# Project Report

**By: Priyanka Joshi**

## Analyze the Healthcare cost and Utilization in Wisconsin hospitals

### Business Scenario

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 the healthcare costs and their utilization.

### Expectations or goals:

The goals of this project are:

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.
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 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 the hospital costs.

### Source Code:

```
getwd()
#importing datasets
mydata<-read.csv('HospitalCosts.csv')
str(mydata)
```

```
#section 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.
head(mydata)
summary(mydata)
hist(mydata$AGE,xlab = "Age of Patients",ylab = "Frequency of visiting hosptital",col = "blue")
summary(as.factor(mydata$AGE))
#to find the maximum expenditure based on age category
s<-tapply(mydata$TOTCHG,mydata$AGE,sum)
barplot(s,xlab = "age",ylab = "hospital cost")
#We can infer from the histogram that infants have the maximum visits to the hospital and has the maximum
expenditure

#section 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.
s1<-summary(as.factor(mydata$APRDRG))
s1
barplot(s1)
which.max(summary(as.factor(mydata$APRDRG)))
d<-tapply(mydata$TOTCHG,mydata$APRDRG,sum)
d
barplot(d,xlab = "diagnosis related group",ylab="hospitalization cost")
which.max(tapply(mydata$TOTCHG,mydata$APRDRG,sum))
#From the results we can see that the category 640 has the maximum entries of hospitalization
#and also has the highest total hospitalization cost (437978).

#section 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.
#we can use ANOVA test to find the relationship between cost(numerical variable)and race(categorical variable)
#H0:Race is related to cost
#H1:There is no relation
mydata$RACE=as.factor(mydata$RACE)
summary(mydata$RACE)
#omit the NA values from the dataset
hospdata<-na.omit(mydata)
summary(hospdata$RACE)
modelannova<- aov(TOTCHG~RACE,data = hospdata)
modelannova
summary(modelannova)
#we can infer that p value is very high around 94% this means that we can reject the null hypothesis also
#so race of the patients is not related to the hospitalization costs
#also we can observe that the race 1 have 484 patients which is higher than other groups so data is skewed and it will
#affect the results by ANOVA
```

#section 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.

#to find whether variable gender and age have impact on hospital costs we can use linear regression model

modelLGM<-lm(TOTCHG~AGE+FEMALE,data = hospdata)

modelLGM

summary(modelLGM)

#from the results we can infer that pvalue for age is very less this means it is a  important factor in the hospital costs as seen

#by the significance levels and p-values

#gender has also less p value means it is also having the impact on cost and same with intercept

#section-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.

modelLGM1<-lm(LOS~AGE+FEMALE+RACE,data = hospdata)

modelLGM1

summary(modelLGM1)

#from the results of mdel we can infer that except for the intercept.

#The very high p-value signifies that there is no linear relationship between the given variables.

#That is, with just the age, gender, and race, it is not possible to predict the los of a patient

#Section-6

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

modelLGM2<-lm(TOTCHG~ .,data = hospdata)

modelLGM2

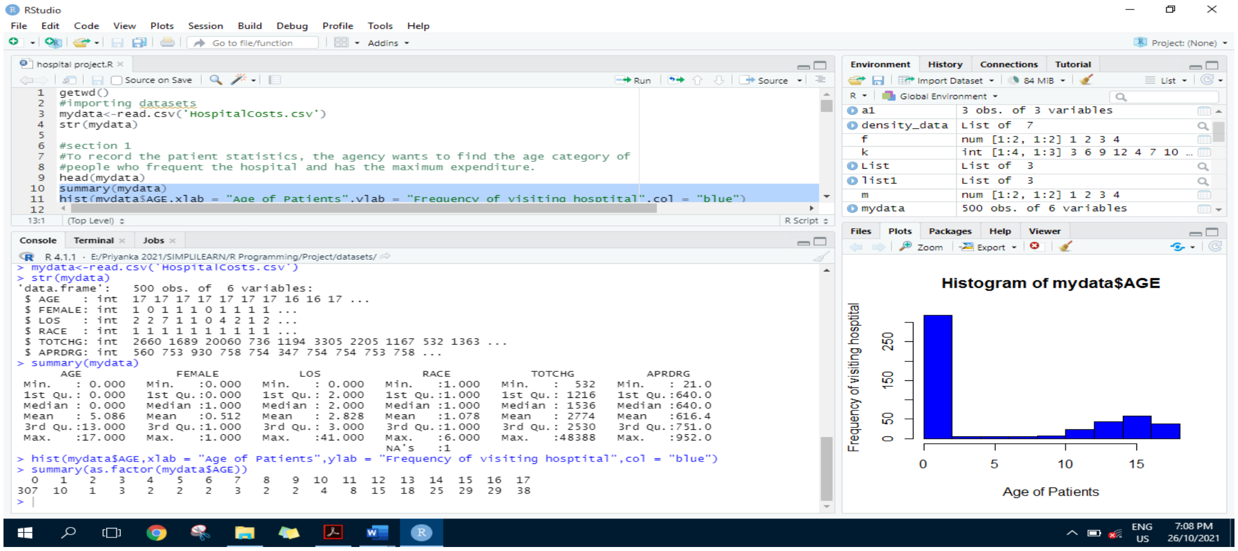
summary(modelLGM2)

