STA130 Rstudio Homework

Problem Set 3

[Xuanqi Wei] ([1009353209]), 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 in the `oly12` data set loaded above represents the information of each athlete who participated in the 2012 Olymic Games.

(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.

```
oly12 %>% filter(Country == "Canada" | Country == "United States of America") %>% nrow()
```

```
## [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.

```
oly12 %>%
  filter(Sport == "Gymnastics - Artistic" | Sport == "Gymnastics - Rhythmic" |
         Sport == "Diving" | Sport == "Swimming") %>% nrow()
```

```
## [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.

```
allGymnastics <- oly12 %>% c("Gymnastics - Artistic", "Gymnastics - Rhythmic", "Trampoline", "Diving",
filter(oly12, Sport %in% allGymnastics) %>% nrow()
```

```
## [1] 1446
```

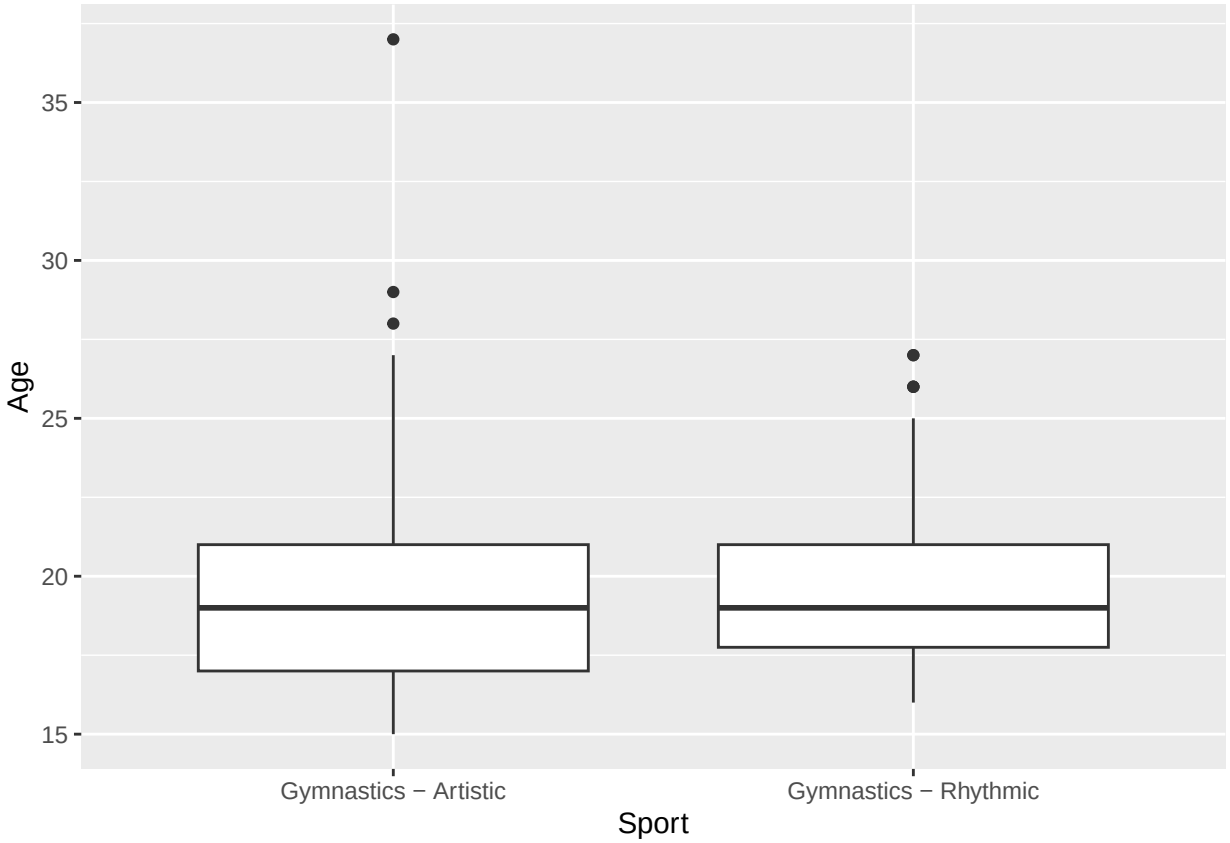
(e) Create the data subset `oly12_FemaleArtisticRhythmicGymnasts` that contains all female Olympic athletes who competed in artistic gymnastics or rhythmic gymnastics.

