Summer Training Project Report

Data Science with R (via Simplilearn)

Name of the Student: Gibran Khan Tareen

Registration Number: 12100173

Institution: Lovely Professional University

Program: BTech CSE (Lateral Entry)

Topic Chosen for Project: Healthcare Cost Analysis

DESCRIPTION-

Background and Objective:

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.

Domain: Healthcare

(Please Turn Over)

Dataset Description:

Here is a detailed description of the given dataset:

Attribute	Description
Age	Age of the patient discharged
Female	A binary variable that indicates if the patient is female
Los	Length of stay in days
Race	Race of the patient (specified numerically)
Totchg	Hospital discharge costs
Aprdrg	All Patient Refined Diagnosis Related Groups

Dataset for the Given Question can be download from here:- HospitalCosts.csv

Analysis to be done (Goals):

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

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.

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.

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.

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.

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

My Solution-

Since this course is based on R programming so I will be using R Studio to solve it. Now to start with the solution, the very first step will be to import our dataset provided to us in the question (HospitalCosts.csv) and mount it as table in R studio.

We can simply do that using:
gkthospital<-read.csv("Location of file\\HospitalCosts.csv")

(**gkt** stands for Gibran Khan Tareen)

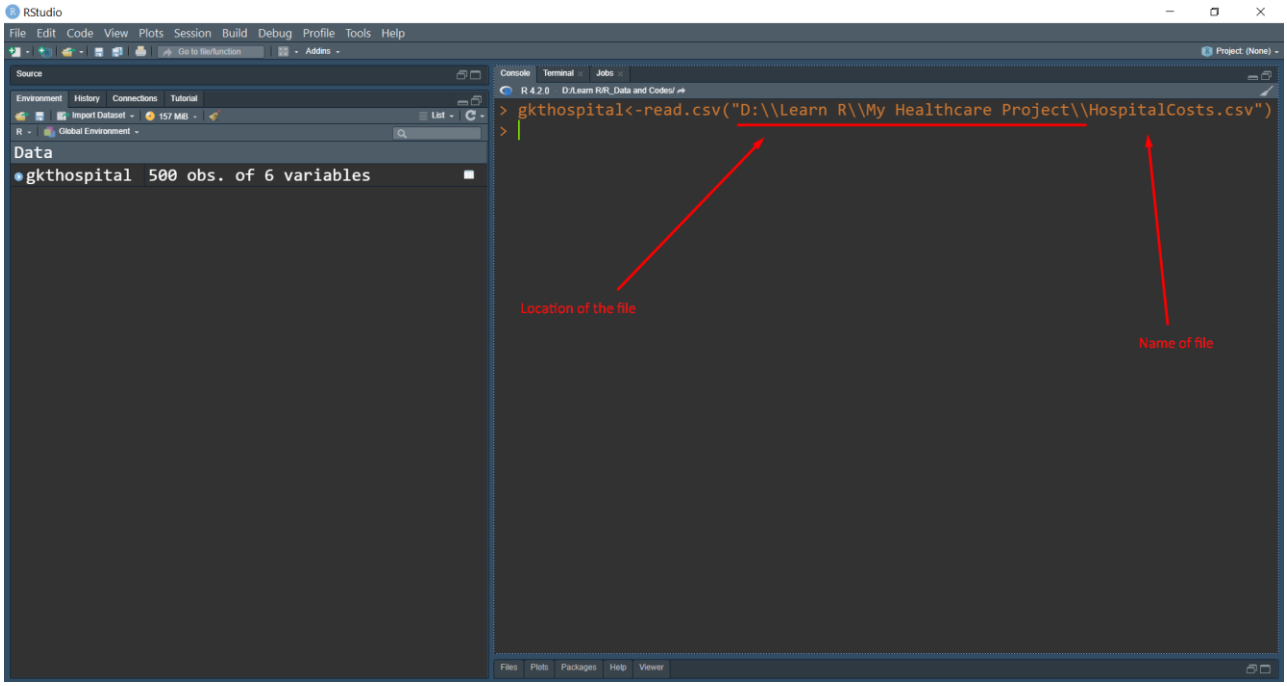


Figure 1: Reading Dataset

(Please Turn Over)

Then we can simply check if the dataset is imported successfully by simply entering: ***gkthospital***

