STA130 Rstudio Homework

Problem Set 3

Daniel Sun (1008992609), with Josh Speagle & Scott Schwartz

Instructions

Complete the exercises in this `.Rmd` file and submit your `.Rmd` and knitted `.pdf` output through Quercus by 11:59 pm E.T. on Thursday, February 2.

Question 1: 2012 Olympics

The code below uses `names()` to show all the column names of the `oly12` data set and then `glimpse()` to provide a preview the entire data set. Note that the `oly12` data set is *not the same* as the `olympics` data set shown in class.

```
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~
```

(a) During our class meeting this week, 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.). In that data set, which we called `olympics`, there was one observation (i.e. one row) for each participating country.

What does each row in the `oly12` data set (loaded above) represent?

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). Alternately, you can search for "oly12" in the Help tab.

Each row represents an athlete

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

Hint: Applying the `filter()` function to the `Country` column of the `oly12` dataset will be much easier than sorting through each entry one at a time.

```
CAUSA<-filter(oly12, Country== "Canada" | Country == "United States of America")
nrow(CAUSA)
```

```
## [1] 792
```

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

Hint: You can see all the possible values for the `Sport` variable by applying the `levels()` function to the `oly12$Sport` column. You can count the number of possible levels using the `nlevels()` function.

```
a<-nrow(filter(oly12, Sport == "Gymnastics - Artistic" )) #182
b<-nrow(filter(oly12, Sport == "Gymnastics - Rhythmic")) #92
c<-nrow(filter(oly12, Sport == "Diving")) #133
d<-nrow(filter(oly12, Sport == "Swimming")) #907
sum(a,b,c,d)
```

```
## [1] 1314
```

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

Hint: The `%in%` comparision operator could be useful here, which allows us to determine if a value `x` matches with an entry within a vector `v`. If we define `allGymnastics <- c("Gymnastics - Artistic", "Gymnastics - Rhythmic", "Trampoline")`, for instance, then `filter(Sport %in% allGymnastics)` would return entries that matched any of the categories in `allGymnastics`. See this stackoverflow post for additional discussion.

```
sport <- c("Gymnastics - Artistic", "Gymnastics - Rhythmic", "Trampoline", "Diving", "Swimming", "Synch
nrow(filter(oly12, Sport %in% sport ))
```

```
## [1] 1695
```

