**Health Care Project with R**

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.**

**R Code:**

Hosp\_cost = read.csv("HospitalCosts.csv")

View(Hosp\_cost)

Age\_cat=c(Hosp\_cost$AGE)

View(Age\_cat)

Frequency = summary(as.factor(Hosp\_cost$AGE))

View(Frequency)

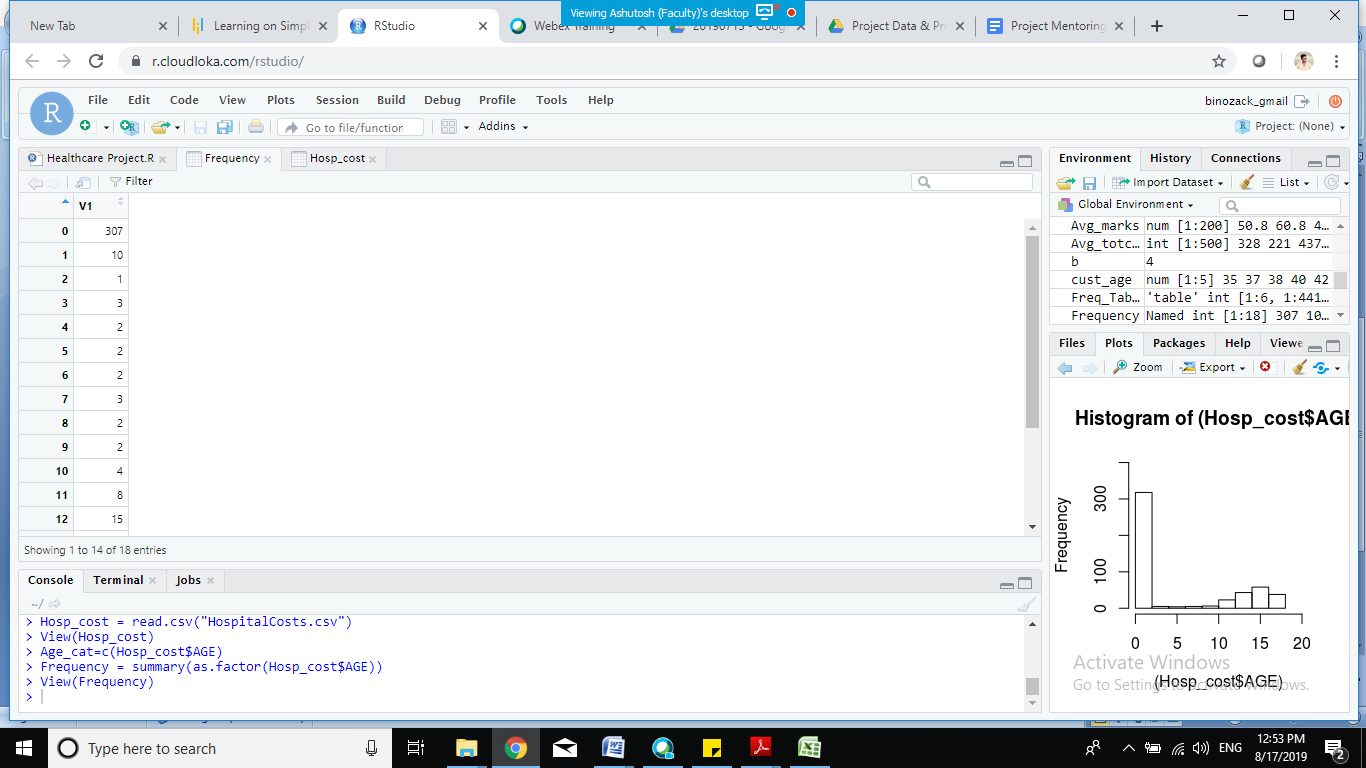
hist((Hosp\_cost$AGE),xlim=c(0,20), ylim=c(0,400))

