STA130H1S – Winter 2020

Week 3 Practice Problems - Sample Answers

L. Bolton and N. Moon

Instructions

How do I hand in these problems for the January 23 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/138992/assignments/284427) by 11:59PM, on January 23. Late problem sets or problems submitted another way (e.g., by email) are *not* accepted.

What should I bring to tutorial on January 23?

R output (e.g., plots) for Question 1. 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] 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(VGAMdata)
glimpse(oly12)
```

```
## Observations: 10,384
## Variables: 14
## $ Name    <fct> Lamusi A, A G Kruger, Jamale Aarrass, Abdelhak Aatakni, Maria...
## $ Country <fct> People's Republic of China, United States of America, France,...
## $ Age     <int> 23, 33, 30, 24, 26, 27, 30, 23, 27, 19, 37, 28, 28, 28, 22, 1...
## $ 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, ...
## $ 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-...
## $ PlaceOB <fct> NEIMONGGOL (CHN), Sheldon (USA), BEZONS (FRA), AIN SEBAA (MAR...
## $ 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, Rowi...
## $ Event   <fct> "Men's -60kg", "Men's Hammer Throw", "Men's 1500m", "Men's Li...
```

(a) In class 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.), 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)

In the `oly12` dataset, each row corresponds to one athlete who participated in the 2012 Olympic Games.

(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 class.

```
# Type your code here

# Using filter to keep only canadian athletes,
# then glimpse to view the number of observations
oly12 %>% filter(Country == "Canada") %>%
  glimpse()
```

```
## Observations: 274
## Variables: 14
## $ Name    <fct> Jennifer Abel, Natalie Achonwa, Mohammed Ahmed, Dylan Armstro...
## $ Country <fct> Canada, Canada, Canada, Canada, Canada, Canada, Canada, Canad...
## $ Age     <int> 20, 19, 21, 31, 28, 24, 20, 28, 23, 22, 21, 56, 29, 24, 23, 2...
## $ Height  <dbl> 1.60, 1.92, 1.90, 1.93, 1.85, 1.83, 1.68, 1.86, 1.86, 1.68, 1...
## $ Weight  <int> 62, 83, 60, 139, 82, 78, 150, 90, 80, 58, 75, 78, 98, 48, 69,...
## $ Sex     <fct> F, F, M, M, F, F, M, M, M, F, M, M, M, F, F, F, M, M, F, F, M...
## $ DOB     <date> NA, NA, 1991-05-01, NA, NA, 1988-06-05, 1992-11-03, NA, NA, ...
## $ PlaceOB <fct> Montreal (CAN), , Mogadishu (SOM), Kamloops (CAN), , , Westmo...
## $ 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, 1...
## $ Bronze  <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0...
## $ Total   <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1...
## $ Sport   <fct> Diving, Basketball, Athletics, Athletics, Basketball, Basketb...
## $ Event   <fct> "Women's 3m Springboard, Women's Synchronised 3m Springboard"...
```

```
# Using filter to keep only canadian athletes,
# then count the number of rows in the resulting data frame
oly12 %>% filter(Country == "Canada") %>%
  nrow()
```

```
## [1] 274
```

```
# Use summarise to calculate the number of athletes for each country,
# then filter to keep only the row for Canada
oly12 %>% group_by(Country) %>%
  summarise(team_size = n()) %>%
```