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

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

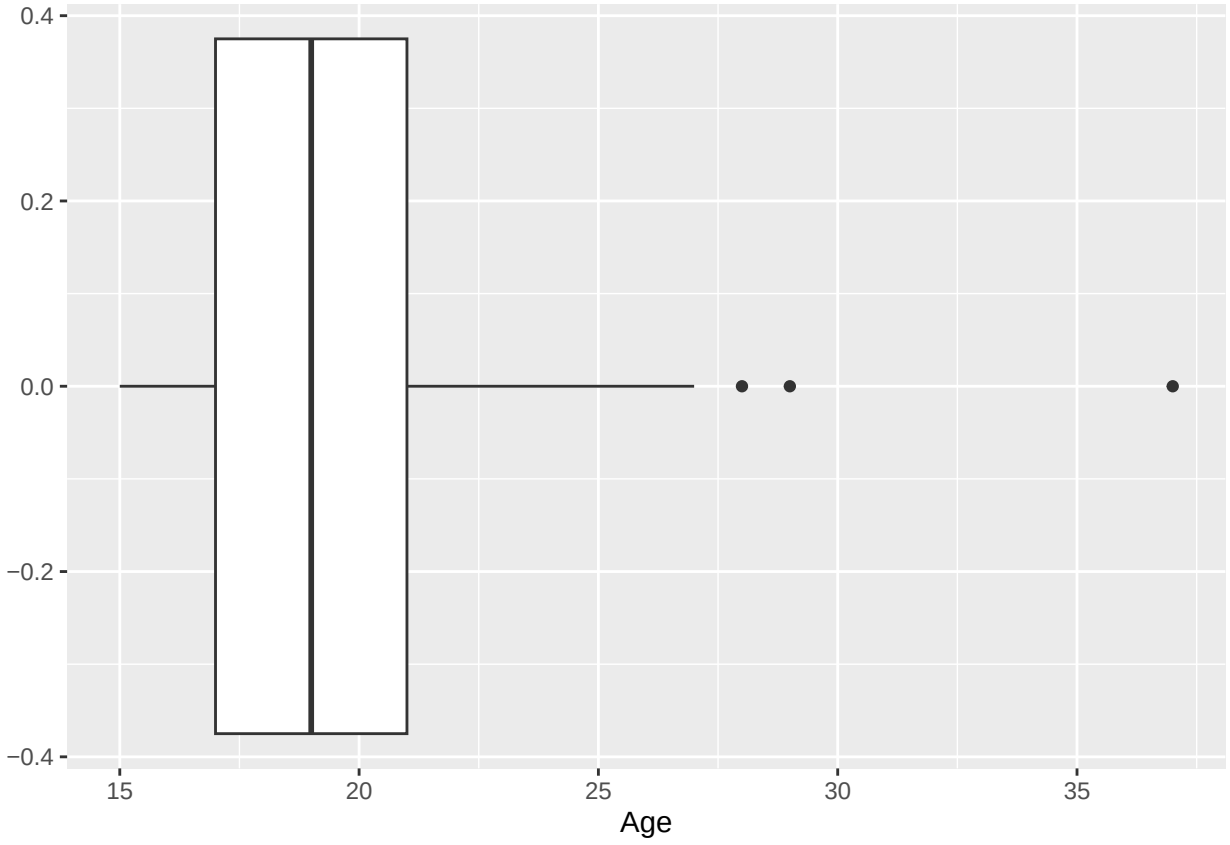
```
sport <- c("Gymnastics - Artistic", "Gymnastics - Rhythmic")
x<-filter(oly12, Sport %in% sport )
sex <- c("F")
oly12_FemaleArtisticRhythmicGymnasts<-subset(filter(x, Sex %in% sex))
#oly12_FemaleArtisticRhythmicGymnasts cannot run this otherwise pdf is too long
```

(f) Use `oly12_FemaleArtisticRhythmicGymnasts` and `ggplot2` to create both boxplots and histograms to compare (1) the age distribution of female Olympic athletes competing in artistic gymnastics to (2) the age distribution of female Olympic athletes competing in rhythmic gymnastics.