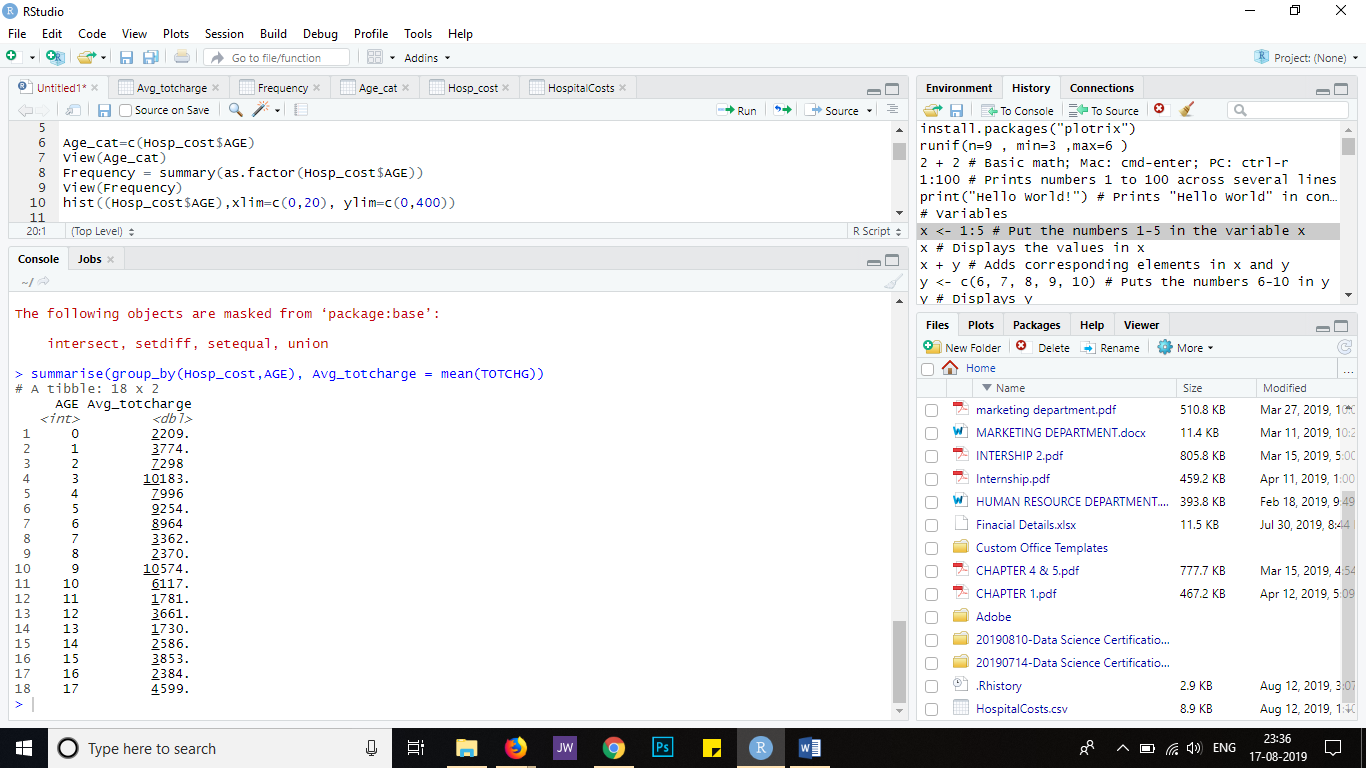
Avg\_totcharge = tapply(Hosp\_cost$AGE,Hosp\_cost$TOTCHG)

View(Avg\_totcharge)

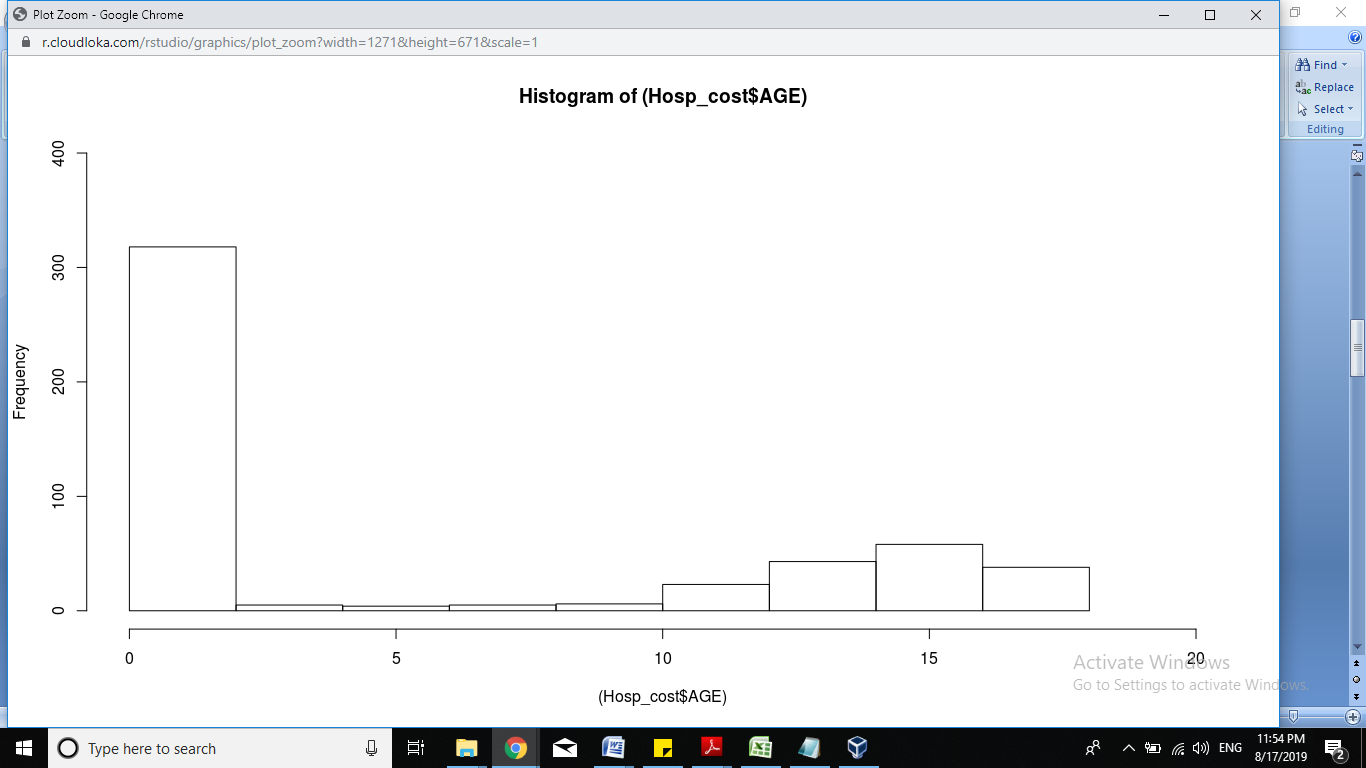
aggregate(TOTCHG~AGE,Hosp\_cost,mean)

