BADM 372 Applied Analytics

BADM 372

2022-01-06

Contents

Chapter 1

About this course

This website serves as headquarters for **BADM 372 Intro to Applied Analytics**.

Content here will be updated with any changes made during the semester, so if at any point you are told there was a change in the assignment, you can come here to get the updated version.

Also, this book has benefited greatly from lots of free, readily available resources posted on the web and we leverage these extensively. I would encourage you to review these resources in your analytics journey. Some that we specifically use with great frequency are these (and I say loud THANK YOU to the authors!):

- R for Data Science
- An Introduction to Statistical Learning with Applications in R
- Data Science in a Box
- stackoverflow.com, for example

Chapter 2

Syllabus

Instructor: Tobin Turner

Office Hours: mutually convenient time arranged by email e-mail: jtturner@presby.edu

2.1 Course Objectives and Learning Outcomes

This course is designed to introduce to data science. Students will apply statistical knowledge and techniques to both business and non-business contexts.

At the end of this course students should be able to:

By the end of the course, you will be able to… - gain insight from data - gain insight from data, reproducibly - gain insight from data, reproducibly, using modern programming tools and techniques - gain insight from data, reproducibly and collaboratively, using modern programming tools and techniques - gain insight from data, reproducibly (with literate programming and version control) and collaboratively, using modern programming tools and techniques - communicate results effectively

This course will be focused on both understanding and applying key business analytical concepts. Although the text serves as a useful foundation for the concepts covered in the class, simple memorization of the material in the text will not be sufficient. Class participation, discussion, and application are critical.

2.2 Text and Resources

- This course website (primary resource)

- R for Data Science
- An Introduction to Statistical Learning with Applications in R
- Data Science in a Box
- stackoverflow.com, for example
- Other free, publicly available datasets and publications.

2.3 Performance Evaluation (Grading)

- Quizzes and Assignments - 40%
- Exam 1 - 20%
- Exam 2 - 20%
- Final Exam - 20%

2.3.1 Exams

Exams will cover assigned chapters in the textbook, other assigned readings, lectures, class exercises, class discussions, videos, and guest speakers. I will typically allocate time prior to each exam to clearly identify the body of knowledge each test will cover and to answer questions about the format and objectives of the exam.

2.3.2 Quizzes – DON"T MISS CLASS

- The average of all quizzes and assignments will comprise the Quizzes and Assignments - 40% portion of your final grade
- Quizzes and Assignments are designed to prepare you for your exams and to ensure you stay up with the course material
- **Missed Quizzes and Assignments cannot be made up later. Be present.**

Quizzes rule. **LISTEN.**

2.3.3 Final Average

- Final Average Grade
  - 90-100 A
  - 88-89 B+
  - 82-87 B+
  - 80-81 B-
  - 78-79 C+
  - 72-77 C+

- 70-71 C-
- 60-69 D
- 59 and below F

2.4 Class Participation:

I will frequently give readings or assignments for you to complete prior to the next class meeting. I expect you to fully engage the material: answer questions, pose questions, provide insightful observations. Keep in mind that quality is an important component in "participation." Periodic cold calls will take place. I will also put students in the "hot seat" on occasion. In these class sessions, I may select a random group of students to lead us in the discussion and debate. Because the selection of participants will not be announced until class begins, everyone will be expected to prepare for the discussion. Reading the assigned chapters and articles are the best way to prepare for the discussion. If you have concerns about being called on in class, please see me to discuss. The purpose of the "hot seat" is not to stress or embarrass students, but to encourage students to actively engage the material.

2.5 Phones

Phones are not allowed to be used in class without the instructor's prior consent. If you have a need of a phone during class please let me know before class. Unauthorized use of electronic devices may result in the lowering of the grade or dismissal from the class. **I mean this.**

2.6 Attendance

You are expected to be regular and punctual in your class attendance. Students are responsible for all the material missed and homework assignments made. If class is missed, notes/homework should be obtained from another student. If I am more than 15 minutes late, class is considered cancelled. No more than 4 absences are allowed during a semester. Exceeding the absence policy may result in receiving an F for the course. The professors roll is the official roll and students not present when roll is taken will be counted as absent. If a student must miss an exam, she or he must work out an agreeable time with the instructor to take the test prior to the exam being given. If a student misses a test due to an emergency, the student must inform the instructor as soon as is possible. In special cases, the instructor may allow the student to take a make-up exam.

2.7 Accommodations

Presbyterian College is committed to providing reasonable accommodations for all students with documented disabilities. If you are seeking academic accommodations under the Americans with Disabilities Act, you must register with the Academic Success Office, located on 5th Avenue (beside Campus Police). To receive these accommodations, please obtain the proper Accommodations Approval Form from that office, and then meet with me at the beginning of the semester to discuss how we may deliver your approved accommodations. I especially encourage you to meet with me well in advance of the actual accommodations being provided, as it may not be feasible to offer immediate accommodations without sufficient advance notice (such as in the case of tests). I can assure you that all discussions will remain confidential. Disability Services information is located at this link http://bit.ly/PCdisabilityservices

Additionally, it is the student's responsibility to give the instructor one week's notice prior to each instance where accommodation will be required.

2.8 Honor Code and Plagiarism:

All assignments/exams must be your own work. Any copying or use of unauthorized assistance will be treated as a violation of PC's Honor Code. If you are unsure of what resources are allowed, please ask. Please note that all text longer than 7 words taken from ANY other source must be placed in quotations and cited. Also, summarizing ANY other source must also be cited. Using ANY other source and showing work to be your own is a violation of plagiarism and the honor code.

2.9 First-Generation Version:

I am a Presby First+ Advocate. I am here to support our current first-generation students. At Presbyterian College, first-generation students are those in which neither parent nor legal guardian graduated from a four-year higher education institution with a bachelor's degree. If you are a first-generation college student, please contact me. For more information about support for first-generation college students on our campus visit our Presby First+ webpage.

2.10 Continuing Advocate Version

I am a Presby First+ Advocate. I am committed to supporting first-generation students at Presbyterian College. At Presbyterian College, first-generation students are those in which neither parent nor legal guardian graduated from a

four-year higher education institution with a bachelor's degree. If you are a first-generation college student, please contact me anytime or visit me during my office hours. For more information about support for first-generation college students on our campus visit our Presby First+ webpage.

Chapter 3

Schedule

This is a tentative schedule, and it will change. **BUT** I will do my very best to review this often so that we all stay on the same page and so that you may plan accordingly!

