Health care Cost Analysis

Question 1. To record the patient statistics, the agency wants to find the age category of people who frequently visits the hospital and has the maximum expenditure.

Solution 1.

The as. factor () is called to make sure that the categories are not treated as numbers. The package “ggplot2” is used to display the histogram.

Code:

library(readxl) # To read an excel file.

Hospital <- read_excel ("D:/Simplilearn/Project Data Sets/7/Hospital.xlsx")

Location of excel file.

View (Hospital) # To view the inputted dataset.

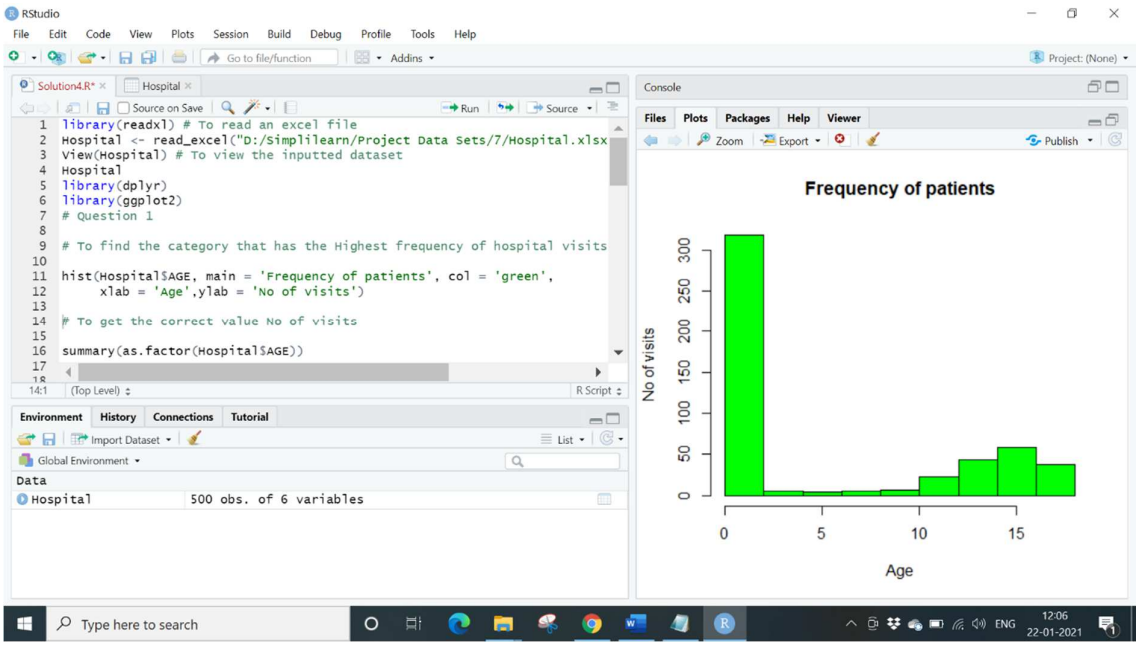
Hospital # To get the tibble 6*6 in the console.

library(dplyr)

library(ggplot2)

