Homework 2 (50 Points)

Shreya Rao sr3843

May 11, 2021

Part 1 (Iris)

Background

The R data description follows:

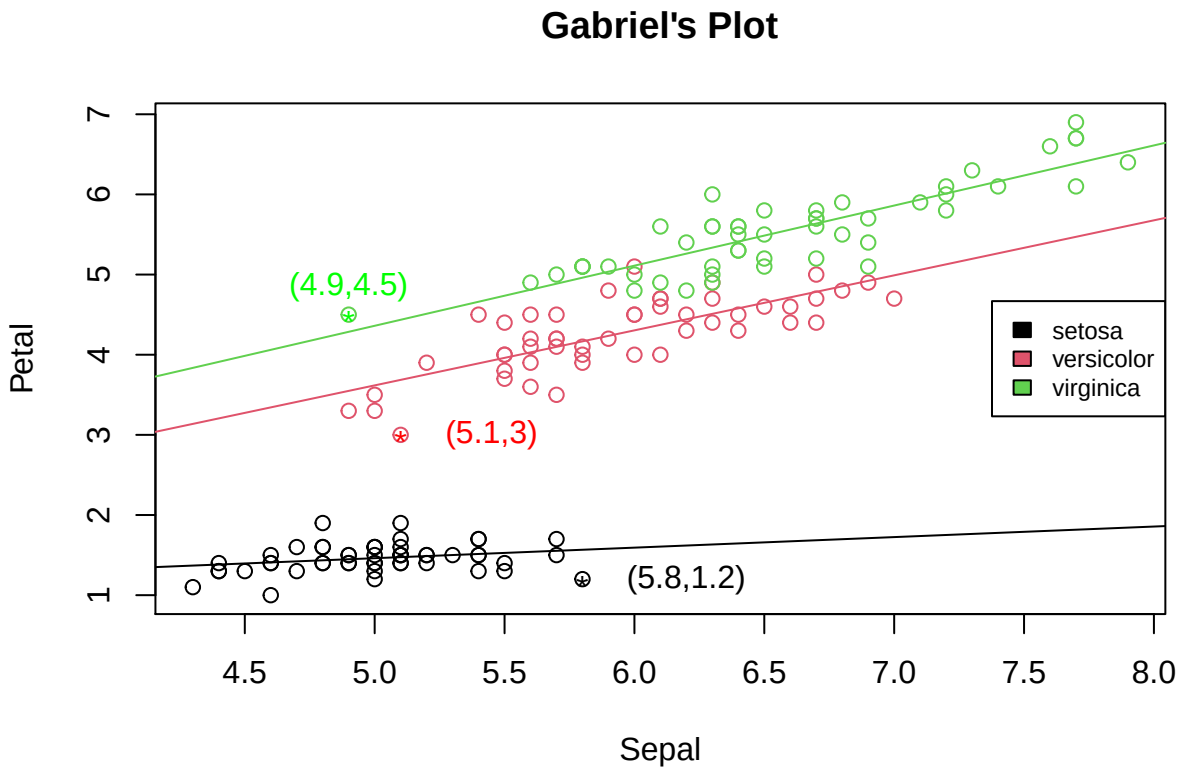
This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

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.

