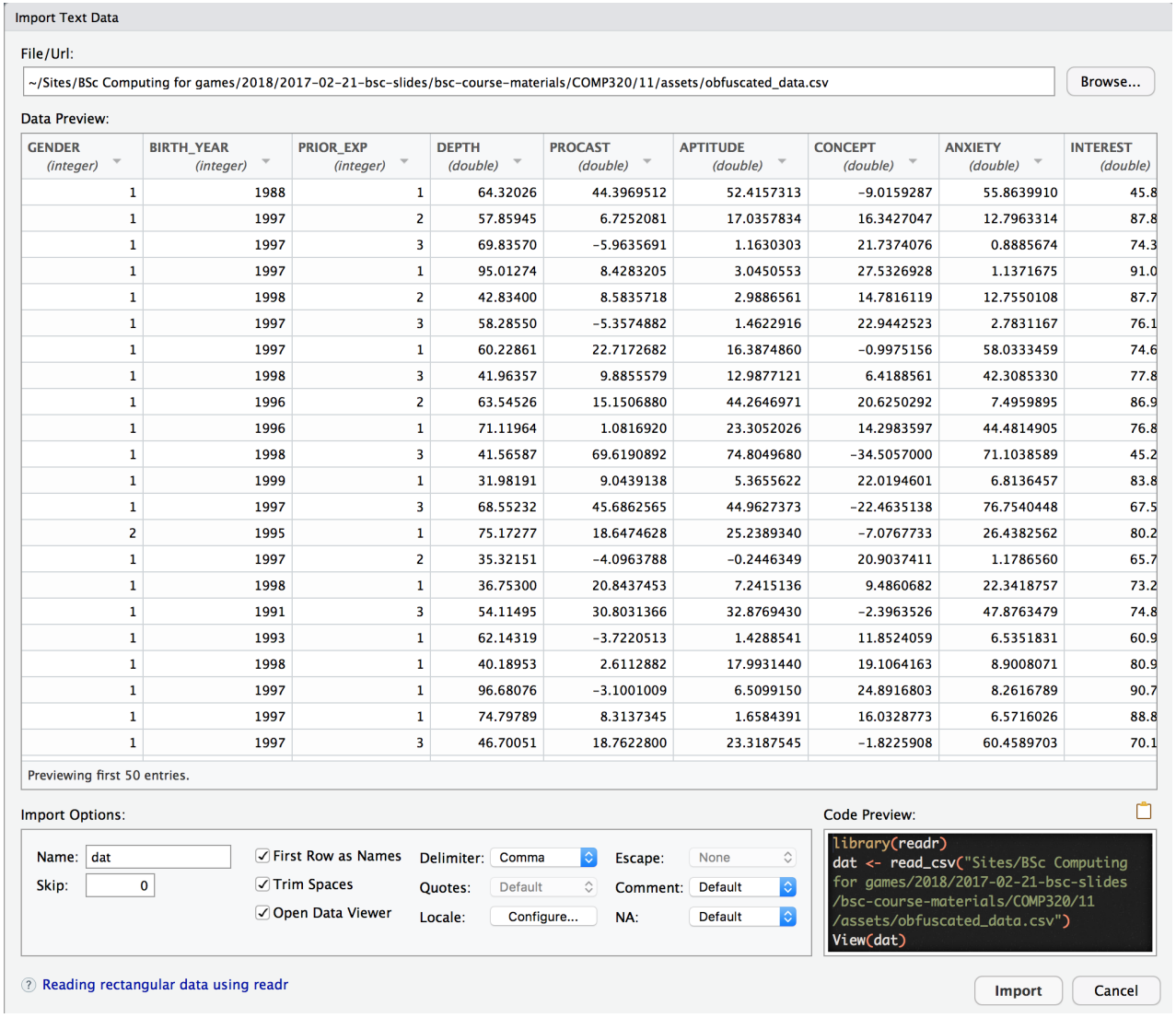
```
> dat <- read_csv('assets/obfuscated_data.csv')
> dat
```

```
## # A tibble: 159 x 14
##    GENDER BIRTH_YEAR PRIOR_EXP DEPTH PROCAST APTITUDE CONCEPT ANXIETY
##     <int>      <int>     <int> <dbl>   <dbl>    <dbl>   <dbl>   <dbl>
##  1      1       1988         1  64.3   44.4     52.4   -9.02   55.9
##  2      1       1997         2  57.9    6.73    17.0   16.3    12.8
##  3      1       1997         3  69.8   -5.96     1.16  21.7     0.889
##  4      1       1997         1  95.0    8.43     3.05  27.5     1.14
##  5      1       1998         2  42.8    8.58     2.99  14.8    12.8
##  6      1       1997         3  58.3   -5.36     1.46  22.9     2.78
##  7      1       1997         1  60.2   22.7     16.4   -0.998  58.0
##  8      1       1998         3  42.0    9.89    13.0    6.42   42.3
##  9      1       1996         2  63.5   15.2     44.3   20.6     7.50
## 10      1       1996         1  71.1    1.08    23.3   14.3    44.5
## # ... with 149 more rows, and 6 more variables: INTEREST <dbl>,
## #   EXTRAVERSION <int>, AGREEABLENESS <int>, CONSCIENTIOUSNESS <int>,
## #   NEUROTICISM <int>, OPENNESS <int>
```

Loading data (GUI) STEP 1



Loading data (GUI) STEP 2



Import Text Data

File/Url:

~/Sites/BSc Computing for games/2018/2017-02-21-bsc-slides/bsc-course-materials/COMP320/11/assets/obfuscated_data.csv

Browse...

