Stats 101C - Homework 1

Instructions

Summer 2023

Homework 1 Requirements

You will submit two files.

The files you submit will be:

1. `101C_HW1_First_Last.Rmd` Take the provided R Markdown file and make the necessary edits so that it generates the requested output.

2. `101C_HW1_First_Last.pdf` Your output file. This must be a PDF. This is the primary file that will be graded. **Make sure all requested output is visible in the output file.**

Academic Integrity

Include the following statement after modifying it with your name.

"By including this statement, I, **Luke Villanueva**, declare that all of the work in this assignment is my own original work. At no time did I look at the code of other students nor did I search for code solutions online. I understand that plagiarism on any single part of this assignment will result in a 0 for the entire assignment and that I will be referred to the dean of students."

Reading:

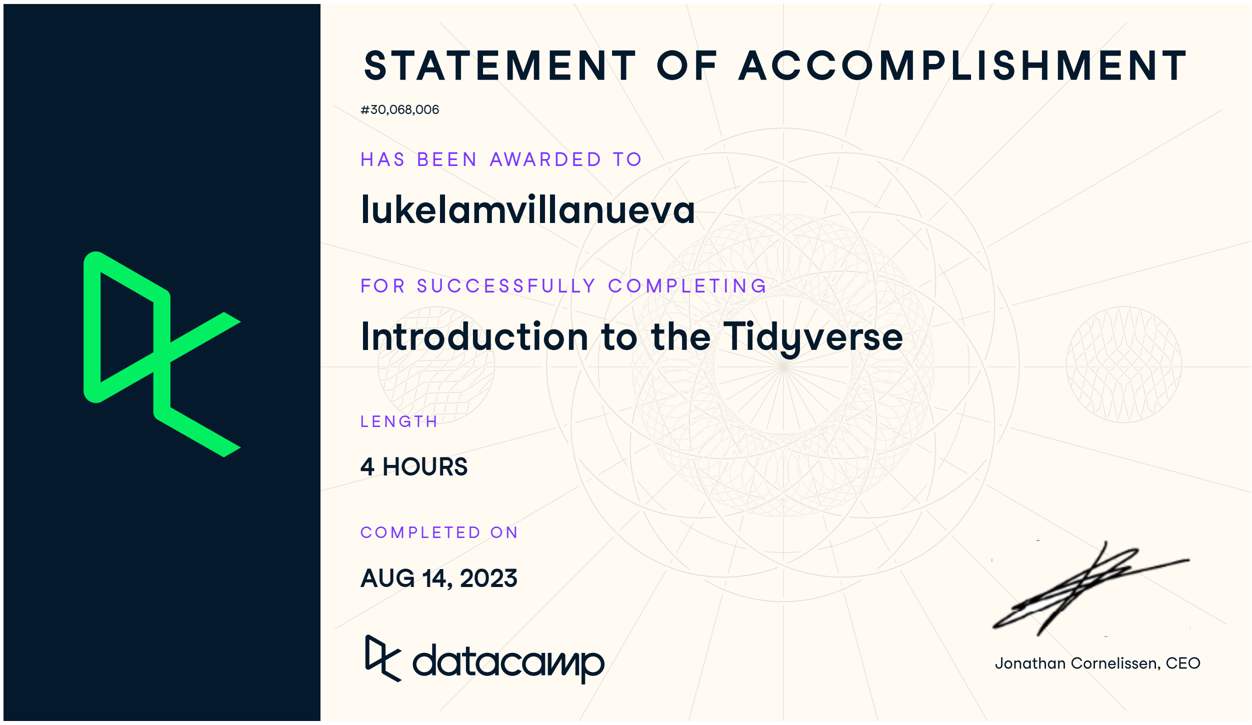
a. Tidy Modeling with R: Chapters 2 through Chapter 4
b. Introduction to Statistical Learning: Chapter 2

DataCamp Homework

- Course: DataCamp Introduction to the Tidyverse:
- https://app.datacamp.com/learn/courses/introduction-to-the-tidyverse

Include certificate of completion here (30 pts):

```
include_graphics("certificate.pdf")
```



STATEMENT OF ACCOMPLISHMENT

#30,068,006

HAS BEEN AWARDED TO

lukelamvillanueva

FOR SUCCESSFULLY COMPLETING

Introduction to the Tidyverse

LENGTH

4 HOURS

COMPLETED ON

AUG 14, 2023

datacamp

Jonathan Cornelissen, CEO

Introduction to Statistical Learning Chapter 2 Exercises.

Exercises begin on page 52 of the textbook.

