**Data Science with R** Assessment

**Project –** *Healthcare Cost Analysis*

**To analyze the data to research on healthcare costs and their utilization.**

**Objective of the Project:**

A nationwide survey of hospital costs conducted by the US Agency for Healthcare consists of hospital records of inpatient samples. The given data is restricted to the city of Wisconsin and relates to patients in the age group 0-17 years. The agency wants to analyze the data to research on healthcare costs and their utilization.

**Task Analysis:**

**Task1:**

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

**Code:**

summary(Hcosts)

hist(Hcosts$AGE)

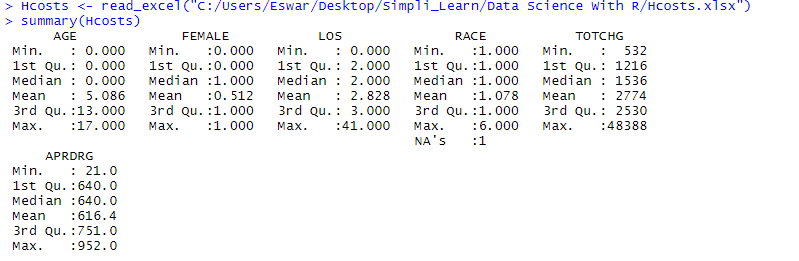
age<-as.factor(Hcosts$AGE)

summary(age)

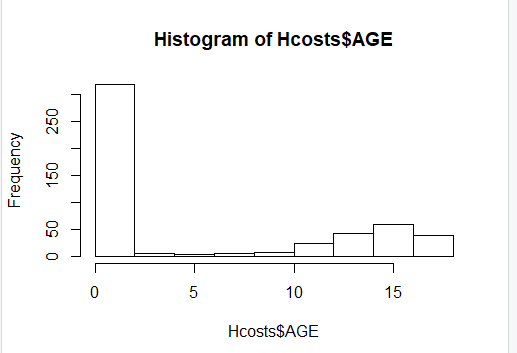
max(Hcosts$TOTCHG)

tapply(Hcosts$TOTCHG,Hcosts$AGE,sum)

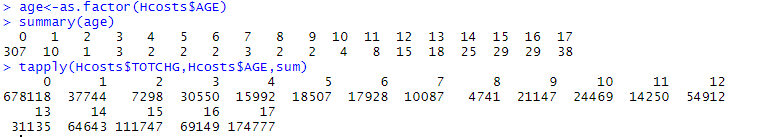
**Result:**



From summarizing the given dataset, the minimum age of the people discharged is 0 and maximum is 17 years. By this we can clearly say that it is a children Hospital. The maximum hospital discharges costs is 48388 and minimum is 532. The people who are consulting the doctor must pay a minimum of 532 and maximum depends on the diagnosis required.



From the histogram we can conclude that, The infants who was having age below 1 year are more taking treatment compared all other age groups. Here each bar have an interval of 2. So we can say the children having age 0-2 are frequent in taking diagnosis.



From here we can say that the children of age below 1 year are more frequent. The highest number of entries can be seen from the children of age below 1year, 15 and 17 years.

**Task2:**

 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.

**Code:**

model1<-as.factor(Hcosts$APRDRG)

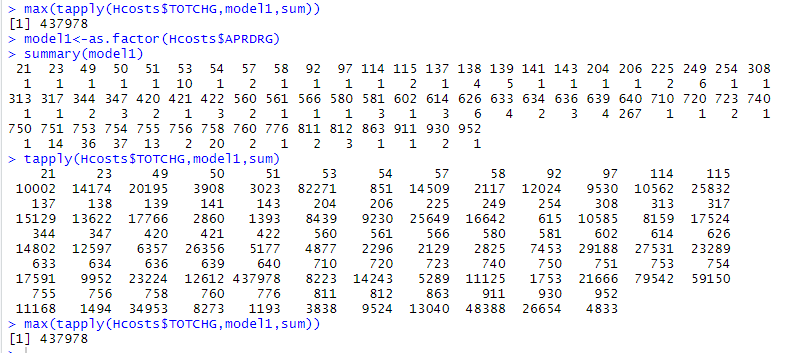
summary(model1)

tapply(Hcosts$TOTCHG,model1,sum)

max(tapply(Hcosts$TOTCHG,model1,sum))

which.max(tapply(Hcosts$TOTCHG,model1,sum))