```
  filter(Country=="Canada")
```

```
## # A tibble: 1 x 2
##   Country team_size
##   <fct>       <int>
## 1 Canada        274
```

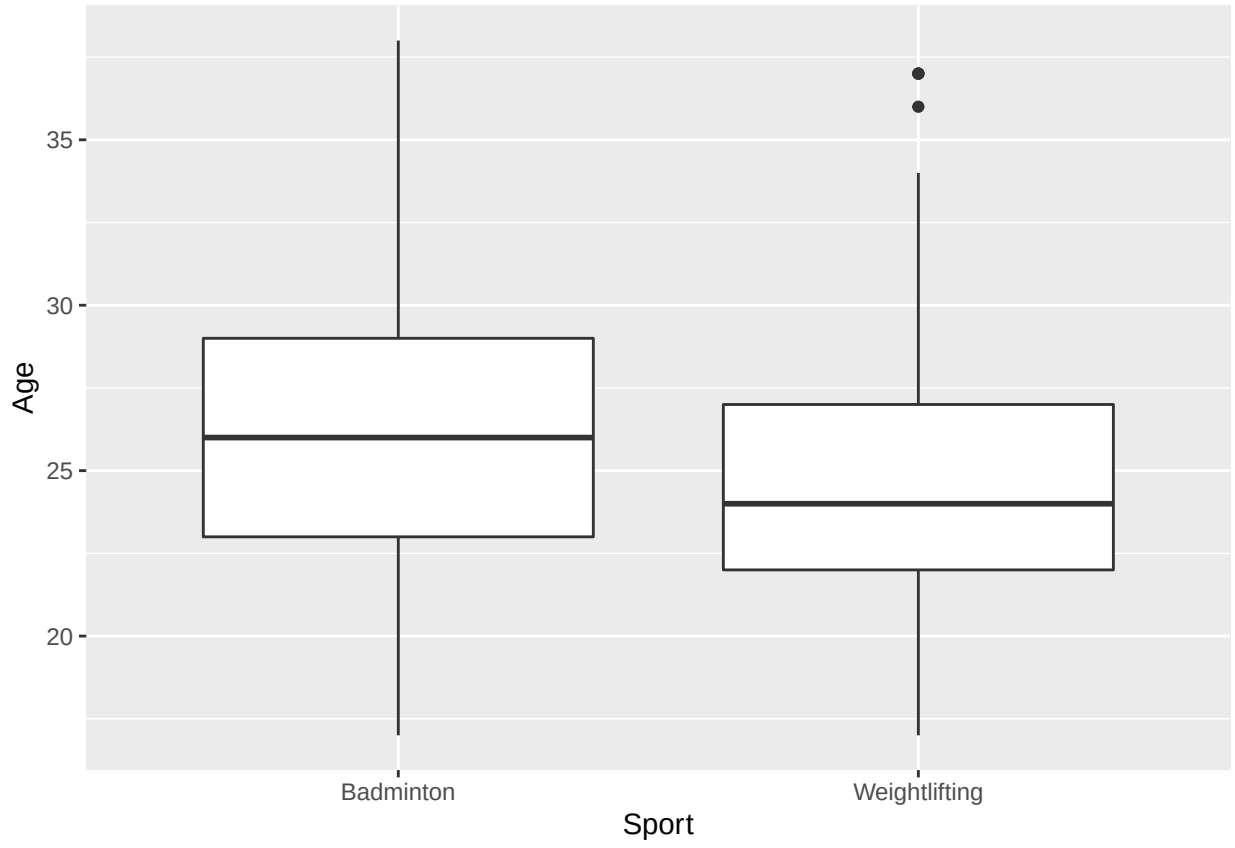
274 athletes represented Canada at the 2012 Olympic Games.

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

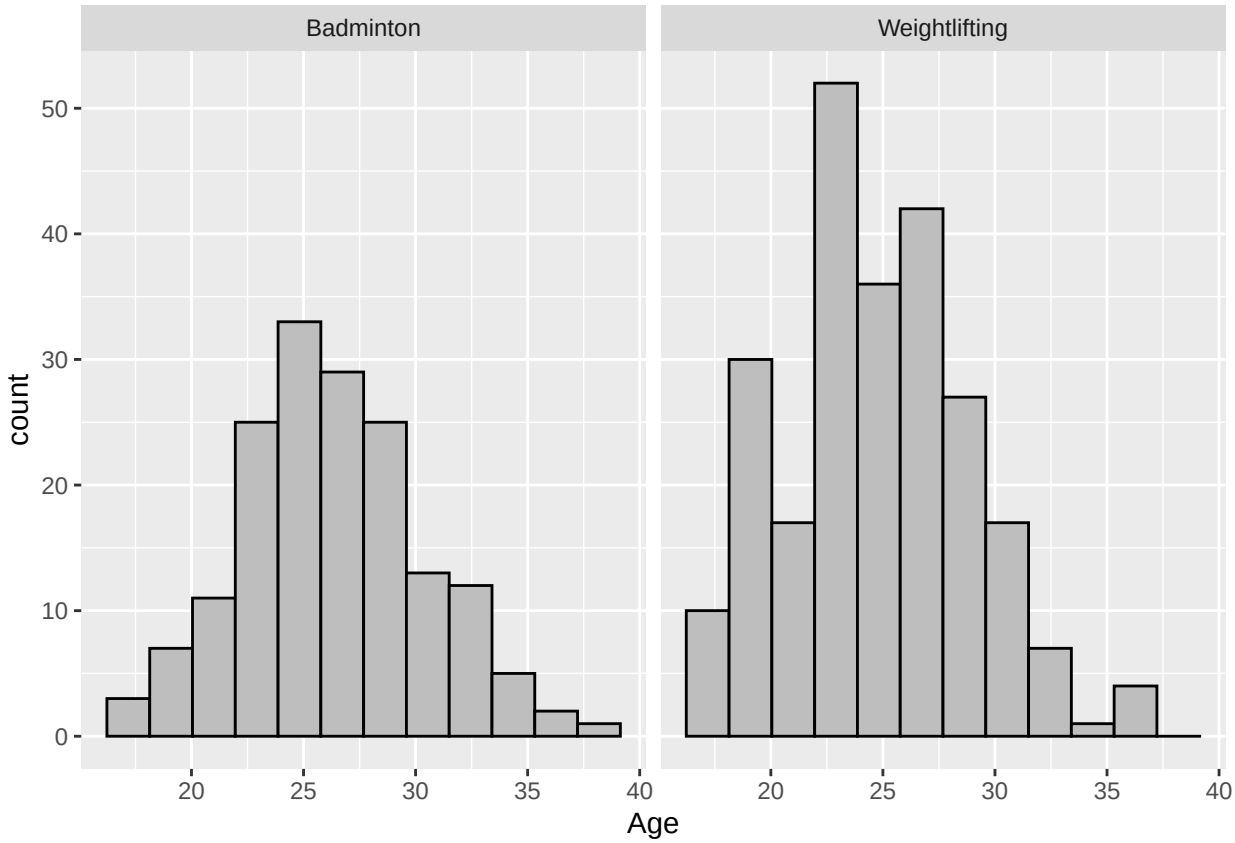
```
oly12_selectedSports <- oly12 %>% filter(Sport == "Weightlifting" | Sport == "Badminton")
```

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

```
oly12_selectedSports %>% ggplot(aes(x=Sport, y=Age)) +
  geom_boxplot()
```




```
oly12_selectedSports %>% ggplot(aes(x=Age)) +
  geom_histogram(bins=12, color="black", fill="gray") + facet_wrap(~Sport)
```



(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?

From the histograms, we can see that the age distribution of badminton players is approximately symmetric but the age distribution of weightlifters is slightly skewed to the right. This can also be seen in the boxplots of the age distributions - in particular, we see there are two outliers in the right tail of the age distribution of weightlifters, corresponding to two weightlifters who are much older than most of the weightlifters.

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

From the boxplots, we can see that the median age of badminton players is higher than the median age of weightlifters (~26 vs ~24).

(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.

I predict that the standard deviatiation of ages for badminton players will be a little bit larger than that of weightlifters since the IQR and whiskers are both a bit longer. However, the range of the age distributions (max - min) are similar for both sports.

(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)

```
oly12_selectedSports %>% 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 Badminton        17    38  26.2     26  4.12
## 2 Weightlifting    17    37  24.6     24  4.06
```

As predicted in (e-iii) the standard deviation of ages is slightly higher for badminton players than for weightlifters (4.12 vs 4.06), but they are very similar.

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

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