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

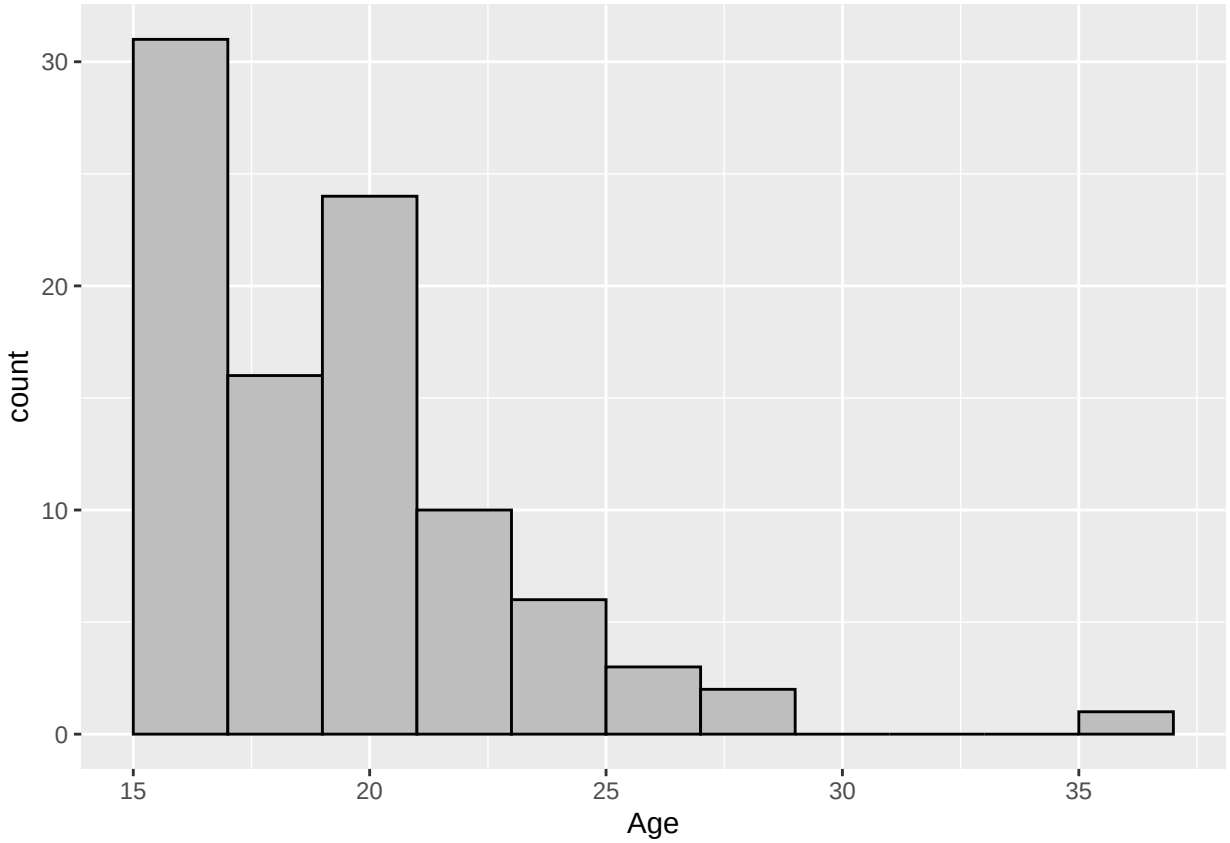
```
oly12_FemaleArtisticRhythmicGymnasts <- oly12 %>%
  filter(Sex=="F") %>%
  filter(Sport == "Gymnastics - Rhythmic" | Sport == "Gymnastics - Artistic")
```

(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.

