Homework 8

Hwasoo Shin

2019 10 29

Problem 3

We can first read the file and clean up the data.

```r
edstat<-read_csv("C:/Users/pc/Desktop/HWASOO/STUDY/StatPackage/Homework8/EdStats_csv/EdStatsData.csv")
summary(edstat)
#We can see that the last column is totally not available. Therefore, we will erase it
dim(edstat)
edstat<-edstat[,-70]
mastered<-edstat #Store the raw data file
checkna<-function(x){
  tf<-sum(is.na(x))<65
  return(tf)
}
idxed<-apply(edstat,1,checkna)
table(idxed) #We can see that the number of rows that have at least one value on year column is 354575.
edstat<-edstat[idxed,]
dim(edstat) #We will only get the data that have valid values.
table(edstat[,1])
edstatmex<-edstat%>%filter(`Country Name`=="Mexico") #Data of Mexico
edstatcan<-edstat%>%filter(`Country Name`=="Canada") #Data of Canada
edstatcom<-rbind(edstatcan,edstatmex) #combine two datasets
```

```
## [1] 886930     69
```

```
## [1] 354575     69
```

Table 1: Brief summary of Candada

Indicator Name	1970	1971	1972	1973
Length:1809	Min. :0.000e+00	Min. :0.000e+00	Min. :0.000e+00	Min. :0.000e+00
Class :character	1st Qu.:4.000e+00	1st Qu.:1.200e+01	1st Qu.:1.200e+01	1st Qu.:1.200e+01
Mode :character	Median :1.200e+01	Median :2.806e+04	Median :7.939e+03	Median :3.151e+04
NA	Mean :1.639e+09	Mean :5.594e+09	Mean :5.936e+09	Mean :6.599e+09
NA	3rd Qu.:4.700e+01	3rd Qu.:6.265e+05	3rd Qu.:5.630e+05	3rd Qu.:6.657e+05
NA	Max. :5.252e+11	Max. :5.468e+11	Max. :5.766e+11	Max. :6.168e+11
NA	NA's :1382	NA's :1676	NA's :1674	NA's :1676

Table 2: Brief summary of Mexico

Indicator Name	1970	1971	1972	1973
Length:1809	Min. :0.000e+00	Min. :0.000e+00	Min. :0.000e+00	Min. :0.000e+00
Class :character	1st Qu.:4.000e+00	1st Qu.:1.200e+01	1st Qu.:1.200e+01	1st Qu.:1.200e+01

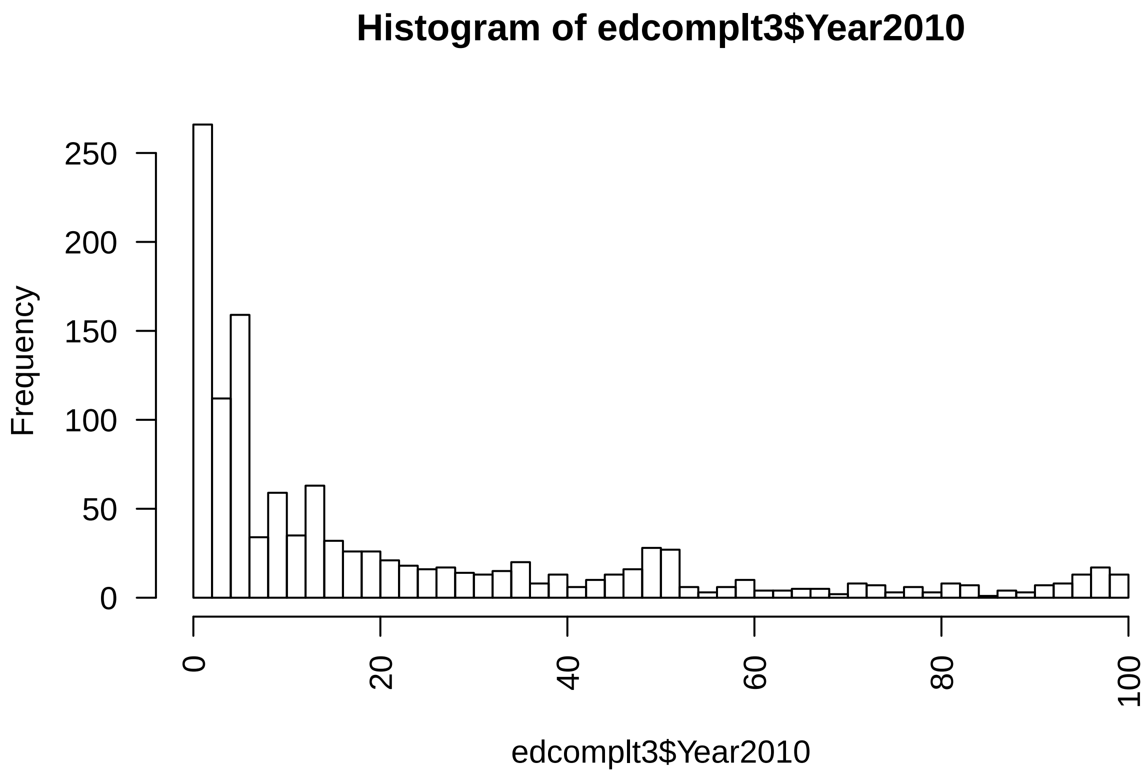
Indicator Name	1970	1971	1972	1973
Mode :character	Median :1.200e+01	Median :2.806e+04	Median :7.939e+03	Median :3.151e+04
NA	Mean :1.639e+09	Mean :5.594e+09	Mean :5.936e+09	Mean :6.599e+09
NA	3rd Qu.:4.700e+01	3rd Qu.:6.265e+05	3rd Qu.:5.630e+05	3rd Qu.:6.657e+05
NA	Max. :5.252e+11	Max. :5.468e+11	Max. :5.766e+11	Max. :6.168e+11
NA	NA’s :1382	NA’s :1676	NA’s :1674	NA’s :1676