```
##                Name Age      Sport                                        Event
## 1   Hiroshi Hoketsu  71 Equestrian                Individual Dressage, WHISPER
## 2 Afanasijs Kuzmins  65   Shooting                 Men's 25m Rapid Fire Pistol
## 3        Ian Millar  65 Equestrian Individual Jumping, Team Jumping, STAR POWER
## 4    Carl Bouckaert  58 Equestrian Individual Eventing, Team Eventing, CYRANO Z
## 5  Andrei Kavalenka  57   Shooting                                    Men's Trap
## 6        Mary Hanna  57 Equestrian Individual Dressage, Team Dressage, SANCETTE
```

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

```
oly12  %>%
  filter(Gold > 0) %>%
  arrange(desc(Age)) %>%
  head() %>%
  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
```

(h) 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 = Gold + Silver + Bronze) %>%
  arrange(desc(total_medals)) %>%
```

```
  head() %>%
  select(Name, Country, Sport, total_medals)
```

```
##              Name                  Country    Sport total_medals
## 1     Ryan Lochte United States of America Swimming            5
## 2   Alicia Coutts                Australia Swimming            4
## 3  Michael Phelps United States of America Swimming            4
## 4 Allison Schmitt United States of America Swimming            4
## 5   Yannick Agnel                   France Swimming            3
## 6  Missy Franklin United States of America Swimming            3
```

[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)
```

```
## Observations: 2,208
## Variables: 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", "Southampto...
## $ Class        <chr> "3", "Crew", "3", "3", "3", "3", "3", "2", "2", "3", "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", "A...
## $ Sex          <chr> "Male", "Male", "Male", "Male", "Female", "Female", "Mal...
## $ Paid         <dbl> 7.550000, NA, 20.250000, 20.250000, 20.250000, 7.650000,...
## $ Ticket_No    <chr> "5547", NA, "CA2673", "CA2673", "CA2673", "348125", "348...
## $ Boat_or_Body <chr> NA, NA, NA, "[190]", "A", "16", "A", NA, "10", "15", "C"...
## $ Job          <chr> "Blacksmith", "Lounge Pantry Steward", "Scholar", "Jewel...
## $ Class_Dept   <chr> "3rd Class Passenger", "Victualling Crew", "3rd Class Pa...
## $ 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`.

```
titanic <- titanic %>%
  select(Name, Survived, Boarded, Class, MWC, Age, Adut_or_Chld, Sex, Paid, Job)
```

(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`.

```
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 survivors (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"),
            proportion=n_died / n)
```

```
## # A tibble: 3 x 4
##   Man_Woman_or_Child     n n_died proportion
##   <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 %>%
  filter(Age >= 18 & Age <= 25) %>%
  summarise(n=n(),
            n_died = sum(Survived=="Dead"),
            proportion=n_died / n)
```

```
## # A tibble: 1 x 3
##       n n_died proportion
##   <int>  <int>      <dbl>
## 1   635    438      0.690
```

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

```
titanic %>%
  filter(Paid > 30) %>%
  group_by(Man_Woman_or_Child) %>%
  summarise(n=n(),
            n_died = sum(Survived=="Dead"),
            proportion=n_died / n)
```

```
## # A tibble: 3 x 4
##   Man_Woman_or_Child     n n_died proportion
##   <chr>              <int>  <int>      <dbl>
## 1 Child                 43     25      0.581
## 2 Man                  149    109      0.732
## 3 Woman                150     12      0.08
```

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

Survival rates on the Titantic were associated with whether the passenger was a man, woman or child and the cost of their ticket. About 24% of all women passengers on the Titantic died Unfortunately men and

children passengers had considerably higher death rates (0.81 and 0.48 respectively). Amoung the passengers who paid more for their tickets, death rates were lower for the adult passengers since only 8% of these women and 73% of these men died, but higher for children (58% of these children died).

(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.

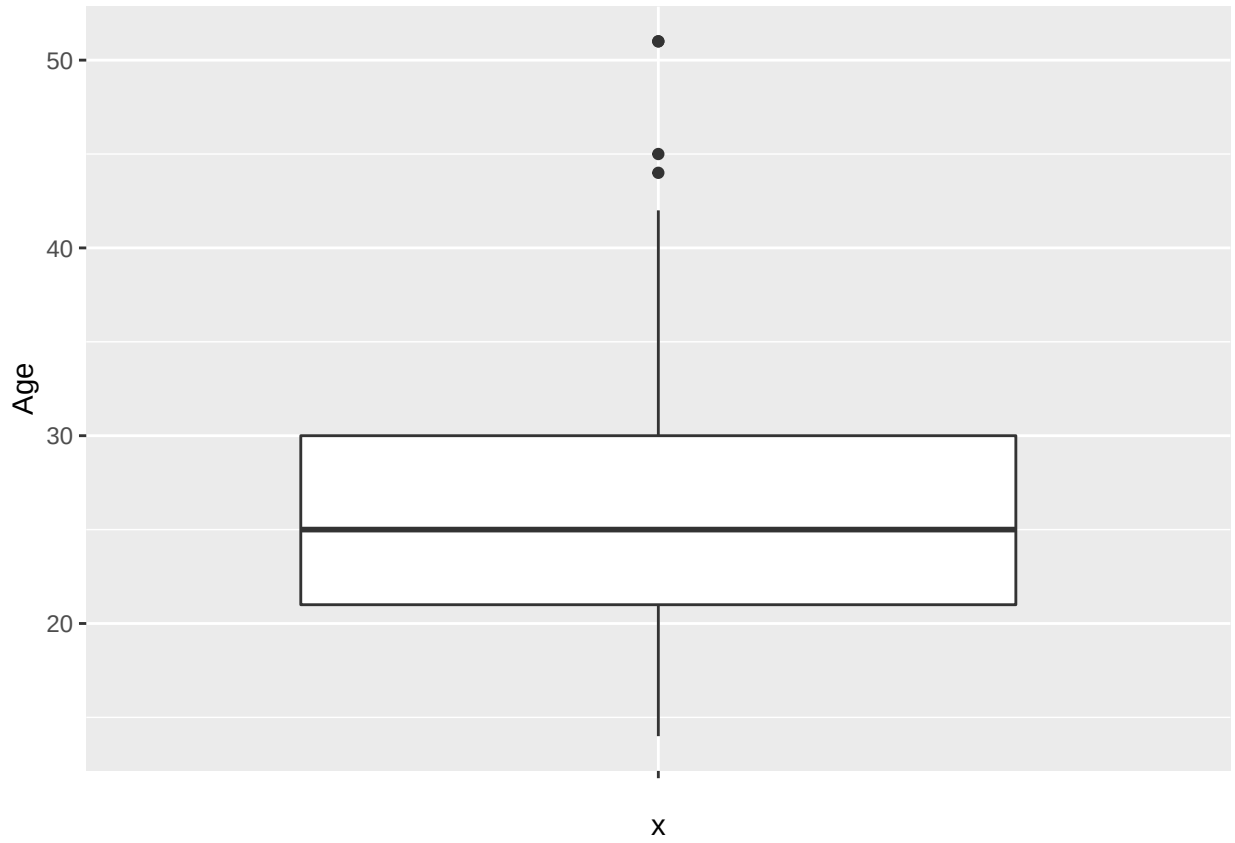
```
titanic %>% group_by(Job) %>%
  summarise(n=n()) %>%
  arrange(desc(n))
```

```
## # 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
```