Exercise 1

For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

a. The sample size n is extremely large, and the number of predictors p is small.
   - answer: Flexible would be better than inflexible because flexible models tend to be more accurate with more observations
b. The number of predictors p is extremely large, and the number of observations n is small.
   - answer: Flexible would be worse than inflexible because inflexible models with highly specific predictor variables will provide more accuracy with a small sample rather than a flexible model
c. The relationship between the predictors and response is highly non-linear.
   - answer: Flexible would be better than inflexible because the MSE for flexible models would generally be lower than assuming a simple inflexible model onto the data. However, this is only to a certain point because the flexible model past a certain flexibility threshold will start rising in MSE again.
d. The variance of the error terms, i.e. $\sigma^2 = Var(\epsilon)$, is extremely high.
   - answer: Flexible would be worse because the true behavior of the data would not be accurate described by flexible models since they tend to overfit the data, especially when the variance is high.

Exercise 2

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

a. We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.
    - answer: Regression. Inference. N = 500. P = profit, employee_number, industry_type
b. We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.
    - answer: Classification. Prediction. N = 20. P = Price_product, Marketing_budget, competition_price, 10 other variables
c. We are interesting in predicting the % change in the US dollar in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the dollar, the % change in the US market, the % change in the British market, and the % change in the German market.
    - answer: Regression. Prediction. N = 52. P = percent_change_dollar, percent_change_us, percent_change_british, percent_change_german

Exercise 5

What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

```
- answer:
Some advantages of a flexible approach vs inflexible would be that there would be much less rigorous tra
A more flexible approach would be preffered if there needed to be less bias on the model or if the predi
```

Exercise 6

Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

```
- answer:
Parametric approach focuses on creating the model's form first (i.e. linear, nonlinear, etc.) and then t
Some advantages to using a parametric approach would be that it would be easier to interpret the model i
```

Exercise 7

The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

(Table Omited. Please check the textbook.)

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.

- answer:

```
#1
sqrt((0)^2+(3)^2+(0)^2)
```

```
## [1] 3
```

```
#2
sqrt((2)^2+(0)^2+(0)^2)
```

```
## [1] 2
```

```
#3
sqrt((0)^2+(1)^2+(3)^2)
```

```
## [1] 3.162278
```

```
#4
sqrt((0)^2+(1)^2+(2)^2)
```

```
## [1] 2.236068
```

```
#5
sqrt((-1)^2+(0)^2+(1)^2)
```

```
## [1] 1.414214
```

```
#6
sqrt((1)^2+(1)^2+(1)^2)
```

```
## [1] 1.732051
```

1. 3
2. 2
3. 3.16
4. 2.236
5. 1.41
6. 1.73

(b) What is our prediction with K = 1? Why?

- answer: For having only 1 neighbor, the closest neighbor would be observation 5 because of the distance being 1.41. So, the prediction for K = 1 is green.

(c) What is our prediction with K = 3? Why?

- answer: For having 3 neighbors, the 3 closest neighbors would be observations 5, 6, and 2, with distance values of 1.41, 1.73, and 2 respectively. Their colors are green, red, red, respectively. So, the prediction for K = 3 is red.

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

- answer: A highly nonlinear boundary implies that there needs to be a more selective approach on which are the most associated neighbors, so a smaller K value would be more applicable.

No problems from Chapter 2 “Applied” Section of the textbook.

Tidymodeling with R chapter 2: A tidyverse Primer

https://www.tmwr.org/tidyverse

Read through chapter 2 of Tidymodeling with R.

In the following code chunk, retype the necessary lines of code from the chapter to reproduce the following output using the function `map2_dbl()` which calculates the log of the ratio between `$mpg` and `$wt` in `mtcars`. (It is okay if your output is rounded.)

```
#> [1] 2.081348 1.988470 2.285193 1.895564 1.693052 1.654643
```

```
map2_dbl(mtcars$mpg, mtcars$wt, function(x,y) log(x/y))
```

```
##  [1] 2.0813481 1.9884698 2.2851934 1.8955636 1.6930521 1.6546433 1.3876939
##  [8] 2.0345622 1.9793581 1.7194388 1.6437270 1.3936383 1.5342983 1.3915714
## [15] 0.6835777 0.6509723 1.0116859 2.6897011 2.9351077 2.9163705 2.1658611
## [22] 1.4823790 1.4872785 1.2422917 1.6081367 2.6467794 2.4972907 3.0003482
## [29] 1.6062784 1.9617713 1.4354846 2.0409400
```

In chapter 2, the authors write the following code. It's not necessary to run, but you are allowed to run it. Warning, the csv file is huge (over 1 million records).

