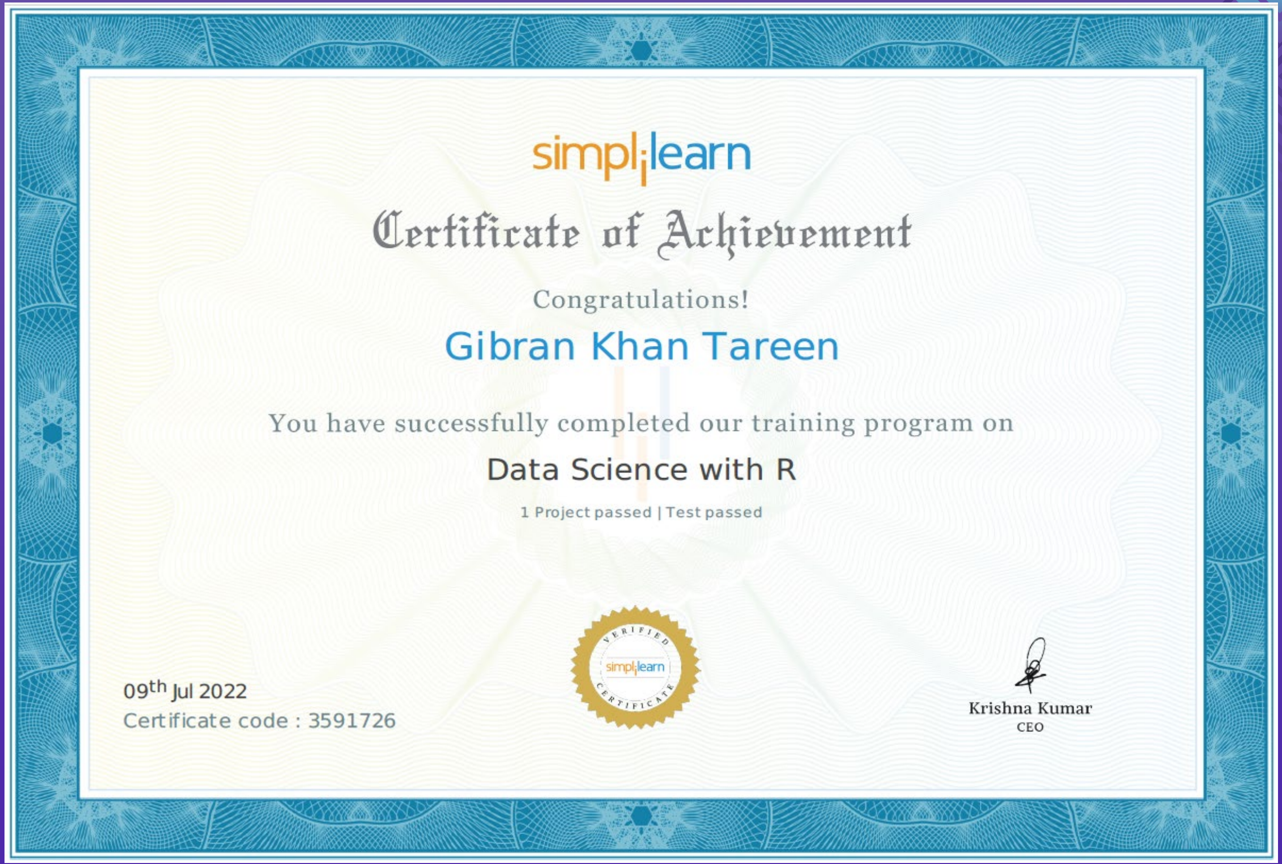
Data Science With R

Summer Training ETP VIVA

Made by Gibran Khan Tareen
Registration Number: 12100173

Certificate from The Institution



simplilearn

Certificate of Achievement

Congratulations!

Gibran Khan Tareen

You have successfully completed our training program on

Data Science with R

1 Project passed | Test passed

09th Jul 2022

Certificate code : 3591726

VERIFIED simplilearn CERTIFICATE

Krishna Kumar
CEO

Table of contents

01 A Brief Introduction

About the main objectives of our project and language used

02 Important Concepts

Discuss the important concepts used in the project

03 Few Glimpses

Of important outcomes which we found during the Project

04 Presenting Project Report

Finally I will show my project report, source code and execution

> “Torture the data,
> and it will **confess**
> to anything.”

