Data Science with R: Project 3

By: Bhupendra Nirgun

Contents

[1. Read the csv file : 3](#_Toc523817909)

[2. Visualize the data 3](#_Toc523817910)

[3. Goals of the Project 6](#_Toc523817911)

[Problem 1: 6](#_Toc523817912)

[Problem 2: 7](#_Toc523817913)

[Problem 3: 7](#_Toc523817914)

[Problem 4: 8](#_Toc523817915)

[Problem 5: 9](#_Toc523817916)

[Problem 6: 10](#_Toc523817917)

[4. Complete Code File 11](#_Toc523817918)

[5. Complete Output file 11](#_Toc523817919)

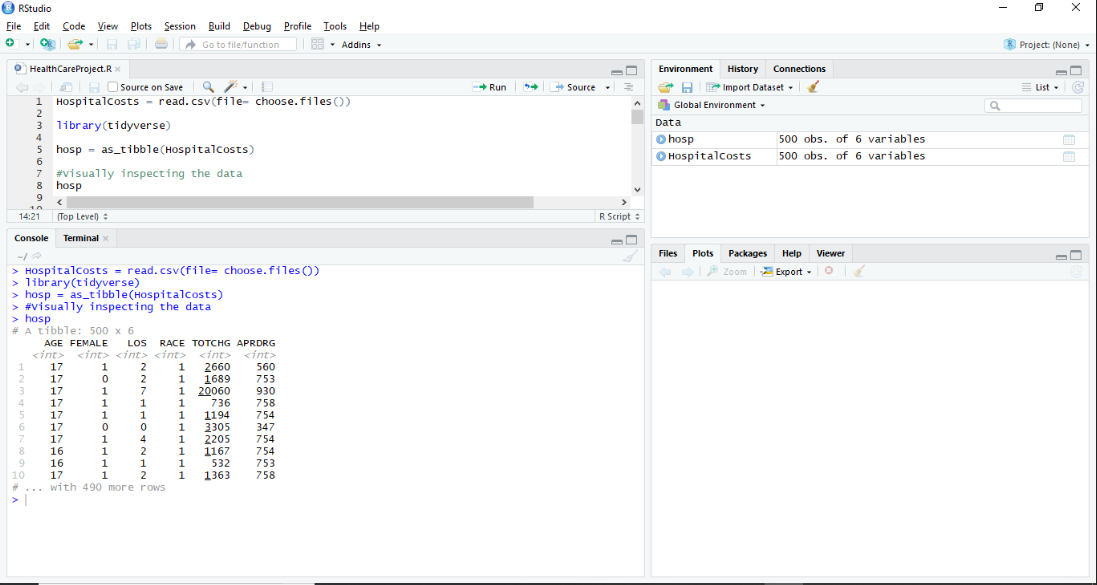
## Read the csv file :