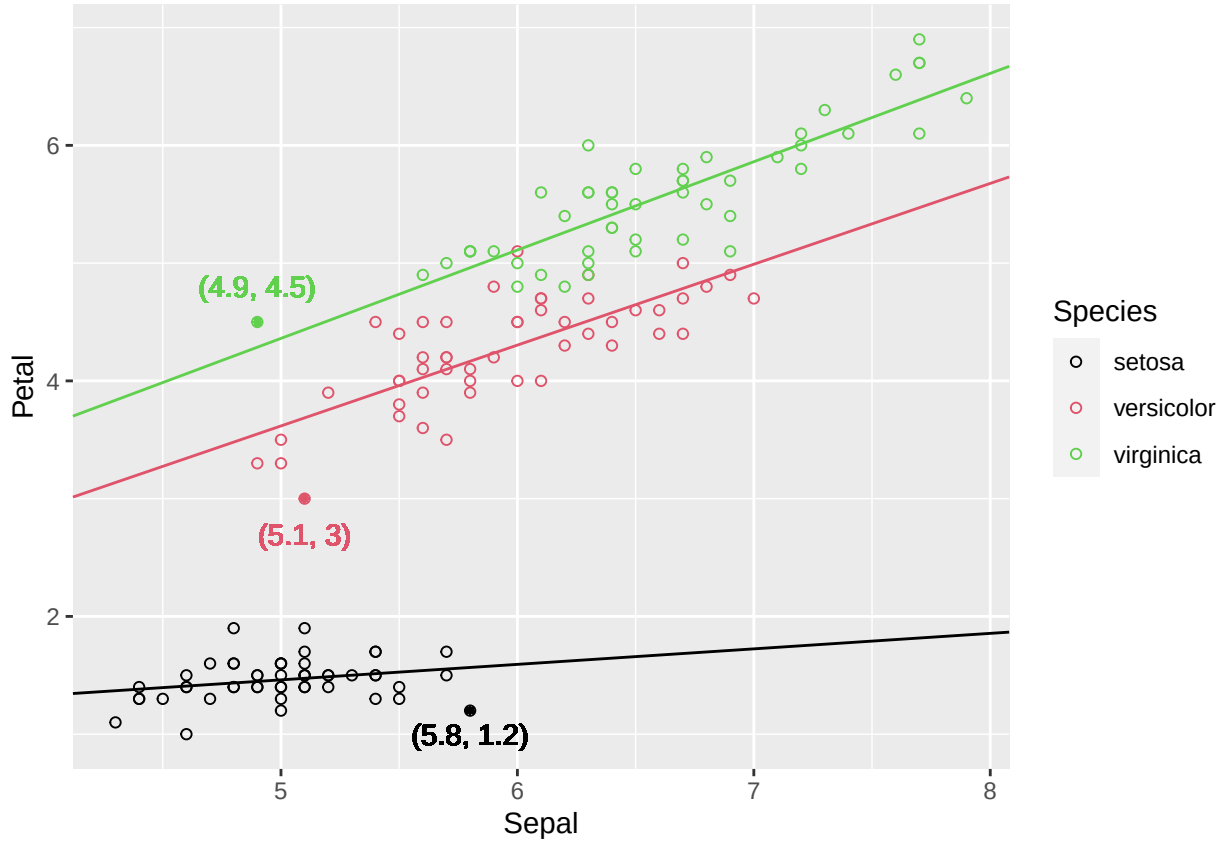
```
# Base plot
plot(iris$Sepal.Length, iris$Petal.Length, col = iris$Species,
    xlab = "Sepal", ylab = "Petal", main = "Gabriel's Plot")

# 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 = 0.75)

# Add points and text
points(iris$Sepal.Length[15], iris$Petal.Length[15], pch = "*",
    col = "black")
text(iris$Sepal.Length[15] + 0.4, iris$Petal.Length[15], "(5.8,1.2)",
    col = "black")
points(iris$Sepal.Length[99], iris$Petal.Length[99], pch = "*",
```

```
    col = "red")
text(iris$Sepal.Length[99] + 0.35, iris$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] + 0.35, "(4.9,4.5)",
    col = "green")
```




```
library(ggplot2)

rl_s <- lm(iris$Petal.Length[iris$Species == "setosa"] ~ iris$Sepal.Length[iris$Species ==
    "setosa"])
rl_ver <- lm(iris$Petal.Length[iris$Species == "versicolor"] ~
    iris$Sepal.Length[iris$Species == "versicolor"])
rl_vir <- lm(iris$Petal.Length[iris$Species == "virginica"] ~
    iris$Sepal.Length[iris$Species == "virginica"])

# ggplot(iris) + geom_point(aes(x=Sepal.Length,
# y=Petal.Length, col=Species)) +
# geom_smooth(aes(x=Sepal.Length, y=Petal.Length,
# col=Species), method='lm', se=0, fullrange=T)

# ggplot(iris) + geom_point(aes(x=Sepal.Length,
# y=Petal.Length, col=Species), shape=1) + geom_abline(slope
# = coef(rl_s)[2], intercept = coef(rl_s)[1], color = 'red')
```

```
# + geom_abline(slope = coef(rl_ver)[2], intercept =
# coef(rl_ver)[1], color = 'green') + geom_abline(slope =
# coef(rl_vir)[2], intercept = coef(rl_vir)[1], color =
# 'blue') + geom_text(aes(x=5.8, y=1, label='(5.8, 1.2)'),
# col='red') + geom_point(aes(x=5.8, y=1.2), col='red',
# shape=8, size=0.5) + geom_text(aes(x=5.1, y=2.7,
# label='(5.1, 3)'), col='green') + geom_point(aes(x=5.1,
# y=3), col='green', shape=8, size=0.5) +
# geom_text(aes(x=4.9, y=4.5, label='(4.9, 4.5)'),
# col='blue') + geom_point(aes(x=4.9, y=4.5), col='blue',
# shape=8, size=0.5)

ggplot(iris) + geom_point(aes(x = Sepal.Length, y = Petal.Length,
    col = Species), shape = 1) + geom_abline(slope = coef(rl_s)[2],
    intercept = coef(rl_s)[1], color = 1) + geom_abline(slope = coef(rl_ver)[2],
    intercept = coef(rl_ver)[1], color = 2) + geom_abline(slope = coef(rl_vir)[2],
    intercept = coef(rl_vir)[1], color = 3) + geom_text(aes(x = 5.8,
    y = 1, label = "(5.8, 1.2)"), col = 1) + geom_point(aes(x = 5.8,
    y = 1.2), col = 1, shape = 8, size = 0.5) + geom_text(aes(x = 5.1,
    y = 2.7, label = "(5.1, 3)"), col = 2) + geom_point(aes(x = 5.1,
    y = 3), col = 2, shape = 8, size = 0.5) + geom_text(aes(x = 4.9,
    y = 4.8, label = "(4.9, 4.5)"), col = 3) + geom_point(aes(x = 4.9,
    y = 4.5), col = 3, shape = 8, size = 0.5) + scale_color_manual(values = c(1,
    2, 3)) + labs(x = "Sepal", y = "Petal")
```