Data Preview:

GENDER (integer)	BIRTH_YEAR (integer)	PRIOR_EXP (integer)	DEPTH (double)	PROCAST (double)	APTITUDE (double)	CONCEPT (double)	ANXIETY (double)	INTEREST (double)
1	1988	1	64.32026	44.3969512	52.4157313	-9.0159287	55.8639910	45.8
1	1997	2	57.85945	6.7252081	17.0357834	16.3427047	12.7963314	87.8
1	1997	3	69.83570	-5.9635691	1.1630303	21.7374076	0.8885674	74.3
1	1997	1	95.01274	8.4283205	3.0450553	27.5326928	1.1371675	91.0
1	1998	2	42.83400	8.5835718	2.9886561	14.7816119	12.7550108	87.7
1	1997	3	58.28550	-5.3574882	1.4622916	22.9442523	2.7831167	76.1
1	1997	1	60.22861	22.7172682	16.3874860	-0.9975156	58.0333459	74.6
1	1998	3	41.96357	9.8855579	12.9877121	6.4188561	42.3085330	77.8
1	1996	2	63.54526	15.1506880	44.2646971	20.6250292	7.4959895	86.9
1	1996	1	71.11964	1.0816920	23.3052026	14.2983597	44.4814905	76.8
1	1998	3	41.56587	69.6190892	74.8049680	-34.5057000	71.1038589	45.2
1	1999	1	31.98191	9.0439138	5.3655622	22.0194601	6.8136457	83.8
1	1997	3	68.55232	45.6862565	44.9627373	-22.4635138	76.7540448	67.5
2	1995	1	75.17277	18.6474628	25.2389340	-7.0767733	26.4382562	80.2
1	1997	2	35.32151	-4.0963788	-0.2446349	20.9037411	1.1786560	65.7
1	1998	1	36.75300	20.8437453	7.2415136	9.4860682	22.3418757	73.2
1	1991	3	54.11495	30.8031366	32.8769430	-2.3963526	47.8763479	74.8
1	1993	1	62.14319	-3.7220513	1.4288541	11.8524059	6.5351831	60.9
1	1998	1	40.18953	2.6112882	17.9931440	19.1064163	8.9008071	80.9
1	1997	1	96.68076	-3.1001009	6.5099150	24.8916803	8.2616789	90.7
1	1997	1	74.79789	8.3137345	1.6584391	16.0328773	6.5716026	88.8
1	1997	3	46.70051	18.7622800	23.3187545	-1.8225908	60.4589703	70.1

Previewing first 50 entries.

Import Options:

Name: dat

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)
dat <- read_csv("Sites/BSc Computing
for games/2018/2017-02-21-bsc-slides
/bsc-course-materials/COMP320/11
/assets/obfuscated_data.csv")
View(dat)
```

Reading rectangular data using readr

Import

Cancel

Summary

```
> summary(dat)
```

```
##      GENDER        BIRTH_YEAR     PRIOR_EXP         DEPTH
##  Min.   :1.000   Min.   :1976   Min.   :1.000   Min.   :-170.12
##  1st Qu.:1.000   1st Qu.:1996   1st Qu.:1.000   1st Qu.:  46.56
##  Median :1.000   Median :1997   Median :1.000   Median :  59.51
##  Mean   :1.119   Mean   :1996   Mean   :1.704   Mean   :  57.39
##  3rd Qu.:1.000   3rd Qu.:1998   3rd Qu.:3.000   3rd Qu.:  71.48
##  Max.   :2.000   Max.   :1999   Max.   :3.000   Max.   :  96.68
##     PROCAST          APTITUDE          CONCEPT            ANXIETY
##  Min.   :-5.964   Min.   :-170.23   Min.   :-35.6758   Min.   : 0.8036
##  1st Qu.:12.083   1st Qu.:  17.58   1st Qu.:-10.1707   1st Qu.:19.3574
##  Median :26.658   Median :  25.24   Median :  1.7783   Median :37.0401
##  Mean   :26.062   Mean   :  29.24   Mean   :  0.4195   Mean   :37.6957
##  3rd Qu.:39.166   3rd Qu.:  44.76   3rd Qu.: 11.8288   3rd Qu.:53.8863
##  Max.   :69.619   Max.   :  95.28   Max.   : 28.7092   Max.   :92.0369
##     INTEREST      EXTRAVERSION    AGREEABLENESS   CONSCIENTIOUSNESS
##  Min.   :12.53   Min.   :  0.00   Min.   : 20.0   Min.   : 20.0
##  1st Qu.:60.50   1st Qu.: 62.50   1st Qu.:100.0   1st Qu.: 91.0
##  Median :71.46   Median :100.00   Median :120.0   Median :120.0
##  Mean   :68.71   Mean   : 94.44   Mean   :121.4   Mean   :115.6
##  3rd Qu.:83.57   3rd Qu.:120.00   3rd Qu.:140.5   3rd Qu.:140.0
##  Max.   :91.06   Max.   :188.00   Max.   :200.0   Max.   :200.0
##   NEUROTICISM       OPENNESS
##  Min.   :  0.0   Min.   : 24.0
##  1st Qu.: 60.0   1st Qu.:100.0
##  Median : 98.0   Median :120.0
##  Mean   : 94.4   Mean   :117.8
##  3rd Qu.:126.5   3rd Qu.:140.0
##  Max.   :200.0   Max.   :200.0
```

Describe

```
> describe(dat$PROCAST)
```

```
##    vars   n  mean    sd median trimmed   mad   min   max range skew
## X1    1 159 26.06 17.36  26.66   25.75 20.67 -5.96 69.62 75.58 0.18
##    kurtosis   se
## X1    -0.72 1.38
```

Correlation

```
cor(x, y)
cor.test(x, y, method)
```

```
> cor.test(dat$PROCAST, dat$APTITUDE, method="pears
```

```
##
##  Pearson's product-moment correlation
##
## data:  dat$PROCAST and dat$APTITUDE
## t = 9.7917, df = 157, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5088690 0.7039297
## sample estimates:
##       cor
## 0.6157466
```

P-Value

What does this value mean?

$$< 2.2e - 16$$

P-Value

What does this value mean?

$$p < 2.2 * 10 - 16$$

numerically undistinguishable from 0

Correlation Results

Null Hypothesis: There is no **relationship** between PROCAST and APTITUDE

Result: **Refute** the null hypothesis and **accept** the alternative hypothesis

- Correlation Coefficient: 0.6156
- P-Value: 2.2 X 10 ^ -16

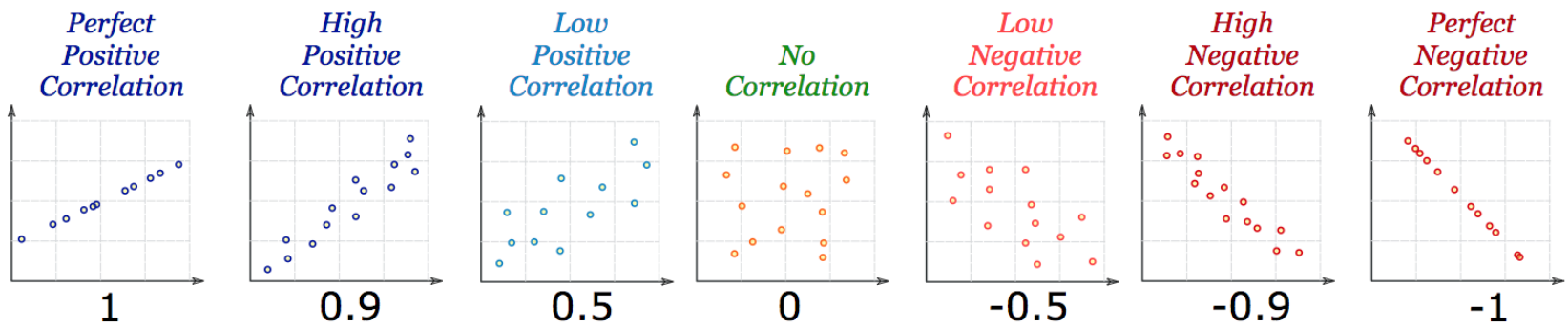


Figure 6:

T-Test

