Homework 2 (50 Points)

Shuyuan Wang, sw3449

October 4, 2019

Instructions

Make sure that you upload an RMarkdown file to the canvas page (this should have a .Rmd extension) as well as the PDF or HTML output after you have knitted the file (this will have a .pdf or .html extension). Note that since you have already knitted this file, you should see both a **HW2.pdf** and a **HW2.Rmd** file in your GU4206/GR5206 folder. Click on the **Files** tab to the right to see this. The files you upload to the Canvas page should be updated with commands you provide to answer each of the questions below. You can edit this file directly to produce your final solutions. HW 2 is due 11:59 pm on Friday, October 11th.

Part 1 (Iris)

Background

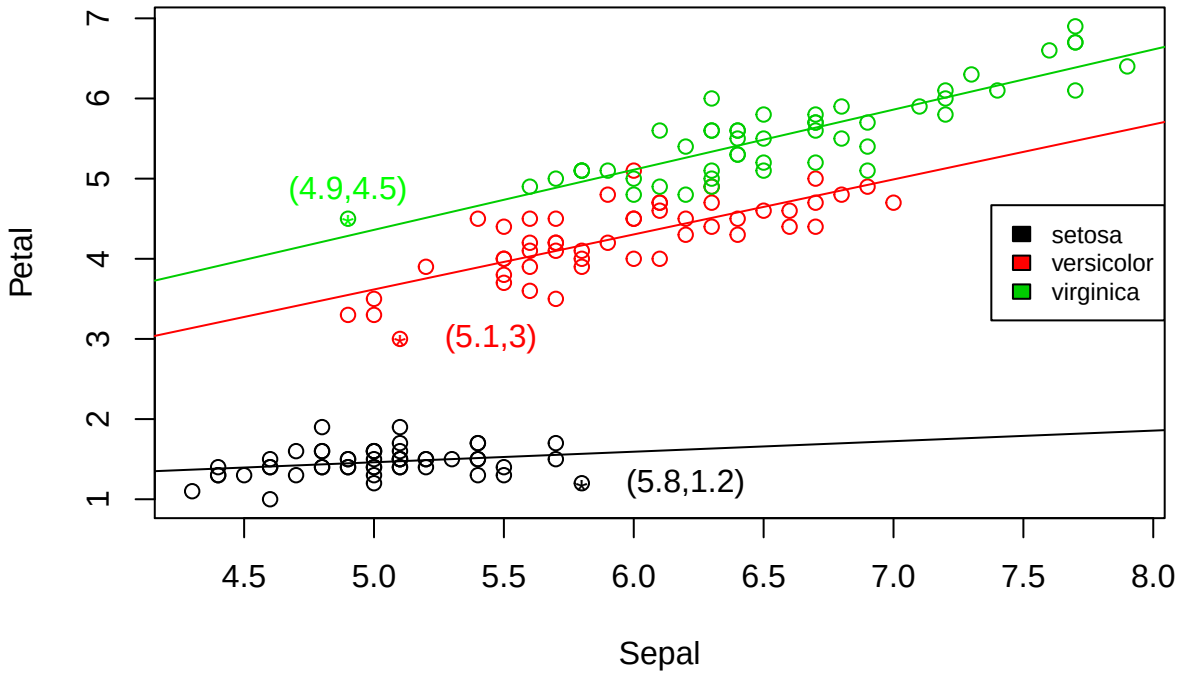
The R data description follows:

This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

Task

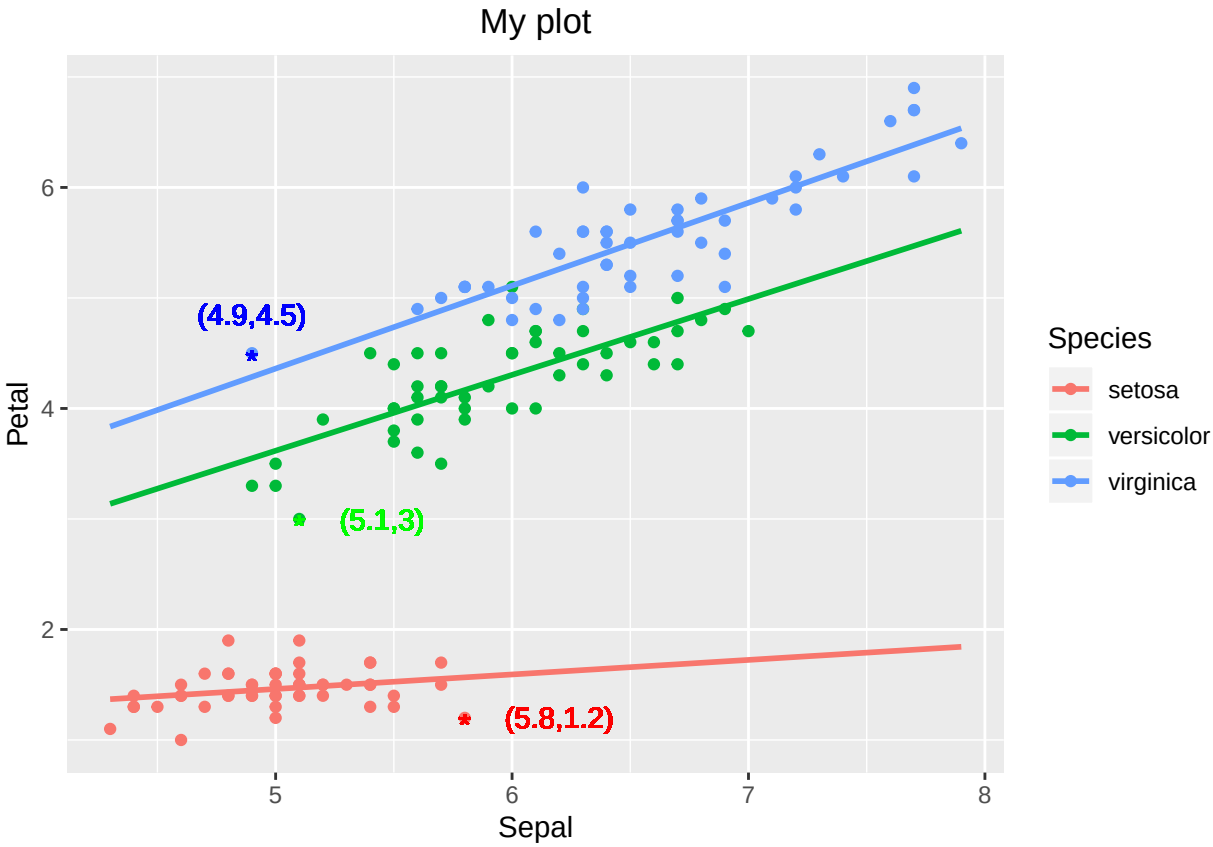
1) Using `ggplot`, as apposed to `Base R`, produce the same plot constructed by the following code. That is, plot **Petal Length** versus **Sepal Length** split by **Species**. The colors of the points should be split according to **Species**. Also overlay three regression lines on the plot, one for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

```
# Base plot
plot(iris$Sepal.Length,iris$Petal.Length,col=iris$Species,xlab="Sepal",ylab="Petal",main="Gabriel's Plo

# loop to construct each LOBF
for (i in 1:length(levels(iris$Species))) {
  extract <- iris$Species==levels(iris$Species)[i]
  abline(lm(iris$Petal.Length[extract]~iris$Sepal.Length[extract]),col=i)
  }

# Legend
legend("right",legend=levels(iris$Species),fill = 1:length(levels(iris$Species)), cex = .75)

# Add points and text
points(iris$Sepal.Length[15],iris$Petal.Length[15],  pch = "*", col = "black")
text(iris$Sepal.Length[15]+.4,iris$Petal.Length[15],"(5.8,1.2)",col="black")
points(iris$Sepal.Length[99],iris$Petal.Length[99],  pch = "*", col = "red")
text(iris$Sepal.Length[99]+.35,iris$Petal.Length[99],"(5.1,3)",col = "red")
points(iris$Sepal.Length[107],iris$Petal.Length[107],pch = "*", col = "green")
text(iris$Sepal.Length[107],iris$Petal.Length[107]+.35,"(4.9,4.5)",col = "green")
```