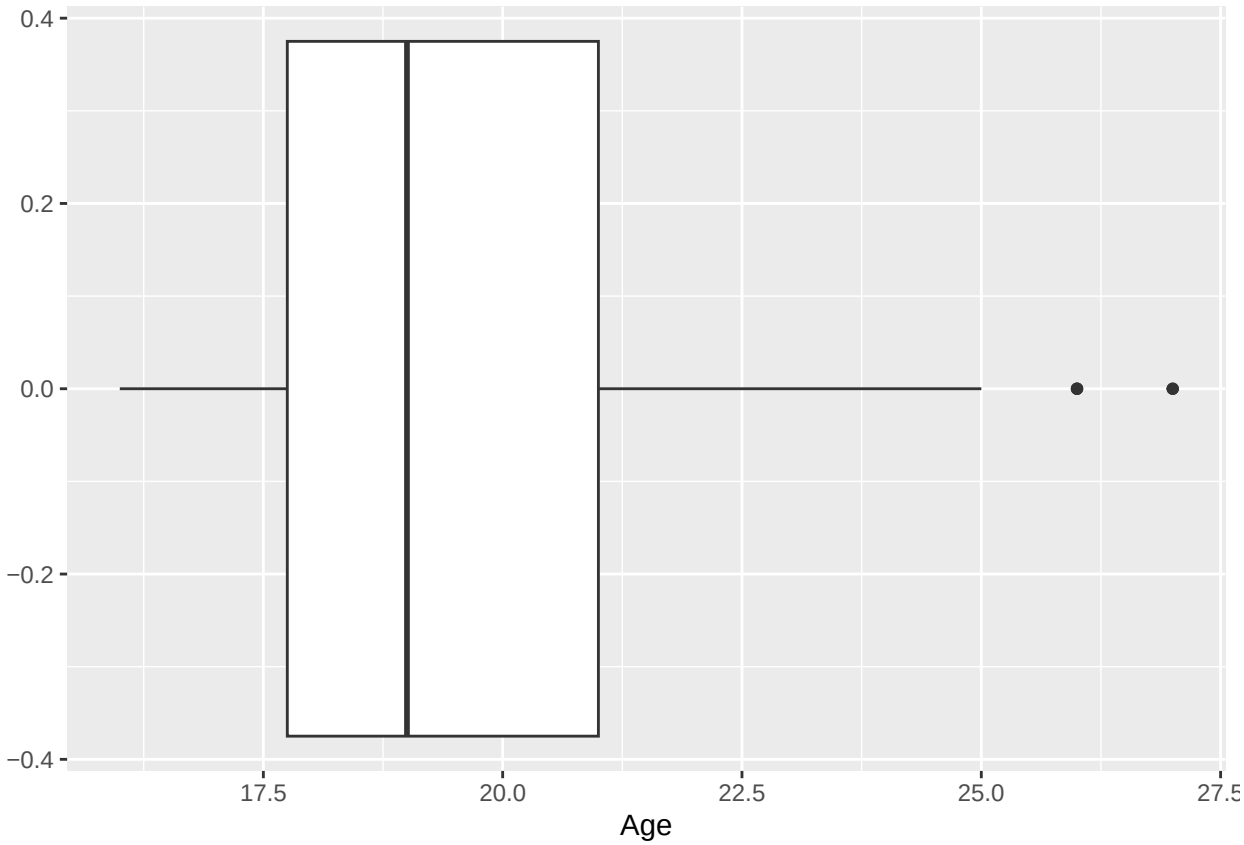
```
g <-"Gymnastics - Artistic"
y <-filter(oly12_FemaleArtisticRhythmicGymnasts, Sport %in% g)
```

```
a <-"Gymnastics - Rhythmic"
b <-filter(oly12_FemaleArtisticRhythmicGymnasts, Sport %in% a)


ggplot(data=y, aes(x= Age)) + geom_boxplot()
```