```
> t.test(dat$ANXIETY~dat$GENDER)


##
##  Welch Two Sample t-test
##
## data:  dat$ANXIETY by dat$GENDER
## t = -0.97505, df = 26.122, p-value = 0.3385
## alternative hypothesis: true difference in means
## 95 percent confidence interval:
##  -13.655452   4.867173
## sample estimates:
## mean in group 1 mean in group 2
##        37.17062        41.56476
```

T-Test Results

Null Hypothesis: There is no **relationship** between GENDER and ANXIETY

Result: **Accept** the null hypothesis

- P-Value: 0.3385

Information Presentation

- There are various techniques for reformatting and reducing data to make the analysis more interpretable or to illustrate a key point
- Graphical representations will also assist in decision making and reinforce the justification for those decisions e.g., has a hypothesis been falsified? To what extent is it clearly falsified?
- An overall picture of the data can be gleaned and initial conclusions drawn
- It is important to select the most effective ways to illustrate your findings in the dissertation

Information Presentation

- Your communication skills are under assessment; keep all graphical depiction meaningful to justifying your analysis and/or your intellectual decisions
- Provides an overall picture of the data underlying your findings to reach and support your conclusions
- Be wary of delegating charts solely to important data:
  - Depictions can distort message of original data
  - Concise, but often lacks precision
  - Ensure adequate support in body of text
  - Leverage explicit references (e.g., "as shown in Figure 1")

- Bar Chart
- Histogram
- Frequency Polygon
- Cummaltive Frequency Polygon (Ogive)
- Pie Chart
- Scatter Plot
- Box Plot

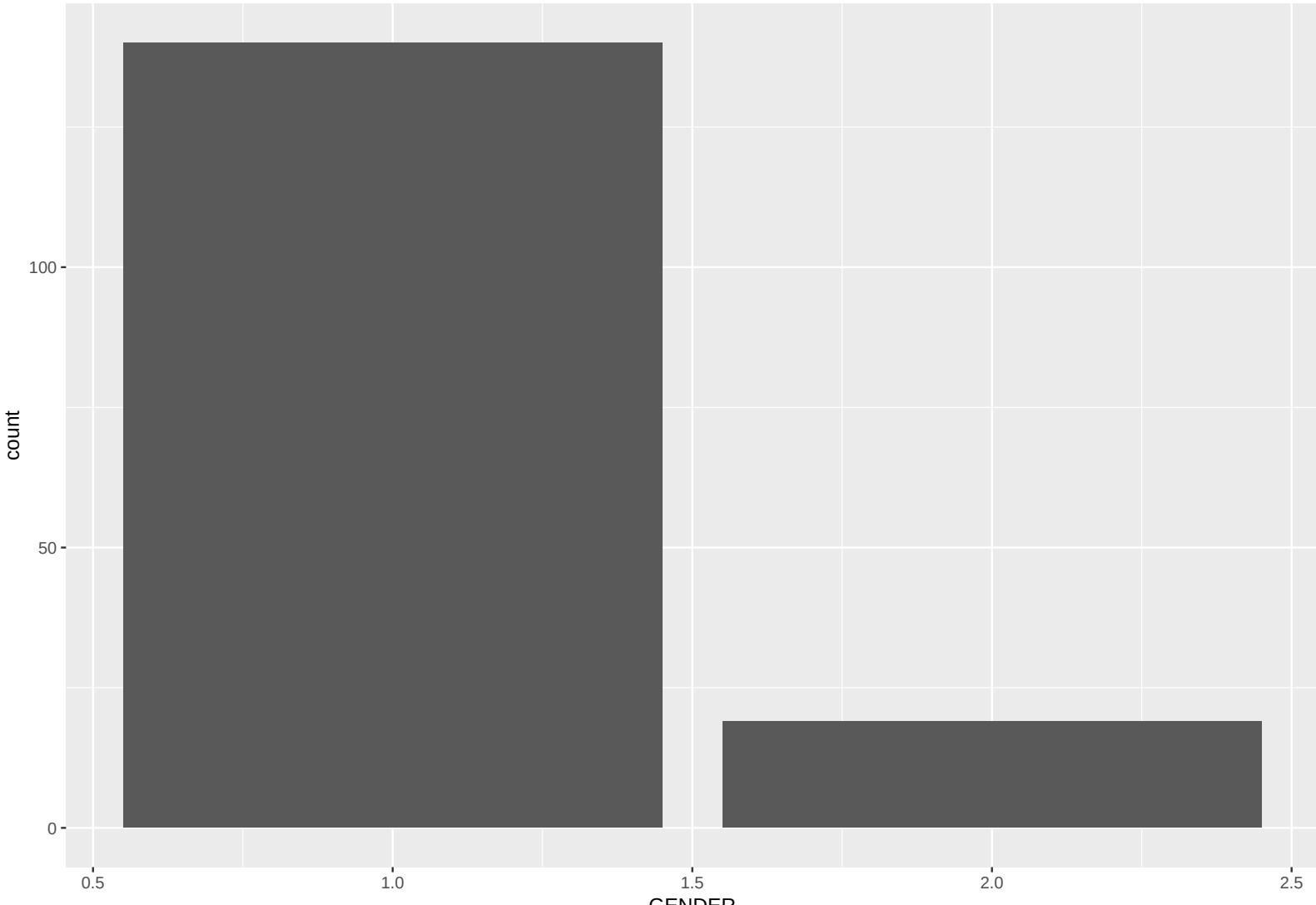
ggplot

