



LOVELY
PROFESSIONAL
UNIVERSITY

**SIX WEEKS SUMMER TRAINING
REPORT**

on

DATA SCIENCE WITH R PROGRAMMING

Submitted by

Gibran Khan Tareen

Registration Number: 12100173

Program Name: BTech CSE

**School of Computer Science & Engineering,
Lovely Professional University, Phagwara**

(June-July 2022)

Declaration

I hereby declare that I have completed my six weeks summer training course named **DATA SCIENCE WITH R PROGRAMMING** from **SIMPLILEARN platform**. I declare that I have worked with full dedication during these six weeks of training and my learning outcomes fulfil the requirements of training for the award of degree of **Bachelor of Technology (Computer Science & Engineering)**, Lovely Professional University, Phagwara.



(Signature of Student)

Gibran Khan Tareen

Registration Number: 12100173

Date: 12 July 2022

Acknowledgement

I had a great experience learning in this six-week summer training and even got to learn plenty of new skills. First, I would like to thank Almighty Allah who made it possible for me to do anything. After that I would like to thank my Late Grandfather- Mohd Nayeem Khan, my Grandma, my Parents, my little sister, and my maternal Grandparents because without their kind support and help I would not have been able to complete my project. I would also like to express my sole gratitude to my university who gave me this golden opportunity to do this training. The project making during the training, helped me to learn how to do proper Research and I learned about many new things while doing my training.



(Signature of Student)

Gibran Khan Tareen

Registration Number: 12100173

Date: 12 July 2022

Summer Training Certificate



Table of Contents

S No.	Title	Page No.
1	Cover Page	Page 1
2	Declaration	Page 2
3	Acknowledgement	Page 3
4	Summer Training Certificate	Page 4
5	List of Content	Page 5
6	Introduction	Page 6
7	Profile of the Problem & Problem Analysis	Page 7
8	Existing System	Page 9
10	Software Requirement Analysis	Page 9
11	Design	Page 10
12	Implementation	Page 11
13	Gantt Chart (Timeline)	Page 25
14	Project Legacy	Page 26
15	Bibliography	Page 27

Introduction

1.1 Context

This project has been done as part of my Six Week Summer Training for B.Tech CSE at Lovely Professional University. I had Six weeks to fulfil the requirements in order to complete the training.

1.2 Motivations

Since very start, I have been extremely interested in everything which is related to Data Science. Extracting information from huge junks of data, this has been one of my favourites. This summer training was a great occasion to give me the time to learn and confirm my interest for this field. The fact that I can use DS in the field of Healthcare Sector for Cost analysis by using data science concepts helped me more to enhance my interest in the field of Data Science. That's why I decided to do summer training around Data Science.

1.3 Idea

Back in Kashmir, few relatives of mine run a hospital named Khanam's Hospital. Many a times when I went to their home, I often found them debating on hospital costs, which age groups has most patients and many other things. These parameters helped them to efficiently track the revenue of the hospital and also to analyse which wards (infants or adults) needs to be increased in numbers when they analyse the number of visits of the age groups of patients so that there is no deficit in number of wards available to patients. This was the moment at which the idea struck my mind that I should program such a Realtime system that completely analyses the data of the hospital with help of using the concepts of Data Science. The aim of my project was to develop such a system that helps all hospitals to analyse their data and costs so that they could easily understand their data and then do the necessary actions (like increase the number of beds for infant patients if frequency number of infant patients is high etc). So thus I chose to take this Healthcare Cost Analysis as my project.

Profile of the Problem & Problem Analysis

Topic: Healthcare Cost Analysis (Given in the Simplilearn course itself)

Description:

1. 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.

2. Domain: Cost Analysis in Healthcare

3. Dataset Description:

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

Table 1. Dataset Description

Dataset for my Project (given by Simplilearn) :- [HospitalCosts.csv](#)

3. Goals to accomplish:

- a) 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.
- b) 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.
- c) 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.
- d) 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.
- e) 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.
- f) To perform a complete analysis, the agency wants to find the variable that mainly affects hospital costs.

Existing System

This project is developed for my Summer Training Course: Data Science with R Programming. Upon finding I found that there are some pre-existing softwares for analysing costs of Business Environment etc (like CostPerform, JMP etc) but I didn't find any system specifically for this Healthcare Cost analysis problem of ours. So therefore I made this project using my own learning and understanding of the Data Analysis which I gained during my 6-week summer training course.

Software Requirement Analysis

This project is developed for my Summer Training Course: Data Science with R Programming. We have some requirements for it to run.

1. Hardware Requirements

- 4+ CPUs (Recommended)
- 50+ GB of disk storage (Recommended)
- At least 256 MB of RAM, a mouse, and enough disk space for recovered files, image files, etc.
- An Intel-compatible platform running Windows 11, 10 /8.1/8 /7 /Vista /XP /2000 Windows Server 2022, 2019 /2016 /2012 /2008 /2003

2. Software Requirements

- R Language
- R Studio

Design (Flowchart)