**Result:**

**C:\Users\Eswar\Desktop\Simpli_Learn\Data Science With R\Hospital Costs Project\whichmax.PNG**

The diagnosis related group which has maximum hospitalization and expenditure is 640.It has the highest expenditure of 437978 compared to all diagnosis related groups.

**Task 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.

**Code:**

Hcosts1<-na.omit(Hcosts)

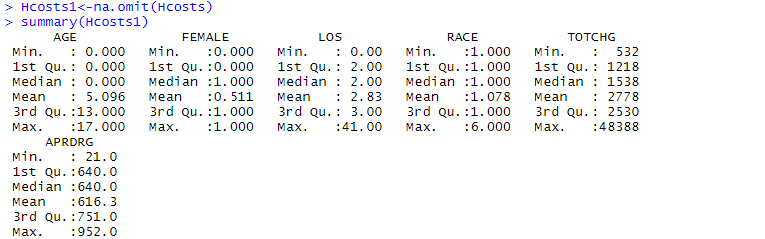
summary(Hcosts1)

model2<-lm(RACE~TOTCHG,data=Hcosts1)

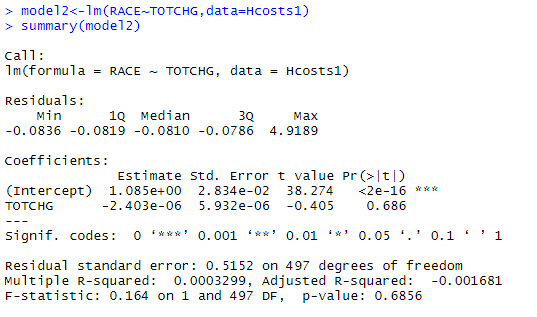
summary(model2)

cor(Hcosts1$TOTCHG,Hcosts1$RACE)

**Result:**



As we saw there is one missing value in Race we cannot perform analysis. So ,without performing missing values operations iam just eliminating that row. Though, I omit that the results wouldn’t effect.



From here we can say that, the hospitality discharge costs are not at all related based on race of the patient. The p value is high (it is least significant). So the hospitalization costs are not at all dependent on race of the patient.

C:\Users\Eswar\Desktop\Simpli_Learn\Data Science With R\Hospital Costs Project\cor.PNG

By finding correlation also we can say whether the two variables are dependent or not. So iam just verifying again whether my analysis is correct or not. The correlation between hospitalization costs and race is approximately 0. So the hospitality costs are not related to the race of the patients.

**Task 4:**

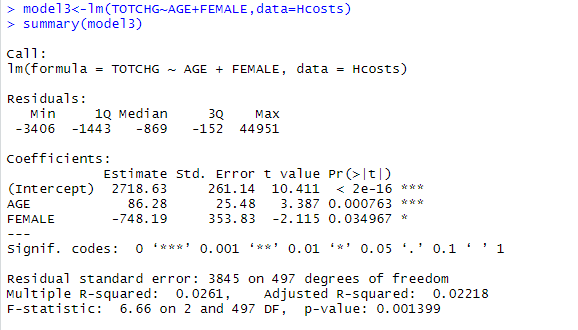
To properly utilize the costs, the agency has to analyze the severity of the hospital costs by age and gender for the proper allocation of resources.

