**Analyze the Healthcare cost and Utilization in Wisconsin hospitals**

**Question**

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 the healthcare costs and their utilization.

**The goals of this project are:**

**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.

To find the category that has the highest frequency of hospital visit, we can use graphical analysis. A histogram would display the number of occurrences of each age category. The as.factor() is called to make sure that the categories are not treated as numbers.

Outlier treatments: None

**Code:**

Hospital=HospitalCosts

str(Hospital)

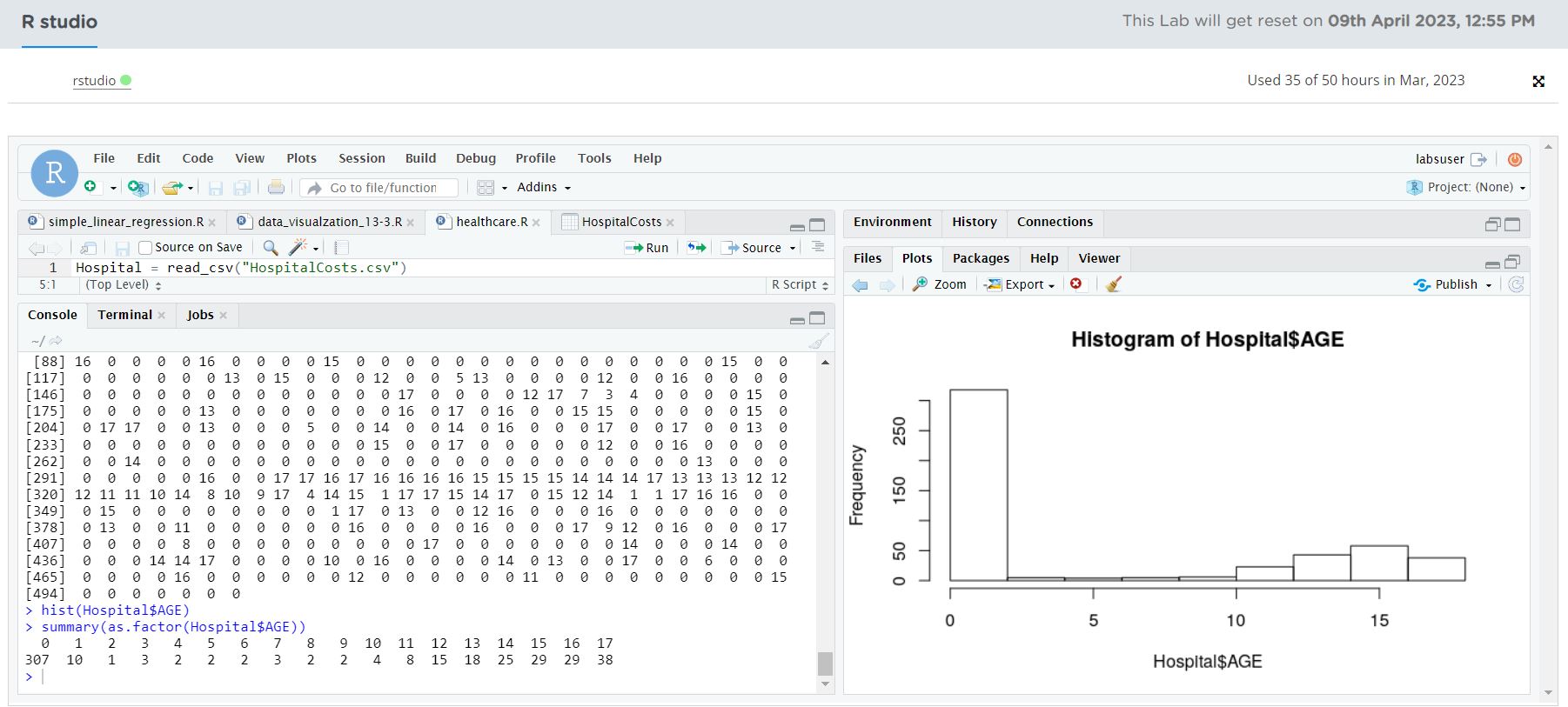
Hospital$AGE

hist(Hospital$AGE)

summary(as.factor(Hospital$AGE))

**Result:** From the graph that is displayed, we can see that infants have the maximum frequency of hospital visit, going above 300. The summary of AGE attribute gives the numerical output (after converting the age from numeric to factor) – and we can see that there are 307 entries for those in the range of 0-1 year.

