ggplot Practice

Part 1 (Iris)

Background

Background: Edgar Anderson's Iris Data

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

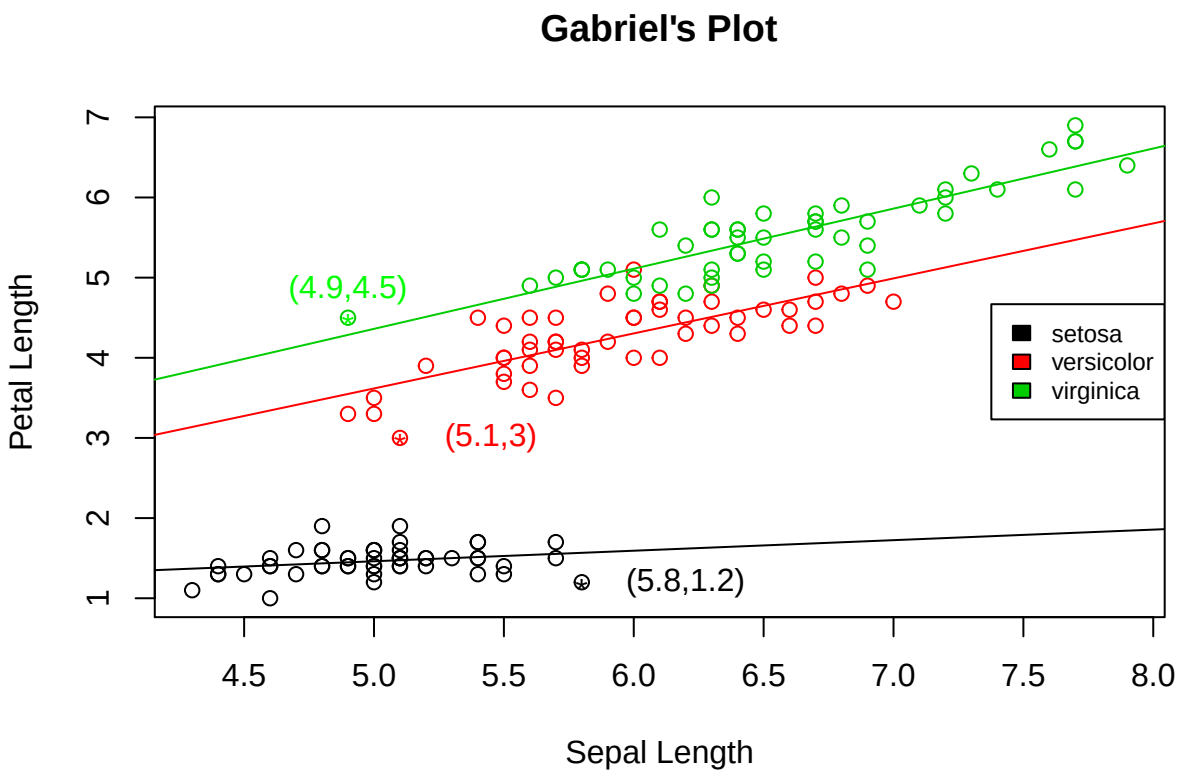
The purpose of this task is to construct a complex plot using both base **R** graphics and **ggplot**. Consider the follwoing base **R** plot.

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

# 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")
```



1) Produce the exact same plot from above using `ggplot` as opposed to Base **R** graphics. 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.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.1.3
```

```
## Plot.
```

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://topincomes.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.

```
setwd("~/Desktop/Data")
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) Plot the three percentile levels against time using `ggplot`. 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. Remember `library(ggplot2)`.