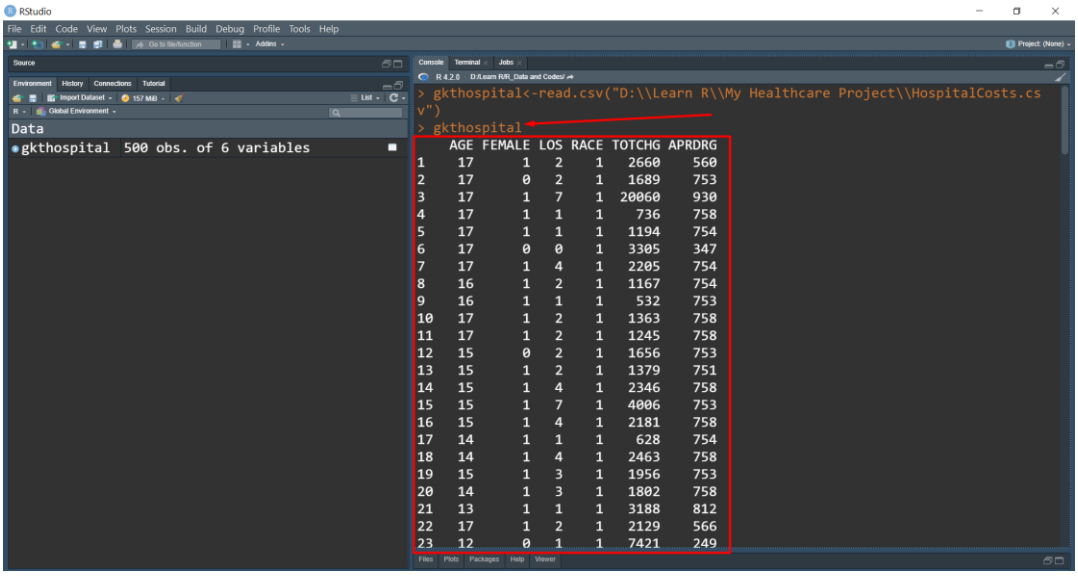


Figure 2: Successfully Imported

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

If we read carefully, we are asking to find 2 things in this Goal-

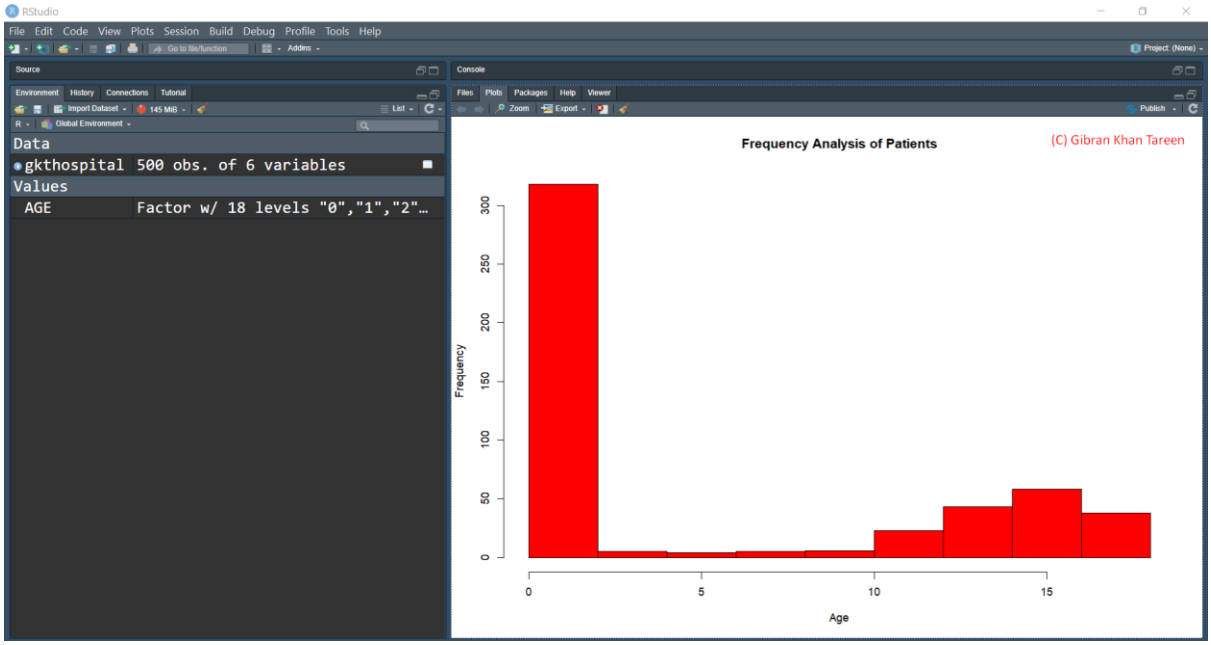
I. The age category of people who frequently visit the hospital

II. The age category that has maximum expenditure

Now first, to find the category with the maximum frequency of hospital visits we will have to visualise the whole data to get an overview of all the categories. The best way to present this data for frequency analysis we can use a Histogram.

(Please Turn Over)

We can simply do that by:
hist(gkthospital$AGE,main = "Frequency Analysis of Patients",col = "red",xlab = "Age")



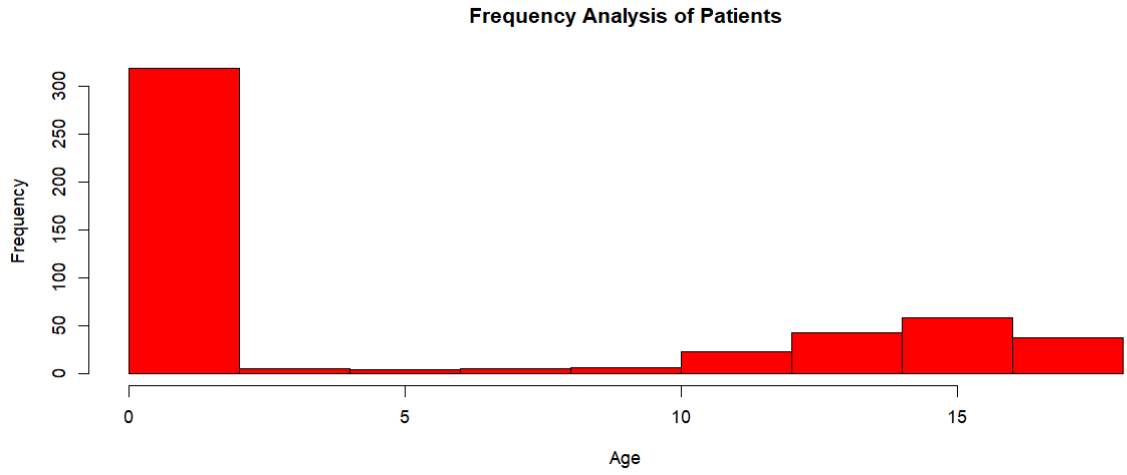


Figure 3: Frequency Analysis of Patients

Now, to get a detailed information from our histogram we will use “factor” function for the “AGE” column and ‘summary’ function for the detailed summary of the data.

We can simply do that by:
attach(gkthospital) ***[attach is used to access the variables present in the data framework without calling the data frame]***

AGE<-as.factor(AGE)
summary(AGE)

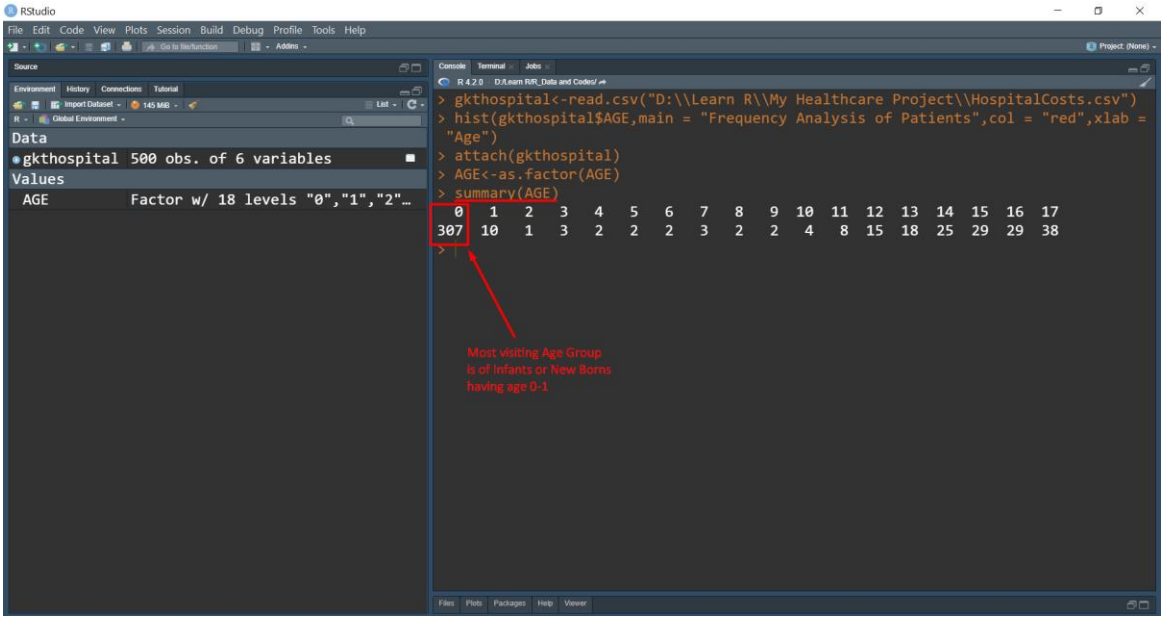


Figure 4: Summarized Analysis

Conclusion: From the above results we determine that infant category or new born category has the maximum hospital visits (300+ visits). **Thus (AGE 0 or <1) patients have the maximum hospital visits followed by Ages 15,16,17**.