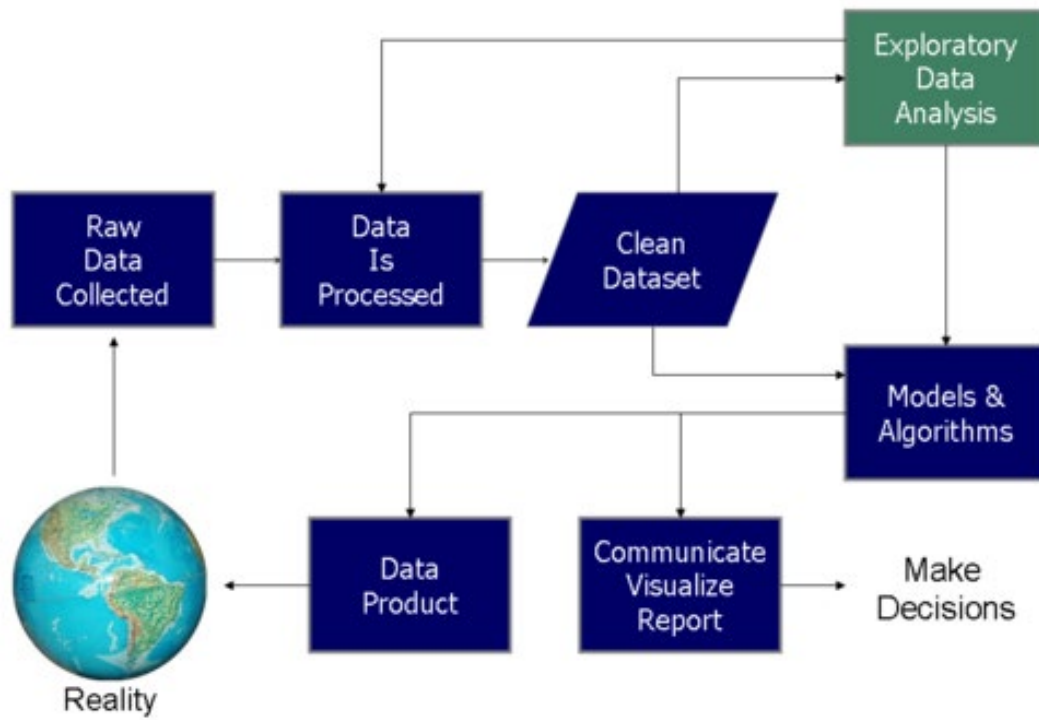


Figure 1: Flowchart of Data Analysis Project

Implementation

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 by using:

```
gkthospital<-read.csv("Location of file\\HospitalCosts.csv")
```

(gkt stands for Gibran Khan Tareen)

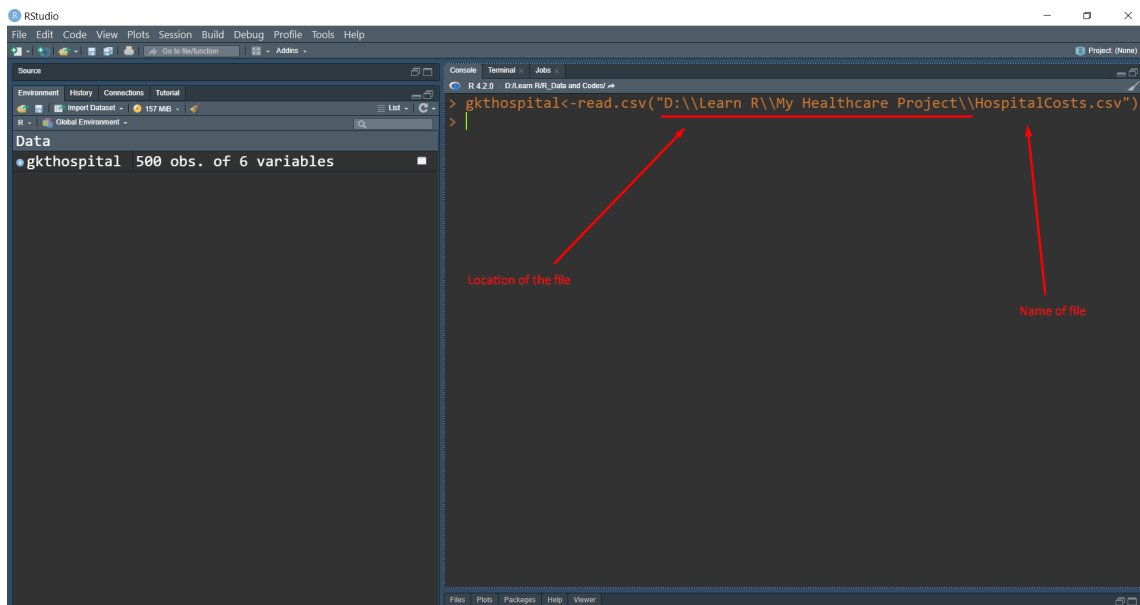


Figure 2: Reading Dataset

(Please Turn Over)

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

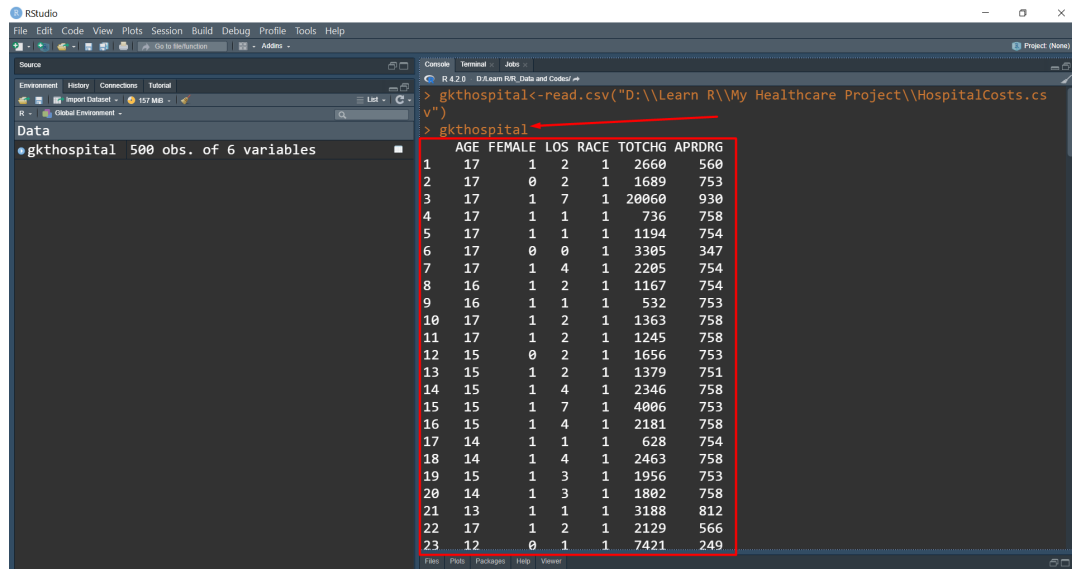


Figure 3: Successfully Imported

4.1.1. Goal 1 (given in question):

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 asked to find **2 things** in this Goal-

