GY7702 Assignment

MSc Satellite Data Science at the University of Leicester

Elliot Greatrix

04/11/2020

Introduction

This document was created to meet the requirements of GY7702 R for Data Science at University of Leicester. It was designed and created in R Markdown, a markup language that allows users to create documents that can be formatted to embed a variety of data such as code blocks, images and hyperlinks. When the R Markdown file is compiled, the markup language is hidden and the document is displayed in plain text.

This content was created using R, Rstudio, RMarkdown and GitHub

Materials

The libraries used in this assignment:

```r
library(tidyverse)
library(palmerpenguins)
library(dplyr)
library(knitr)
library(readr)
library(tinytex)
```

For further information regarding the source code, data and libraries used in this assignment, please see the projects GitHub page.

References

Elliot would like to acknowledge that this document includes teaching materials from Dr Stefano De Sabbato for the module GY7702 R for Data Science. Dr Stefanos teaching materials can be found here

R for Data Science by Garrett Grolemund and Hadley Wickham, O'Reilly Media, 2016. See online book

Note to reader

Where a function is mentioned for the **first time**, a brief explanation will be given before the code block. Thereafter, this explanation will be dismissed.

Question 1

A vector of 25 numbers between 1 and 7 are listed below. These values represent answers to a survey question

NA 3 4 4 5 2 4 NA 6 3 5 4 0 5 7 5 NA 5 2 4 NA 3 3 5 NA

- 1 = Completely disagree
- 2 = Disagree
- 3 = Somehow disagree
- 4 = Neither agree or disagree
- 5 = Somehow agree
- 6 = Agree
- 7 = Completely agree
- NA = missing value

Question 1.1

Write the code necessary to check whether all participants to the survey either completely disagree or completely agree, once the missing values are excluded

```r
# First, let's create the vector.
num <- c(NA, 3, 4, 4, 5, 2, 4, NA, 6, 3, 5, 4, 0, 5, 7, 5, NA, 5, 2, 4, NA, 3, 3, 5, NA)

# Then, remove missing values.
num <- na.omit(num)

# And finally check whether all students either completely agree (= 7).
students_agree_complete <- all(num == 7)

# Or, completely disagree (=1).
students_disagree_complete <- all(num == 1)
```

Question 1.2

Write the code necessary to extract the indexes related to the participants in the survey who at least somehow agree or more.

In the vector `num`, `which` students selected values between 5 and 7.

```r
students_agree <- which(num %in% c(5:7))
```

Question 2

Question 2.2

Write the code necessary to create a table showing species, island, bill length and body mass of the 10 Gentoo penguins in the penguins table with the highest body mass

The library `palmerpenguins` was created by Dr. Kristen Gorman and holds 2 data sets. For this assignment, we will work with `penguins`, a table containing 8 measured variables from 344 penguins!

Some of the penguins in `penguins` hold NA values for certain variables. Due to the analysis that we are going to do, these need to be removed

```
pen <- na.omit(penguins)
```

As this task requires multiple steps, it will be more efficient to start a pipe. Lets send the *na omitted* table, `pen` down the pipe.

It is important to note that the functions in the pipe operator do not require the first argument, `pen` because it is being 'fed down' the pipe. For more information on pipes, click here

```
pen %>%
  # First, let's select all the variables that the question asks for
  select(species, island, bill_length_mm, bill_depth_mm, body_mass_g)%>%

  # and select all the 'Gentoo' penguins.
  filter(species == "Gentoo")%>%

  # Order body mass in ascending order
  arrange(-body_mass_g)%>%

  # Return the 10 heaviest Gentoo penguins
  slice_head(n = 10)%>%
kable()
```

species	island	bill_length_mm	bill_depth_mm	body_mass_g
Gentoo	Biscoe	49.2	15.2	6300
Gentoo	Biscoe	59.6	17.0	6050
Gentoo	Biscoe	51.1	16.3	6000
Gentoo	Biscoe	48.8	16.2	6000
Gentoo	Biscoe	45.2	16.4	5950
Gentoo	Biscoe	49.8	15.9	5950
Gentoo	Biscoe	48.4	14.6	5850
Gentoo	Biscoe	49.3	15.7	5850
Gentoo	Biscoe	55.1	16.0	5850
Gentoo	Biscoe	49.5	16.2	5800

Question 2.3

write the code necessary to create a table showing the average bill length per island, ordered by average bill length.

