STA130H1S – Winter 2024

Week 3 Practice Problems

N. Moon and J. Speagle and Christopher Li

Instructions

How do I hand in my solutions and how do I check my work

You will submit your solutions (.Rmd and .pdf) on MarkUs at the following link: https://markus4.teach.cs.toronto.edu/2024-01/courses/1 Submissions are due at 5pm on Thursdays; see Quercus page for the specific deadline for each problem set.

Usually when you do an assignment, you don't find out whether your answers are correct until *after* the deadline, when you get your grade back. However, using MarkUs, you can submit your work before the deadline and run tests to check your solutions!

Note: Some parts of some questions may not be covered by tests in MarkUs, but you're still responsible for reviewing the posted solutions and make sure you understand them. Some of these parts will also be graded in some weeks.

Some of the optional questions have tests in MarkUs, but you won't be penalized for not completing these (or failing the tests for these parts) when we grade your work after the submission deadline. The tests for these parts are provided for your guidance.

What to do if a test fails on MarkUs

- Take a deep breath! Your work won't really be graded until the deadline, so start early to make sure you have lots of time to resolve issues before the deadline.
- Read the message to get hints about what the problem is. For example "variable X not present" means that you may have a typo in the name of the variable we're looking for - re-read the question carefully and make sure you're following the instructions.
- Search on Piazza to see if other classmates have encountered a similar error (and if not, consider posting a screenshot of the error message).
- Come to TA or instructor office hours with your issue.

[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, 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~
```

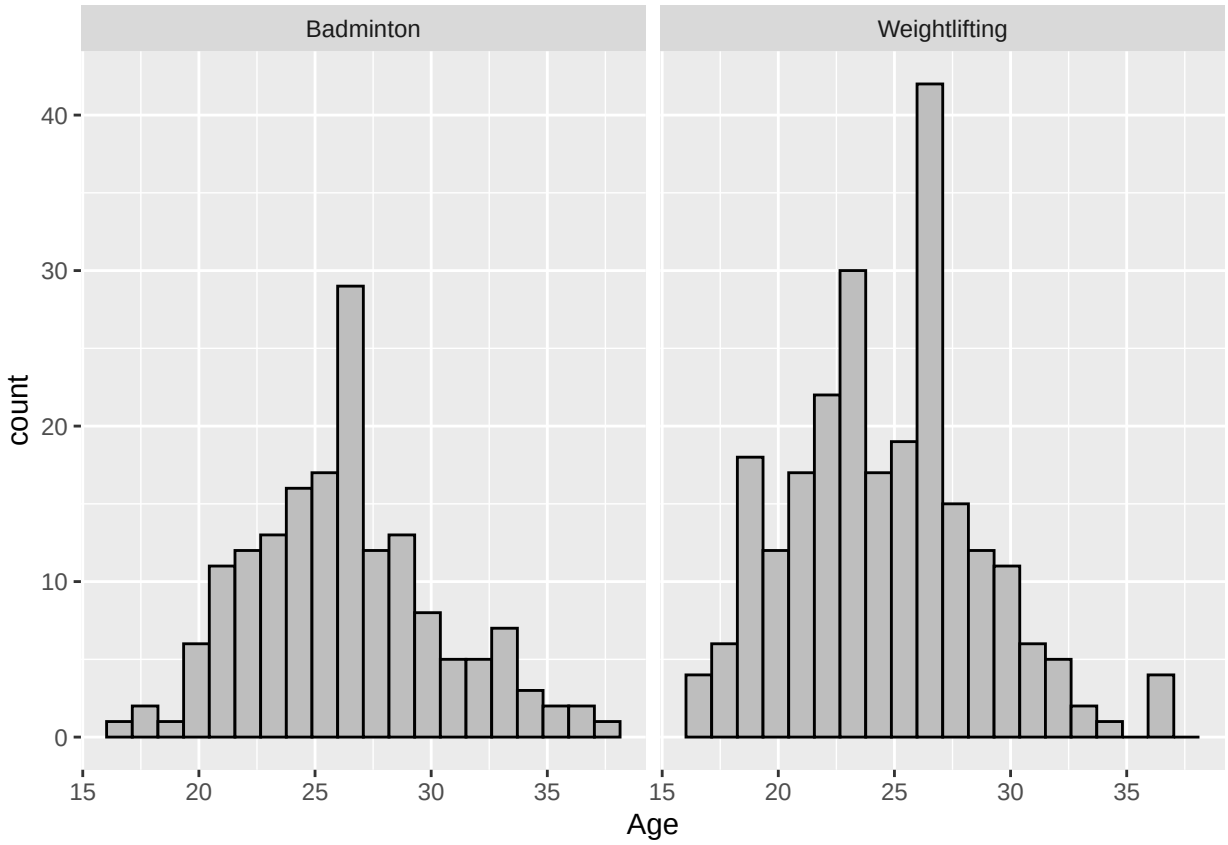
```
## $ Height  <dbl> 1.70, 1.93, 1.87, NA, 1.78, 1.82, 1.82, 1.87, 1.90, 1.70, NA, ~
## $ Weight  <int> 60, 125, 76, NA, 85, 80, 73, 75, 80, NA, NA, NA, 60, 64, 62, N~
## $ Sex     <fct> M, M, M, M, F, M, F, M, M, M, M, M, F, F, M, F, M, M, M, M, F,~
## $ DOB     <date> 1989-02-06, NA, NA, 1988-09-02, 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) Create a new tibble called `oly12_selectedSports` which contains only data for athletes who competed in either Weightlifting or Badminton (look at values of the `Sport` variable). Your new dataset should contain all the same columns as the original `oly12` tibble, and you shouldn't sort the data. In the code chunk below, replace NULL with the code required to produce the tibble described above.