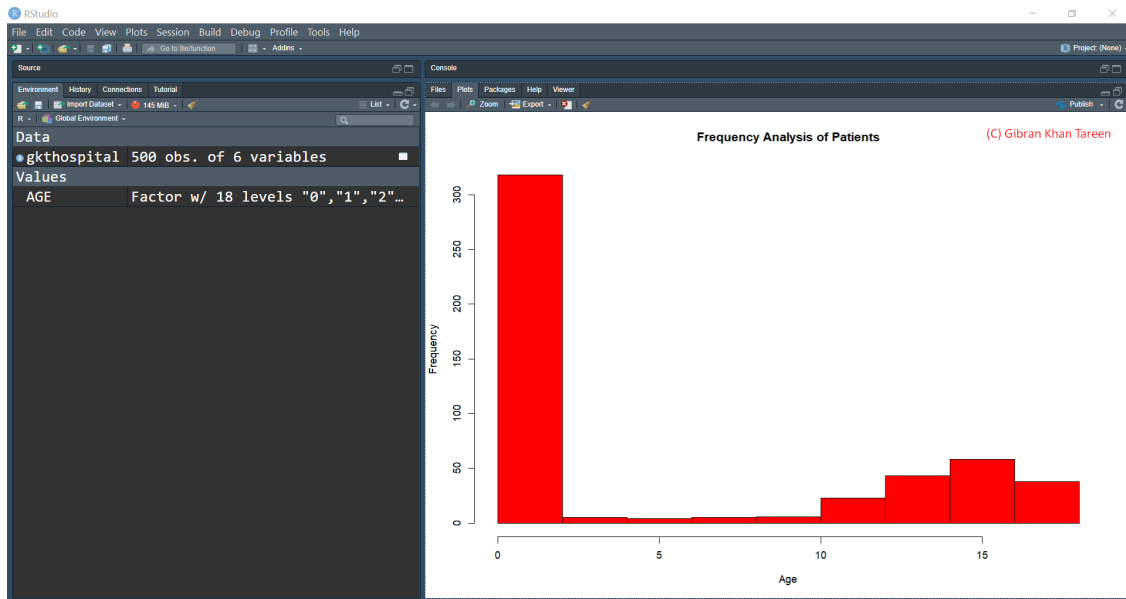
- I. The age category of people who frequently visit the hospital
- II. The age category that has maximum expenditure

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")



Frequency Analysis of Patients

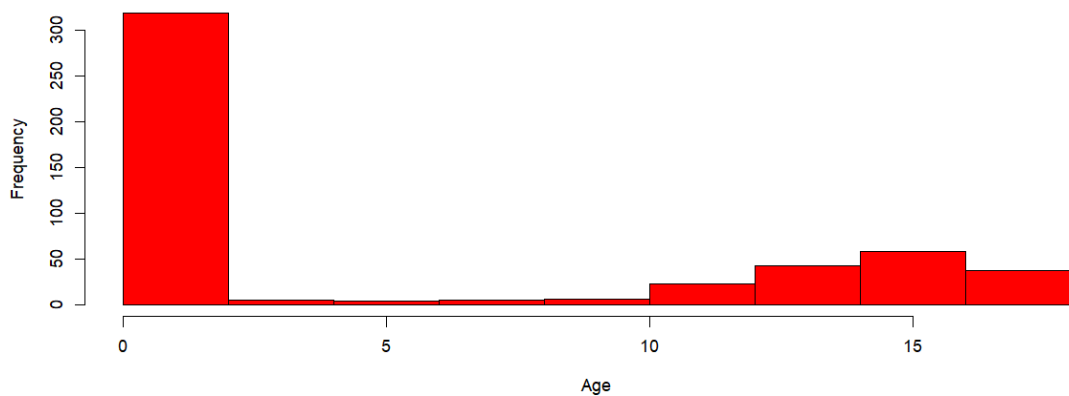


Figure 4: Frequency Analysis of Patients

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)
```

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

