STA130H1S – Winter 2021

Week 3 Practice Problems - Sample Answers

N. Moon and S. Caetano Yixing Xu

Instructions

How do I hand in these problems for the January 28th deadline ?

Your complete .Rmd file that you create for these practice problems AND the resulting pdf (i.e., the one you 'Knit to PDF' from your .Rmd file) must be uploaded into a Quercus assignment (link: https://q.utoronto.ca/courses/206597/assignments/541106) by 11:59AM ET, on Thursday, January 28th. Late problem sets are not accepted.

Problem set grading

There are two parts to your problem set. One is largely R-based with short written answers and the other is more focused on writing. We recommend you use a word processing software like Microsoft Word to check for grammar errors in your written work. Note: there can be issues copying from Word to R Markdown so it may be easier to write in this file first and then copy the text to Word. Then you can make any changes flagged in Word directly in this file.

Part 1

[Question 1] The code below loads the `VGAMdata` package (so you can access the datasets it contains) and the `tidyverse` package (so you can use the functions it contains) and glimpses the `oly12` dataset, which you will use for this question

```
library(tidyverse)
library(VGAMdata)
glimpse(oly12)
```

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

(a) In this week's slides/videos, 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? Hint: Type `?oly12` in the console (bottom left) to view the help file for the `oly12` dataset (it will appear in the Help tab in the bottom right corner of RStudio)

Each row represents an athlete participating in the Olympics, and the columns are their descriptions/attributes.

(b) Use the `oly12` dataset to determine the number of athletes who represented Canada in the 2012 Olympic Games. Note: there is more than one way to do this, but you need to use the `oly12` dataset for this question, not the dataset from the slides/videos.

```
oly12 %>% group_by(Country) %>% summarise(n = n())
```

```
## # A tibble: 205 x 2
##    Country                 n
##  * <fct>               <int>
##  1 Afghanistan             6
##  2 Albania                10
##  3 Algeria                37
##  4 American Samoa          5
##  5 Andorra                 4
##  6 Angola                 33
##  7 Antigua and Barbuda     4
##  8 Argentina             137
##  9 Armenia                25
## 10 Aruba                   4
## # ... with 195 more rows
```

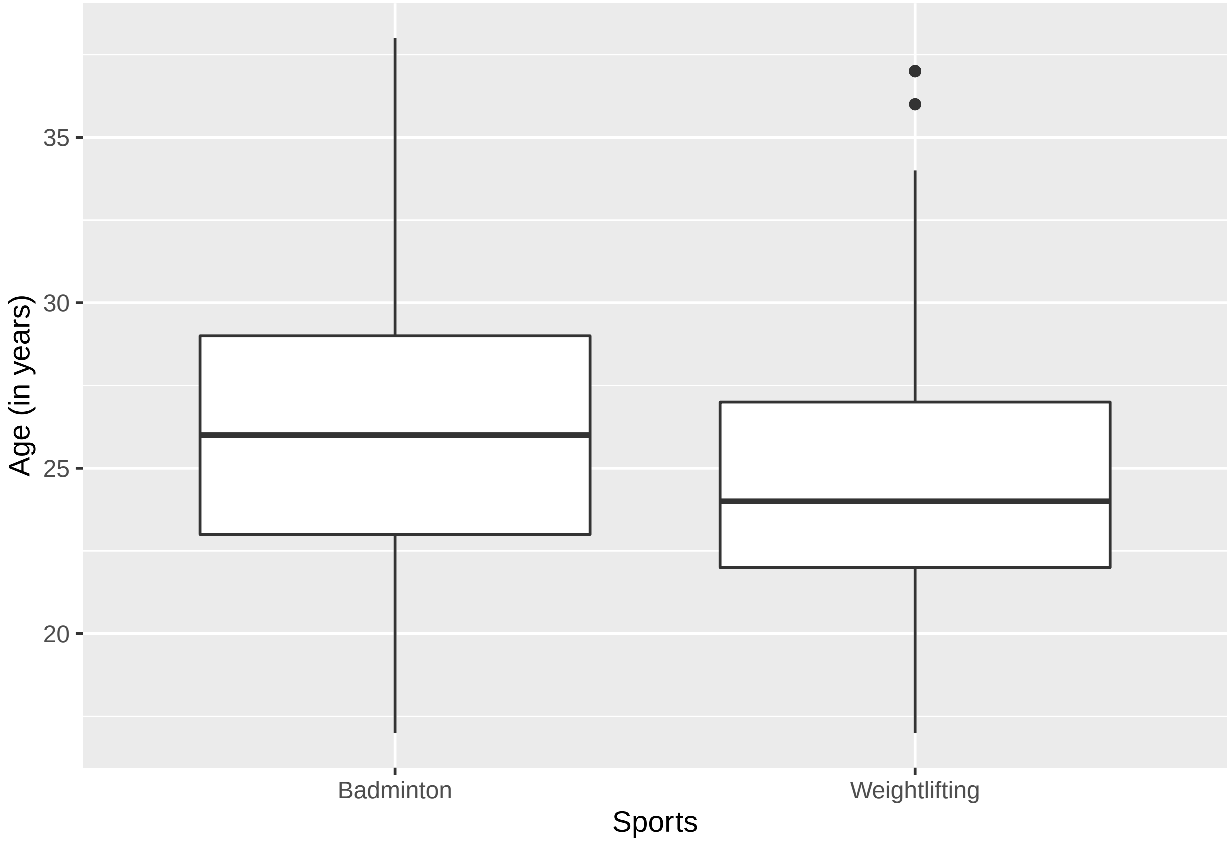
274 Canadian athletes participated in the 2012 Olympics.

(c) Create a new dataframe called `oly12_selectedSports` which contains only data for athletes who competed in Weightlifting or Badminton (look at values of the `Sport` variable).