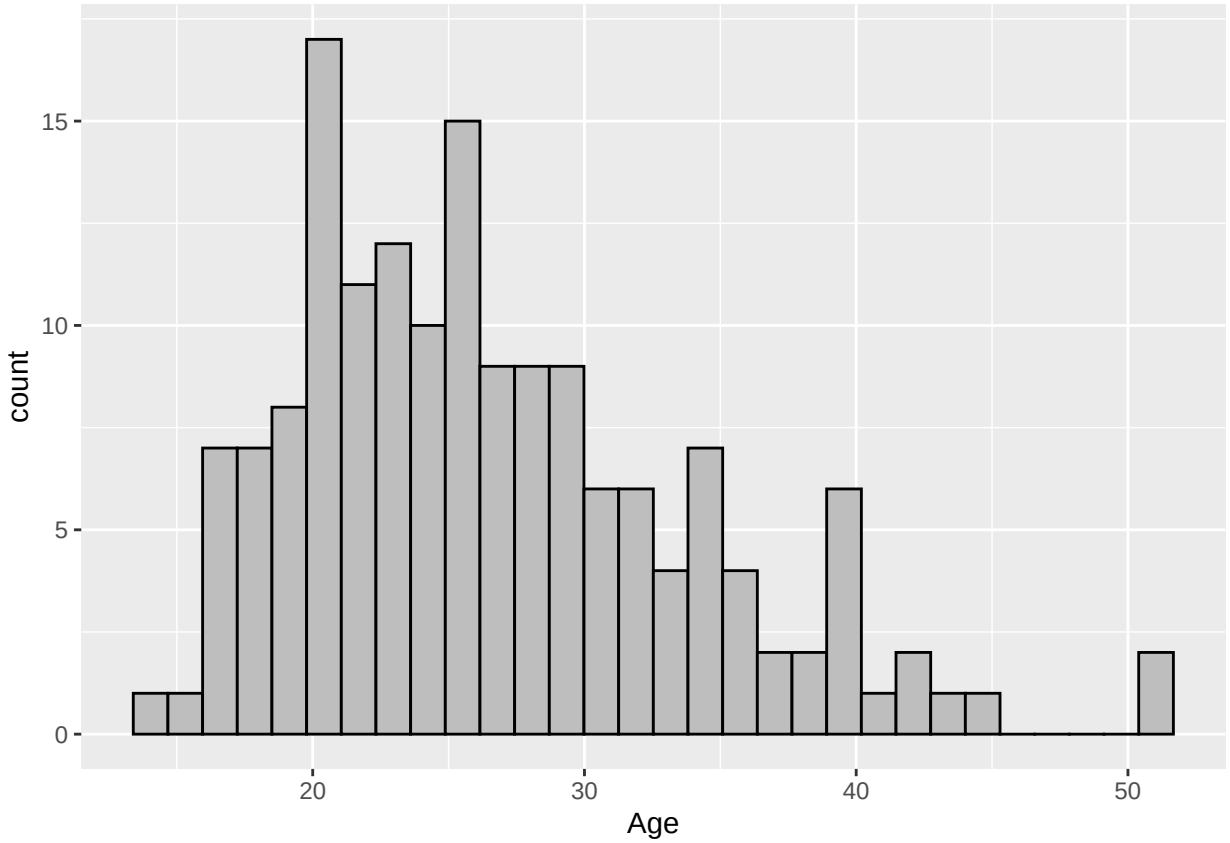
631 of the passengers do not have a job listed (NA). The job recorded for the largest number of passengers is "General Labourer" (162), although there were also 161 firemen.

