STATS5206 Homework 2

Maxime Grossman (UNI: mmg2240)

5/20/2021

Part 1

Task 1

Using ggplot, as opposed to Base R, produce the same plot constructed by the following code. That is, plot Petal Length versus Sepal Length split by Species. The colors of the points should be split according to Species. Also overlay three regression lines on the plot, one for each Species level. Makesure to include an appropriate legend and labels to the plot. Note: The function coef() extracts the intercept and the slope of an estimated line.

```
library(ggstar)
```

```
## Warning: package 'ggstar' was built under R version 3.6.3
```

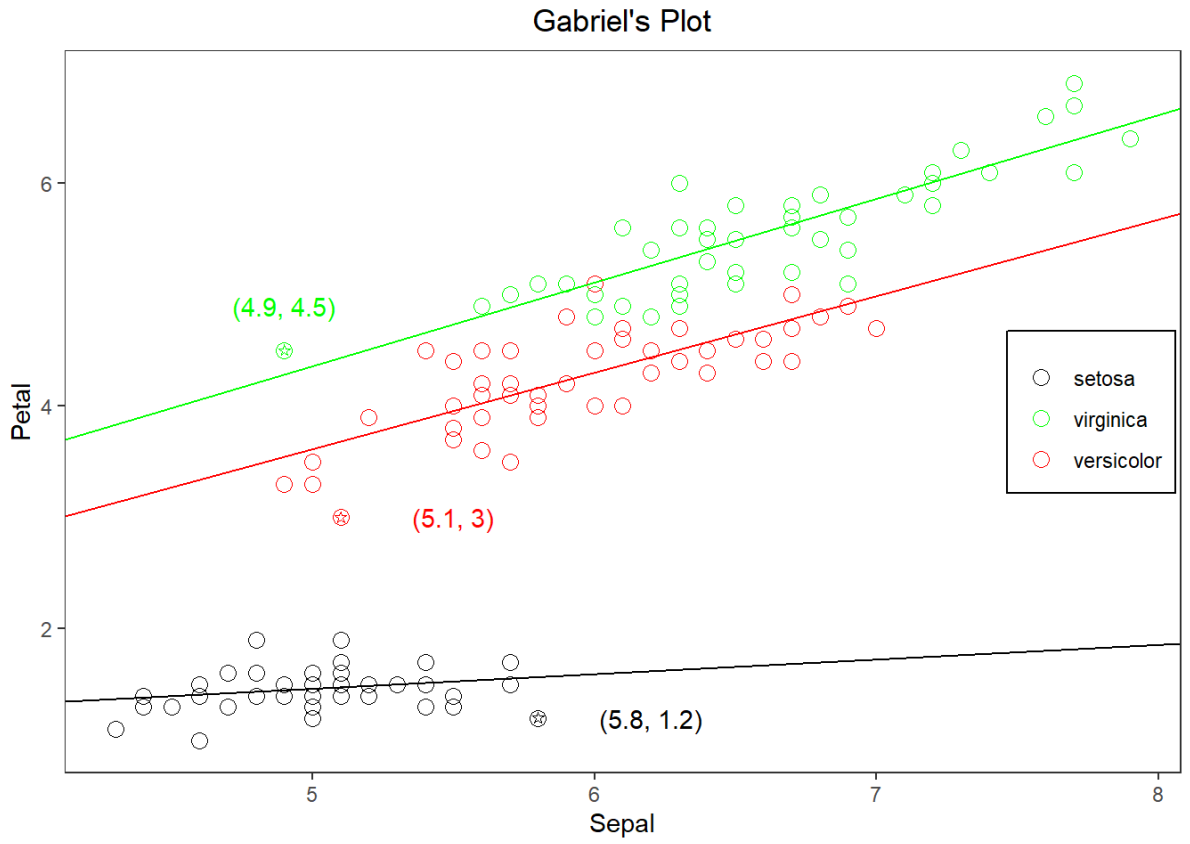
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
# find out which rows contain which species

whichsetosa <- which(iris$Species == "setosa")
whichvirginica <- which(iris$Species == "virginica")
whichversicolor <- which(iris$Species == "versicolor")

# regress each specie with its petal length and sepal length and take the coefficients of the regression line
# we will use this for plotting the regression line

lmsetosa <- coef(lm(iris$Petal.Length[whichsetosa]~iris$Sepal.Length[whichsetosa], data = iris))
lmvirginica <- coef(lm(iris$Petal.Length[whichvirginica]~iris$Sepal.Length[whichvirginica], data = iris))
lmversicolor <- coef(lm(iris$Petal.Length[whichversicolor]~iris$Sepal.Length[whichversicolor], data = iris))
```