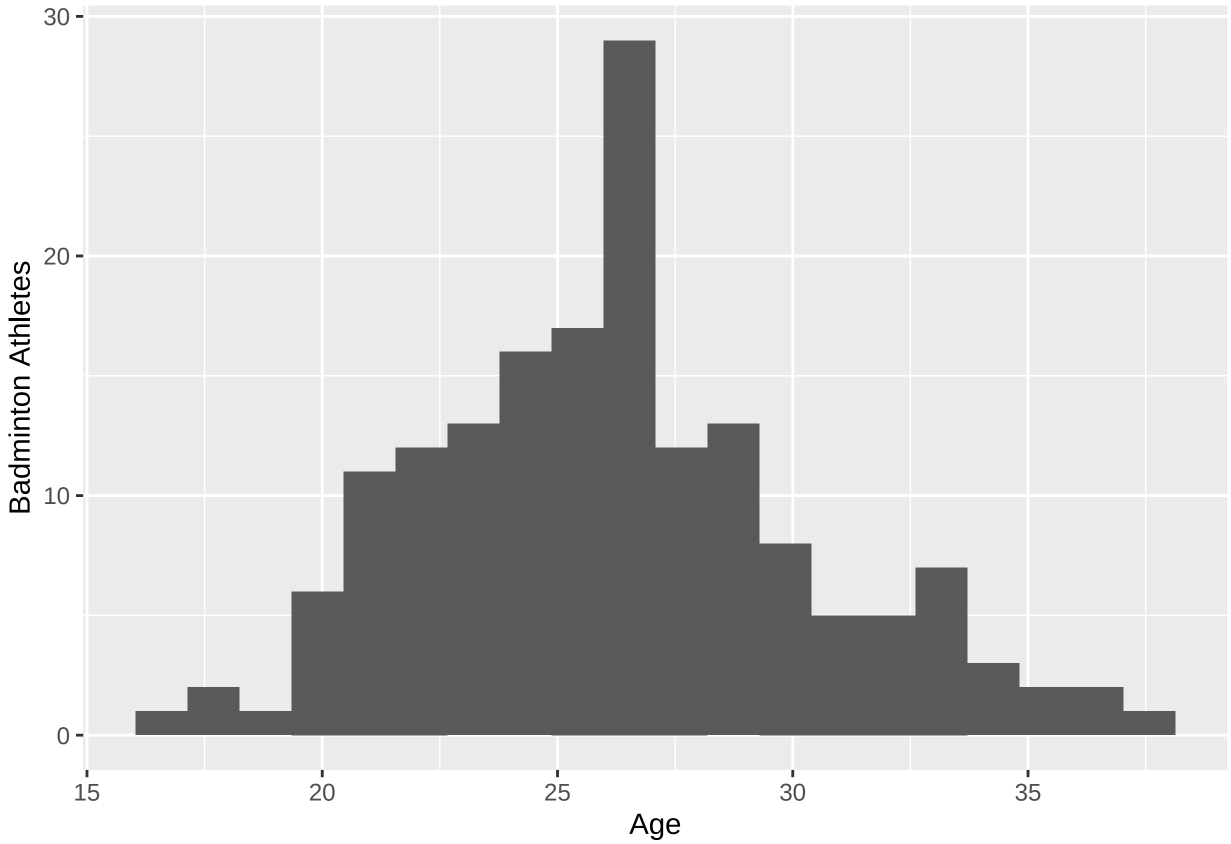
```
oly12_selectedSports <- oly12 %>%
   filter(Sport == 'Weightlifting'|Sport == 'Badminton')
```

(d) Compare the age distribution for olympic athletes competing in weightlifting to the age distribution of olympic athletes competing in badminton using both boxplots and histograms.

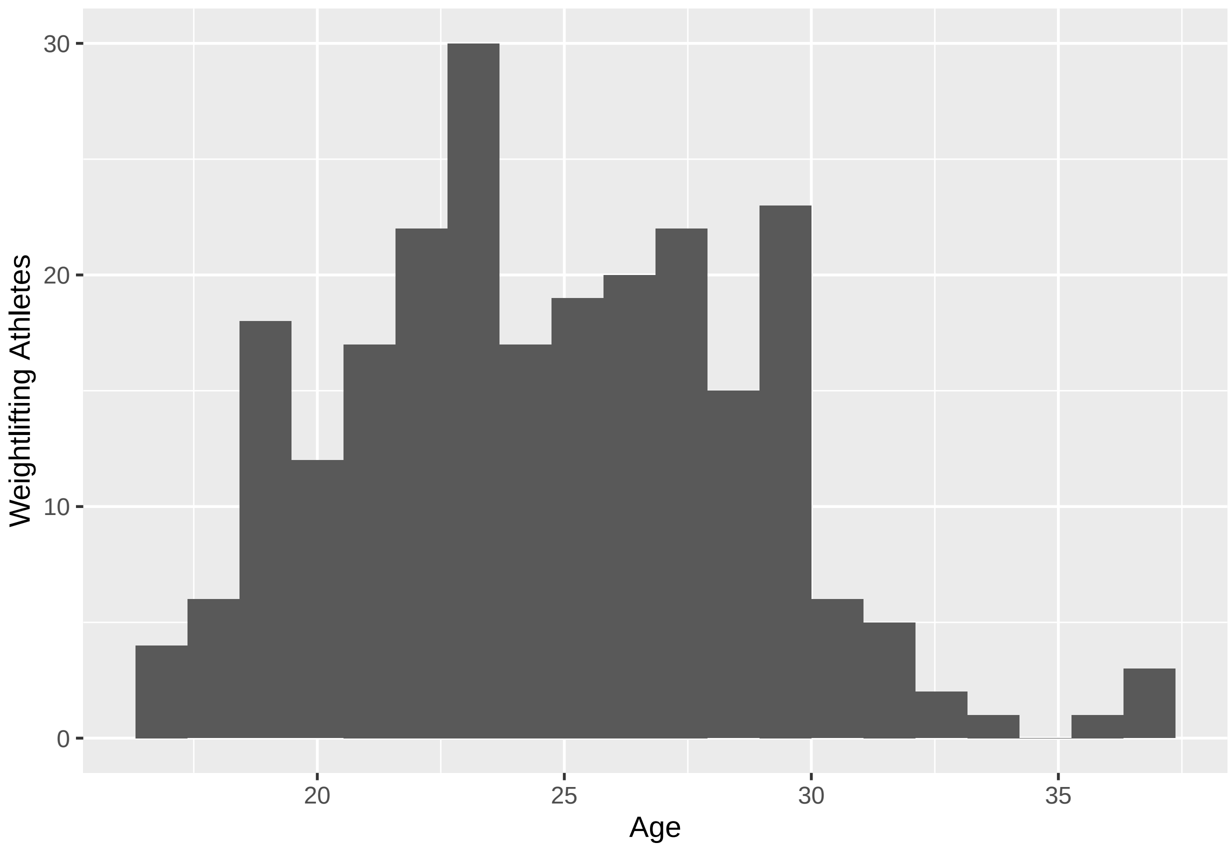
```
#the boxplot for Badminton vs. Weightlifting
oly12_selectedSports %>% ggplot(aes(x = Sport, y = Age)) +
   geom_boxplot() +
   labs(x = 'Sports', y = 'Age (in years)')
```



```
#the histograms for Badminton vs. Weightlifting
oly12_selectedSports %>% filter(Sport == 'Badminton') %>%
   ggplot(aes(x = Age)) +
   geom_histogram(bins = 20) +
   labs(y = 'Badminton Athletes', x = 'Age')
```



```
oly12_selectedSports %>% filter(Sport == 'Weightlifting') %>%
   ggplot(aes(x = Age)) +
   geom_histogram(bins = 20) +
   labs(y = 'Weightlifting Athletes', x = 'Age')
```



(e) Based on the plots you created in (d), answer the following questions:

(i) Are the age distributions of badminton players and weightlifters symmetrical or skewed?