Read the Hospitals cost.csv

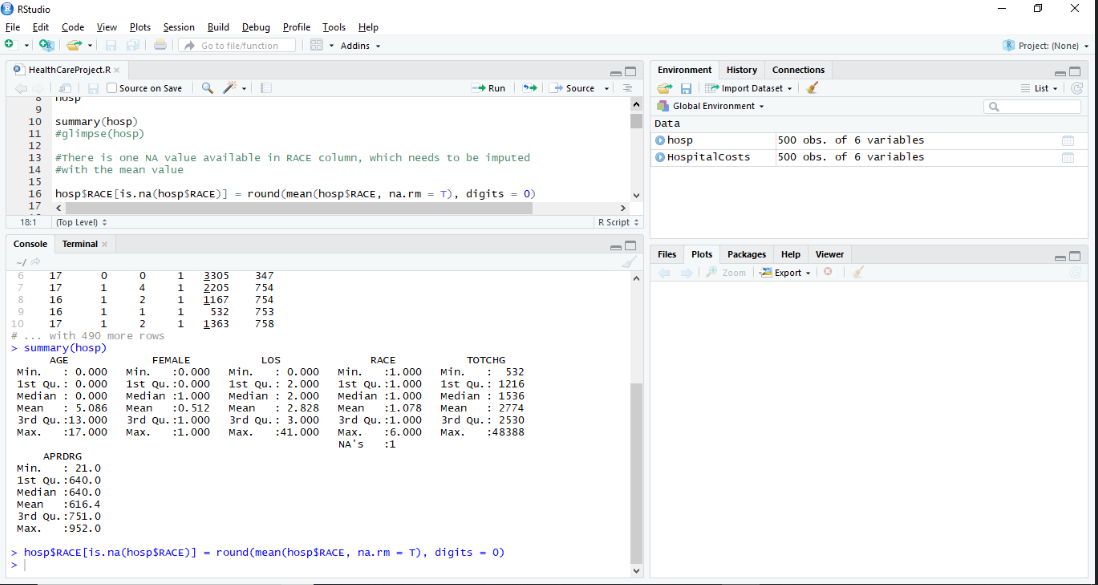
## Visualize the data

This can be visualized using various forms some of them are

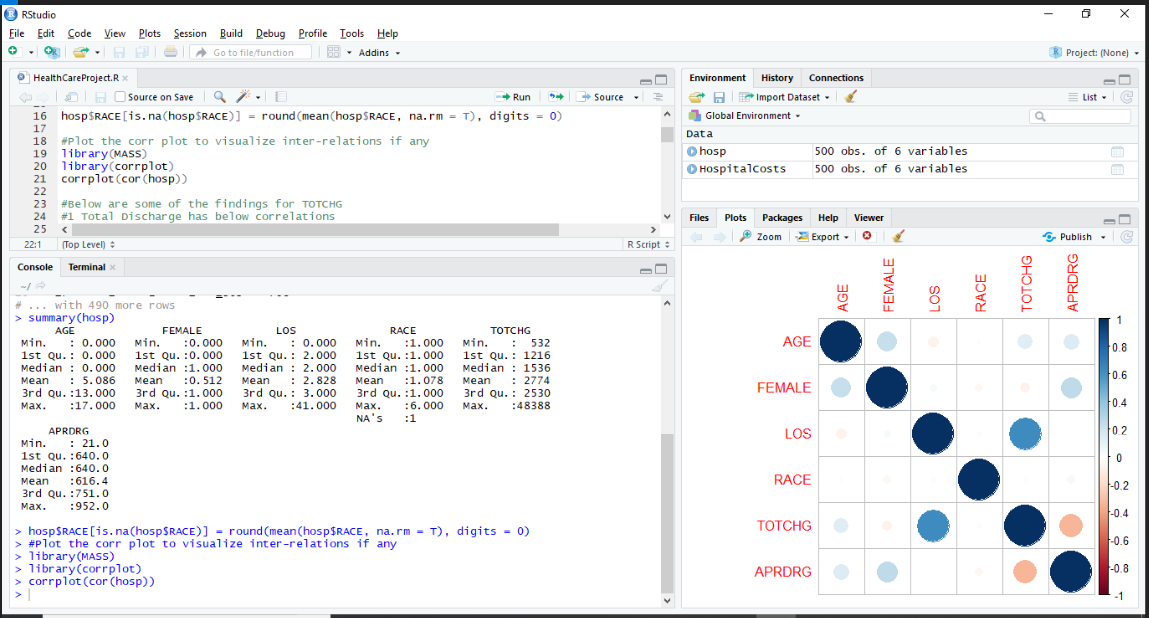
Structure, summary, glimpse, class, head; I have used tibble and summary and below are the screenshots of the same.

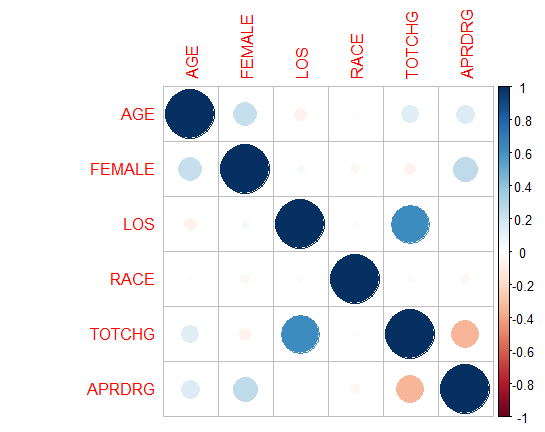


Total charge only looks to be the continuous variable, rest of the columns looks to be categorical variables.