**Output Screenshot:**





**Histogram**:



**Insights:**

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.

1. **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.**

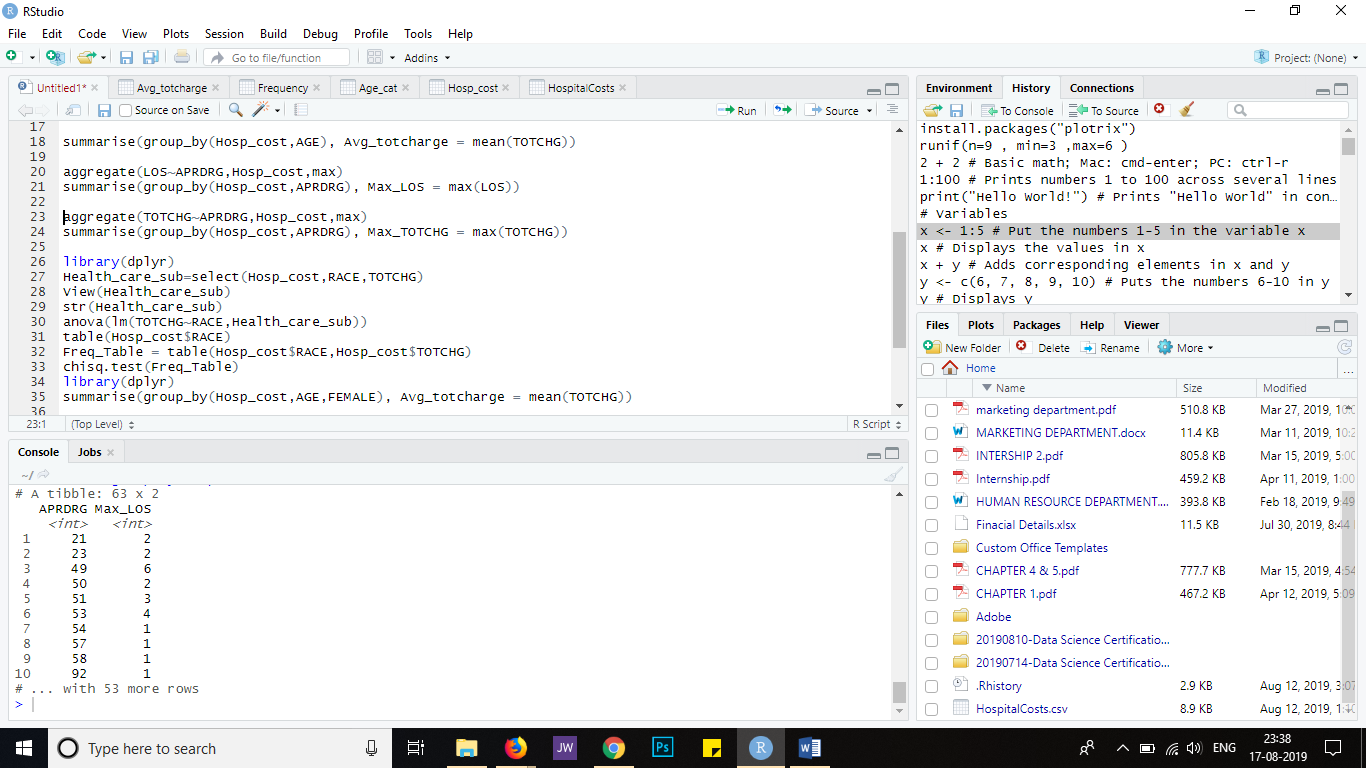
**R code:**

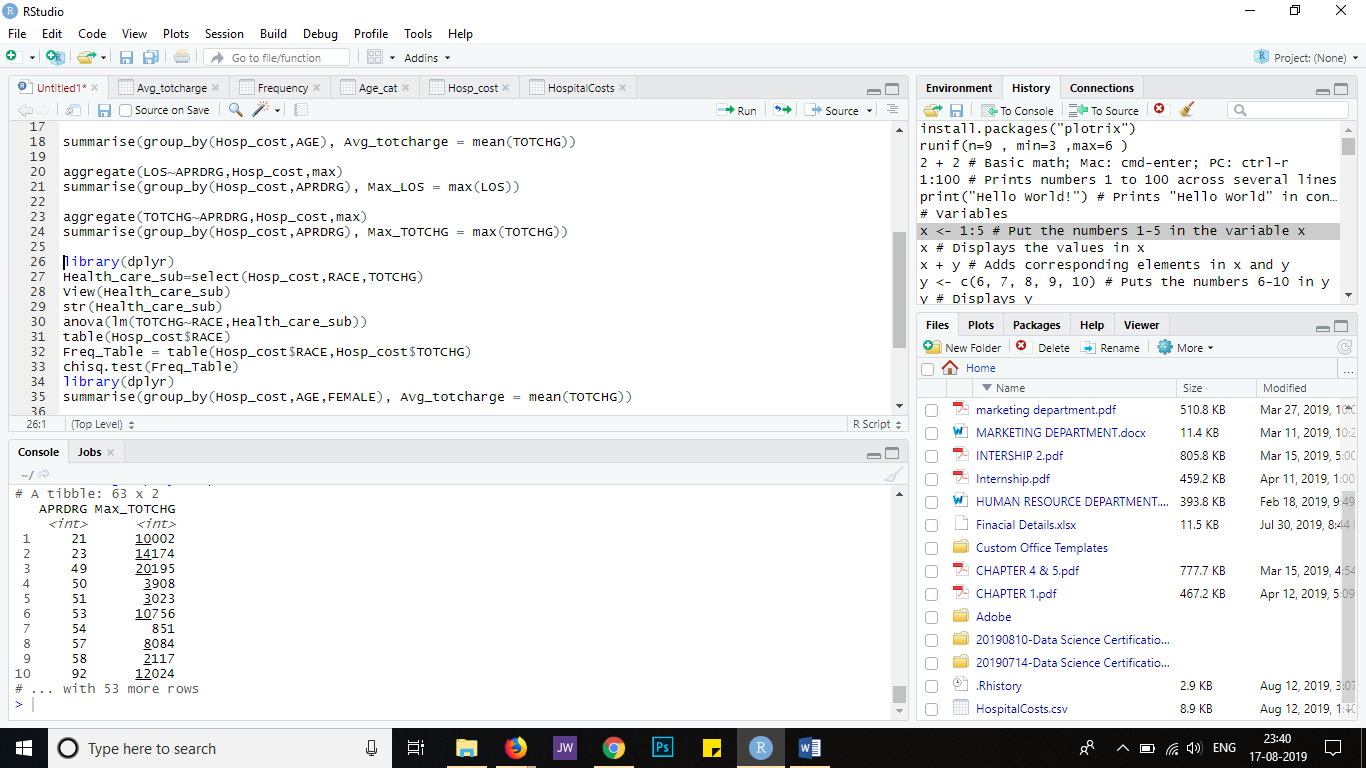
library(dplyr)

summarise(group\_by(Hosp\_cost,AGE), Avg\_totcharge = mean(TOTCHG))

summarise(group\_by(Hosp\_cost,APRDRG), Max\_LOS = max(LOS))

**Output Screenshot:**





1. **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.**

**R code:**

library(dplyr)

Health\_care\_sub=select(Hosp\_cost,RACE,TOTCHG)

View(Health\_care\_sub)

str(Health\_care\_sub)