(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="", y=Age)) + geom_boxplot()
```




```
titanic %>% filter(Job=="General Labourer") %>%
  ggplot(aes(x=Age)) + geom_histogram(bins=30, color="black", fill="gray")
```

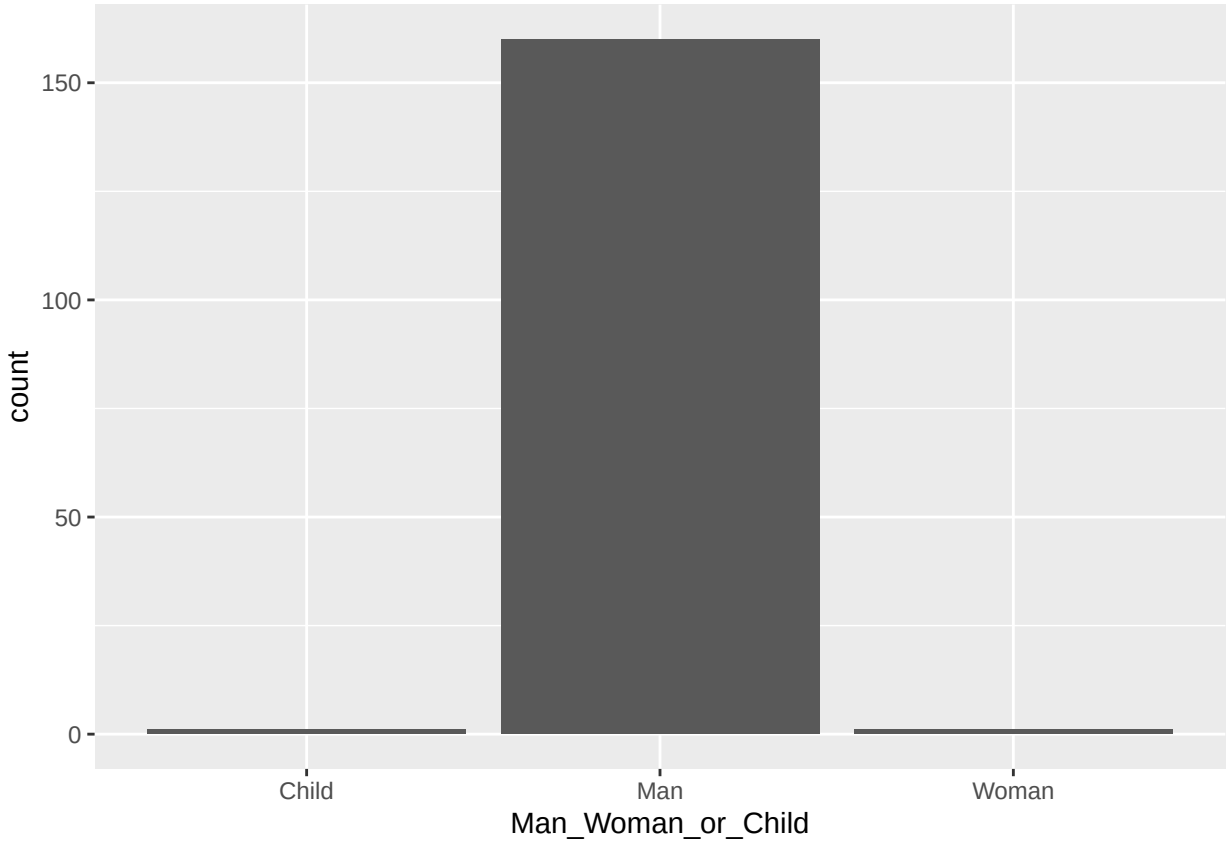


General labourers on the Titanic ranged in age from uner 15 to just over 50. The age distribution is slightly right skewed, with a few outliers in the right tail corresponding to older individuals (over age 43). The median age of general labourers on the Titanic is close to 25 years, with an interquartile range of approximately 9 years (21 to 30 years).

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

```
# there are several ways to do this

titanic %>% filter(Job=="General Labourer") %>%
  ggplot(aes(x=Man_Woman_or_Child)) + geom_bar()
```




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

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

```
titanic %>% filter(Job=="General Labourer" & Sex == "Female")
```

```
## # A tibble: 1 x 10
##   Name  Survived Boarded Class Man_Woman_or_Ch...   Age Adult_or_Child Sex    Paid
##   <chr> <chr>    <chr>   <chr> <chr>            <dbl> <chr>          <chr> <dbl>
## 1 HAAS... Dead     Southa... 3     Woman               24 Adult          Fema...  8.85
## # ... with 1 more variable: Job <chr>
```

Of the 162 general labourers on the Titanic, 160 were men, 1 was a child and 1 was a woman.

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