```
## Plot
```

Part 3 (Titanic)

Background

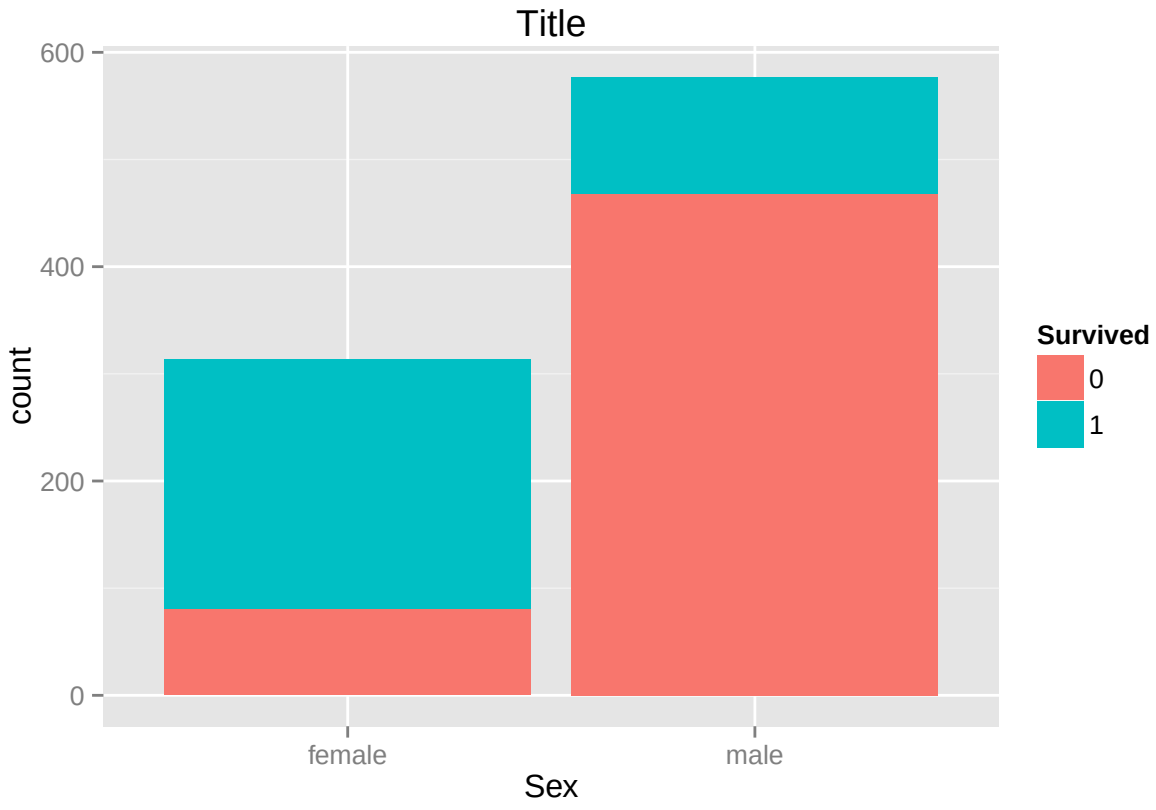
In this part we'll be studying a data set which provides information on the survival rates of passengers on the fatal voyage of the ocean liner *Titanic*. The dataset provides information on each passenger including, for example, economic status, sex, age, cabin, name, and survival status. This is a 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

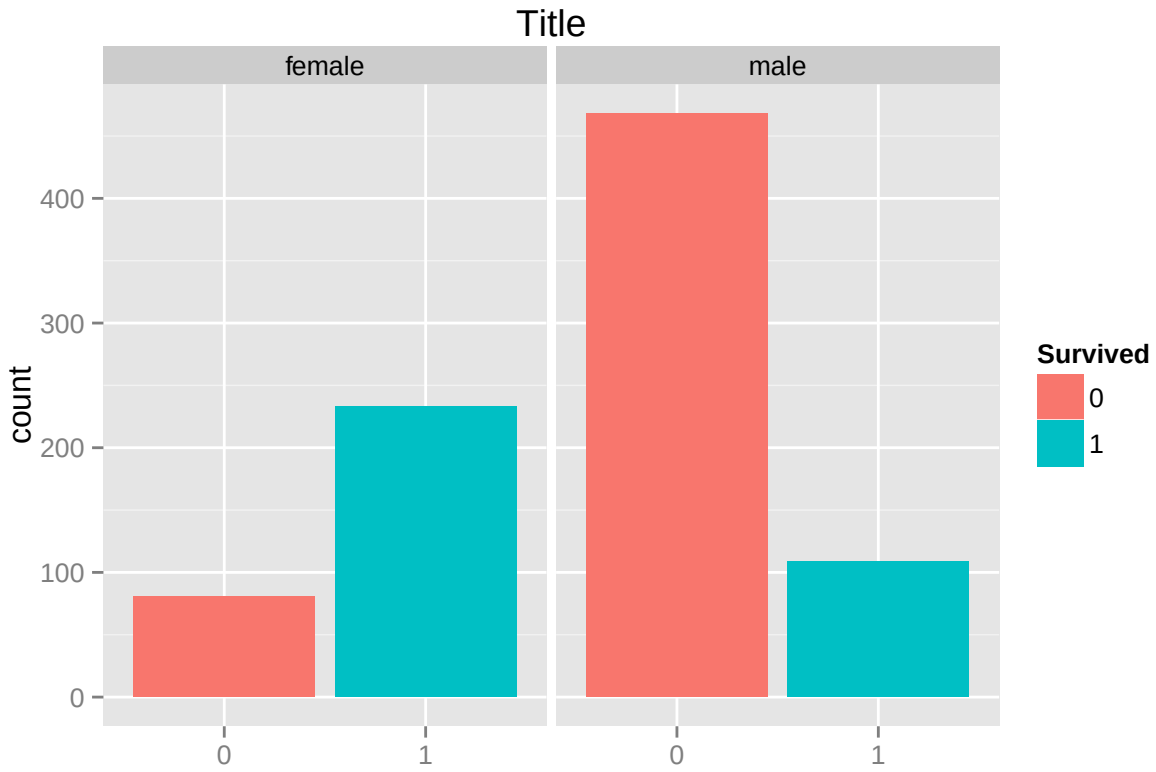
4) Run the following code and describe what the two plots are producing

```
# Read in data
setwd("~/Desktop/Data")
titanic <- read.csv("Titanic.txt", header = TRUE, as.is = TRUE)

# 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="")
```



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.

```
# Plots
```

6) Create a new variable of the **titanic** dataframe called **Title** that gives the appropriate title of each passenger. The variable **Title** should have 5 levels: **Miss**, **Mrs**, **Mr**, **Master**,and **Other**. Display the first 10 entries of the new variable **Title**.

```
# Solution
```

7) Create one more plot of that shows survival rates with the passenger's **Title**.

```
# Plots
```