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.gmond.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.

```
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")

wtid$P99[wtid$Year == 1993]
```

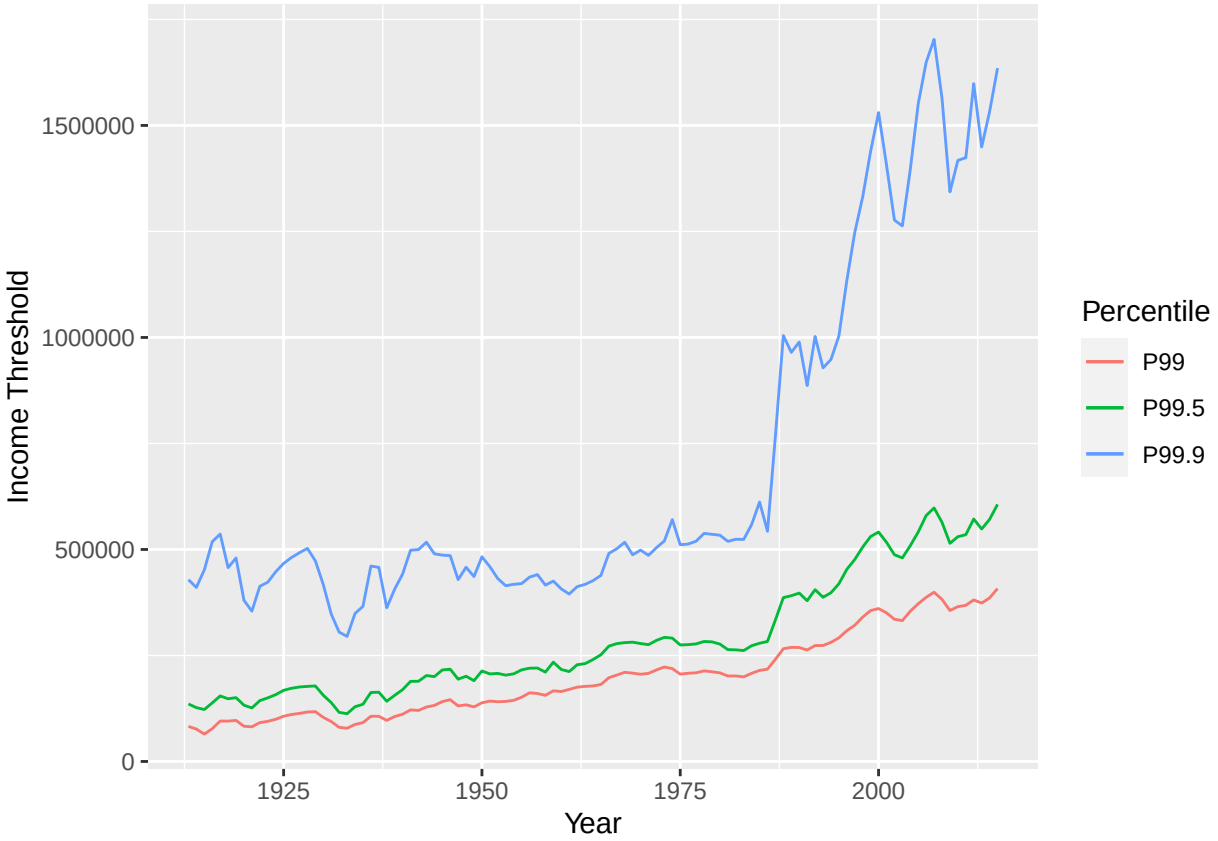
```
## [1] 273534.9
```

```
wtid$P99.5[wtid$Year == 1942]
```