As this question states `per island`, we will use `group_by` to perform an average bill length for each island. We will also see `mean()` to determine average. This is a function built into R.

After the data set is grouped by island, average bill length will be calculated. The result will be appended to `pen` in a column called `avg_bill_length` using the `mutate` function.

```
pen %>%

  # Select the appropriate attributes
  select(species, island, bill_length_mm, bill_depth_mm,)%>%
```

```r
  # Order by bill_length_mm. By default arrange sorts in ascending order.
  arrange(bill_length_mm)%>%

  # Use group_by to execute the average for each island
  group_by(island)%>%

  # Append the average bill_length_mm to the table pen
  mutate(avg_bill_length_mm = mean(bill_length_mm))%>%

  # Display 3 example average_bill_lengths for each island
  slice_head(n = 3)%>%
kable()
```

species	island	bill_length_mm	bill_depth_mm	avg_bill_length_mm
Adelie	Biscoe	34.5	18.1	45.24847
Adelie	Biscoe	35.0	17.9	45.24847
Adelie	Biscoe	35.0	17.9	45.24847
Adelie	Dream	32.1	15.5	44.22195
Adelie	Dream	33.1	16.1	44.22195
Adelie	Dream	34.0	17.1	44.22195
Adelie	Torgersen	33.5	19.0	39.03830
Adelie	Torgersen	34.4	18.4	39.03830
Adelie	Torgersen	34.6	21.1	39.03830

Question 2.4

Write the code necessary to create a table showing the minimum, median and maximum proportion between bill length and bill depth by species.

The *proportion between bill length and bill depth* will be calculated as `bill_length_mm / bill_depth_mm`. This will return a unit-less measure which will be called `bl_bd`

R also contains inbuilt functions for obtaining the minimum, median and maximum values from a vector. So by example after `select()` and `group_by()`, lets take the minimum `bill_length_mm` and minimum `bill_depth_mm` and divide them to get the unit-less measurement `min_bl_bd`.

```r
  pen %>%
  # Select the necessary attributes
  select(species, island, bill_length_mm, bill_depth_mm,)%>%

  # Execute the proportions for each species
  group_by(species)%>%

  # Append the proportions to the table pen, using mutate
  mutate(min_bl_bd = min(bill_length_mm) / min(bill_depth_mm))%>%
```

Then repeat this for the median and maximum values

```r
  mutate(med_bl_bd = median(bill_length_mm) / median(bill_depth_mm))%>%
  mutate(max_bl_bd = max(bill_length_mm) / max(bill_depth_mm))%>%
```

```
  # Drop bill_length_mm and bill_depth_mm
  subset(select = -c(bill_length_mm, bill_depth_mm))%>%

slice_head(n = 3)%>%
kable()
```

species	island	min_bl_bd	med_bl_bd	max_bl_bd
Adelie	Torgersen	2.070968	2.111413	2.139535
Adelie	Torgersen	2.070968	2.111413	2.139535
Adelie	Torgersen	2.070968	2.111413	2.139535
Chinstrap	Dream	2.493902	2.685637	2.788461
Chinstrap	Dream	2.493902	2.685637	2.788461
Chinstrap	Dream	2.493902	2.685637	2.788461
Gentoo	Biscoe	3.122137	3.160000	3.445087
Gentoo	Biscoe	3.122137	3.160000	3.445087
Gentoo	Biscoe	3.122137	3.160000	3.445087

Question 3

Question 3.1

Write the code necessary to load the data from *covid19_cases_20200301_20201017.csv* to a variable named covid_data

The library `readr` contains the function `read_csv()` which, unsurprisingly allows us to read .csv (comma-separated value) files into R.

```
# Reading the COVID-19 .csv from a local directory
covid_data <- read_csv("C:/.../data/covid19_cases_20200301_20201017.csv")
```