```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

Figure 7: The anatompy of a ggplot command

Bar Chart in R

```
> ggplot(dat, aes(GENDER)) + geom_bar()
```



Complex Bar Chart

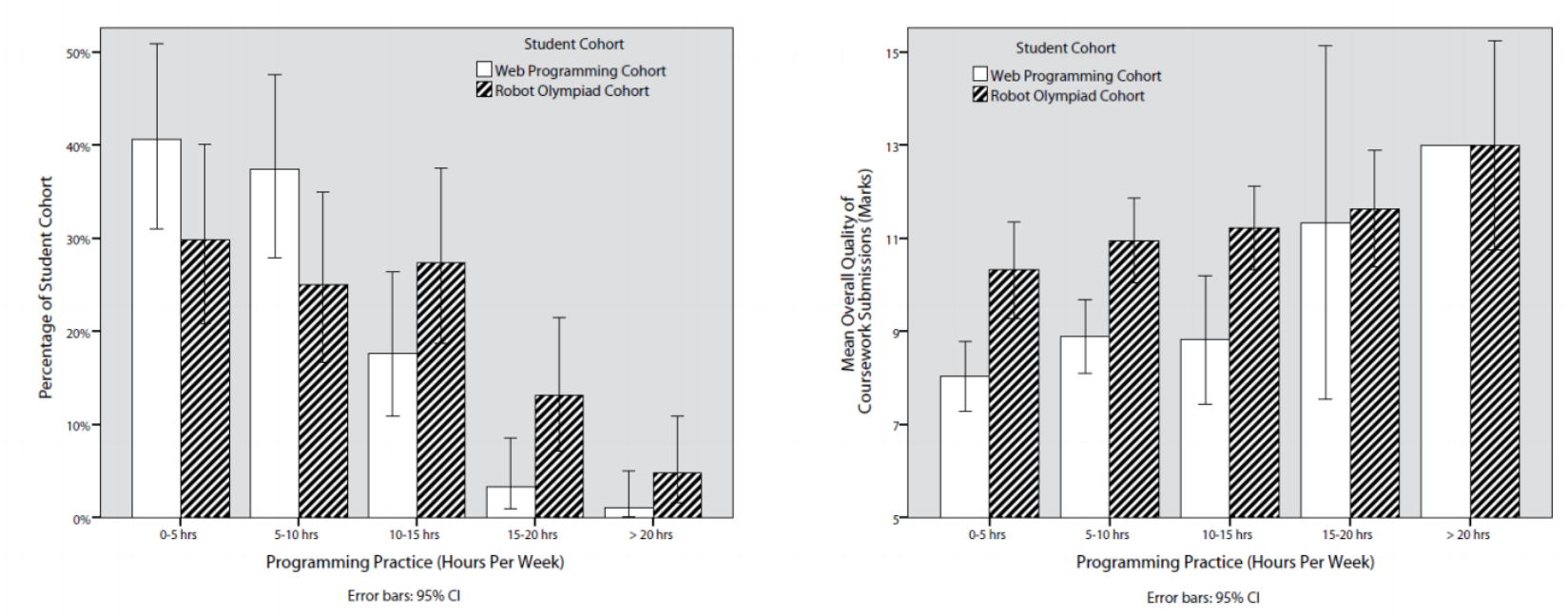


Figure 8: Docs: Bar and line graphs (ggplot2)