Symmetrical for badminton athletes, and right skewed for weightlifters.

(ii) Is the median age higher for badminton players or weightlifters?

Higher for badminton athletes.

(iii) Based only on the histogram and boxplots, predict whether the standard deviation of the ages is similar or different. Justify your answer in 2-3 sentences.

The standard deviation is most likely higher for weightlifters, because the variance is higher, and standard deviation is dependent on variance. We know variance is higher, because the graph is right-skewed, so there is a greater difference/variability in the range of ages of weightlifting athletes.

(f) Create a summary table reporting the minimum, maximum, mean, median, and standard deviation of ages for badminton players and weightlifters. Compare these values to the prediction you made in (e-iii)

```
#  Type your code here
oly12_selectedSports %>% group_by(Sport) %>%
   summarise(n = n(), min(Age),
                                max(Age),
                                mean(Age),
                                median(Age),
                                sd(Age))
```

```
## # A tibble: 2 x 7
##   Sport             n `min(Age)` `max(Age)` `mean(Age)` `median(Age)` `sd(Age)`
## * <fct>         <int>      <int>      <int>       <dbl>         <dbl>     <dbl>
## 1 Badminton       166         17         38        26.2            26      4.12
## 2 Weightlifting   243         17         37        24.6            24      4.06
```

The median age of Badminton athletes were indeed higher, so I was correct about that. However, I was incorrect about the standard deviation of the ages. The standard deviation for Badminton atheletes' ages was actually higher.

(g) Use the arrange function to find the name and age of the 6 youngest athletes who competed in the 2012 Olympics.

```
oly12 %>% arrange(Age) %>% select(Name, Age) %>% head(6)
```

```
##                         Name Age
## 1               Adzo Kpossi  13
## 2         Aurelie Fanchette  14
## 3                   Suji Kim  14
## 4 Nafissatou Moussa Adamou  14
## 5  Lea Melissa Moutoussamy  14
## 6                  Yuhan Qiu  14
```

(h) Modify your code from (g) to find the name, Age, and event for the 6 youngest competitors who won gold medals at the 2012 olympics

```
oly12 %>% arrange(Age) %>% filter(Gold >= 1) %>%
   select(Name, Age, Event) %>% head(6)
```

```
##                 Name Age
## 1    Ruta Meilutyte  15
## 2          Kyla Ross  15
## 3 Gabrielle Douglas  16
## 4       Yolane Kukla  16
## 5  Mc Kayla Maroney  16
## 6          Shiwen Ye  16
##                                                                                                     Event
## 1                        Women's 50m Freestyle, Women's 100m Freestyle, Women's 100m Breaststroke
## 2                                                                   Women's Team, Women's Qualification
## 3                               Women's Individual All-Around, Women's Team, Women's Qualification
## 4                                                                  Women’s 4x100m Freestyle Relay
## 5                                                             Women’s Team, Women’s Qualification
## 6 Women’s 200m Individual Medley, Women’s 400m Individual Medley, Women’s 4x200m Freestyle Relay
```

(i) Create a new variable called `total_medals` and find the name of the athlete who won the most medals at the 2012 Olympics.

```
oly12 %>% mutate(total_medals = Total) %>%
   arrange(desc(total_medals)) %>%
   select(Name) %>% head(1)
```

```
##          Name
## 1 Ryan Lochte
```

[Question 2] 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")
glimpse(titanic)
```

```
## Rows: 2,208
## Columns: 14
## $ Name         <chr> "ABBING, Mr Anthony", "ABBOTT, Mr Ernest Owen", "ABBOT...
## $ Survived     <chr> "Dead", "Dead", "Dead", "Dead", "Alive", "Alive", "Ali...
## $ Boarded      <chr> "Southampton", "Southampton", "Southampton", "Southamp...
## $ Class        <chr> "3", "Crew", "3", "3", "3", "3", "3", "2", "2", "3", "...
## $ MWC          <chr> "Man", "Man", "Child", "Man", "Woman", "Woman", "Man",...
## $ Age          <dbl> 42.00, 21.00, 14.00, 16.00, 39.00, 16.00, 25.00, 30.00...
## $ Adut_or_Chld <chr> "Adult", "Adult", "Child", "Adult", "Adult", "Adult", ...
## $ Sex          <chr> "Male", "Male", "Male", "Male", "Female", "Female", "M...
## $ Paid         <dbl> 7.550000, NA, 20.250000, 20.250000, 20.250000, 7.65000...
## $ Ticket_No    <chr> "5547", NA, "CA2673", "CA2673", "CA2673", "348125", "3...
## $ Boat_or_Body <chr> NA, NA, NA, "[190]", "A", "16", "A", NA, "10", "15", "...
## $ Job          <chr> "Blacksmith", "Lounge Pantry Steward", "Scholar", "Jew...
## $ Class_Dept   <chr> "3rd Class Passenger", "Victualling Crew", "3rd Class ...
## $ Class_Full   <chr> "3", "V", "3", "3", "3", "3", "3", "2", "2", "3", "3",...
```

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

(i) 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.>
titanic <- titanic %>% select(!Ticket_No &
                       !Boat_or_Body &
                       !Class_Dept &
                       !Class_Full)
```

(ii) The variable `Adut_or_Chld` indicates which passengers were adults and which were children. Change the name of this variable to `Adult_or_Child`. `MWC` is a little more specific, recording whether the passenger was a man, woman or child. To make this variable name clearer, change the name of `MWC` to `Man_Woman_or_Child`. Hint: the use `rename()` function from the `dplyr` library to change the name of an existing variable. For example, the following code would change the name of the "PlaceOB" variable in the `oly12` dataset to "Place_of_birth":

```
# Note 1: Don't forget to overwrite the original tibble (i.e. save the modified data in place of the or
# Note 2: When using the rename function, put the new variable name on the left of the equals sign, and
oly12 <- oly12 %>% rename(Place_of_birth = PlaceOB)
```

```
#  Type your code here.
titanic <- titanic %>% rename(Adult_or_Child = Adut_or_Chld, Man_Woman_or_Child = MWC)
```

(b) Create a summary table reporting the number of passengers on the Titanic (n), the number of passengers who died (n_died), and the proportion of passengers who died (prop_died).

```
titanic %>% summarise(n = n(),
                      n_died = sum(Survived == "Dead"),
                      prop_died = n_died/n)
```

```
## # A tibble: 1 x 3
##       n n_died prop_died
##   <int>  <int>     <dbl>
## 1  2208   1496     0.678
```

(c) Calculate the proportion of deaths for the following groups of passengers. Note that there is more than one way to do this in each of the parts below.