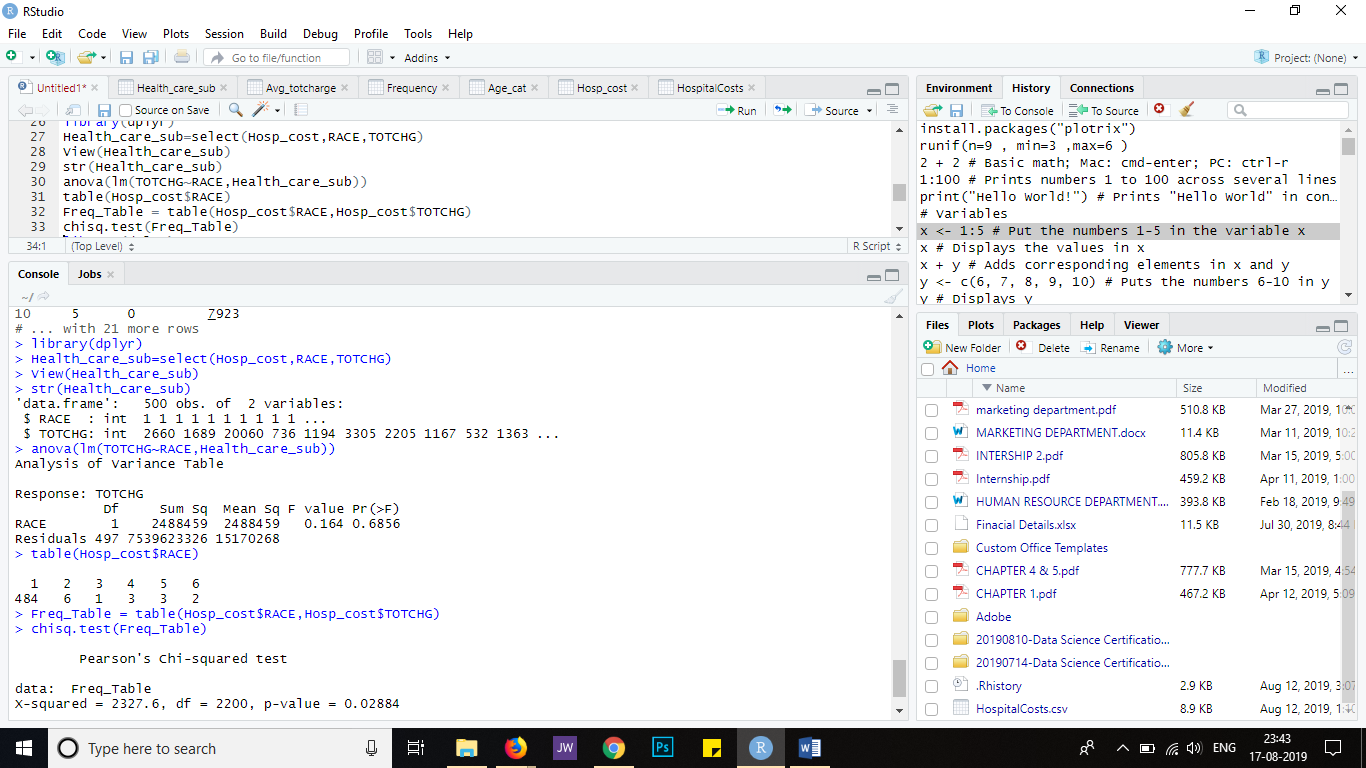
anova(lm(TOTCHG~RACE,Health\_care\_sub))

table(Hosp\_cost$RACE)

Freq\_Table = table(Hosp\_cost$RACE,Hosp\_cost$TOTCHG)

chisq.test(Freq\_Table)

**Output Screenshot:**



**Insights:**

By using the anova function and the chi squared test we came to know that hospital charges more from one race of the patient than the other.

