STA130H1S – Fall 2022

Problem Set 2

() and STA130 Professors

Instructions

Complete the exercises in this .Rmd file and submit your .Rmd and .pdf output through Quercus on September 29 by 5:00 p.m. ET.

Part 1: More Olympics Data

The code below loads the `VGAMdata` package (so you can access the data sets it contains) and the `tidyverse` package (so you can use the functions it contains) and glimpses the `oly12` data set, which you will use for this question. **Do not use the `olympics` data set from class to answer the prompts in this question**.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(VGAMdata) # install.packages("VGAMdata")
```

```
## Loading required package: VGAM

## Loading required package: stats4

## Loading required package: splines
```

```
names(oly12) # convenient function to quickly glance at data set column names
```

```
##  [1] "Name"    "Country" "Age"     "Height"  "Weight"  "Sex"     "DOB"
##  [8] "PlaceOB" "Gold"    "Silver"  "Bronze"  "Total"   "Sport"   "Event"
```

```
glimpse(oly12)
```

```
## Rows: 10,384
## Columns: 14
## $ Name    <fct> Lamusi A, A G Kruger, Jamale Aarrass, Abdelhak Aatakni, Maria ~
## $ Country <fct> "People's Republic of China", "United States of America", "Fra~
## $ Age     <int> 23, 33, 30, 24, 26, 27, 30, 23, 27, 19, 37, 28, 28, 28, 22, 19~
## $ Height  <dbl> 1.70, 1.93, 1.87, NA, 1.78, 1.82, 1.82, 1.87, 1.90, 1.70, NA, ~
## $ Weight  <int> 60, 125, 76, NA, 85, 80, 73, 75, 80, NA, NA, NA, 60, 64, 62, N~
## $ Sex     <fct> M, M, M, M, F, M, F, M, M, M, M, M, F, F, M, F, M, M, M, M, F,~
## $ DOB     <date> 1989-02-06, NA, NA, 1988-09-02, NA, 1984-06-09, NA, 1989-03-0~
## $ PlaceOB <fct> "NEIMONGGOL (CHN)", "Sheldon (USA)", "BEZONS (FRA)", "AIN SEBA~
## $ Gold    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ Silver  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ Bronze  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ Total   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ Sport   <fct> "Judo", "Athletics", "Athletics", "Boxing", "Athletics", "Hand~
## $ Event   <fct> "Men's -60kg", "Men's Hammer Throw", "Men's 1500m", "Men's Lig~
```

Question 1: Practice with `filter()`

(a) In this week's class, we looked at data for each country which participated in the 2012 Olympics (e.g. size of each country's Olympic team, number of medals won, etc.), and there was one observation (i.e. one row) for each participating country. What does each row in the `oly12` dataset represent?

REPLACE THIS TEXT WITH YOUR ANSWER

Hint: Type `?oly12` or `help(oly12)` in the console (on the bottom left corner) to view the help file for the `oly12` dataset in the Help tab (on the bottom right corner) of RStudio); or, just search for "oly12" in the Help tab.

(b) Determine the number of athletes who represented Canada (`Canada`) or the United States (`United States of America`) in the 2012 Olympic Games.

```
# Type your code here
```

Hint: Apply the `filter()` function to the `Country` column of the `oly12` dataset

(c) Determine the number of athletes who competed in classical gymnastics (`Gymnastics - Artistic` and `Gymnastics - Rhythmic`) or classical pool sports (`Diving` and `Swimming`).

```
# Type your code here
```

Hint: You can see all the possible values for the `Sport` variable with `levels(oly12$Sport)`, and count the number of possible levels with `nlevels(oly12$Sport)`.

(d) Determine the number of athletes who competed in ANY gymnastic (`Gymnastics - Artistic`, `Gymnastics - Rhythmic`, `Trampoline`) or ANY pool sports (`Diving`, `Swimming`, `Synchronised Swimming`, and `Water Polo`)

Hint: As indicated on stackoverflow, the `%in%` comparision operator could be useful here with `allGymnastics <- c("Gymnastics - Artistic", "Gymnastics - Rhythmic", "Trampoline")` and `allWaterPool <- c("Diving", "Swimming", "Synchronised Swimming", "Water Polo")` and `filter(Sport %in% allGymnastics | Sport %in% allWaterPool)`.

(e) Create the data subset `oly12_FemaleArtisticRhythmicGymnasts` which contains all female olympic athletes who competed in artistic gymnastics or rhythmic gymnastics.

