**R script for set 3 data set**

I had the dataset as an Excel file. To read an excel file in R, we need to import the below libraries

1)Import the library to read an excel file

library(xlsx)

library(readxl)

2) Create the data frame in R

set3 <- readxl::read\_xlsx(file.choose())

3) Clean the data set by omitting the dropout rate = “-1”

set3 <- set3[set3$`Dropout Rate`>-1,]

4) Create two subsets of type = “ST” , “DT” i.e.. state and district.

typest <- set3[set3$Type=="ST",]

typedi <- set3[set3$Type=="DI",]

5) Ran initial descriptive statistics for state and district subsets.

summary(typest)

Type School Year County District Agency Name

Length:4 Min. :20112012 Min. :0 Min. :0 Length:4 Class :character 1st Qu.:20127014 1st Qu.:0 1st Qu.:0 Class:character

Mode :character Median :20137014 Median :0 Median :0 Mode :character

Mean :20134514 Mean :0 Mean :0

3rd Qu.:20144515 3rd Qu.:0 3rd Qu.:0

Max. :20152016 Max. :0 Max. :0

Percentage of Teachers with Master’s Degrees Dropout Rate

Min. :0.4746 Min. :0.01090

1st Qu.:0.5066 1st Qu.:0.01105

Median :0.5194 Median :0.01155

Mean :0.5111 Mean :0.01215

3rd Qu.:0.5239 3rd Qu.:0.01265

Max. :0.5307 Max. :0.01460

summary(typedi)

Type School Year County District Agency Name Length:86 Min.:20112012 Min.: 1.00 Min.: 1.00 Length:86 Class:character 1st Qu.:20117012 1st Qu.:24.75 1st Qu.: 1.00 Class :character

Mode:character Median :20132014 Median :34.00 Median : 7.00 Mode :character

Mean :20133991 Mean :43.31 Mean : 25.83

3rd Qu.:20142015 3rd Qu.:68.25 3rd Qu.: 24.75

Max. :20152016 Max. :90.00 Max. :505.00

Percentage of Teachers with Master’s Degrees Dropout Rate

Min. :0.2452 Min. :0.002400

1st Qu.:0.4714 1st Qu.:0.008475

Median :0.5705 Median :0.014500

Mean :0.5344 Mean :0.019065

3rd Qu.:0.6099 3rd Qu.:0.020125

Max. :0.7213 Max. :0.179100

6) Found the max and min value on dropout rate for state and district

mxdropst <- typest[typest$`Dropout Rate`==max(typest$`Dropout Rate`)|typest$`Dropout Rate`==min(typest$`Dropout Rate`),]

mxdropdi <- typedi[typedi$`Dropout Rate`==max(typedi$`Dropout Rate`)|typedi$`Dropout Rate`==min(typedi$`Dropout Rate`),]

7) boxplot

boxplot(typedi$`Agency Name`~typedi$`Dropout Rate`)

boxplot(typedi$District~typedi$`Dropout Rate`)

8) aov – This function is used to know if there is any significant difference between the percentage of teachers with master’s degree and the dropout rate of students

aov(typedi)

aov(typedi$`Percentage of Teachers with Master’s Degrees`~typedi$`Dropout Rate`)

Call:

aov(formula = typedi$`Percentage of Teachers with Master’s Degrees` ~

typedi$`Dropout Rate`)

Terms:

typedi$`Dropout Rate` Residuals

Sum of Squares 0.1723423 0.7953346

Deg. of Freedom 1 84

Residual standard error: 0.09730503

Estimated effects may be unbalanced

9) The function **summary.aov**() is used to summarize the analysis of variance model. The output includes the columns *F value* and *Pr(>F)* corresponding to the p-value of the test

summary(aovdi)

Df Sum Sq Mean Sq F value Pr(>F)

typedi$`Dropout Rate` 1 0.1723 0.17234 18.2 5.19e-05 \*\*\*

Residuals 84 0.7953 0.00947

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

10) Regression: We performed the linear regression (lm) to predict the percentage of master’s degree based on dropout rate. And also predicted the dropout rate based on the percentage of teacher’s with master’s degree.

resultdi <- lm(typedi$`Percentage of Teachers with Master’s Degrees`~typedi$`Dropout Rate`)

Call:

lm(formula = typedi$`Percentage of Teachers with Master’s Degrees` ~

typedi$`Dropout Rate`)

Coefficients:

(Intercept) typedi$`Dropout Rate`

0.5698 -1.8560

resultdi <- lm(typedi$`Dropout Rate`~typedi$`Percentage of Teachers with Master’s Degrees`)

Call:

lm(formula = typedi$`Dropout Rate` ~ typedi$`Percentage of Teachers with Master’s Degrees`)

Coefficients:

(Intercept) typedi$`Percentage of Teachers with Master’s Degrees`

0.07035 -0.09596

11) Split data by county

countydi <- split(typedi,typedi$County)

summary(countydi[[1]]$`Dropout Rate`)

12) attach the data frame to perform graphical analysis. Histogram plot on various county list. We did histogram plot on few of the data based on the countywise.