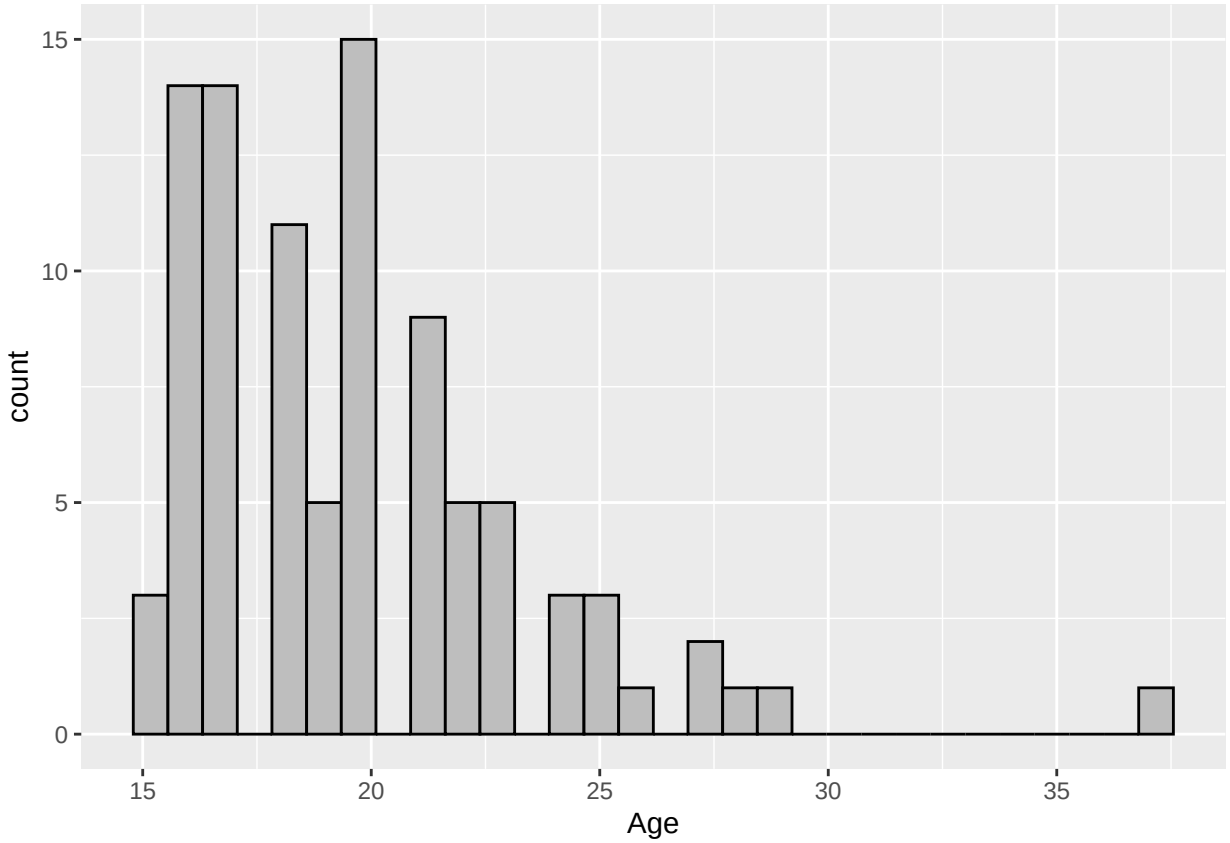



```
ggplot(data=b, aes(x= Age)) + geom_boxplot()
```