Spring 2022

Date	Topic
Monday, January 10, 2022	Intro and A1 review
Wednesday, January 12, 2022	A1 Review & Look ahead to trees & forests
Friday, January 14, 2022	QUIZ 1
Monday, January 17, 2022	MLK Holiday
Wednesday, January 19, 2022	RMARKDOWN
Friday, January 21, 2022	RMARKOWN WITH SLIDES ASSIGNMENT
Monday, January 24, 2022	GGPLOT
Wednesday, January 26, 2022	GGPLOT
Friday, January 28, 2022	GG PLOT QUIZ
Monday, January 31, 2022	TIDY SPREAD AND GATHER (R4DS CH 9 DPLYR)
Wednesday, February 2, 2022	RELATIONAL DATA (R4DS CH 10 DPLYR)
Friday, February 4, 2022	QUIZ
Monday, February 7, 2022	STRINGS (R4DS CH 11 stringr)
Wednesday, February 9, 2022	STRINGS (R4DS CH 12 factors)
Friday, February 11, 2022	QUIZ
Monday, February 14, 2022	Dates and Times
Wednesday, February 16, 2022	Dates and Times
Friday, February 18, 2022	QUIZ
Monday, February 21, 2022	Functions
Wednesday, February 23, 2022	Functions

Date	Topic
Friday, February 25, 2022	QUIZ
Monday, February 28, 2022	Itertation
Wednesday, March 2, 2022	Itertation
Friday, March 4, 2022	QUIZ
Monday, March 7, 2022	LAUNCH PROJECT
Wednesday, March 9, 2022	INDEPENDENT PROJECT
Friday, March 11, 2022	INDEPENDENT PROJECT
Monday, March 14, 2022	SPRING BREAK
Wednesday, March 16, 2022	SPRING BREAK
Friday, March 18, 2022	SPRING BREAK
Monday, March 21, 2022	INDEPENDENT PROJECT
Wednesday, March 23, 2022	PRESENTATIONS
Friday, March 25, 2022	PRESENTATIONS
Monday, March 28, 2022	Model Builiding/ADVISING WEEK
Wednesday, March 30, 2022	Model Builiding/ADVISING WEEK
Friday, April 1, 2022	QUIZ
Monday, April 4, 2022	regrssion
Wednesday, April 6, 2022	stepwise addition/deletion
Friday, April 8, 2022	QUIZ
Monday, April 11, 2022	logistic regression
Wednesday, April 13, 2022	trees & forests
Friday, April 15, 2022	Easter Holidays
Monday, April 18, 2022	Easter Holidays
Wednesday, April 20, 2022	Model Builiding
Friday, April 22, 2022	QUIZ
Monday, April 25, 2022	PRESENTATIONS
Wednesday, April 27, 2022	PRESENTATIONS
Friday, April 29, 2022	LAST DAY
Monday, May 2, 2022	Final Exam 8:30 p.m. – F period

Chapter 4

Applied Analytics Overview

Analytics is the systematic computational analysis of data or statistics.

It is used for the discovery, interpretation, and communication of meaningful patterns in data. It also entails applying data patterns towards effective decision-making.

It can be valuable in areas rich with recorded information; analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance.

Organizations may apply analytics to business data to describe, predict, and improve business performance.

Specifically, areas within analytics include predictive analytics, prescriptive analytics, enterprise decision management, descriptive analytics, cognitive analytics, Big Data Analytics, retail analytics, supply chain analytics, store assortment and stock-keeping unit optimization, marketing optimization and marketing mix modeling, web analytics, call analytics, speech analytics, sales force sizing and optimization, price and promotion modeling, predictive science, graph analytics, credit risk analysis, and fraud analytics.

Since analytics can require extensive computation (think: big data), the algorithms and software used for analytics harness the most current methods in computer science, statistics, and mathematics.

4.1 Feel good about Analytics 1?

- R basics
- Data wrangling
- Modeling (lm, glm, etc.)

  - lm
  - glm
  - test & training data
  - measures of fit, confusion matrix

4.2 Learning goals

By the end of the course, you will be able to…

Data Science Model

4.3 Toolkit for reproducibility

- Scriptability → R
- Literate programming (code, narrative, output in one place) → R Markdown
- Version control → Git / GitHub

4.4 Resources We'll Use (R4DS, mostly)

1) Explore
2) Wrangle
3) Program
4) Model
5) Communicate

4.5 Explore

3 Data visualisation 4 Workflow: basics 5 Data transformation 6 Workflow: scripts 7 Exploratory Data Analysis 8 Workflow: projects

4.6 Wrangle

9 Introduction 10 Tibbles 11 Data import 12 Tidy data 13 Relational data 14 Strings 15 Factors 16 Dates and times

4.7 Program

17 Introduction 18 Pipes 19 Functions 20 Vectors 21 Iteration

4.8 Model

22 Introduction 23 Model basics 24 Model building 25 Many models

4.9 Communicate

26 Introduction 27 R Markdown 28 Graphics for communication 29 R Markdown formats 30 R Markdown workflow

4.10 Learning goals

By the end of the course, you will be able to…

- gain insight from data
- gain insight from data, **reproducibly**
- gain insight from data, reproducibly, **using modern programming tools and techniques**
- gain insight from data, reproducibly **and collaboratively**, using modern programming tools and techniques
- gain insight from data, reproducibly **(with literate programming and version control)** and collaboratively, using modern programming tools and techniques

Chapter 5

Data and visualisation

Chapter 6

What is in a dataset?

6.1 Dataset terminology

- Each row is an **observation**
- Each column is a **variable**

```
starwars
#> # A tibble: 87 x 14
#>    name        height  mass hair_color  skin_color eye_color
#>    <chr>        <int> <dbl> <chr>       <chr>      <chr>
#>  1 Luke Skywa~    172    77 blond       fair       blue
#>  2 C-3PO          167    75 <NA>        gold       yellow
#>  3 R2-D2           96    32 <NA>        white, bl~ red
#>  4 Darth Vader    202   136 none        white      yellow
#>  5 Leia Organa    150    49 brown       light      brown
#>  6 Owen Lars      178   120 brown, grey light      blue
#>  7 Beru White~    165    75 brown       light      blue
#>  8 R5-D4           97    32 <NA>        white, red red
#>  9 Biggs Dark~    183    84 black       light      brown
#> 10 Obi-Wan Ke~    182    77 auburn, wh~ fair       blue-gray
#> # ... with 77 more rows, and 8 more variables:
#> #   birth_year <dbl>, sex <chr>, gender <chr>,
#> #   homeworld <chr>, species <chr>, films <list>,
#> #   vehicles <list>, starships <list>
```