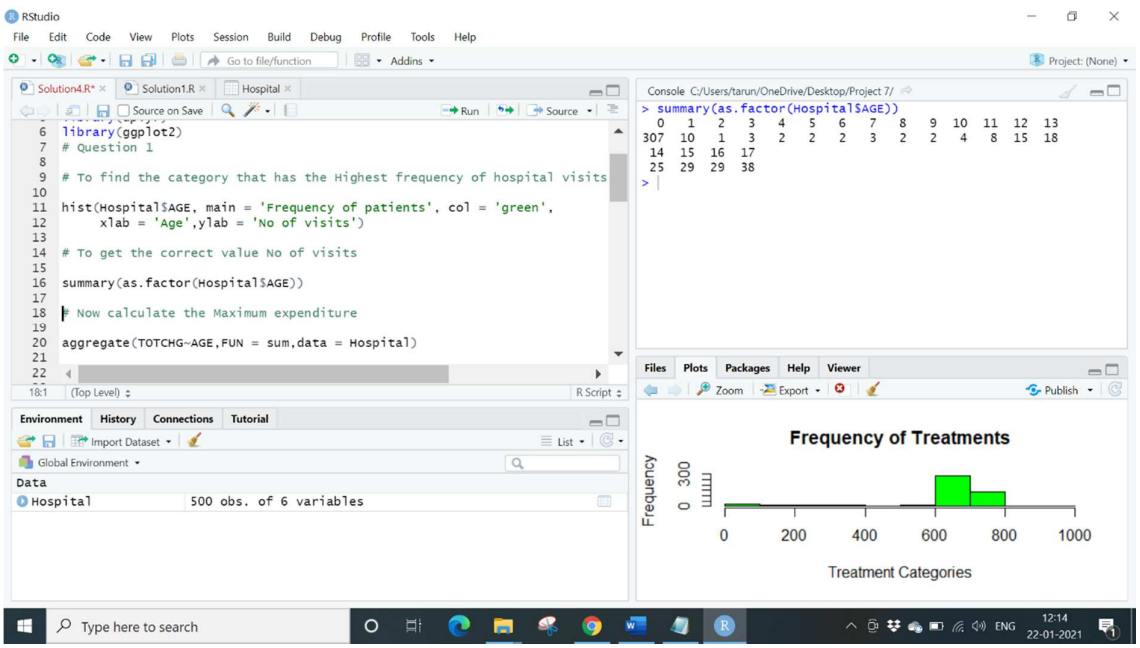
To find the category that has the Highest frequency of hospital visits

hist (Hospital$AGE, main = “Frequency of patients”, col = “green”, xlab = “Age”, ylab = “No of visits”)



To get the correct value No of visits

summary (as. factor (Hospital$AGE))



Conclusion:

From the graph that is displayed, we can see that (0) infants have the maximum frequency of hospital visit.

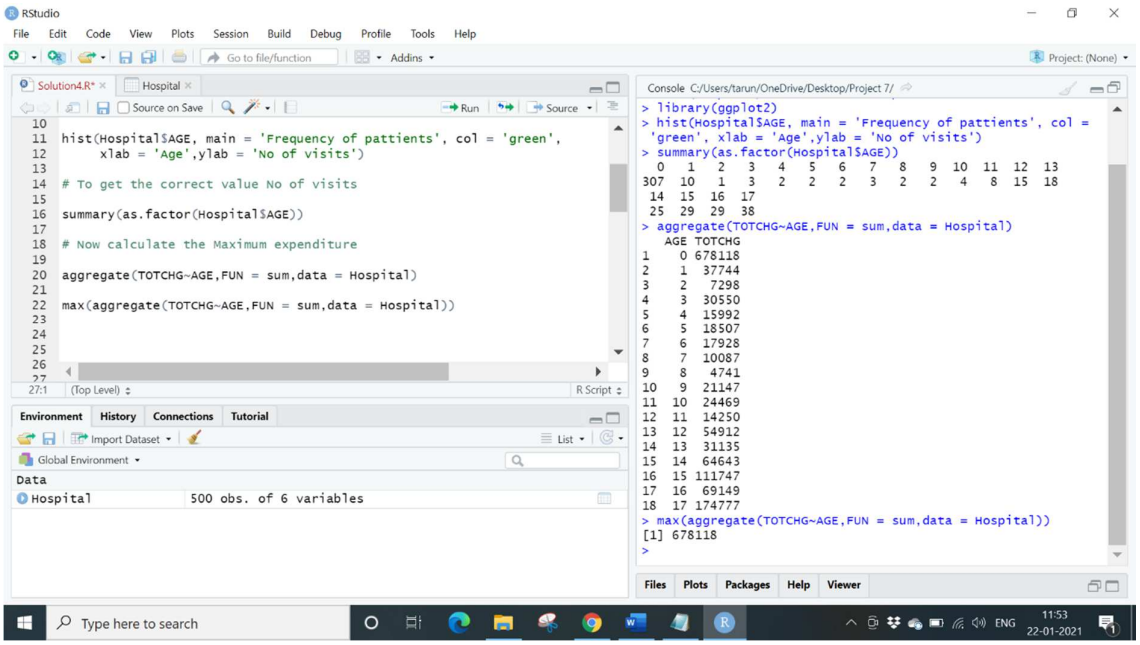
Now to calculate maximum expenditure:

Aggregate function is used to add the expenditure from each age and then max function used to find highest costs.

Code:

aggregate (TOTCHG~AGE, FUN=sum, data = Hospital)

max (aggregate (TOTCHG~AGE, FUN=sum, data = Hospital))



Conclusion:

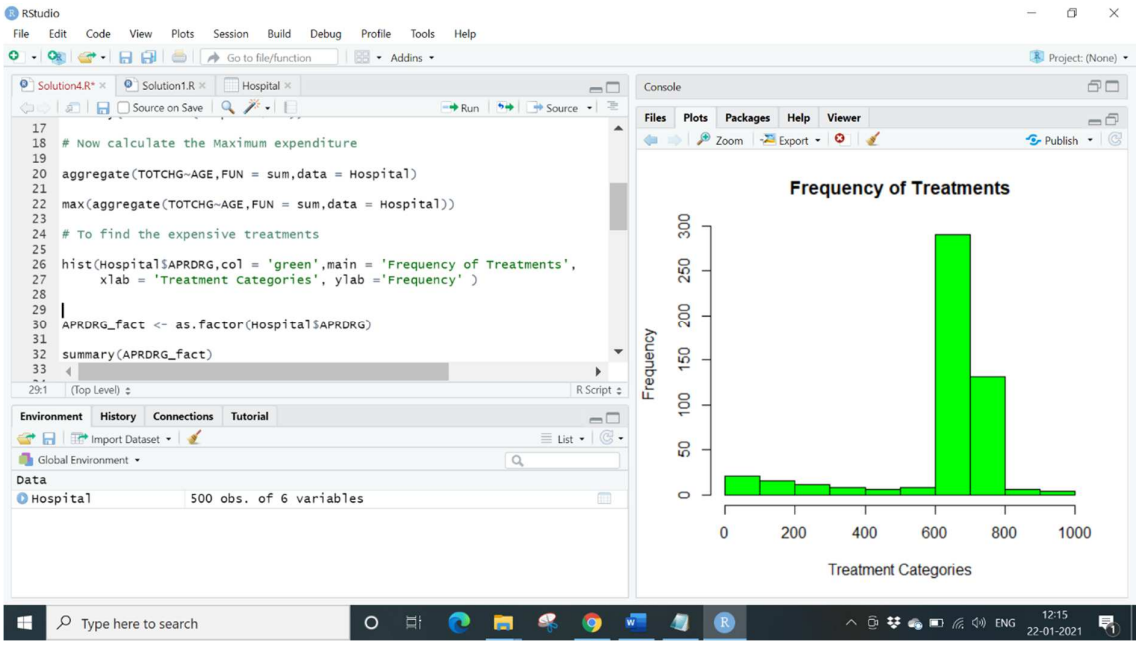
So again, result is age group 0 (infant) for maximum expenditure.

Question 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.

Solution 2.

Code:

```
hist (Hospital$APRDRG, col = 'green', main = 'Frequency of Treatments',
                xlab =    'Treatment Categories', ylab ='Frequency')
```



The as. factor () is called to make sure that the categories are not treated as numbers.

