STA130H1F – Fall 2019

Week 2 Practice Problems Answers

N. Moon and B. White [Xinyi (Lisa) Chen 1005825136]

Instructions

How do I hand in these problems for the September 19th 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/115817/assignments/198292) by 11:59PM, on September 19th. Late problem sets or problems submitted another way (e.g., by email) are *not* accepted.

What should I bring to tutorial on September 20?

R output (e.g., plots) for Questions 1 AND 2. You can either bring a hardcopy or bring your laptop with the output.

Tutorial Grading

Tutorial grades will be assigned according to the following marking scheme.

	Mark
Completion of required problems (due on Quercus the day before your tutorial)	1
Attendance for the entire tutorial	1
In-class exercises	4
Total	6

Practice Problems

[Question 1] Recall the `AutoClaims` dataset that we explored this week in class.

```
library(insuranceData)
data(AutoClaims)
glimpse(AutoClaims)
```

```
## Observations: 6,773
## Variables: 5
## $ STATE  <fct> STATE 14, STATE 15, STATE 15, STATE 15, STATE 15, STATE...
## $ CLASS  <fct> C6 , C6 , C11, F6 , F6 , F6 , C11, C6 , C11, C11, C6 , ...
## $ GENDER <fct> M, M, M, F, M, M, M, M, M, M, M, M, M, M, F, F, F, M, F...
## $ AGE    <int> 97, 96, 95, 95, 95, 95, 94, 94, 93, 93, 93, 93, 92, 92,...
## $ PAID   <dbl> 1134.44, 3761.24, 7842.31, 2384.67, 650.00, 391.12, 377...
```

(a) What R data type describes the claim payment data stored in `PAID`?

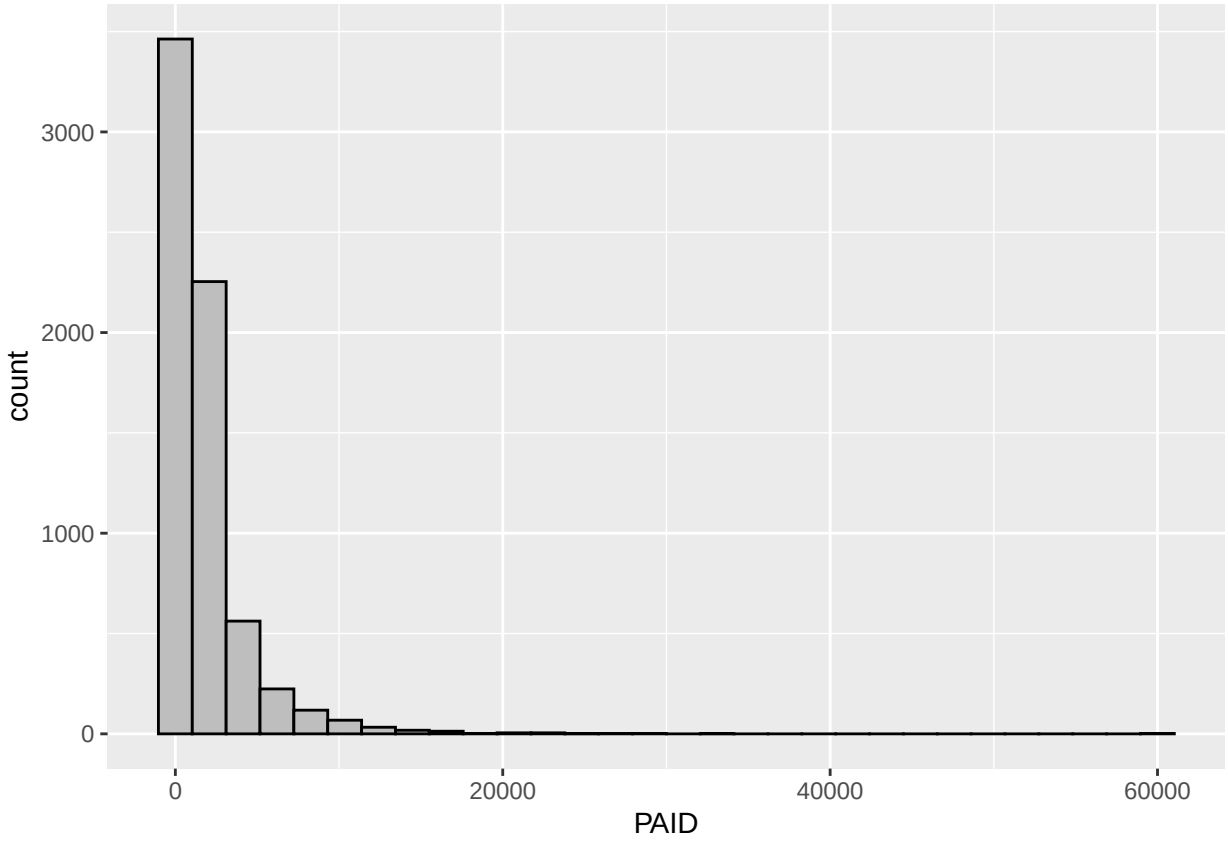
```
# Type your code here
# Note: You can add a new R code chunk by clicking on the Insert button above and choose R
```