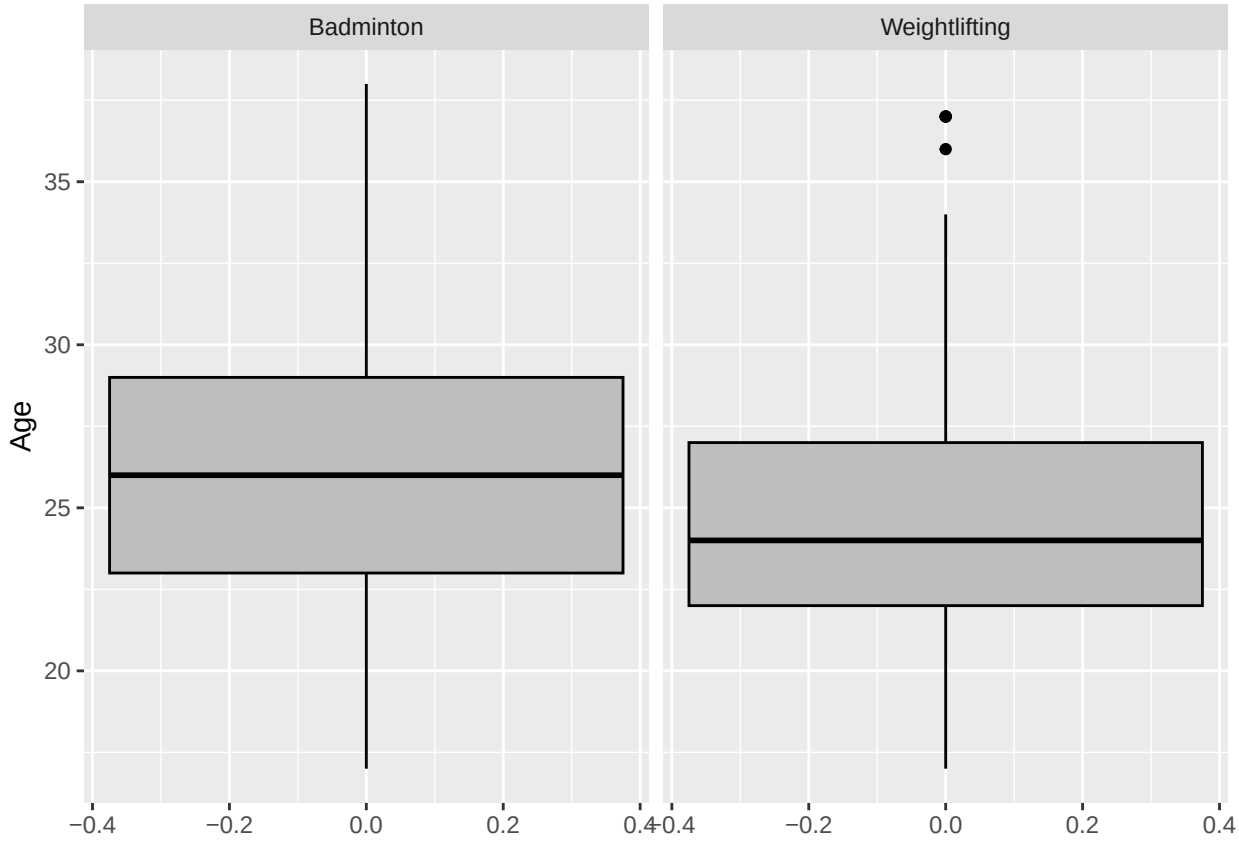
```
oly12_selectedSports <- oly12 %>% filter(Sport %in% c("Weightlifting", "Badminton"))
```