```
titanic %>%
  arrange(desc(Paid)) %>%
  select(Name, Paid, Survived)
```

```
## # A tibble: 2,208 x 3
```

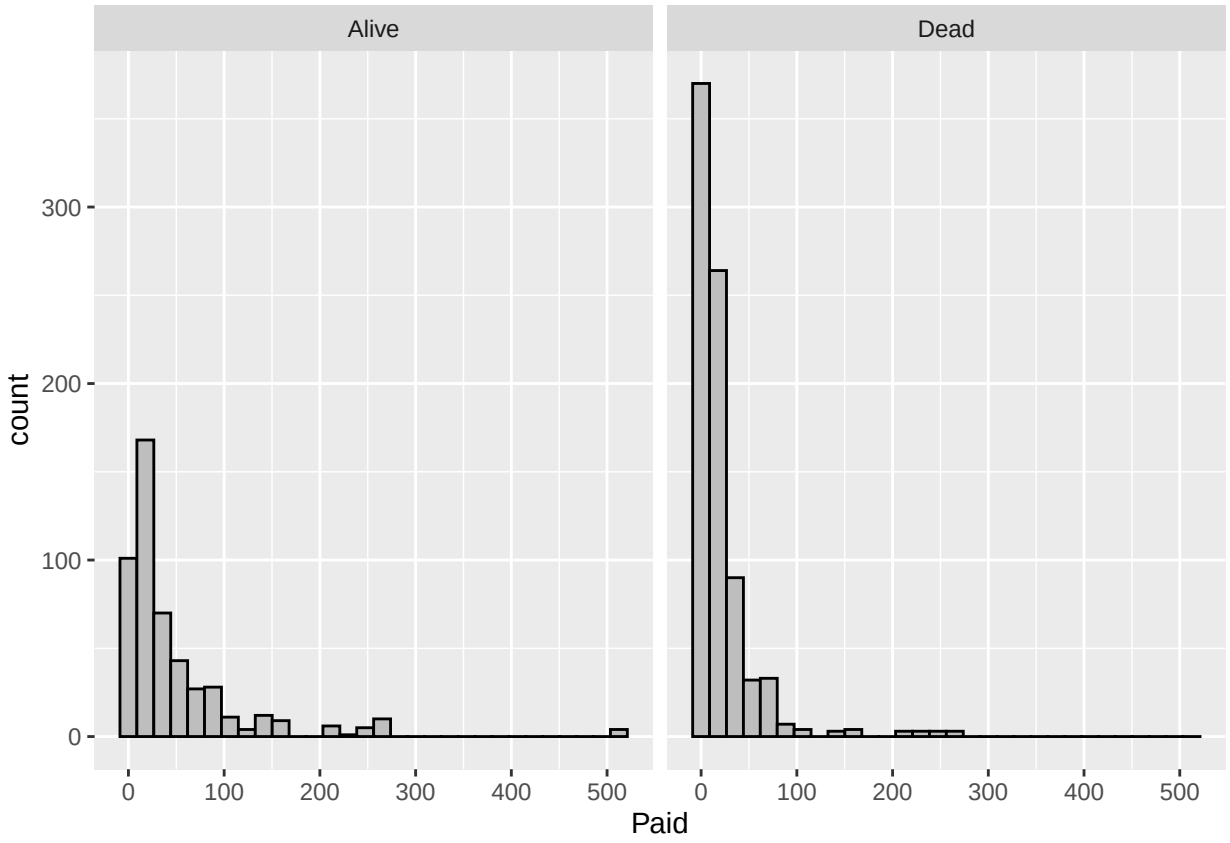
```
##    Name                                Paid Survived
##    <chr>                              <dbl> <chr>
##  1 CARDEZA, Mr Thomas Drake Martinez  512. Alive
##  2 CARDEZA, Mrs Charlotte Wardle      512. Alive
##  3 LESUEUR, Mr Gustave J.             512. Alive
##  4 WARD, Miss Annie Moore             512. Alive
##  5 FORTUNE, Miss Alice Elizabeth      263  Alive
##  6 FORTUNE, Miss Ethel Flora          263  Alive
##  7 FORTUNE, Miss Mabel Helen          263  Alive
##  8 FORTUNE, Mr Charles Alexander      263  Dead
##  9 FORTUNE, Mr Mark                   263  Dead
## 10 FORTUNE, Mrs Mary                  263  Alive
## # ... with 2,198 more rows
```

The most expensive tickets were sold to: - Mr Thomas Drake Martinez CARDEZA - Mrs Charlotte Wardle CARDEZA - Mr Custave J. LESUEUR - Miss Annie Moore WARD All four of these passengers paid 512.32 British pounds for their tickets and they all survived the accident.

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

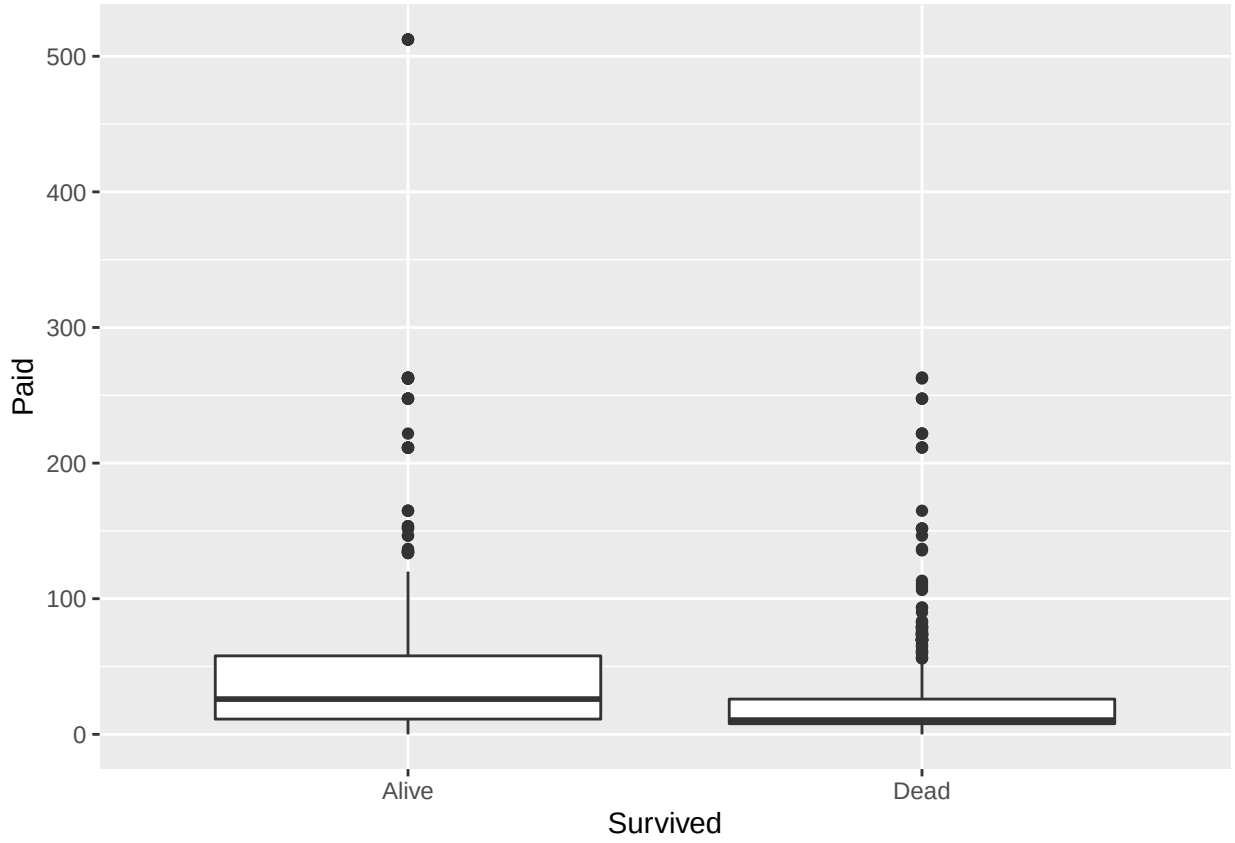
```
titanic %>% ggplot(aes(x=Paid)) +
  geom_histogram(color="black", fill="gray", bins=30) + facet_wrap(~Survived)
```



The distribution of ticket prices is very right-skewed for both the survivors and those who perished; while most of the tickets cost less than 100 pounds, some of the survivors paid over 500 pounds for their tickets. The first bar in the histogram (corresponding to the lowest range of fares) is much taller in the distribution of non-survivors than survivors, so we see that most of the individuals who bought these low-cost tickets did not survive the accident.

(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 %>% ggplot(aes(x=Survived, y=Paid)) + geom_boxplot()
```