Solution goes below:

```
library(ggplot2)
ggplot(data=iris)+
   geom_point(mapping=aes(x=Sepal.Length,y=Petal.Length,color=Species))+
   geom_smooth(mapping=aes(x=Sepal.Length,y=Petal.Length,color=Species),method=lm,se=F,fullrange=T)+
   theme(plot.title = element_text(hjust = .5))+
   labs(title = "My plot",x="Sepal",y="Petal")+
   geom_point(mapping = aes(x=Sepal.Length[15], y=Petal.Length[15]),shape='*',size=4, color='red') +
   geom_text(mapping = aes(x=Sepal.Length[15]+.4, y=Petal.Length[15], label = "(5.8,1.2)"), size=4,colo
   geom_point(mapping = aes(x=Sepal.Length[99], y=Petal.Length[99]),shape='*',size=4, color='green') +
   geom_text(mapping = aes(x=Sepal.Length[99]+.35, y=Petal.Length[99], label = "(5.1,3)"), size=4,color
   geom_point(mapping = aes(x=Sepal.Length[107], y=Petal.Length[107]),shape='*',size=4, color='blue') +
   geom_text(mapping = aes(x=Sepal.Length[107], y=Petal.Length[107]+.35, label = "(4.9,4.5)"), size=4,c
```



Part 2 (World's Richest)

Background

We consider a data set containing information about the world's richest people. The data set us taken form the World Top Incomes Database (WTID) hosted by the Paris School of Economics [http://top-incomes.g-mond.parisschoolofeconomics.eu]. This is derived from income tax reports, and compiles information about the very highest incomes in various countries over time, trying as hard as possible to produce numbers that are comparable across time and space.

Tasks

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.

Solution goes below:

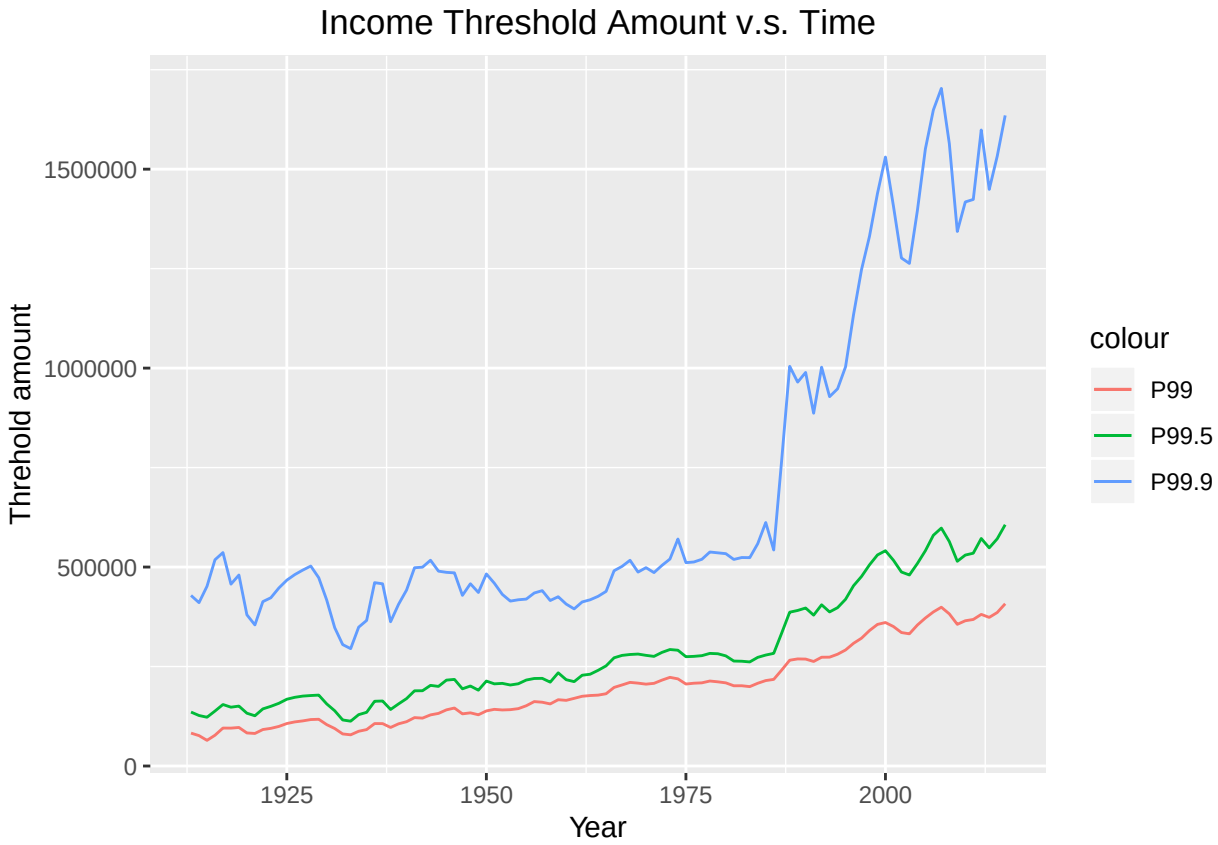
```
setwd("/Users/wangshuyuan/Desktop/GR5206-Stat Comp & DS/hw2")
wtid <- read.csv("wtid-report.csv", as.is = TRUE)
wtid <- wtid[, c("Year", "P99.income.threshold","P99.5.income.threshold", "P99.9.income.threshold")]
names(wtid) <- c("Year", "P99", "P99.5", "P99.9")
```