—**Ronald Coase,** *Economics, Nobel prize Laureate*

01

A Brief Introduction

What are the Project Objectives and Language We Used In It

Basic Structure of Our Dataset

Attributes	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)
TOTCH	Hospital discharge costs
APRDRG	All Patient Refined Diagnosis Related Groups

Main Objectives of Our Project

Record Patient Stats

Find which age category most frequently visit the hospital and has the maximum expenditure

Check For Malpractice

Find if there is any malpractice going on by analyzing the race of the patient is related to the hospitalization costs.

Allocation of resources

To properly utilize the costs, the agency has to analyze the severity of the hospital costs by age and gender.

Find Main Cost Factors

To perform a complete analysis, the agency wants to find the variable(s) that mainly affect hospital costs.

Language Used for our Project: R

R is the most popular language in the world of Data Science. It is heavily used in analyzing data that is both structured and unstructured. This has made R, the standard language for performing statistical operations. R allows various features that set it apart from other Data Science languages.

02

Important Concepts

Now we will Discuss the important concepts used in the project

Important Concepts Used

Hypothesis Test

Its is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis.

Anova Testing

Analysis of variance is used to investigate relations between categorical variables and continuous variable

Linear Regression

It is a statistical approach for modeling the relationship between a dependent variable and a given set of independent variables.

Where did I Implement These Concepts?

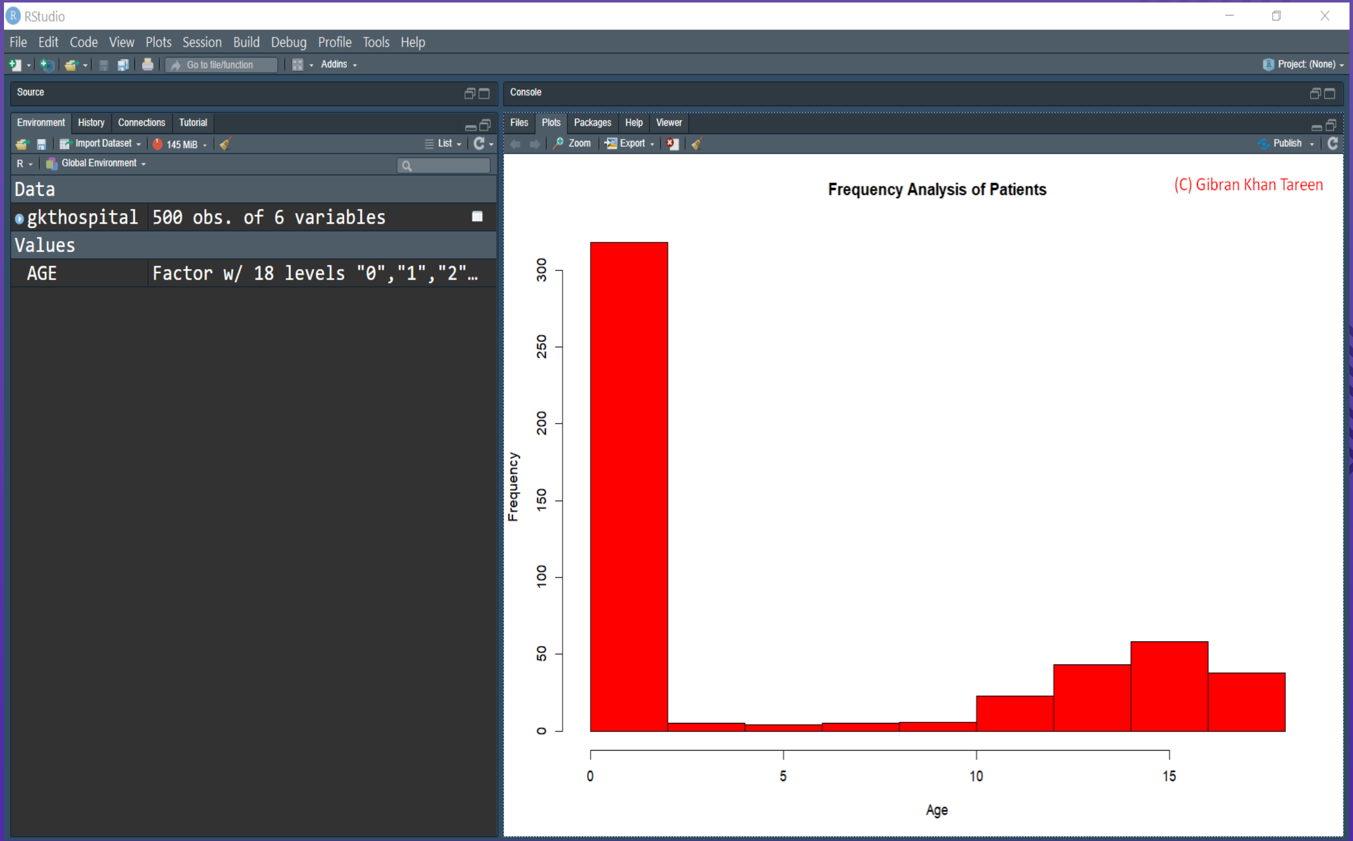
	Description	Implemented In
Hypothesis Test	An act in statistics whereby an ana tests an assumption regarding a population parameter. It provides evidence concerning the plausibili the hypothesis, given the data.	We used concept of Hypothesis testing using ANNOVA Testing our Goal 3
Annova Testing	It is a type of hypothesis testing fo population variance used to find th relations between categorical variables and continuous variable Programming	We implemented ANNOVA Tes in our Goal 3
Linear Regression	It is a commonly used type of predictive analysis. It is a statistica method that allows us to summari and study relationships between t continuous (quantitative) variables	We implemented Linear Regression by using our own Model in Goal 4, 5 and 6