The data type for the claim payment data stored in 'PAID' is double.

(b) Suppose we are interested in the distribution of the claim payments. Consider each of the following graphical summaries. If it will show the distribution of claims payments, use R to produce the graph (note: you can click on the Insert button above and choose R to add an R chunk where you can produce your graph). If the graph will not show the distribution of claims payments, do not produce a graph and instead explain why that summary would not be appropriate visualization of the distribution of claims payments.

(i) Histogram

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



(ii) Bar Plot

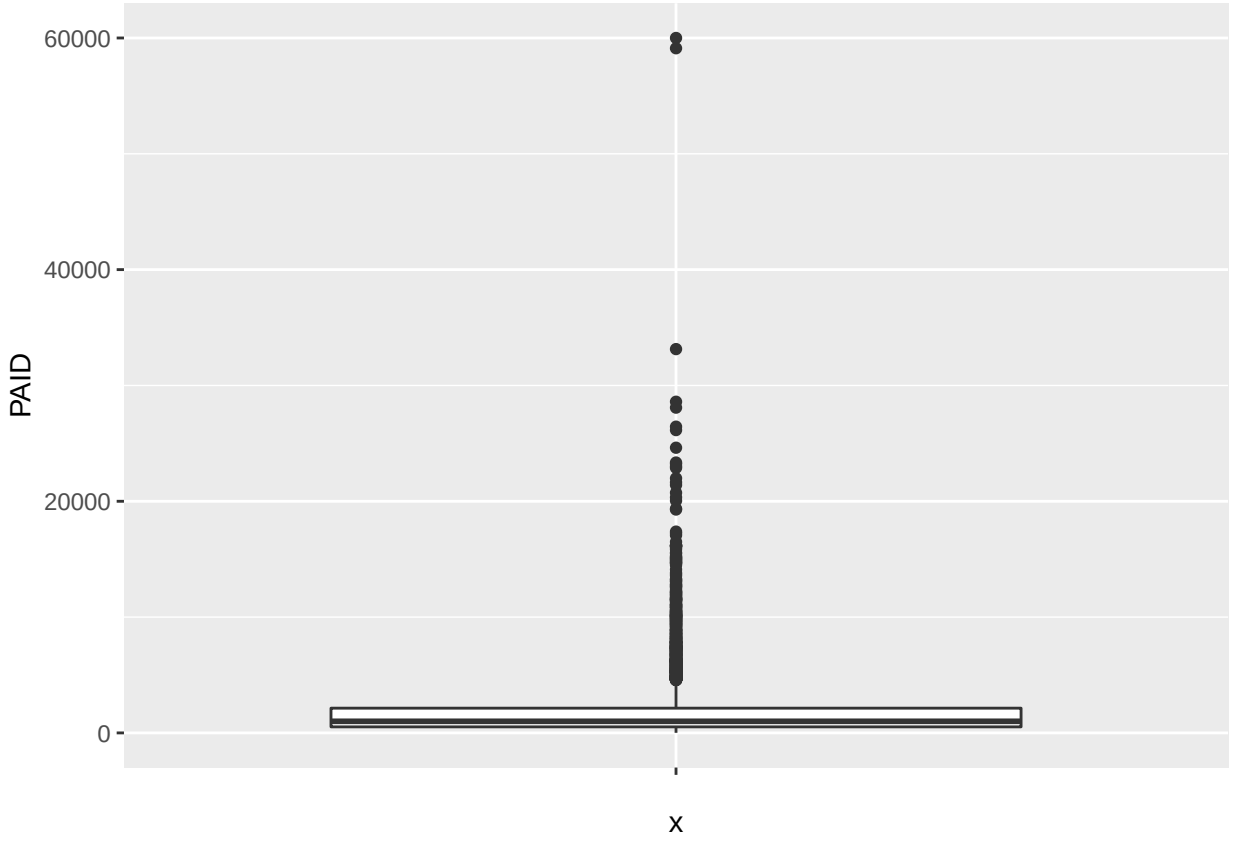
A bar plot shows the distribution of a categorical variable. The variable 'PAID' however, is a numerical variable, so it is not appropriate to show the distribution of the claim payments with a bar plot.