Code:

APRDRG_fact<-as. factor (Hospital$APRDRG)

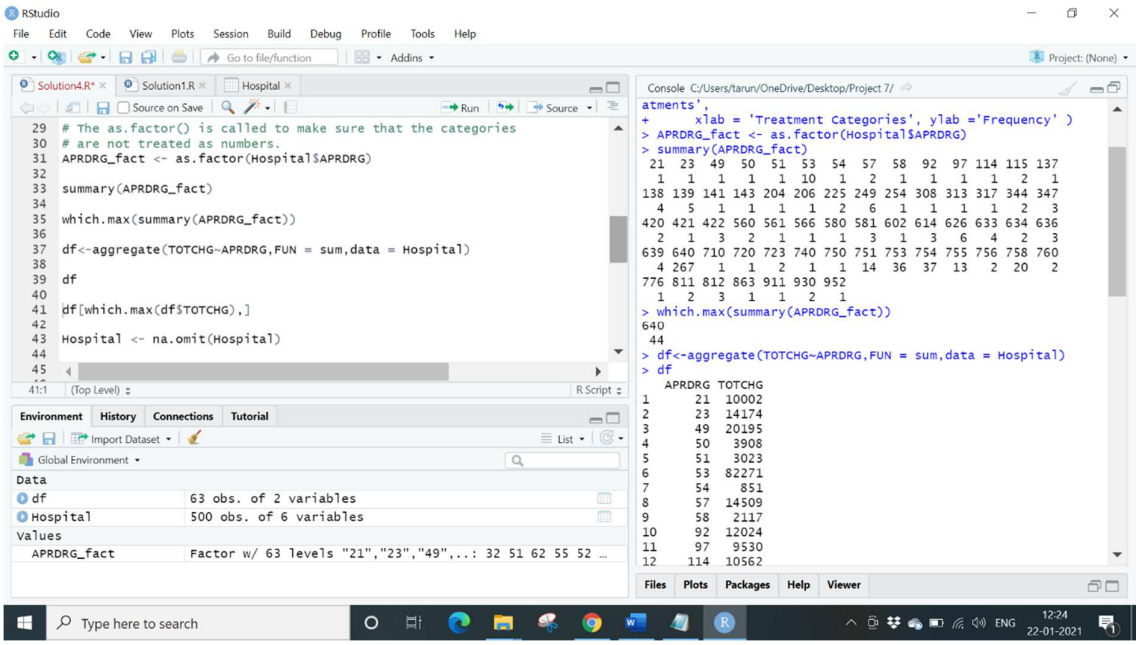
summary (APRDRG_fact)

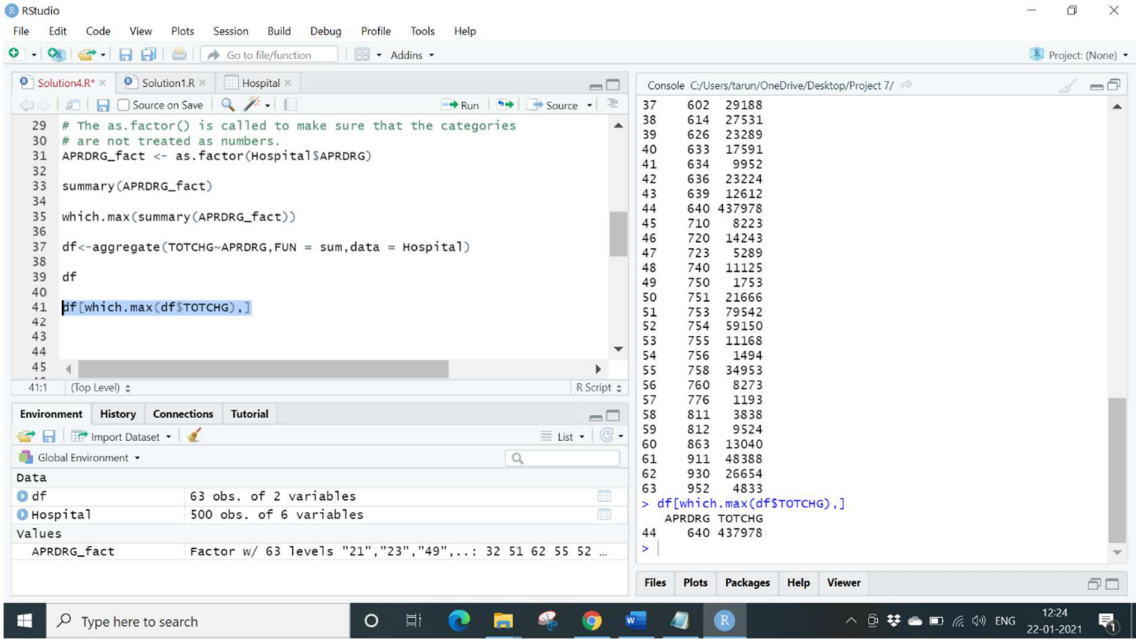
which.max (summary (APRDRG_fact))