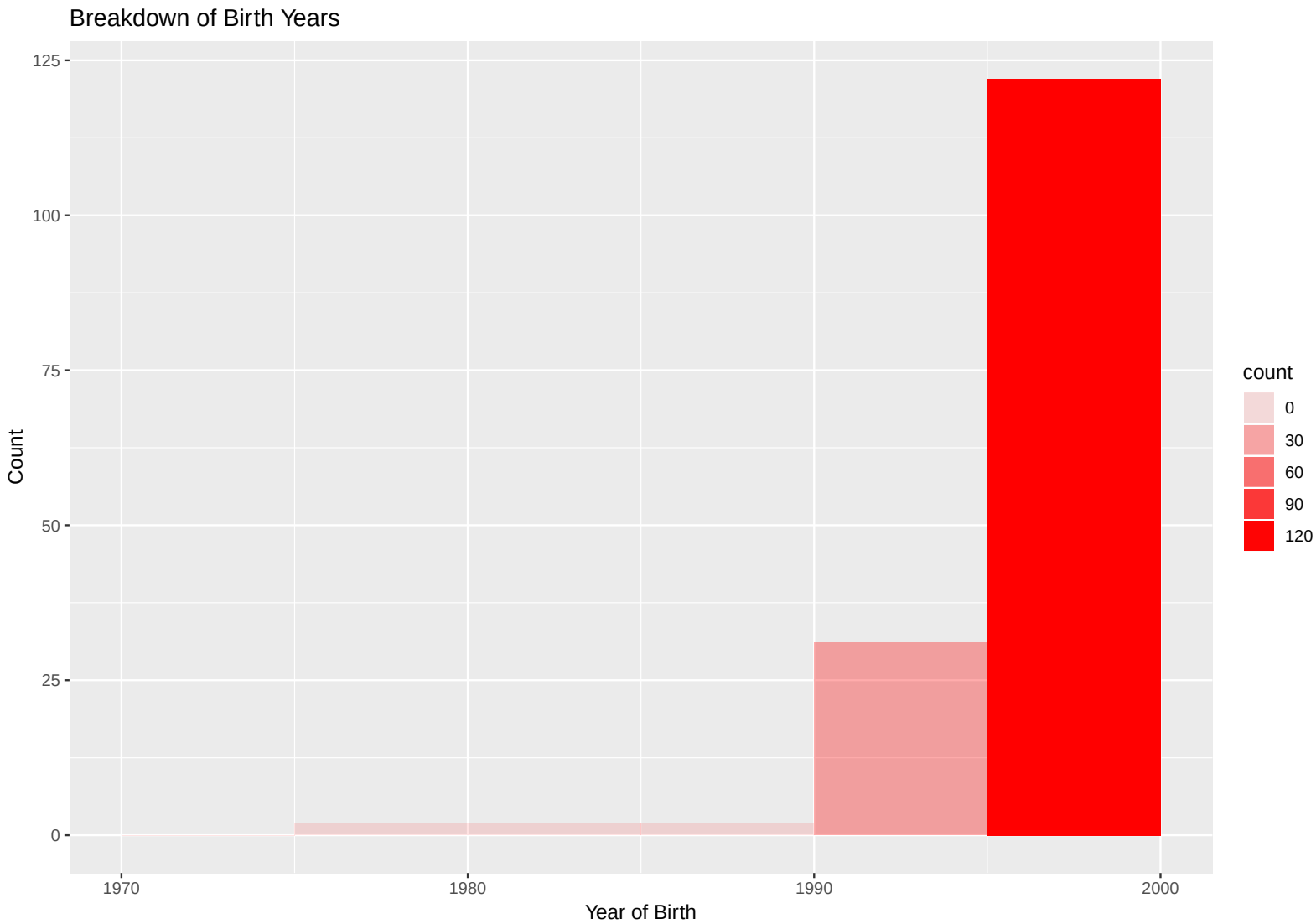
Histogram

- A type of vertical bar chart used to depict a frequency distribution
- Construction steps:
  - Label the **x** axis with the class endpoints
  - Label the **y** axis with the frequencies
  - Label the chart with an appropriate title, i.e. not `bar chart'
- A quick look at the histogram reveals which class intervals produce the highest frequency totals E.g. which age group most often enrols in undergraduate computing courses?

Histogram in R

```
> ggplot(dat, aes(BIRTH_YEAR)) + geom_histogram(
  breaks=seq(1970, 2000, by =5),
  fill="red",
  aes(alpha = ..count..)) +
  labs(x = "Year of Birth",
    y = "Count",
    title = "Breakdown of Birth Years")
```

Histogram Output



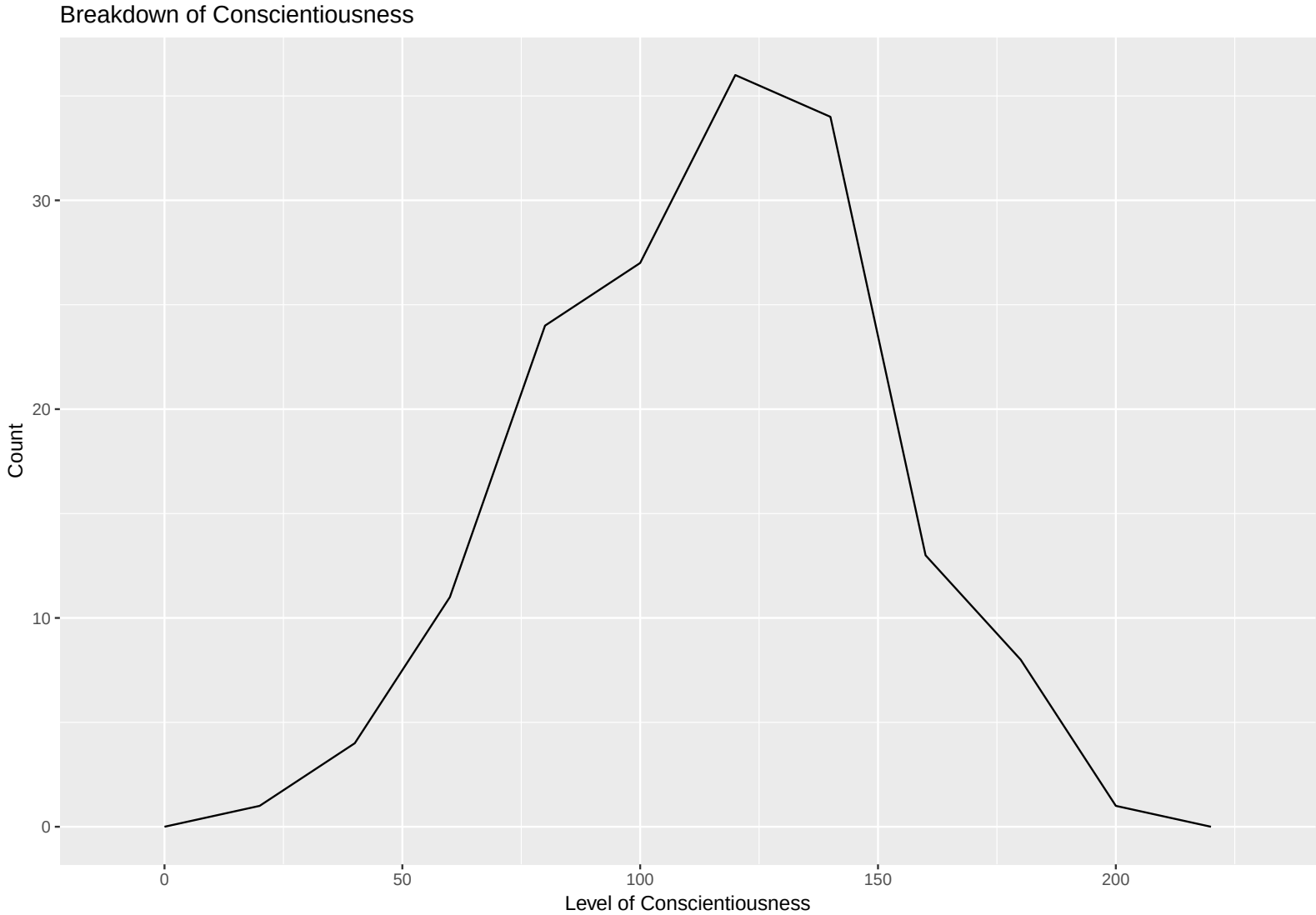
Frequency Polygon

- A graph in which line segments connecting the dots depict a frequency distribution
- Construction steps:
  - Label the **x** axis with the class endpoints
  - Label the **y** axis with the frequencies
  - Plot a dot for the frequency value at the midpoint of each class interval
  - Connect the dots with a line

Frequency Polygon in R

```
> ggplot(dat, aes(CONSCIENTIOUSNESS), stat="count")
  + geom_freqpoly(binwidth = 20)
  + labs(
    x = "Year of Birth",
    y = "Count",
    title = "Breakdown of Birth Years")
```

Frequency Polygon Output



OGive

- A Cumulative Frequency (CF) polygon
- Construction steps:
  - Label the x axis with the class endpoints
  - Label the y axis with the cumulative frequencies
  - A dot of `0' is placed at the beginning of the first class
  - Mark a dot for the CF value at the end of each class interval
  - Connect these dots with a line