6.2 What's in the Star Wars data?

Take a `glimpse` at the data:

```
#> Rows: 87
#> Columns: 14
#> $ name       <chr> "Luke Skywalker", "C-3PO", "R2-D2", "Da~
#> $ height     <int> 172, 167, 96, 202, 150, 178, 165, 97, 1~
#> $ mass       <dbl> 77, 75, 32, 136, 49, 120, 75, 32, 84, 7~
#> $ hair_color <chr> "blond", NA, NA, "none", "brown", "brow~
#> $ skin_color <chr> "fair", "gold", "white, blue", "white",~
#> $ eye_color  <chr> "blue", "yellow", "red", "yellow", "bro~
#> $ birth_year <dbl> 19.0, 112.0, 33.0, 41.9, 19.0, 52.0, 47~
#> $ sex        <chr> "male", "none", "none", "male", "female~
#> $ gender     <chr> "masculine", "masculine", "masculine", ~
#> $ homeworld  <chr> "Tatooine", "Tatooine", "Naboo", "Tatoo~
#> $ species    <chr> "Human", "Droid", "Droid", "Human", "Hu~
#> $ films      <list> <"The Empire Strikes Back", "Revenge o~
#> $ vehicles   <list> <"Snowspeeder", "Imperial Speeder Bike~
#> $ starships  <list> <"X-wing", "Imperial shuttle">, <>, <>~
```

6.3 Questions

- How many rows and columns does this dataset have?
- What does each row represent?
- What does each column represent?

6.4 Questions

```
?starwars
```

6.5 How many rows and columns does this dataset have?

```
nrow(starwars) # number of rows
#> [1] 87
ncol(starwars) # number of columns
#> [1] 14
dim(starwars)  # dimensions (row column)
#> [1] 87 14
```

Chapter 7

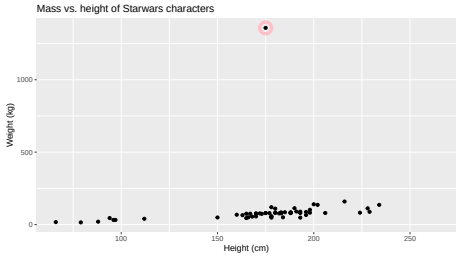
Exploratory data analysis

7.1 What is EDA?

- Exploratory data analysis (EDA) is an approach to analysing data sets to summarize its main characteristics
- Often, this is visual – this is what we'll focus on first
- But we might also calculate summary statistics and perform data wrangling/manipulation/transformation at (or before) this stage of the analysis – this is what we'll focus on next

7.2 Mass vs. height

- How would you describe the relationship between mass and height of Starwars characters?
- What other variables would help us understand data points that don't follow the overall trend?
- Who is the not so tall but really chubby character?



7.3 Jabba!

Chapter 8

Data visualization

8.1 Data visualization

> *"The simple graph has brought more information to the data analyst's mind than any other device." — John Tukey*

- Data visualization is the creation and study of the visual representation of data
- Many tools for visualizing data – R is one of them
- Many approaches/systems within R for making data visualizations – **ggplot2** is one of them, and that's what we're going to use

8.2 ggplot2 $\in$ tidyverse

8.3 ggplot2

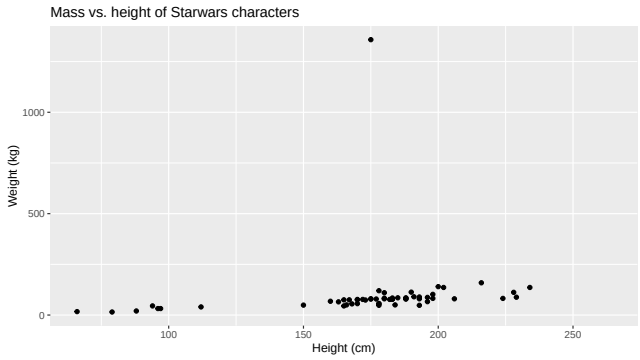
- **ggplot2** is tidyverse's data visualization package
- `gg` in "ggplot2" stands for Grammar of Graphics
- Inspired by the book **Grammar of Graphics** by Leland Wilkinson

8.4 Grammar of Graphics

A grammar of graphics is a tool that enables us to concisely describe the components of a graphic

8.5 Mass vs. height

```
ggplot(data = starwars, mapping = aes(x = height, y = mass)) +
  geom_point() +
  labs(title = "Mass vs. height of Starwars characters",
       x = "Height (cm)", y = "Weight (kg)")
#> Warning: Removed 28 rows containing missing values
#> (geom_point).
```



8.6 Questions

- What are the functions doing the plotting?
- What is the dataset being plotted?
- Which variables map to which features (aesthetics) of the plot?
- What does the warning mean?+

```
ggplot(data = starwars, mapping = aes(x = height, y = mass)) +
  geom_point() +
  labs(title = "Mass vs. height of Starwars characters",
       x = "Height (cm)", y = "Weight (kg)")
#> Warning: Removed 28 rows containing missing values
#> (geom_point).
```

8.7 Hello ggplot2!

- `ggplot()` is the main function in ggplot2
- Plots are constructed in layers
- Structure of the code for plots can be summarized as

```
ggplot(data = [dataset],
       mapping = aes(x = [x-variable], y = [y-variable])) +
   geom_xxx() +
   other options
```

- The ggplot2 package comes with the tidyverse

```
library(tidyverse)
```

- For help with ggplot2, see ggplot2.tidyverse.org

Chapter 9

Why do we visualize?

9.1 Anscombe's quartet

```
#>    set  x     y
#> 1    I 10  8.04
#> 2    I  8  6.95
#> 3    I 13  7.58
#> 4    I  9  8.81
#> 5    I 11  8.33
#> 6    I 14  9.96
#> 7    I  6  7.24
#> 8    I  4  4.26
#> 9    I 12 10.84
#> 10   I  7  4.82
#> 11   I  5  5.68
#> 12  II 10  9.14
#> 13  II  8  8.14
#> 14  II 13  8.74
#> 15  II  9  8.77
#> 16  II 11  9.26
#> 17  II 14  8.10
#> 18  II  6  6.13
#> 19  II  4  3.10
#> 20  II 12  9.13
#> 21  II  7  7.26
#> 22  II  5  4.74
#> 23 III 10  7.46
#> 24 III  8  6.77
#> 25 III 13 12.74
#> 26 III  9  7.11
```

```
#> 27 III 11  7.81
#> 28 III 14  8.84
#> 29 III  6  6.08
#> 30 III  4  5.39
#> 31 III 12  8.15
#> 32 III  7  6.42
#> 33 III  5  5.73
#> 34  IV  8  6.58
#> 35  IV  8  5.76
#> 36  IV  8  7.71
#> 37  IV  8  8.84
#> 38  IV  8  8.47
#> 39  IV  8  7.04
#> 40  IV  8  5.25
#> 41  IV 19 12.50
#> 42  IV  8  5.56
#> 43  IV  8  7.91
#> 44  IV  8  6.89
```

```
#>    set  x     y
#> 1    I 10  8.04
#> 2    I  8  6.95
#> 3    I 13  7.58
#> 4    I  9  8.81
#> 5    I 11  8.33
#> 6    I 14  9.96
#> 7    I  6  7.24
#> 8    I  4  4.26
#> 9    I 12 10.84
#> 10   I  7  4.82
#> 11   I  5  5.68
#> 12  II 10  9.14
#> 13  II  8  8.14
#> 14  II 13  8.74
#> 15  II  9  8.77
#> 16  II 11  9.26
#> 17  II 14  8.10
#> 18  II  6  6.13
#> 19  II  4  3.10
#> 20  II 12  9.13
#> 21  II  7  7.26
#> 22  II  5  4.74
```