df <-aggregate (TOTCHG~APRDRG, FUN = sum, data=hospital)

df

df[which.max(df$TOTCHG),]





Conclusion: So, the category 640 has the maximum hospitalizations along with this it also has the highest hospitalization cost.

Question 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.

Solution 3.

Remove the “NA” values from our database, then by using as.factor() the Race variable to generate a summary to verify whether race made an impact on the hospital costs we will use ANOVA function with TOTCHG as dependent variable and RACE as grouping variable.

Code:

Hospital <-na. omit(hospital) #first we remove “NA” values

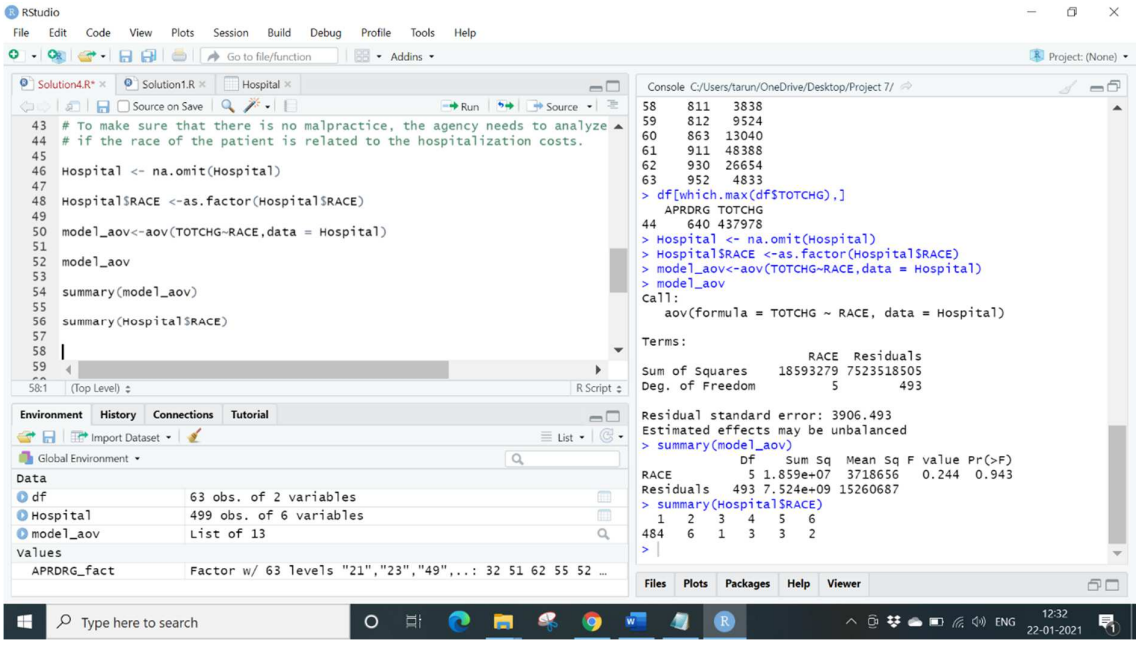
Hospital$RACE<-as. factor (Hospital$RACE)

model_aov<-aov (TOTCHG~RACE, data = Hospital)

model_aov #ANOVA RESULTS

summary(model_aov)

summary (Hospital$RACE)



Conclusion: The result shows that there is no relationship between race and hospital costs, thereby accepting the Null hypothesis.