```
summary(AGE)
```

[attach is used to access the variables present in the data framework without calling the data frame]

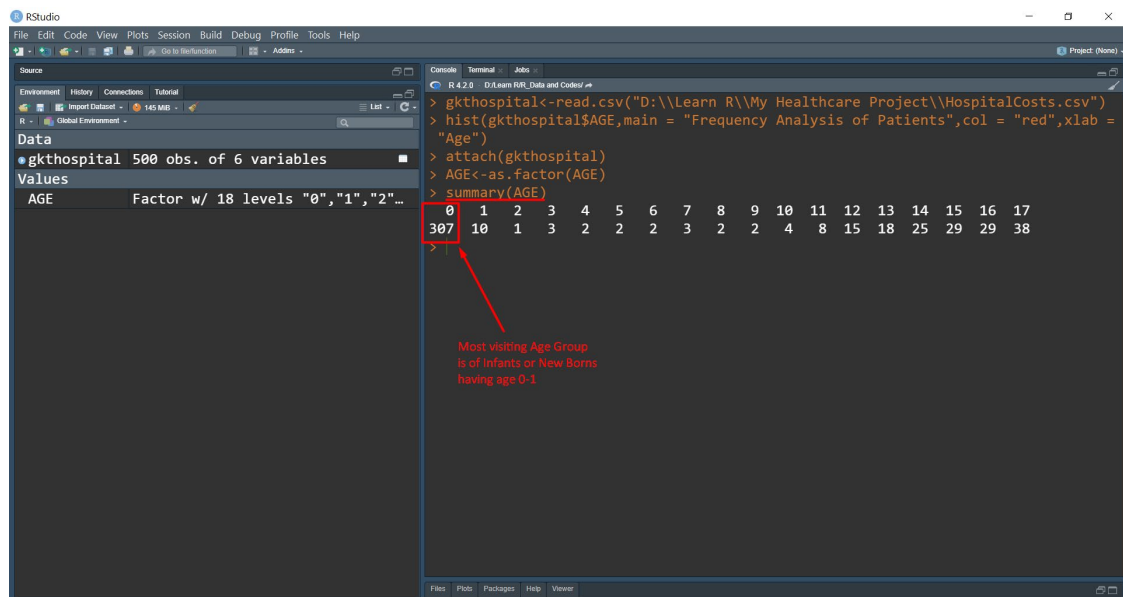


Figure 5: Summarized Analysis

4.1.2. 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}

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}

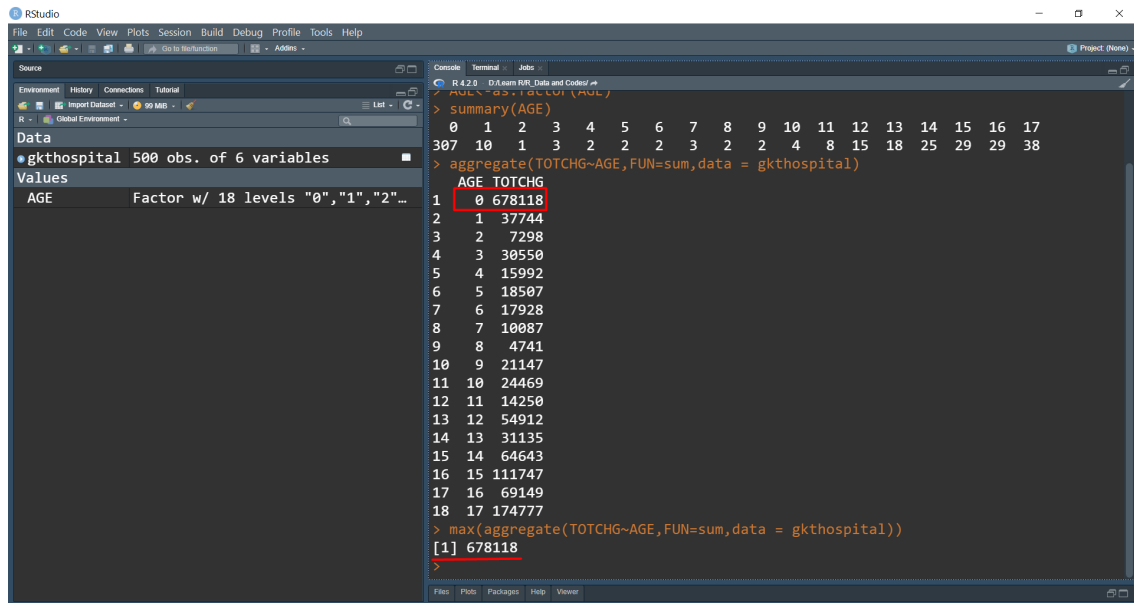


Figure 6: Subgoal 2

4.2.1. Goal 2 (given in question):

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}

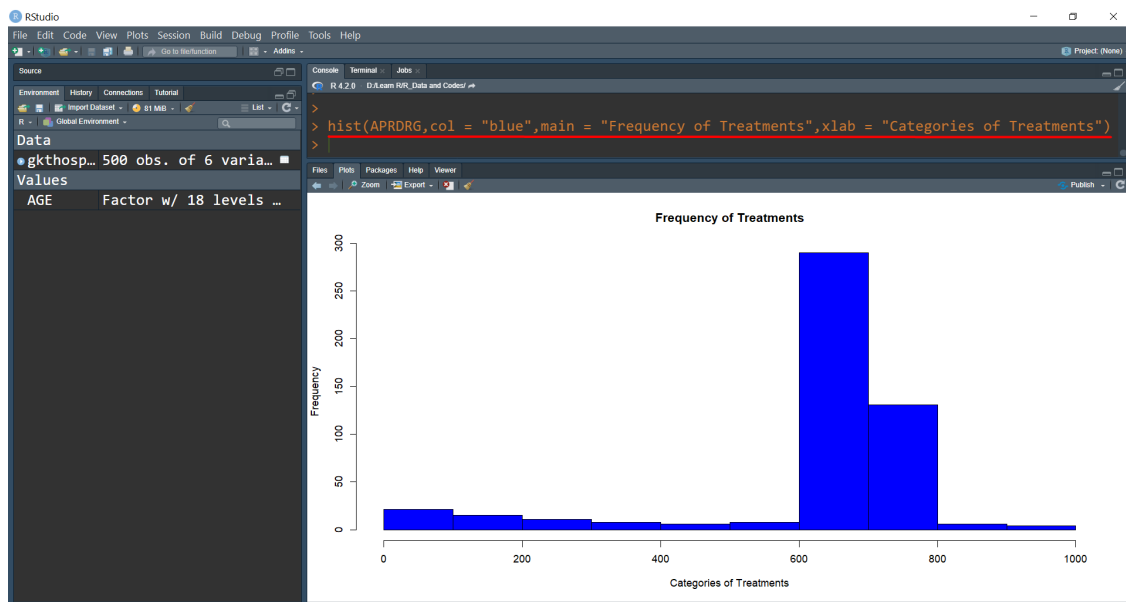


Figure 7: Histogram

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_ensure))

{to get max from Summary analysis}

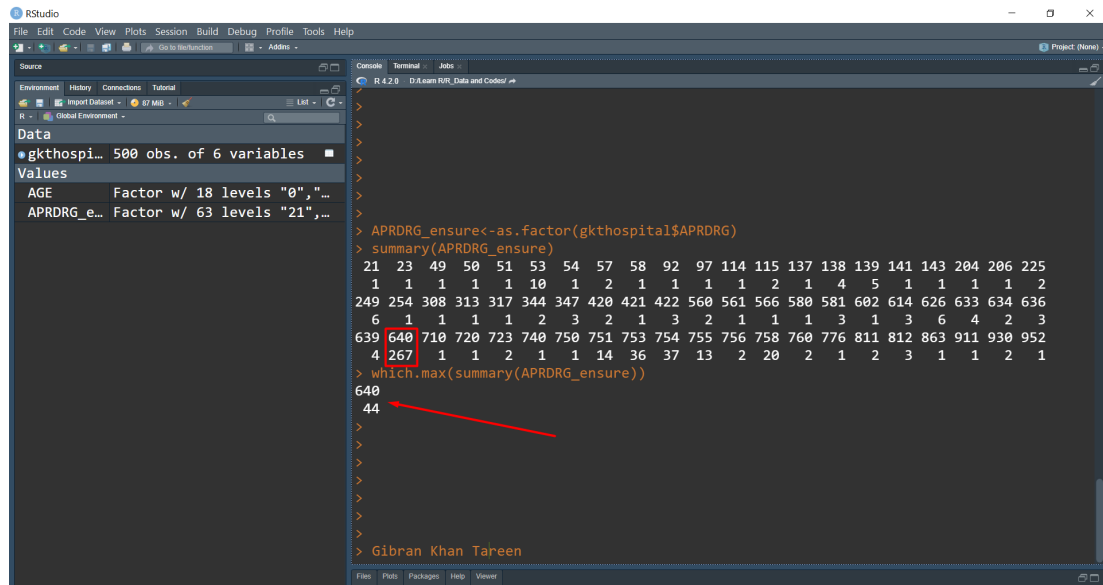


Figure 8: Result of Summary

4.2.2. 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)

```

> gkt<-aggregate(TOTCHG~APRDRG,FUN = sum,data=gkthospital)
> gkt
  APRDRG TOTCHG
1      21  10002
2      23  14174
3      49  20195
4      50   3908
5      51   3023
6      53  82271
7      54    851
8      57 145009
9      58   2117
10     92  12024
11     97   9530
12    114  10562
13    115  25832
14    137  15129
15    138  13622
16    139  17766
17    141   2860
18    143   1393
19    204   8439
20    206   9230
21    225  25649
22    249 166642
23    254    615

```

```

20    206   9230
21    225  25649
22    249 166642
23    254    615
24    308  10585
25    313   8159
26    317  17524
27    344  14802
28    347  12597
29    420   6357
30    421  26356
31    422   5177
32    560   4877
33    561   2296
34    566   2129
35    580   2825
36    581   7453
37    602  29188
38    614  27531
39    626  23289
40    633  17591
41    634   9952
42    636  23224
43    639  12612
44    640 437978
45    710   8223
46    720 14243

```