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

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.

```
n <- nrow(wtid)
wtid_new <- data.frame(Year = rep(wtid$Year, 3), Income = c(wtid$P99,
    wtid$P99.5, wtid$P99.9), Income_Threshold = c(rep("P99",
    n), rep("P99.5", n), rep("P99.9", n)))

# ggplot(wtid_new) + geom_line(aes(x=Year, y=Income,
# col=Income_Threshold)) + labs(y='Income Threshold',
# col='Percentile')

ggplot(wtid) + geom_line(aes(x = Year, y = P99, col = "P99")) +
    geom_line(aes(x = Year, y = P99.5, col = "P99.5")) + geom_line(aes(x = Year,
    y = P99.9, col = "P99.9")) + labs(y = "Income Threshold",
    col = "Percentile")
```



The income threshold has increased gradually for the 99th and 99.5th percentile from the years 1913 to 2015 showing a slightly higher increase rate from around 1987. For the income threshold at the 99.9th percentile, the income threshold was sporadic with a slight increase from the years 1913 to 1987, but drastically increased from then to 2015.

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.

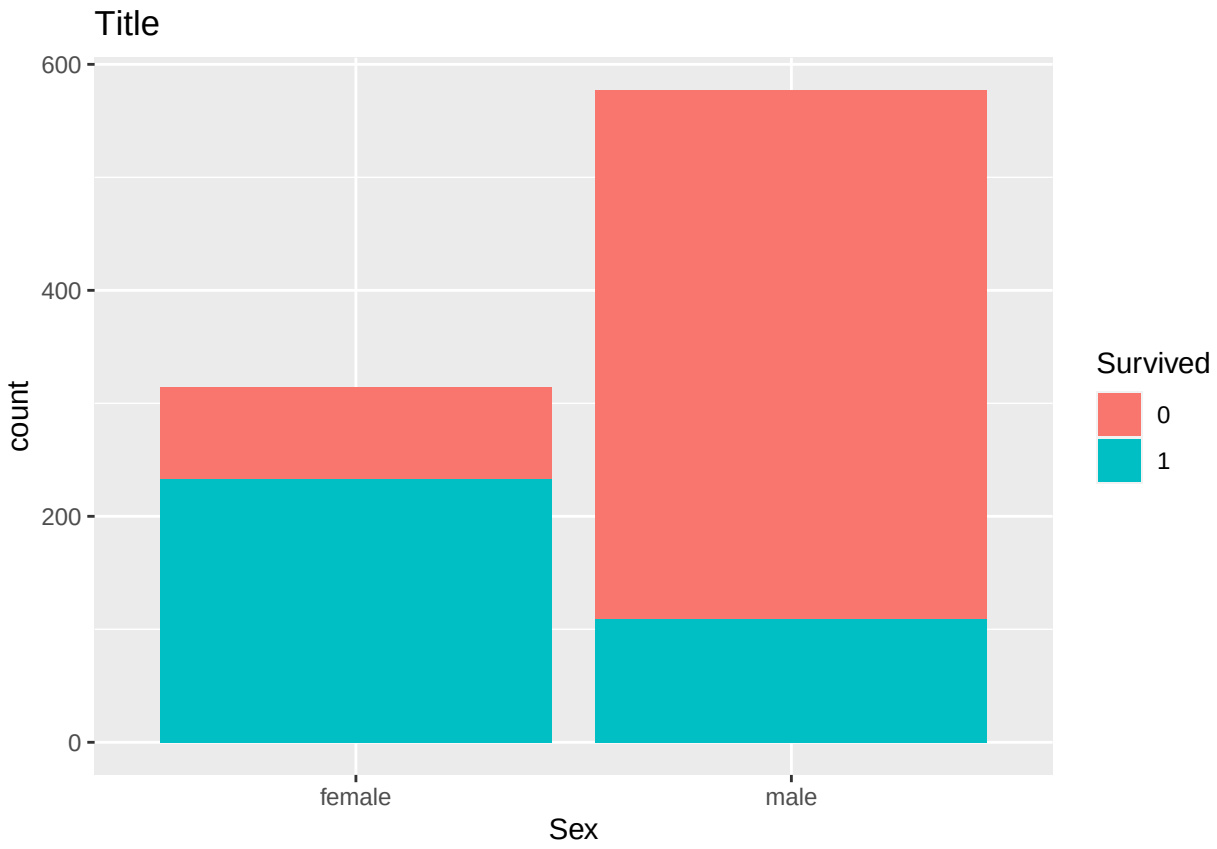
Tasks

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

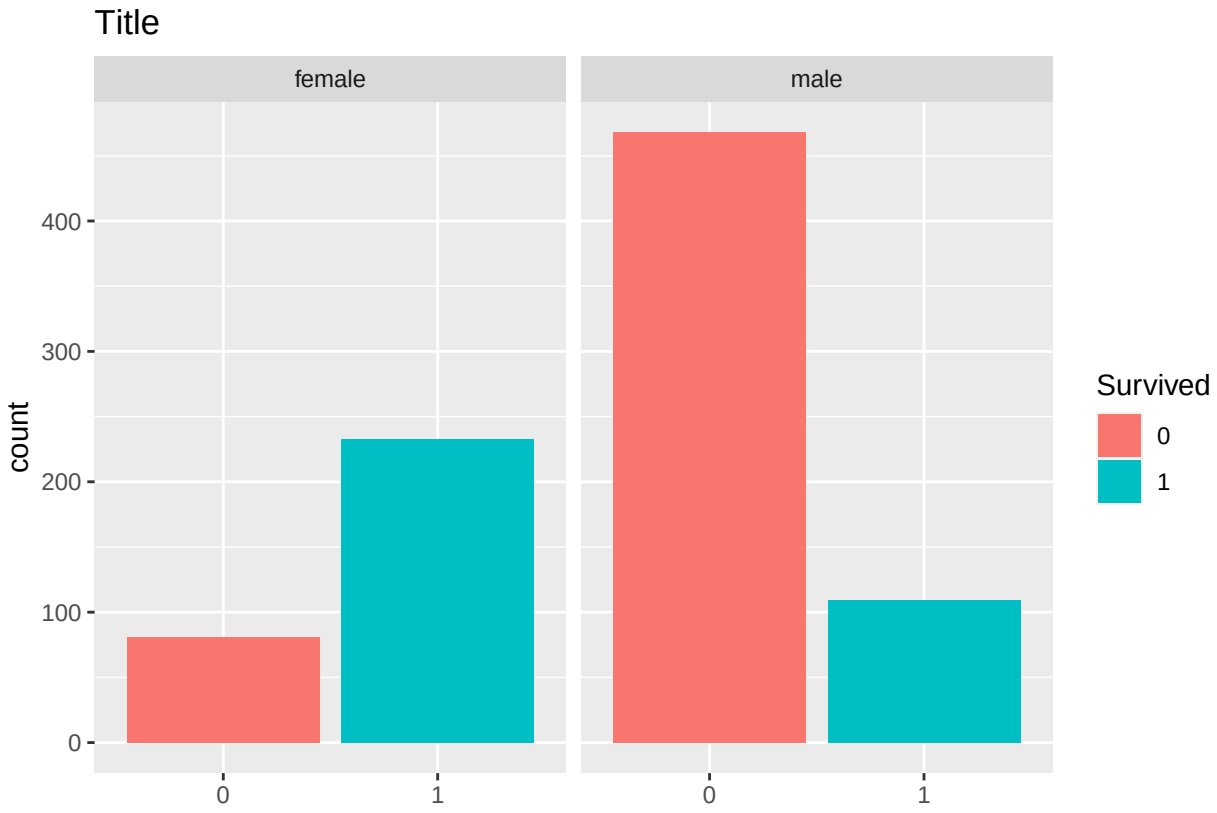
```
# Read in data
titanic <- read.table("Titanic.txt", header = TRUE, as.is = TRUE)
head(titanic)
```

```
##   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 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")
```