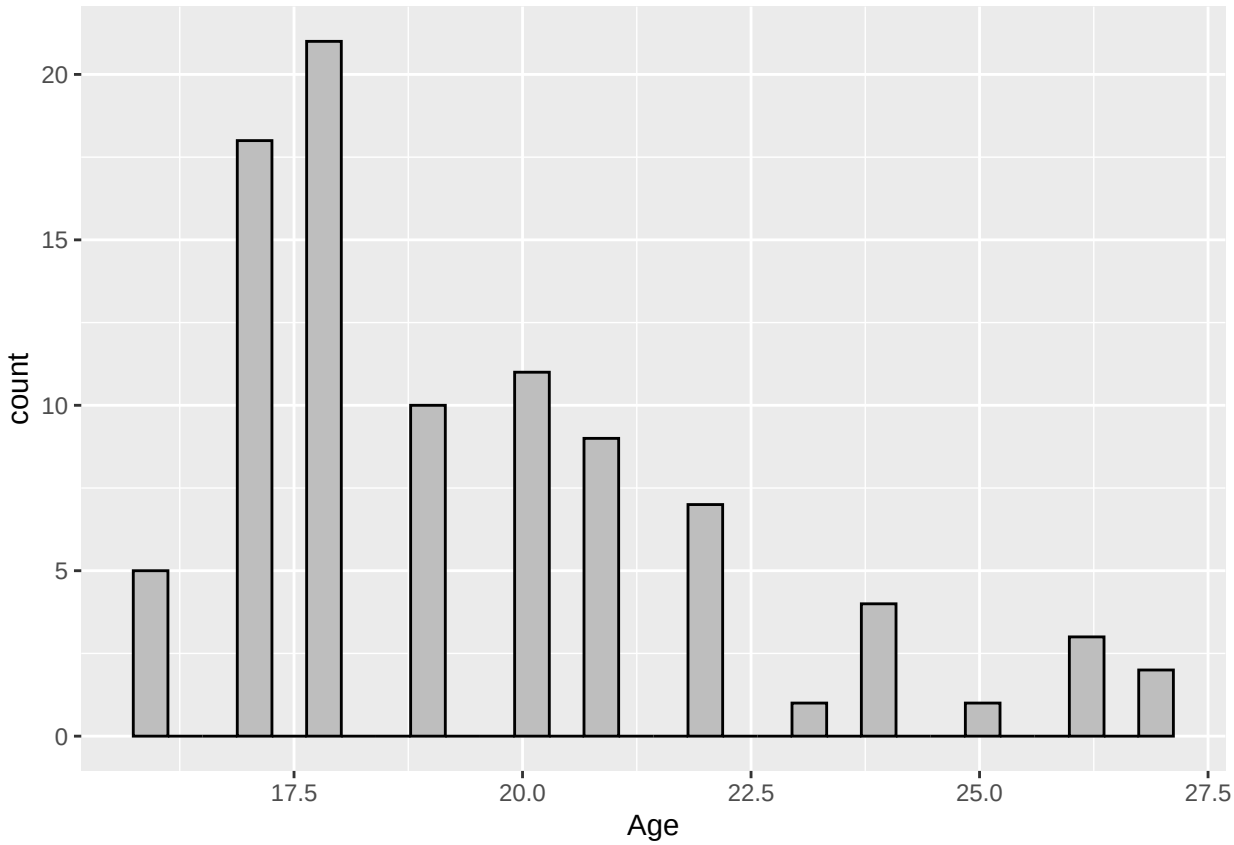

```
ggplot(data=y, aes(x=Age)) + geom_histogram(colour = "black", fill = "grey")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(data=b, aes(x=Age)) + geom_histogram(colour = "black", fill = "grey")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



(g) Answer the following questions in 1-2 sentences based on the plots you created in (d).

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

   The distributions are right skewed

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

   The difference between the median and the 75th percentile is much larger for rhythmic gymnasts than artists gymnasts indicating that the general age is probably older. This is reflected as the general plot shifts to the right for the artistic gymnasts. By extension the rhythmic plot is a left shift over showing that their youngest are probably much younger than the artistic gymnasts The median is also closer for rhythmic meaning that there is an even spread of ages in that age range.

3. Based only on the histograms and boxplots, predict whether the standard deviation of the ages is similar or different and justify your reasoning.

   As the spread of 25th-75th percentile spread is different, the standard deviation is different.

(h) Use `summarise()` to create a summary table of `oly12_FemaleArtisticRhythmicGymnasts` that report the following statistics based on the ages for female rhythmic gymnasts and female artistic gymnasts:

- the minimum (`min`),
- the maximum (`max`),
- the mean (`mean`),
- the median (`median`), and

- the standard deviation (`sd`).

Hint: Running ***`group_by()`*** *over the relevant column before running* ***`summarise()`*** *will simultaneously generate summaries over both groups.*

```
a <-"Gymnastics - Rhythmic"
b <-filter(oly12_FemaleArtisticRhythmicGymnasts, Sport %in% a)
c<-group_by(b, Age)
summarize(c)
```

```
## # A tibble: 12 x 1
##      Age
##    <int>
##  1    16
##  2    17
##  3    18
##  4    19
##  5    20
##  6    21
##  7    22
##  8    23
##  9    24
## 10    25
## 11    26
## 12    27
```

```
x<-"Gymnastics - Artistic"
y <-filter(oly12_FemaleArtisticRhythmicGymnasts, Sport %in% x)
z<-group_by(y, Age)
summarize(z)
```

