GY7702: Coursework 1

James A. Hardwick (179001497)

04/11/2020

Introduction

This document has been created to increase the **reproducibility** of this coursework assignment, written in RMarkdown. To support the reproducibility of the document please refer to the *GitHub data repository* for the commits that document the development of this Coursework 1

Libraries

This coursework use the library **tidyverse**

```
library(tidyverse)
```

Also the library **knitr**

```
library(knitr)
```

Other libraries are also used for specific question for instance in question 2 the library **palmerpenguins** these specific libraries will be referred to within each question

Questions

Question 1:

Question 1 deals with a vector of 25 numbers between 1 and 7, with each value representing answers to survey questions. Some values are missing. #### Question 1.1:

Question 2:

Question 2 looks data from Adélie, Chinstrap, and Gentoo penguins observed on islands in the Palmer Archipelago near Palmer Station, Antarctica. Palmerpenguins library can be found at *Palmerpenguins GitHub Repositry*

Question 2.1 Question 2.1 ask for the library (palmerpenguins) to be installed and loaded

```
library(tidyverse)
library(knitr)
# install.packages("palmerpenguins")
library(palmerpenguins)
```

Question 2.2 Question 2.2 creates a table showing *species, island, bill length and body mass* of the 10 Gentoo penguins in the penguins table with the highest body mass

```
# Starts from the entire palmerpenguins libraries
palmerpenguins::penguins %>%
  # Selects only the necessary columns
```

```
  dplyr::select(species, island, bill_length_mm, body_mass_g
  ) %>%
  # Retain only rows representing the Gentoo species
  dplyr::filter(species == "Gentoo"
  ) %>%
  # Sort by descending body mass in g
  dplyr::arrange(desc(body_mass_g))
```

```
## # A tibble: 124 x 4
##    species island bill_length_mm body_mass_g
##    <fct>   <fct>           <dbl>       <int>
##  1 Gentoo  Biscoe           49.2        6300
##  2 Gentoo  Biscoe           59.6        6050
##  3 Gentoo  Biscoe           51.1        6000
##  4 Gentoo  Biscoe           48.8        6000
##  5 Gentoo  Biscoe           45.2        5950
##  6 Gentoo  Biscoe           49.8        5950
##  7 Gentoo  Biscoe           48.4        5850
##  8 Gentoo  Biscoe           49.3        5850
##  9 Gentoo  Biscoe           55.1        5850
## 10 Gentoo  Biscoe           49.5        5800
## # ... with 114 more rows
```

Question 2.3 Question 2.3 creates a table with *average bill length per island*, ordered by *average bill length*

```
# Starts from the entire palmerpenguins libraries
palmerpenguins::penguins %>%
  # Selects only the necessary columns
  dplyr::select(bill_length_mm, island) %>%
  # Grouped by island
  dplyr::group_by(island) %>%
  # Drops rows containing NAs in the bill_length_mm column
  # otherwise the mean function will return NA
  dplyr::filter(!is.na(bill_length_mm)) %>%
  # Calculates the average of bill_length_mm
  dplyr::summarise(average_bill_length = mean(bill_length_mm)) %>%
  # Ordered by descending average_bill_length
  dplyr::arrange(desc(average_bill_length)) %>%
  # kable improves tibble format
  knitr::kable()
```

island	average_bill_length
Biscoe	45.25749
Dream	44.16774
Torgersen	38.95098

Question 2.4 Question 2.4 creates a table showing the *minimum, median and maximum* proportion between *bill length and bill depth by species*

```
# Starts from the entire palmerpenguins libraries
palmerpenguins::penguins %>%
  # Selects only the necessary columns
  dplyr::select(species, bill_length_mm, bill_depth_mm) %>%
```

```r
# Grouped by species
dplyr::group_by(species) %>%
# Drops rows containing NAs in the bill_length_mm column
# otherwise the mean function will return NA
dplyr::filter(!is.na(bill_length_mm))%>%
# Drops rows containing NAs in the bill_depth_mm column
# otherwise the mean function will return NA
dplyr::filter(!is.na(bill_depth_mm)) %>%
# Calculates the bill length to bill depth ratio
dplyr::summarise(Proportion=
                   (bill_length_mm/bill_depth_mm)) %>%
 # using summariase again the minimum, median and maximum for each species can be calculated
dplyr::summarise(min(Proportion),
                 median(Proportion),
                 max(Proportion)) %>%
# Using the function kable formats the table
knitr::kable()
```

species	min(Proportion)	median(Proportion)	max(Proportion)
Adelie	1.639810	2.136842	2.450000
Chinstrap	2.350516	2.661577	3.258427
Gentoo	2.566474	3.166667	3.612676

Question 3:

Question 3 looks at a topical data set of new and cumulative **COVD19** cases in the UK between March 1st and October 17th 2020. **COVID19** data is sourced from the HM Government Coronavirus in the UK

Question 3.1 Question 3.1 asks for the data covid19 cases to be loaded

```r
#using readr (part of tidyverse)
library(readr)
# reads the .CSV file with the correct directory
# Imports covid19_cases_20200301_20201017.csv and assings to a new variable
#covid_data
covid_data <-readr::read_csv("covid19_cases_20200301_20201017.csv")
```

Question 3.1