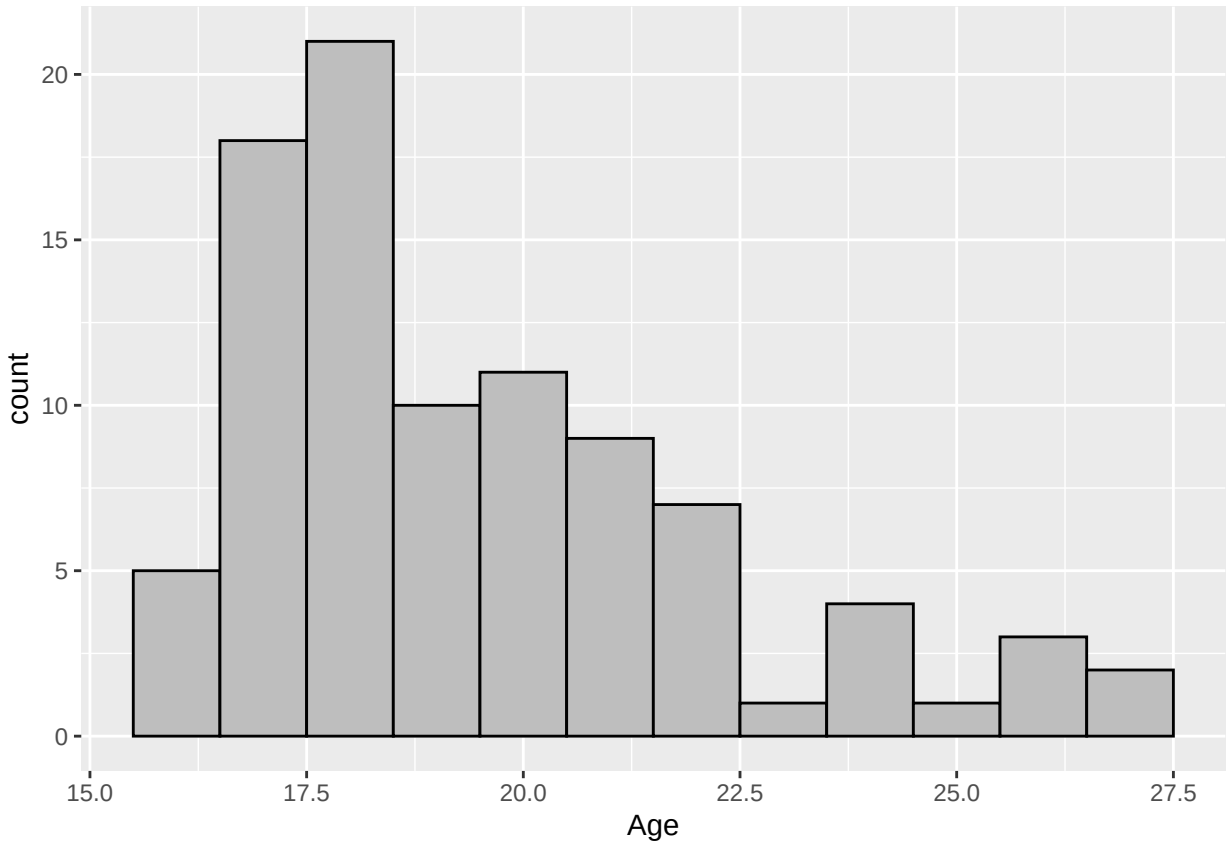
```
ggplot(oly12_FemaleArtisticRhythmicGymnasts, aes(x=Sport, y=Age)) +
  geom_boxplot()
```




```
ggplot(oly12_FemaleArtisticRhythmicGymnasts %>% filter(Sport == "Gymnastics - Artistic"), aes(x=Age)) +
  geom_histogram(bins=12, color="black", fill="gray")
```




```
ggplot(oly12_FemaleArtisticRhythmicGymnasts %>% filter(Sport == "Gymnastics - Rhythmic"), aes(x=Age)) +
  geom_histogram(bins=12, color="black", fill="gray")
```



(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?

   > According to the boxplots and histograms, we can obtain that the age distributions of female rhythimic gymnasts is symmetric, and a little right skew; and, the age distributions of female artistic gymnasts follows right skewed.

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

   > We can obtian that the median age for female whythmic gymnasts and female artistics gymnasts is similar. However, the 25th percentile age of female rhythmic gymnasts is slightly higher than the 25th percentile age of female artistics ymnasts according to the boxplots. And, as for the 75th percentile of ages of female rhythmic gymnasts and female artistic gymnasts is similar too.