```

47    723   5289
48    740  11125
49    750   1753
50    751  21666
51    753  79542
52    754  59150
53    755  11168
54    756   1494
55    758  34953
56    760   8273
57    776   1193
58    811   3838
59    812   9524
60    863  13040
61    911  48388
62    930  26654
63    952   4833
> gkt[which.max(gkt$TOTCHG),]
  APRDRG TOTCHG
44    640 437978

```

Figure 9: Result of the Data Analysis

4.3. Goal 3 (given in question):

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}
```

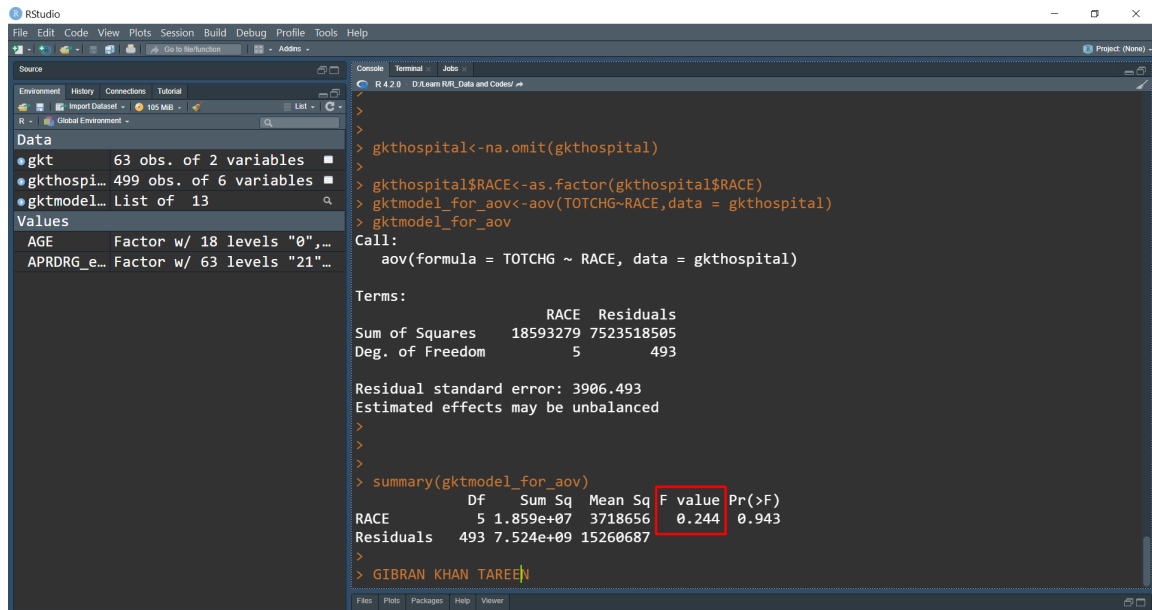


Figure 10: Results of the Summary

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)

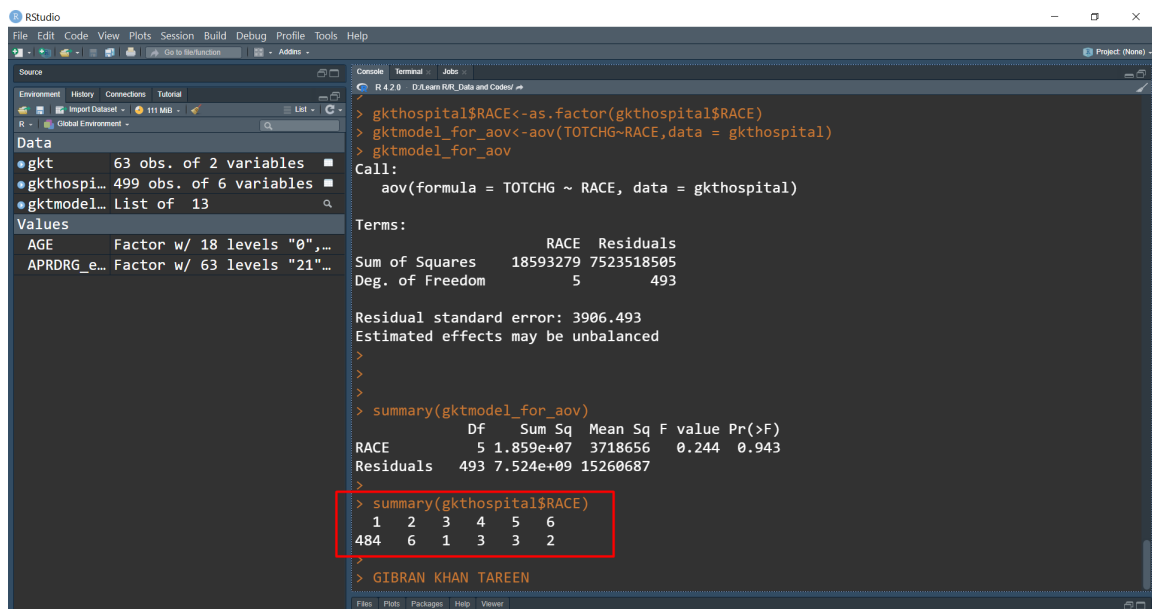


Figure 11: Result of ‘RACE’ Column Summary

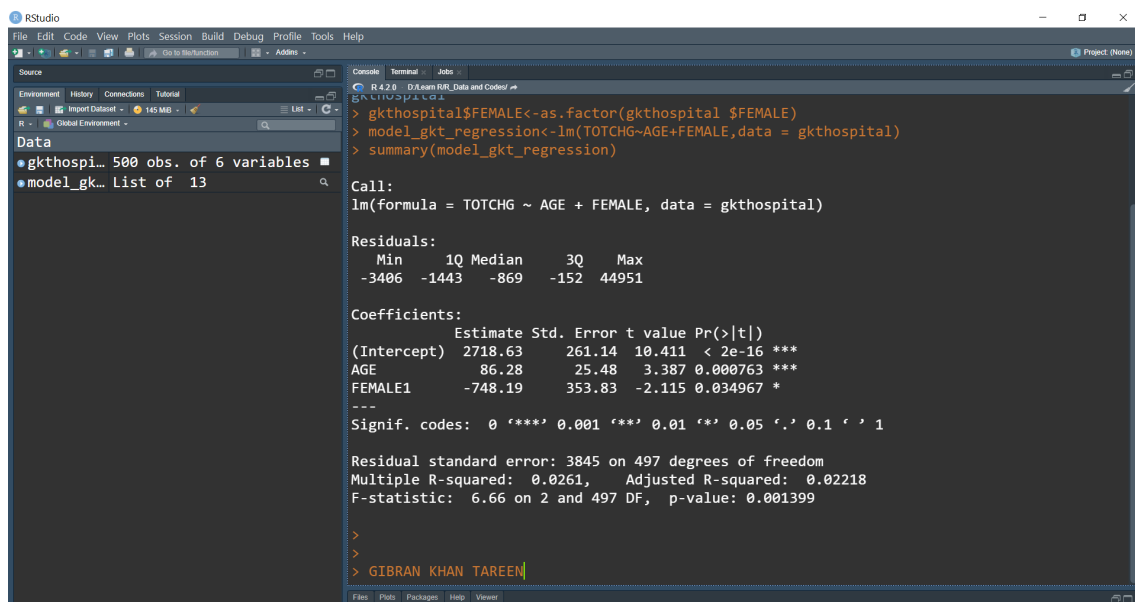
4.4. Goal 4 (given in question):