```
ggplot(iris) +
  geom_point(mapping = aes(x=Sepal.Length, y = Petal.Length, color = Species), shape=1, size=3)+

  geom_abline(intercept = lmsetosa[1], slope = lmsetosa[2], color="black")+ #, color="black") +
  geom_abline(intercept = lmvirginica[1], slope = lmvirginica[2], color="green") + #, color="green") +
  geom_abline(intercept = lmversicolor[1], slope = lmversicolor[2], color="red") +

  labs(title = "Gabriel's Plot", x = "Sepal", y = "Petal") +  # labs for labels

  theme(plot.title = element_text(hjust=0.5))+ # center title

  scale_color_manual(breaks = c("setosa", "virginica", "versicolor"),  # color the points
                       values=c("black", "green", "red")) +

  geom_star(mapping=aes(x=5.8, y = 1.2), shape=1, size=1.2)+#, color = "black", shape=11)+
  geom_star(mapping=aes(x=5.1, y = 3), shape=1, size=1.2, color="red")+#, color = "purple", shape=11)+
  geom_star(mapping=aes(x=4.9, y = 4.5), shape=1, size=1.2, color="green")+  #, color = "purple", shape=11)

  annotate(geom="text", x=6.2, y=1.2, label = "(5.8, 1.2)", color="black") +
  annotate(geom="text", x=5.5, y=3, label = "(5.1, 3)", color="red") +
  annotate(geom="text", x=4.9, y=4.9, label = "(4.9, 4.5)", color="green") +

  theme(legend.position = c(0.92, 0.50), legend.background = element_rect(fill = "white", color = "black", li
netype = 1), legend.title = element_blank()) +

  theme_bw() +
  theme(panel.grid = element_blank()) +

  theme(legend.position = c(0.92, 0.50), legend.background = element_rect(fill = "white", color = "black", li
netype = 1), legend.title = element_blank()) +

    theme(plot.title = element_text(hjust=0.5))  # center the title
```

```
## Warning: Ignoring unknown parameters: shape

## Warning: Ignoring unknown parameters: shape

## Warning: Ignoring unknown parameters: shape
```




```
# Here is something else we could've done which is much simpler, but it wouldn't have plotted regression line
s that run indefinitely


#ggplot(iris, aes(Sepal.Length, Petal.Length, color=Species)) +
#  geom_point() +
# geom_smooth(method = "lm", se=FALSE)+
#  labs(title = "Sepal Length vs. Petal Length for All Species", x = "Sepal Length", y = "Petal Length") +
#  geom_abline(intercept = 0, slope = 2)
```

Part 2

Task 2

Open the file and make a new variable (dataframe) containing only the year, “P99”, “P99.5” and “P99.9” variables; these are the income levels which put someone at the 99th, 99.5th, and 99.9th, percentile of income. What was P99 in 1993? P99.5 in 1942? You must identify these using your code rather than looking up the values manually. The code for this part is given below.

```
wtid_report <- read.csv("wtid-report.csv")

wtid_report2 <- wtid_report[,c("Year", "P99.income.threshold","P99.5.income.threshold", "P99.9.income.thresho
ld")]

head(wtid_report2)
```

```
##   Year P99.income.threshold P99.5.income.threshold P99.9.income.threshold
## 1 1913             82677.22               135583.5               428630.4
## 2 1914             76405.62               126910.5               410528.7
## 3 1915             64409.44               122555.7               451668.3
## 4 1916             77289.78               138102.3               518327.4
## 5 1917             95326.69               154537.8               536356.5
## 6 1918             95202.66               147850.1               457045.0
```