03

Few Glimpses of Outcomes

We will now see some important outcomes which we found during the Project

Screenshot: Patient Statistics

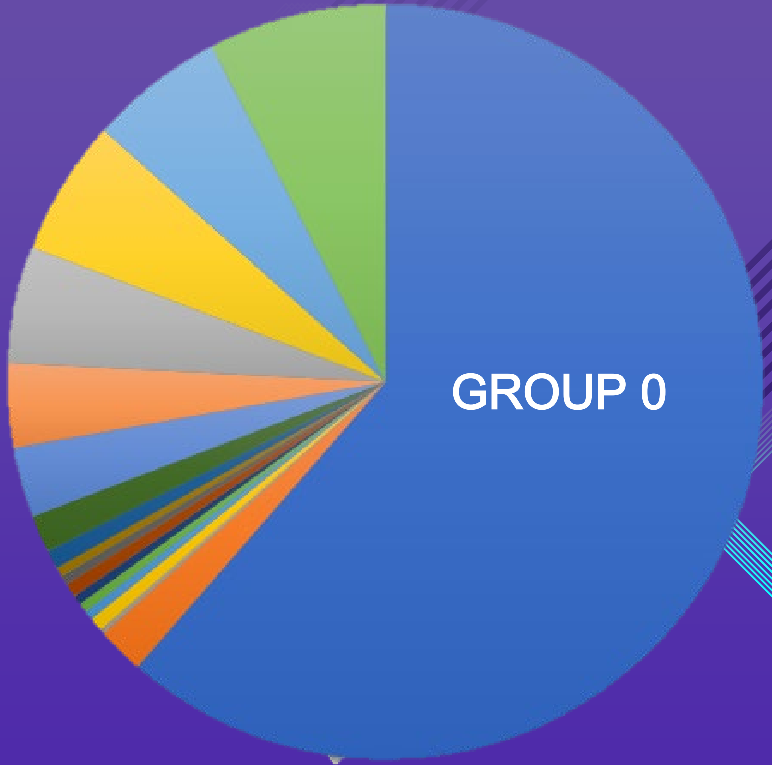


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
Console
Environment History Connections Tutorial
Import Dataset
145 MiB
List
R
Global Environment
Files Plots Packages Help Viewer
Zoom
Export
Publish