Question 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.

Solution 4.

To analyze the costs we will use linear regression with TOTCHG(Cost) and independent variable along with AGE and Female as dependent variables

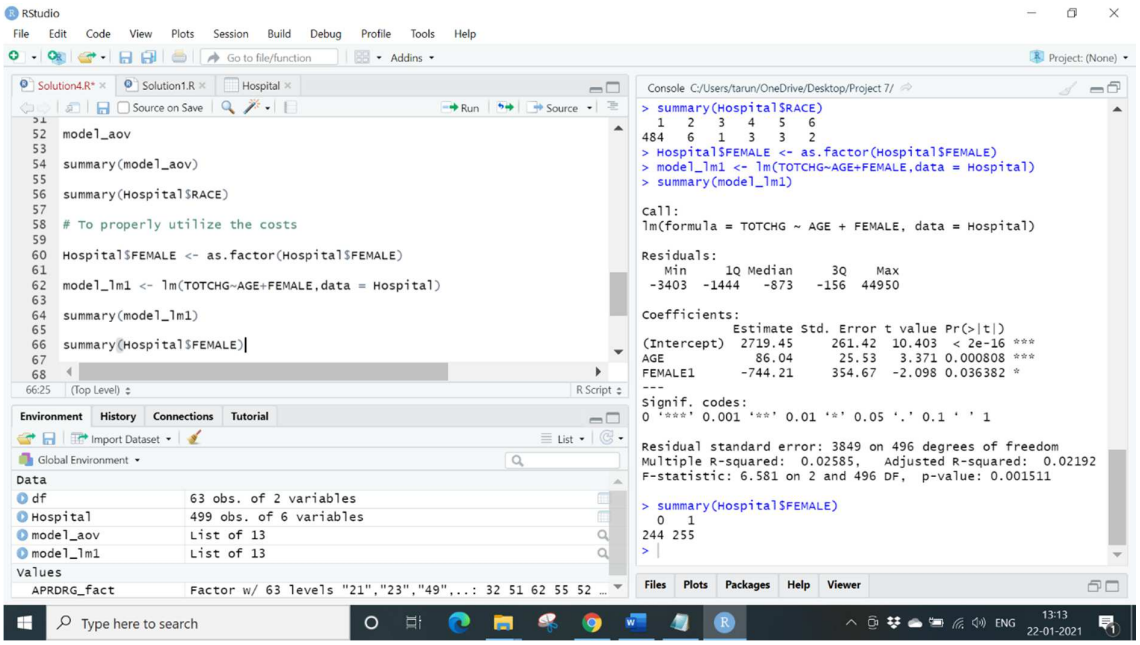
Code:

Hospital$FEMALE <- as. factor (Hospital$FEMALE)

model_lm1 <- lm (TOTCHG~AGE+FEMALE, data =Hospital) #calling Regression function

summary(model_lm1)

summary(hosp$FEMALE) #comapring genders



Conclusion: There are equal number of Females and Males and on an average (based on the negative coefficient values) females makes lesser hospital costs than males.

Question 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.

Solution 5.