OGive in R

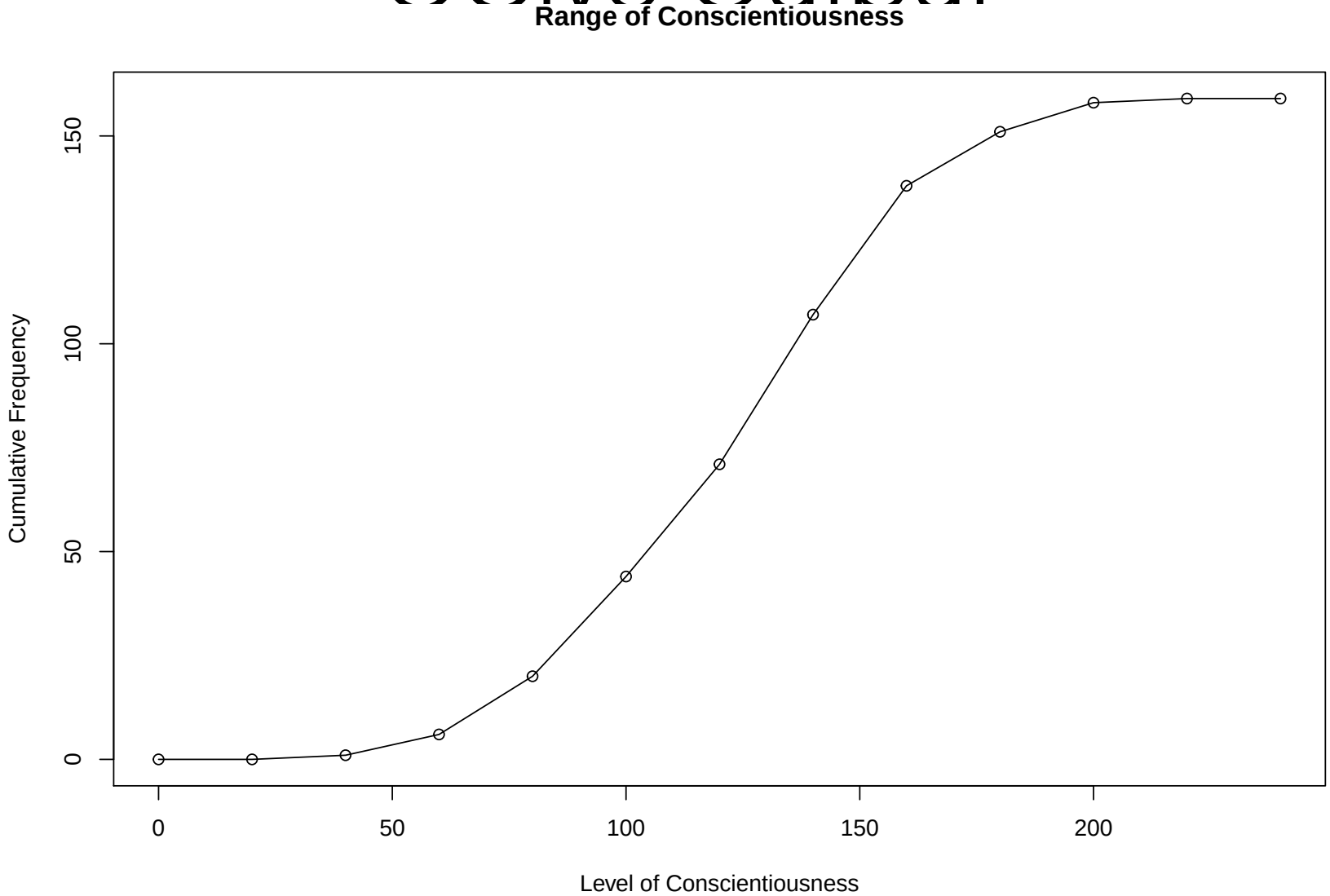
To construct an ogive, you need to format the data into cumulative frequencies:

```
cf <- c(0, cumsum(
  table(
    cut(dat$CONSCIENTIOUSNESS, seq(0, 240, by=20),
    right=FALSE))))
```

Then plot the chart based on this data:

```
plot(seq(0, 240, by=20),
  cf,
  main="Range of Conscientiousness",
  xlab="Level of Conscientiousness",
  ylab="Cumulative Frequency")
    + lines(seq(0, 240, by=20), cf)
```

OGive Output




```
## integer(0)
```

Pie Chart

- A circular depiction of data where the area of the whole pie = 100% of the data being studied.
- Slices represent a % breakdown of each of the values
- Business uses: e.g. for depicting budget categories, market share, time and resource allocation
- Generally more difficult to interpret the size of the slices compared to the bars in a histogram. But- usage of '%' can clarify slice size