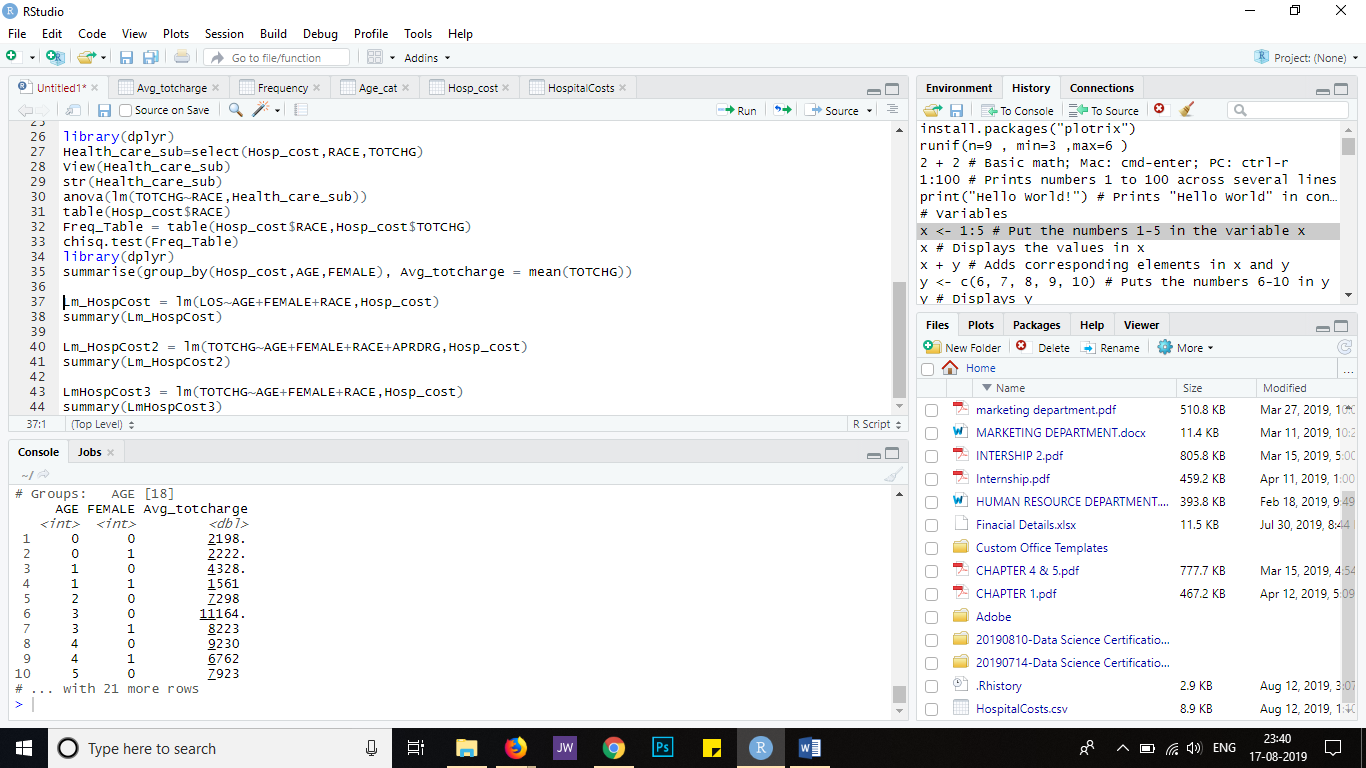
1. **To properly utilize the costs, the agency has to analyze the severity of the hospital costs by age and gender for proper allocation of resources.**

**R code:**

library(dplyr)

summarise(group\_by(Hosp\_cost,AGE,FEMALE), Avg\_totcharge = mean(TOTCHG))

**Output Screenshot:**

****

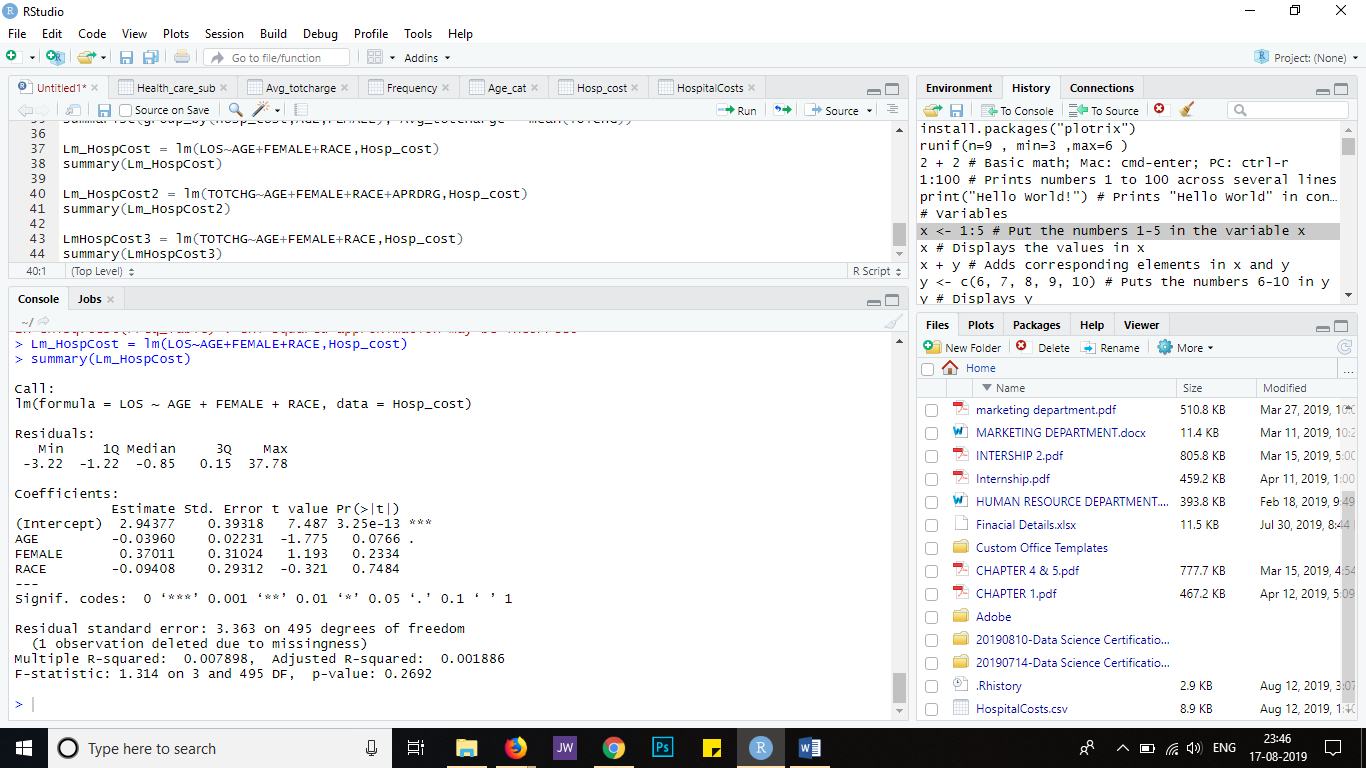
1. **Since the length of stay is the crucial factor for inpatients, the agency wants to find if the length of stay can be predicted from age, gender, and race.**