Again, we see that both distributions are highly right skewed. From the boxplots, it is clear that the median fare paid by surviving passengers was higher than that paid by the non-survivors and the interquartile range of ticket prices is much wider among survivors than non-survivors (IQR of approximately 50 pounds for survivors and less than 25 pounds for non-survivors). The distribution of ticket prices is more right-skewed among non-survivors than among survivors, as the median appears to be very close to the first quartile.

(iii) Construct a summary table with the minimum, median, mean, and maximum ticket price for survivors and non-survivors. Write 2-3 sentences comparing the two distributions based on this summary table.

```
titanic %>% group_by(Survived) %>%
  summarise(n=n(),  min=min(Paid, na.rm=TRUE),
            first_quartile=quantile(Paid, 0.25, na.rm=TRUE),
            median=median(Paid, na.rm=TRUE), mean=mean(Paid, na.rm=TRUE),
            third_quartile=quantile(Paid, 0.75, na.rm=TRUE),
            max=max(Paid, na.rm=TRUE))
```

```
## # A tibble: 2 x 8
##   Survived     n   min first_quartile median  mean third_quartile   max
##   <chr>    <int> <dbl>          <dbl>  <dbl> <dbl>          <dbl> <dbl>
## 1 Alive      712     0          11.3    26    49.6           57.9  512.
## 2 Dead      1496     0           7.85   10.5  22.9            26    263
```

The minimum ticket price among both survivors and non-survivors is 0, which is strange; more investigation is required to determine whether this is an error in the data or if some passengers in fact received complimentary tickets. From the summary table, we see that the median ticket price among survivors was more than twice as high as the median ticket price among non-survivors.
Among survivors, 75% paid less than 58 pounds for their tickets, while 75% of the non-survivors paid less than 26 pounds. The mean ticket price is also much higher among survivors, but this is pulled up by the particularly high prices paid by a small number of passengers.

As a side note, we can take a closer look at passenger with 0-pound tickets

```
titanic %>% filter(Paid == 0) %>% group_by(Class) %>% summarise(n())
```

```
## # A tibble: 3 x 2
##   Class `n()`
##   <chr> <int>
## 1 1         8
## 2 2        14
## 3 3         6
```

```
titanic %>% filter(Paid == 0) %>% group_by(Boarded) %>% summarise(n())
```

```
## # A tibble: 4 x 2
##   Boarded     `n()`
##   <chr>       <int>
## 1 Belfast         9
## 2 Cherbourg       1
## 3 Queenstown      1
## 4 Southampton    17
```

```
titanic %>% filter(Paid == 0) %>% group_by(Survived) %>% summarise(n())
```

```
## # A tibble: 2 x 2
##   Survived `n()`
##   <chr>    <int>
## 1 Alive        4
## 2 Dead        24
```

There is no obvious pattern connecting individuals with recorded 0-pound tickets. It is not clear whether this is an error or not, but since only 28 out of 2208 observations are affected, these are not expected to have a large impact on the comparison.

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

Histograms: The paired histograms give us a good overall impression of the distribution of ticket prices among survivors and non-survivors, but it is difficult to extract estimates of the mean, median, and quantiles, as well as the bounds of each bin.

Boxplots: The boxplots make it easy for us to compare the medians, quartiles, IQR (and outliers) of ticket prices across the two groups, although we cannot easily extract exact values for these. Also, since boxplots only display a small number of summary statistics, we lose information about the shape of the distributions.

Summary table: The summary table makes it easy to compare numerical values of key statistics. It is only from the summary table that we noticed that some passengers were recorded to have paid 0 pounds for their tickets. However, it is more difficult to get a quick sense of the overall shape of the distributions from these summary statistics alone, although these could be used to sketch a pair of boxplots.

[Question 3] The `CIACountries` data set is available in the `mdsr` library. Countries can be categorized by gross domestic product (GDP) - "a monetary measure of the market value of all the final goods and services produced in a period of time, often yearly or quarterly" (see Wikipedia article). The `gdp` variable contains data on GDP per person (or per capita).

```
library(mdsr)
glimpse(CIACountries)
```

```
## Observations: 236
## Variables: 8
## $ country   <chr> "Afghanistan", "Albania", "Algeria", "American Samoa", "And...
## $ pop       <dbl> 32564342, 3029278, 39542166, 54343, 85580, 19625353, 16418,...
## $ area      <dbl> 652230, 28748, 2381741, 199, 468, 1246700, 91, 443, 2780400...
## $ oil_prod  <dbl> 0, 20510, 1420000, 0, NA, 1742000, NA, 0, 532100, 0, 0, 354...
## $ gdp       <dbl> 1900, 11900, 14500, 13000, 37200, 7300, 12200, 23600, 22600...
## $ educ      <dbl> NA, 3.3, 4.3, NA, NA, 3.5, 2.8, 2.4, 6.3, 3.3, 6.0, 5.6, 5....
## $ roadways  <dbl> 0.06462444, 0.62613051, 0.04771929, 1.21105528, 0.68376068,...
## $ net_users <fct> >5%, >35%, >15%, NA, >60%, >15%, >15%, >60%, >35%, >35%, >6...
```

(a) Modify the `CIACountries` data frame to do the following:

- create a new variable called `gdp_cat` which takes the value "high" if the GDP is at least $50,000 and "med-low" otherwise
- exclude all observations with missing values for `gdp_cat`
- keep only the `country`, `pop`, `roadways`, and `gdp_cat` categories.