(i) For men, women, and children:

```
titanic %>% group_by(Man_Woman_or_Child) %>%
  summarise(n = n(),
                      n_died = sum(Survived == "Dead"),
                      prop_died = n_died/n)
```

```
## # A tibble: 3 x 4
##   Man_Woman_or_Child     n n_died prop_died
## * <chr>              <int>  <int>     <dbl>
## 1 Child                124     60     0.484
## 2 Man                 1652   1331     0.806
## 3 Woman                432    105     0.243
```

(ii) For passengers aged between 18-25 years of age:

```
titanic %>%
  mutate(AgeGroups = case_when(Age >= 18 & Age <= 25 ~ 'Ages18_25',
                               !(Age >= 18 & Age <= 25) ~ 'Others')) %>%
  group_by(AgeGroups) %>%
  summarise(n = n(),
            n_died = sum(Survived == "Dead"),
            prop_died = n_died/n)
```

```
## # A tibble: 3 x 4
##   AgeGroups     n n_died prop_died
## * <chr>     <int>  <int>     <dbl>
## 1 Ages18_25   635    438     0.690
## 2 Others     1570   1055     0.672
## 3 <NA>          3      3     1
```

(iii) For men, women, and children among the passengers who paid more than 30 British pounds for their tickets:

```
titanic %>% mutate(Paid = case_when(Paid > 30 ~ 'Paid_More_Than_30', Paid <= 30 ~ 'Paid_Less_Than_30'))
  group_by(Man_Woman_or_Child, Paid) %>%
  summarise(n = n(),
            n_died = sum(Survived == "Dead"),
            prop_died = n_died/n)
```

```
## # A tibble: 9 x 5
## # Groups:   Man_Woman_or_Child [3]
##   Man_Woman_or_Child Paid                  n n_died prop_died
##   <chr>              <chr>             <int>  <int>     <dbl>
## 1 Child              Paid_Less_Than_30    80     34     0.425
## 2 Child              Paid_More_Than_30    43     25     0.581
## 3 Child              <NA>                  1      1     1
## 4 Man                Paid_Less_Than_30   637    549     0.862
## 5 Man                Paid_More_Than_30   149    109     0.732
## 6 Man                <NA>                866    673     0.777
## 7 Woman              Paid_Less_Than_30   259     90     0.347
## 8 Woman              Paid_More_Than_30   150     12     0.08
## 9 Woman              <NA>                 23      3     0.130
```

(iv) Write several sentences interpreting the summary tables you created in parts (i)-(iii) of this question.

The proportion of women who died was approximately 0.24, while the proportion of men who died was approximately 0.81, and the proportion of children who died was approximately 0.48. The proportion of women who died is significantly lower than the proportion of men or children who died, which indicates that women and children were offered the greatest amount of protection on the Titanic. A reason for why more children may have died, despite being offered greater protections, is because of how fragile they are.

However, when we split these groups up into those who paid more than 30 British Pounds vs those who paid less, we find that the proportion of men and women who paid more and died was less than the proportion of those who paid less. This difference is even more significantly reflected in the proportion of women who paid more and died, as the difference is 0.26.

As for the age groups, the proportion of passengers of the age group of 18-25 year olds who died was similar to the proportion of passengers who died and did not belong to that age group. One reason for this similarity is the elderly may have been extended greater protections but were fragile or did not even make it to the boats, while the able-bodied/younger adults aided the pursuit of safety.

(d) What was the most common job among passengers of the Titanic? Write 1-2 sentences explaining your answer. Hint: create a summary table reporting the number of passengers with each job title, and sort it from most common to least common job.

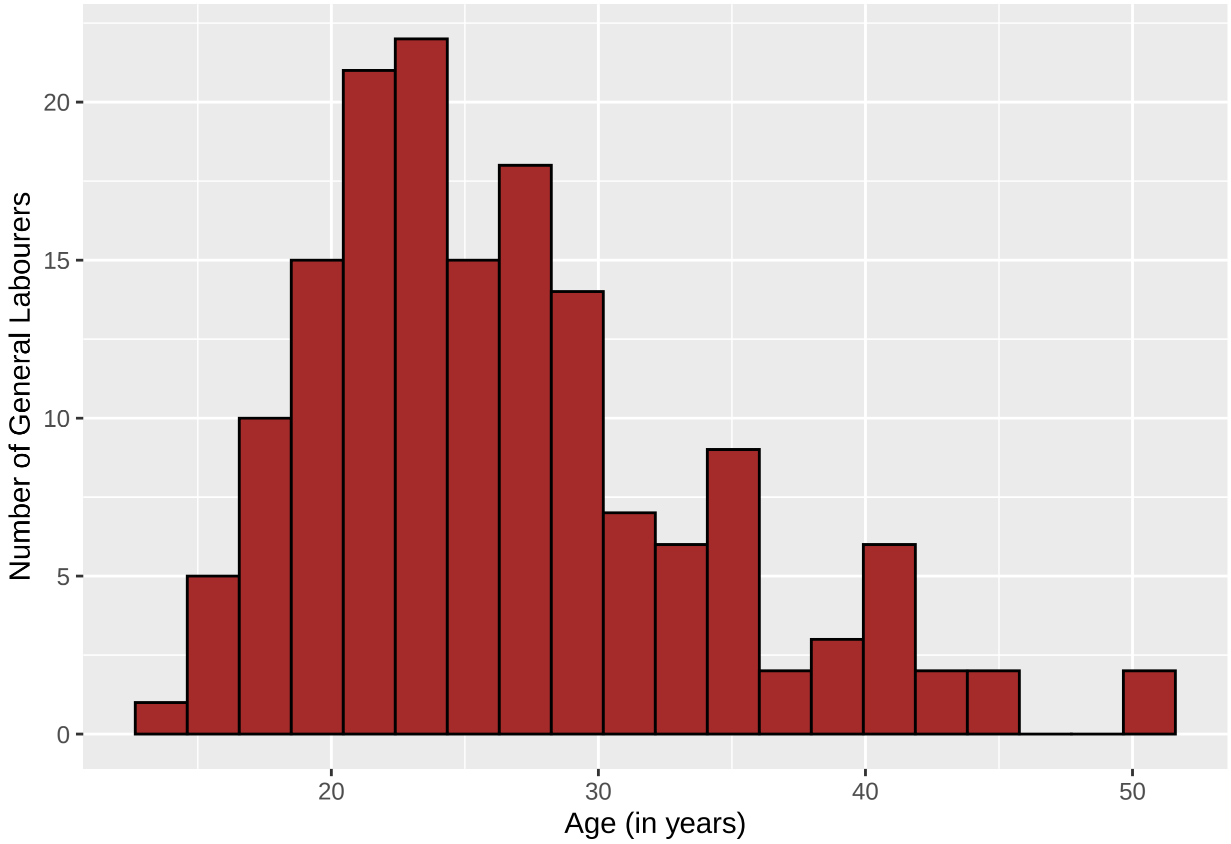
```
titanic %>% group_by(Job) %>% summarise(n = n()) %>% arrange(desc(n))
```

```
## # A tibble: 358 x 2
##    Job                            n
##    <chr>                      <int>
##  1 <NA>                         631
##  2 General Labourer             162
##  3 Fireman                      161
##  4 Trimmer                       73
##  5 Saloon Steward                56
##  6 Farm Labourer                 49
##  7 Farmer                        48
##  8 Saloon Steward (1st class)    48
##  9 Greaser                       33
## 10 Able Seaman                   28
## # ... with 348 more rows
```

Those without jobs recorded or with missing data on their occupation/job were the highest. General labourers and firemen were the two most popular jobs. Unfortunately, there were only 28 Able Seamen, which is ironic, because one would have expected more seamen on the Titanic to aid the voyage.

