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. Vector was defined by the question:

```
# vector survey_responses contains 25 elements
survey_responses <- 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)
```

```

# specify the survey_response variable
survey_responses %>%
  # using na.omit removes NA values from the vector
  na.omit() %>%
  # is. element will return logical output, if 7 or completely agree was answered in the
  # survey it will be returned as TRUE
  is.element(7) %>%
  # is. element will return logical output, if 1 or completely agree was answered in the
  # survey it will be returned as TRUE
  is.element(1)
```

Question 1.1:

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

Question 1.2 Extract the indexes who at least somewhat agreed or more (values 5 to 7)

```
# Function which see if any value between 5 and 7 and return the index of each
# survey participant
which(survey_responses %in% 5:7)
```

```
## [1]  5  9 11 14 15 16 18 24
```

###**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) %>%
  # 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 3.2 This question asked for an area specific table to be generated. Here **Brentwood** (Essex) COVID19 is presented. Brentwood Borough Council *(COVID19 response and information)[https://www.brentwood.gov.uk/index.php?cid=2937]*

```r
# Creates a new table named Brentwood_complete_covid_data from manipulation of
# imported national covid_data
Brentwood_complete_covid_data <- covid_data %>%
  # desired columns (specimen_date, area_name, newCasesBySpecimenDate,
  # cumCasesBySpecimenDate) are selected
  dplyr::select(specimen_date, area_name, newCasesBySpecimenDate,
                cumCasesBySpecimenDate) %>%
  # table then ordered based on specimen_date and area_name
  dplyr::arrange(specimen_date, area_name) %>%
  # fill replace NA values with the values from the previous row
  # default direction is down
  tidyr::fill(newCasesBySpecimenDate, cumCasesBySpecimenDate) %>%
  # replace_na replaces any reamining NA with 0
  tidyr::replace_na() %>%
  # filter data to select on values when the area_name equals Bentwood
  dplyr::filter(area_name == "Brentwood") %>%
  # drops column area_name using select
  dplyr::select(-area_name)%>%
  # slice_head to print a representive number of rows
  dplyr::slice_head(n= 8)
```

Question 3.3

load library lubridate

Initally, library lubdridate is loaded to aid with date format.

```r
library("lubridate")
```

Second part of this question converts day_before into character from following year month and day format and reformats the the table

```r
# Creates a new table Brentwood_day_before from manipulation of
# Brentwood_complete_covid_data
Brentwood_day_before <- Brentwood_complete_covid_data %>%
  # using lubridate, a new column day_before is created
  # day_before is in character data type in the format year month day
  # with specimen_date from Brentwood_complete_covid_data - 1
  dplyr::mutate(day_before = as.character(ymd(specimen_date -1))) %>%
  # specimen_date and cumCasesBySpecimenDate are droped from the tble
  dplyr::select(-specimen_date, -cumCasesBySpecimenDate) %>%
  #newCasesBySpecimenDate are renamed to newCases_day_before
  dplyr::rename(newCases_day_before = newCasesBySpecimenDate) %>%
  # slice_head to print a representive number of rows
  dplyr::slice_head(n= 8)
```

Initially, when I tried to join the tables *Brentwood_complete_covid_data* with *Brentwood_day_before* this didn't work as each column in the join was in a different data type and . So I tried: 1. converting column specimen_date in *Brentwood_complete_covid_data* + this resulted in an error when attempted to convert the date to character as specified by the question in the table *Brentwood_day_before* 2. Making a new table *Brentwood_complete_covid_data_converted* before the table join + This was successful in enabling the table join + However it is a messy way round of executing this column data type conversion

```
# New table Brentwood_complete_covid_data_converted created
Brentwood_complete_covid_data_converted <- Brentwood_complete_covid_data %>%
  # using mutate with the function across allowed the date value to be
  # converted to character without the creation of a new column
  dplyr::mutate(across(where(is.Date), as.character))
```

Lastly, Brentwood_day_before and Brentwood_complete_covid_data are joined using a left join. With day_before being equal to specimen_date both in character data type. Then a new column is added showing the daily percentage change of cases. Stored in a table called *Brentwood_covid_development*

```
# Left table_join where specimen_date and day_before are equal
Brentwood_covid_development <- dplyr::left_join(Brentwood_complete_covid_data_converted,
                                                Brentwood_day_before,
                                                by =c("specimen_date" = "day_before")) %>%
  # using mutate new column was created showing the percentage day on day change
  dplyr::mutate(percentage_new_cases =
                  ((newCasesBySpecimenDate / newCases_day_before)*100))
```

Question 3.4 495 COVID-19 cases were recorded in the Brentwood area between 2020-03-02 to 2020-10-09. Development of cases was not linear instead three distinct phases can be observed. Phase-1 began between the period 2020-03-03 to 2020-03-16, only two cases where recorded. Between 2020-03-17 to 2020-05-14 cases increased rapidly with percentage of new cases of day before peaking at 800%. During phase-3, new cases began to slow and stabilised overall between 2020-05-15 and 2020-08-19. Daily percentage change fell over time (e.g. 200% to 0%). Within this general trend there was some short-scale variability (e.g. 2020-05-17 to 2020-05-21 of 0% to 25% to 133% before returning to 0%). Stabilisation of cases, during phase-2, was short lived with phase-3 mirroring the initial phase-1 development of new cases starting from 2020-08-20. During Phase-3 percentage of new cases was rapidly rising averaging 100.71% daily increase, a higher magnitude than phase-1 (i.e 84%).

###**Question 4**

Importing population per local authority and COVID19 case data

```
# import lad19_population data and assigned to a new variable LAD_pop
LAD_pop<- readr::read_csv("lad19_population.csv")
# import COVID19 case data and assigned to a new variable LAD_covid_cases
LAD_covid_cases <- readr::read_csv("covid19_cases_20200301_20201017.csv")
```

Carrying out table join for LAD_pop and LAD_covid_cases

```
# need to rename LAD_pop column lad19_area_name to area_name to match
# LAD_covid_cases table
covid_cases_lad_pop <- dplyr::rename(LAD_pop,area_name = lad19_area_name) %>%
# table join between LAD_pop and LAD_covid_cases
dplyr::left_join(LAD_covid_cases, LAD_pop, by = "area_name")
```

####Description and Interpretation