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

   > I'd like to predict that the standard deviation of ages for female rhythmic gymnasts will be lower than the standard deviation of ages for female artistic gymnasts as the range are smaller for the shythmic gymnast group rather than the artistic gymnasts group and the data are more concreated.

(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.

```
oly12_FemaleArtisticRhythmicGymnasts %>% group_by(Sport) %>%
  summarise(min=min(Age), max=max(Age), mean=mean(Age),
            median=median(Age), sd=sd(Age))
```

```
## # A tibble: 2 x 6
##   Sport                   min   max  mean median    sd
##   <fct>                 <int> <int> <dbl>  <dbl> <dbl>
## 1 Gymnastics - Artistic    15    37  19.7     19  3.66
## 2 Gymnastics - Rhythmic    16    27  19.5     19  2.68
```

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

> Yes, the standard deviation of ages for female rhythmic gymnasts is lower than the standard deviation of ages for female artistic gymnast.

(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.

```
oly12 <- mutate(oly12, medal_points = 3*Gold + 2*Silver + Bronze)
oly12_OneMedalClub <- oly12 %>%
  mutate(total_medals=Gold+Silver+Bronze) %>%
  filter(total_medals==1)
glimpse(oly12_OneMedalClub)
```

```
## Rows: 457
## Columns: 16
## $ Name         <fct> Jennifer Abel, Alaaeldin Abouelkassem, Chantal Achterberg~
## $ Country      <fct> "Canada", "Egypt", "Netherlands", "Germany", "Great Brita~
## $ Age          <int> 20, 21, 27, 29, 23, 20, 21, 23, 41, 37, 26, 32, 21, 38, 3~
## $ Height       <dbl> 1.60, 1.88, 1.71, 1.89, 1.79, 1.58, 1.78, 1.83, 1.78, 1.6~
## $ Weight       <int> 62, 82, 72, 90, 70, NA, 78, 80, 70, 55, 70, 52, 64, 58, 5~
## $ Sex          <fct> F, M, F, M, F, F, F, M, M, F, F, F, F, F, F, F, M, F, F, ~
## $ DOB          <date> NA, NA, NA, 1983-05-01, NA, NA, 1991-03-08, NA, NA, 1974~
## $ PlaceOB      <fct> "Montreal (CAN)", "", "", "", "Mansfield (GBR)", "Tula (R~
## $ Gold         <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, ~
## $ Silver       <int> 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, ~
## $ Bronze       <int> 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, ~
## $ Total        <int> 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", "Gym~
## $ Event        <fct> "Women's 3m Springboard, Women's Synchronised 3m Springbo~
## $ medal_points <dbl> 1, 2, 1, 3, 1, 2, 1, 1, 1, 2, 1, 3, 3, 3, 2, 1, 1, 2, 2, ~
## $ total_medals <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
```

(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)) %>%
```

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

```
##                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) %>%
  head(6) %>%
  select(Name, Age, Sport)
```

```
##                       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 > 0) %>%
  arrange(Age) %>%
  head(6) %>%
  select(Name, Age, Sport, Event)
```

```
##                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 > 0) %>%
  arrange(desc(Age)) %>%
  head(6) %>%
```

```
  select(Name, Age, Sport, Event)
```

```
##                 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.