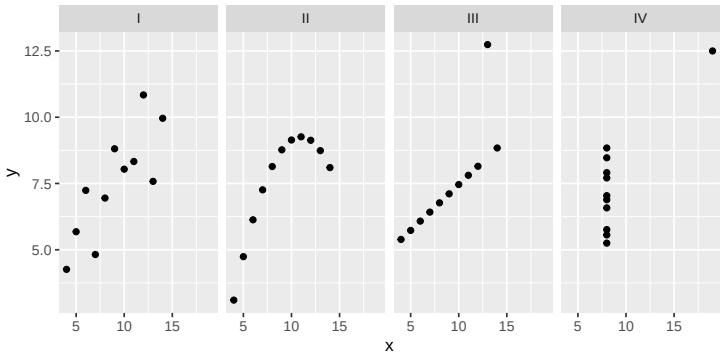
```
#>    set  x     y
#> 23 III 10  7.46
```

```
#> 24 III  8  6.77
#> 25 III 13 12.74
#> 26 III  9  7.11
#> 27 III 11  7.81
#> 28 III 14  8.84
#> 29 III  6  6.08
#> 30 III  4  5.39
#> 31 III 12  8.15
#> 32 III  7  6.42
#> 33 III  5  5.73
#> 34  IV  8  6.58
#> 35  IV  8  5.76
#> 36  IV  8  7.71
#> 37  IV  8  8.84
#> 38  IV  8  8.47
#> 39  IV  8  7.04
#> 40  IV  8  5.25
#> 41  IV 19 12.50
#> 42  IV  8  5.56
#> 43  IV  8  7.91
#> 44  IV  8  6.89
```

9.2 Summarising Anscombe's quartet

```
quartet %>%
  group_by(set) %>%
  summarise(
    mean_x = mean(x),
    mean_y = mean(y),
    sd_x = sd(x),
    sd_y = sd(y),
    r = cor(x, y))
#> # A tibble: 4 x 6
#>   set   mean_x mean_y  sd_x  sd_y     r
#>   <fct>  <dbl>  <dbl> <dbl> <dbl> <dbl>
#> 1 I          9   7.50  3.32  2.03 0.816
#> 2 II         9   7.50  3.32  2.03 0.816
#> 3 III        9   7.5   3.32  2.03 0.816
#> 4 IV         9   7.50  3.32  2.03 0.817
```

9.3 Visualizing Anscombe's quartet

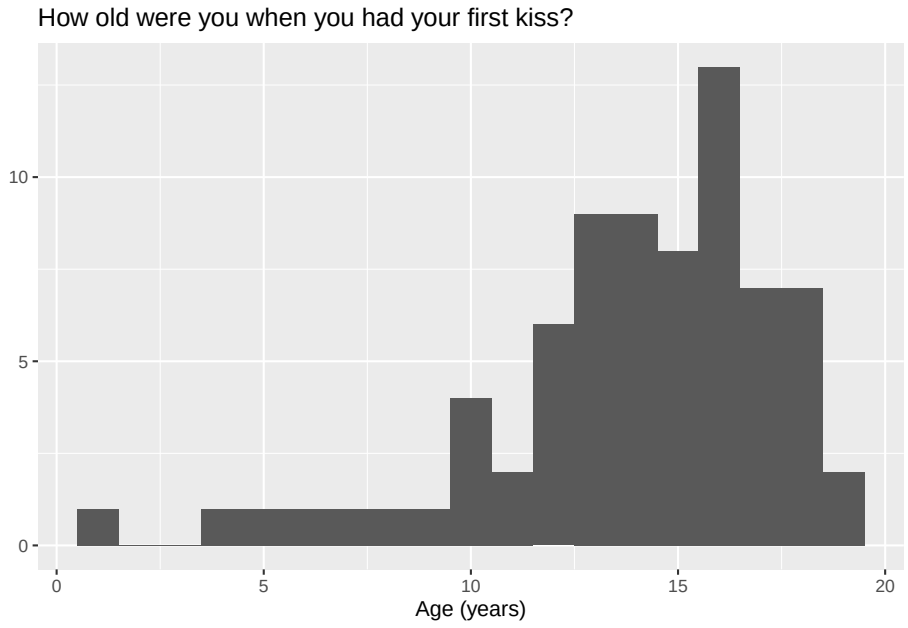


9.4 About Anscombe's quartet

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are *ROUGH*."

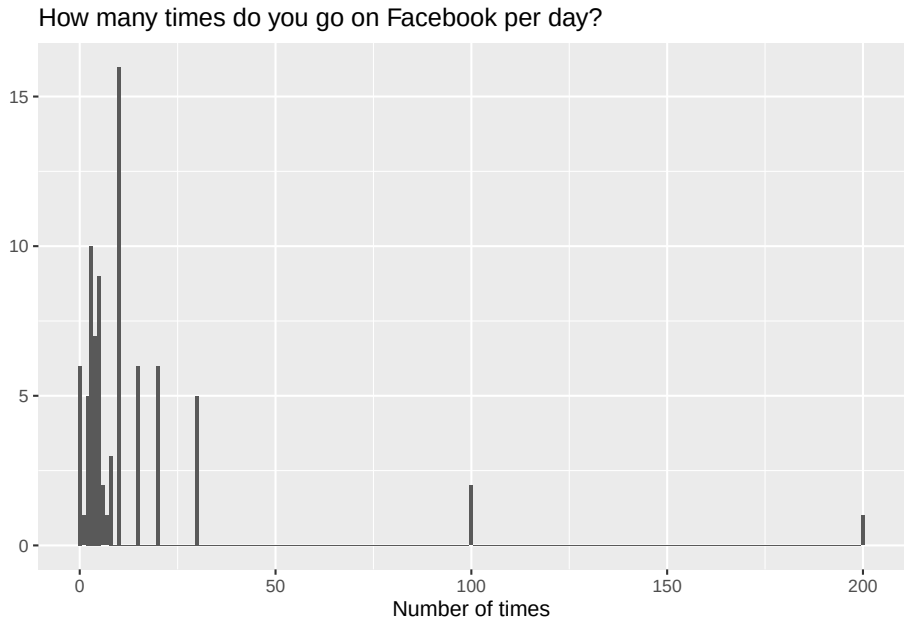
9.5 Age at first kiss

- Do you see anything out of the ordinary?