More information about the source of data can be found at: https://data.cityofchicago.org/Transportation/CTA-Ridership-L-Station-Entries-Daily-Totals/5neh-572f

```
library(tidyverse)
library(lubridate)

url <- "https://data.cityofchicago.org/api/views/5neh-572f/rows.csv?accessType=DOWNLOAD&bom=true&format=

all_stations <-
  # Step 1: Read in the data.
  read_csv(url) %>%
  # Step 2: filter columns and rename stationname
  dplyr::select(station = stationname, date, rides) %>%
  # Step 3: Convert the character date field to a date encoding.
  # Also, put the data in units of 1K rides
  mutate(date = mdy(date), rides = rides / 1000) %>%
  # Step 4: Summarize the multiple records using the maximum.
  group_by(date, station) %>%
  summarize(rides = max(rides), .groups = "drop")
```

Explain what the results in `all_stations` tells us. What do the values in the column "rides" signify?

- answer: all_stations have the columns date, station, rides. The dataset tells us the maximum number of rides on each day in each station in units of 1000 rides (i.e. 5 in column "rides" means 5 thousand).

Tidymodeling with R chapter 3

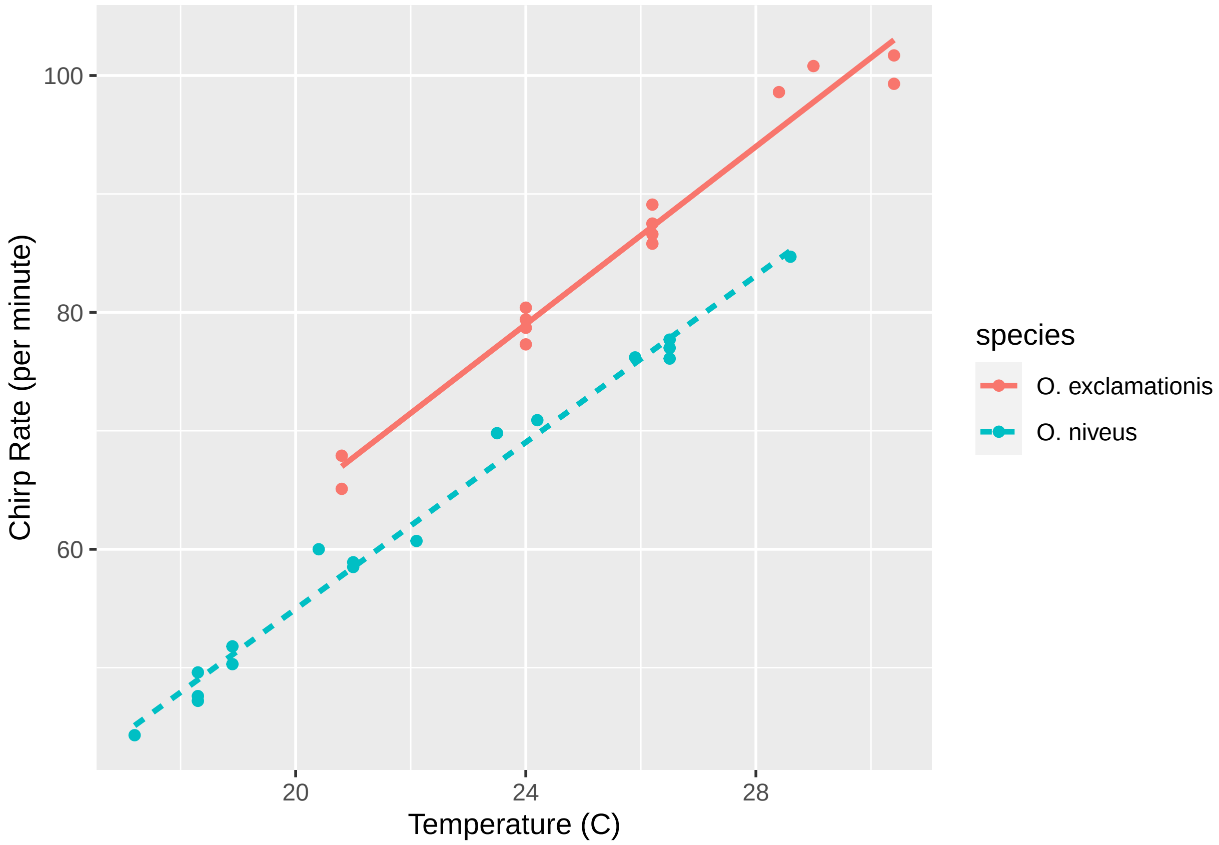
https://www.tmwr.org/base-r

Read through chapter 3 of Tidymodeling with R.

In the following code chunk, retype the necessary lines of code from the chapter to reproduce figure 3.1 showing the relationship between chirp rate and temperature for two different species of crickets.