Subgoal 2: To find Age group that has Maximum expenditure

To do this we need to get the summary statistics of a specific group / column's data (we need for the column TOTCHG ie. Total expense) in our dataset. For this task we can simply use "AGGREGATE" function. We need total expenditures (or the sum total of the expense) for the age groups so we will use "FUNCTION" attribute = "SUM" for AGGREGATE function.

{TOTCHG is given in Dataset as Hospital discharge costs}

(Please Turn Over)

We can simply do that by:
aggregate(TOTCHG~AGE,FUN=sum,data = gkthospital)
max(aggregate(TOTCHG~AGE,FUN=sum,data = gkthospital)) -> {To get MAX Expenditure}

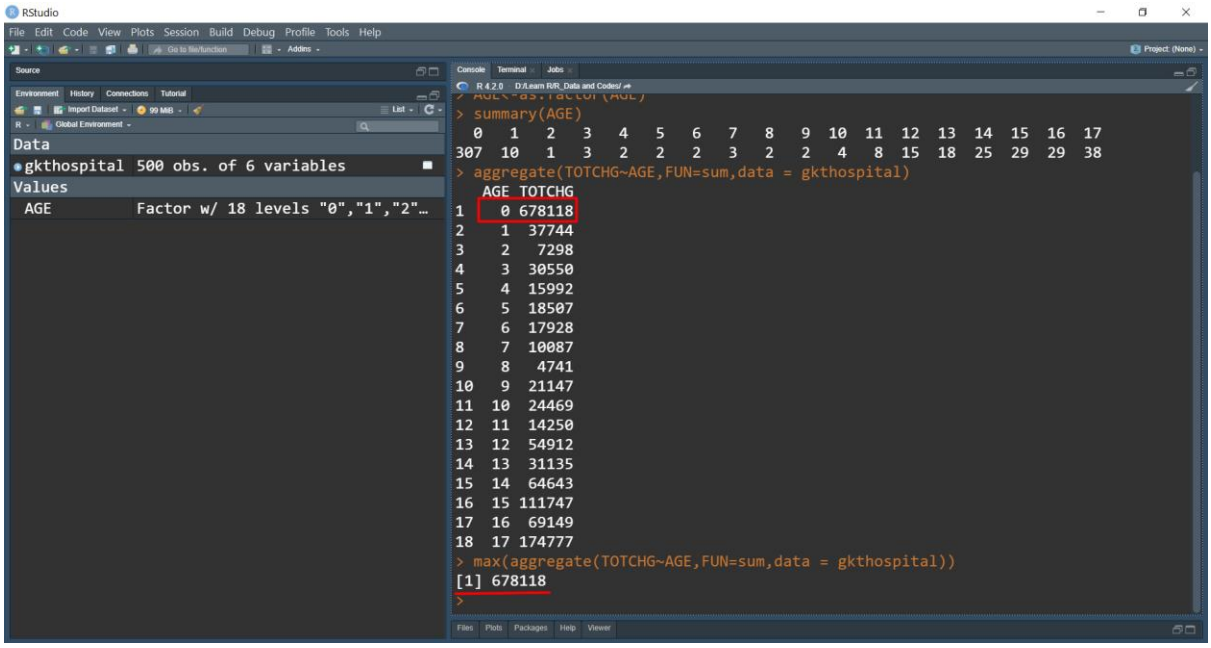


Figure 5: Subgoal 2

Conclusion: From the above results we determine that here also the Infant category or new born category **(AGE group 0 or < 1) has the maximum Hospital expenditure** of Rs 678118

Hence Goal #1 Accomplished ✔

Goal 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.

If we read carefully here also, we can divide the goal into 2 subgoals-

I. The diagnosis related group that has maximum hospitalization
II. The diagnosis related group that has maximum expenditure.

Now first to find the diagnosis related group with the maximum hospitalization visits we will have to visualise the whole data on basis of their frequency to get an overview all the categories. Same as last time, The best way to present this data for frequency analysis we will be using a Histogram.

We can simply do that by:
hist(APRDRG,col = "blue",main = "Frequency of Treatments",xlab = "Treatment Categories")