9.6 Facebook visits

- How are people reporting lower vs. higher values of FB visits?


How many times do you go on Facebook per day?
15
10
5
0
0
50
100
150
200
Number of times

Chapter 10

Reproducible data analysis

What is reproducible research? Wy is it important?

> Reproducible research is the idea that data analyses, and more generally, scientific claims, are published with their data and software code so that others may verify the findings and build upon them. The need for reproducibility is increasing dramatically as data analyses become more complex, involving larger datasets and more sophisticated computations. Reproducibility allows for people to focus on the actual content of a data analysis, rather than on superficial details reported in a written summary. In addition, reproducibility makes an analysis more useful to others because the data and code that actually conducted the analysis are available. This course will focus on literate statistical analysis tools which allow one to publish data analyses in a single document that allows others to easily execute the same analysis to obtain the same results.

10.1 Reproducibility checklist

10.1.1 Near-term goals:

- Are the tables and figures reproducible from the code and data?
- Does the code actually do what you think it does?
- In addition to what was done, is it clear *why* it was done?

10.1.2 Long-term goals:

- Can the code be used for other data?

- Can you extend the code to do other things?

Chapter 11

Data and visualisation

Chapter 12

What is in a dataset?

12.1 Dataset terminology

- Each row is an **observation**
- Each column is a **variable**

```
starwars
#> # A tibble: 87 x 14
#>    name        height  mass hair_color  skin_color eye_color
#>    <chr>        <int> <dbl> <chr>       <chr>      <chr>
#>  1 Luke Skywa~    172    77 blond       fair       blue
#>  2 C-3PO          167    75 <NA>        gold       yellow
#>  3 R2-D2           96    32 <NA>        white, bl~ red
#>  4 Darth Vader    202   136 none        white      yellow
#>  5 Leia Organa    150    49 brown       light      brown
#>  6 Owen Lars      178   120 brown, grey light      blue
#>  7 Beru White~    165    75 brown       light      blue
#>  8 R5-D4           97    32 <NA>        white, red red
#>  9 Biggs Dark~    183    84 black       light      brown
#> 10 Obi-Wan Ke~    182    77 auburn, wh~ fair       blue-gray
#> # ... with 77 more rows, and 8 more variables:
#> #   birth_year <dbl>, sex <chr>, gender <chr>,
#> #   homeworld <chr>, species <chr>, films <list>,
#> #   vehicles <list>, starships <list>
```

12.2 What's in the Star Wars data?

Take a `glimpse` at the data:

```
#> Rows: 87
#> Columns: 14
#> $ name       <chr> "Luke Skywalker", "C-3PO", "R2-D2", "Da~
#> $ height     <int> 172, 167, 96, 202, 150, 178, 165, 97, 1~
#> $ mass       <dbl> 77, 75, 32, 136, 49, 120, 75, 32, 84, 7~
#> $ hair_color <chr> "blond", NA, NA, "none", "brown", "brow~
#> $ skin_color <chr> "fair", "gold", "white, blue", "white",~
#> $ eye_color  <chr> "blue", "yellow", "red", "yellow", "bro~
#> $ birth_year <dbl> 19.0, 112.0, 33.0, 41.9, 19.0, 52.0, 47~
#> $ sex        <chr> "male", "none", "none", "male", "female~
#> $ gender     <chr> "masculine", "masculine", "masculine", ~
#> $ homeworld  <chr> "Tatooine", "Tatooine", "Naboo", "Tatoo~
#> $ species    <chr> "Human", "Droid", "Droid", "Human", "Hu~
#> $ films      <list> <"The Empire Strikes Back", "Revenge o~
#> $ vehicles   <list> <"Snowspeeder", "Imperial Speeder Bike~
#> $ starships  <list> <"X-wing", "Imperial shuttle">, <>, <>~
```

12.3 Questions

- How many rows and columns does this dataset have?
- What does each row represent?
- What does each column represent?

12.4 Questions

```
?starwars
```

12.5 How many rows and columns does this dataset have?

```
nrow(starwars) # number of rows
#> [1] 87
ncol(starwars) # number of columns
#> [1] 14
dim(starwars)  # dimensions (row column)
#> [1] 87 14
```

Chapter 13

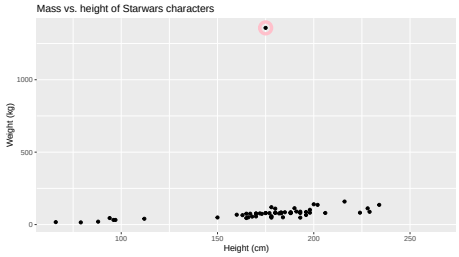
Exploratory data analysis

13.1 What is EDA?

- Exploratory data analysis (EDA) is an approach to analysing data sets to summarize its main characteristics
- Often, this is visual – this is what we'll focus on first
- But we might also calculate summary statistics and perform data wrangling/manipulation/transformation at (or before) this stage of the analysis – this is what we'll focus on next

13.2 Mass vs. height

- How would you describe the relationship between mass and height of Starwars characters?
- What other variables would help us understand data points that don't follow the overall trend?
- Who is the not so tall but really chubby character?



13.3 Jabba!

Chapter 14

Data visualization

14.1 Data visualization

> *"The simple graph has brought more information to the data analyst's mind than any other device." — John Tukey*

- Data visualization is the creation and study of the visual representation of data
- Many tools for visualizing data – R is one of them
- Many approaches/systems within R for making data visualizations – **ggplot2** is one of them, and that's what we're going to use

14.2 ggplot2 $\in$ tidyverse

14.3 ggplot2

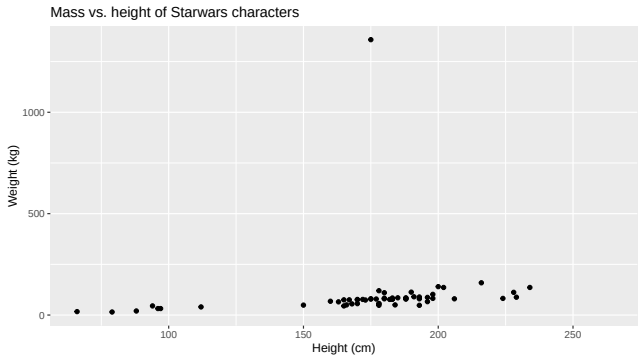
- **ggplot2** is tidyverse's data visualization package
- `gg` in "ggplot2" stands for Grammar of Graphics
- Inspired by the book **Grammar of Graphics** by Leland Wilkinson