(iii) Scatterplot

The scatterplot shows the association between two variables. We cannot use a scatterplot to show the distribution of one variable– claim payments.

(iv) Boxplot

```
ggplot(data=AutoClaims,
       aes(x=" ", y=PAID)) +
  geom_boxplot()
```



(c) Answer the following questions based on only on the graphs you produced in part (b).

(i) Make a prediction about how the mean and median of claim payments compare. Refer to appropriate graph(s) in part (b) to justify your answer.

The mean will be higher than the median of the claim payments. This is because from the histogram, we can tell that the distribution of the claim payments is right skewed, meaning that there are some extreme values for claim payments that are much larger than the rest of the data. These values will likely pull the mean up and result in a higher mean than the median.

(ii) Would the standard deviation or interquartile range be a more appropriate measure of spread (or variation) in claim payments? Refer to appropriate graph(s) in part (b) to justify your answer.

As we can see in the histogram, the graph is right skewed. Since the standard deviation measurers the spread of the data around the mean and the mean is pulled much higher, it would not be an accurate representation of how spread most of the data truly is. Instead, the interquartile range would be a more appropriate measure of spread in claim payments since it shows the spread of the central data points without getting influenced by the extreme values.

(iii) Suppose we were interested in the mode of the claims payments distribution. Which of the graphs you produced in (b) would be most useful? Justify your answer and refer to the most useful graph to comment on the mode of paid claims.

The histogram would be the most useful for determining the mode of the claims payments distribution. This is because with a histrogram, you can easily tell how frequent a numerical value apprears in a vector. From the histogram in (b), we can tell that highest bin (value with most frequency) is 0. This means that most people had claim payments of $0 on average.

(d) In class we talked about the `summarise` function to compute numerical summaries of a variable. We looked at functions including `mean()`, `median()`, `n()`, `min()`, `max()` and `sd()`. There is also a built-in R function to compute quartiles: `quantile(data,percentile)`. Here is an example:

```
# For instance if you wanted to compute the first quartile of the claim payments
# in this data set, you would use the following command:

summarise(AutoClaims, Q1=quantile(PAID, 0.25))
```

```
##       Q1
## 1 523.73
```

```
# This would also work:

quantile(AutoClaims$PAID, 0.25)
```

```
##    25%
## 523.73
```

Use the summarise function to display the number of claim payments, the mimimum, first quartile, median, 2nd quartile and maximum, mean and standard deviation of the claim payments in this data set. What is an advantage of this summary over the graphs you produced in part (b)?

```
summarise(AutoClaims,
          n=n(),
          min = min(PAID),
          Q1=quantile(PAID, 0.25),
          median = median(PAID),
          Q2=quantile(PAID, 0.50),
          max = max(PAID),
```

```
          mean = mean(PAID),
          sd = sd(PAID))
```

```
##      n min     Q1 median     Q2   max     mean       sd
## 1 6773 9.5 523.73 1001.7 1001.7 60000 1853.035 2646.909
```

The advantage of this summary is that it gives you the precise values of each important numerical value, whereas with graphs you can only see the general shape and spread of the data.