Construction steps

1. Convert each toothpaste brand amount to a proportion by dividing each individual amount by the total

$$102/200 = 0.51$$

2. Convert each proportion to degrees by multiplying by 360°

$$0.51 * 360 = 183.6$$

Pie Chart in R

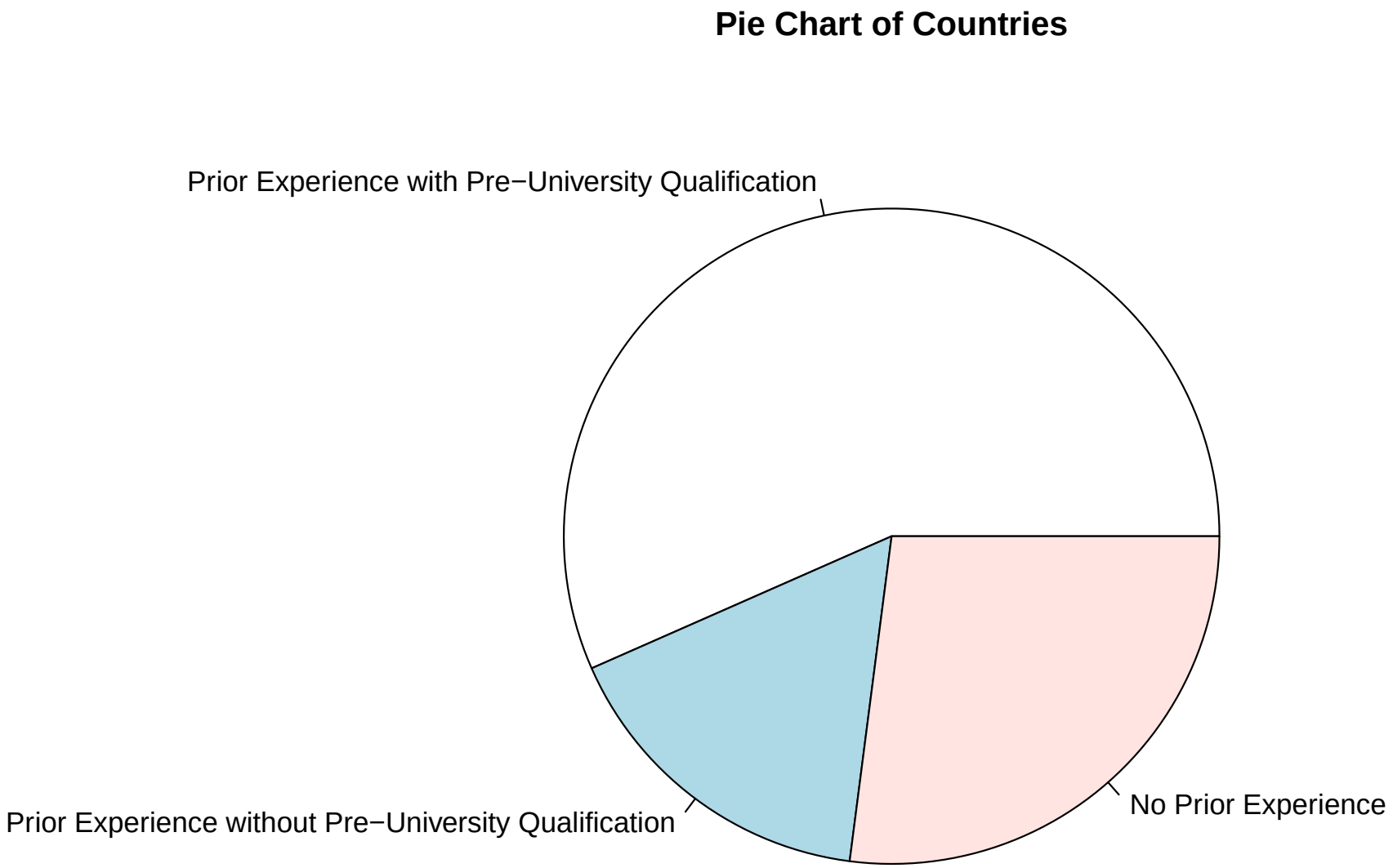
```
Lbls <- c(
  "Prior Experience with Pre-University Qualificati
  "Prior Experience without Pre-University Qualific
  "No Prior Experience")

pieValues <- as.data.frame(table(dat$PRIOR_EXP))

pieValues$labels = Lbls

pie(
  pieValues$Freq,
  labels = pieValues$labels,
  main="Pie Chart of Countries")
```

Pie Chart Output



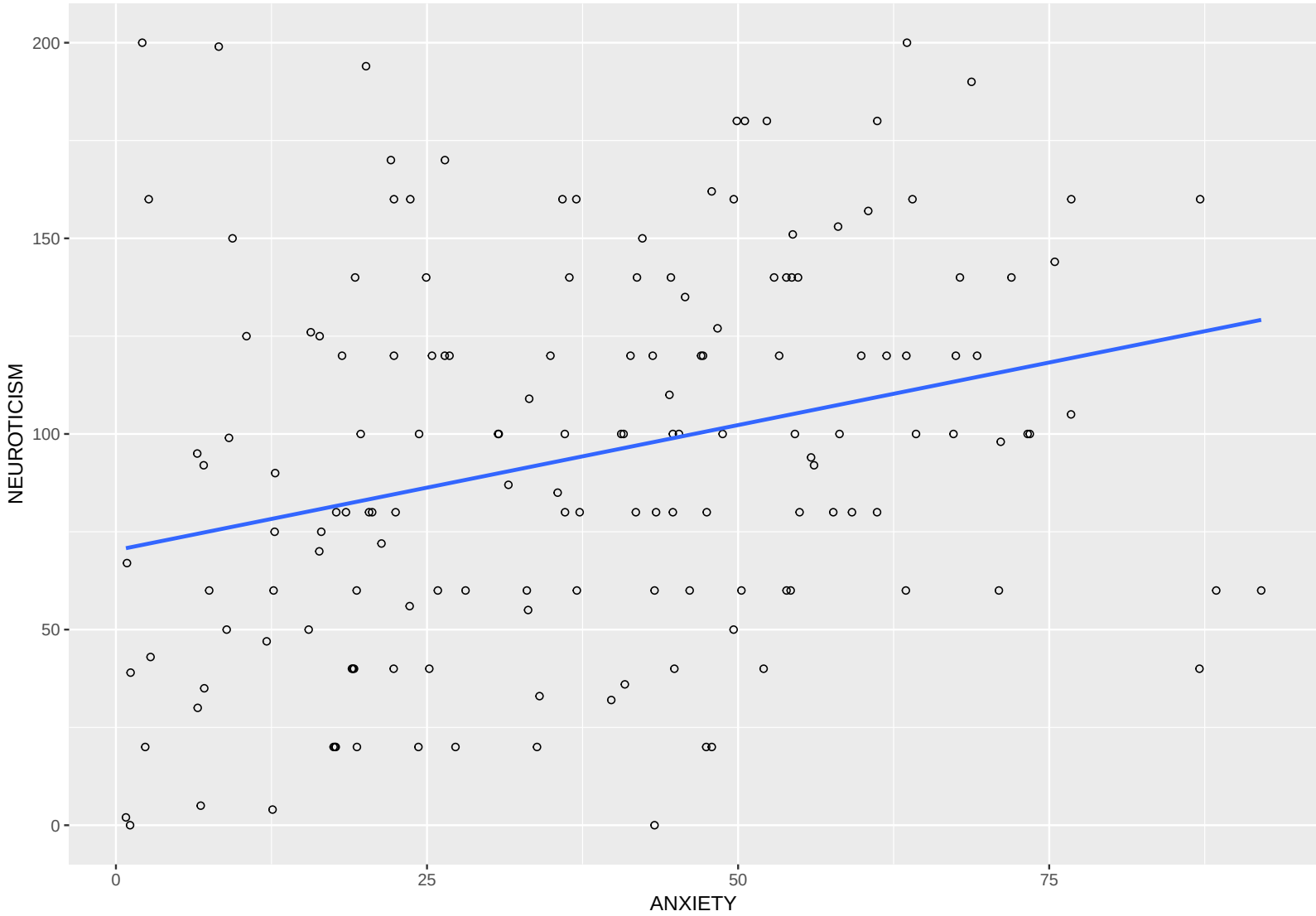
Scatter Plot

- Illustrates the relationship between two variables based on its underlying data points
- E.g. the link between neurotic personality traits and programming anxiety
- Scatter graph - a two-dimensional graph plot of pairs of points from two variables
- Relationships will vary in strength, line of best fit used to indicate magnitude through slope