14.4 Grammar of Graphics

A grammar of graphics is a tool that enables us to concisely describe the components of a graphic

14.5 Mass vs. height

```
ggplot(data = starwars, mapping = aes(x = height, y = mass)) +
  geom_point() +
  labs(title = "Mass vs. height of Starwars characters",
       x = "Height (cm)", y = "Weight (kg)")
#> Warning: Removed 28 rows containing missing values
#> (geom_point).
```



14.6 Questions

- What are the functions doing the plotting?
- What is the dataset being plotted?
- Which variables map to which features (aesthetics) of the plot?
- What does the warning mean?+

```
ggplot(data = starwars, mapping = aes(x = height, y = mass)) +
  geom_point() +
  labs(title = "Mass vs. height of Starwars characters",
       x = "Height (cm)", y = "Weight (kg)")
#> Warning: Removed 28 rows containing missing values
#> (geom_point).
```

14.7 Hello ggplot2!

- `ggplot()` is the main function in ggplot2
- Plots are constructed in layers
- Structure of the code for plots can be summarized as

```
ggplot(data = [dataset],
       mapping = aes(x = [x-variable], y = [y-variable])) +
   geom_xxx() +
   other options
```

- The ggplot2 package comes with the tidyverse

```
library(tidyverse)
```

- For help with ggplot2, see ggplot2.tidyverse.org

Chapter 15

Why do we visualize?

15.1 Anscombe's quartet

```
#>    set  x     y
#> 1    I 10  8.04
#> 2    I  8  6.95
#> 3    I 13  7.58
#> 4    I  9  8.81
#> 5    I 11  8.33
#> 6    I 14  9.96
#> 7    I  6  7.24
#> 8    I  4  4.26
#> 9    I 12 10.84
#> 10   I  7  4.82
#> 11   I  5  5.68
#> 12  II 10  9.14
#> 13  II  8  8.14
#> 14  II 13  8.74
#> 15  II  9  8.77
#> 16  II 11  9.26
#> 17  II 14  8.10
#> 18  II  6  6.13
#> 19  II  4  3.10
#> 20  II 12  9.13
#> 21  II  7  7.26
#> 22  II  5  4.74
#> 23 III 10  7.46
#> 24 III  8  6.77
#> 25 III 13 12.74
#> 26 III  9  7.11
```

```
#> 27 III 11  7.81
#> 28 III 14  8.84
#> 29 III  6  6.08
#> 30 III  4  5.39
#> 31 III 12  8.15
#> 32 III  7  6.42
#> 33 III  5  5.73
#> 34  IV  8  6.58
#> 35  IV  8  5.76
#> 36  IV  8  7.71
#> 37  IV  8  8.84
#> 38  IV  8  8.47
#> 39  IV  8  7.04
#> 40  IV  8  5.25
#> 41  IV 19 12.50
#> 42  IV  8  5.56
#> 43  IV  8  7.91
#> 44  IV  8  6.89
```

```
#>    set  x     y
#> 1    I 10  8.04
#> 2    I  8  6.95
#> 3    I 13  7.58
#> 4    I  9  8.81
#> 5    I 11  8.33
#> 6    I 14  9.96
#> 7    I  6  7.24
#> 8    I  4  4.26
#> 9    I 12 10.84
#> 10   I  7  4.82
#> 11   I  5  5.68
#> 12  II 10  9.14
#> 13  II  8  8.14
#> 14  II 13  8.74
#> 15  II  9  8.77
#> 16  II 11  9.26
#> 17  II 14  8.10
#> 18  II  6  6.13
#> 19  II  4  3.10
#> 20  II 12  9.13
#> 21  II  7  7.26
#> 22  II  5  4.74
```

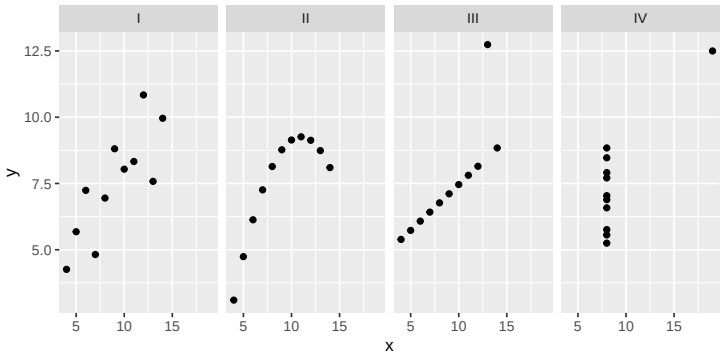
```
#>    set  x     y
#> 23 III 10  7.46
```

```
#> 24 III  8  6.77
#> 25 III 13 12.74
#> 26 III  9  7.11
#> 27 III 11  7.81
#> 28 III 14  8.84
#> 29 III  6  6.08
#> 30 III  4  5.39
#> 31 III 12  8.15
#> 32 III  7  6.42
#> 33 III  5  5.73
#> 34  IV  8  6.58
#> 35  IV  8  5.76
#> 36  IV  8  7.71
#> 37  IV  8  8.84
#> 38  IV  8  8.47
#> 39  IV  8  7.04
#> 40  IV  8  5.25
#> 41  IV 19 12.50
#> 42  IV  8  5.56
#> 43  IV  8  7.91
#> 44  IV  8  6.89
```

15.2 Summarising Anscombe's quartet

```
quartet %>%
  group_by(set) %>%
  summarise(
    mean_x = mean(x),
    mean_y = mean(y),
    sd_x = sd(x),
    sd_y = sd(y),
    r = cor(x, y))
#> # A tibble: 4 x 6
#>   set   mean_x mean_y  sd_x  sd_y     r
#>   <fct>  <dbl>  <dbl> <dbl> <dbl> <dbl>
#> 1 I          9   7.50  3.32  2.03 0.816
#> 2 II         9   7.50  3.32  2.03 0.816
#> 3 III        9   7.5   3.32  2.03 0.816
#> 4 IV         9   7.50  3.32  2.03 0.817
```

