Homework 5_Zhiyuan

Zhiyuan Du

10/22/2020

Problem 3

Firstly, let's import the data into R and clean the data.

```
# Import the data
#unzip("Edstats_csv.zip", exdir = "D:/VT/Rstudio/directory")
Edstat=read.csv("D:/VT/Rstudio/directory/EdStatsData.csv",header = TRUE)

#clean the columns
Edstat <- Edstat[,apply(!is.na(Edstat), 2, sum) != 0]
#clean the rows
Edstat<- Edstat[apply(!is.na(Edstat), 1, sum) != 4, ]
```

Here, I deleted the the rows and columns with all missing value. There were 886930 data points in the original data, and there are 357405 currently. Now let's choose China and United States and count the indicators of them.

```
#Count the length of indicator
China=Edstat[Edstat$ï..Country.Name=="China",]
US=Edstat[Edstat$ï..Country.Name=="United States",]

#create table
tablep1=matrix(c(length(na.omit(China$Indicator.Name)),length(na.omit(US$Indicator.Name))),ncol=2)
colnames(tablep1)=c("China","United States")
tablep1
```

```
##      China United States
## [1,]  1703          1870
```

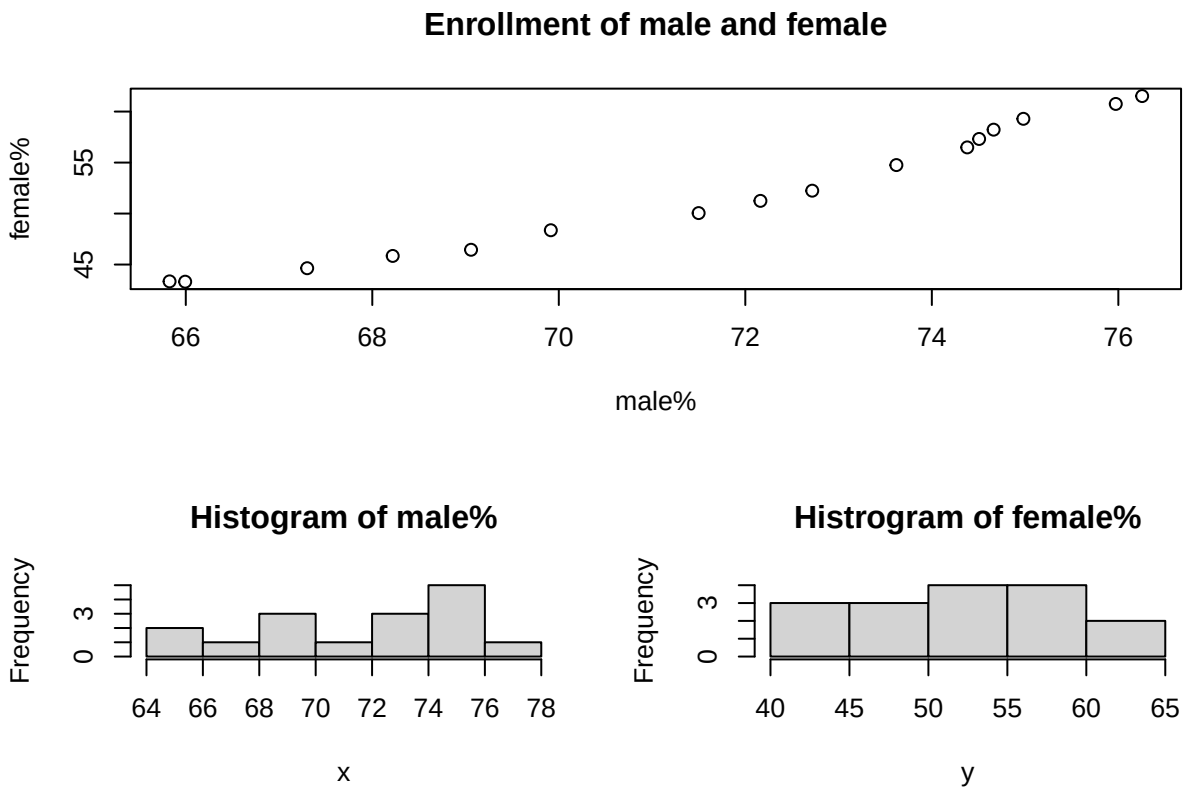
Problem 4

Let's make the plots for two varibales, one is "Adjusted net enrolment rate, primary, male (%)" in "Arab World" and one is"Adjusted net enrolment rate, primary, female (%)" in "Arab World".

```
#assign the value
x=as.numeric(Edstat[Edstat$ï..Country.Name=="Arab World"&Edstat$Indicator.Name
                    =="Adjusted net enrolment rate, primary, male (%)", 5:20 ])
y=as.numeric(Edstat[Edstat$ï..Country.Name=="Arab World"&Edstat$Indicator.Name
                    =="Adjusted net enrolment rate, primary, female (%)",5:20])
```

Now let's make the plots.

```
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE),widths=c(1,1), heights=c(2,1.5))
plot(x,y,xlab = "male%",ylab = "female%",main="Enrollment of male and female")
hist(x,main="Histogram of male%")
hist(y,main="Histrogram of female%")
```



Problem 5

Let's make the plots by ggplot2 functions.

```
datap5=data.frame(x,y)

#scatter plot
sc_plot=ggplot(datap5,mapping = aes(x=x,y=y))+
  geom_point()+
  labs(x="male%")+
  labs(y="female%")+
  labs(title="Enrollment of male and female")

#histogram1
histx=ggplot(datap5,aes(x=x))+
  geom_histogram()+
 labs(x="male%")
```

```r
#histogram2
histy=ggplot(datap5,aes(x=y))+
  geom_histogram()+
 labs(x="male%")

ggarrange(sc_plot,                                # First row with scatter plot
          ggarrange(histx, histy, ncol = 2), # Second row with histograms
          nrow = 2
          )
```