(b) [Optional] 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. Produce at least two relevant visualizations to compare these distributions.

```
oly12_selectedSports %>% ggplot(aes(x = Age)) + geom_histogram(color="black", fill="gray", bins=20) + fa
```



```
oly12_selectedSports %>% ggplot(aes(x = Age)) + geom_boxplot(color="black", fill="gray") + coord_flip()
```



(c) [Optional] Based on the plots you created in (b), answer the following questions:

(i) Are the age distributions of badminton players and weightlifters symmetrical or skewed? Specify which aspects of the plot(s) you used to assess this.

These 2 distributions are approximately symmetrical with a very slight skew right. For each graph, the peaks seem to be in the center of each distribution, but there does seem to be a tail in each graph.

(ii) Is the median age higher for badminton players or weightlifters? Specify which aspect(s) of the plot(s) you used to assess this.

The median for badminton is higher than that of weightlifting. When comparing the median line in each box plot, the median of Badminton is higher than that of Weightlifting.

(iii) Based only on the histogram and boxplots (no calculations), predict whether the standard deviation of the ages is similar or different. Justify your answer in 2-3 sentences, making specific reference to aspect(s) of the plot(s) you used to assess this.

The standard deviation for weightlifting is most likely lower than that of badminton. The range for badminton is approx ~25y while the range for weightlifting is ~20y. Both distributions look similar so the larger range of badminton will result in a larger standard deviation.

(d) Create a summary table reporting the minimum, maximum, mean, median, and standard deviation of ages for badminton players and weightlifters. The columns of your summary table should be "min", "max", "mean", "median", and "sd" (in this order). Write a sentence compare these values to the prediction you made in (e-iii). In the code chunk below, replace NULL with the code required to produce the tibble described above.

```
oly12_part_d <- oly12_selectedSports %>% group_by(Sport) %>% summarize(
  min = min(Age),
  max = max(Age),
  mean = mean(Age),
  median = median(Age),
  sd = sd(Age),
)
```

(e) Use the `arrange` function to find the name and age of the 6 youngest athletes who competed in the 2012 Olympics. Your answer should be a tibble with 6 rows (sorted by age, with the youngest athlete in the first row) and the following 4 columns in this order: "Name", "Age", "Sport", "Event". In the code chunk below, replace NULL with the code required to produce the tibble described above.

```
oly12_part_e <- oly12 %>% select(Name, Age, Sport, Event) %>% arrange(Age) %>% head()
```

(f) 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. Your answer should be a tibble sorted by age, with the youngest athlete in the first row and the following 4 columns in this order: "Name", "Age", "Sport", "Event". In the code chunk below, replace NULL with the code required to produce the tibble described above.

```
oly12_part_f <- oly12 %>% filter(Gold > 0) %>% select(Name, Age, Sport, Event) %>% arrange(Age) %>% hea
```