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

```
titanic %>% filter(Job == "General Labourer") %>%
  ggplot(aes(x = Age)) +
  geom_histogram(bins = 20, color = "black", fill = "brown") +
  labs(y = "Number of General Labourers", x = "Age (in years)")
```



This distribution is right skewed and unimodal, and we can see that the peak/modal is around 23-24 years of age. This means that more general labourers were 23-24 years old than any other age, and most of the general labourers were concentrated at around 19-30 years of age.

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

```
titanic %>% filter(Job == "General Labourer") %>%
  group_by(Man_Woman_or_Child) %>%
  summarize(n=n())
```

```
## # A tibble: 3 x 2
##   Man_Woman_or_Child     n
## * <chr>              <int>
## 1 Child                  1
## 2 Man                  160
## 3 Woman                  1
```

There was only one woman who was a general labourer.

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

```
titanic %>% summarise(Name = Name, n = n(), Paid = Paid, Survived = Survived) %>% arrange(desc(Paid)) %
```

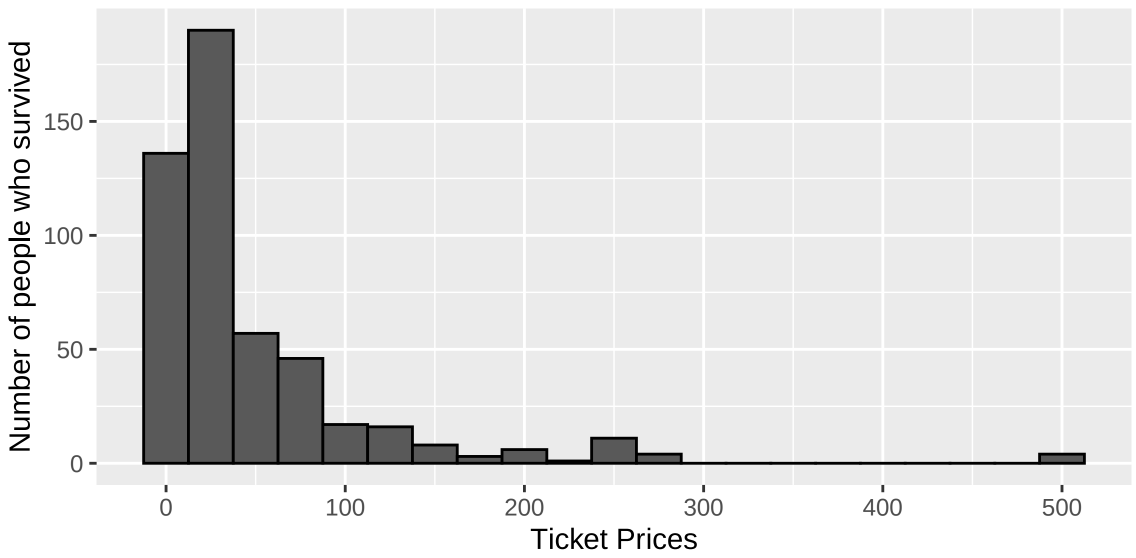
```
## # A tibble: 4 x 4
##   Name                                   n  Paid Survived
##   <chr>                              <int> <dbl> <chr>
## 1 CARDEZA, Mr Thomas Drake Martinez   2208  512. Alive
## 2 CARDEZA, Mrs Charlotte Wardle       2208  512. Alive
## 3 LESUEUR, Mr Gustave J.              2208  512. Alive
## 4 WARD, Miss Annie Moore              2208  512. Alive
```

The names of the passengers with the top 4 most expensive tickets were Mr. Thomas Drake Martinez, Mrs, Charlotte Wardle, Mr. Gustave J., and Miss Annie Moore, and they all survived.

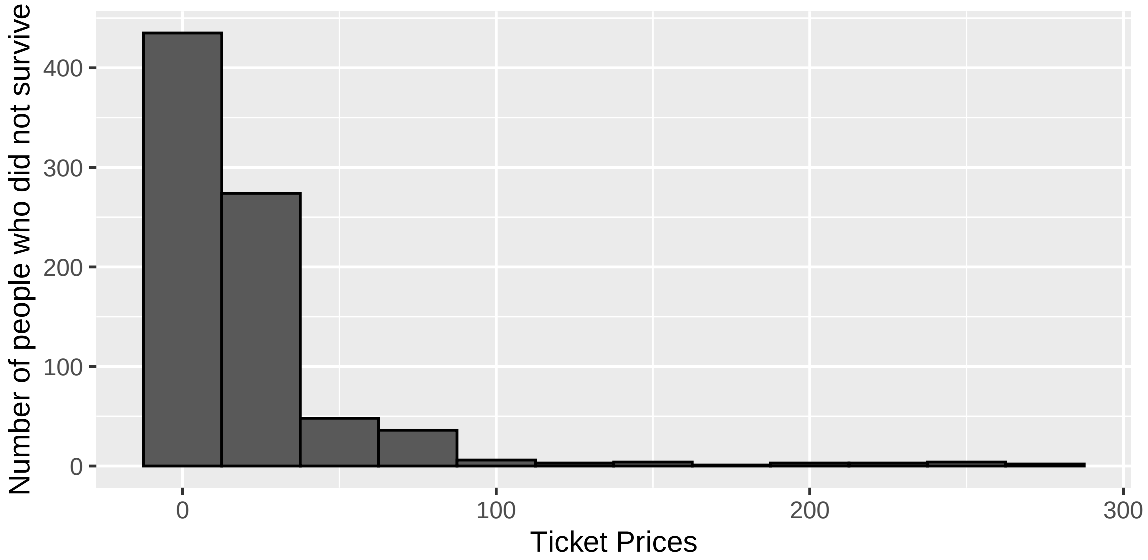
(h) In this question, you will compare the distribution of ticket prices for survivors and non-survivors of the Titanic using both visualizations and summary tables.