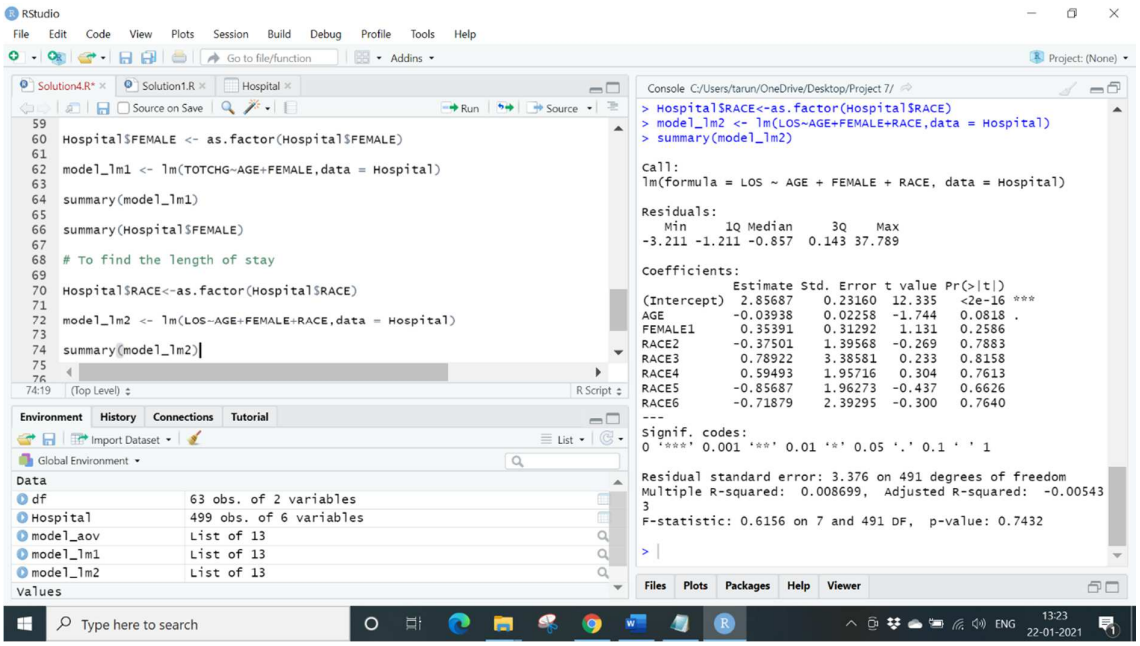
To find the length of stay we are using linear regression. Here length of stay is the dependent variable and age, gender and race are independent variables

Code:

Hospital$RACE <- as. factor (Hospital$RACE)

model_lm2 <- lm (LOS~AGE+FEMALE+RACE, data = Hospital)

summary(model_lm2)



Conclusion: All independent variables are quite high thus signifying that there is no linear relationship between the given variables, so we can't predict length of stay of a patient based on age, gender and race.

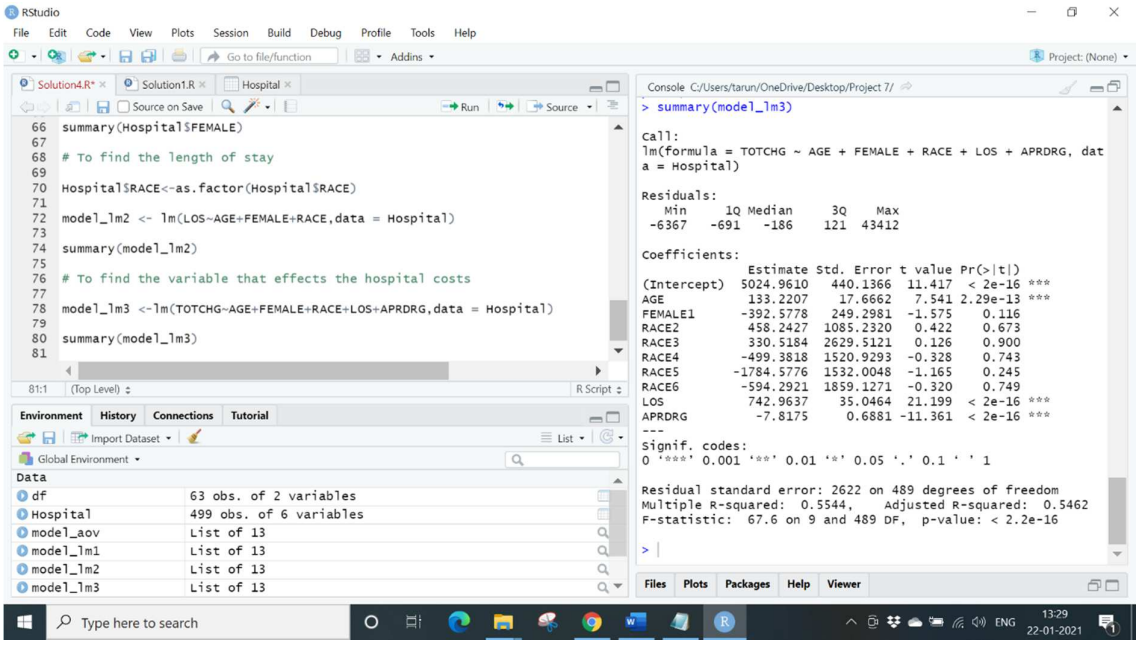
Question 6. To perform a complete analysis, the agency wants to find the variable that mainly affects the hospital costs.