{ADPRDG is given in Dataset: All Patient Refined Diagnosis Related Groups}

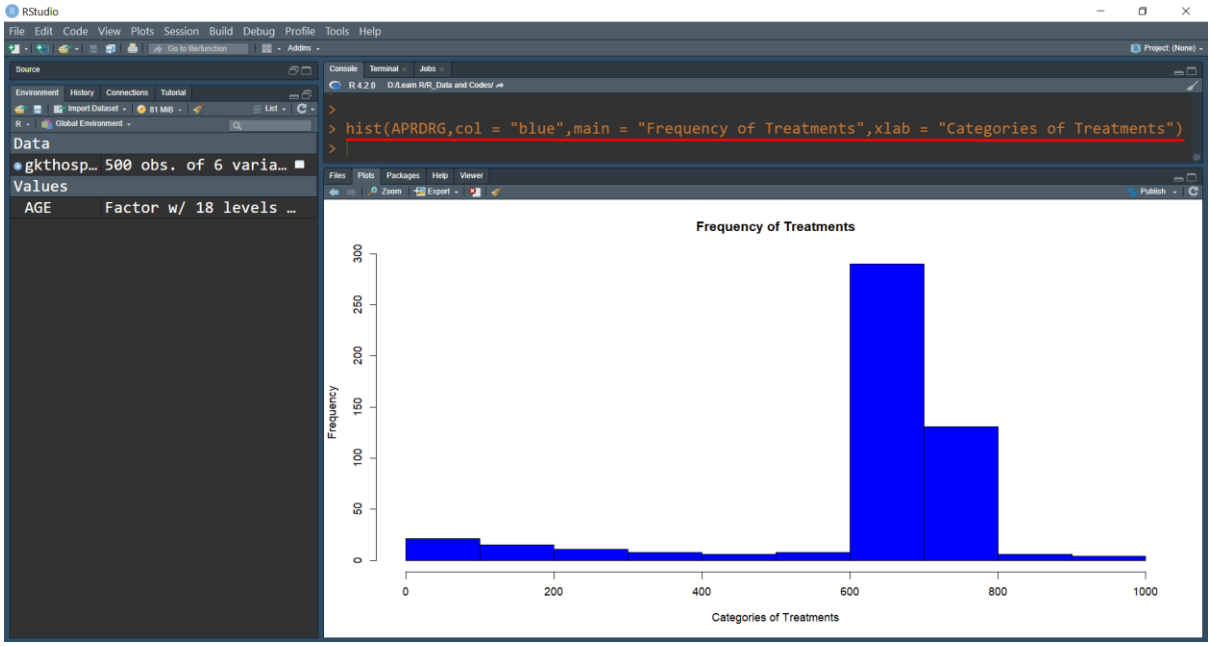


Figure 6: Histogram

Now before proceeding further we will have to make sure that category column(“APRDRG”) is numerical. After that we will generate a summary for the data along with “which.max” function to determine which category of treatment has max expense. This will be followed by aggregate function used in a similar way as above.

We can simply do that by:
APRDRG_ensure<-as.factor(gkthospital$APRDRG)
summary(APRDRG_ensure)
which.max(summary(APRDRG_fact)) {to get max from Summary analysis}

(Please Turn Over)

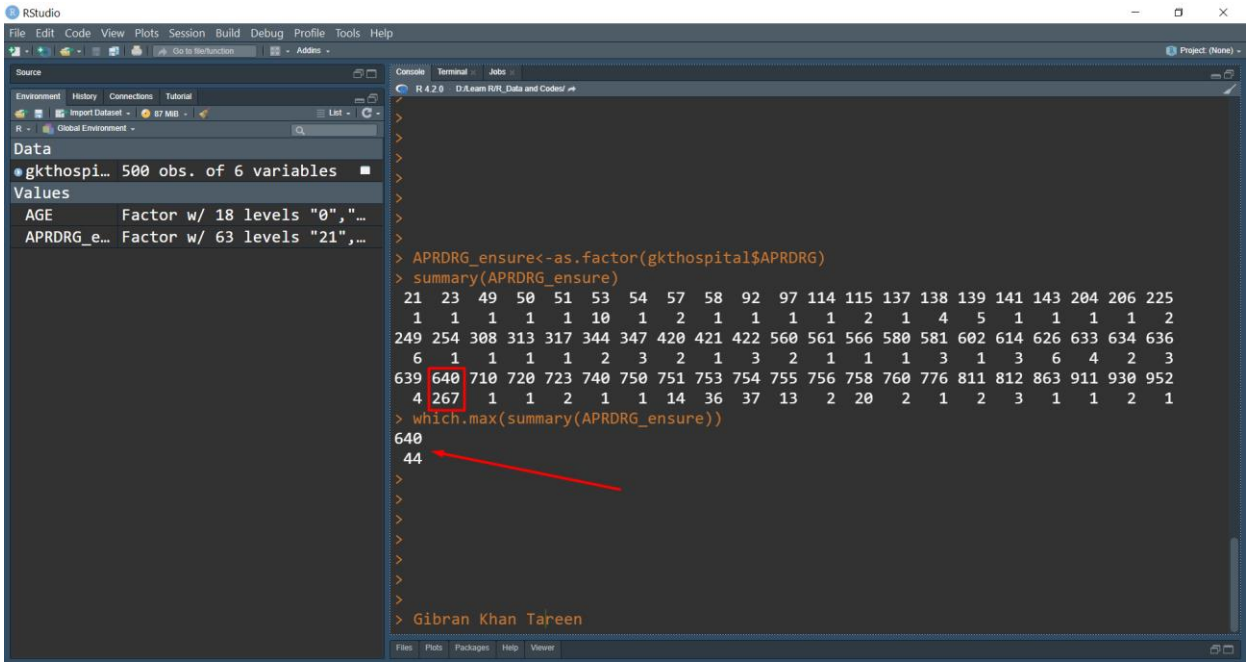


Figure 7: Result of Summary

Conclusion: From the above summary results we determine that **"640" is the diagnosis-related group which has the maximum hospitalization frequency of 267**.

Subgoal 2: Find the Diagnosis-related group with Maximum expenditure

To do this, we can use the approach we did in previous goal. We will have to get the sum total or aggregate of two columns (TOTCHG and APRDRG). So for this task again we will use "AGGREGATE" function.

We can simply do that by:

gkt<-aggregate(TOTCHG~APRDRG,FUN = sum,data=gkthospital)
gkt
gkt [which.max(gkt$TOTCHG),] (To get the APRDRG group with max expense)

(Please Turn Over)