**R code:**

Lm\_HospCost = lm(LOS~AGE+FEMALE+RACE,Hosp\_cost)

summary(Lm\_HospCost)

**Output Screenshot:**

****

**Insights:**

By using the linear model function we came to know that the length of stay can be predicted from Age, Race & Gender.

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

**R code:**

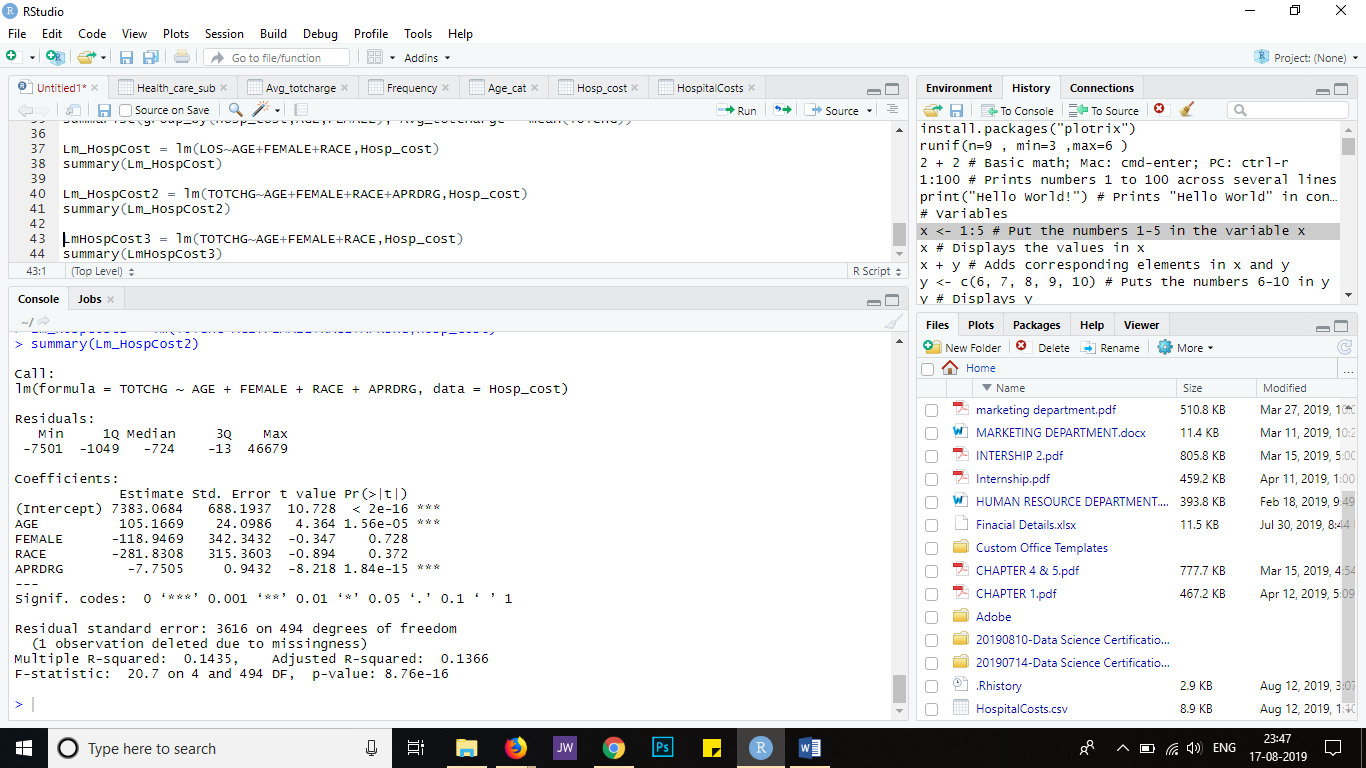
Lm\_HospCost2 = lm(TOTCHG~AGE+FEMALE+RACE+APRDRG,Hosp\_cost)