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 in order 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)
```



```
RStudio  
File Edit Code View Plots Session Build Debug Profile Tools Help  
Source  
Environment History Connections Tutorial  
Data  
gkthospi... 500 obs. of 6 variables  
model_gk... List of 13  
Console  
R 4.2.0 D:\Learn R\R_Data and Codes\...  
> gkthospital$FEMALE<-as.factor(gkthospital $FEMALE)  
> model_gkt_regression<-lm(TOTCHG~AGE+FEMALE,data = gkthospital)  
> summary(model_gkt_regression)  
Call:  
lm(formula = TOTCHG ~ AGE + FEMALE, data = gkthospital)  
Residuals:  
    Min       1Q   Median       3Q      Max  
-3406  -1443   -869   -152  44951  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  2718.63    261.14   10.411 < 2e-16 ***  
AGE           86.28     25.48    3.387 0.000763 ***  
FEMALE1     -748.19    353.83   -2.115 0.034967 *  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
Residual standard error: 3845 on 497 degrees of freedom  
Multiple R-squared:  0.0261,    Adjusted R-squared:  0.02218  
F-statistic:  6.66 on 2 and 497 DF,  p-value: 0.001399  
>  
>  
> GIBRAN KHAN TAREEN
```

Figure 12 - Part 1 of the Analysis

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

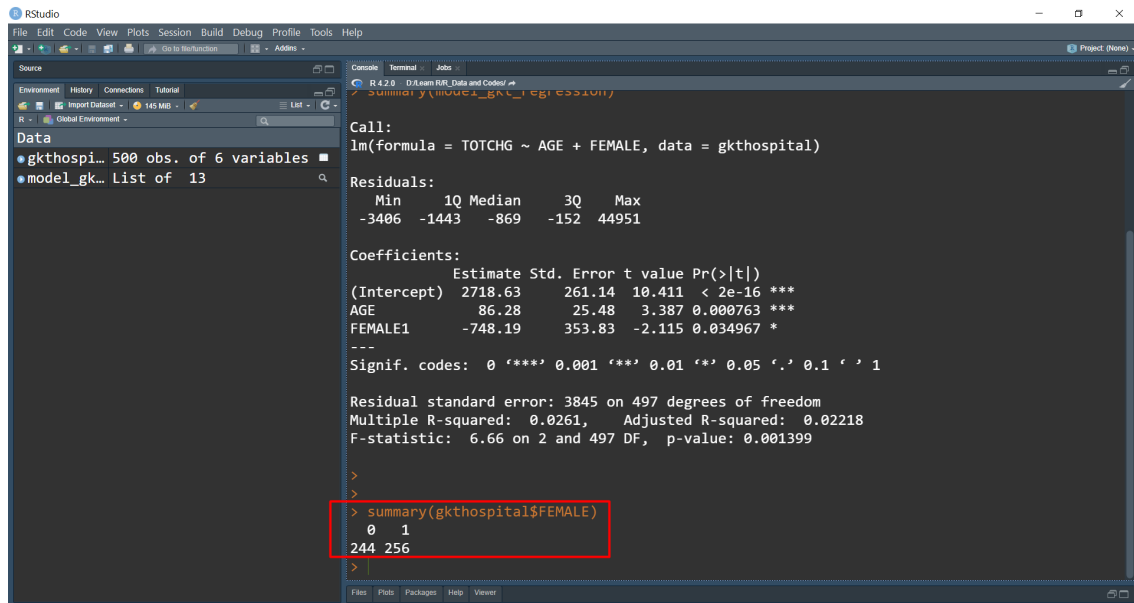


Figure 13: Result of FEMALE Summary

4.5. Goal 5 (given in question):

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)
```