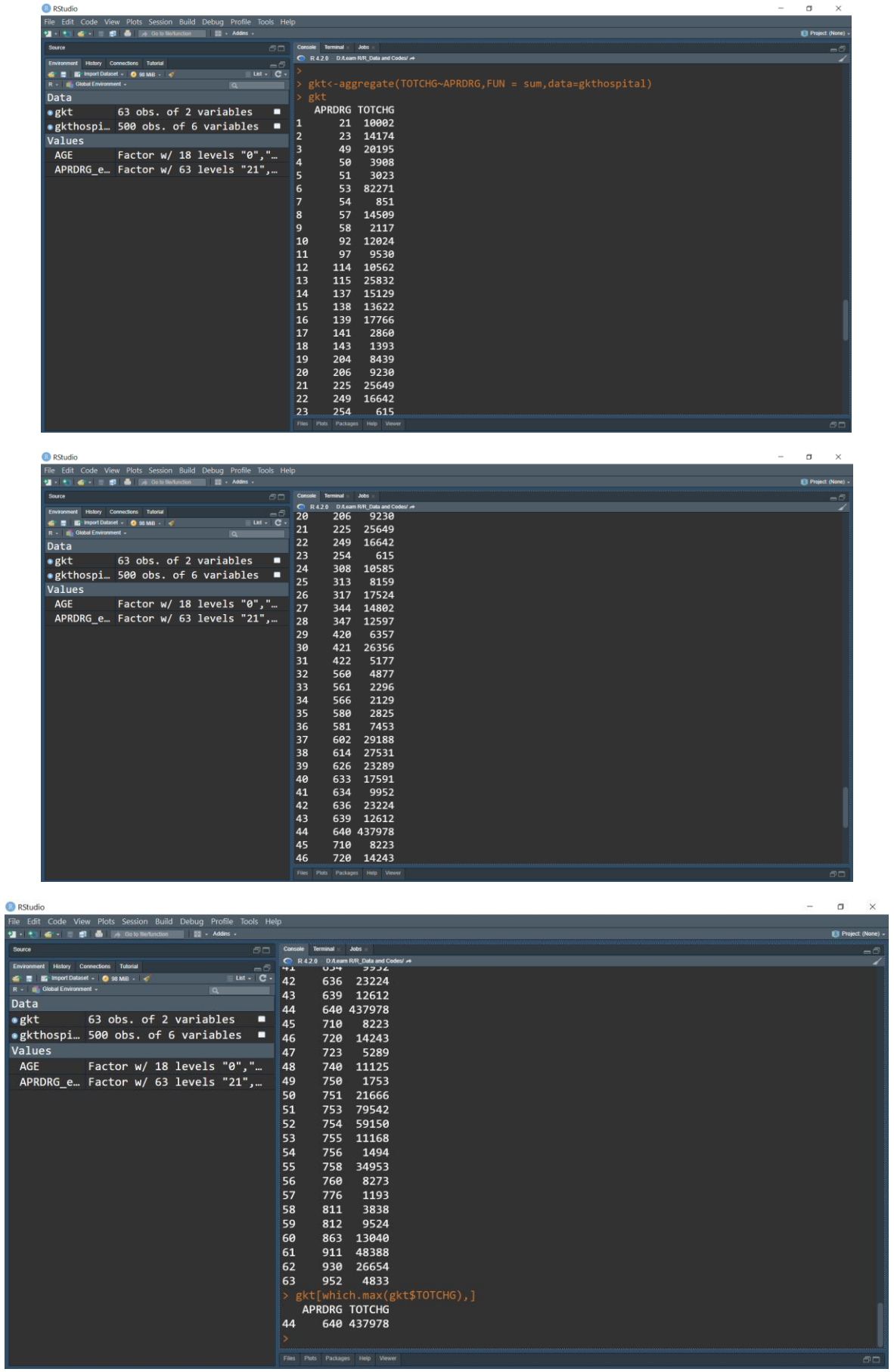


Figure 8: Result of the Data Analysis

Conclusion: From the above detailed results we determine that **"640" is the diagnosis-related group which has the maximum hospitalization frequency as well as the Maximum treatment expenditure of Rs 437978**

Hence Goal #2 Accomplished ✔

Goal 3- To make sure that there is no malpractice, the agency needs to analyse if the race of the patient is related to the hospitalization costs.

For this question we will be using the 'RACE' attribute given to us in the dataset for the analysis. To solve this goal, we first need to understand it. We can determine that there is no malpractice of racism in terms of pricing with the patients only if, the patient's race will make no impact on the hospital expense pricing. So in order to proceed with it, first we need to remove all the "NA" values from the dataset. Once we achieve that, then we will factorize the "RACE" column of the dataset and summarize it. Now if we see carefully, the process of verifying the impact of people's race on pricing, its more of a type of hypothesis testing for population variance and also we need to investigate relations between categorical variables and continuous variable. Therefore, we will use the concept of Hypothesis testing by implementing ANOVA Test. We will make a null hypothesis that "There is no RACISM in terms of Hospitalization costs". To solve these requirements, we will use "ANOVA" function with 'TOTCHG' as dependent variable and 'RACE' as a grouping variable.

{Note: aov() performs 1 way ANOVA. The generic function anova() is used to compute the analysis of variance (or deviance) tables for one or more fitted model objects}

We can simply do that by:

```
gkthospital<-na.omit(gkthospital)
gkthospital$RACE<-as.factor(gkthospital$RACE)
gktmodel_for_aov<-aov(TOTCHG~RACE,data = gkthospital)
gktmodel_for_aov
summary(gktmodel_for_aov) = {To get the detailed summary for our model}
```

(Please Turn Over)

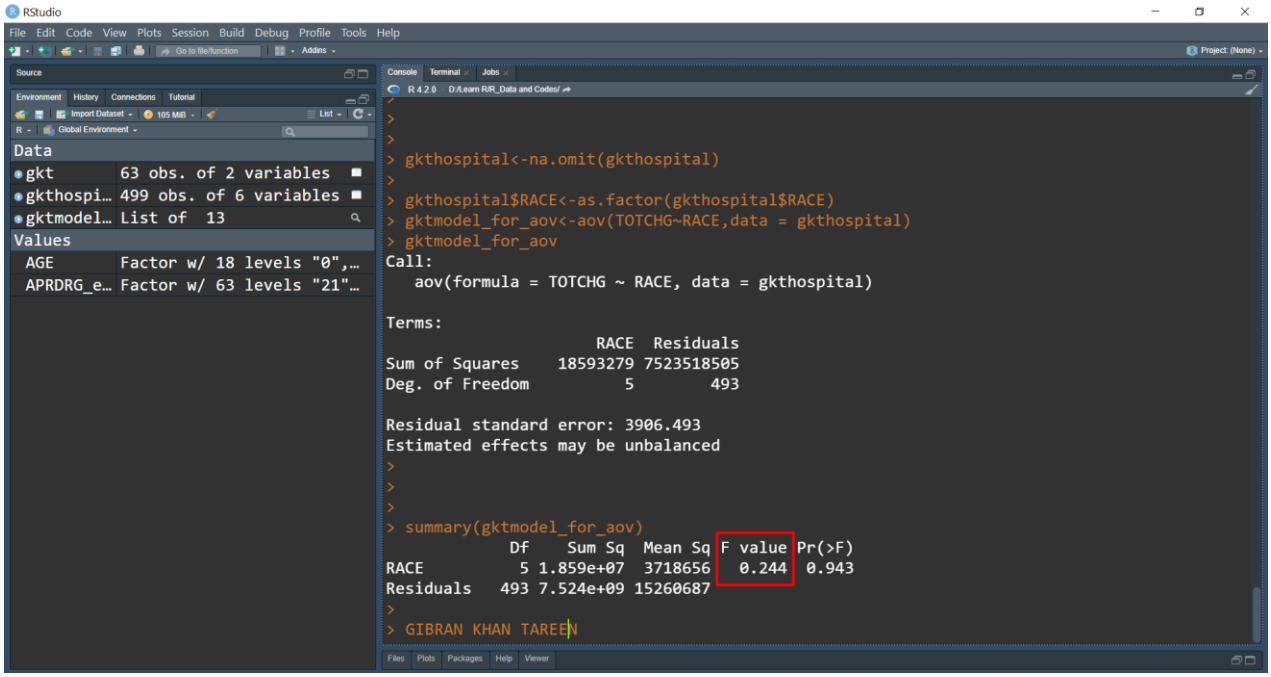


Figure 9: Results of the Summary

We can also check the summary for “RACE” column in our dataset to get the summarized analysis of the maximum cost of hospitalization per race of the people.