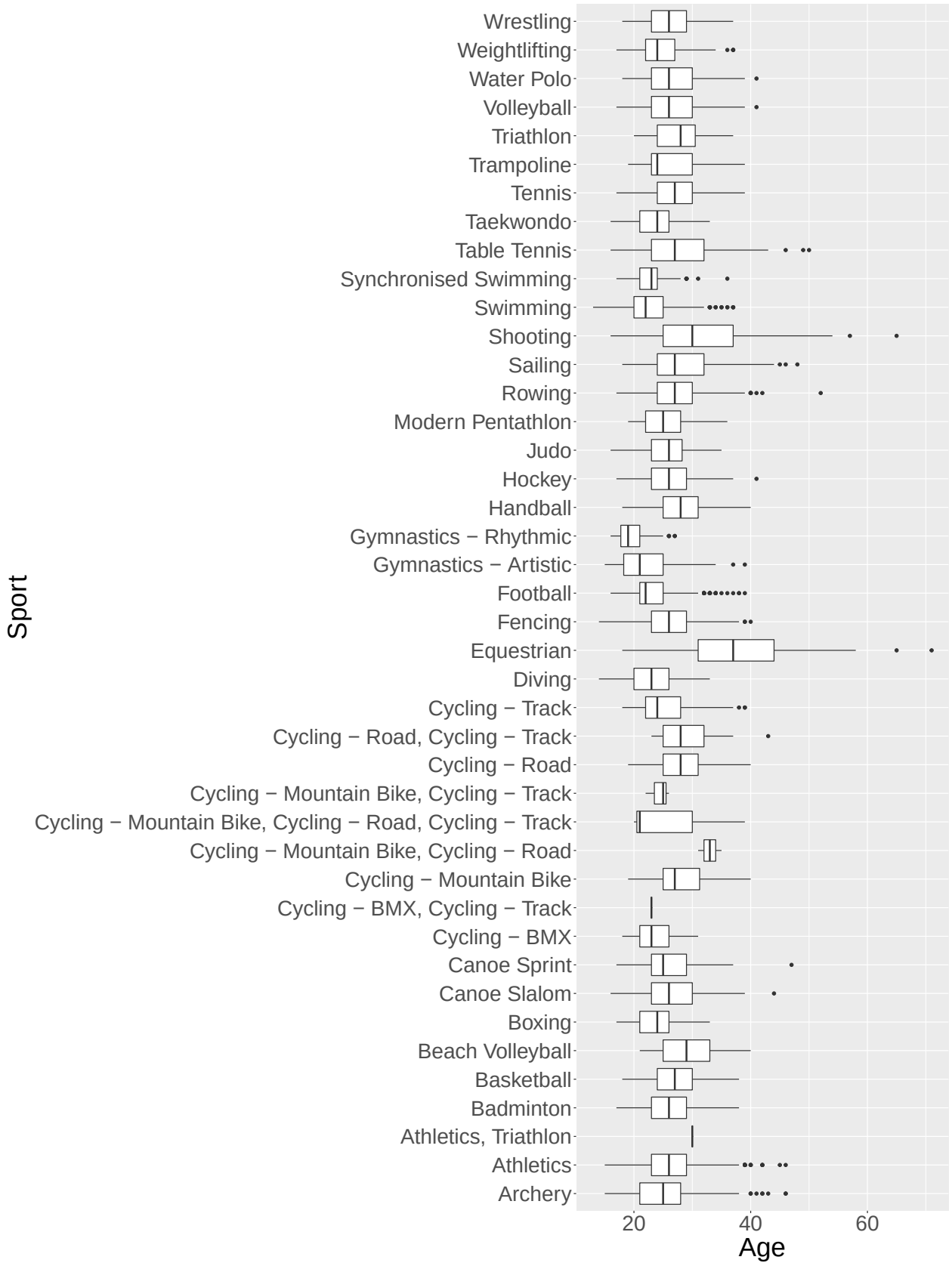
[Question 2] In this question, you'll again use the `oly12` data from last week, which contains information about all athletes who participated in the 2012 Olympics in London. Recall that the `oly12` dataset is part of the `VGAMdata` package, which you can load by running the first R code chunk at the top of this document. Remember: You need to install the packages before you can load them. We pre-installed the packages you need in the RStudio Cloud version of these problems. If you are working on your own version of RStudio and the above R code chunk does not work, you need to use `install.packages()` first.