Summary of the dataframe hosp gives the summary; Note that the Race column contains one NA value which has been imputed with the mean value.





Below are some of the observations related to the corr plot.

Below are some of the findings for TOTCHG

Total Discharge has below correlations

high +ve co-relation with Length Of Stay

low -ve co-relation with Diagnosis Groups

3 low +ve co-relation with Age

There are also co-relations exists with

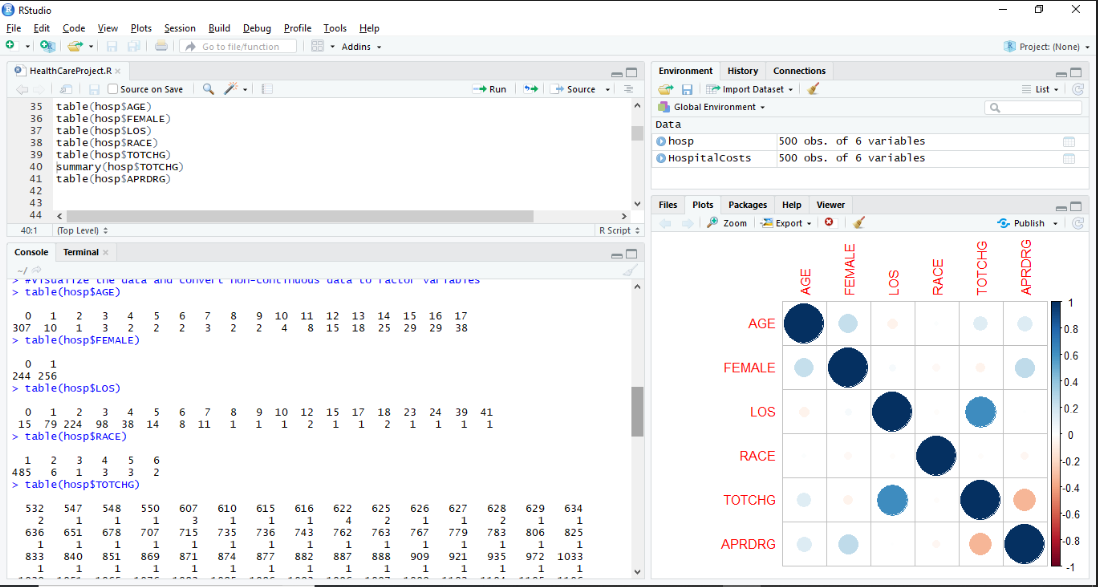
High +ve co-relation between Age and Gender

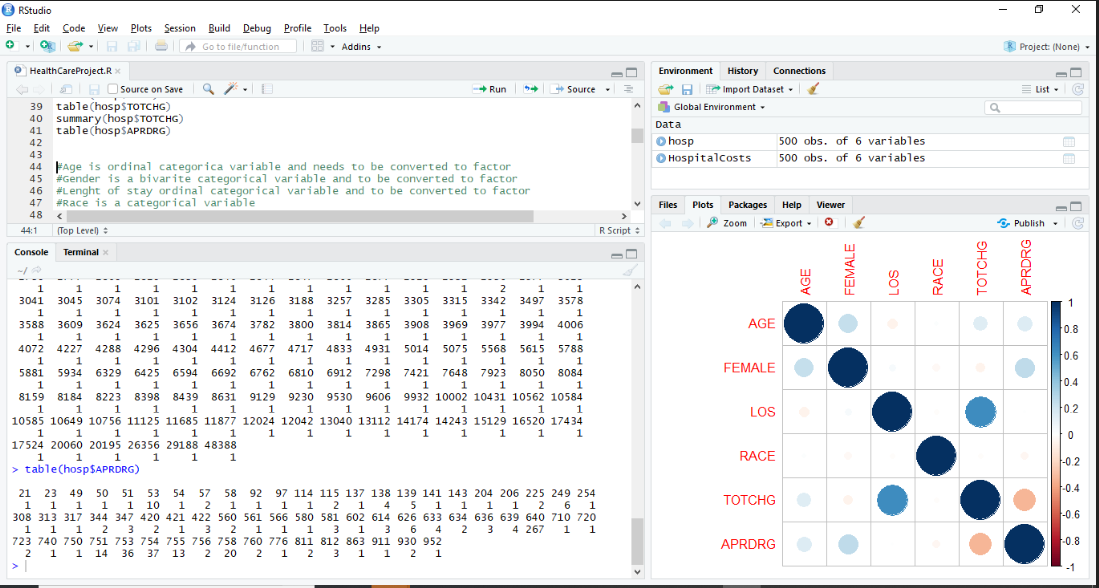
Low -ve co-relation between Age and Length of Stay