15.3 Visualizing Anscombe's quartet

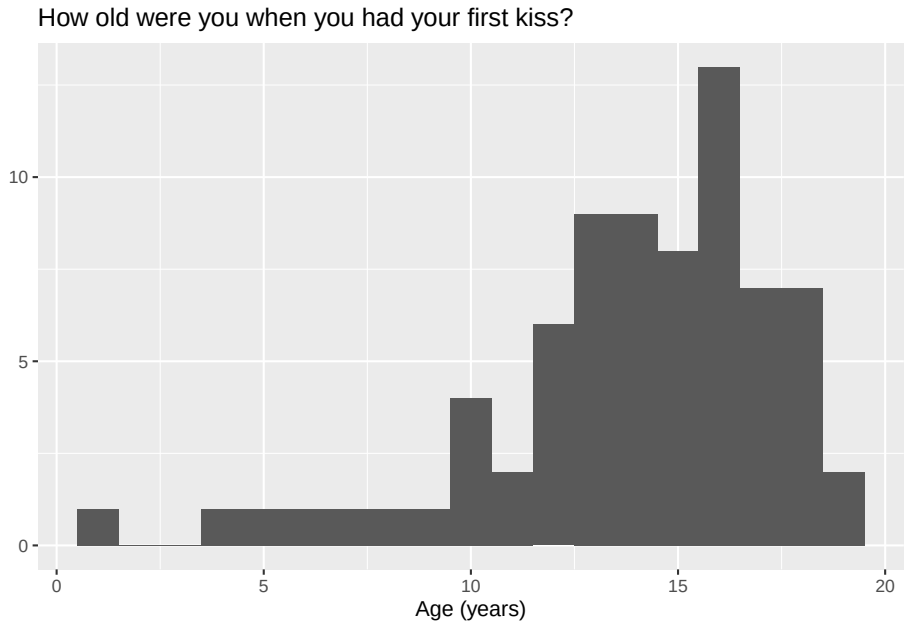


15.4 About Anscombe's quartet

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are *ROUGH*."

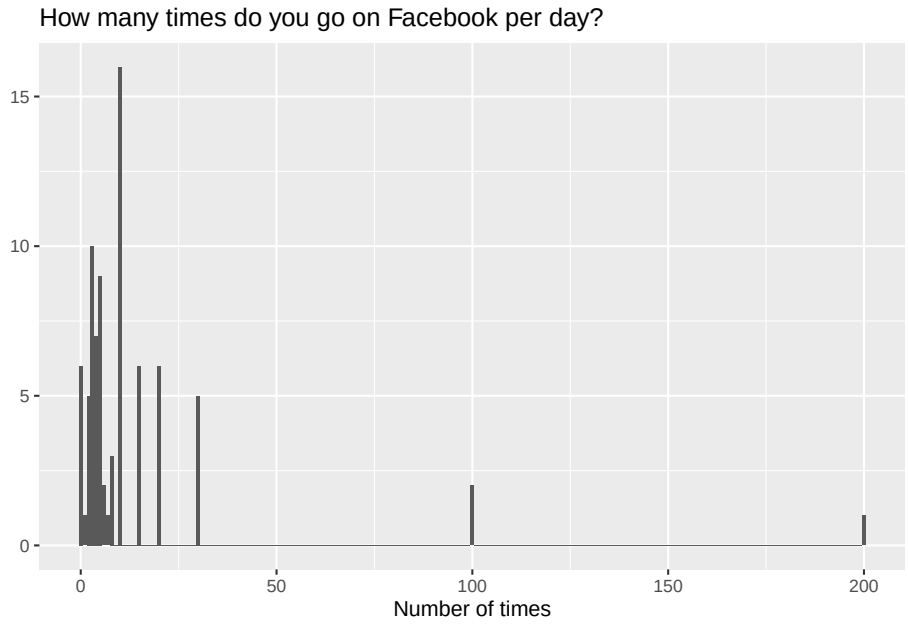
15.5 Age at first kiss

- Do you see anything out of the ordinary?



15.6 Facebook visits

- How are people reporting lower vs. higher values of FB visits?



15.7 ggplot

15.8 For next class

- Flip (or tab!) through R4DS and be able to answer (I **WILL** call on you next class) broadly-speaking what each chapter of the book covers
- Pay particular attention to the visualization chapter and be able to answer (I **WILL** call on you next class) what these terms refer to:
  - mapping
  - data
  - geom
  - stat
  - position

15.9 ggplot2 components

15.10 Quiz Next Class

- Quiz: explain:

- mapping
- data
- geom
- stat
- position

link

15.11 mapping

A set of aesthetic mappings, specified using the aes() function and combined with the plot defaults as described in aesthetic mappings. If NULL, uses the default mapping set in ggplot().

15.12 data

A dataset which overrides the default plot dataset. It is usually omitted (set to NULL), in which case the layer will use the default data specified in ggplot(). The requirements for data are explained in more detail in data.

15.13 Geoms

The name of the geometric object to use to draw each observation. Geoms are discussed in more detail in geom, and the toolbox explores their use in more depth.

15.14 Geoms

can have additional arguments. All geoms take aesthetics as parameters. If you supply an aesthetic (e.g. colour) as a parameter, it will not be scaled, allowing you to control the appearance of the plot, as described in setting vs. mapping. You can pass params in … (in which case stat and geom parameters are automatically teased apart), or in a list passed to geom_params.

15.15 stat

The name of the statistical tranformation to use. A statistical transformation performs some useful statistical summary is key to histograms and smoothes.

To keep the data as is, use the "identity" stat. Learn more in statistical transformations.

You only need to set one of stat and geom: every geom has a default stat, and every stat a default geom.

Most stats take additional parameters to specify the details of statistical transformation. You can supply params either in … (in which case stat and geom parameters are automatically teased apart), or in a list called stat_params.

15.16 position

The method used to adjusting overlapping objects, like jittering, stacking or dodging. More details in position.

Chapter 16

About

Originally from the default index.

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports; for example, a math equation $a^2 + b^2 = c^2$.

16.1 Usage

Each **bookdown** chapter is an .Rmd file, and each .Rmd file can contain one (and only one) chapter. A chapter *must* start with a first-level heading: `# A good chapter`, and can contain one (and only one) first-level heading.

Use second-level and higher headings within chapters like: `## A short section` or `### An even shorter section`.

The `index.Rmd` file is required, and is also your first book chapter. It will be the homepage when you render the book.

16.2 Render book

You can render the HTML version of this example book without changing anything:

1. Find the **Build** pane in the RStudio IDE, and
2. Click on **Build Book**, then select your output format, or select "All formats" if you'd like to use multiple formats from the same book source files.

Or build the book from the R console:

```
bookdown::render_book()
```

To render this example to PDF as a `bookdown::pdf_book`, you'll need to install XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.org/tinytex/.

16.3 Preview book

As you work, you may start a local server to live preview this HTML book. This preview will update as you edit the book when you save individual .Rmd files. You can start the server in a work session by using the RStudio add-in "Preview book", or from the R console:

```
bookdown::serve_book()
```

Chapter 17

Hello bookdown

All chapters start with a first-level heading followed by your chapter title, like the line above. There should be only one first-level heading (`#`) per .Rmd file.

17.1 A section

All chapter sections start with a second-level (`##`) or higher heading followed by your section title, like the sections above and below here. You can have as many as you want within a chapter.

An unnumbered section

Chapters and sections are numbered by default. To un-number a heading, add a `{.unnumbered}` or the shorter `{-}` at the end of the heading, like in this section.