We can simply do that by:
summary(gkthospital$RACE)

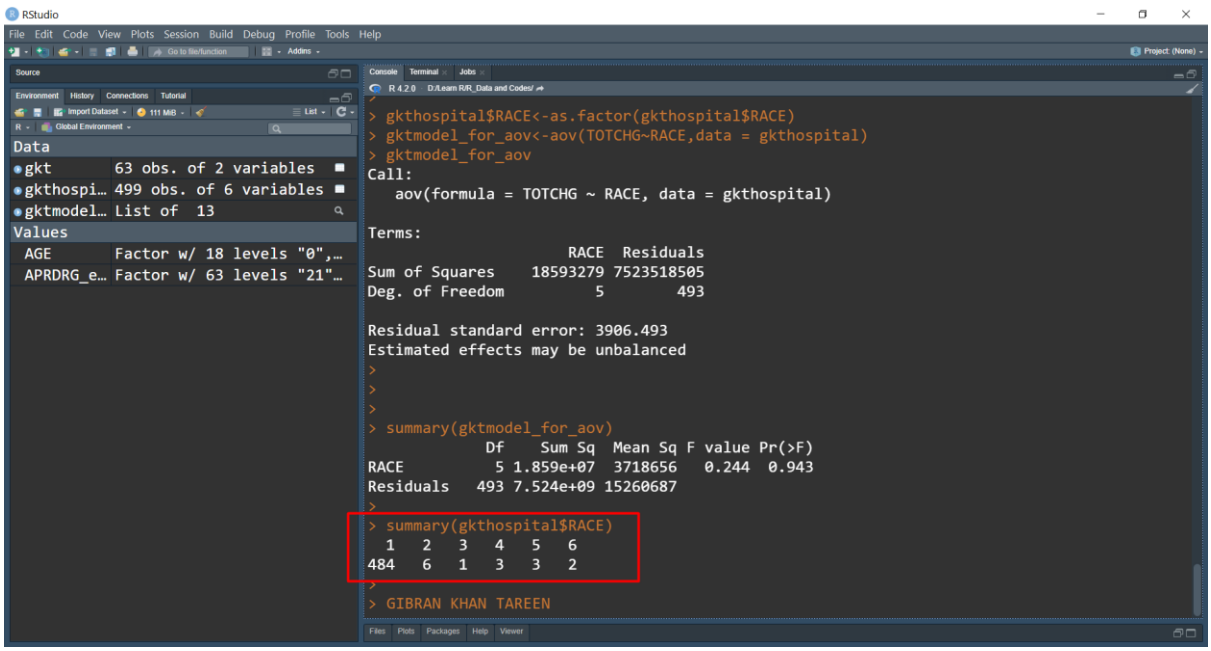


Figure 10: Result of ‘RACE’ Column Summary

Conclusion: From the above results, we can obtain the following conclusions:

1. In the summarized analysis of our "gktmodel_for_aov" we found that the "F value" is quite low (i.e 0.244). This clearly indicates that the variation between hospitalization costs among different races is very small as compared to the variation of hospitalization costs within each race. The "P value" was quite high, so both of these observations indicate that there is no relationship between race and hospital costs, thereby accepting the Null hypothesis.

2. Additionally, at last when we took out the summary for "RACE" column in our dataset, we observed that we have more data for RACE 1 (484 out of 500 patients) in comparison to all other races. This makes the observations biased. At last, all we can say is that "There is Insufficient data or we don't have enough data to verify whether a patient's race affects hospitalization costs."

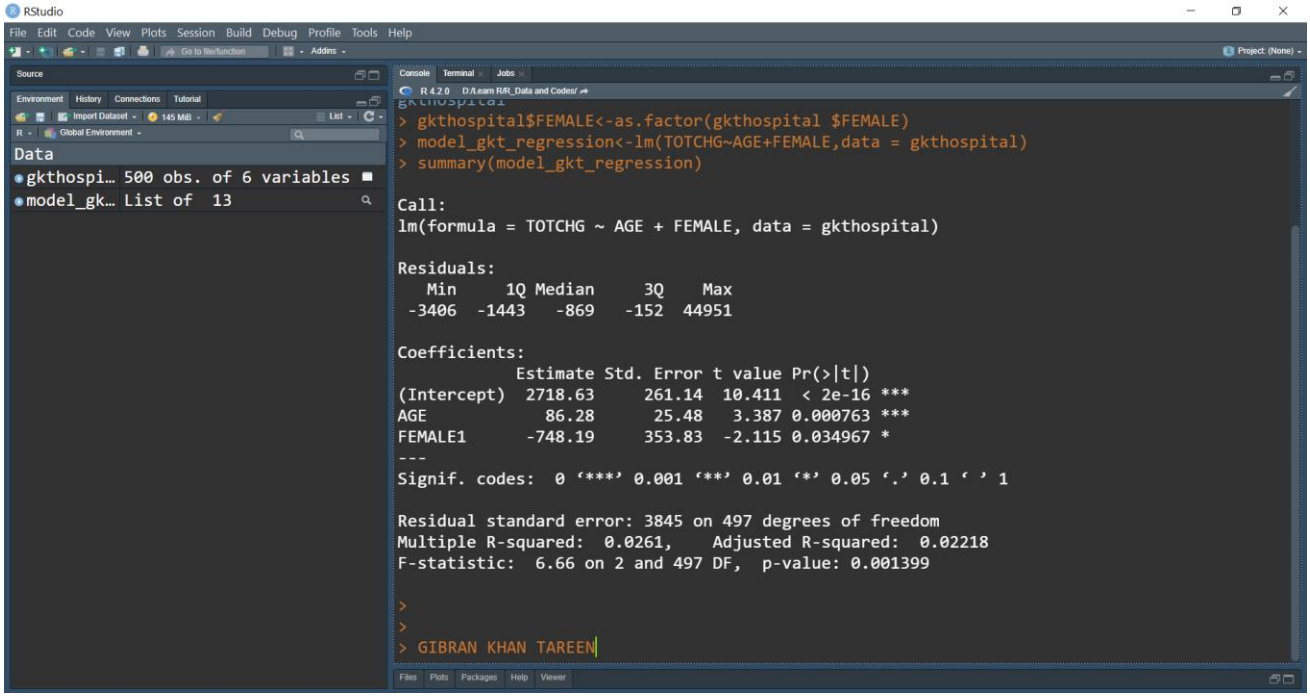
Hence Goal #3 Accomplished ✔

Goal 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.