High +ve co-relation between Age and Diagnosis Groups

There is +ve co-relation between Gender and Diagnosis Groups

Verify the frequency of the categorical variables





Below is the general observation

Age is ordinal categorical variable and needs to be converted to factor

Gender is a bivariate categorical variable and to be converted to factor

Length of stay ordinal categorical variable and to be converted to factor

Race is a categorical variable

Discharge cost is continuous variable

Diagnosis Group is categorical variable

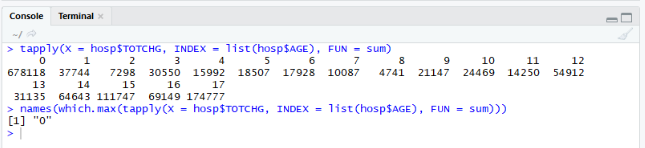
## Goals of the Project

### Problem 1:

To record the patient statistics, the agency wants to find the age category of people who frequent the hospital and has the maximum expenditure.

tapply(X = hosp$TOTCHG, INDEX = list(hosp$AGE), FUN = sum)

names(which.max(tapply(X = hosp$TOTCHG, INDEX = list(hosp$AGE), FUN = sum)))



#-------------------------------------------------------------------------------------

#Conclusion

Infants with the Age category “0” spend the max amount of 678118 for the hospitalization

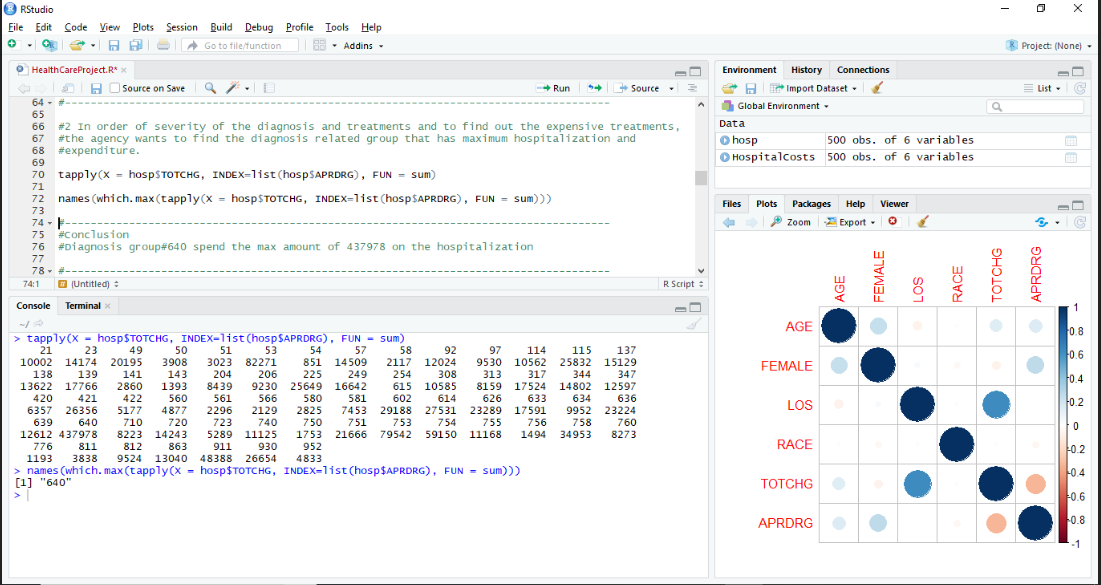
#-------------------------------------------------------------------------------------