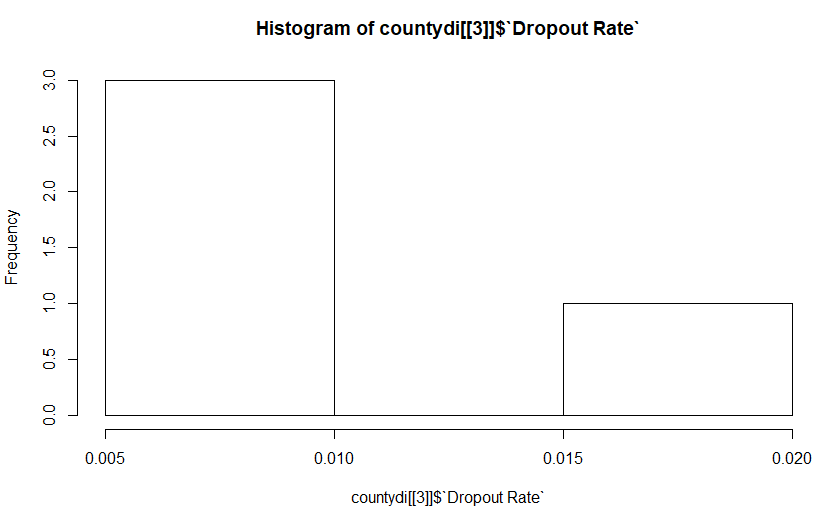
attach(countydi)

attach(typedi)

hist(countydi[[1]]$`Dropout Rate`)

hist(countydi[[2]]$`Dropout Rate`)

hist(countydi[[3]]$`Dropout Rate`)



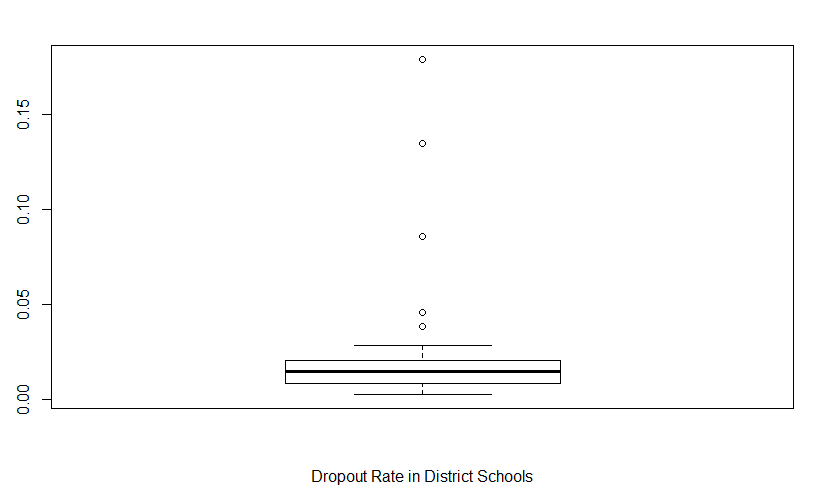
hist(countydi[[1]]$`Percentage of Teachers with Master’s Degrees`)

hist(countydi[[2]]$`Percentage of Teachers with Master’s Degrees`)

hist(countydi[[3]]$`Percentage of Teachers with Master’s Degrees`)

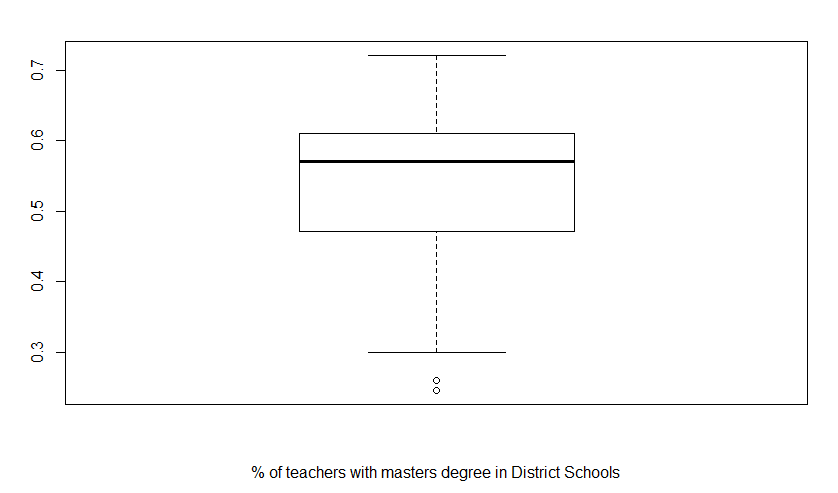
13) We did boxplot on the dropout rate for district level schools

boxplot(typedi$`Dropout Rate`,xlab="Dropout Rate in District Schools")