For this goal we will be using the 'FEMALE' and 'AGE' attribute given to us in the dataset. It is important to note that 'FEMALE' attribute given in Dataset is a "binary variable that indicates if the patient is female". So since 'MALE' patients value is based on Dataset's "FEMALE" attribute's binary value, also then the "COST" will also be depended on that. So inorder to predict the value of a variable based on the value of another variable we will use Linear Regression. In our question we will use linear regression with "TOTCHG" (Cost) as independent variable (variable that is tested to see if they predict the outcome) along with "AGE" and "FEMALE" as dependent variables (the dependent variable represents the output or response).

We can simply do that by:

```
gkthospital$FEMALE<-as.factor(gkthospital $FEMALE)
model_gkt_regression<-lm(TOTCHG~AGE+FEMALE,data = gkthospital)
summary(model_gkt_regression)
```



Then we can take out the summary of the "***gkthospital$FEMALE***" to get an overview of "FEMALE" and "MALE" count

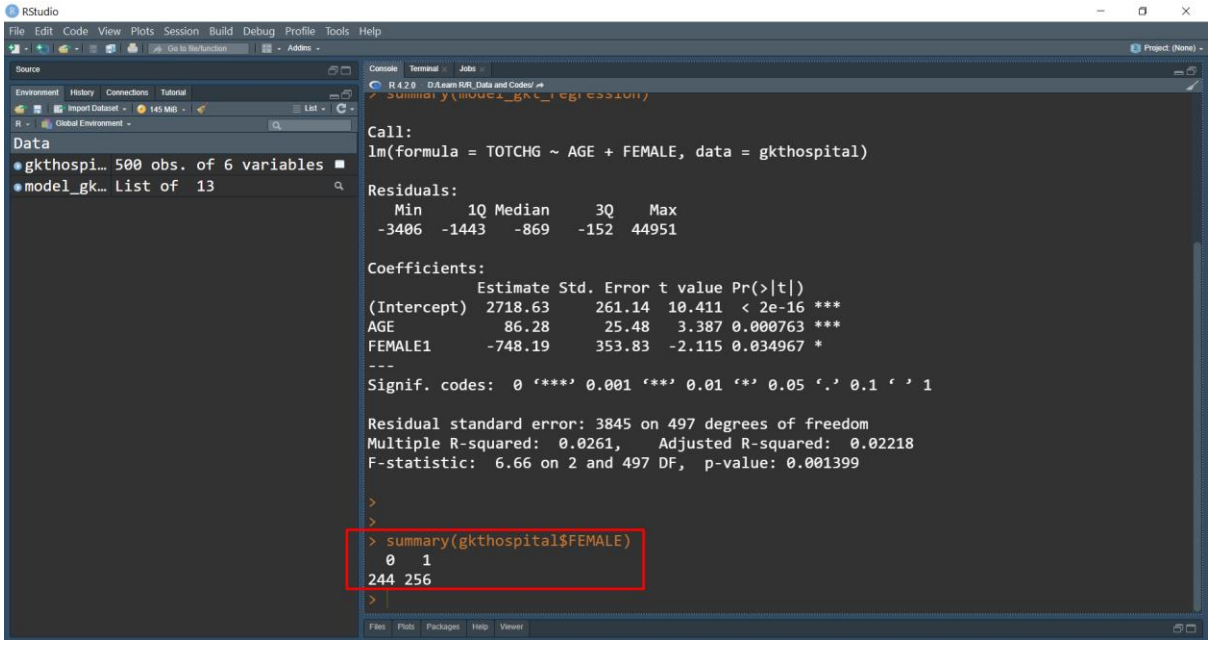


Figure 11: Result of FEMALE Summary

(Please Turn Over)

Conclusion: From the above results, we can obtain the following conclusions:

1. If we see the results carefully, on analysing P-values and other significant parameter levels, they all indicate that AGE has more impact on severity of hospital costs than gender.

2. There are almost equal number of Females and Males.

3. On an average (based on -ve coefficient values), Females incur lesser hospital costs than males.

Hence Goal #4 Accomplished ✔

Goal 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.

For this goal we will be using the 'LOS', 'AGE', 'FEMALE' and 'RACE' attributes given to us in the dataset. To show whether length of stay is dependent on age, gender or race we will be using Linear Regression again. This time we will keep 'LOS' as the dependent variable and keep 'AGE', 'FEMALE' and 'RACE' as independent variables.

We can simply do that by:
gkthospital$RACE<-as.factor(gkthospital$RACE)
model_gkt_regression2<-lm(LOS~AGE+FEMALE+RACE,data = ***gkthospital)*** ***{To call the regression function}***
summary(model_gkt_regression2)

(Please Turn Over)