Solution 6. To find the variable that mainly affects the hospital costs we use linear regression thus TOTCHG becomes dependent rest all becomes independent.

Code:

```
model_lm3 <- lm (TOTCHG~AGE+FEMALE+RACE+LOS+APRDRG, data = Hospital)

summary(model_lm3)
```



Conclusion: Age and Length of stay affect the total hospital costs.

Code:

```
library(readxl) # To read an excel file
Hospital <- read_excel ("D:/Simplilearn/Project Data Sets/7/Hospital.xlsx")
# Location of excel file
View (Hospital) # To view the inputted dataset
Hospital
library(dplyr)
library(ggplot2)
# Question 1
# To find the category that has the highest frequency of hospital visits and has
the maximum expenditure.
hist (Hospital$AGE, main = 'Frequency of patients', col = 'green',
    xlab = ‘Age’, ylab = 'No of visits')
# To get the correct value No of visits.
summary (as. factor (Hospital$AGE))
# Now calculate the Maximum expenditure.
aggregate (TOTCHG~AGE, FUN = sum, data = Hospital)
max (aggregate (TOTCHG~AGE, FUN = sum, data = Hospital))
# Question 2
# To find out the expensive treatments
hist (Hospital$APRDRG, col = ‘green’, main = 'Frequency of Treatments',
    xlab = 'Treatment Categories', ylab ='Frequency’)
# The as. factor () is called to make sure that the categories are not treated as
numbers.
APRDRG_fact <- as. factor (Hospital$APRDRG)
summary (APRDRG_fact)
which.max (summary (APRDRG_fact))
df <- aggregate (TOTCHG~APRDRG, FUN = sum, data = Hospital)
```

```
df

df[which.max(df$TOTCHG),]

# Question 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.

Hospital <- na. omit (Hospital)

Hospital$RACE <-as. factor (Hospital$RACE)

model_aov<-aov (TOTCHG~RACE, data = Hospital)

model_aov

summary(model_aov)

summary (Hospital$RACE)

# Question 4

# To properly utilize the costs

Hospital$FEMALE <- as. factor (Hospital$FEMALE)

model_lm1 <- lm (TOTCHG~AGE+FEMALE, data = Hospital)

summary(model_lm1)

summary (Hospital$FEMALE)

# Question 5

# To find the length of stay

Hospital$RACE<-as.factor (Hospital$RACE)

model_lm2 <- lm (LOS~AGE+FEMALE+RACE, data = Hospital)

summary(model_lm2)

# Question 6

# To find the variable that effects the hospital costs

model_lm3 <- lm (TOTCHG~AGE+FEMALE+RACE+LOS+APRDRG, data = Hospital)

summary(model_lm3)
```

(End of Project)