```
## # A tibble: 16 x 1
##      Age
##    <int>
##  1    15
##  2    16
##  3    17
##  4    18
##  5    19
##  6    20
##  7    21
##  8    22
##  9    23
## 10    24
## 11    25
## 12    26
## 13    27
## 14    28
## 15    29
## 16    37
```

Were you correct in your guess about the standard deviation in part (g) of the last question?

I believe so

(i) Use `mutate()` to create a new variable called `medal_points` that awards 3 points for a gold, 2 for a silver, and 1 for a bronze. Then, create a new tibble called `oly12_OneMedalClub` that contains athletes who won *exactly* one medal at the 2012 olympics. Finally, use the `glimpse()` function to verify the properties of your

tibble.

```
newoly12 <- mutate(oly12, medal_points = Gold*3 + Silver*2 + Bronze*1)# %>%
#arrange(desc(medal_points)) if someone wanted them arranged uncomment ^ this too
#newoly12 i have to comment this otherwise it will print the whole data frame

oly12_OneMedalClub <-filter(oly12, Total == 1)
glimpse(oly12_OneMedalClub)
```

```
## Rows: 457
## Columns: 14
## $ Name    <fct> Jennifer Abel, Alaaeldin Abouelkassem, Chantal Achterberg, Fil~
## $ Country <fct> "Canada", "Egypt", "Netherlands", "Germany", "Great Britain", ~
## $ Age     <int> 20, 21, 27, 29, 23, 20, 21, 23, 41, 37, 26, 32, 21, 38, 36, 18~
## $ Height  <dbl> 1.60, 1.88, 1.71, 1.89, 1.79, 1.58, 1.78, 1.83, 1.78, 1.62, 1.~
## $ Weight  <int> 62, 82, 72, 90, 70, NA, 78, 80, 70, 55, 70, 52, 64, 58, 59, 61~
## $ Sex     <fct> F, M, F, M, F, F, F, M, M, F, F, F, F, F, F, F, M, F, F, M, F,~
## $ DOB     <date> NA, NA, NA, 1983-05-01, NA, NA, 1991-03-08, NA, NA, 1974-08-1~
## $ PlaceOB <fct> "Montreal (CAN)", "", "", "", "Mansfield (GBR)", "Tula (RUS)",~
## $ Gold    <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0,~
## $ Silver  <int> 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1,~
## $ Bronze  <int> 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0,~
## $ Total   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ Sport   <fct> "Diving", "Fencing", "Rowing", "Rowing", "Swimming", "Gymnasti~
## $ Event   <fct> "Women's 3m Springboard, Women's Synchronised 3m Springboard",~
```

(j) Use a combination of `select()`, `arrange()`, `desc()`, and/or `filter()` to:

1. Find the `Name` and `Age` variables of the six oldest athletes who competed in the 2012 Olympics.

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

```
##                Name Age
## 1   Hiroshi Hoketsu  71
## 2 Afanasijs Kuzmins  65
## 3        Ian Millar  65
## 4    Carl Bouckaert  58
## 5  Andrei Kavalenka  57
## 6        Mary Hanna  57
```

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

```
oly12 %>%
filter(Sex == "F") %>%
  arrange(Age) %>%
  select(Name, Age, Sport) %>%
  head()
```

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

3. Find the `Name`, `Age`, `Sport`, and `Event` for the 6 youngest and 6 oldest competitors who won gold medals at the 2012 Olympics.

Note that this can be run as two pieces of code rather than one piece of combined code.

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

```
##                Name Age                Sport
## 1    Ruta Meilutyte  15             Swimming
## 2         Kyla Ross  15 Gymnastics - Artistic
## 3 Gabrielle Douglas  16 Gymnastics - Artistic
## 4      Yolane Kukla  16             Swimming
## 5  Mc Kayla Maroney  16 Gymnastics - Artistic
## 6         Shiwen Ye  16             Swimming
##                                                                                                  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
```

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