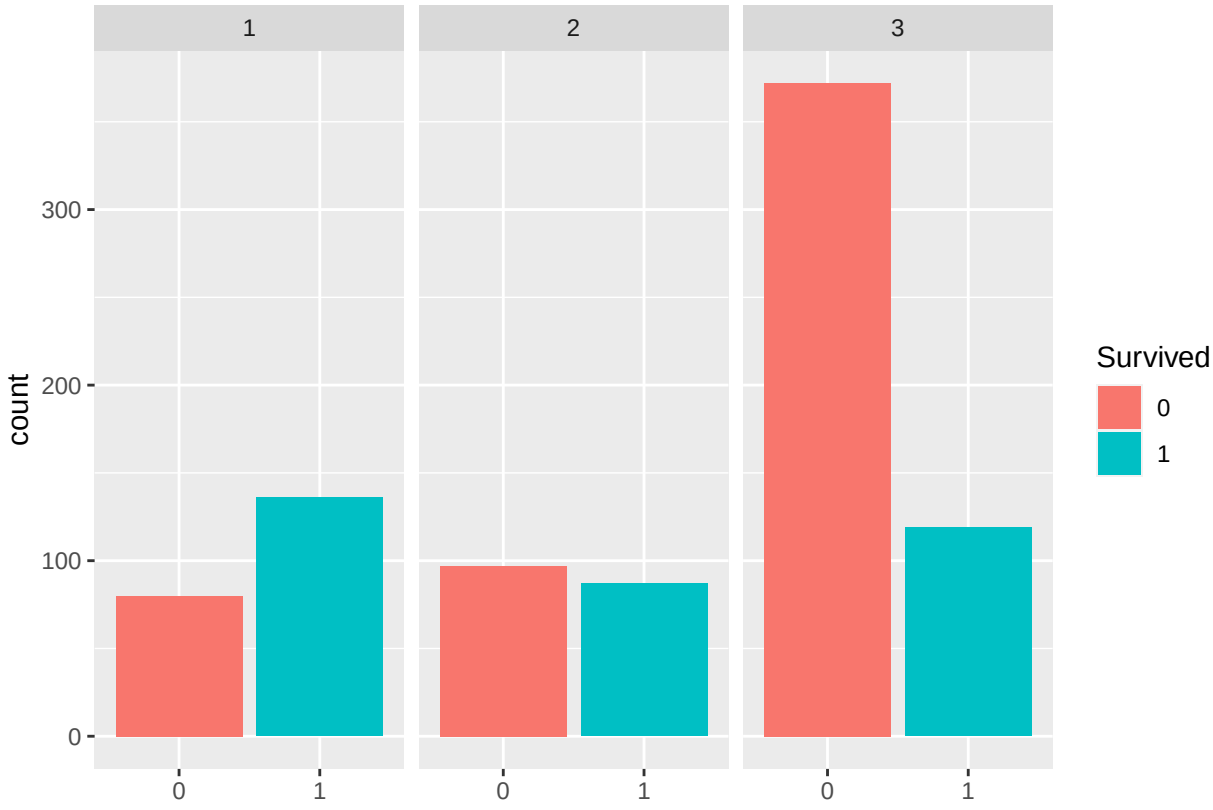
```
# plot 2
ggplot(data = titanic) + geom_bar(aes(x = factor(Survived), fill = factor(Survived))) +
    facet_grid(~Sex) + labs(title = "Title", fill = "Survived",
    x = "")
```



Plot 1 is displaying the number of people that survived and did not survive the Titanic in the same bar with two different colors - blue for those who did and red for those who didn't - and divided by Sex into two bars. Plot 2 is showing the same infomrtaion as Plot 1, but it separates the number of casualties and survivors into separate bars.