#from the output we an say that Age and LOS(Length of stay) affects the hospital costs as higher length

#of stay of the patients will result in higher hospital costs

## Output Screenshot

1)

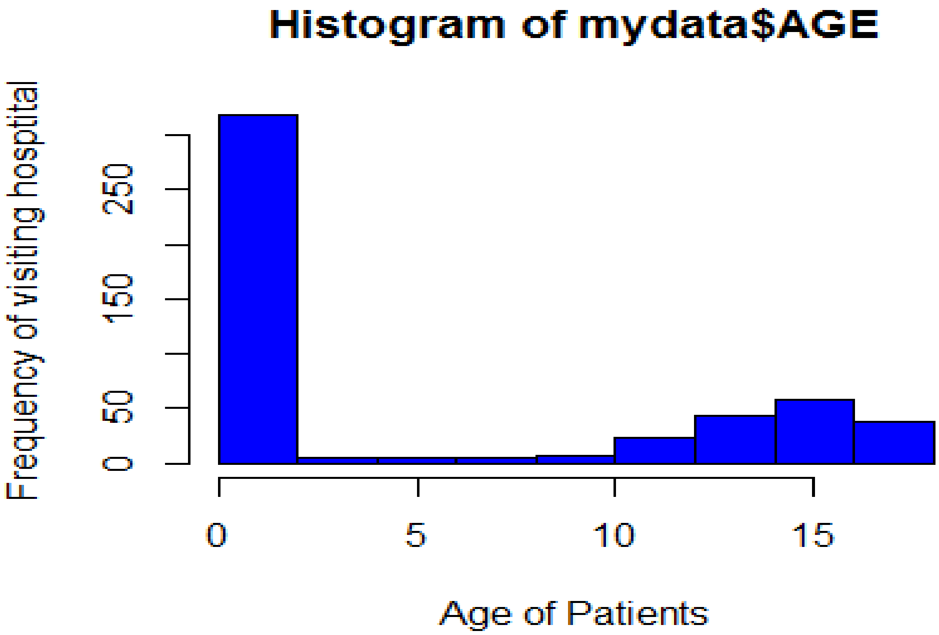


```
getwd()
#importing datasets
mydata<-read.csv('HospitalCosts.csv')
str(mydata)

#section 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.
head(mydata)
summary(mydata)
hist(mydata$AGE,xlab = "Age of Patients",ylab = "Frequency of visiting hosptital",col = "blue")
```

```
> mydata<-read.csv('HospitalCosts.csv')
> str(mydata)
'data.frame':	500 obs. of  6 variables:
 $ AGE   : int  17 17 17 17 17 17 17 16 16 17 ...
 $ FEMALE: int  1 0 1 1 1 0 1 1 1 1 ...
 $ LOS   : int  2 2 7 1 1 0 4 2 1 2 ...
 $ RACE  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ TOTCHG: int  2660 1689 20060 736 1194 3305 2205 1167 532 1363 ...
 $ APRDRG: int  560 753 930 758 754 347 754 754 753 758 ...
> summary(mydata)
      AGE             FEMALE           LOS              RACE           TOTCHG          APRDRG     
 Min.   : 0.000   Min.   :0.000   Min.   : 0.000   Min.   :1.000   Min.   :  532   Min.   : 21.0  
 1st Qu.: 0.000   1st Qu.:0.000   1st Qu.: 2.000   1st Qu.:1.000   1st Qu.: 1216   1st Qu.:640.0  
 Median : 0.000   Median :1.000   Median : 2.000   Median :1.000   Median : 1536   Median :640.0  
 Mean   : 5.086   Mean   :0.512   Mean   : 2.828   Mean   :1.078   Mean   : 2774   Mean   :616.4  
 3rd Qu.:13.000   3rd Qu.:1.000   3rd Qu.: 3.000   3rd Qu.:1.000   3rd Qu.: 2530   3rd Qu.:751.0  
 Max.   :17.000   Max.   :1.000   Max.   :41.000   Max.   :6.000   Max.   :48388   Max.   :952.0  
                                                   NA's   :1                                      
> hist(mydata$AGE,xlab = "Age of Patients",ylab = "Frequency of visiting hosptital",col = "blue")
> summary(as.factor(mydata$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 
> 
```