**Output:**



**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.

**Code:**

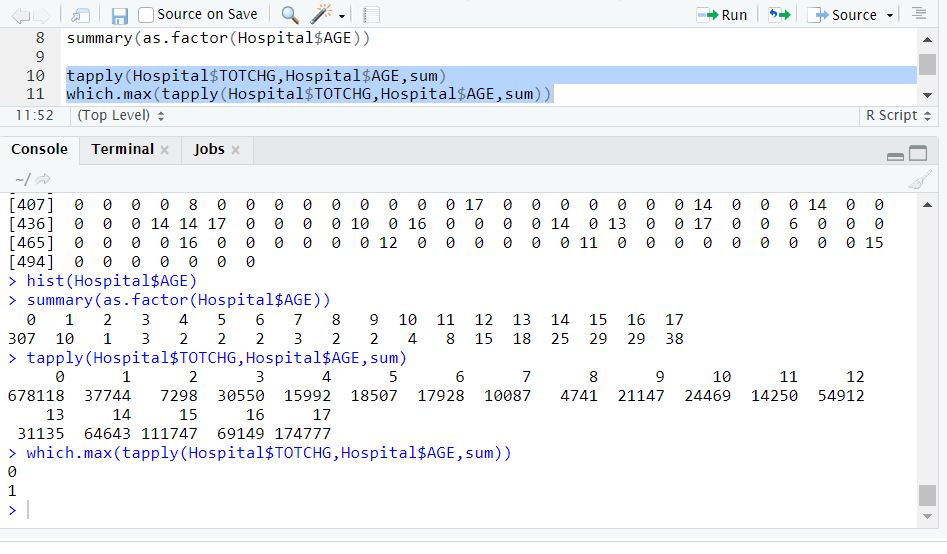
tapply(Hospital$TOTCHG,Hospital$AGE,sum)

which.max(tapply(Hospital$TOTCHG,Hospital$AGE,sum))

**Result:**

From the above output we can clearly find that the infants of age group 0-1 have the maximum hospitalization and expenditure.

**Output:**



**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:**

anov=aov(TOTCHG~RACE,data=Hospital)

anov

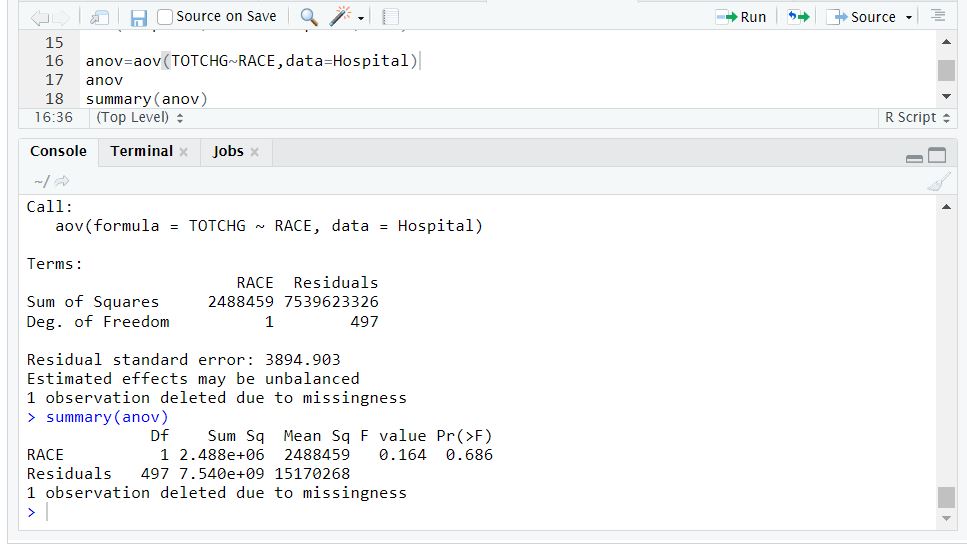
summary(anov)

**Result:**

To make sure that there is no malpractice if we analyze using Anova we can clearly see the findings

in summary as relation of race of the patient to the hospitalization costs.

**Output:**



**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.

**Code:**

str(Hospital)

names(Hospital)

cor(Hospital$TOTCHG,Hospital$AGE+Hospital$FEMALE)

plot(Hospital$TOTCHG,Hospital$AGE+Hospital$FEMALE)