(i) 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.

```
titanic %>% filter(Survived == "Alive") %>%
  ggplot(aes(x = Paid)) +
  geom_histogram(color = "black", binwidth = 25) +
  labs(x = "Ticket Prices", y = "Number of people who survived")
```



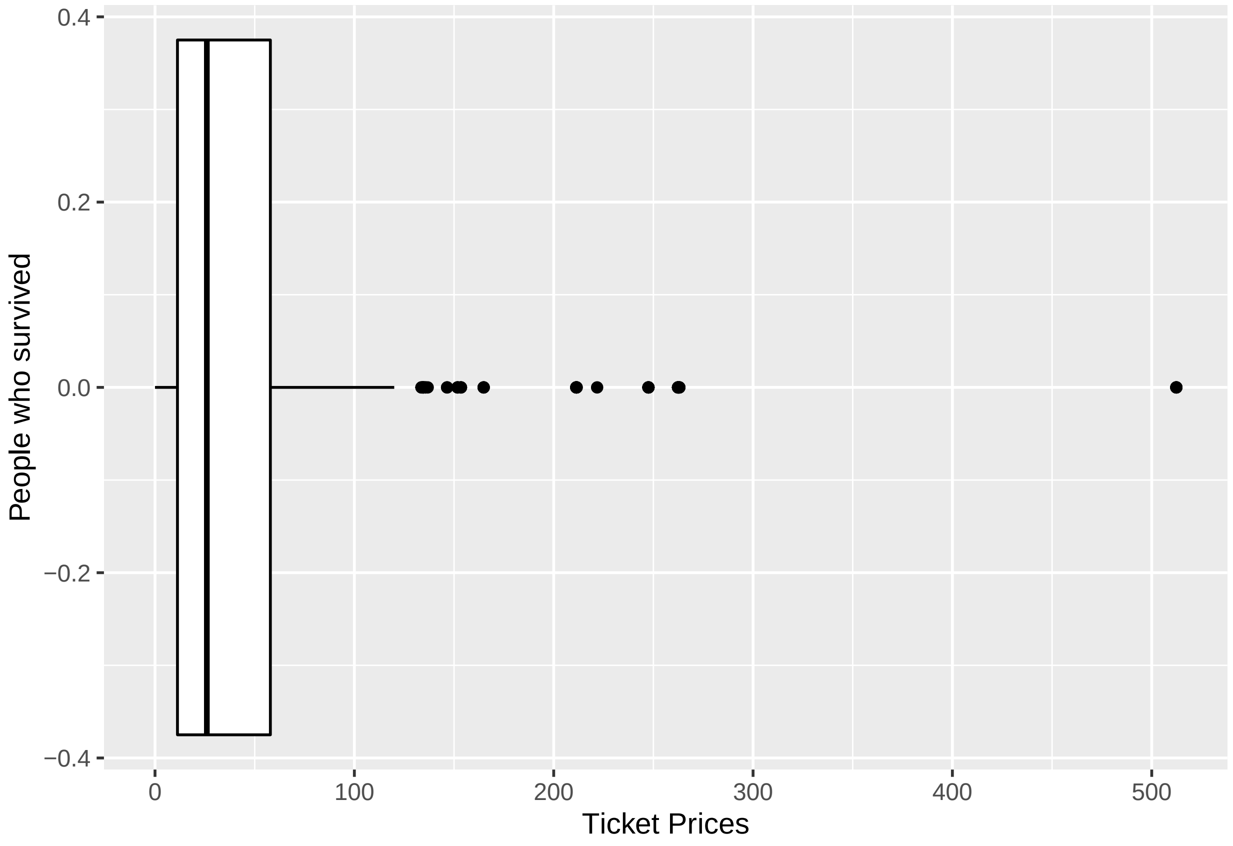
```
titanic %>% filter(Survived == "Dead") %>%
  ggplot(aes(x = Paid)) +
  geom_histogram(color = "black", binwidth = 25) +
  labs(x = "Ticket Prices", y = "Number of people who did not survive")
```



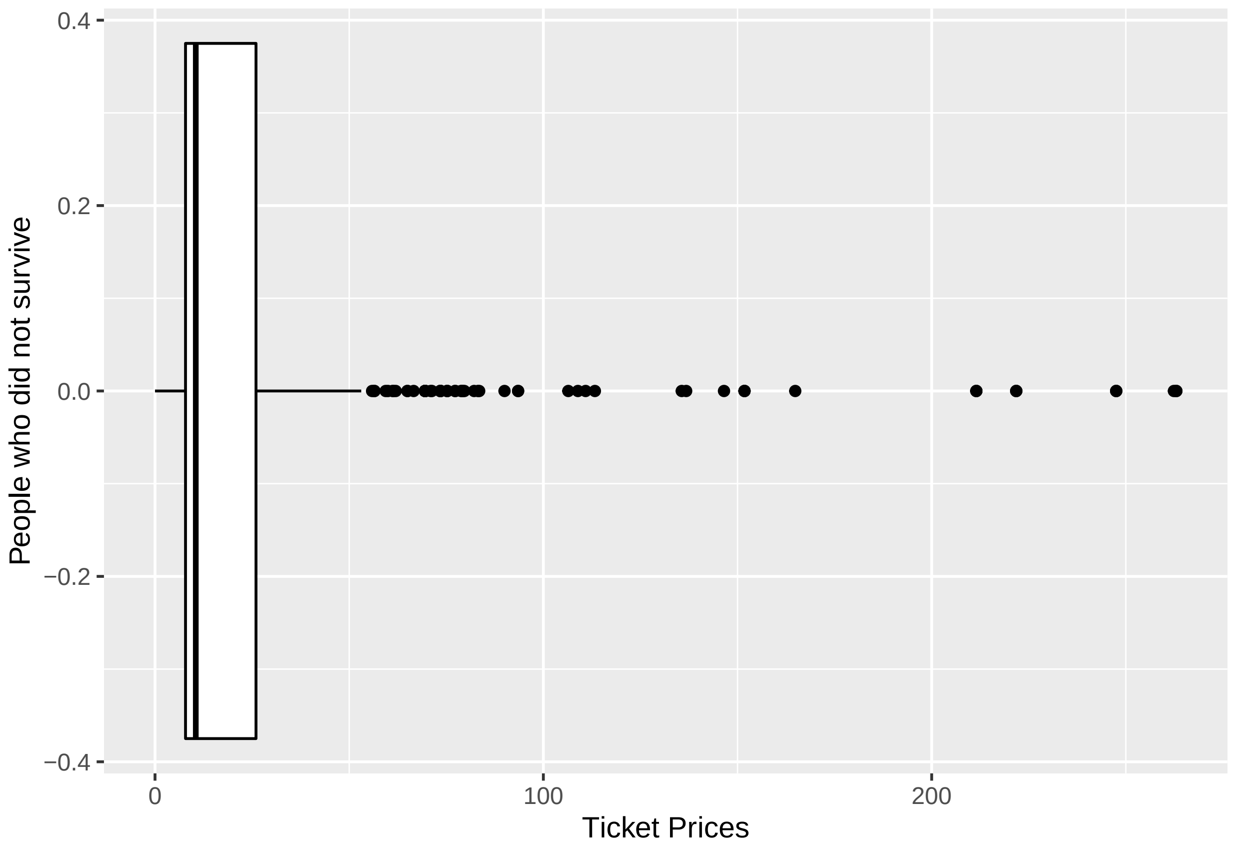
The distributions are pretty similar in shape, as they are both right skewed and unimodal. However, we can see that the number of people who did not survive and paid more is far fewer than those who paid more and survived.

