Homework 5: ggplot Practice

Due 11:59pm, November 14, 2018

Instructions: Please submit the pdf file on Gradescope and Rmd file on Canvas.

Part 1 (Iris)

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

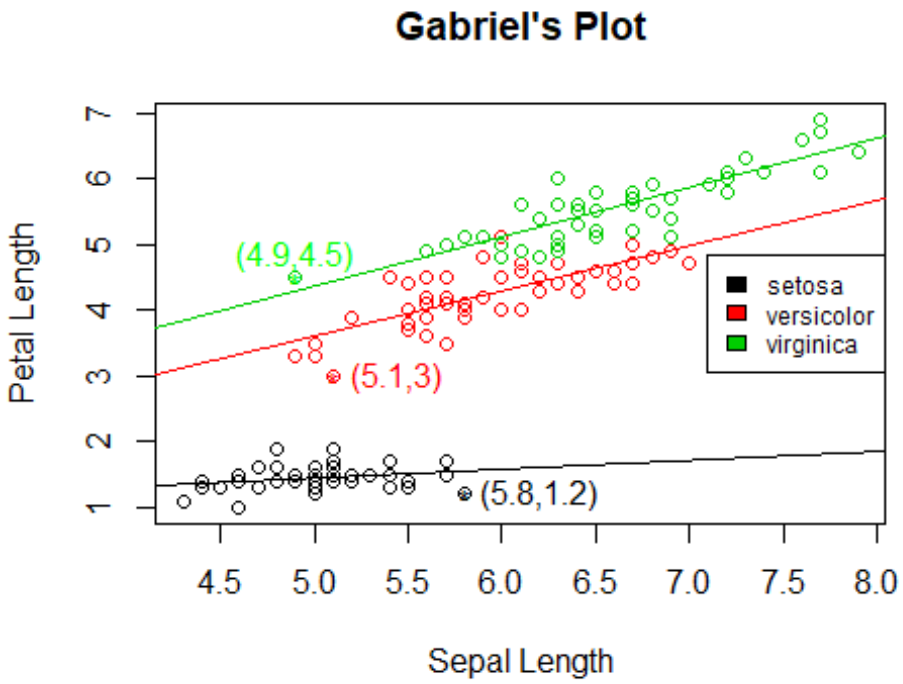
The purpose of this task is to construct a complex plot using both base **R** graphics and **ggplot**. Consider the follwoing base **R** plot.

```
# Base plot
plot(iris$Sepal.Length,iris$Petal.Length,col=iris$Species,xlab="Sepal Length"
,ylab="Petal Length",main="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$Species),fill = 1:length(levels(iris$Specie
s)), 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 = "gre
en")
```



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.

```
### your code goes here
lm1 <- lm(iris[iris[,"Species"]=="setosa","Petal.Length"]~iris[iris[,"Species
"]=="setosa","Sepal.Length"] )
lm2 <- lm(iris[iris[,"Species"]=="versicolor","Petal.Length"]~iris[iris[,"Spe
cies"]=="versicolor","Sepal.Length"])
lm3 <- lm(iris[iris[,"Species"]=="virginica","Petal.Length"]~iris[iris[,"Spec
ies"]=="virginica","Sepal.Length"])


library('ggplot2')
ggplot(data = iris) +
  geom_point(mapping = aes(x=iris$Sepal.Length,y=iris$Petal.Length,col=iris$S
pecies),
             fill = "white", size = 2, shape = 21) +
  geom_abline(intercept = coef(lm1)[[1]], slope =coef(lm1)[[2]], color="orang
e", size =0.5) +
  geom_abline(intercept = coef(lm2)[[1]], slope =coef(lm2)[[2]], color="green
", size =0.5) +
  geom_abline(intercept = coef(lm3)[[1]], slope =coef(lm3)[[2]], color="blue"
, size =0.5) +
```

```
  geom_point(mapping = aes(x=iris$Sepal.Length[15], y=iris$Petal.Length[15]),
             shape = '*', color = "orange", size =5) +
  geom_point(mapping = aes(x=iris$Sepal.Length[99], y=iris$Petal.Length[99]),
             shape = '*', color = "green", size =5) +
  geom_point(mapping = aes(x=iris$Sepal.Length[107], y=iris$Petal.Length[107]
),
             shape = '*', color = "blue", size =5) +
  geom_text(mapping = aes(x=iris$Sepal.Length[15]+.4,y=iris$Petal.Length[15])
,label = "(5.8,1.2)",
             color="orange", size = 4)+
  geom_text(mapping = aes(x=iris$Sepal.Length[99]+.35,y=iris$Petal.Length[99]
),label = "(5.1,3)",
             color="green", size = 4)+
  geom_text(mapping = aes(x=iris$Sepal.Length[107],y=iris$Petal.Length[107]+.
35),label = "(4.9,4.5)",
             color="blue", size = 4) +
  # set legend manully
  scale_color_manual(name="Name of Species",
                     labels=c("setosa","versicolior","viginica"),
                     values=c("orange","green","blue")) +
  labs(title = "scatter plot and regression lines for species",
       x = "Sepal Length", y = "Petal Length")
```