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

```
ggplot(titanic) + geom_bar(aes(x = factor(Survived), fill = factor(Survived))) +
    facet_grid(~Pclass) + labs(x = "", fill = "Survived")
```



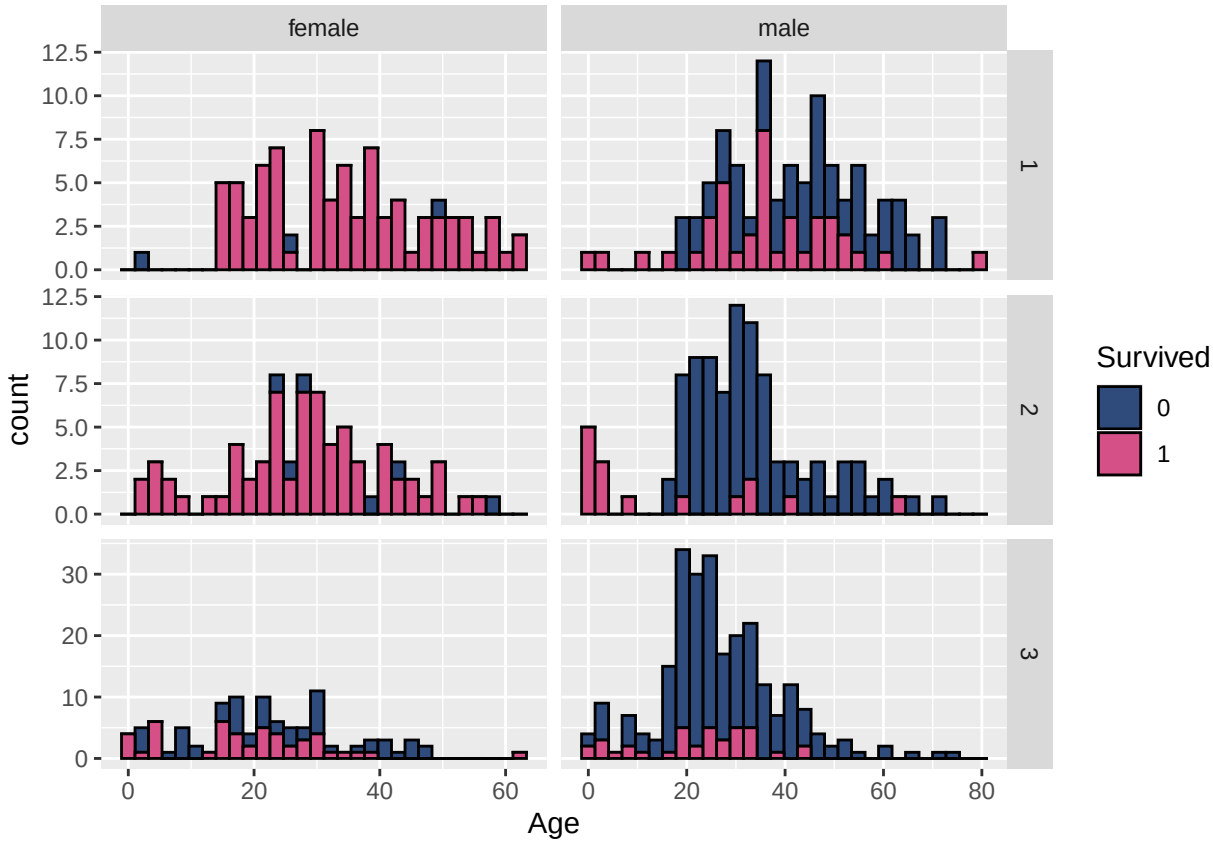

```
# ggplot(titanic) + geom_bar(aes(x=Sex,
# fill=factor(Survived))) + facet_grid(~Pclass) + labs(x='',
# fill='Survived')
```

This plot shows the number of people that survived and did not survive the titanic in each passenger class (1, 2 and 3).