```
##                 Name Age          Sport
## 1      Peter Thomsen  51     Equestrian
## 2      Ingrid Klimke  44     Equestrian
## 3    Sergei Martynov  44       Shooting
## 4  Kristin Armstrong  38 Cycling - Road
## 5  Valentina Vezzali  38        Fencing
## 6 Alexandr Vinokurov  38 Cycling - Road
##                                                Event
## 1          Individual Eventing, Team Eventing, BARNY
## 2 Individual Eventing, Team Eventing, BUTTS ABRAXXAS
## 3                              Men's 50m Rifle Prone
## 4   Women's Individual Time Trial, Women's Road Race
## 5         Women's Individual Foil, Women's Team Foil
## 6       Men's Individual Time Trial, Men's Road Race
```

Question 2: The Data Consultant

You have just been hired by a consultancy company. Congratulations!

Your new employer is 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.

In addition, you happen to know that your new boss' favourite sports are badminton and weightlifting. You conclude that addressing these sports specifically might be an easy way to capture their attention. However, you also are aware that the report as a whole needs to describe all types of athletes and events within the

2012 Olympics. And, of course, you want to include appealing and informative plots and tables that your clients can easily understand and learn from. The more interesting the better!

Remember: - This is meant to be a quick report for your boss, so use full sentences and communicate in a clear and professional manner (so don't use slang or emojis). - Grammar isn't the main focus of this assessment, although readability is important. - **Avoid "Analysis Paralysis"**: This is envisioned as a **30-60 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 focus on quickly finding something you can communicate and write about rather than worrying too much about the exact argument.

(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 is often used with respect to the strength of the argument, the limitations of data, and the generalizability of the conclusions.

> *Completed the video.*

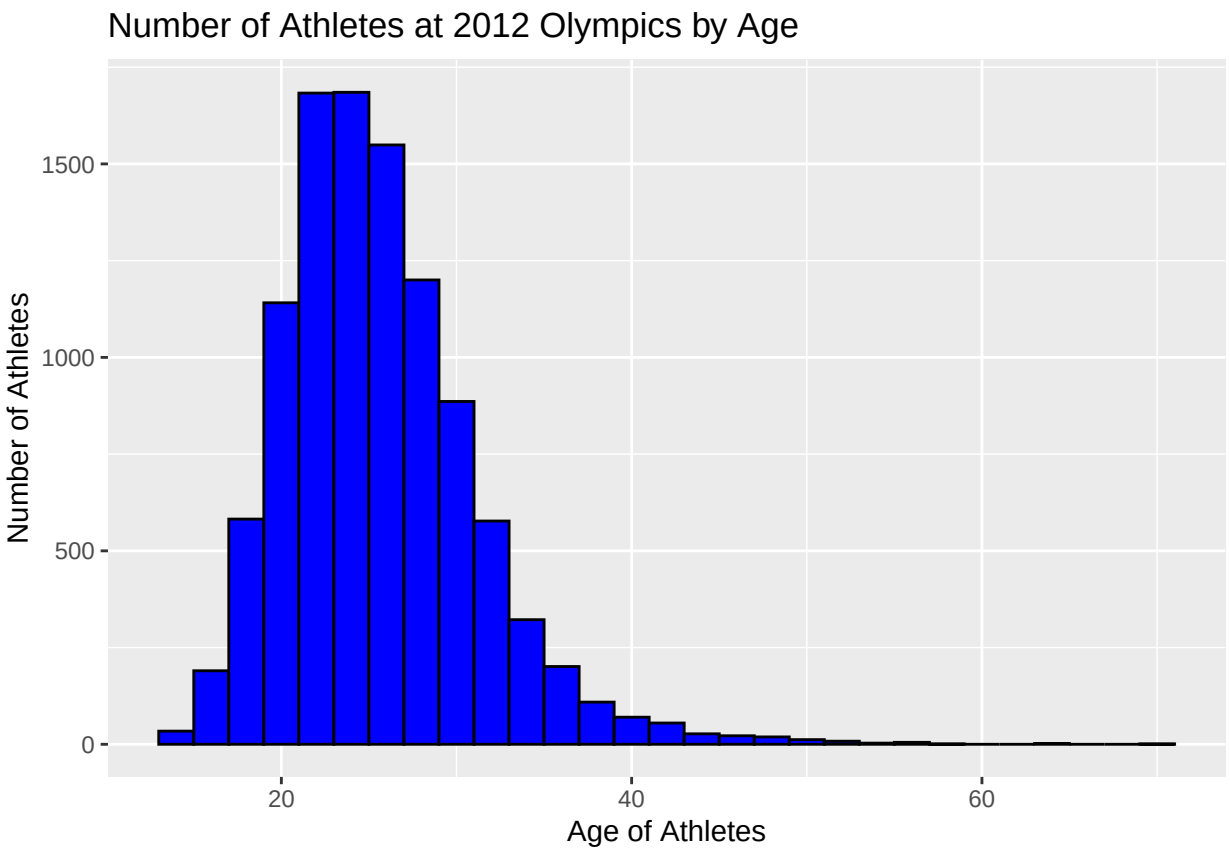
(b) Provide a small introduction of 1-2 sentences to draw your reader in and explain what you'll be discussing. Be definitive about what your data is, and use hedging to highlight the limitations of the data.

> *This report will provide a general analysis of popular sports such as badminton and weightlifting from the 2012 Olympics. I should address there is some limitations with the data such of that of sample size of each sport being dependent on the athletes present.*

(c) Provide 1-2 clearly titled and labeled figures addressing interesting features of the 2012 Olympic athletes' ages.