### Problem 2:

In order of severity of the diagnosis and treatments and to find out the expensive treatments, the agency wants to find the diagnosis related group that has maximum hospitalization and expenditure.

tapply(X = hosp$TOTCHG, INDEX=list(hosp$APRDRG), FUN = sum)

names(which.max(tapply(X = hosp$TOTCHG, INDEX=list(hosp$APRDRG), FUN = sum)))



#-------------------------------------------------------------------------------------

#Conclusion

#Diagnosis group#640 spend the max amount of 437978 on the hospitalization

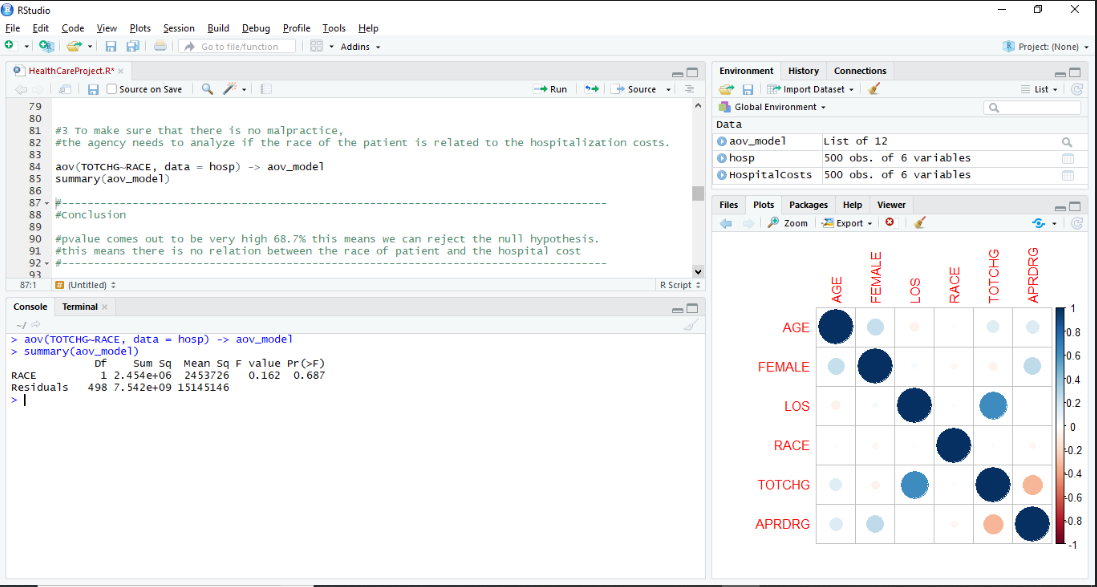
#-------------------------------------------------------------------------------------

### Problem 3:

To make sure that there is no malpractice, the agency needs to analyze if the race of the patient is related to the hospitalization costs.

aov(TOTCHG~RACE, data = hosp) -> aov\_model

summary(aov\_model)



#-------------------------------------------------------------------------------------

#Conclusion

#pvalue comes out to be very high 68.7% this means we can reject the null hypothesis.

#this means there is no relation between the race of patient and the hospital cost

#-------------------------------------------------------------------------------------