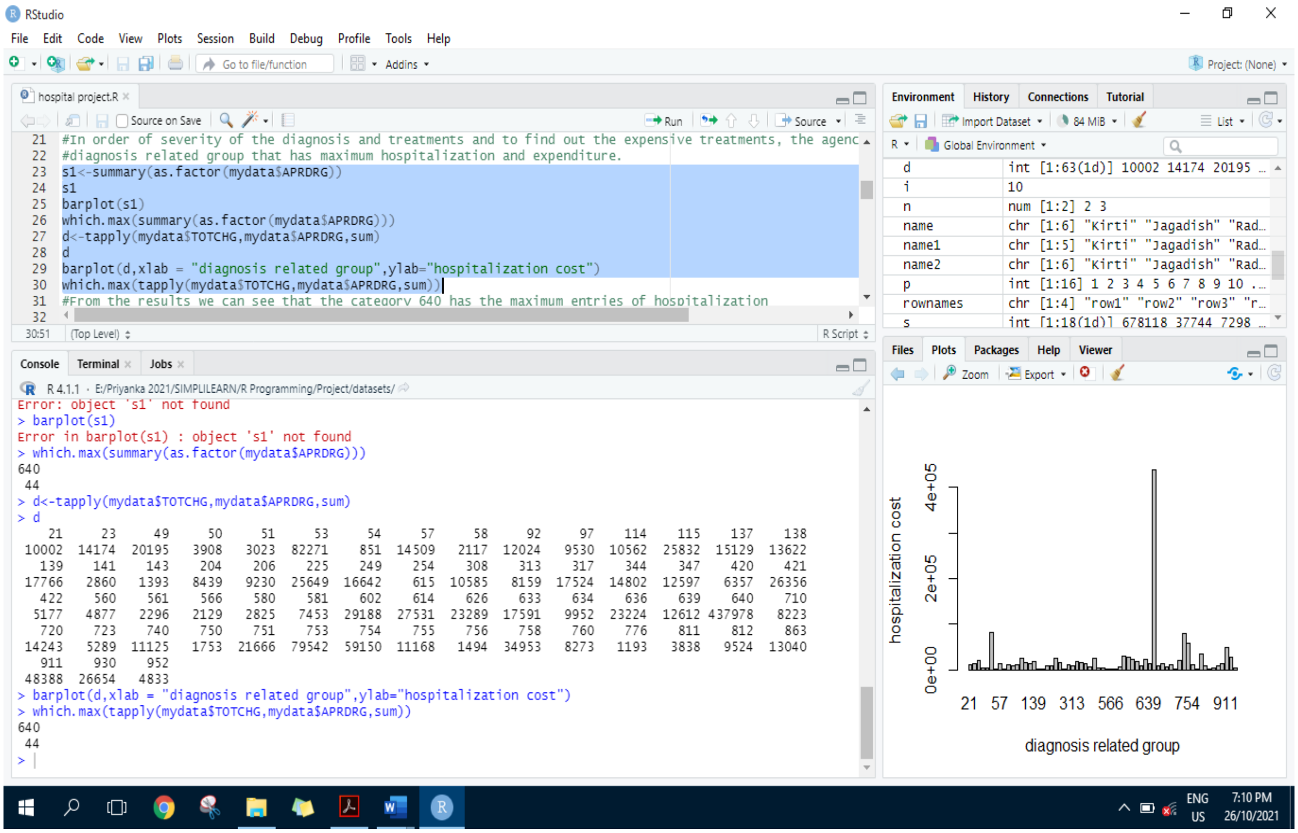
| | |
|---|---|
| a1 | 3 obs. of 3 variables |
| density_data | List of  7 |
| f | num [1:2, 1:2] 1 2 3 4 |
| k | int [1:4, 1:3] 3 6 9 12 4 7 10 ... |
| List | List of  3 |
| list1 | List of  3 |
| m | num [1:2, 1:2] 1 2 3 4 |
| mydata | 500 obs. of 6 variables |