```
ggplot(data= oly12, aes(x= Age)) + geom_histogram(colour = "black", fill = "blue") + xlab("Age of Athle
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



(d) Provide one or two clearly labeled summary tables addressing interesting features of the 2012 Olympic athletes' ages.

```
data <- oly12
summarize(data, mean(Age))
```

```
##   mean(Age)
## 1  26.06886
```

(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 the important topic of **plagarism** and as an exercise to start getting into the right referencing habits. Incorporating proper citations and references can be easy and natural, and almost always makes your writing better. It also helps you avoid potentially serious academic integrity violations!

Video Completed.

(f) Describe the interesting features of the 2012 Olympic 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. However, remember that your boss isn't a statistician so you will need to clearly define and explain the vocabulary you use.

Vocabulary:

- Location/Center (mean, median, mode)
- Scale/Spread (range, IQR, var, sd, minimum, maximum)
  - *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)

You may also find the following phrases helpful:

- Cleaning data
- Missing data (`NA`)
- Filtering data (`filter`)
- Selecting data (`select`)
- Sorting data (`arrange`, `desc`)
- Grouping data (`group_by`)
- Selecting a subset of variables (`select`)
- Defining new variables (`mutate`)
- Renaming variables (`rename`)
- Producing new data frames
- Creating summary tables (`summarise`)

> *When we look at the histogram(first graph) we can see that the data's shape is right skewed meaning that most of the data is clustered along the left, making a tail on the right. This means that the general bulk and the median of the graph is towards the left side of the frame as logically most athletes are probably pretty young and in their physical prime in their 20s as represented. After filtering the data, what's interesting however is that the mean (average) age of the athletes is 26 which isn't in the direct middle of the longest bin but slightly to the right of it. This is probably due to there being more older athletes than younger, due to perhaps schooling/time available, pulling the average age up.*

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

> *To recap, the average age of the athletes is around 26 for all sports; however this finding is affected by the larger number of older athletes than the median, around early twenties, than younger the median. Some would draw conclusions here that the average age for athletes is around their mid-twenties, however, the data collected here is from slightly over a decade ago and on those who attended the summer Olympics in London.*