```
# P99 in 1993

wtid_report2[wtid_report2$Year=="1993", "P99.income.threshold"]
```

```
## [1] 273534.9
```

```
# P99.5 in 1942

wtid_report2[wtid_report2$Year=="1942", "P99.5.income.threshold"]
```

```
## [1] 189140.6
```

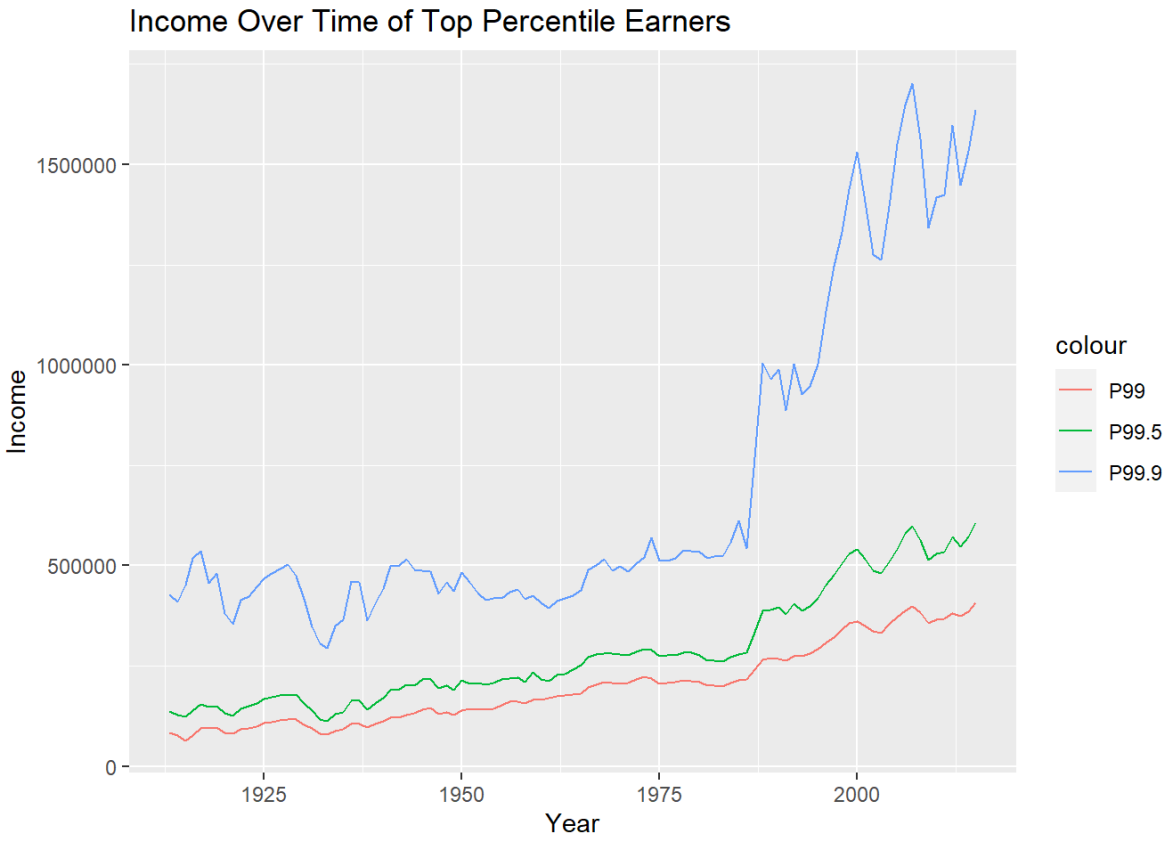
P99 in 1993 was 273,534.9

P99.5 in 1942 was 189,140.6

Task 3

Using ggplot, display three line plots on the same graph showing the income threshold amount against time for each group, P99, P99.5 and P99.9. Make sure the axes are labeled appropriately, and in particular that the horizontal axis is labeled with years between 1913 and 2012, not just numbers from 1 to 100. Also make sure a legend is displayed that describes the multiple time series plot. Write one or two sentences describing how income inequality has changed throughout time.

```
fac <- factor((subset(wtid_report2, select = c("P99.income.threshold","P99.5.income.threshold", "P99.9.incom
e.threshold"))))


plot <- ggplot(data = wtid_report2) +
  geom_line(mapping = aes(x=Year, y = P99.income.threshold, color="P99"))+
  geom_line(mapping = aes(x=Year, y = P99.5.income.threshold, color="P99.5"))+
  geom_line(mapping = aes(x=Year, y = P99.9.income.threshold, color="P99.9"))+
  labs(title = "Income Over Time of Top Percentile Earners", x = "Year", y = "Income")