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

```
titanic %>% filter(Survived == "Alive") %>%
  ggplot(aes(x = Paid)) +
  geom_boxplot(color = "black") +
  labs(x = "Ticket Prices", y = "People who survived")
```



```
titanic %>% filter(Survived == "Dead") %>%
  ggplot(aes(x = Paid)) +
  geom_boxplot(color = "black") +
  labs(x = "Ticket Prices", y = "People who did not survive")
```



We can see that both boxplots are similar in shape again, as they are both right skewed. However, with the boxplot for the distribution of ticket prices among those who survived, we can see that the median is higher than the median for the boxplot representing those who did not survive.

(iii) Construct a summary table with the minimum, first quartile, median, mean, third quartile, and maximum ticket price for survivors and non-survivors. Hint: The code below gives an example of the `quantile` function, which you'll use to calculate Q1 and Q3, as well as the na.rm=TRUE option:

```
#### 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
```

```
titanic %>% filter(!is.na(Paid)) %>% group_by(Survived) %>%
  summarize(n = n(), min = min(Paid),
            Q1 = quantile(Paid, probs = 0.25, na.rm = TRUE),
            mean = mean(Paid),
            Q3 = quantile(Paid, probs = 0.75, na.rm = TRUE),
            max = max(Paid))
```

```
## # A tibble: 2 x 7
##   Survived     n   min    Q1  mean    Q3   max
## * <chr>    <int> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Alive      499     0 11.3   49.6  57.9  512.
## 2 Dead       819     0  7.85  22.9  26    263
```

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

The average amount of money paid for tickets among those who survived was twice as high as the average ticket price for those who died. And this is the same for the Q1 and Q3 values as well.

<Type your answer here.>

<Type your answer here.>

(iv) Comment on the strengths and weaknesses of each of the visualizations and summary table you constructed in parts (i), (ii), and (iii)

The strength of the summary table is that I can see the exact calculated values displayed for the main data points of a distribution. However, I cannot see the shape of the distribution like I can with the boxplot and histogram visualizations. The boxplot visualization is almost a bit of both the summary table and the histogram, because one can see the shape and also estimate the values of the median, mean, and quartiles. Nonetheless, it will still not be as visually specific as the histogram or as numerically specific as the summary table.

[Question 3] The code below reads in data about books sold on Amazon (https://dasl.datadescription.com/datafile/amazon-books/). The data frame containing the book data is named `books`. Note that the height (`Height`), width (`Width`) and thickness (`Thick`) of books in this data frame are measured in inches.

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

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

```
glimpse(books)
```

```
## Rows: 325
## Columns: 13
## $ Title         <chr> "1,001 Facts that Will Scare the S#*t Out of You: The...
## $ Author        <chr> "Cary McNeal", "Ben Mezrich", "Smith", "Gavin Menzies...
## $ List.Price    <dbl> 12.95, 15.00, 1.50, 15.99, 30.50, 28.95, 20.00, 15.00...
## $ Amazon.Price  <dbl> 5.18, 10.20, 1.50, 10.87, 16.77, 16.44, 13.46, 8.44, ...
## $ Hard_or_Paper <chr> "P", "P", "P", "P", "P", "H", "H", "P", "H", "H", "P"...
## $ NumPages      <int> 304, 273, 96, 672, 720, 460, 336, 405, NA, 304, 624, ...
## $ Publisher     <chr> "Adams Media", "Free Press", "Dover Publications", "H...
## $ Pub.year      <int> 2010, 2008, 1995, 2008, 2011, 2011, 2010, 1987, 2011,...
## $ ISBN.10       <chr> "1605506249", "1416564195", "486285537", "61564893", ...
## $ Height        <dbl> 7.8, 8.4, 8.3, 8.8, 8.0, 8.9, 7.8, 8.2, 9.6, 9.6, 7.7...
## $ Width         <dbl> 5.5, 5.5, 5.2, 6.0, 5.2, 6.3, 5.3, 5.3, 6.5, 6.4, 5.1...
## $ Thick         <dbl> 0.8, 0.7, 0.3, 1.6, 1.4, 1.7, 1.2, 0.8, 2.1, 1.1, 1.7...
## $ Weight_oz     <dbl> 11.2, 7.2, 4.0, 28.8, 22.4, 32.0, 15.5, 11.2, NA, 19....
```

```
books %>% select(Title, NumPages) %>%
  arrange(desc(NumPages)) %>%
  head(1)
```

```
##              Title NumPages
## 1 Andrew Carnegie      896
```

The name of the book is Andrew Carnegie and it has 896 pages.

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

```
books %>% group_by(Author) %>%
  summarise(NumBooks =n(),
            AvgPages = mean(NumPages))
```

```
## # A tibble: 256 x 3
##    Author              NumBooks AvgPages
##   * <chr>                <int>    <dbl>
##  1 ""                        1      432
##  2 "Abraham Verghese"        1      667
##  3 "Adam Goodheart"          1      460
##  4 "Adam Hochschild"         1      480
##  5 "Adam Mansbach"           1       32
##  6 "Alaa Aswany"             1      255
##  7 "Alice Munro"             2      320
##  8 "Alice Schroeder"         1      832
##  9 "Allen, Toorawa"          1      200
## 10 "Andrea Warren"           1      160
## # ... with 246 more rows
```

(c) Modify your code from (b) so to create a new summary table which contains only information for authors who wrote 5 or more books, and sort them in decreasing order of number of books written.

```
books %>%
  group_by(Author) %>%
  summarise(NumBooks =n(),
            AvgPages = mean(NumPages)) %>%
  filter(NumBooks >= 5)
```

```
## # A tibble: 2 x 3
##   Author           NumBooks AvgPages
##   <chr>               <int>    <dbl>
## 1 Jodi Picoult            7     414.
## 2 Vladimir Nabokov        7     316
```

Part 2

Writing prompt

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 (since your boss' favourite sports are badminton and weightlifting, you know she will be happy if your summary talks about these sports specifically, but you can talk about other interesting of features athletes' ages which can be learned from your plots and tables.)