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

```
# colnames(titanic)
ggplot(titanic) + geom_histogram(aes(x = Age, y = , fill = factor(Survived)),
    col = 1, bins = 30) + facet_grid(Pclass ~ Sex, scales = "free") +
    labs(fill = "Survived") + scale_fill_manual(values = c("#2f4b7c",
    "#d45087", "#ff7c43"))
```

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

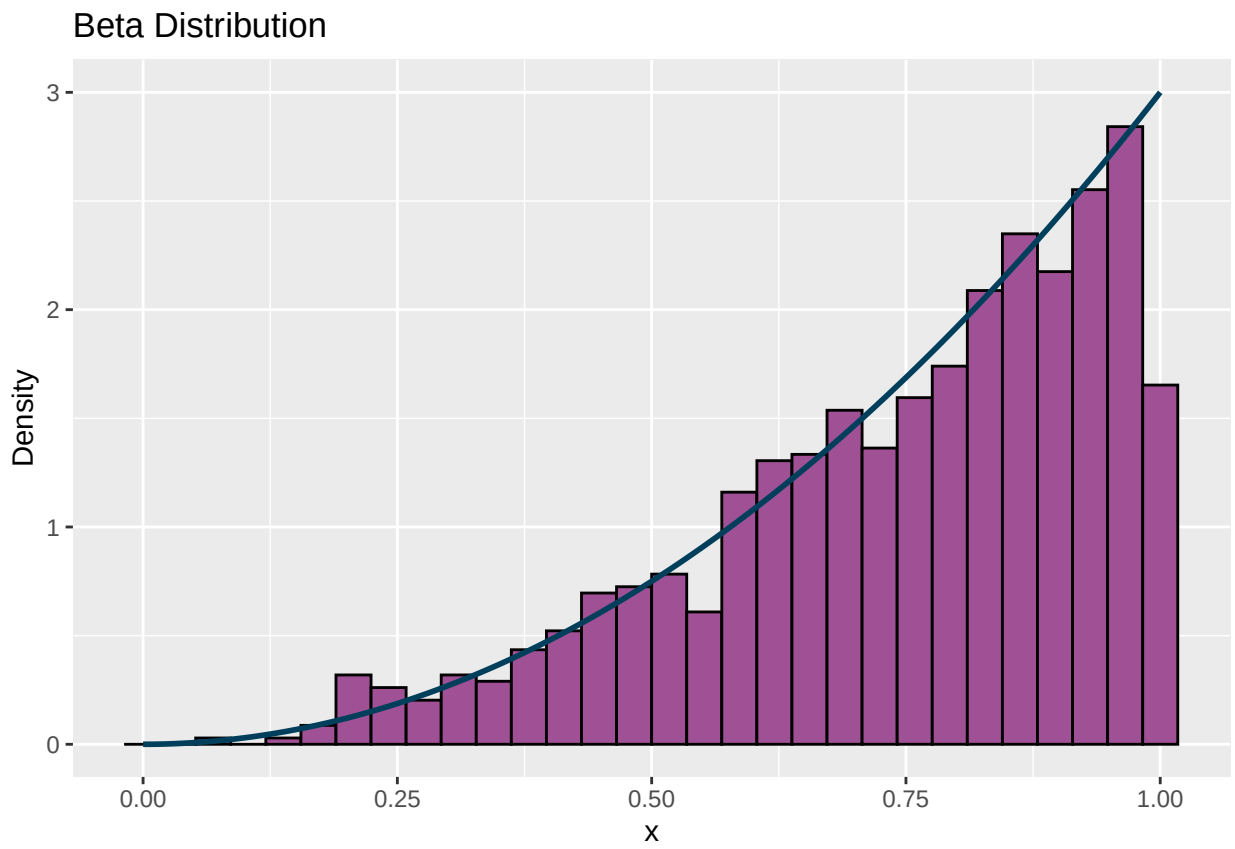


The graph shows the number of people that survived and did not survive for each age grop divided by Pclass and Sex. The number of Survivors is indicated by blue (0) and those who didn't is indicated by pink (1). The plots indicate that more number of females suvived than men across all classes. For both men and women, most of the casualties belonged to class 3.

Part 4 (Simulating and Graphing Probability Density)

7) Simulate a $n = 1000$ random draws from a beta distribution with parameters $\alpha = 3$ and $\beta = 1$. 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 = 0.01)
sim_beta = data.frame(x.var = rbeta(1000, 3, 1))
sim_beta_density = data.frame(x = x, f = dbeta(x, 3, 1))

ggplot(sim_beta) + geom_histogram(aes(x = x.var, y = ..density..),
    fill = "#a05195", col = 1, bins = 30) + geom_line(sim_beta_density,
    mapping = aes(x = x, y = f), col = "#003f5c", size = 1) +
    labs(x = "x", y = "Density", title = "Beta Distribution")
```


Beta Distribution
Density
3
2
1
0
0.00
0.25
0.50
0.75
1.00
x