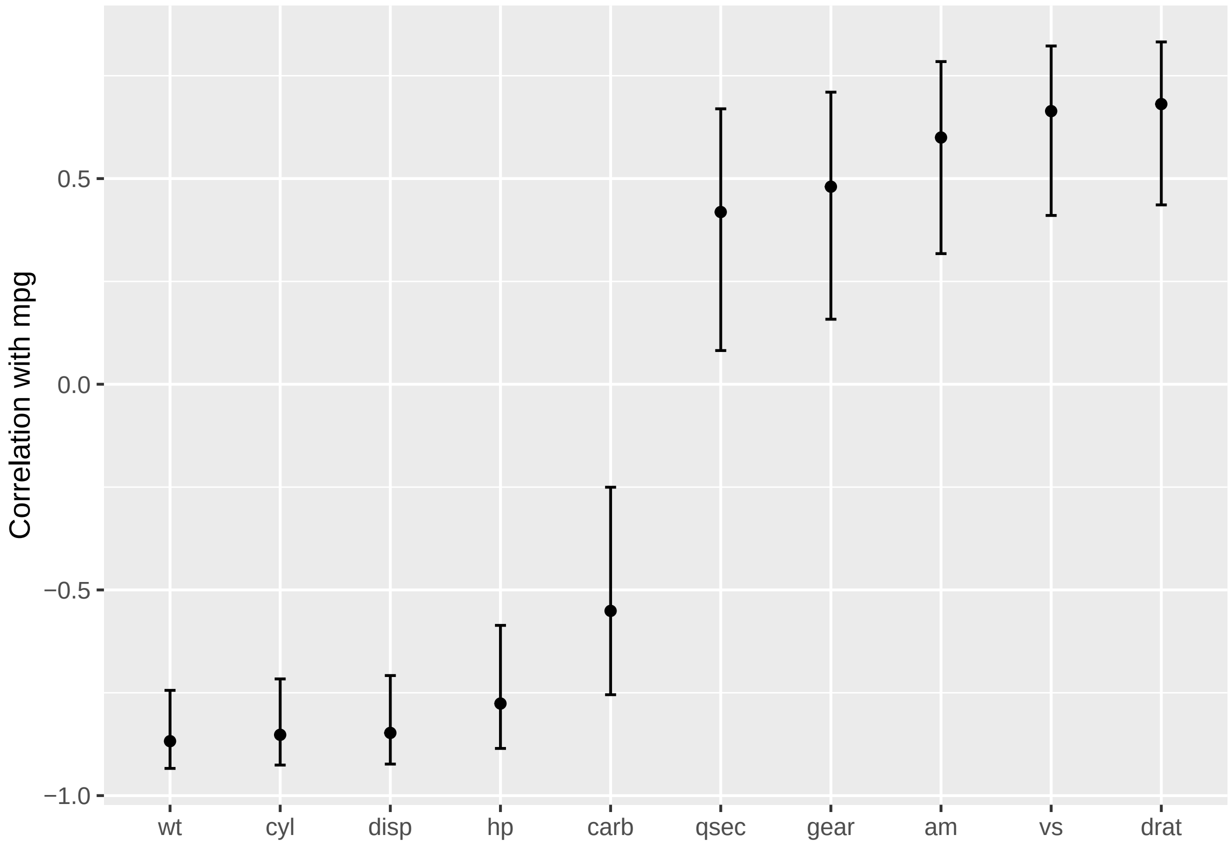
```r
data(crickets, package = "modeldata")
ggplot(crickets, aes(temp,rate, color = species, lty = species)) + geom_point() + geom_smooth(method = l
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



In the following code chunk, retype the necessary lines of code from the chapter to reproduce figure 3.3 showing the correlation between mpg and the other variables in mtcars.

```r
map(mtcars %>% select(-mpg), cor.test, y = mtcars$mpg) %>% map_dfr(tidy, .id = "predictor") %>%
  ggplot(aes(x = fct_reorder(predictor, estimate))) +
  geom_point(aes(y = estimate)) +
  geom_errorbar(aes(ymin = conf.low, ymax = conf.high), width = .1) +
  labs(x = NULL, y = "Correlation with mpg")
```



Question: what does the line `map_dfr(tidy, .id = "predictor")` do? Hint: try only running the first couple lines of code `corr_res %>% map_dfr(tidy, .id = "predictor")` and study the output.

- answer: The line `map_dfr(tidy, .id = "predictor")` takes the cor.test object's values and places the important data in a tidy format, where each observation is a predictor variable in the original data, and the columns are the predictor's hypothesis test results.