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.

Solution goes below:

```
ggplot(data=wtid)+
   geom_line(mapping=aes(x=Year,y=P99,color='P99'))+
   geom_line(mapping=aes(x=Year,y=P99.5,color='P99.5'))+
   geom_line(mapping=aes(x=Year,y=P99.9,color='P99.9'))+
   theme(plot.title = element_text(hjust = .5))+
   labs(y='Threhold amount',x='Year',title='Income Threshold Amount v.s. Time')
```



After around 1985, the difference between 99.9% percentile and the other two percentiles is significantly larger than before. The income inequality is enlarging as time goes by.

Part 3 (Titanic)

Background

In this part we'll be studying a data set which provides information on the survival rates of passengers on the fatal voyage of the ocean liner *Titanic*. The dataset provides information on each passenger including, for example, economic status, sex, age, cabin, name, and survival status. This is a training dataset taken from the Kaggle competition website; for more information on Kaggle competitions, please refer to https://www.kaggle.com. Students should download the data set on Canvas.

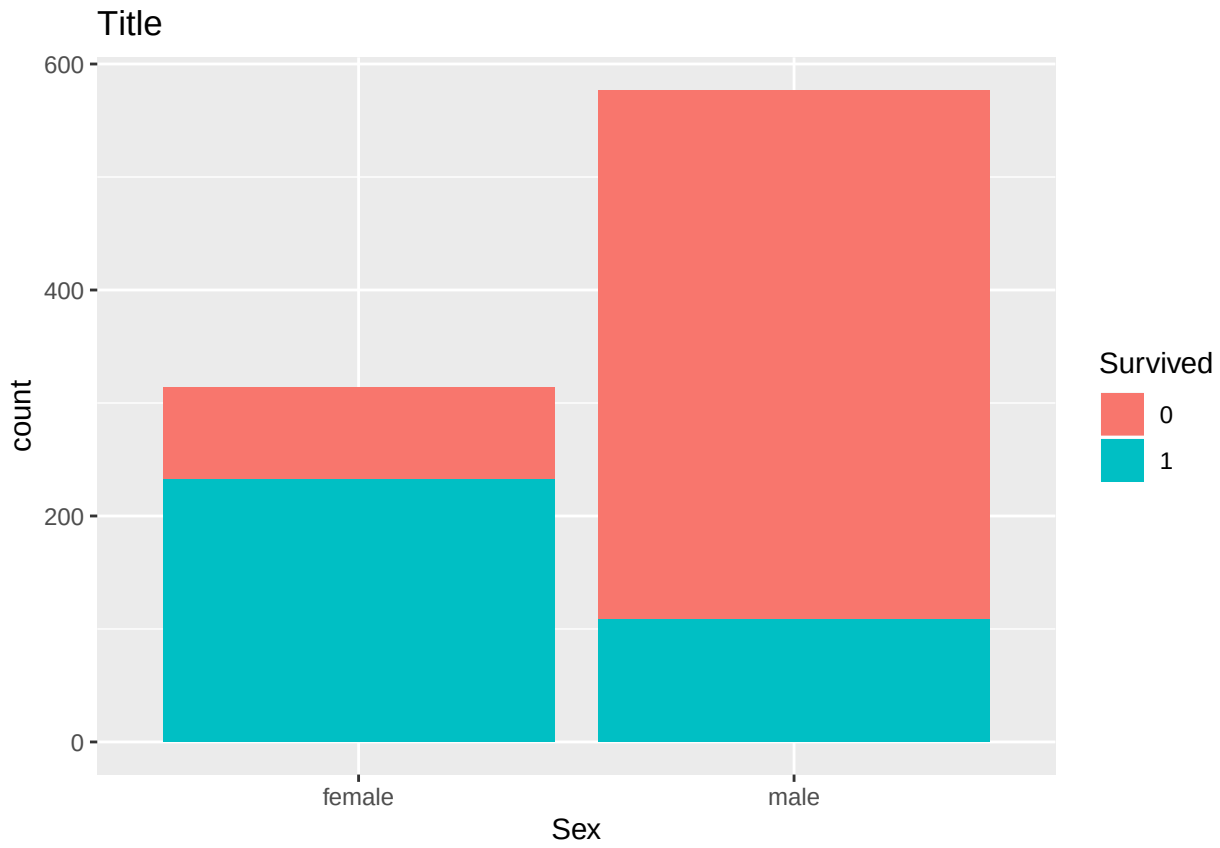
Tasks

4) Run the following code and describe what the two plots are producing

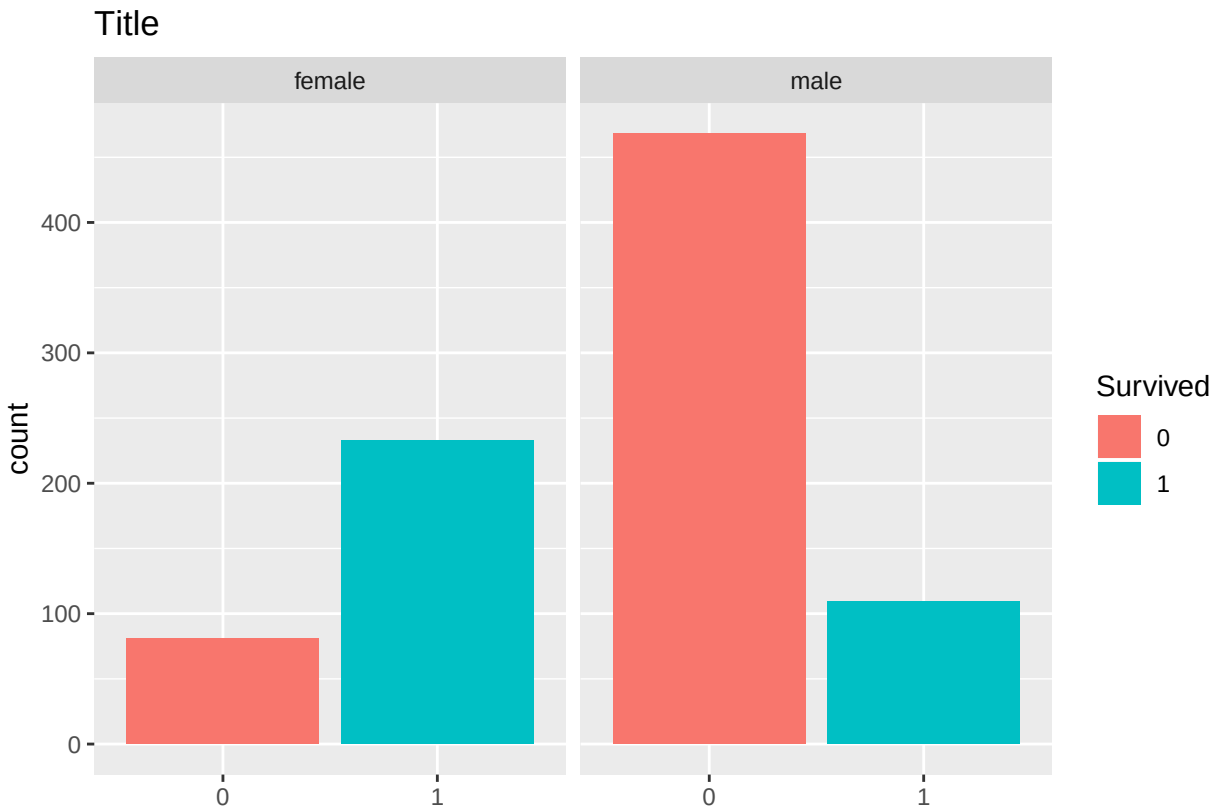
```
# Read in data
titanic <- read.table("Titanic.txt", header = TRUE, as.is = TRUE)
head(titanic)
```