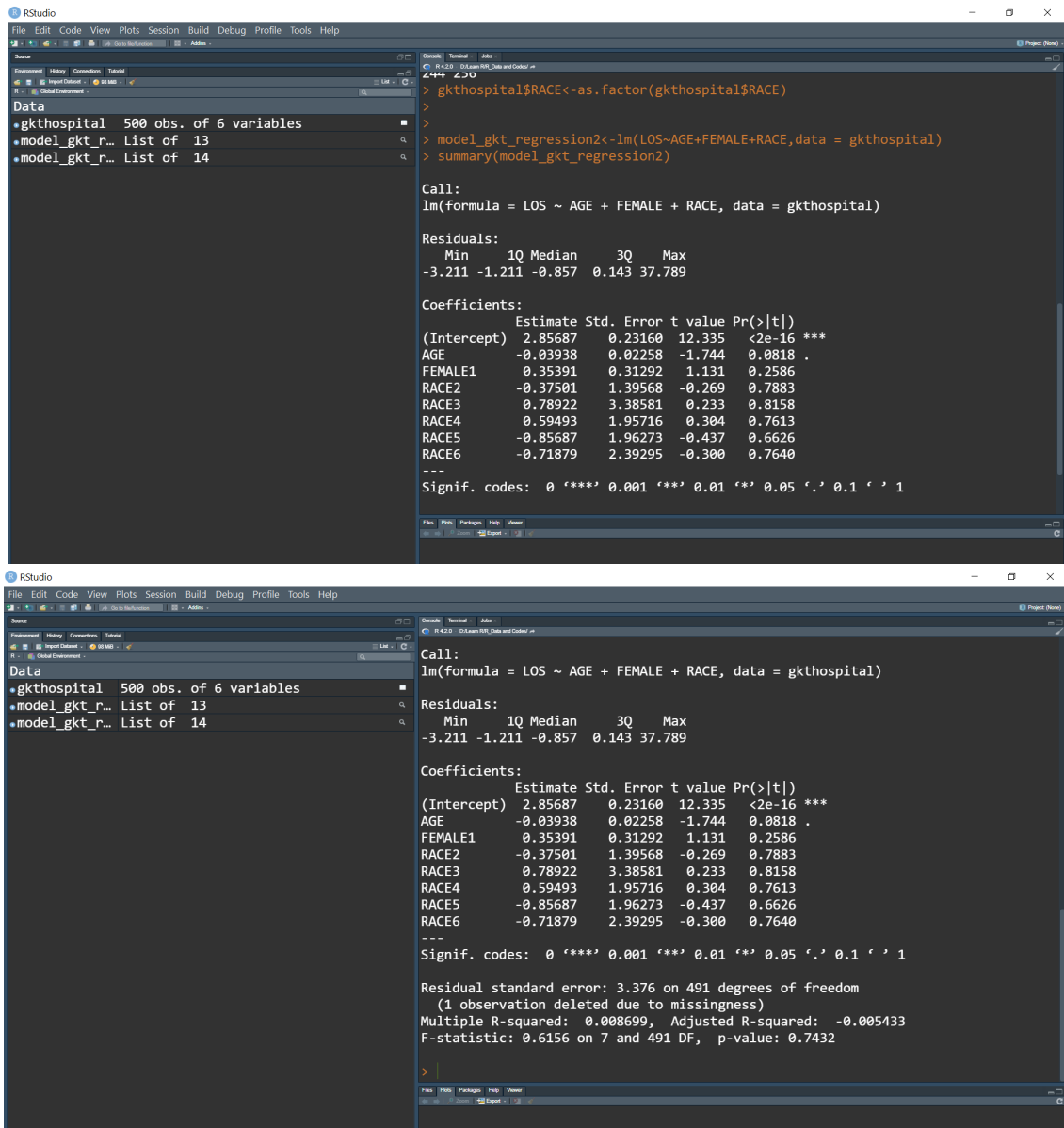


Figure 14: Result of the Analysis

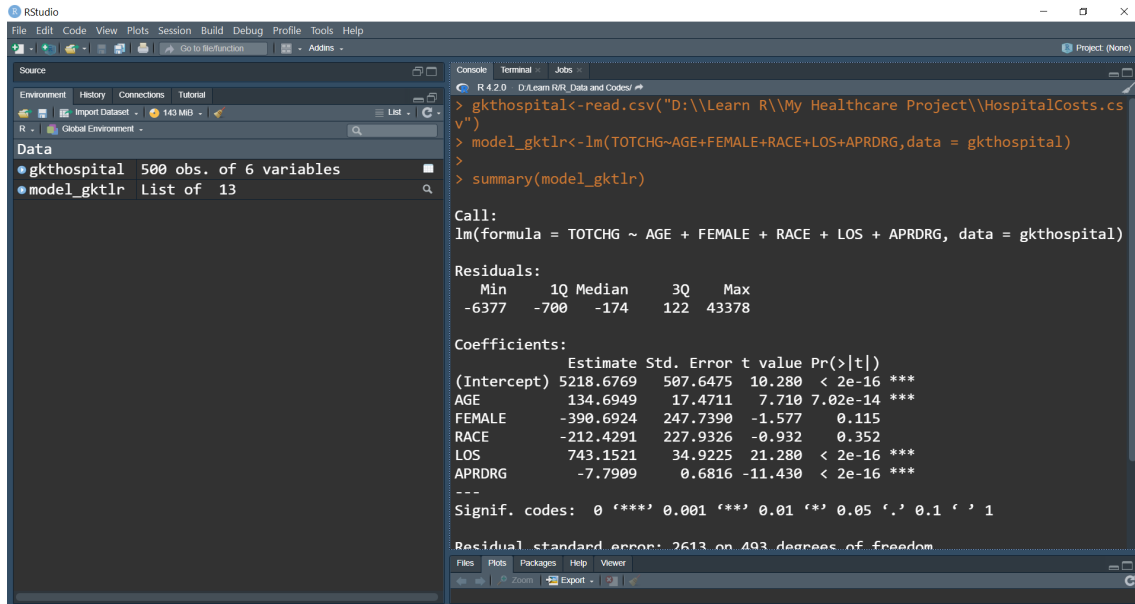
4.6. Goal 6 (given in question):

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)
```



The screenshot shows the RStudio interface. The Environment pane on the left lists 'gkthospital' (500 obs. of 6 variables) and 'model_gktlr' (List of 13). The Console pane on the right displays the following R code and its output:

```
> gkthospital<-read.csv("D:\\Learn R\\My Healthcare Project\\HospitalCosts.csv")  
> model_gktlr<-lm(TOTCHG~AGE+FEMALE+RACE+LOS+APRDRG,data = gkthospital)  
> summary(model_gktlr)
```

Call:
lm(formula = TOTCHG ~ AGE + FEMALE + RACE + LOS + APRDRG, data = gkthospital)

Residuals:

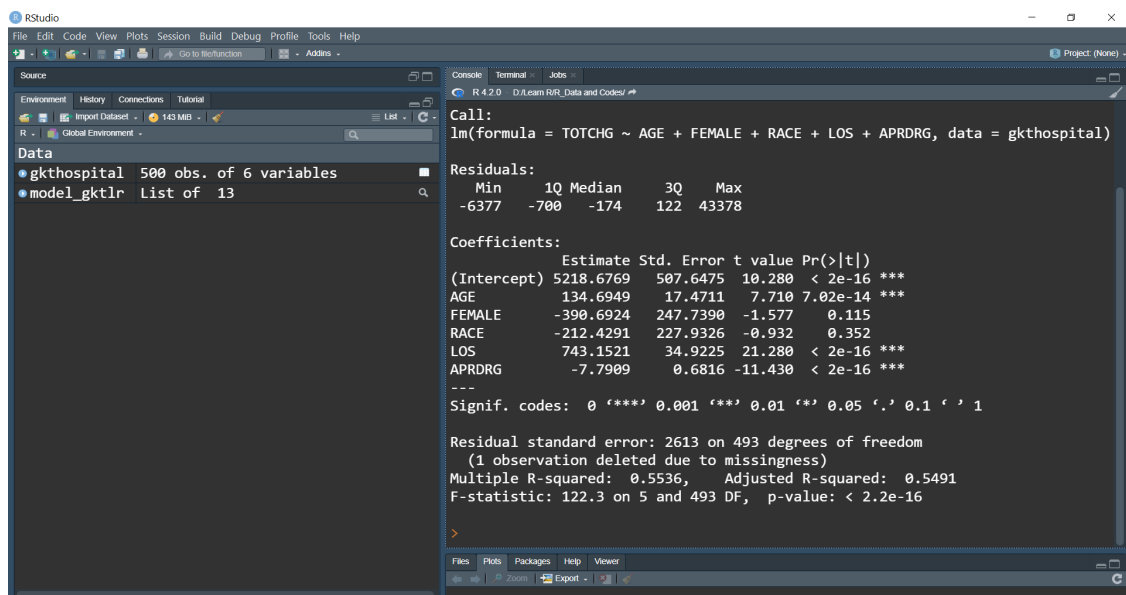
Min	1Q	Median	3Q	Max
-6377	-700	-174	122	43378

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5218.6769	507.6475	10.280	< 2e-16 ***
AGE	134.6949	17.4711	7.710	7.02e-14 ***
FEMALE	-390.6924	247.7390	-1.577	0.115
RACE	-212.4291	227.9326	-0.932	0.352
LOS	743.1521	34.9225	21.280	< 2e-16 ***
APRDRG	-7.7909	0.6816	-11.430	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2613 on 493 degrees of freedom



This screenshot shows the same RStudio interface as the previous one, but the Console pane displays additional summary statistics at the bottom of the output:

```
> summary(model_gktlr)
```

Call:
lm(formula = TOTCHG ~ AGE + FEMALE + RACE + LOS + APRDRG, data = gkthospital)

Residuals:

Min	1Q	Median	3Q	Max
-6377	-700	-174	122	43378

Coefficients:

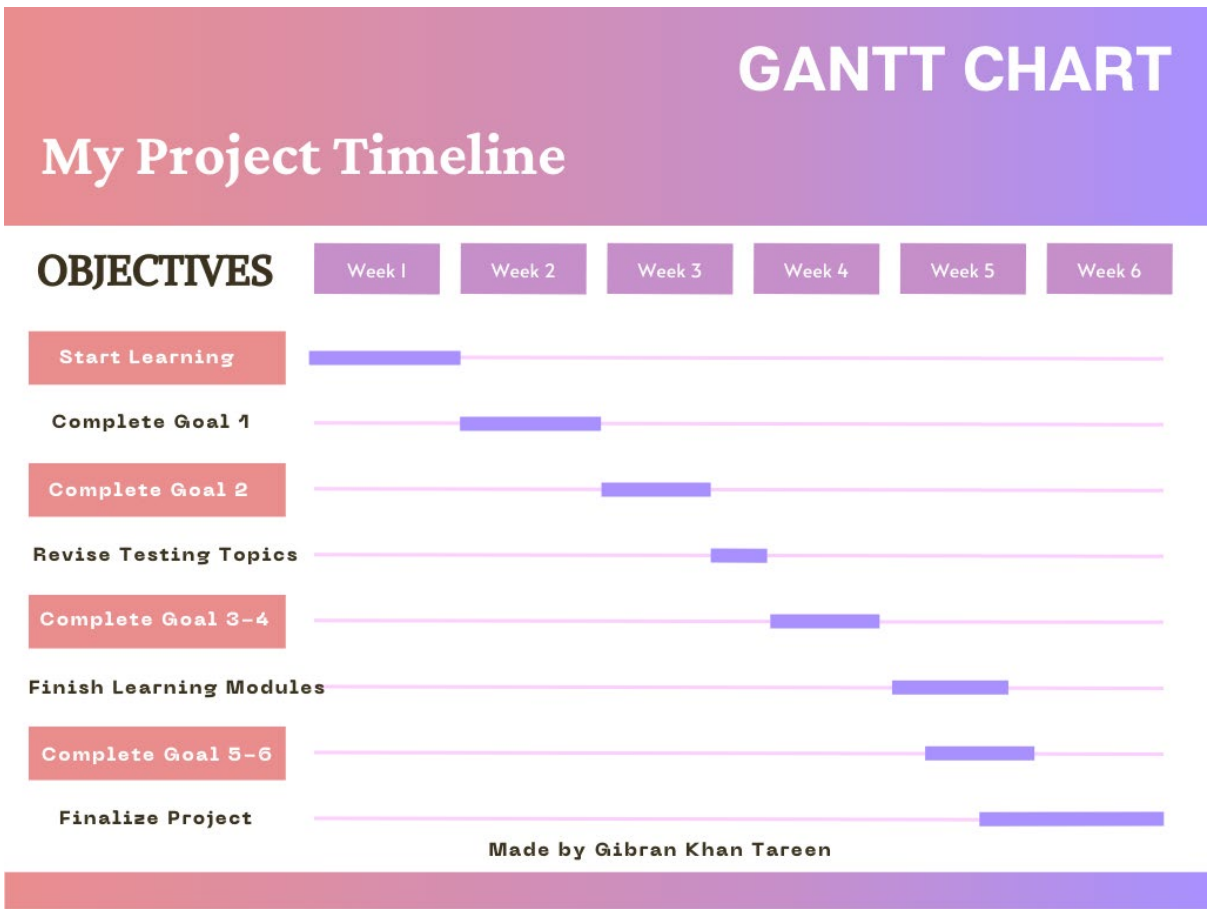
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5218.6769	507.6475	10.280	< 2e-16 ***
AGE	134.6949	17.4711	7.710	7.02e-14 ***
FEMALE	-390.6924	247.7390	-1.577	0.115
RACE	-212.4291	227.9326	-0.932	0.352
LOS	743.1521	34.9225	21.280	< 2e-16 ***
APRDRG	-7.7909	0.6816	-11.430	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2613 on 493 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared: 0.5536, Adjusted R-squared: 0.5491
F-statistic: 122.3 on 5 and 493 DF, p-value: < 2.2e-16

Figure 15: Result

Gantt Chart



Project Legacy

This project is developed for my Summer Training. I would love to have some future improvements in the project as well.

1. Increase the Scalability

I want to develop this project more, for handling even larger and more diverse datasets.

2. Make a GUI for the system

I would love to develop a GUI for my project where any user can simply import his/her dataset and with a few clicks he/she can get any insight from the huge chunk of data.

Bibliography

1. <https://www.geeksforgeeks.org/one-way-anova/>
2. <https://www.geeksforgeeks.org/anova-test-in-r-programming/>
3. Hadley Wickham and Garrett Gorlemund, R for Data Science (Latest Edition), **2017**, pp. 345–372.
4. <https://www.youtube.com/watch?v=aC009Px4tEg>
5. https://www.youtube.com/watch?v=Xg8_iSkJpAE
6. John A. Rice, Mathematical Statistics and Data Analysis (3rd Edition), **2007**, pp. 329–420.