(g) Create a new tibble called `oly12_partg` which contains the following columns: "Name", "Country", "Sport", and "total_medals", in this order. The new variable "total_medals" should be calculated based on values of existing variables in `oly12`.

```
oly12_part_g <- oly12 %>% mutate(total_medals = Gold + Silver + Bronze) %>% select(Name, Country, Sport
```

(h) [Optional] Use `arrange`, `head(n=1)` and `select` to extract name of the athlete who won the most total medals (you should get a tibble with one row and one column). Hint: `head(n=1)` keeps only the top row of a tibble. In the code chunk below, replace NULL with the code required to extract the name of the athlete you're looking for

```
oly12_part_h_optional <- oly12_part_g %>% arrange(desc(total_medals)) %>% head(n=1) %>% select(Name)
```

[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", col_types = "cccccdccdccccc")
glimpse(titanic)
```

```
## Rows: 2,208
## Columns: 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", "Southampton~
## $ Class        <chr> "3", "Crew", "3", "3", "3", "3", "3", "2", "2", "3", "3",~
## $ MWC          <chr> "Man", "Man", "Child", "Man", "Woman", "Woman", "Man", "M~
## $ Age          <dbl> 42.00, 21.00, 14.00, 16.00, 39.00, 16.00, 25.00, 30.00, 2~
## $ Adut_or_Chld <chr> "Adult", "Adult", "Child", "Adult", "Adult", "Adult", "Ad~
## $ Sex          <chr> "Male", "Male", "Male", "Male", "Female", "Female", "Male~
## $ Paid         <dbl> 7.550000, NA, 20.250000, 20.250000, 20.250000, 7.650000, ~
## $ Ticket_No    <chr> "5547", NA, "CA2673", "CA2673", "CA2673", "348125", "3481~
## $ Boat_or_Body <chr> NA, NA, NA, "[190]", "A", "16", "A", NA, "10", "15", "C",~
## $ Job          <chr> "Blacksmith", "Lounge Pantry Steward", "Scholar", "Jewell~
## $ Class_Dept   <chr> "3rd Class Passenger", "Victualling Crew", "3rd Class Pas~
## $ Class_Full   <chr> "3", "V", "3", "3", "3", "3", "3", "2", "2", "3", "3", "E~
```

(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`. Don't change the order of the other variables in the `titanic` tibble. In the code chunk below, replace NULL with the code required to produce the tibble described above.

```
titanic_part_a_i <- 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. Starting with the dataset you created in the previous part (`titanic_part_a_i`), 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: When using the rename function, put the new variable name on the left of the equals sign, and t
oly12 <- oly12 %>% rename(Place_of_birth = PlaceOB)
```

```
titanic_part_a_ii <- titanic_part_a_i %>% rename(Man_Woman_or_Child = MWC)
```

```
# for the rest of this question, you'll be using the dataset you created in the previous part (with the
```

```
# Do not change the code on the line below
titanic_clean <- titanic_part_a_ii
titanic_clean
```

```
## # A tibble: 2,208 x 10
##    Name       Survived Boarded Class Man_Woman_or_Child   Age Adut_or_Chld Sex
##    <chr>      <chr>    <chr>   <chr> <chr>               <dbl> <chr>        <chr>
##  1 ABBING, M~ Dead     Southa~ 3     Man                    42 Adult        Male
##  2 ABBOTT, M~ Dead     Southa~ Crew  Man                    21 Adult        Male
##  3 ABBOTT, M~ Dead     Southa~ 3     Child                  14 Child        Male
##  4 ABBOTT, M~ Dead     Southa~ 3     Man                    16 Adult        Male
##  5 ABBOTT, M~ Alive    Southa~ 3     Woman                  39 Adult        Fema~
##  6 ABELSETH,~ Alive    Southa~ 3     Woman                  16 Adult        Fema~
##  7 ABELSETH,~ Alive    Southa~ 3     Man                    25 Adult        Male
##  8 ABELSON, ~ Dead     Cherbo~ 2     Man                    30 Adult        Male
##  9 ABELSON, ~ Alive    Cherbo~ 2     Woman                  28 Adult        Fema~
## 10 ABRAHAMSS~ Alive    Southa~ 3     Man                    20 Adult        Male
## # i 2,198 more rows
## # i 2 more variables: Paid <dbl>, Job <chr>
```

(b) Starting with the clean dataset you created in the previous part (`titanic_clean`), 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). The names of the columns in your summary table should be `n`, `n_died` and `prop_died`, in this order. In the code chunk below, replace NULL with the code required to produce the tibble described above.