plot
```



Income disparity seems to have increased in recent years. We can see that the 99.9 percentile has seen its income level exponentially increase since the late 1980s at a dramatic rate, while the 99.5 and 99 percentiles have seen a steady, somewhat linear increase.

Part 3

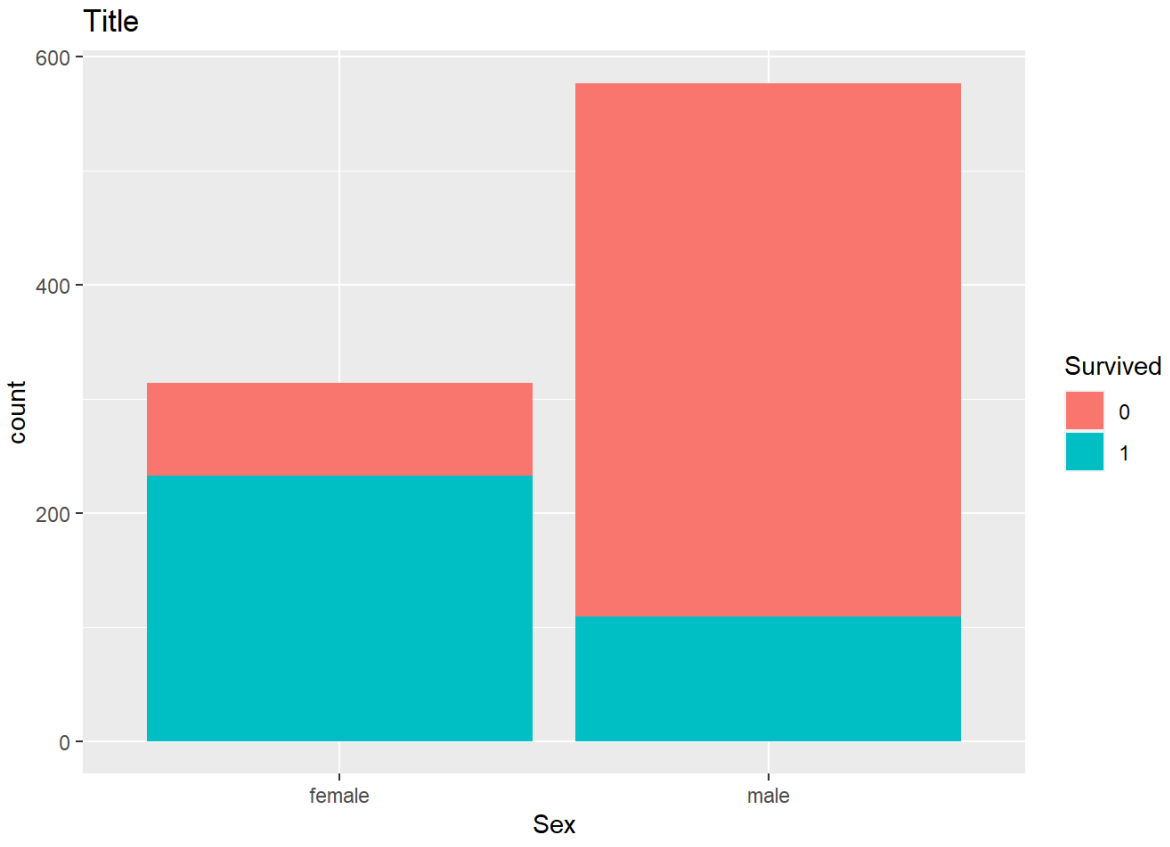
Task 4

Run the following code and describe what the two plots are producing.

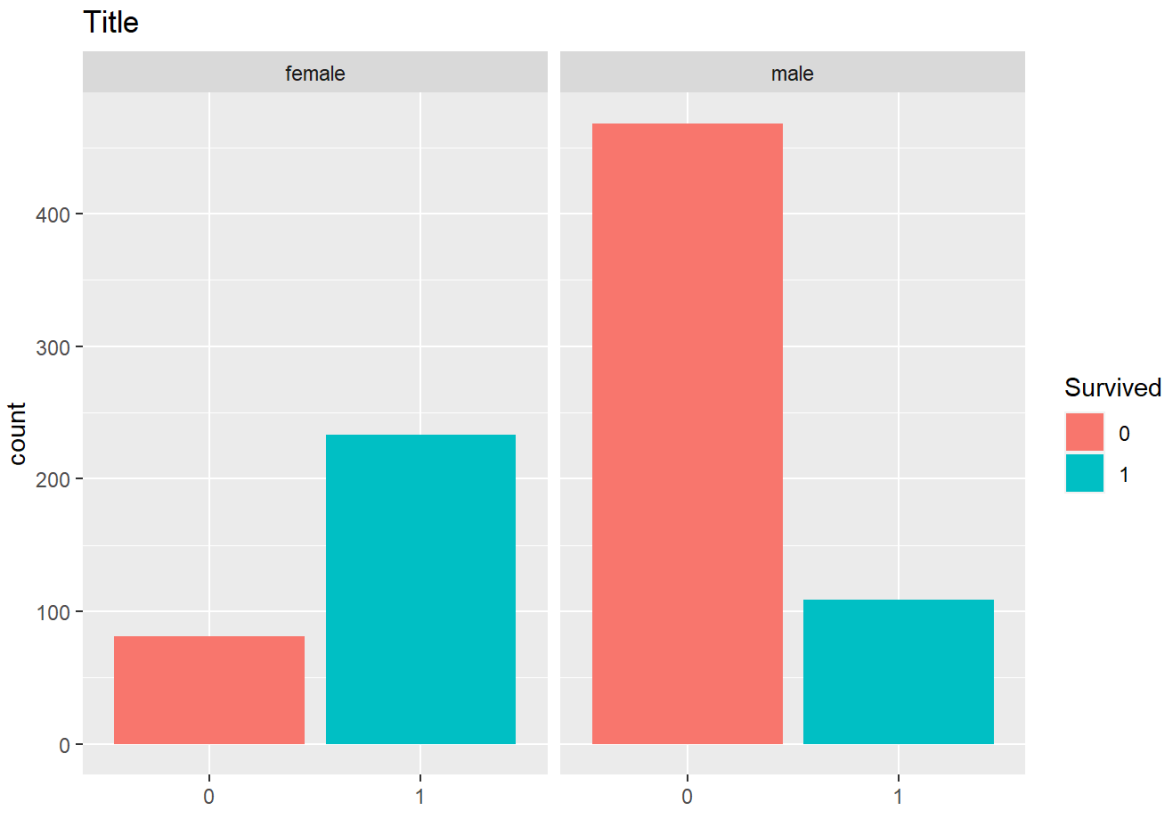
```
# Read in data
titanic <- read.table("Titanic.txt", header = TRUE, as.is = TRUE)
head(titanic)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp Parch
## 1                             Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                              Heikkinen, Miss. Laina female  26     0     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                            Allen, Mr. William Henry   male  35     0     0
## 6                                    Moran, Mr. James   male  NA     0     0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500              S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

```r
library(ggplot2)
# Plot 1
ggplot(data=titanic) +
geom_bar(aes(x=Sex,fill=factor(Survived)))+
labs(title = "Title",fill="Survived")
```



```
# plot 2
ggplot(data=titanic) +
geom_bar(aes(x=factor(Survived),fill=factor(Survived)))+
facet_grid(~Sex)+
labs(title = "Title",fill="Survived",x="")
```