```
# Type your code here
```

Hint: `names(oly12)` shows all the column names of the data set.

(f) Use `oly12_FemaleArtisticRhythmicGymnasts` and `ggplot2` to compare the age distribution of female olympic athletes competing in artistic gymnastics to the age distribution of female olympic athletes competing in rhythmic gymnastics using both boxplots and histrograms.

```
# Type your code here
```

Hint: don't forget `aes()` and to use `+` rather than `%>%`.

(g) Answer the following questions based on the plots you created in (d).

- Are the age distributions of female rhythmic gymnasts and female artistic gymnasts symmetrical or skewed?

REPLACE THIS TEXT WITH YOUR ANSWER

- How do the medians, 25th percentiles, and 75th percentiles for ages of female rhythmic gymnasts and female artistic gymnasts compare?

REPLACE THIS TEXT WITH YOUR ANSWER

- Based only on the histogram and boxplots, predict whether the standard deviation of the ages is similar or different. Justify your answer in 1-2 sentences.

REPLACE THIS TEXT WITH YOUR ANSWER

Question 2: Practice with `summarise()`, `group_by()`, and `mutate()`

(a) Create a summary table of `oly12_FemaleArtisticRhythmicGymnasts` reporting the minimum (`min`), maximum (`min`), `mean`, `median`, and standard deviation (`sd`) of ages for female rhythmic gymnasts and female artistic gymnasts. Were you correct in your guess about the standard deviation in part (g) of the last question?

```
# Type your code here
```

REPLACE THIS TEXT WITH YOUR ANSWER

(b) Create a new variable called `total_medals` and create a new tibble called `oly12_OneMedalClub` that contains athletes who won exactly one medal at the 2012 olympics.

```
# Type your code here
```

(c) Uncomment the code below and run the glimpse of the data created in part (c).

```
# glimpse(oly12_OneMedalClub)
```

Question 3: Practice with `select()`, `arrange()`, `desc()`, and `filter()`

(b) Find the `Name` and `Age` of the 6 oldest athletes who competed in the 2012 Olympics.

```
# Type your code here
```

(b) Find the `Name`, `Age` and `Sport` of the 6 youngest female athletes who competed in the 2012 Olympics.

```
# Type your code here
```

(c) Find the `Name`, `Age`, `Sport`, and `Event` for the 6 youngest and 6 oldest competitors who won gold medals at the 2012 olympics.

```
# Type your code here
```

Question 4: The Data Consultant

You have just been hired by a consultancy company. Congratulations! They are doing a report on each Olympics for the past 10 years. Given your recent experience in STA130, you ask to be responsible for the 2012 summary. Write a short report to your boss on information that can be gleaned about the ages of the athletes across sports. As it turns out, you happen to know that your new boss' favourite sports are badminton and weightlifting, so addressing these sports specifically might be an easy way to capture their attention; but, other features athletes' ages which can be learned from your plots and tables will of course be appreciated, too. The more interesting the better!

Question Constraints This is a quick report for your boss, so use full sentences and communicate in a clear and professional manner. Grammar isn't the main focus of the assessment, but don't use slang or emojis.

- ***Avoid Analysis Paralysis***: this is envisioned as a 30 minute exercise, so you don't have time to exhaustively explore every aspect of the data set.
- ***Avoid Writer's Block***: this is envisioned as a 200-400 word exercise, so quickly find something you can communicate and write about.

(a) Watch this 7-minute video introduction to hedging.

Hedging is helpful whenever you can't say something is 100% one way or another, as is often the case. In statistics, hedging should always be used with respect to the limitations of data and the strength and generalizability of the conclusions.

(b) Provide a small introduction of one or two sentences to draw your reader in and then explain what you'll be discussing. Be definitive about what your data is, and use *hedging* to caveat the limitations of the data.

(c) Provide one or two clearly titled and labeled figures addressing interesting features of athletes' ages.

(d) Provide one or two clearly labeled summary tables addressing interesting features of athletes' ages.

(e) Watch this 8-minute video introduction to plagiarism.

You don't need to cite any outside references for your report to your boss, but you will be referring to your own created figures and tables. We'll use this as an excuse to get started early thinking about this important topic, and also use it as an exercise to start getting into the right referencing habits. It's easy and natural and makes your writing better (not mention avoids potential serious academic integrity violations. . .)

(f) Describe the interesting features of athletes' ages that you've found, referencing the figures and summary tables created in (c) and (d) just above. Use at least two of the vocabulary words listed below; but, your boss isn't a statistician, so make sure to clearly define and explain the vocabulary you use.

(f) Finish with a conclusion to remind your boss of the key take home points from your summary about the athletes' ages. Be definitive about what your findings are, but use *hedging* to caveat the limitations of the conclusion more generally.

Vocabulary

- Cleaning data
- Tidy data
- Handling missing values (NAs)
- Removing a column
- Extracting a subset of variables
- Filtering a tibble based on a condition (e.g. based on the values in one or more of the variables/columns)
- Sorting data based on the values of a variable
- Defining new variables
- Renaming the variables
- Producing new data frames
- Grouping categories
- Creating summary tables

You may also find these vocabulary words from last week useful with your writing this week

- location/center (mean, median, mode) and scale/spread (range, IQR, var, sd)
  - note: interpreting center and spread relative to each other can be helpful
- shape (symmetric, left-skewed, right-skewed, unimodal, bimodal, multimodal, uniform)
- outliers/extreme values
  - note: this can be related to the tails of a distribution (heavy-tailed, thin-tailed)
- frequency (most, least, pattern tendencies)

Part 2: OPTIONAL but Recommended

You may complete these questions for practice if you wish. ***You are not required to complete these questions as they ARE NOT included as part of your mark.***