(a) Create a series of boxplots to visualize the distribution of athletes' ages competing in each sport. Make sure that the Sport names can be read clearly.

```
glimpse(oly12)
```

```
## Observations: 10,384
## Variables: 14
## $ Name    <fct> Lamusi A, A G Kruger, Jamale Aarrass, Abdelhak Aatakni...
## $ Country <fct> People's Republic of China, United States of America, ...
## $ Age     <int> 23, 33, 30, 24, 26, 27, 30, 23, 27, 19, 37, 28, 28, 28...
## $ Height  <dbl> 1.70, 1.93, 1.87, NA, 1.78, 1.82, 1.82, 1.87, 1.90, 1....
## $ Weight  <int> 60, 125, 76, NA, 85, 80, 73, 75, 80, NA, NA, NA, 60, 6...
## $ Sex     <fct> M, M, M, M, F, M, F, M, M, M, M, M, F, F, M, F, M, M, ...
## $ DOB     <date> 1989-02-06, NA, NA, 1988-09-02, NA, 1984-06-09, NA, 1...
## $ PlaceOB <fct> NEIMONGGOL (CHN), Sheldon (USA), BEZONS (FRA), AIN SEB...
## $ Gold    <int> 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, ...
## $ Bronze  <int> 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, ...
## $ Sport   <fct> Judo, Athletics, Athletics, Boxing, Athletics, Handbal...
## $ Event   <fct> "Men's -60kg", "Men's Hammer Throw", "Men's 1500m", "M...
```

```
ggplot(data=oly12,
       aes(x= Sport, y=Age)) +
  geom_boxplot()+
  coord_flip()+
  theme(text =element_text(size=30))
```


Sport
Wrestling
Weightlifting
Water Polo
Volleyball
Triathlon
Trampoline
Tennis
Taekwondo
Table Tennis
Synchronised Swimming
Swimming
Shooting
Sailing
Rowing
Modern Pentathlon
Judo
Hockey
Handball
Gymnastics − Rhythmic
Gymnastics − Artistic
Football
Fencing
Equestrian
Diving
Cycling − Track
Cycling − Road, Cycling − Track
Cycling − Road
Cycling − Mountain Bike, Cycling − Track
Cycling − Mountain Bike, Cycling − Road, Cycling − Track
Cycling − Mountain Bike, Cycling − Road
Cycling − Mountain Bike
Cycling − BMX, Cycling − Track
Cycling − BMX
Canoe Sprint
Canoe Slalom
Boxing
Beach Volleyball
Basketball
Badminton
Athletics, Triathlon
Athletics
Archery
20
40
60
Age

(b) For each of the following questions, write one or two sentences either answering the question or explaining why you cannot answer the question, based *only* on your output from part (a).

(i) In which sport did the oldest athlete compete?

Since the data furtherest to the right across the Age variable corresponds to an outlier in the category Equestrian, the oldest athlete competed in the sport Equestrian.

(ii) How old was the second oldest athlete at the 2012 olympics?

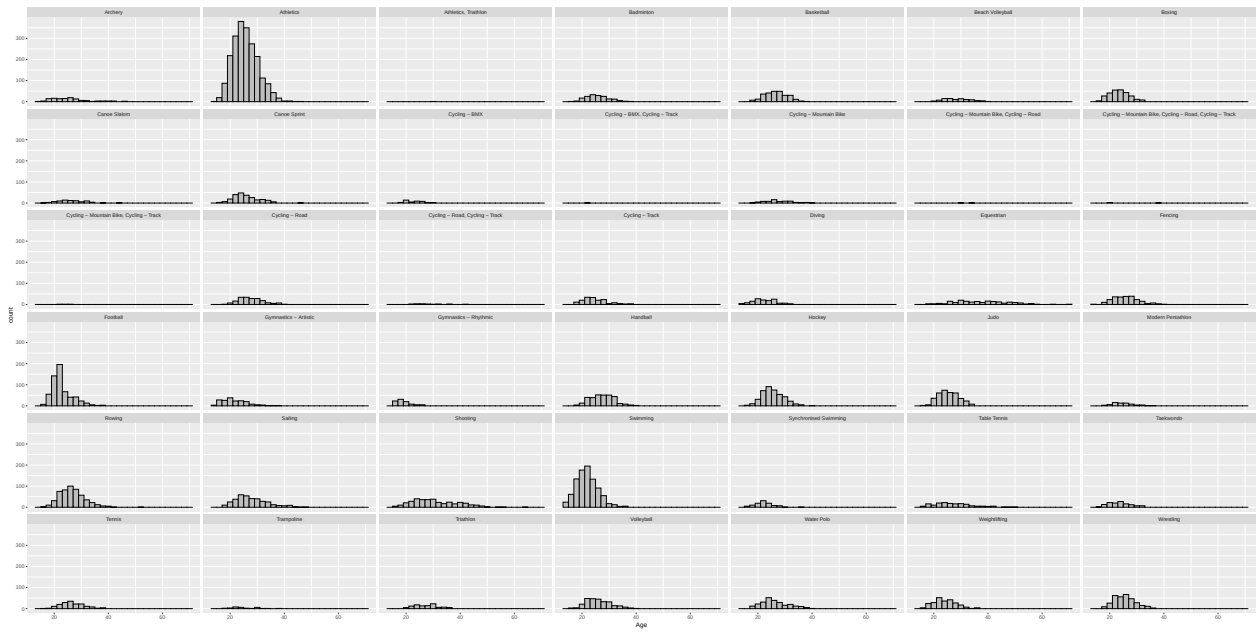
It is hard to determine the second furthest point across the x-axis's precise corresponding age since there is a wide range between each unit on the x-axis(20 years old).

