GY7702 Assignment 1

Ben Coombs

14/11/2020

R for Data Science - GY7702 Assignment 1

Loading libraries

Throughout this assignment, I will be using the libraries tidyverse and knitr. Therefore it is a good idea to load them straight away.

```
library(tidyverse)
library(knitr)
```

Question 1

Q1.1

```
#Create the vector of 25 numbers listed on the question paper
nums <- 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)
#Create a new vector of the same numbers, this time with missing values (NA
#values omitted)
nums_new <- nums[!is.na(nums)]
#Check if all responses are strongly agree or strongly disagree
all(nums_new %>% is.element(c(1,7)))
```

```
## [1] FALSE
```

A return of FALSE indicates that there were participants in the survey that did not either completely agree or completely disagree.

Q1.2

```
#Return the positions in the vector of elements that are participants responding
#somehow agree or stronger (5 or greater).
which(nums_new >= 5)
```

```
## [1]  4  7  9 12 13 14 15 20
```

Question 2

Q2.1

```
# Load the installed library "palmerpenguins"
library(palmerpenguins)
```

Q2.2

```
#Create a table to show species, island, bill length and body mass of the 10
#Gentoo penguins with the highest body mass
gentoo_penguins <- penguins %>%
  #Only want the table to show species, island, bill length and body mass
  select(species, island, bill_length_mm, body_mass_g) %>%
  #Only display rows of Gentoo penguins
  filter(species == "Gentoo") %>%
  #Arrange the rows by body mass largest to smallest
  arrange(-body_mass_g) %>%
  #Take the first 10 values, i.e. the 10 Gentoo penguins with the largest body
  #mass
  slice_head(n = 10)

# Display the table
knitr::kable(gentoo_penguins)
```

species	island	bill_length_mm	body_mass_g
Gentoo	Biscoe	49.2	6300
Gentoo	Biscoe	59.6	6050
Gentoo	Biscoe	51.1	6000
Gentoo	Biscoe	48.8	6000
Gentoo	Biscoe	45.2	5950
Gentoo	Biscoe	49.8	5950
Gentoo	Biscoe	48.4	5850
Gentoo	Biscoe	49.3	5850
Gentoo	Biscoe	55.1	5850
Gentoo	Biscoe	49.5	5800

Q2.3

```
#Create a table of the average bill length per island ordered by average bill
#length
island_bill_length <- penguins %>%
  #Only considering island and bill length
  select(island, bill_length_mm) %>%
  #Remove NA values to correctly summarise
  filter(!is.na(bill_length_mm)) %>%
  #Group rows by island
```

```
  group_by(island) %>%
  #Summarise the islands by the average bill length
  summarise(
    avg_bill_length_mm = mean(bill_length_mm)
  ) %>%
  #Arrange the table by highest to lowest average bill length
  arrange(-avg_bill_length_mm)

knitr::kable(island_bill_length)
```

island	avg_bill_length_mm
Biscoe	45.25749
Dream	44.16774
Torgersen	38.95098

Q2.4

```
#Create a table to show the maximum, median and minimum proportion between bill
#length and bill depth by species
bill_ratios <- penguins %>%
  #Only considering species, bill length and bill depth
  select(species, bill_length_mm, bill_depth_mm) %>%
  #Remove NA values to correctly summarise
  filter(!is.na(bill_length_mm), !is.na(bill_depth_mm)) %>%
  #Calculate and create a column for the bill ratio of each penguin
  mutate(
    bill_ratio = bill_length_mm / bill_depth_mm
  ) %>%
  #Group rows by species
  group_by(species) %>%
  #Summarise each species with their minimum, median, and maximum bill ratios
  summarise(
    min_proportion = min(bill_ratio),
    median_proportion = median(bill_ratio),
    max_proportion = max(bill_ratio)
  )

knitr::kable(bill_ratios)
```