```
titanic_part_b <- titanic_clean %>% summarize(n = n(), n_died = sum(Survived == "Dead"), prop_died = su
```

(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_clean %>% group_by(Man_Woman_or_Child) %>% summarize(prop_died = sum(Survived == "Dead") / n())
```

```
## # A tibble: 3 x 2
##   Man_Woman_or_Child prop_died
##   <chr>                   <dbl>
## 1 Child                   0.484
## 2 Man                     0.806
## 3 Woman                   0.243
```

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

```
titanic_clean %>% filter(18 <= Age & Age <= 25) %>% summarize(prop_died = sum(Survived == "Dead") / n()
```

```
## # A tibble: 1 x 1
##   prop_died
##       <dbl>
## 1     0.690
```

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

```
titanic_clean %>% filter(30 <= Paid) %>% summarize(prop_died = sum(Survived == "Dead") / n())
```

```
## # A tibble: 1 x 1
##   prop_died
##       <dbl>
## 1     0.424
```

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

The proportion of men who died is much higher that either woman or children. However, the proportion of richer passengers (i.e. paid > 30 pounds) had a higher chance of survival, though the deaths of people within 18-25 was pretty representative of the total deaths.

(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. In the code chunk below, replace NULL with the code required to produce the tibble described above.

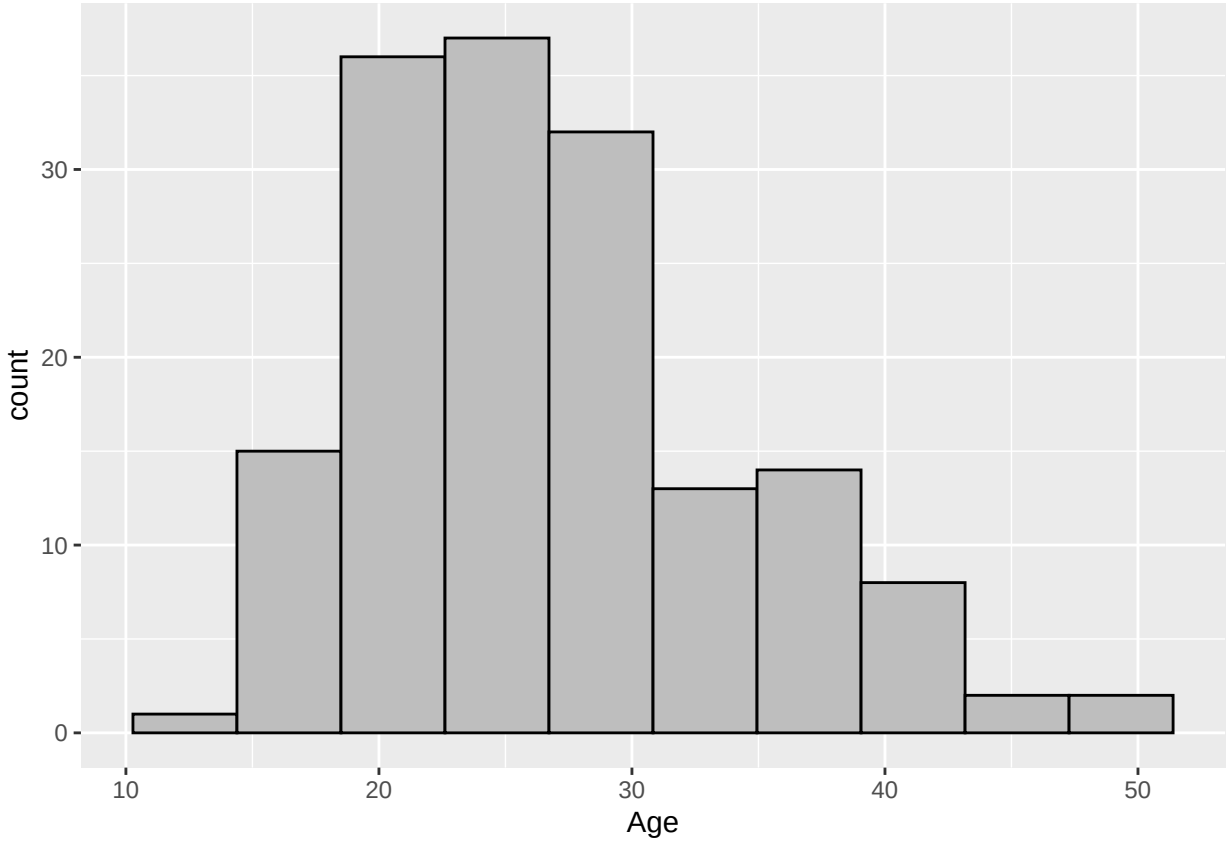
```
titanic_part_d <- titanic_clean %>% group_by(Job) %>% summarize(n = n()) %>% arrange()


titanic_part_d
```

```
## # A tibble: 358 x 2
##    Job                            n
##    <chr>                      <int>
##  1 1st Class Bedroom Steward     12
##  2 1st Class Clerk                1
##  3 1st class baggage steward      1
##  4 1st. Officer                   1
##  5 1st. Saloon