Problem 4

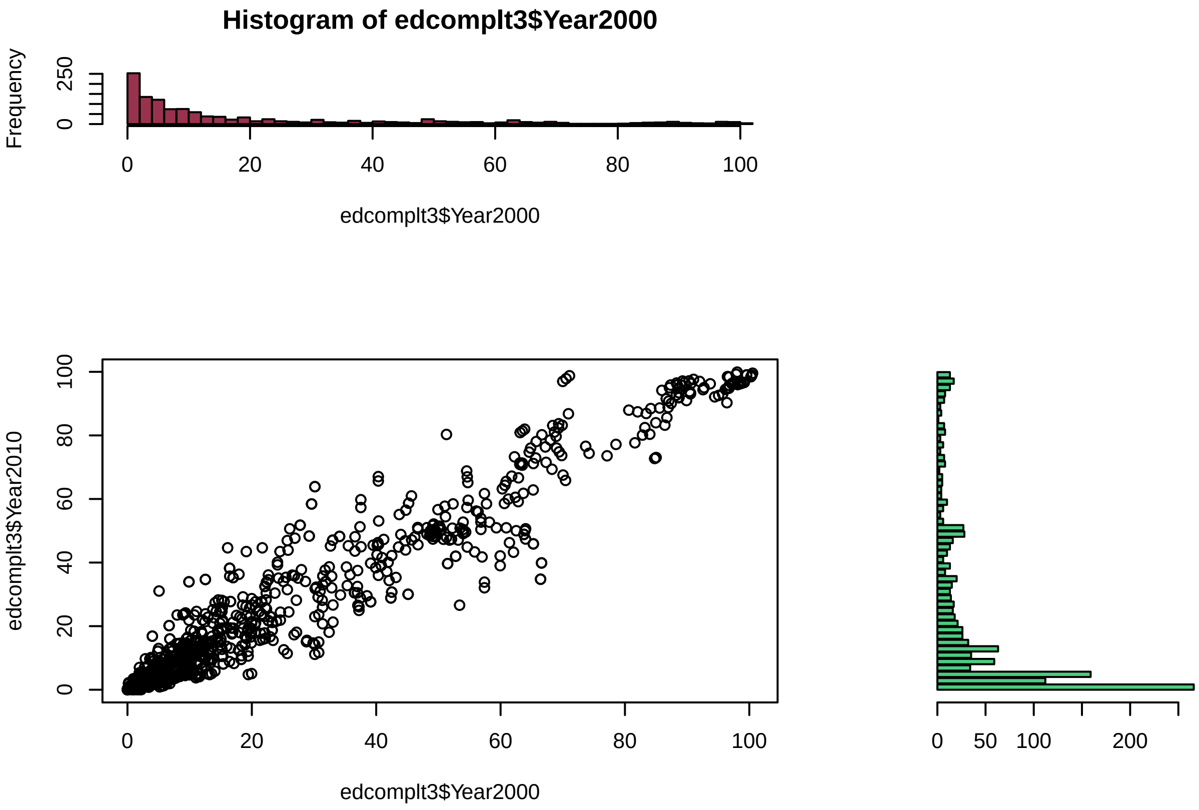
```
edcomplt<-edstatcom[,c("2000","2010")]
edcomplt2<-edcomplt[which(!is.na(edcomplt[,1])),]
edcomplt3<-edcomplt2[which(!is.na(edcomplt2[,2])),]
colnames(edcomplt3)<-c("Year2000","Year2010")
summary(edcomplt3$Year2010)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.000e+00 6.000e+00 4.200e+01 8.919e+09 3.515e+05 1.823e+12
```

```
edcomplt3<-edcomplt3 %>% filter(Year2010<100)
k<-hist(edcomplt3$Year2010,breaks=50,las=2)
```



```r
layout(rbind(c(2,2,0),c(1,1,3),c(1,1,3)))
plot(edcomplt3$Year2000,edcomplt3$Year2010)
hist(edcomplt3$Year2000,breaks=50,col=rgb(0.6,0.2,0.3))
barplot(k$counts,horiz=TRUE,col=rgb(0.3,0.8,0.5))
```