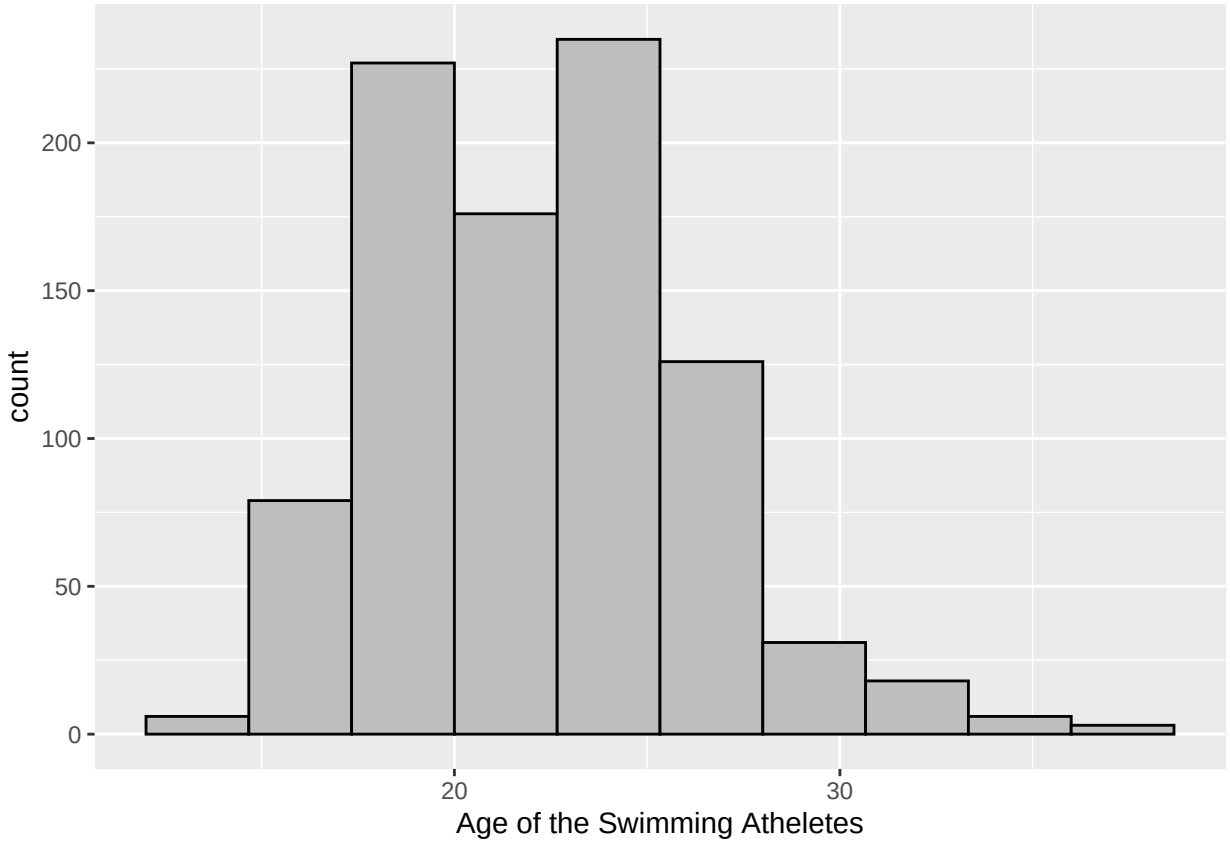
> I've watched the video.

(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.

> According to the data set of 2012 Summer Olympic Games, we will intorduce the possible relationship between athletes age and the sports' competition they attended by comparing two histrograms.

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

```
ggplot(oly12 %>% filter(Sport == "Swimming"), aes(x = Age)) + geom_histogram(bins=10, color="black", fi
```



A histogram can represent the age of the swimming athletes. The age of attending the Swimming competition is centralized at from 18 to 25 years old.

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

```
oly12 %>% filter(Sport == "Swimming") %>% group_by(Sport) %>%
  summarise(min=min(Age), max=max(Age), mean=mean(Age),
            median=median(Age), sd=sd(Age))
```

```
## # A tibble: 1 x 6
##   Sport      min   max  mean median    sd
##   <fct>    <int> <int> <dbl>  <int> <dbl>
## 1 Swimming    13    37  22.4     22  3.98
```

(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!

Understood!

(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`)

According to the data set of 2012 Summer Olympics, we will discuss the possible relationship between athletes age and the swimming competition the atheletes chose to compete at. It's apparent that the graph is right skewed and unimodal, meaning most players are under 30, exspecially between 18 and 25 years old. The mode of the ages for swimming atheletes that attended the 2012 Olympic Games is around 24 years old, which means, among all ages, the athelete who aged around 24 has the greatest number of attendence in the swimming competition according to the graph in part (c).

(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 summarize, most of the atheletes that attended in the swimming competition in 2012 Olymic Games is aged between 18 and 25, which is understandable for this period is the most energetic period in the common sense.