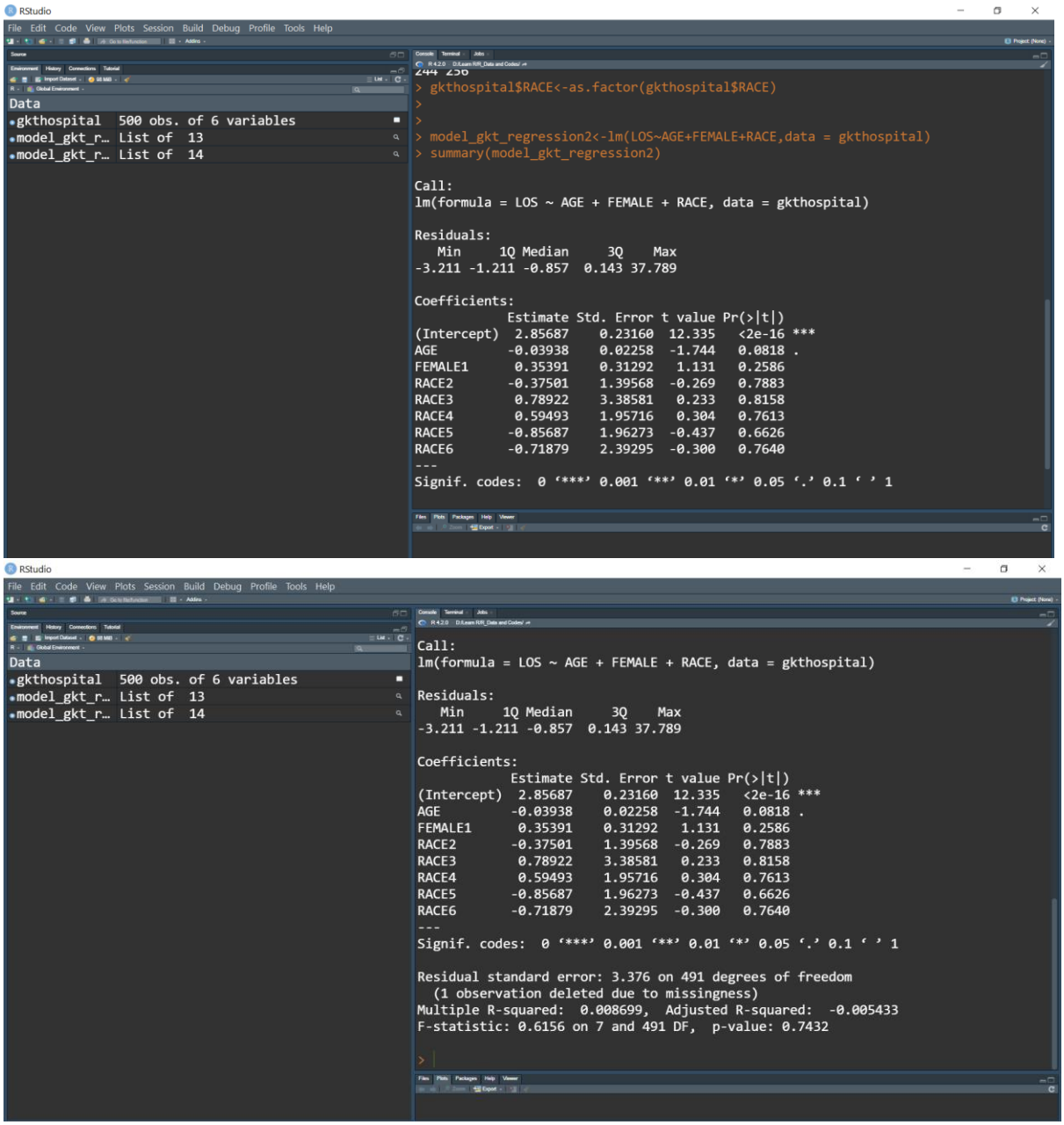


Figure 12: Result of the Analysis

Conclusion: From the above results, we can obtain the following conclusions:

1. If we see the results carefully, the P-values for all independent variables are quite high. This indicates that there is no linear relationship between the given variables.
2. The final conclusion is the fact that “We can’t predict length of stay of a patient based on age, gender and race”.

Hence Goal #5 Accomplished ✔

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

This goal is quite easy. We only need to determine which variable effects the “TOTCHG” or the Hospital costs. All this can be easily shown using Linear Regression itself. This time we will be using the ‘TOTCHG’ as dependant variable and all the other ones as independent variables.

We can simply do that by:
model_gktlr<-lm(TOTCHG~AGE+FEMALE+RACE+LOS+APRDRG,data = gkthospital)
summary(model_gktlr)

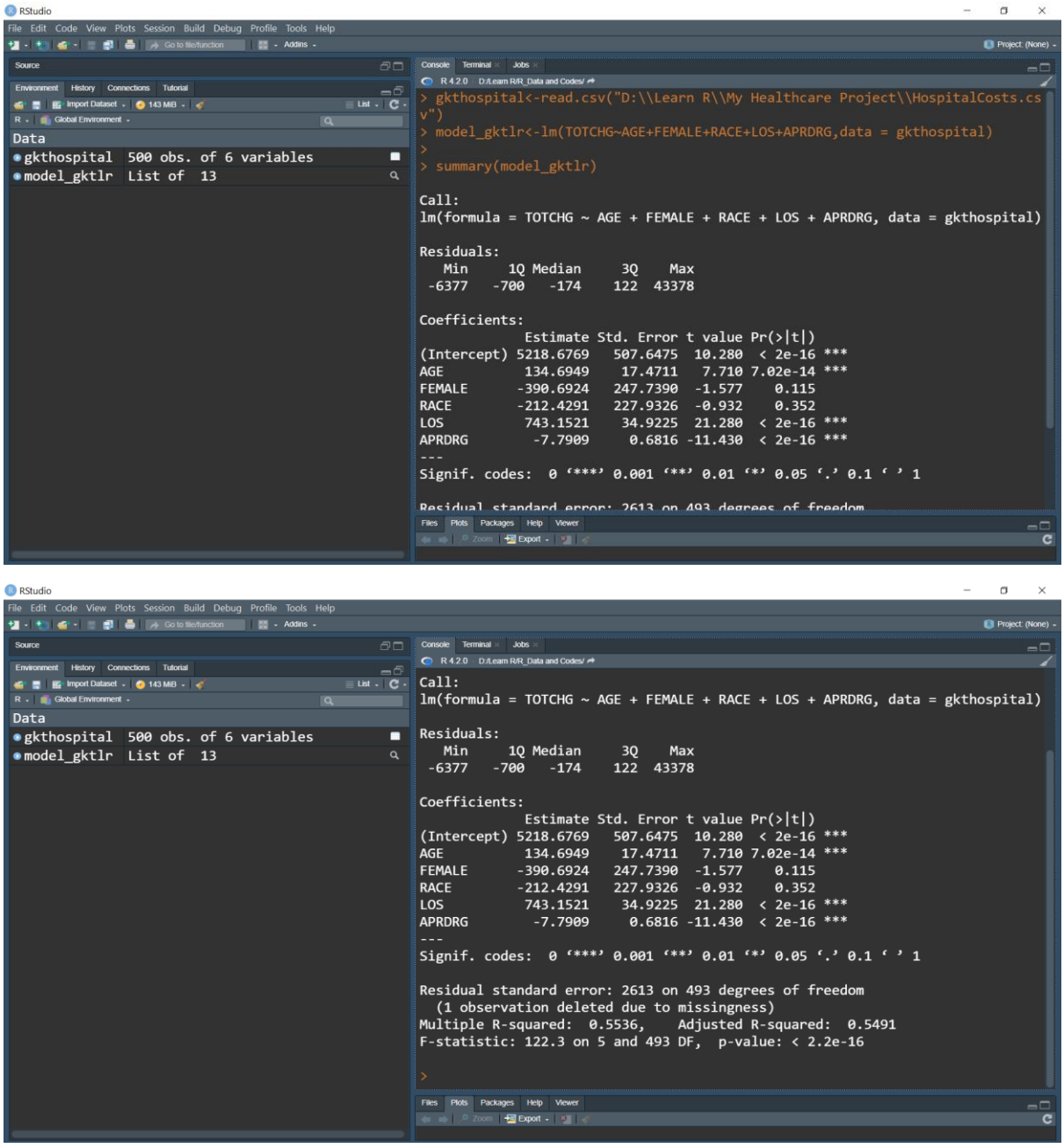


Figure 13: Result

Conclusion: From the above results, we can obtain the following conclusions:

1. AGE and LOS (Length of stay) highly affect the total hospital costs.

2. In addition to that there is positive relationship between LOS (Length of stay) to the TOTCHG (Hospital costs). We can say that if there is an increase of 1 day in LOS, then there is an addition of a value of 742 to the Hospital costs.

Hence all Goals 6/6 are now Accomplished ✔

[End of Project Report]