Analysis :

From all the age groups infants have the maximum visits to 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 agenc
#diagnosis related group that has maximum hospitalization and expenditure.
s1<-summary(as.factor(mydata$APRDRG))
s1
barplot(s1)
which.max(summary(as.factor(mydata$APRDRG)))
d<-tapply(mydata$TOTCHG,mydata$APRDRG,sum)
d
barplot(d,xlab = "diagnosis related group",ylab="hospitalization cost")
which.max(tapply(mydata$TOTCHG,mydata$APRDRG,sum))
#From the results we can see that the category 640 has the maximum entries of hospitalization
```

R 4.1.1 · E:/Priyanka 2021/SIMPLILEARN/R Programming/Project/datasets/

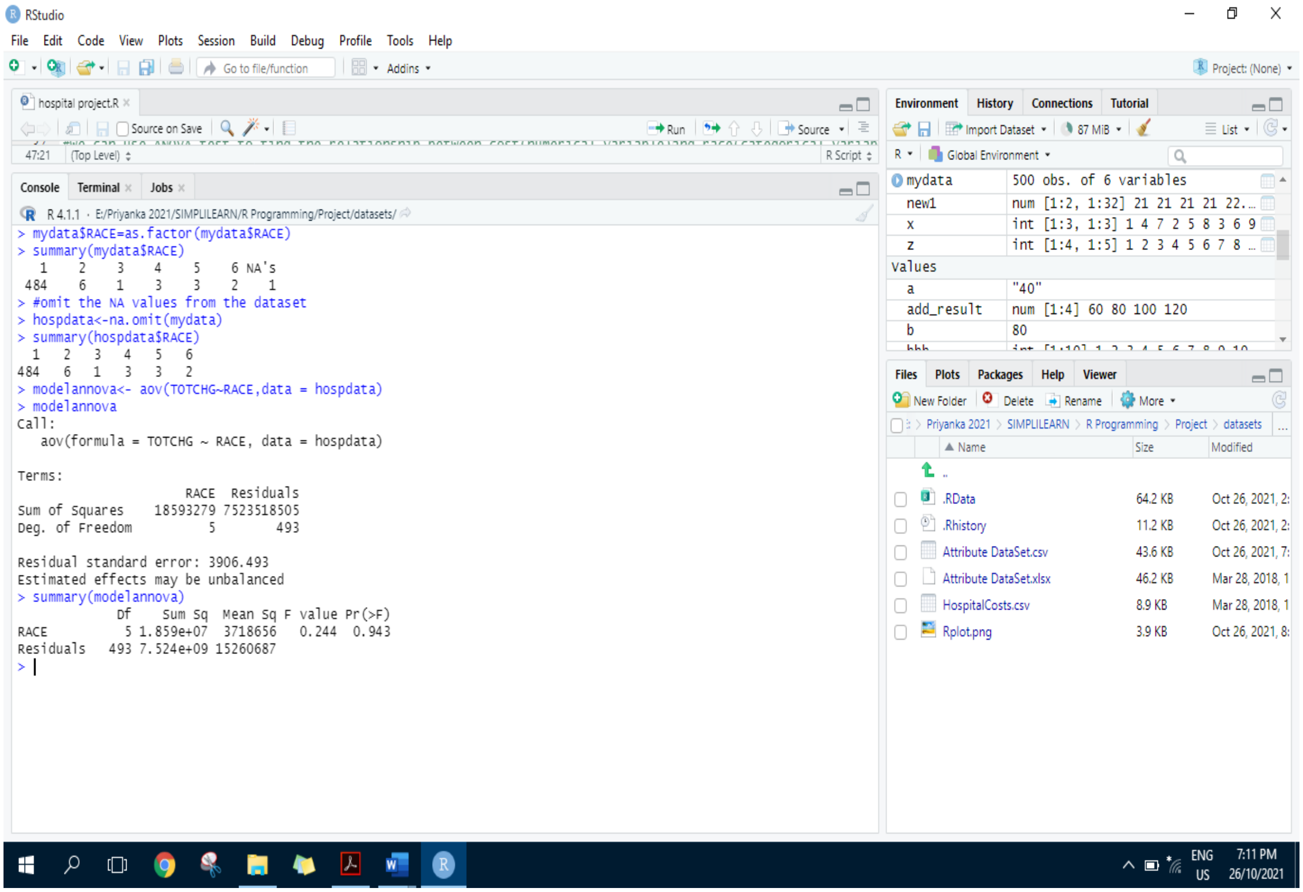
```
Error: object 's1' not found
> barplot(s1)
Error in barplot(s1) : object 's1' not found
> which.max(summary(as.factor(mydata$APRDRG)))
640
 44
> d<-tapply(mydata$TOTCHG,mydata$APRDRG,sum)
> d
    21     23     49     50     51     53     54     57     58     92     97    114    115    137    138
 10002  14174  20195   3908   3023  82271    851  14509   2117  12024   9530  10562  25832  15129  13622
   139    141    143    204    206    225    249    254    308    313    317    344    347    420    421
 17766   2860   1393   8439   9230  25649  16642    615  10585   8159  17524  14802  12597   6357  26356
   422    560    561    566    580    581    602    614    626    633    634    636    639    640    710
  5177   4877   2296   2129   2825   7453  29188  27531  23289  17591   9952  23224  12612 437978   8223
   720    723    740    750    751    753    754    755    756    758    760    776    811    812    863
 14243   5289  11125   1753  21666  79542  59150  11168   1494  34953   8273   1193   3838   9524  13040
   911    930    952
 48388  26654   4833
> barplot(d,xlab = "diagnosis related group",ylab="hospitalization cost")
> which.max(tapply(mydata$TOTCHG,mydata$APRDRG,sum))
640
 44
> |
```