(iii) In which sport did the largest number of athletes compete?

The box plot only shows the spread of the age variable data for each sport. It does not tell you how many data points there are in one sport, so I cannot determine which sport the largest number of athletes compete in.

(c) Create a series of histograms to visualize the distribution of athletes' ages competing in each sport. Hint: Use `facet_wrap()` to create a separate histogram for each sport

```
ggplot(data=oly12) +
       aes(x=Age)+
      geom_histogram(colour = "black",
                     fill = "grey")+
      facet_wrap(~Sport)
```



(d) Answer the following questions based *only* on your output from part (c)

(i) Based only on your histograms, can you easily identify the sport in which the oldest athlete competed at these Olympic Games?

No. Based on the histograms, you cannot see with bare eyes which sport contains the furthest bin to the right on the x-axis with at least one athlete in it.

(ii) Based on these histograms, can you tell if there were more athletes competing in Rowing or Boxing events?

From the histograms, we can see when comparing the Rowing event and the Boxing event, we can see the general shape of the distribution for Rowing is slightly higher than the general shape for Boxing.This implies that we can estimate that there are more athletes in rowing than in boxing.

(d) Create a summary table calculating the minimum, mean, median, and maximum age of athletes for each sport. Save your summary table by giving it a name so that you can click on it in the top right panel and view it as a spreadsheet.

```
oly12_sport <- group_by(oly12, Sport)
athlete_Age <- summarise(oly12_sport,
                         min=min(Age),
                         mean=mean(Age),
                         median= median(Age),
                         max=max(Age))
```

(e) View the summary table you created in (e) as a spreadsheet (by clicking on its name in the top right panel of your RStudio window) and answer the following questions.

(i) How old was the oldest athlete? How old was the youngest athlete?

The oldest athlete was 71 years old. The youngest athelete was 13 years old.

(ii) In which sport is the mean age of athletes the lowest? If you only looked at the boxplots or histograms, could you answer this question? Why or why not?

In Gymnastics - Rhythmic, the mean age of athletes is the lowest. If you only looked at the boxplots or histograms, you would not be able to determine this. This is because boxplots only show the value of medians, not means, and histograms would only show the approximate mean value for a graph. Neither of them compare the precise mean values of ages between sport events.

[Question 3] The `Galton` data set in the `mosaic` library contains data from Francis Galton in the 1880s.

```
glimpse(Galton)
```

```
## Observations: 898
## Variables: 6
## $ family <fct> 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5...
## $ father <dbl> 78.5, 78.5, 78.5, 78.5, 75.5, 75.5, 75.5, 75.5, 75.0, 7...
## $ mother <dbl> 67.0, 67.0, 67.0, 67.0, 66.5, 66.5, 66.5, 66.5, 64.0, 6...
## $ sex    <fct> M, F, F, F, M, M, F, F, M, F, M, M, F, F, F, M, M, M, F...
## $ height <dbl> 73.2, 69.2, 69.0, 69.0, 73.5, 72.5, 65.5, 65.5, 71.0, 6...
## $ nkids  <int> 4, 4, 4, 4, 4, 4, 4, 4, 2, 2, 5, 5, 5, 5, 5, 6, 6, 6, 6...
```

(a) How many children did parents in the `Galton` data set have? Create a data frame called `data` that contains the family id number and the numbers of kids in each family. Note that the number of children is repeated for every member of the families. The data frame you create should not include the repeats.