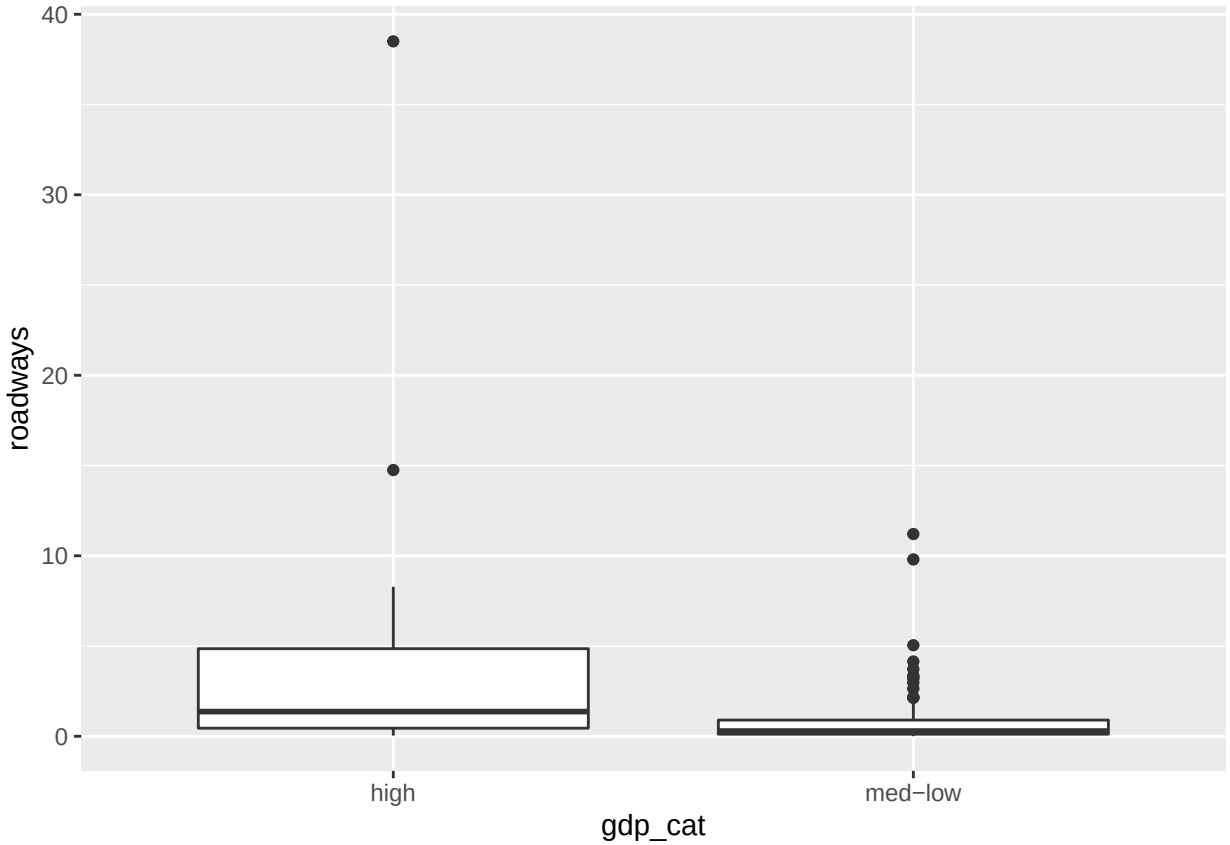
```
CIACountries <- CIACountries %>%
  mutate(gdp_cat = ifelse(gdp >= 50000, "high","med-low")) %>%
  filter(is.na(gdp_cat) == F) %>%
  select(country, pop, roadways, gdp_cat)

glimpse(CIACountries)
```

```
## Observations: 228
## Variables: 4
## $ country  <chr> "Afghanistan", "Albania", "Algeria", "American Samoa", "Ando...
## $ pop      <dbl> 32564342, 3029278, 39542166, 54343, 85580, 19625353, 16418, ...
## $ roadways <dbl> 0.06462444, 0.62613051, 0.04771929, 1.21105528, 0.68376068, ...
## $ gdp_cat  <chr> "med-low", "med-low", "med-low", "med-low", "med-low", "med-...
```

(b) The variable `roadways` records the roadways per unit area (i.e., km of roadways per square km). Use boxplots to compare the distribution of roadways in countries with a GDP of at least $50000 compared to less than $50000? Interpret the boxplots. What conclusions can you draw from this comparison?

```
CIACountries %>%
  ggplot(aes(x = gdp_cat, y= roadways)) + geom_boxplot()
```



From the boxplots, both groups of countries have right-skewed roadways distributions, with some high outliers. Roadways vary from just over 0 km/sq km for both groups of countries, to just under 40 km/sq km for countries with high GDP versus 12 km/sq km for countries with low-medium GDP. We also see that countries with high GDP tend to have more roadways per square kilometer than countries with lower GDP since the median roadways is higer (i.e., around 2 km/sq km) for countries with higher GDP, than countries with lower GDP (around 0.5 km/sq km). This makes sense because wealthier countries (i.e., those with higher GDP) would have the means to spend more money on roadways than poorer countries. Further, countries that produce fewer goods and services would have lower GDP and would not require the same level of transportation to transport goods and services as countries with higher GDP. Or, perhaps the lack of road infrustructure is a barrier to growth in GDP.