Homework 8

Hwasoo Shin

2019 10 29

Problem 3

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

```
edstat<-read_csv("C:/Users/pc/Desktop/HWASOO/STUDY/StatPackage/Homework8/EdStats_csv/EdStatsData.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   `Country Name` = col_character(),
##   `Country Code` = col_character(),
##   `Indicator Name` = col_character(),
##   `Indicator Code` = col_character(),
##   `2015` = col_logical(),
##   `2016` = col_logical(),
##   `2017` = col_logical(),
##   `2020` = col_logical(),
##   `2025` = col_logical(),
##   `2030` = col_logical(),
##   `2035` = col_logical(),
##   `2040` = col_logical(),
##   `2045` = col_logical(),
##   `2050` = col_logical(),
##   `2055` = col_logical(),
##   `2060` = col_logical(),
##   `2065` = col_logical(),
##   `2070` = col_logical(),
##   `2075` = col_logical(),
##   `2080` = col_logical()
##   # ... with 5 more columns
## )
```

```
## See spec(...) for full column specifications.
```

```
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,]
```

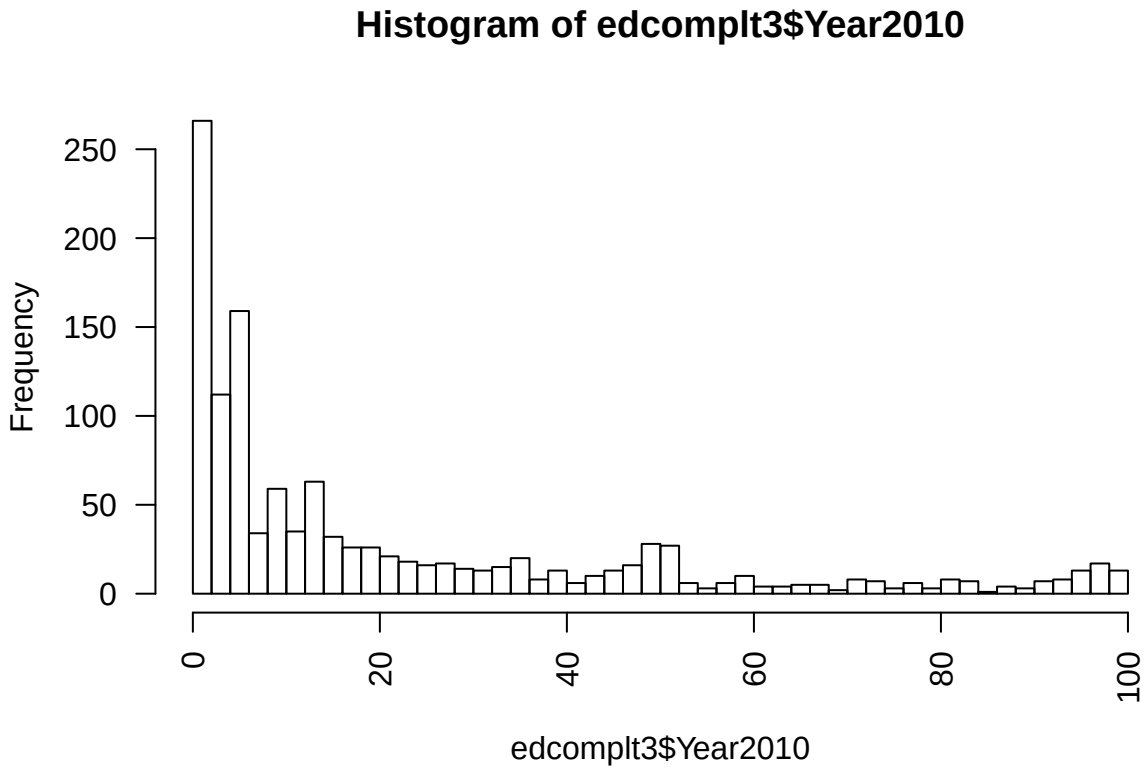
```r
dim(edstat) #We will only get the data that have valid values.
table(edstat[,1])
edstatmex<-edstat%>%filter(`Country Name`=="Mexico") #Data of Mexico
edstatcan<-edstat%>%filter(`Country Name`=="Canada") #Data of Canada
edstatcom<-rbind(edstatcan,edstatmex) #combine two datasets
```

Problem 4

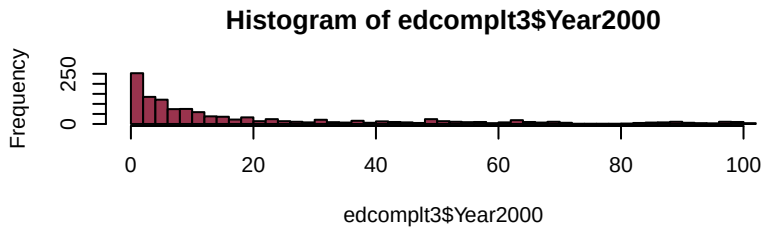
```r
edcomplt<-edstatcom[,c("2000","2010")]
edcomplt2<-edcomplt[which(!is.na(edcomplt[,1])),]
edcomplt3<-edcomplt2[which(!is.na(edcomplt2[,2])),]
colnames(edcomplt3)<-c("Year2000","Year2010")
summary(edcomplt3$Year2010)
```