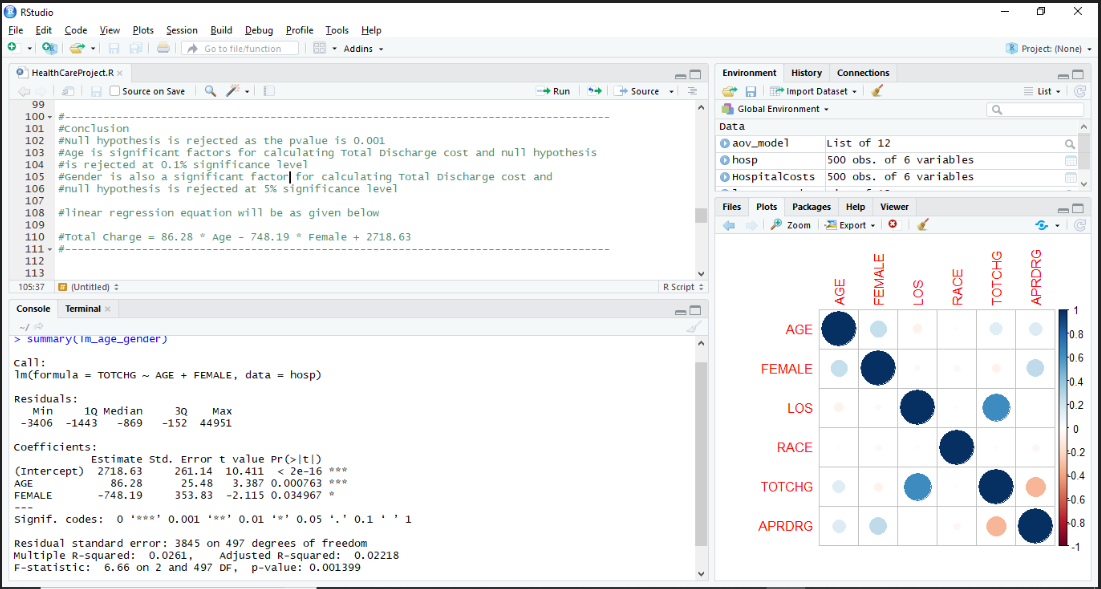
### Problem 4:

To properly utilize the costs, the agency has to analyze the severity of the hospital costs by age and gender for proper allocation of resources.

#Note: Data has not been splitted into Test and Train due to small sample size.

lm\_age\_gender = lm(formula = TOTCHG~AGE + FEMALE ,data = hosp)

summary(lm\_age\_gender)



#-------------------------------------------------------------------------------------

#Conclusion

Null hypothesis is rejected as the pvalue is 0.001 at the significance level of 0.1%

Age is significant factors for calculating Total Discharge cost and null hypothesis is rejected at 0.1% significance level.

Gender is also a significant factor for calculating Total Discharge cost and null hypothesis is rejected at 5% significance level

#linear regression equation will be as given below

#Total Charge = 86.28 \* Age - 748.19 \* Female + 2718.63

#-------------------------------------------------------------------------------------

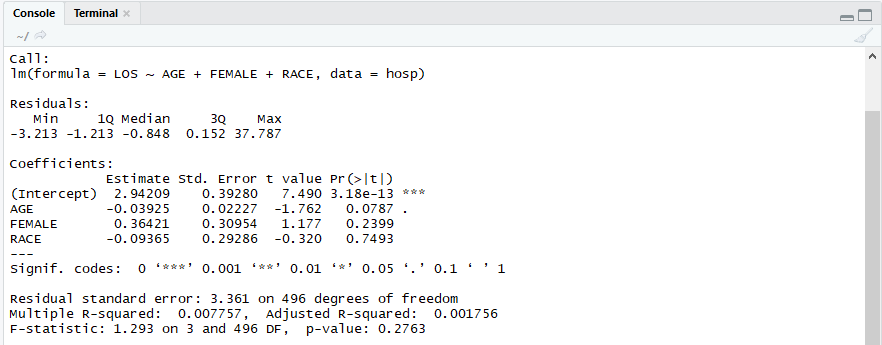
### Problem 5:

Since the length of stay is the crucial factor for inpatients, the agency wants to find if the length of stay can be predicted from age, gender, and race.