```
d          int [1:63(1d)] 10002 14174 20195 ...
i          10
n          num [1:2] 2 3
name       chr [1:6] "Kirti" "Jagadish" "Rad...
name1      chr [1:5] "Kirti" "Jagadish" "Rad...
name2      chr [1:6] "Kirti" "Jagadish" "Rad...
p          int [1:16] 1 2 3 4 5 6 7 8 9 10 ...
rownames   chr [1:4] "row1" "row2" "row3" "r...
s          int [1:18(1d)] 678118 37744 7298 ...
```

Analysis:

From the results we can see that the category 640 has the maximum entries of hospitalization and also has the highest total hospitalization cost (437978).

3)



```
> mydata$RACE=as.factor(mydata$RACE)
> summary(mydata$RACE)
   1    2    3    4    5    6 NA's 
 484    6    1    3    3    2    1 
> #omit the NA values from the dataset
> hospdata<-na.omit(mydata)
> summary(hospdata$RACE)
  1   2   3   4   5   6 
484   6   1   3   3   2 
> modelannova<- aov(TOTCHG~RACE,data = hospdata)
> modelannova
Call:
   aov(formula = TOTCHG ~ RACE, data = hospdata)

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

Residual standard error: 3906.493
Estimated effects may be unbalanced
> summary(modelannova)
             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               
> 
```

| | |
|---|---|
| mydata | 500 obs. of 6 variables |
| new1 | num [1:2, 1:32] 21 21 21 21 22.… |
| x | int [1:3, 1:3] 1 4 7 2 5 8 3 6 9 |
| z | int [1:4, 1:5] 1 2 3 4 5 6 7 8 … |
| values | |
| a | "40" |
| add_result | num [1:4] 60 80 100 120 |
| b | 80 |

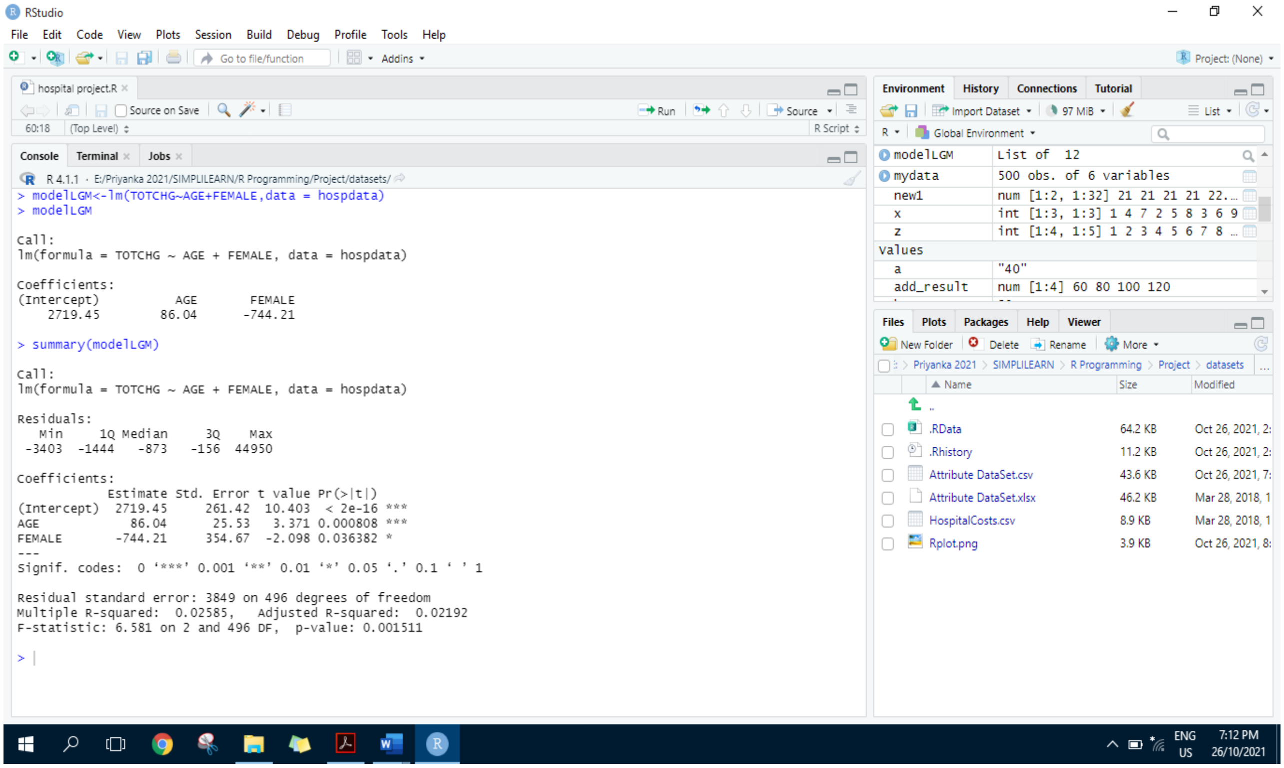
| Name | Size | Modified |
|---|---|---|
| .. | | |
| .RData | 64.2 KB | Oct 26, 2021, 2: |
| .Rhistory | 11.2 KB | Oct 26, 2021, 2: |