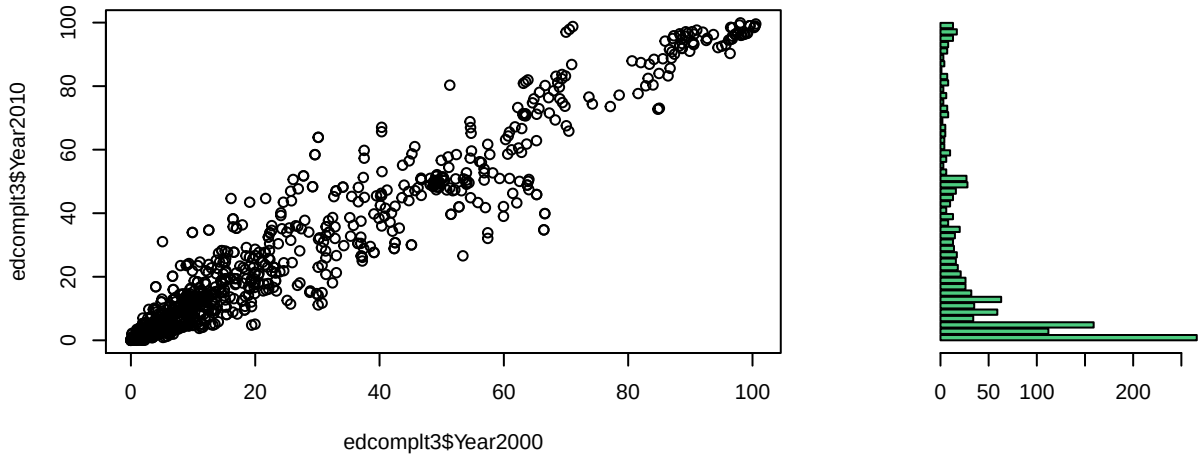
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.000e+00 6.000e+00 4.200e+01 8.919e+09 3.515e+05 1.823e+12
```

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



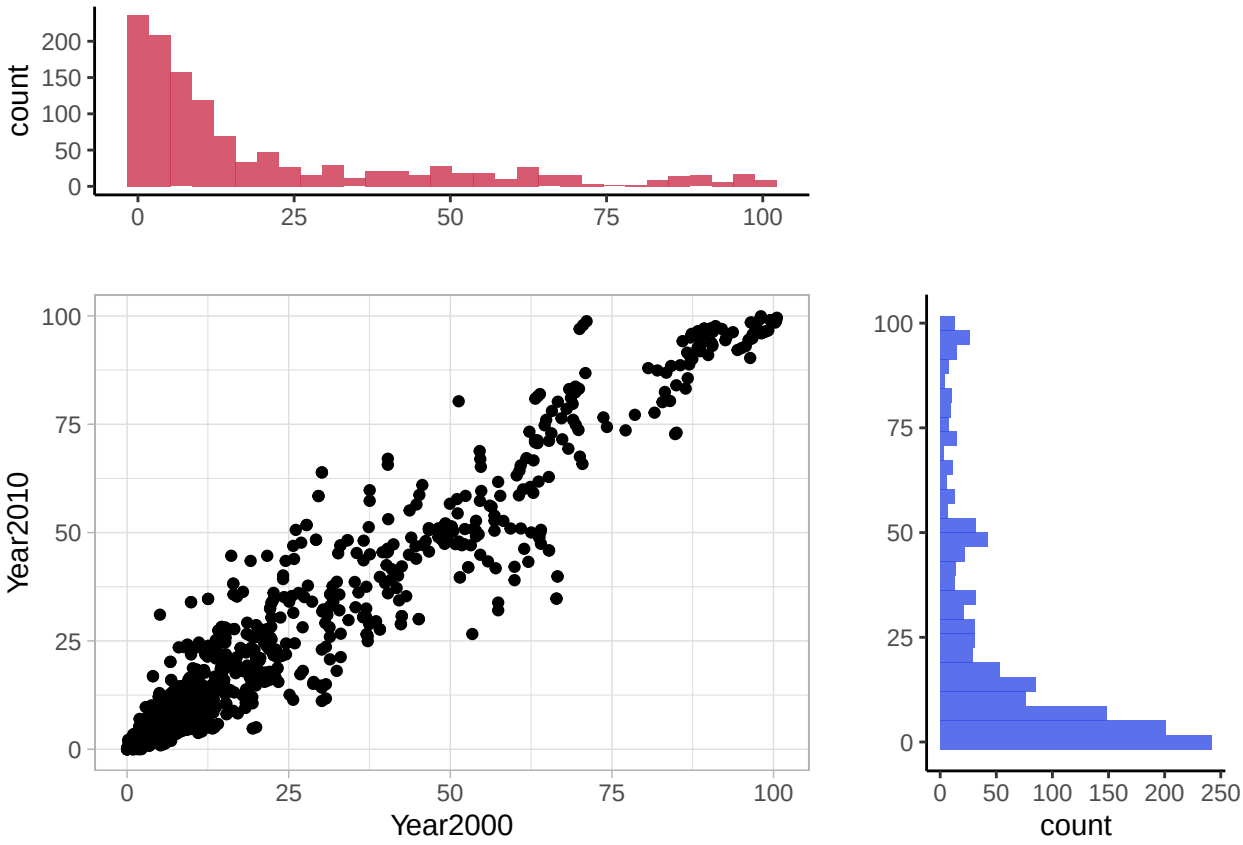
```
layout(rbind(c(2,2,0),c(1,1,3),c(1,1,3)))
plot(edcomplt3$Year2000,edcomplt3$Year2010)
hist(edcomplt3$Year2000,breaks=50,col=rgb(0.6,0.2,0.3))
barplot(k$counts,horiz=TRUE,col=rgb(0.3,0.8,0.5))
```






```
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`.
```


count
200
150
100
50
0
0
25
50
75
100
Year2010
100
75
50
25
0
0
25
50
75
100
Year2000
100
75
50
25
0
0
50
100
150
200
250
count