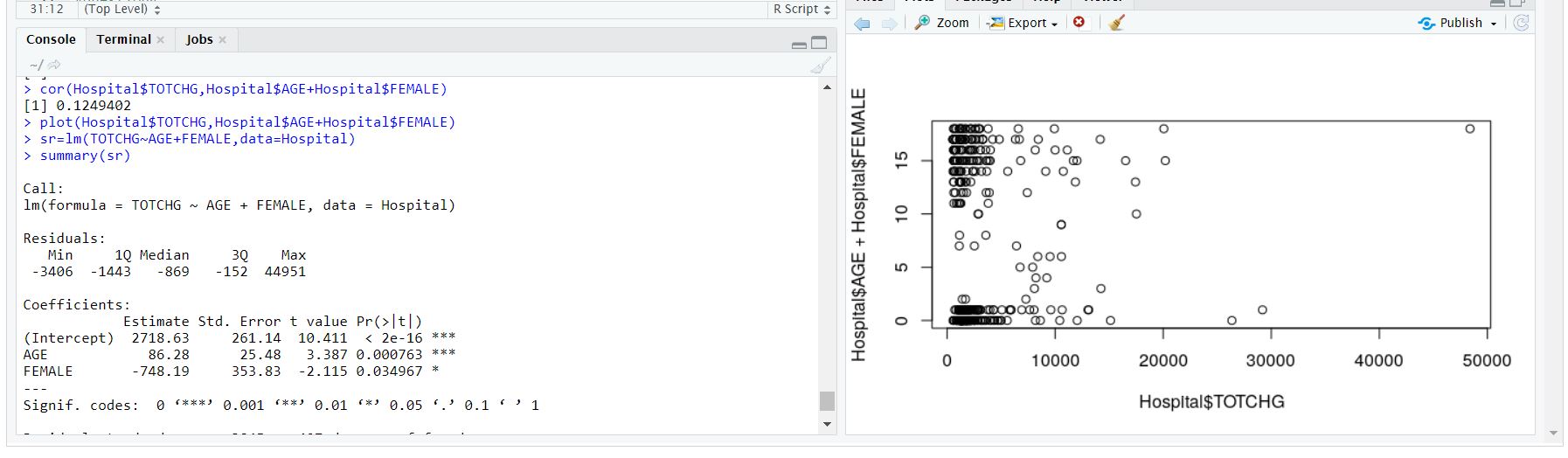
sr=lm(TOTCHG~AGE+FEMALE,data=Hospital)

summary(sr)

**Result:**

To properly utilize the costs of the agency if we deeply analyze the hospital costs by age and hospital costs by gender with field named as in female we observe that there is strong coefficient correlation with intercept between age and it has also correlation with gender as well. All these findings can also be clearly seen in plots which is plotted here for more clear observation .

**Output:**



**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:**

names(Hospital)

colSums(is.na(Hospital))

print(sapply(transform(Hospital,RACE=as.numeric(RACE)),class))

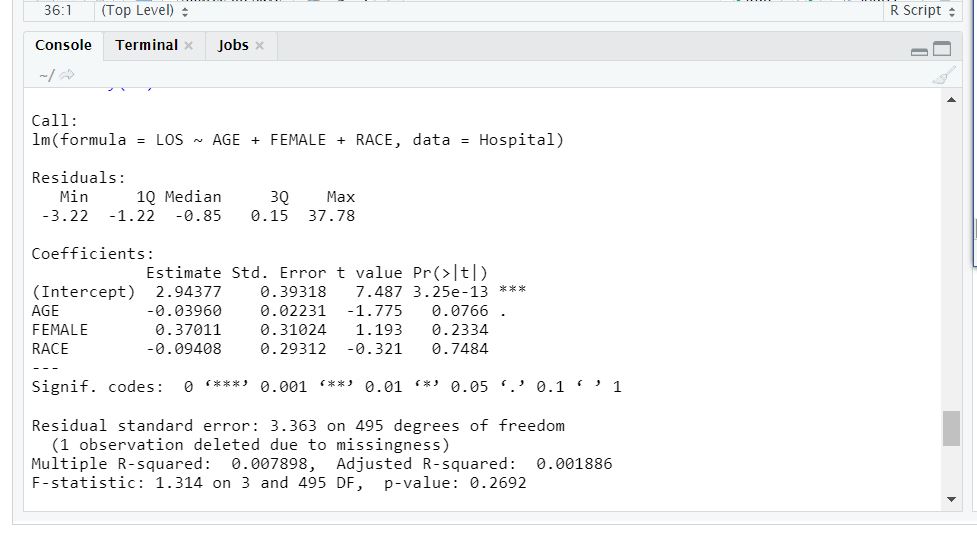
sr=lm(LOS~AGE+FEMALE+RACE,data=Hospital)

vif(sr)

summary(sr)

**Result:**

To predict the length of stay we created the linear model to check against age, female ,race we can clearly see the output that we find none relation between the fields and a weak relation between age.

**Output:** 

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

**Code:**

names(Hospital)

sr1=lm(TOTCHG~., data=Hospital)

summary(sr1)

**Result:**

For complete analysis that what all factors effects the hospital costs we create a linear model to analyze for each factor , where we can clearly see that there is a strong coefficient correlation between intercept with age , los and aprdrg.

**Output:**