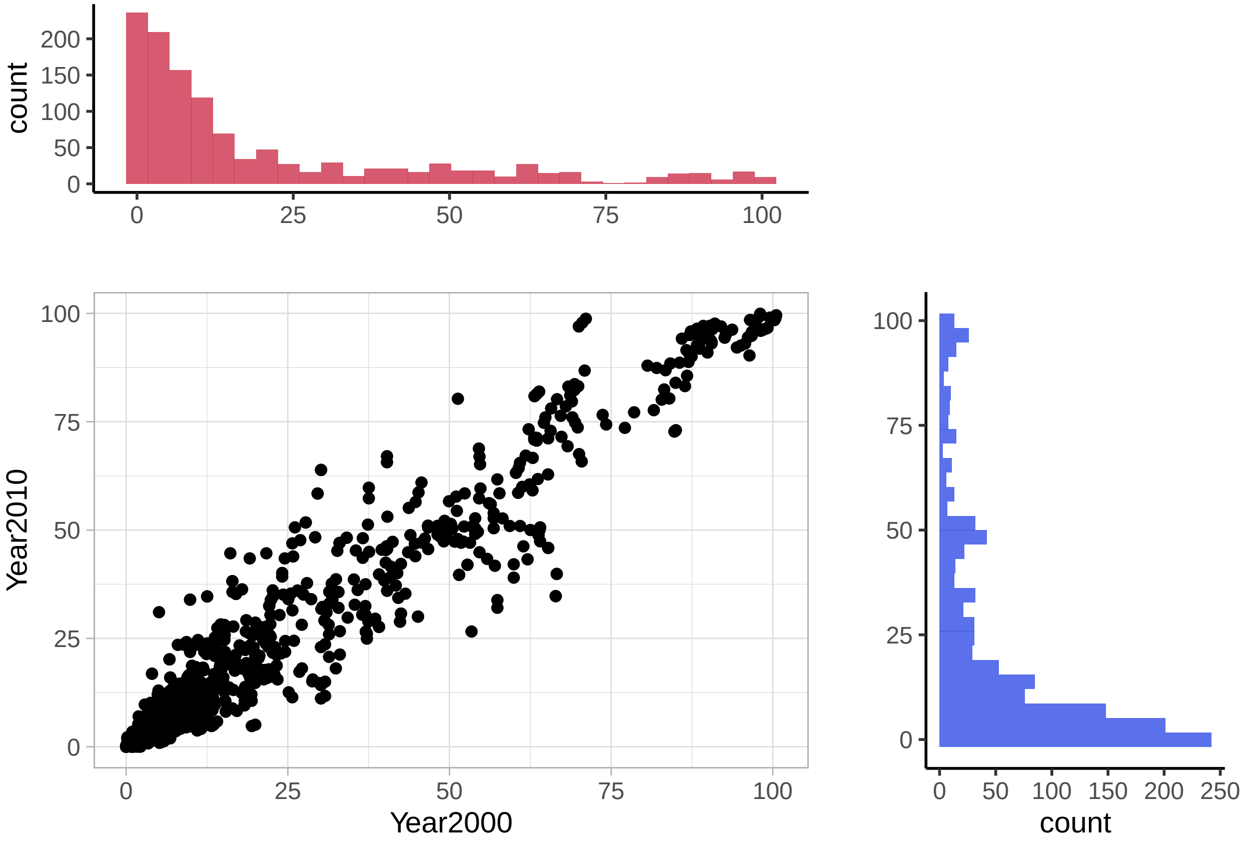


We can make a simple histogram and plots to put altogether. To make the plot clean, I selected values which variable “Year2010” is less than 100. We can see there is a high positive correlation between variables “Year2000” and “Year2010”. Also, both the variables’ values are more concentrated near 0 than other larger values.

Problem 5

```r
p1<-ggplot(data=edcomplt3,aes(x=Year2000,y=Year2010))+geom_point()+theme_light()
p2<-ggplot(data=edcomplt3,aes(x=Year2000))+geom_histogram(fill=rgb(0.8,0.2,0.3,0.8))+xlab("")+theme_cla
p3<-ggplot(data=edcomplt3,aes(x=Year2010))+geom_histogram(fill=rgb(0.2,0.3,0.9,0.8))+xlab("")+coord_fli
grid.arrange(p1,p2,p3,layout_matrix=rbind(c(2,2,NA),c(1,1,3),c(1,1,3)))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We can make a same plot with ggplot package.