Steward            1
##  6 2nd (Assistant) Storekeeper    1
##  7 2nd Baker                      1
##  8 2nd Butcher                    1
##  9 2nd Class Bedroom Steward      5
## 10 2nd Electrician                1
## # i 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) [Optional] Plot the age distribution for passengers with the job "General Labourer", and describe this distribution in 1-2 sentences.

```
titanic_clean %>% filter(Job == "General Labourer" & !is.na(Age)) %>% ggplot(aes(x = Age)) + geom_histo
                  fill = "gray",
                  bins = 10)
```



The distribution is unimodal and skewed right. There is a center at around 25 years of age and no outliers.

(f) [Optional] Were any of the general labourers on the titanic women? If so, how many? Hint: You can either produce a plot or a summary table to answer this question - it is up to you.

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

```
## # A tibble: 1 x 10
##   Name  Survived Boarded Class Man_Woman_or_Child   Age Adut_or_Chld Sex    Paid
##   <chr> <chr>    <chr>   <chr> <chr>              <dbl> <chr>        <chr> <dbl>
## 1 HAAS~ Dead     Southa~ 3     Woman                 24 Adult        Fema~  8.85
## # i 1 more variable: Job <chr>
```

```
# Replace NULL by TRUE or FALSE (no quotation marks), based on what you observe in the plots or summary
titanic_part_f_any_female_labourers_optional <- TRUE
```

(g) What are the names of the passengers with the most expensive ticket? Did these passengers survive the accident? In the code below, replace NULL with the family name of the passenger with the most expensive ticket (write the name in uppercase letters, with quotation marks. For example, if Nathalie Moon had the most expensive ticket, you would write "MOON"); in the titanic tibble, family names are in uppercase letters so they are easy to spot.