Scatter Plot in R

```
ggplot(dat, aes(x=ANXIETY, y=NEUROTICISM)) +
    geom_point(shape=1) +
    geom_smooth(method=lm,
                se=FALSE)
```

Scatter Plot Output

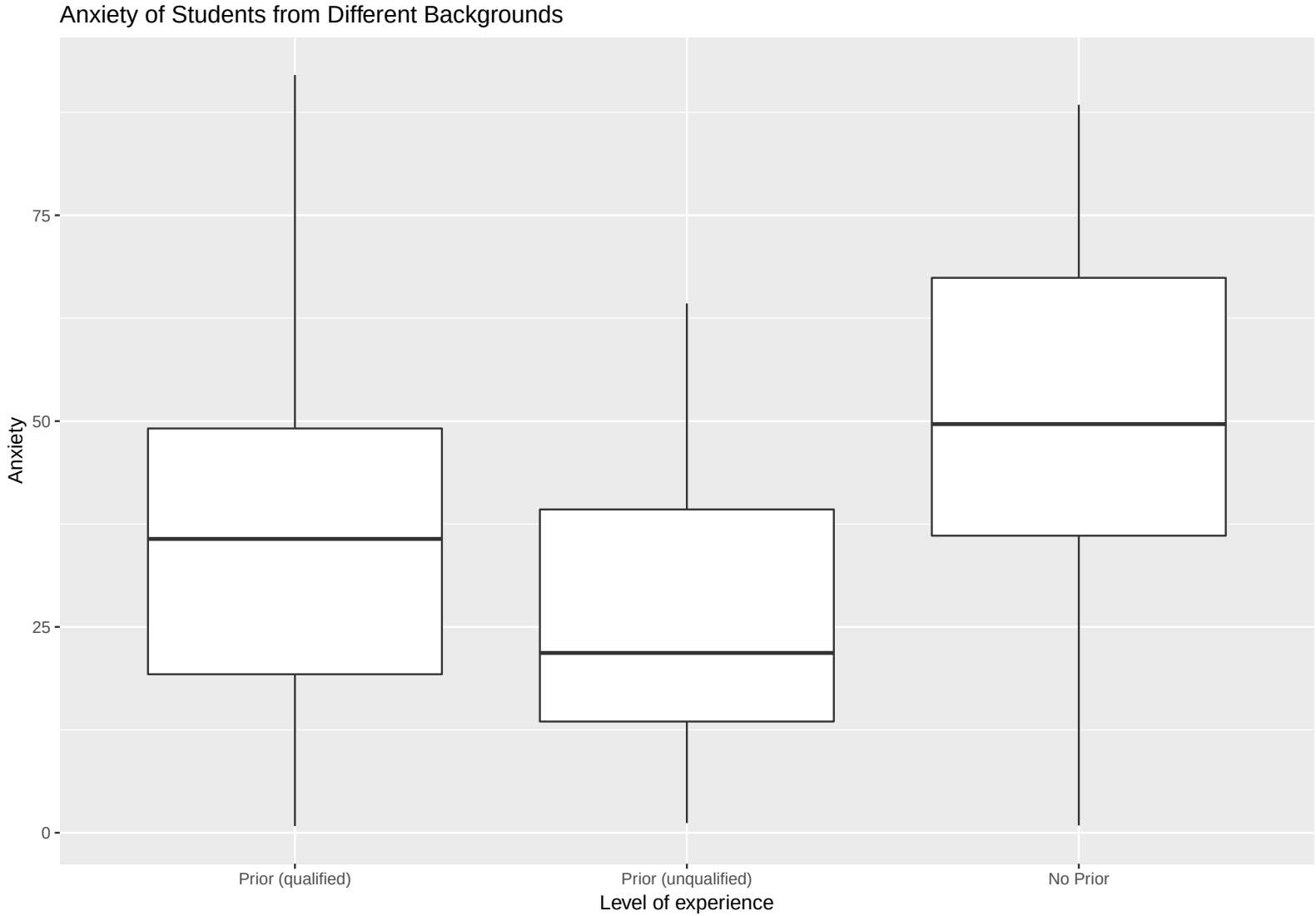


Box Plot in R

```
dat$P_EXP <- factor(
  dat$PRIOR_EXP,
  labels = c(
    "Prior (qualified)",
    "Prior (unqualified)",
    "No Prior"))

ggplot(dat, aes(x = P_EXP, y = ANXIETY)) +
  geom_boxplot() +
  labs(
    x = "Level of experience",
    y = "Anxiety",
    title = "Anxiety of Students")
```

Box Plots Output



Further Reading

- Official Docs
- Stat Methods
- Harvard Tutorial Series
- R Studio Docs
- R Markdown Docs

The Book

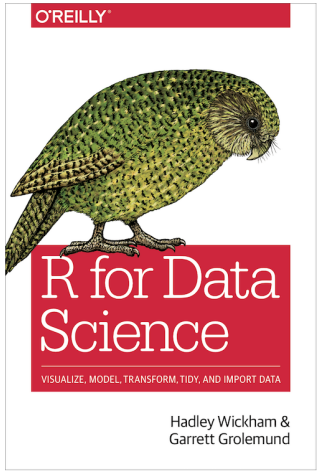
R for Data Science

Garrett Grolemund

Hadley Wickham

Welcome

This is the website for **"R for Data Science"**. This book will teach you how to do data science with R: You'll learn how to get your data into R, get it into the most useful structure, transform it, visualise it and model it. In this book, you will find a practicum of skills for data science. Just as a chemist learns how to clean test tubes and stock a lab, you'll learn how to clean data and draw plots—and many other things besides. These are the skills that allow data science to happen, and here you will find the best practices for doing each of these things with R. You'll learn how to use the grammar of graphics, literate programming, and reproducible research to save time. You'll also learn how to manage cognitive resources to facilitate discoveries when wrangling, visualising, and exploring data.



This website is (and will always be) **free to use**, and is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 License. If you'd like a **physical**

Figure 9: Link to free book: R for Data Science