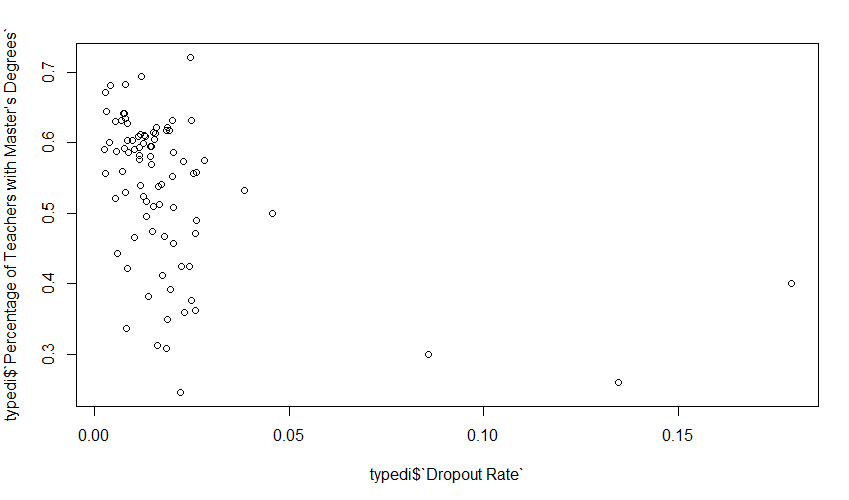
14) Boxplot on the percentage of teachers with master’s degree.

boxplot(typedi$`Percentage of Teachers with Master’s Degrees`,xlab="% of teachers with masters degree in District Schools")



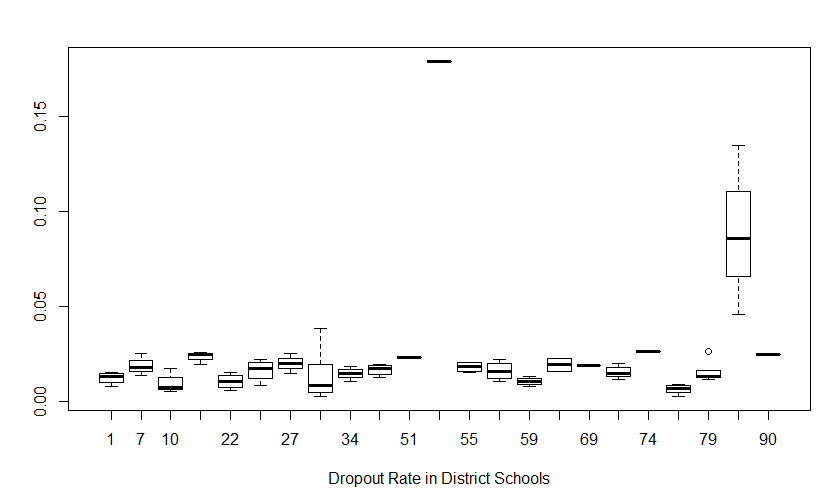
15) Scatterplot on percentage of teachers with master’s degree

plot(typedi$`Percentage of Teachers with Master’s Degrees`~typedi$`Dropout Rate`)



16) Boxplot on dropout rate for each county in a district.

boxplot(typedi$`Dropout Rate`~typedi$County,xlab="Dropout Rate in District Schools")

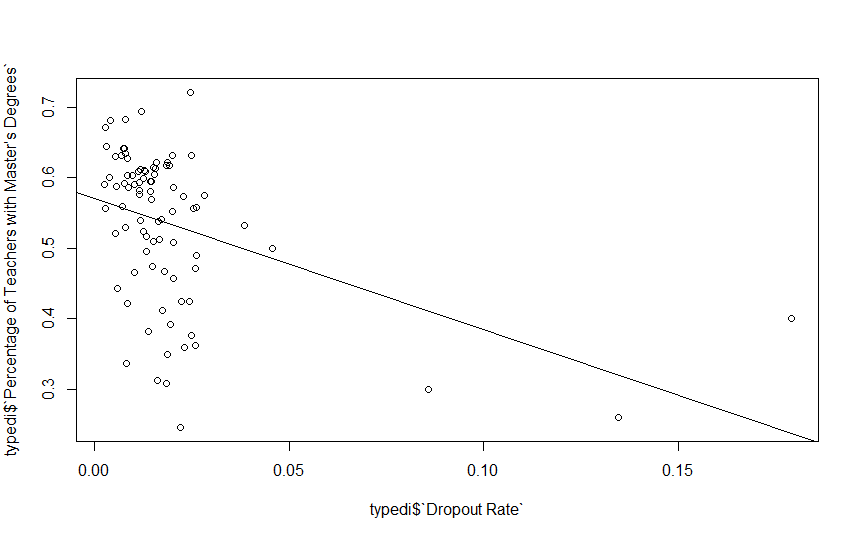


17) The Linear regression line between the percentage of teacher’s with master’s degree and the dropout rate.

plot(typedi$`Percentage of Teachers with Master’s Degrees`~typedi$`Dropout Rate`)

lin <- lm(typedi$`Percentage of Teachers with Master’s Degrees`~typedi$`Dropout Rate`)

abline(lin)



18) Finally save the script to the local folder either as an R file or .txt file.

savehistory("~/set3commands.Rhistory")

savehistory("~/set3comm.txt")