**Code:**

model3<-lm(TOTCHG~AGE+FEMALE,data=Hcosts)

summary(model3)

**Result:**

****

From here we can say that age is having the impact on the hospitalization costs as its p value is very less. Gender is also having impact on the hospitalization costs , but its p value is greater than age. Though, both the factors age and gender are having a impact on Hospitality discharge Costs.

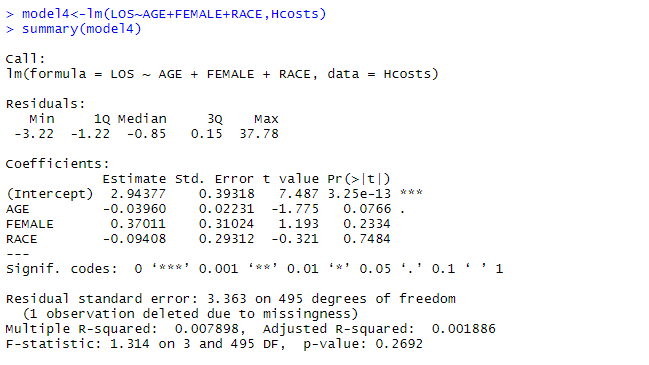
**Task 5:**

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.

**Code:**

model4<-lm(LOS~AGE+FEMALE+RACE,Hcosts)

summary(model4)

**Result: **

From here we can say that the factors age, gender, race does not effect the length of stay. There is no impact on length of stay.

**Task 6:**

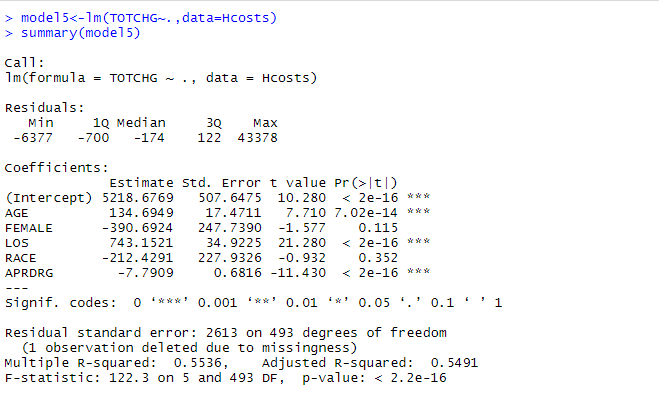
To perform a complete analysis, the agency wants to find the variable that mainly affects hospital costs.

**Code:**

model5<-lm(TOTCHG~.,data=Hcosts)

summary(model5)

**Result:**

****

From here we can say that the factors age and Length of stay effects the hospitalization costs. Only age and length of stay have less p value and significant. All Patient Refined Diagnosis Related Groups also significant but main factors that are impacting the hospital discharge costs are age and length of stay

**Total Source Code:**

library(readxl)

Hcosts <- read\_excel("C:/Users/Eswar/Desktop/Simpli\_Learn/Data Science With R/Hcosts.xlsx")

summary(Hcosts)

hist(Hcosts$AGE)

age<-as.factor(Hcosts$AGE)

summary(age)

max(Hcosts$TOTCHG)

tapply(Hcosts$TOTCHG,Hcosts$AGE,sum)

model1<-as.factor(Hcosts$APRDRG)

summary(model1)

tapply(Hcosts$TOTCHG,model1,sum)

max(tapply(Hcosts$TOTCHG,model1,sum))

which.max(tapply(Hcosts$TOTCHG,model1,sum))

Hcosts1<-na.omit(Hcosts)

summary(Hcosts1)

model2<-lm(RACE~TOTCHG,data=Hcosts1)

summary(model2)

cor(Hcosts1$TOTCHG,Hcosts1$RACE)

model3<-lm(TOTCHG~AGE+FEMALE,data=Hcosts)

summary(model3)

model4<-lm(LOS~AGE+FEMALE+RACE,Hcosts)

summary(model4)

model5<-lm(TOTCHG~.,data=Hcosts)

summary(model5)