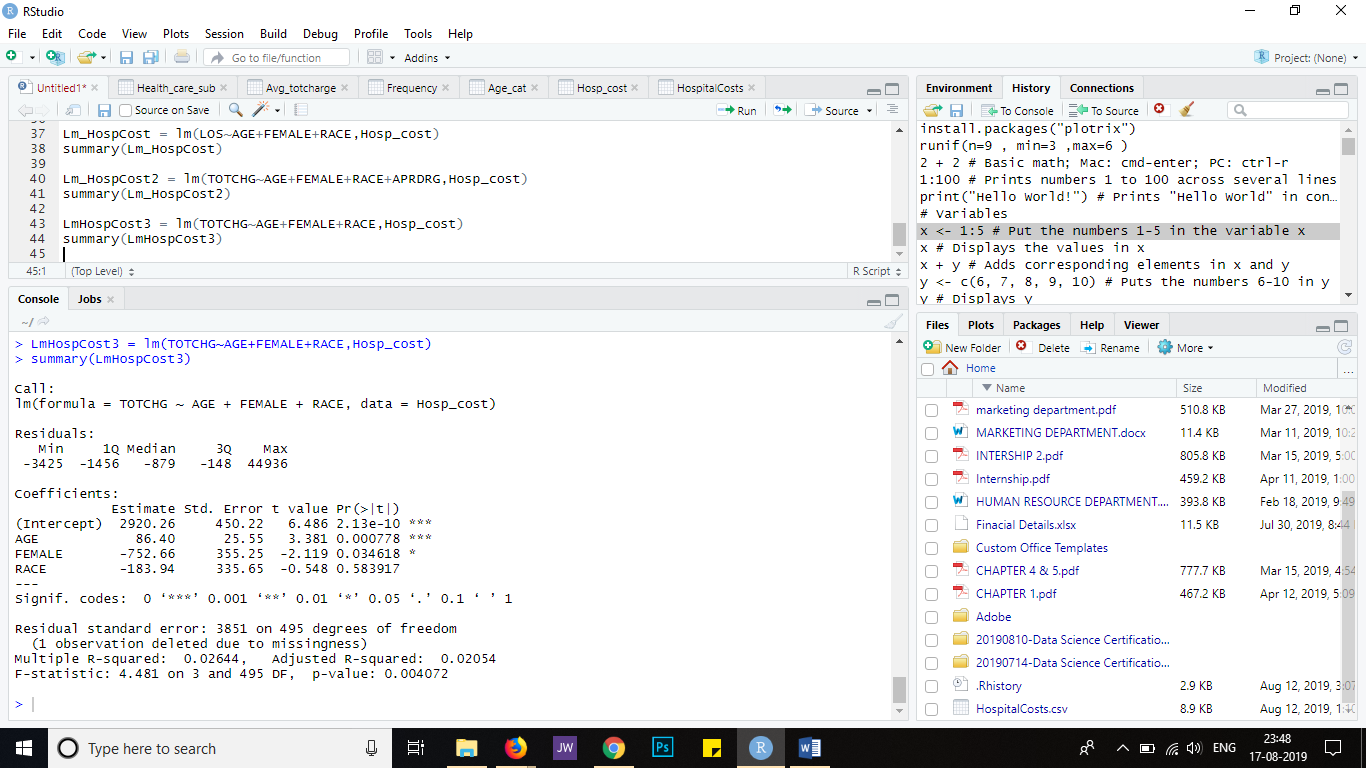
summary(Lm\_HospCost2)

LmHospCost3 = lm(TOTCHG~AGE+FEMALE+RACE,Hosp\_cost)

summary(LmHospCost3)

**Output Screenshot:**

****

****

**Insights:**

After analyzing the report we can come to a conclusion that All Patient Refined Diagnosis Related Groups (APRDRG) was the variable that affects the Hospital costs.