```r
library(tidyverse) # Load the tidyverse package so it is available to use
books <- read.csv("amazonbooks.csv")
```

Question 5: Amazon Books

The code below reads in data about books sold on Amazon.
- Note that the height (`Height`), width (`Width`), and thickness (`Thick`) of books in this data frame are measured in inches.

(a) What is the name of the book(s) with the smallest number of pages in this sample of books, and how many pages does it have?

```r
# Type your code here
```

(b) Create a summary table which reports the total number of books written by each author and the mean and variance of the number of pages per book for each author, for the books represented in this sample of books.

```r
# Type your code here
```

(c) Modify your code from (b) so to create a new summary table which contains only information for authors who wrote more than 2 books, and sort them in decreasing order of number of books written.

```r
# Type your code here
```

Part 3: OPTIONAL for Additional Practice

You may complete these questions for practice if you wish. ***You are not required to complete these questions as they ARE NOT included as part of your mark.***

Question 6: Titanic Data

At the time it departed from England in April 1912, the RMS Titanic was the largest ship in the world. In the night of April 14th to April 15th, the Titanic struck an iceberg and sank approximately 600km south of Newfoundland (a province in eastern Canada). Many people perished in this accident. The code below loads data about the passengers who were on board the Titanic at the time of the accident.

```
titanic <- read_csv("titanic.csv")
```

```
## Rows: 2208 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (12): Name, Survived, Boarded, Class, MWC, Adut_or_Chld, Sex, Ticket_No,...
## dbl  (2): Age, Paid
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(titanic)
```

```
## Rows: 2,208
## Columns: 14
## $ Name         <chr> "ABBING, Mr Anthony", "ABBOTT, Mr Ernest Owen", "ABBOTT, ~
## $ Survived     <chr> "Dead", "Dead", "Dead", "Dead", "Alive", "Alive", "Alive"~
## $ Boarded      <chr> "Southampton", "Southampton", "Southampton", "Southampton~
## $ Class        <chr> "3", "Crew", "3", "3", "3", "3", "3", "2", "2", "3", "3",~
## $ MWC          <chr> "Man", "Man", "Child", "Man", "Woman", "Woman", "Man", "M~
## $ Age          <dbl> 42.00, 21.00, 14.00, 16.00, 39.00, 16.00, 25.00, 30.00, 2~
## $ Adut_or_Chld <chr> "Adult", "Adult", "Child", "Adult", "Adult", "Adult", "Ad~
## $ Sex          <chr> "Male", "Male", "Male", "Male", "Female", "Female", "Male~
## $ Paid         <dbl> 7.550000, NA, 20.250000, 20.250000, 20.250000, 7.650000, ~
## $ Ticket_No    <chr> "5547", NA, "CA2673", "CA2673", "CA2673", "348125", "3481~
## $ Boat_or_Body <chr> NA, NA, NA, "[190]", "A", "16", "A", NA, "10", "15", "C",~
## $ Job          <chr> "Blacksmith", "Lounge Pantry Steward", "Scholar", "Jewell~
## $ Class_Dept   <chr> "3rd Class Passenger", "Victualling Crew", "3rd Class Pas~
## $ Class_Full   <chr> "3", "V", "3", "3", "3", "3", "3", "2", "2", "3", "3", "E~
```

(a) Often, before you start working with a dataset you need to clean it.

- The variable `Adut_or_Chld` indicates which passengers were adults and which were children. Use the `rename()` function to change the name of this variable to `Adult_or_Child`. The variable `MWC` records whether the passenger was a man, woman or child. Use the `rename()` function to change the name of this variable to `Man_Woman_or_Child` to make this clear.