#Note: Data has not been splitted into Test and Train due to small sample size.

lm\_age\_gender\_race = lm(formula = LOS~AGE + FEMALE + RACE ,data = hosp)

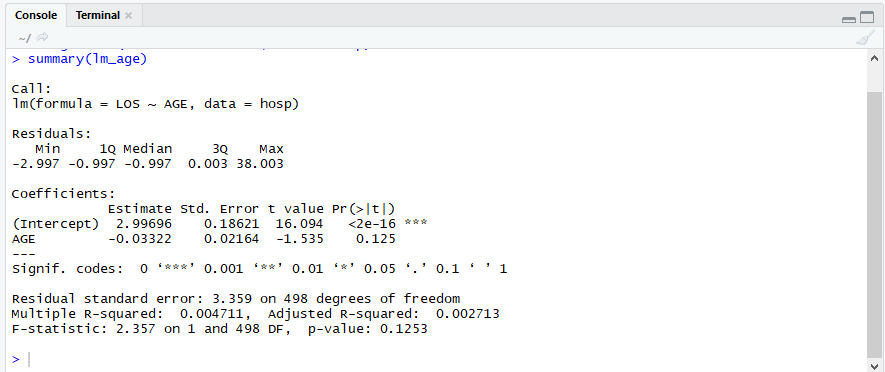
summary(lm\_age\_gender\_race)



Model shows the p\_value is 0.2763 which is very high and relation with the Age at 10%

significance, but as there is no direct corelation model needsd to be verified by

removing Gender and Race parameters



#-------------------------------------------------------------------------------------

#Conclusion

Null hypothesis is retained as the pvalue is 0.2763

Age, Gender are Race are not significant factors for calculating Total Discharge charges

#-------------------------------------------------------------------------------------

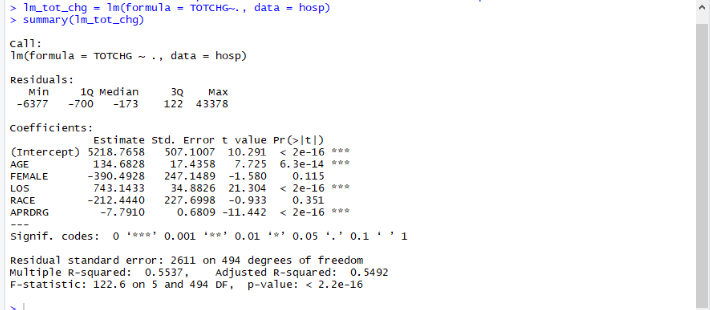
### Problem 6:

To perform a complete analysis, the agency wants to find the variable that mainly affects the hospital costs.

#Note : Data has not been splitted into Test and Train due to small sample size

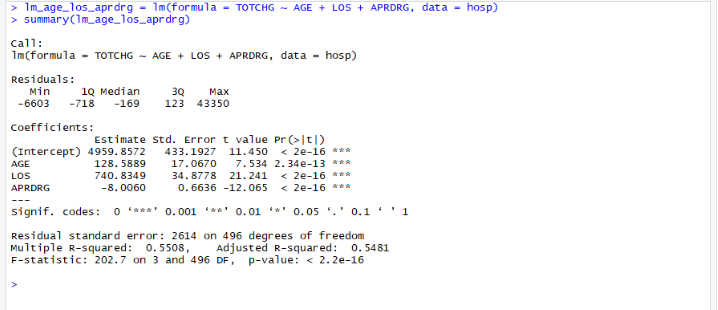
lm\_tot\_chg = lm(formula = TOTCHG~., data = hosp)

summary(lm\_tot\_chg)



lm\_age\_los\_aprdrg = lm(formula = TOTCHG ~ AGE + LOS + APRDRG, data = hosp)

summary(lm\_age\_los\_aprdrg)



#-------------------------------------------------------------------------------------

#Conclusion

Null hypothesis is rejected as the pvalue is very small - 2.2 \* e^-16

Age, Length of stay and diagnosis group are significant factors for calculating Total Discharge cost at the significance level of 0.1%

#The equation is as given below

#Total cost = 128.59 \* Age + 740.83 \* Length of Stay - 8.01 \* Diag Grp fact + 4959.86

#-------------------------------------------------------------------------------------

## Complete Code File

Below attached is the extract of code file.



## Complete Output file

Below attached is the extract of output console.(excludes corrplot)

Console output in doc format Console output in Text format