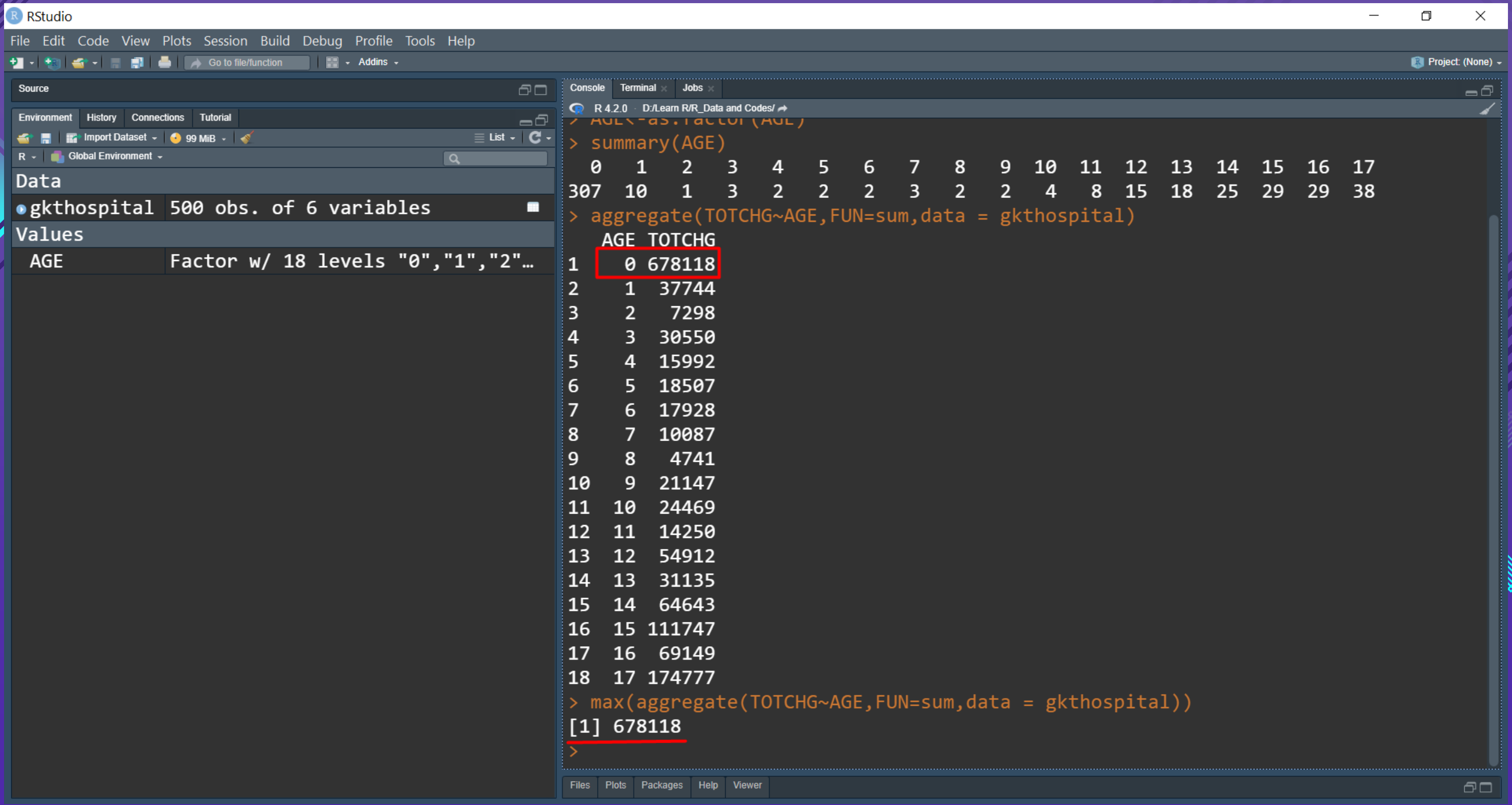
Data	
gkthospital	500 obs. of 6 variables
Values	
AGE	Factor w/ 18 levels "0","1","2"…

Patient Statistics Maximum Patients Top 3 Age Groups

Age GROUP 0 HasMaximumPatients	61%
Age GROUP 1	08%
Age GROUP 2	06%



Screenshot: Patient Statistics


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
Environment History Connections Tutorial
Import Dataset
99 MiB
List
R
Global Environment
Console Terminal Jobs
R 4.2.0 · D:/Learn R/R_Data and Codes/
Files Plots Packages Help Viewer