```
# Type your code here
```

Hint: Unless the transformed tibble is saved into a new object or overwrites the original tibble, like `oly12 <- oly12 %>% rename(Place_of_birth = PlaceOB)`, the changes won't be permanent.

- Since many of their values are missing or unclear, modify the `titanic` data frame by removing the following variables: `Ticket_No`, `Boat_or_Body`, `CLass_Dept`, `Class_Full`.

```
# Type your code here
```

(b) Create a summary table reporting the number of passengers on the Titanic (n), the number of passengers who survied (n_surv), and the proportion of passengers who survived (prop_surv).

```
# Type your code here
```

(c) Calculate the proportion of deaths for the following groups of passengers.

- For men, women, and children:

```
# Type your code here
```

- For passengers aged between 25-40 years of age:

```
# Type your code here
```

- For men, women, and children among the passengers who paid more than 50 British pounds for their tickets:

```
# Type your code here
```

- Write several sentences interpreting the summary tables created for the three groups above.

REPLACE THIS TEXT WITH YOUR ANSWER

(d) What was the most common job among passengers of the Titanic? Write 1-2 sentences explaining your answer.

```
# Type your code here
```

REPLACE THIS TEXT WITH YOUR ANSWER

(e) Plot the age distribution for passengers with the job "General Labourer", and describe this distribution in 1-2 sentences.

```
# Type your code here
```

REPLACE THIS TEXT WITH YOUR ANSWER

(f) Were any of the general labourers on the titanic women? If so, how many?

```
# Type your code here
```

(g) What are the names of the passengers with the top 4 most expensive tickets? Did these passengers survive the accident?

```
# Type your code here
```

(h) In this question, you will compare the distribution of ticket prices for survivors and non-survivors of the Titanic using both visualizations and summary tables.

- Construct two histograms to visualize the distribution of ticket prices for survivors and non-survivors (i.e. one histogram for survivors and one for non-survivors). Write 2-3 sentences comparing the two distributions based on these plots.

```
# Type your code here
```

- Construct a pair of boxplots (in the same figure) to visualize the distribution of ticket prices for survivors and non-survivors. Write 2-3 sentences comparing the two distributions based on these plots.

```
# Type your code here
```

- Construct a summary table with the minimum, first quartile, median, mean, third quartile, and maximum ticket price for survivors and non-survivors.

```
#### Example code to demo quantile() function and is.na ####
x <- c(1,2,3,4,5,6,NA,10)
quantile(x, probs = 0.25, na.rm=TRUE); # Calculate the first quartile (25% quantile), and tell R to exc
```

```
## 25%
## 2.5
```

```
quantile(x, probs = 0.75, na.rm=TRUE); # Calculate the third quartile (75% quantile), and tell R to exc
```

```
## 75%
## 5.5
```

```
# If there are missing values in the vector you're working with (or in one of the columns of a tibble),
mean(x)
```

```
## [1] NA
```

```
mean(x, na.rm=TRUE)
```

```
## [1] 4.428571
```

```
median(x)
```

```
## [1] NA
```

```
median(x, na.rm=TRUE)
```

```
## [1] 4
```

- Write 2-3 sentences comparing the two distributions based on this summary table.

REPLACE THIS TEXT WITH YOUR ANSWER

- Comment on the strengths and weaknesses of each of the visualizations and summary table constructed above.

REPLACE THIS TEXT WITH YOUR ANSWER