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

Q3.1

```
#Load the csv data to a variable
covid_data <- read_csv("covid19_cases_20200301_20201017.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   specimen_date = col_date(format = ""),
##   area_name = col_character(),
##   newCasesBySpecimenDate = col_double(),
##   cumCasesBySpecimenDate = col_double()
## )
```

Q 3.2

```
Ealing_complete_covid_data <- covid_data %>%
  #Create complete table with a row for each day and area
  tidyr::complete(specimen_date, area_name) %>%
  #Arrange the table by area name (A-Z) and specimen date (oldest to newest)
  arrange(area_name, specimen_date) %>%
  #Group by area name to avoid the risk of filling NA values in a later step
  #with values from a different area
  group_by(area_name) %>%
  #Fill NA values of new and cumulative cases with the value of the previous
  #date for that area, if possible
  tidyr::fill(newCasesBySpecimenDate, cumCasesBySpecimenDate) %>%
  #Replace any remaining NA values with 0
  tidyr::replace_na(list(newCasesBySpecimenDate = 0,
                         cumCasesBySpecimenDate = 0)
                    ) %>%
  #Subset the assigned area
  filter(area_name == "Ealing") %>%
  #Ungroup to drop area_name column
  ungroup() %>%
  #Drop area_name column
  select(-area_name)
```

Q 3.3

```
#Copy of Ealing_covid_data as a new variable
Ealing_day_before <- Ealing_complete_covid_data

Ealing_covid_development <- Ealing_day_before %>%
  #Create new column reporting the day after the date reported in Specimen_date
  #column
  mutate(
```

```
    day_to_match = specimen_date + 1
  ) %>%
  #Drop the columns specimen_date and cumCasesBySpecimenDate
  select(-specimen_date, -cumCasesBySpecimenDate) %>%
  #Rename the newCasesBySpecimen date to newCases_day_before
  rename(newCases_day_before = newCasesBySpecimenDate) %>%
  #Join this table with Ealing_complete_covid_data where day_to_match is equal
  #to specimen_date
  full_join(
    Ealing_complete_covid_data,
    by = c("day_to_match" = "specimen_date")
  ) %>%
  #Create new calculated column of new cases as a percentage of the number of
  #cases of the previous day
  mutate(
    As_percentage_of_day_before = (
      newCasesBySpecimenDate/newCases_day_before
      )*100
  )

#Display the first few lines
Ealing_covid_development %>%
slice_head(n = 10) %>%
kable()
```

newCases_day_before	day_to_match	newCasesBySpecimenDate	cumCasesBySpecimenDate	As_percentage_of_day_before
0	2020-03-02	1	2	Inf
1	2020-03-03	1	3	100
1	2020-03-04	1	4	100
1	2020-03-05	1	5	100
1	2020-03-06	0	5	0
0	2020-03-07	1	6	Inf
1	2020-03-08	1	7	100
1	2020-03-09	0	7	0
0	2020-03-10	5	12	Inf
5	2020-03-11	10	22	200

Days reporting 0 new cases lead to "Inf" being the value for As_percentage_of_day_before of the following day as R is being asked to divide by zero. Instead, I will replace these values with NA.

```
#Replace "Inf" values in As_percentage_of_day_before with NA
Ealing_covid_development$As_percentage_of_day_before[
  is.infinite(Ealing_covid_development$As_percentage_of_day_before)
  ] <- NA

#Display the first few lines
Ealing_covid_development %>%
slice_head(n = 10) %>%
kable()
```

newCases_day_before	day_to_match	newCasesBySpecimenDate	CumCasesBySpecimenDate	percentage_of_day_before
0	2020-03-02	1	2	NA
1	2020-03-03	1	3	100
1	2020-03-04	1	4	100
1	2020-03-05	1	5	100
1	2020-03-06	0	5	0
0	2020-03-07	1	6	NA
1	2020-03-08	1	7	100
1	2020-03-09	0	7	0
0	2020-03-10	5	12	NA
5	2020-03-11	10	22	200