R script for Data set2 and analysis.

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

set2 <- read\_excel(file.choose())

View(set2)

set2

3) Clean the data set by omitting the null fields also creating a subset by eliminating the fields that had values “-1”.

setcln <- na.omit(set2)

setcln

set2atrate <- set2[set2$`Attendance Rate` > -1,]

4) Check the summary statistics on the data frame before eliminating the negative fields which had no meaning in the calculation.

summary(set2)

5) Perform statistic on column level.

summary(set2$`School Year`)

summary(set2$`Agency Name`)

summary(set2$District)

summary(set2$`Dropout Rate`)

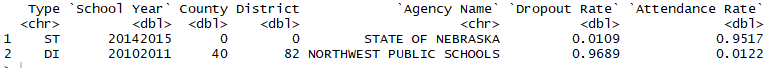
Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0109 0.9433 0.9513 0.9043 0.9561 0.9689

Summary(set2$`Attendance Rate`)

6) Perform the math function like max and min on the dropout rate and display the entire row. This was I could relate the effect of attendance rate on dropout rate.

mx <- set2[set2$`Dropout Rate` == max(set2$`Dropout Rate`) | set2$`Dropout Rate` == min(set2$`Dropout Rate`),]



7) Split the data by year, county. To perform statistics

yr <- split(set2,set2$`School Year`)

summary(yr)

countylist <- split(set2,set2$County)

mxcounty <- set2atrate[set2atrate$`Dropout Rate` == max(set2atrate$`Dropout Rate`) | set2atrate$`Dropout Rate` == min(set2atrate$`Dropout Rate`),]

8) Statistics on the subset data.

summary(typedi)

Type School Year County District Agency Name Dropout Rate

Length:108 Min. :20102011 Min. : 1.00 Min. : 1.00 Length:108 Min. :0.8085

Class :character 1st Qu.:20112012 1st Qu.:26.25 1st Qu.: 1.00 Class :character 1st Qu.:0.9447

Mode :character Median :20132014 Median :34.00 Median : 7.00 Mode :character Median :0.9520

Mean :20127477 Mean :43.32 Mean : 23.94 Mean :0.9456

3rd Qu.:20142015 3rd Qu.:69.50 3rd Qu.: 20.25 3rd Qu.:0.9565

Max. :20152016 Max. :90.00 Max. :505.00 Max. :0.9689

Attendance Rate

Min. :0.00240

1st Qu.:0.00945

Median :0.01495

Mean :0.01944

3rd Qu.:0.02022

Max. :0.17910

9) Perform annova on the dropout rate and the effect of attendance rate.

aov(typedi$`Dropout Rate`~typedi$`Attendance Rate`)

Call:

aov(formula = typedi$`Dropout Rate` ~ typedi$`Attendance Rate`)

Terms:

typedi$`Attendance Rate` Residuals

Sum of Squares 0.04617891 0.02077270

Deg. of Freedom 1 106

Residual standard error: 0.01399889

Estimated effects may be unbalanced

10) Saved the r script to the local folder.

savehistory("D:/Downloads/Fall 2017/DTD/set2.Rhistory")