Important Features to Include

- Start off with a small introduction. You should include 1 or 2 sentences to draw your reader in, and then explain what you will be discussing.
- Make sure to include at least 1 figure to help your reader visualize what you are speaking about.
- You want to show off all the knowledge you gained in STA130 so you must include at least 2 vocabulary words. However, your boss isn't a statistician, therefore you must define any vocabulary terms you used.
- Make sure to finish with a conclusion to remind your boss of the key take home points from your summary about the athletes' ages.

Some general reminders

- Try to not spend more than 20 minutes on the prompt.
- Aim for more than 200 words but less than 400 words.
- Use full sentences.
- Grammar is not the main focus of the assessment, but it is important that you communicate in a clear and professional manner. Remember, this is meant to be a small report to your boss! So you should not include any slang or emojis.
- Remember that you have only conducted a preliminary analysis and therefore you may not have a definitive answer. That is totally ok! It is hard to ever say something is 100% one way or another. Therefore, you will want to try to incorporate some hedging language into your writing. You can find more information about hedging in a short video here: https://q.utoronto.ca/courses/206597/pages/writing-skills-videos?module_item_id=2116409

Vocabulary

- Cleaning data
- Tidy data
- 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
- Renaming the variables
- Grouping categories
- Defining new variables
- Producing new data frames
- Handling missing values (NAs)
- Creating summary tables

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

- Where are the data centered (towards the left, right, middle)
- How much spread (relative to what?)
- Shape: symmetric, left-skewed, right-skewed
- The tails of the distribution (heavy-tailed or thin-tailed)
- Modes: where, how many, unimodal, bimodal, multimodal, uniform
- Outliers, extreme values
- Frequency (which category occurred the most or least often; data concentrated near a particular value or category)
- Mean (average), median, mode
- standard deviation, interquartile range

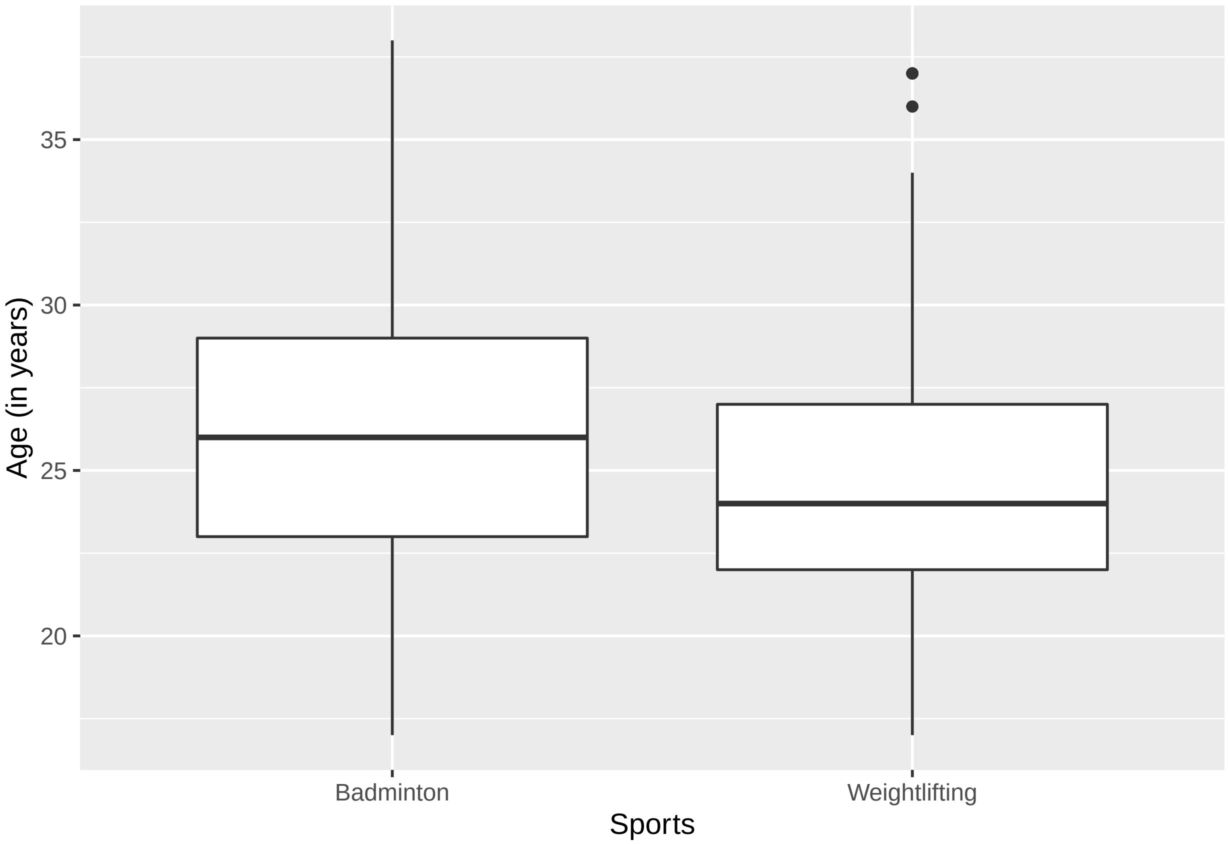
The Olympics in 2012 was a very age-inclusive, with athletes ranging from 13 years of age to 71 years of age. The average age of athletes was around 26 years old, and the median age was close, at around 25 years old. This means that more than half of the Olympic athletes in 2012 were older or younger than 25 years old. More details can be seen in the figure below:

```
oly12 %>% filter(!is.na(Age)) %>%
  select(Age) %>%
  summarise(Youngest = min(Age), AvgAge = mean(Age),
            Q1 = quantile(Age, probs = 0.25),
            MedAge = median(Age),
            Q3 = quantile(Age, probs = 0.75),
            Oldest = max(Age))
```

```
##   Youngest   AvgAge Q1 MedAge Q3 Oldest
## 1       13 26.06886 22     25 29     71
```

After cleaning the data and extracting data on specifically athletes who participated in Badminton or Weightlifting, I produced a new dataframe called oly12_selectedSports. I then used boxplots to visualize the distribution of age for the Olympic athletes who participated in these events, and this resulted in the following figures.

```
oly12_selectedSports %>% ggplot(aes(x = Sport, y = Age)) +
   geom_boxplot() +
   labs(x = 'Sports', y = 'Age (in years)')
```



As we can see in the boxplots, the median age of Badminton athletes were higher than the Weightlifting athletes, which means half of the demographic is older or younger than that age. This means that it is more common for Badminton athletes to be of higher age than Weightlifting athletes. And the variance for Badminton atheletes' ages can also be estimated to be higher, which could indicate that Badminton is more age-inclusive or that it is easier for persons of any age to participate in the sport than in Weightlfiting.

In conclusion, the Olympics of 2012 are a perfect example of how sports is not limited to younger adults. Even older adults, like our 71 year old Olympian in 2012, can participate and prove they are able-bodied enough to engage in competitive sports, and that is the beauty of (the) sports!