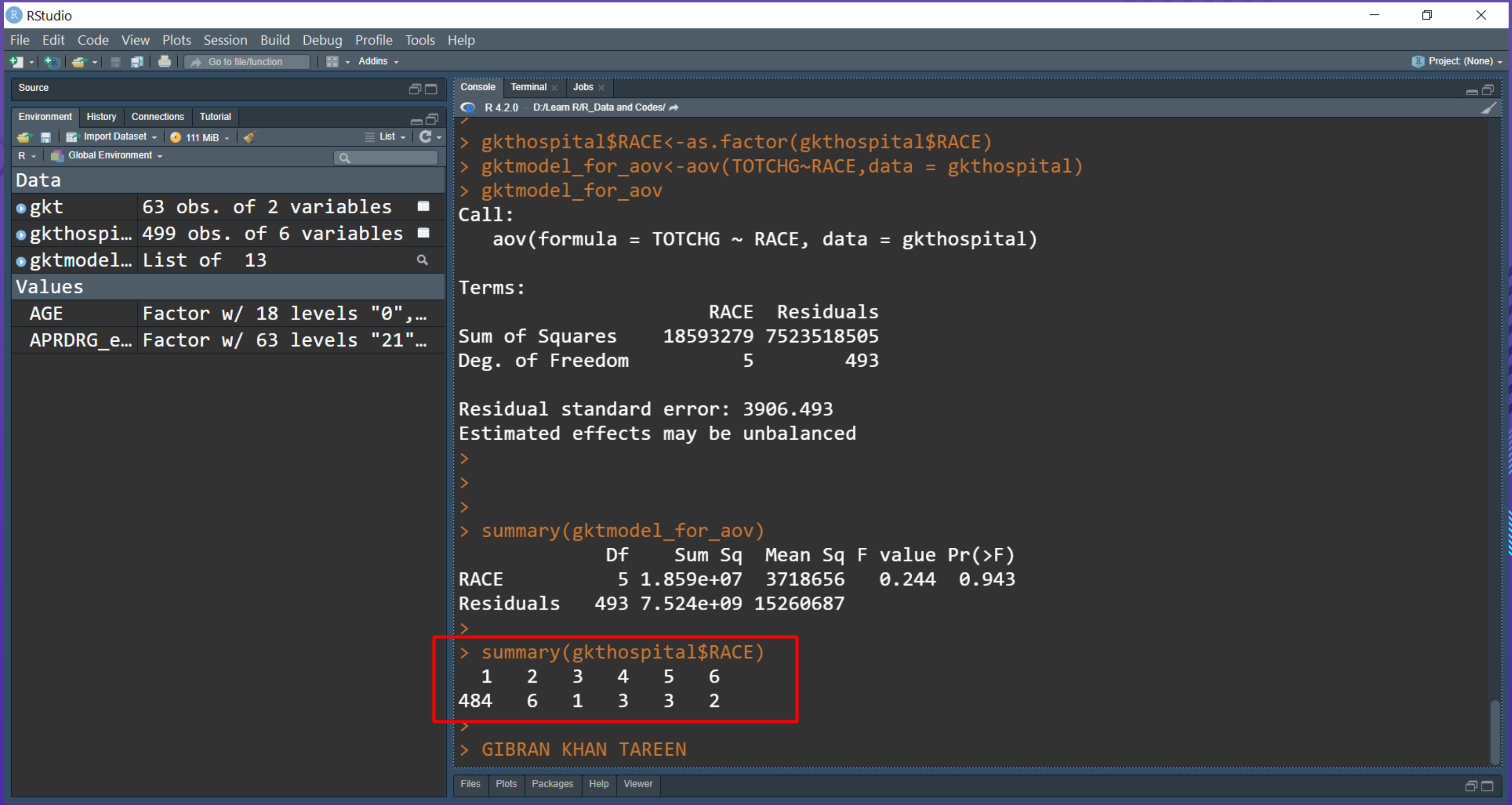
```
Data
gkthospital  500 obs. of 6 variables
Values
  AGE        Factor w/ 18 levels "0","1","2"…
```

```
> AGE<-as.factor(AGE)
> summary(AGE)
  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17 
307  10   1   3   2   2   2   3   2   2   4   8  15  18  25  29  29  38 
> aggregate(TOTCHG~AGE,FUN=sum,data = gkthospital)
   AGE TOTCHG
1    0 678118
2    1  37744
3    2   7298
4    3  30550
5    4  15992
6    5  18507
7    6  17928
8    7  10087
9    8   4741
10   9  21147
11  10  24469
12  11  14250
13  12  54912
14  13  31135
15  14  64643
16  15 111747
17  16  69149
18  17 174777
> max(aggregate(TOTCHG~AGE,FUN=sum,data = gkthospital))
[1] 678118
> 
```