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

```
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="")
```

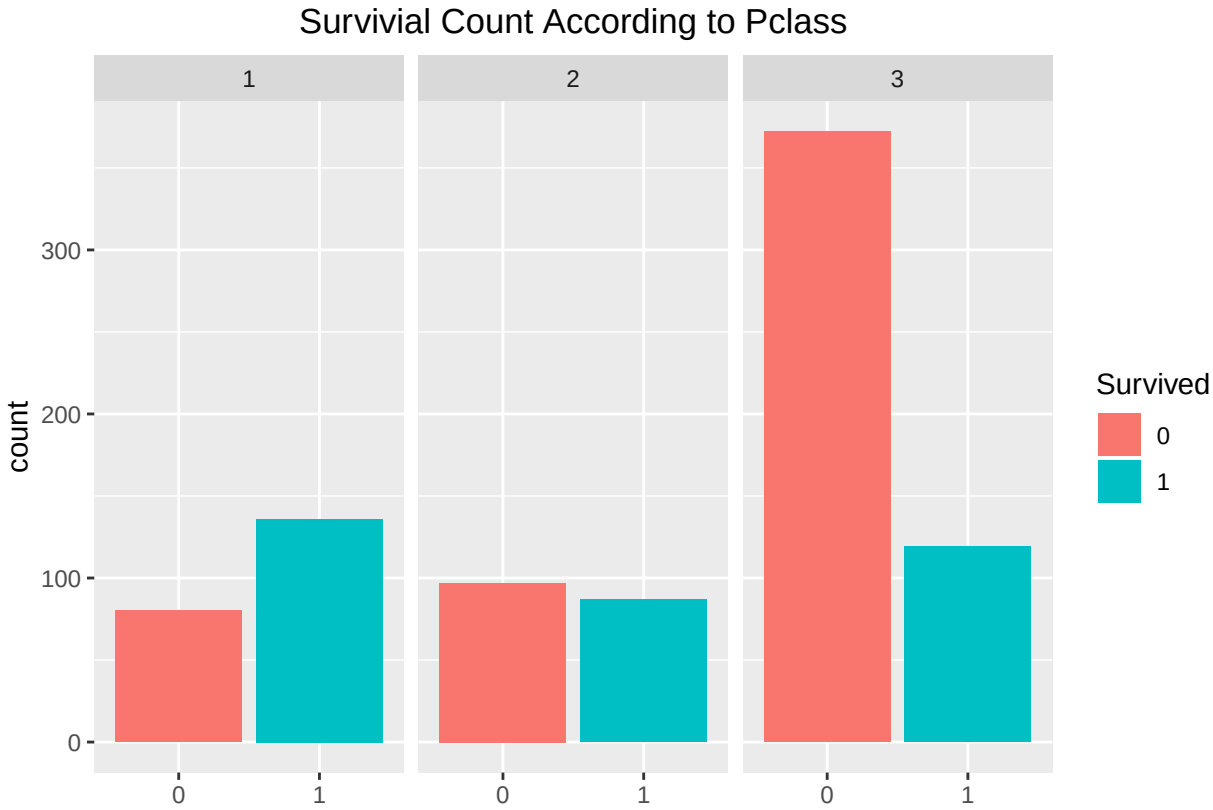


The two plots both producing the bargraph that describes how many females and males were survived or not.

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

Solution goes below:

```
ggplot(data=titanic) +
   geom_bar(aes(x=factor(Survived),fill=factor(Survived)))+
   facet_grid(~Pclass)+
   theme(plot.title = element_text(hjust = .5))+
   labs(title = "Survivial Count According to Pclass",fill="Survived",x="")
```



The first class customers have the highest survival rate and the third class customers have the lowest survival rate.

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

Solution goes below:

```
ggplot(data=titanic) +
   geom_bar(aes(x=factor(Survived),fill=factor(Survived)))+
   facet_grid(~Embarked)+
   theme(plot.title = element_text(hjust = .5))+
   labs(title = "Survivial Count According to Embarked Harbor",fill="Survived",x="")
```

Over 50% passengers embarked from harbor C survived, however, about 30% passengers embarked from harbor S survived. The harbor that passengers embarked from might have effect on their survival.