```
titanic_clean %>% arrange(desc(Paid))
```

```
## # A tibble: 2,208 x 10
##    Name       Survived Boarded Class Man_Woman_or_Child   Age Adut_or_Chld Sex
##    <chr>      <chr>    <chr>   <chr> <chr>              <dbl> <chr>        <chr>
##  1 CARDEZA, ~ Alive    Cherbo~ 1     Man                   36 Adult        Male
```

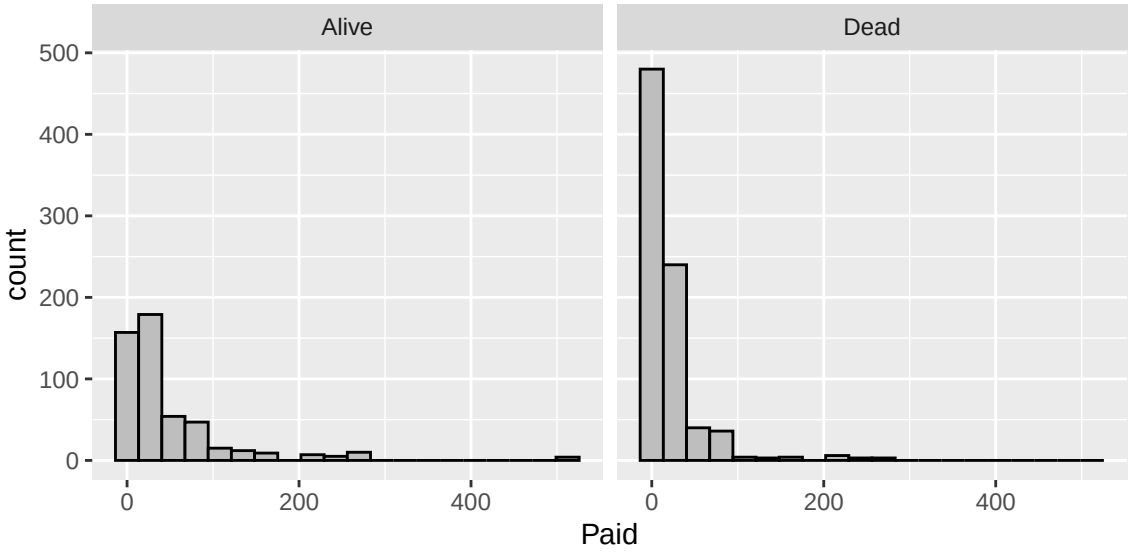
```
##  2 CARDEZA, ~ Alive    Cherbo~ 1     Woman                 58 Adult        Fema~
##  3 LESUEUR, ~ Alive    Cherbo~ 1     Man                   35 Adult        Male
##  4 WARD, Mis~ Alive    Cherbo~ 1     Woman                 35 Adult        Fema~
##  5 FORTUNE, ~ Alive    Southa~ 1     Woman                 24 Adult        Fema~
##  6 FORTUNE, ~ Alive    Southa~ 1     Woman                 28 Adult        Fema~
##  7 FORTUNE, ~ Alive    Southa~ 1     Woman                 23 Adult        Fema~
##  8 FORTUNE, ~ Dead     Southa~ 1     Man                   19 Adult        Male
##  9 FORTUNE, ~ Dead     Southa~ 1     Man                   64 Adult        Male
## 10 FORTUNE, ~ Alive    Southa~ 1     Woman                 60 Adult        Fema~
## # i 2,198 more rows
## # i 2 more variables: Paid <dbl>, Job <chr>
```