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.

```
setwd("C:/Users/Alan_/Desktop/Semester1 Courses/STAT computing/Week 9/assignment5")
wtid <- read.csv("wtid-report.csv", as.is = TRUE)
### your code goes here

# can use number of data to extract columns instead the long column names

new_df <- wtid[,c(2,3,4,5)]
names(new_df) <- c("year","P99","P99.5","P99.9")
head(new_df)

##   year      P99    P99.5    P99.9
## 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_1993 <- new_df[new_df[,"year"] == 1993,"P99"]
P99.5_1942 <- new_df[new_df[,"year"] == 1942,"P99.5"]
P99_1993

## [1] 273534.9

P99.5_1942

## [1] 189140.6
```

P99 in 1993 is 2735.34. P99.5 in 1943 is 189140.6

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

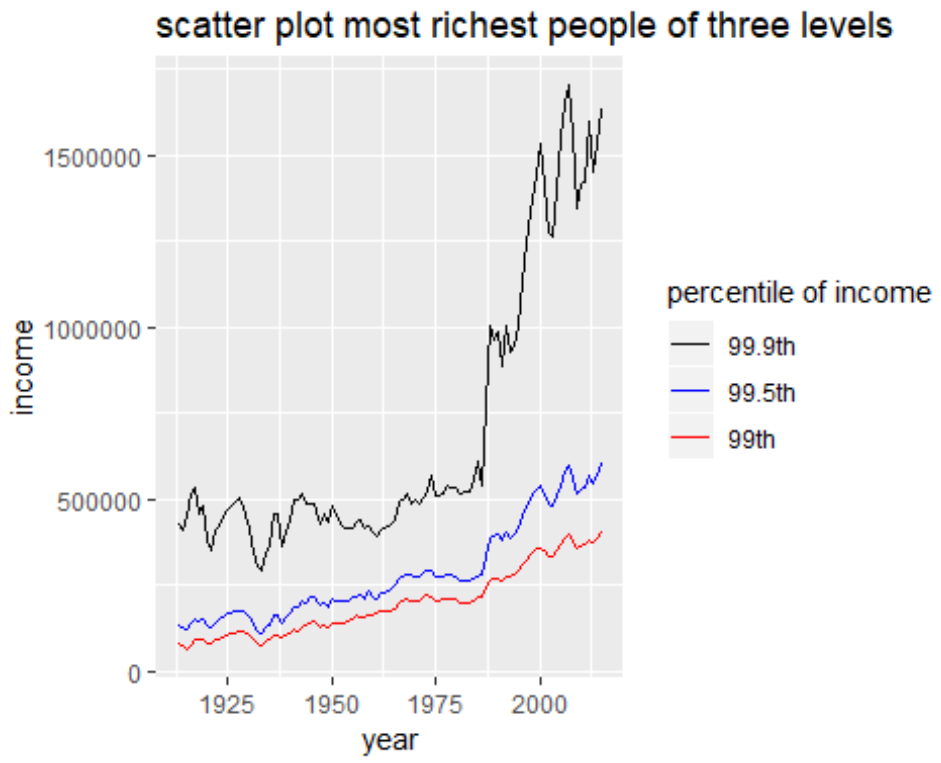
displayed that describes the multiple time series plot. Write one or two sentences describing how income inequality has changed throughout time. Remember `library(ggplot2)`.

```
### your code goes here
library(ggplot2)

# in order to get the legend the color should be put into aes()


ggplot(data = new_df) +
  geom_line(mapping = aes(x=new_df$year,y=new_df$P99,color="red")) +
  geom_line(mapping = aes(x=new_df$year,y=new_df$P99.5,color="blue")) +
  geom_line(mapping = aes(x=new_df$year,y=new_df$P99.9,color="black")) +


   # set legend manully
  scale_color_manual(name="percentile of income",
                     labels=c("99.9th","99.5th","99th"),
                     values=c("black","blue","red")) +
  labs(title = "scatter plot most richest people of three levels",
       x = "year", y = "income")
```



As the year increase, the income of all three types of richest people increase, but after 1985 the increment of people in 99.9th percentile group is much larger than other groups and the differences between this groups and the others are also come larger when the year is increase. It shows that even in the most richest people, whose salary are already Top 1% in the world, the income inequality also becomes larger throughtout time, especially after 1985.