The two plots are showing the count of survivors of the Titanic segmented by sex. We can see that a much higher rate and total value of females survived than males.

Task 5

Create a similar plot with the variable Pclass. The easiest way to produce this plot is to facet by Pclass. Make sure to include appropriate labels and titles. Describe your graph.

```
# mortality by class

ggplot(data=titanic) +
geom_bar(aes(x=factor(Survived),fill=factor(Survived)))+
facet_grid(~Pclass)+
labs(title = "Mortalities on the Titanic",fill="Survived",x="Mortality Counts by 1st, 2nd, & 3rd Class",y="Su
rvivals and Deaths") +
scale_fill_discrete(name = "Mortality", labels = c("0=Died", "1=Survived"))
```


```
# mortality by class and sex

ggplot(data=titanic) +
geom_bar(aes(x=factor(Survived),fill=factor(Survived)))+
facet_grid(Sex~Pclass)+
labs(title = "Mortalities on the Titanic",fill="Survived",x="Mortality Counts by 1st, 2nd, & 3rd Class Facete
d by Gender",y="Survivals and Deaths")+
  scale_fill_discrete(name = "Mortality", labels = c("0=Died", "1=Survived"))
```

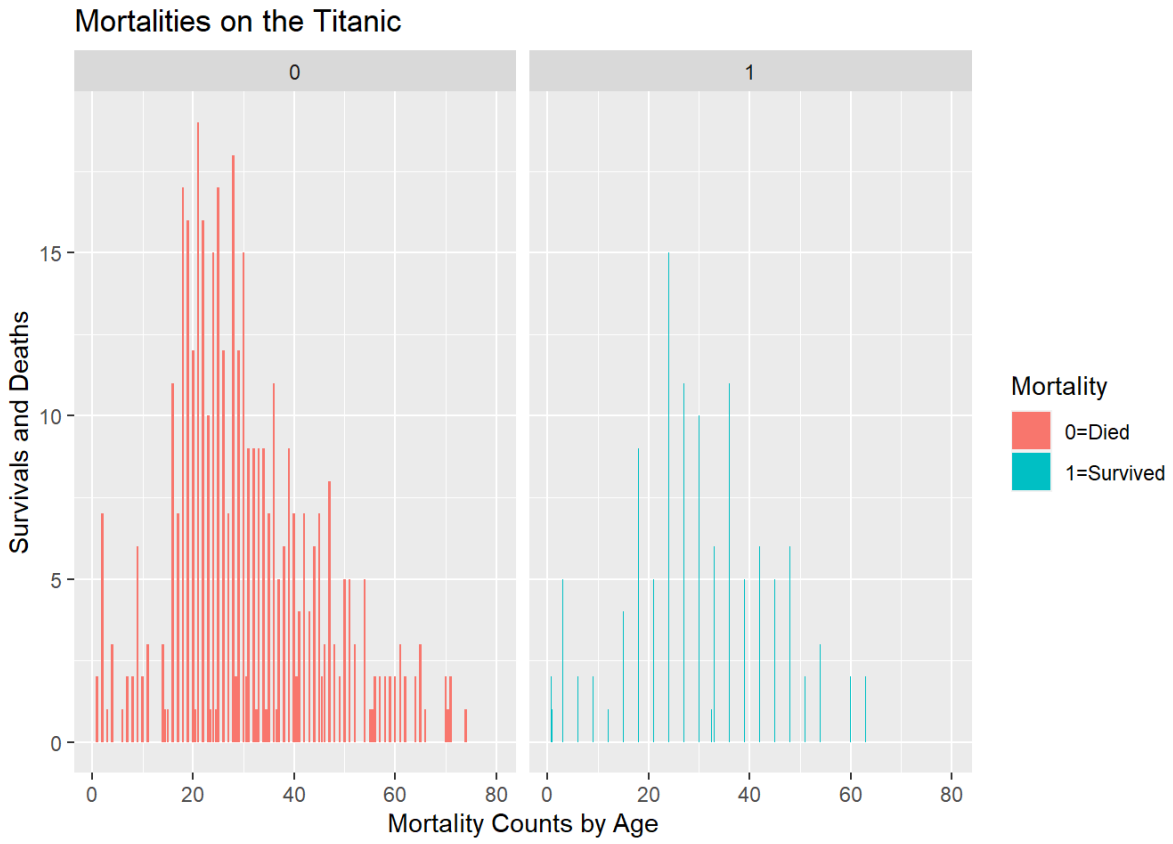
We can clearly observe that many more occupants in the 3rd class perished as compared with the 1st and 2nd class. Also, it's easy to see that many more males perished than females. In the 1st and 2nd class, the number of women who perished is extremely low. The number of women in the 1st class who perished might well be in the single digits.

Task 6

Create one more plot of your choice related to the titanic data set. Describe what information your plot is conveying

```
ggplot(data=titanic) +
geom_bar(aes(x=Age,fill=factor(Survived)))+
facet_grid(~Survived)+
labs(title = "Mortalities on the Titanic",fill="Survived",x="Mortality Counts by Age",y="Survivals and Death
s")+
scale_fill_discrete(name = "Mortality", labels = c("0=Died", "1=Survived"))
```

```
## Warning: Removed 177 rows containing non-finite values (stat_count).
```



This plot is segmenting survivals and deaths by age group. We can see that the majority of deaths occurred in occupants in their twenties and early thirties. We also see a spike in deaths of very young ages ten years and younger. However, we also see the highest survival rate at ages around early 20s. This would lead us to believe that in general, most occupants aboard the Titanic were in their twenties and early thirties.

Part 4

Task 7

Simulate a n = 1000 random draws from a beta distribution with parameters α = 3 and β = 1. Plot a histogram of the simulated cases using ggplot. Also overlay the beta density on the histogram. Hint: look up the beta distribution using ?rbeta.

```
x <- seq(0,1,by=.01)
hist_data <- data.frame(x.var=rbeta(1000, 3,1))  # simulate 1000 standard normals for histogram
plot_data <- data.frame(x=x, f=dbeta(x, 3, 1))  # for plotting pdf


ggplot(hist_data) +
  geom_histogram(mapping = aes(x=x.var, y=..density..),
                 col="blue", fill="white", binwidth=.2) +
  geom_line(plot_data, mapping=aes(x=x,y=f),
            col="red")+
  labs(title="Beta Example", x="x", y = "Density")
```