```
titanic_part_g <- "CARDEZA"
```

(h) [Optional] 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.

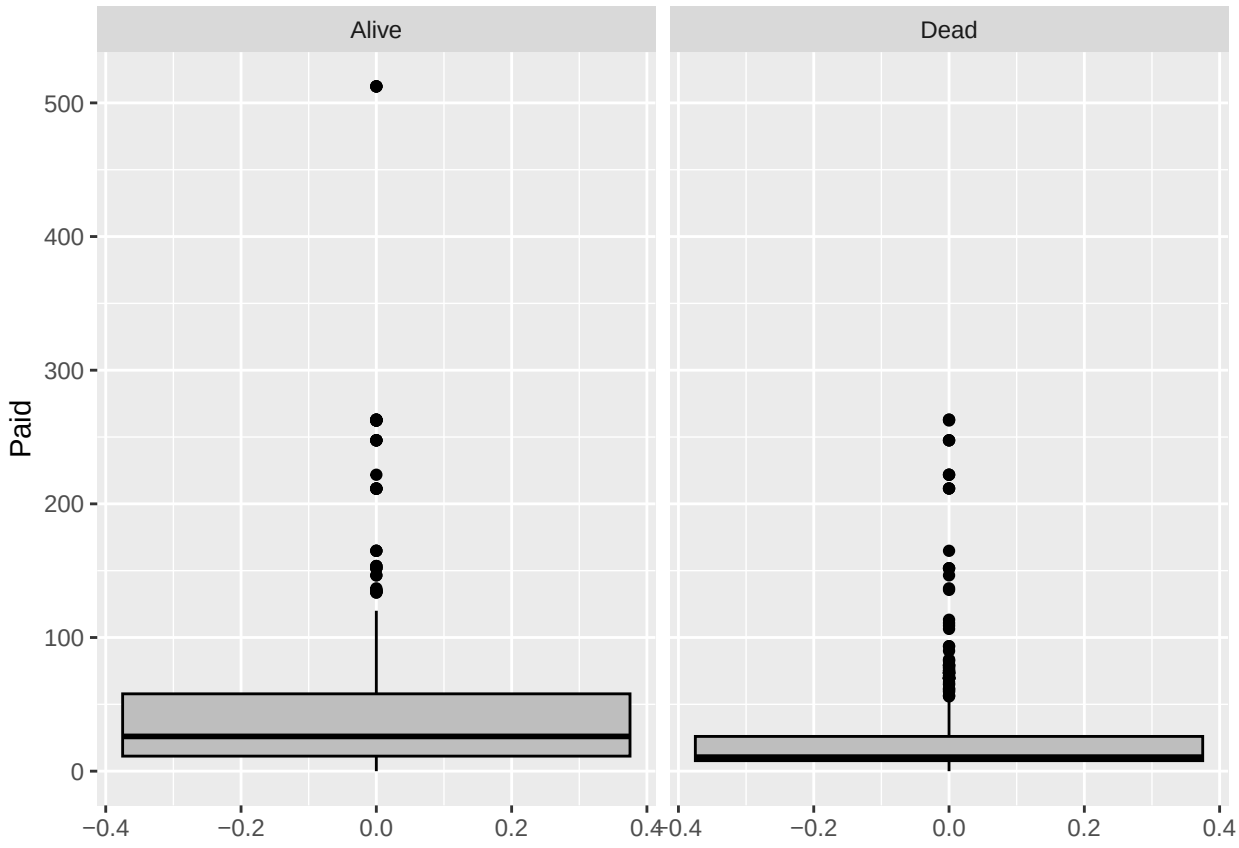
(i) Construct two histograms to visualize the distribution of ticket prices for survivors and non-survivors (i.e. one histogram for survivors and one for non-survivors). Write 2-3 sentences comparing the two distributions based on these plots.

```
titanic_clean %>% filter(!is.na(Paid)) %>% ggplot(aes(x = Paid)) + geom_histogram(color = "black",
                fill = "gray",
                bins = 20) + facet_wrap(~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_clean %>% filter(!is.na(Paid)) %>% ggplot(aes(x = Paid)) + geom_boxplot(color="black", fill="gra
```



(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
```

```r
median(x, na.rm=TRUE)
```

```
## [1] 4
```

```r
#quantile(titanic_clean["Paid"], probs=0.25, na.rm=True)
# quantile(titanic_clean, Paid, probs=0.75, na.rm=True)
```

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

```r
titanic_clean %>% summarize(min = min(Paid, na.rm=TRUE), q1 = quantile(Paid, probs=0.25, na.rm=TRUE), m
```

```
## # A tibble: 1 x 5
##     min    q1 median    q3   max
##   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0  7.90   14.4    31  512.
```

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

Summary tables gives us actual hard numbers, however it's hard (to impossible) to determine the shape of the distribution and using measures of center like mean/median might return bad numbers (i.e. multimodal, skewed, etc). Using a graph allows us to see the shape of the distribution, however we don't get any hard numbers to work with.