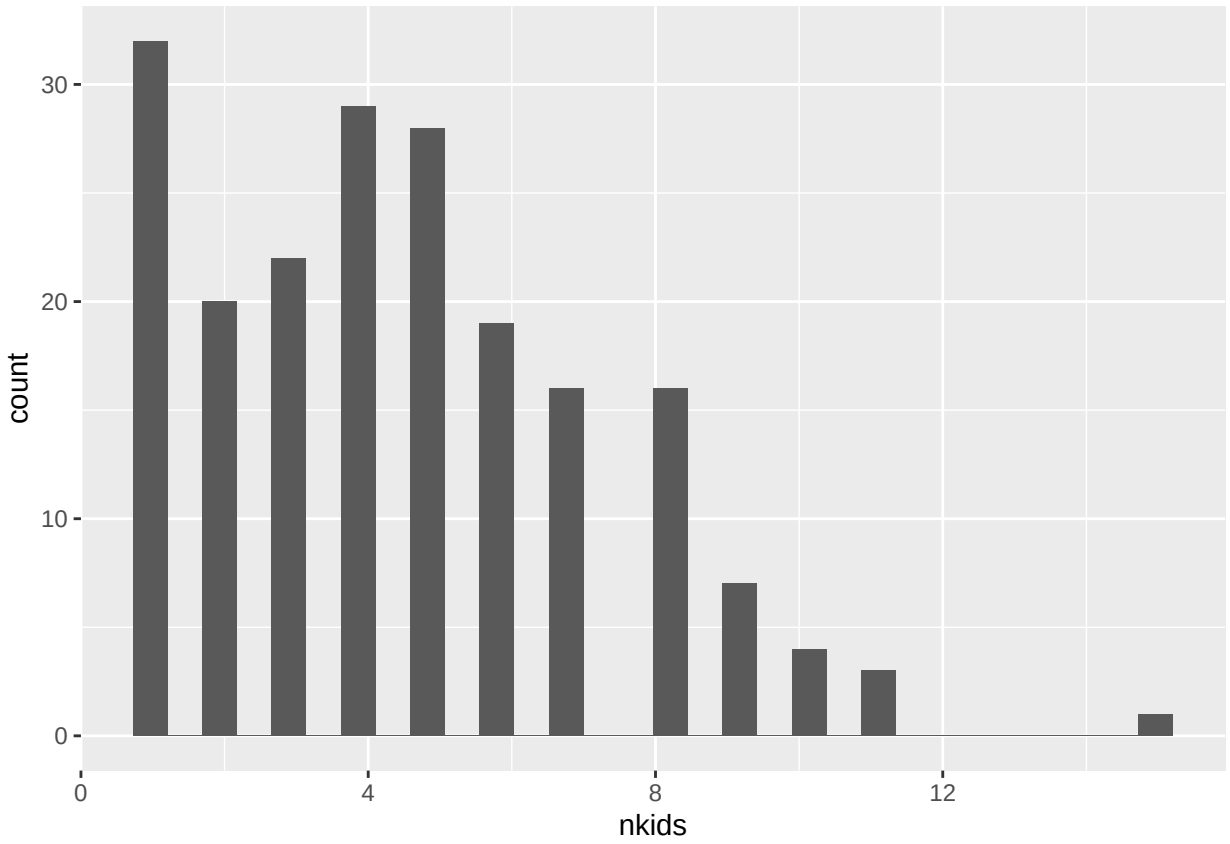
```
data = data.frame(family=c(Galton$family),
                  nkids=c(Galton$nkids))
data= distinct(data)

#or: data <- summarise(group_by(Galton, family), nkids= mean(nkids))
sum(c(data$nkids))
```

```
## [1] 899
```

There are 899 children. ### (b) Graph the distribution of the number of kids in the `Galton` data set families, using the dataframe you created in part (a). Describe (in words) the features of this distribution.

```
ggplot(data, aes(x=nkids)) +
  geom_histogram()
```



The distribution of the number of kids for the families in the "Galton" data set is right skewed. The data has a center at about 4 kids. There is also an unusual data point with over 12 kids.

(c) Just based on the graph you generated in part (b), how do you think the mean and median would compare? Justify your reasoning.

The mean would be slightly higher than the median because the graph is right skewed. In other words, the mean would be pulled higher by the data points by the data point with a large number of kids.

(d) Compute the mean and median of the number of kids in the `Galton` data set families. Does this match what you expected to see in part (e)?

```
summarise(data, mean = mean(nkids),
          median = median(nkids))
```

```
##       mean median
## 1 4.563452      4
```

This matches what I expected to see becuase since the histogram of the distribution of the number of kids is right skewed, the mean should be higher than the median. We can tell from the table that the mean is infact higher than the median.