Chapter 18

Cross-references

Cross-references make it easier for your readers to find and link to elements in your book.

18.1 Chapters and sub-chapters

There are two steps to cross-reference any heading:

1. Label the heading: `# Hello world {#nice-label}`.
    - Leave the label off if you like the automated heading generated based on your heading title: for example, `# Hello world` = `# Hello world {#hello-world}`.
    - To label an un-numbered heading, use: `# Hello world {-#nice-label}` or `{# Hello world .unnumbered}`.
2. Next, reference the labeled heading anywhere in the text using `\@ref(nice-label)`; for example, please see Chapter 18.
    - If you prefer text as the link instead of a numbered reference use: any text you want can go here.

18.2 Captioned figures and tables

Figures and tables *with captions* can also be cross-referenced from elsewhere in your book using `\@ref(fig:chunk-label)` and `\@ref(tab:chunk-label)`, respectively.

See Figure 18.1.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

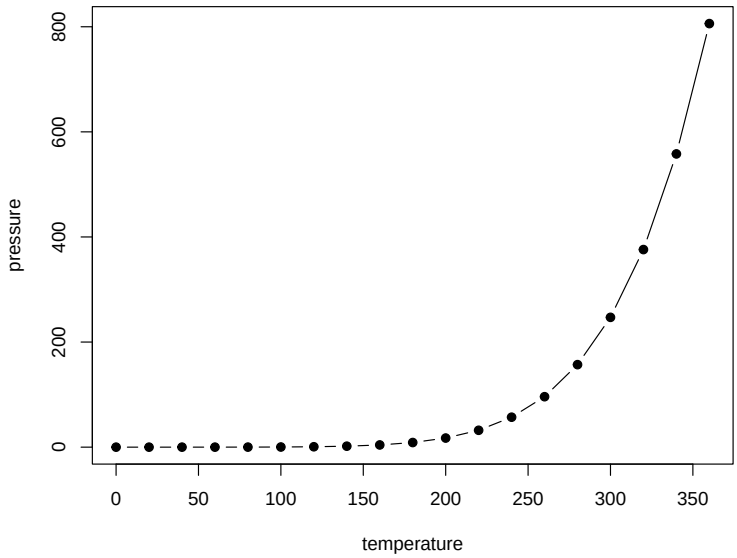


Figure 18.1: Here is a nice figure!

Don't miss Table 18.1.

```
knitr::kable(
  head(pressure, 10), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

Table 18.1: Here is a nice table!

temperature	pressure
0	0.0002
20	0.0012
40	0.0060
60	0.0300
80	0.0900
100	0.2700
120	0.7500
140	1.8500
160	4.2000
180	8.8000

Chapter 19

Parts

You can add parts to organize one or more book chapters together. Parts can be inserted at the top of an .Rmd file, before the first-level chapter heading in that same file.

Add a numbered part: `# (PART) Act one {-}` (followed by `# A chapter`)

Add an unnumbered part: `# (PART\*) Act one {-}` (followed by `# A chapter`)

Add an appendix as a special kind of un-numbered part: `# (APPENDIX) Other stuff {-}` (followed by `# A chapter`). Chapters in an appendix are prepended with letters instead of numbers.

Chapter 20

Footnotes and citations

20.1 Footnotes

Footnotes are put inside the square brackets after a caret `^[]`. Like this one [1].

20.2 Citations

Reference items in your bibliography file(s) using `@key`.

For example, we are using the **bookdown** package (Xie, 2021) (check out the last code chunk in index.Rmd to see how this citation key was added) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015) (this citation was added manually in an external file book.bib). Note that the `.bib` files need to be listed in the index.Rmd with the YAML `bibliography` key.

The `bs4_book` theme makes footnotes appear inline when you click on them. In this example book, we added `csl: chicago-fullnote-bibliography.csl` to the `index.Rmd` YAML, and include the `.csl` file. To download a new style, we recommend: https://www.zotero.org/styles/

The RStudio Visual Markdown Editor can also make it easier to insert citations: https://rstudio.github.io/visual-markdown-editing/#/citations

[1] This is a footnote.

Chapter 21

Blocks

21.1 Equations

Here is an equation.

$$f\left(k\right) = \binom{n}{k} p^k \left(1-p\right)^{n-k} \tag{21.1}$$

You may refer to using `\@ref(eq:binom)`, like see Equation (21.1).

21.2 Theorems and proofs

Labeled theorems can be referenced in text using `\@ref(thm:tri)`, for example, check out this smart theorem 21.1.

Theorem 21.1. *For a right triangle, if c denotes the* length *of the hypotenuse and a and b denote the lengths of the* ***other*** *two sides, we have*

$$a^2 + b^2 = c^2$$

Read more here https://bookdown.org/yihui/bookdown/markdown-extensions-by-bookdown.html.

21.3 Callout blocks

The `bs4_book` theme also includes special callout blocks, like this `.rmdnote`.

You can use **markdown** inside a block.

```
head(beaver1, n = 5)
#>   day time  temp activ
#> 1 346  840 36.33     0
#> 2 346  850 36.34     0
#> 3 346  900 36.35     0
#> 4 346  910 36.42     0
#> 5 346  920 36.55     0
```

It is up to the user to define the appearance of these blocks for LaTeX output.

You may also use: `.rmdcaution`, `.rmdimportant`, `.rmdtip`, or `.rmdwarning` as the block name.

The R Markdown Cookbook provides more help on how to use custom blocks to design your own callouts: https://bookdown.org/yihui/rmarkdown-cookbook/custom-blocks.html

Chapter 22

Sharing your book

22.1 Publishing

HTML books can be published online, see: https://bookdown.org/yihui/bookdown/publishing.html

22.2 404 pages

By default, users will be directed to a 404 page if they try to access a webpage that cannot be found. If you'd like to customize your 404 page instead of using the default, you may add either a `_404.Rmd` or `_404.md` file to your project root and use code and/or Markdown syntax.

22.3 Metadata for sharing

Bookdown HTML books will provide HTML metadata for social sharing on platforms like Twitter, Facebook, and LinkedIn, using information you provide in the `index.Rmd` YAML. To setup, set the `url` for your book and the path to your `cover-image` file. Your book's `title` and `description` are also used.

This `bs4_book` provides enhanced metadata for social sharing, so that each chapter shared will have a unique description, auto-generated based on the content.

Specify your book's source repository on GitHub as the `repo` in the `_output.yml` file, which allows users to view each chapter's source file or suggest an edit. Read more about the features of this output format here:

https://pkgs.rstudio.com/bookdown/reference/bs4_book.html

Or use:

```
?bookdown::bs4_book
```

Bibliography

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2021). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.24.