₹678,118

AGE group 0 has the maximum Hospital expenditure

Screenshot of Check for Malpractice


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
Environment History Connections Tutorial
Import Dataset
111 MiB
List
R
Global Environment
Console Terminal Jobs
R 4.2.0 · D:/Learn R/R_Data and Codes/
Files Plots Packages Help Viewer

```
Data
gkt            63 obs. of 2 variables
gkthospi…      499 obs. of 6 variables
gktmodel…      List of  13
Values
AGE            Factor w/ 18 levels "0",…
APRDRG_e…      Factor w/ 63 levels "21"…
```

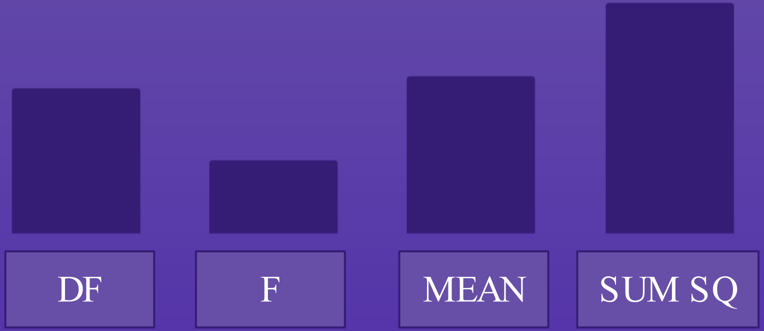
```
> gkthospital$RACE<-as.factor(gkthospital$RACE)
> gktmodel_for_aov<-aov(TOTCHG~RACE,data = gkthospital)
> gktmodel_for_aov
Call:
   aov(formula = TOTCHG ~ RACE, data = gkthospital)

Terms:
                      RACE  Residuals
Sum of Squares    18593279 7523518505
Deg. of Freedom          5        493

Residual standard error: 3906.493
Estimated effects may be unbalanced
>
>
>
> summary(gktmodel_for_aov)
             Df    Sum Sq  Mean Sq F value Pr(>F)
RACE          5 1.859e+07  3718656   0.244  0.943
Residuals   493 7.524e+09 15260687
>
> summary(gkthospital$RACE)
  1   2   3   4   5   6
484   6   1   3   3   2
>
> GIBRAN KHAN TAREEN
```

Check for Malpractice ANNOVA TESTING

Total No. of Races | 6 | Dependent Var. | TOTCH



Change impact

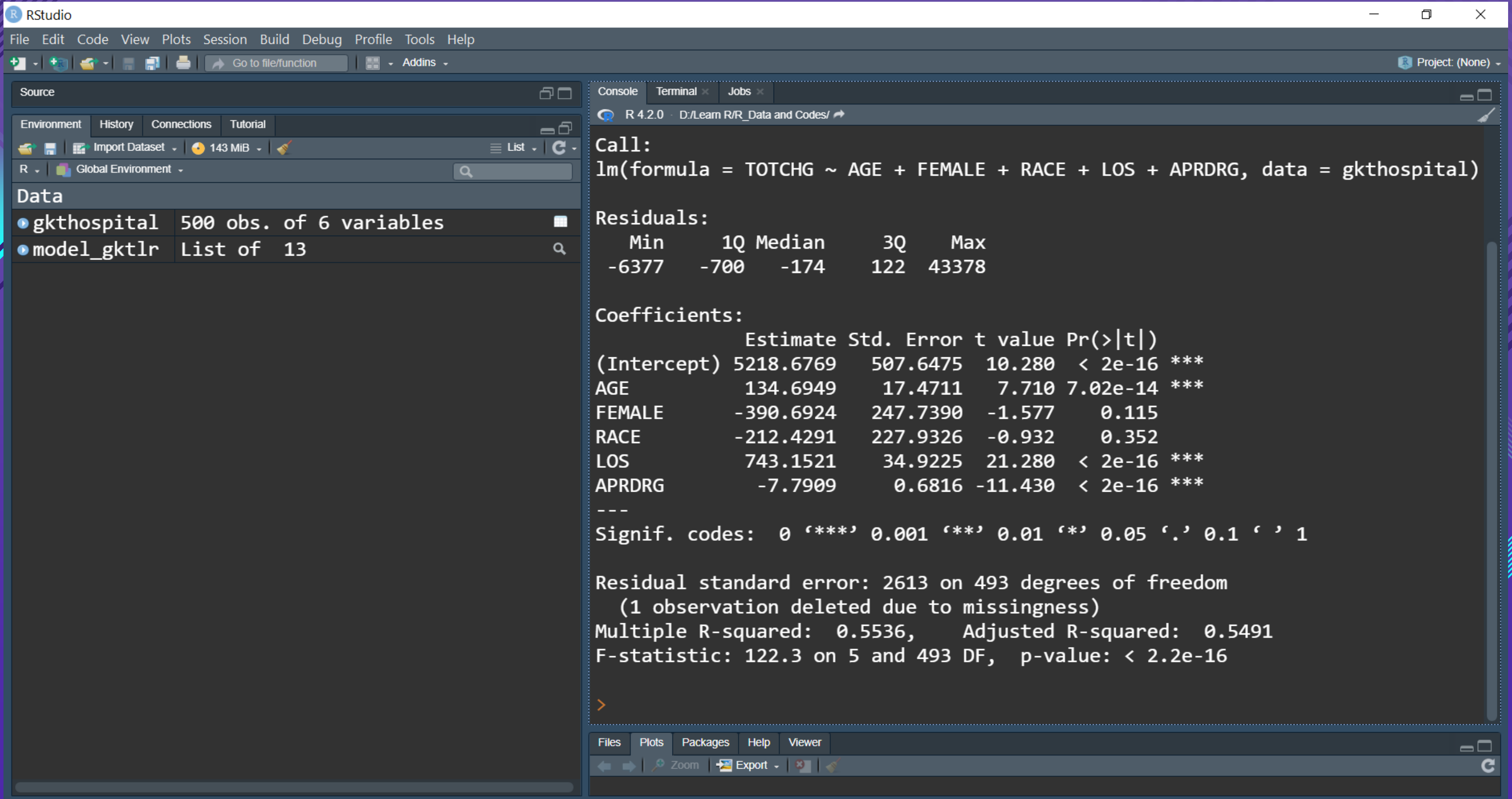
[✓ Low] Medium High

the “F value”is quite low (**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.

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** We conclude by saying “There is Insufficient data to verify if a patient’s race affects his expenditure.”

The p-value (labeled Pr >F) is greater than Significance value ie. Alpha (0.05) and the ‘Residual values”(deviation of the observed values) was quite high, so both of these Observations indicate that there is **no relationship** between race and hospital costs, thereby accepting the Null hypothesis.

Screenshot of Finding Main Cost Factor(s)


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
Environment History Connections Tutorial
Import Dataset
143 MiB
List
R
Global Environment
Console Terminal Jobs
R 4.2.0 · D:/Learn R/R_Data and Codes/
Files Plots Packages Help Viewer
Zoom
Export

```
Data
gkthospital   500 obs. of 6 variables
model_gktlr   List of  13
```

```
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

>
```

Finding Main Variables Which Are Affecting **Cost Factors**

Variable Factor		Impact			Level of Impact
1	AGE	◆✓	◇	◆	Very High
2	FEMALE	◆	◇	◆✓	Very Low
3	LOS (Length of Stay)	◆✓	◇	◆	Very High
4	RACE	◆	◇	◆✓	Very Low
5	APRDRG(All Patient Refined Diagnosis Related Groups)	◆	◇✓	◆	Medium

Thank You!

This Shall be it for the Presentation Part. I will now show the Project Report and Source Code

04

Presenting Project Report

Project report, Source code and execution