| Attribute DataSet.csv | 43.6 KB | Oct 26, 2021, 7: |
| Attribute DataSet.xlsx | 46.2 KB | Mar 28, 2018, 1 |
| HospitalCosts.csv | 8.9 KB | Mar 28, 2018, 1 |
| Rplot.png | 3.9 KB | Oct 26, 2021, 8: |

Analysis:

we can infer that p value is very high around 94% this means that we can reject the null hypothesis also race of the patients is not related to the hospitalization costs

we can observe that the race 1 have 484 patients which is higher than other groups so data is skewed and it will affect the results by ANOVA

4)



```
R 4.1.1 · E:/Priyanka 2021/SIMPLILEARN/R Programming/Project/datasets/
> modelLGM<-lm(TOTCHG~AGE+FEMALE,data = hospdata)
> modelLGM

Call:
lm(formula = TOTCHG ~ AGE + FEMALE, data = hospdata)

Coefficients:
(Intercept)          AGE       FEMALE  
    2719.45        86.04      -744.21  

> summary(modelLGM)

Call:
lm(formula = TOTCHG ~ AGE + FEMALE, data = hospdata)

Residuals:
   Min     1Q Median     3Q    Max 
 -3403  -1444   -873   -156  44950 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2719.45     261.42  10.403  < 2e-16 ***
AGE            86.04      25.53   3.371 0.000808 ***
FEMALE       -744.21     354.67  -2.098 0.036382 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3849 on 496 degrees of freedom
Multiple R-squared:  0.02585,	Adjusted R-squared:  0.02192 
F-statistic: 6.581 on 2 and 496 DF,  p-value: 0.001511

> |
```

```
modelLGM      List of  12
mydata        500 obs. of 6 variables
  new1        num [1:2, 1:32] 21 21 21 21 22....
  x           int [1:3, 1:3] 1 4 7 2 5 8 3 6 9
  z           int [1:4, 1:5] 1 2 3 4 5 6 7 8 ...
Values
  a           "40"
  add_result  num [1:4] 60 80 100 120
```

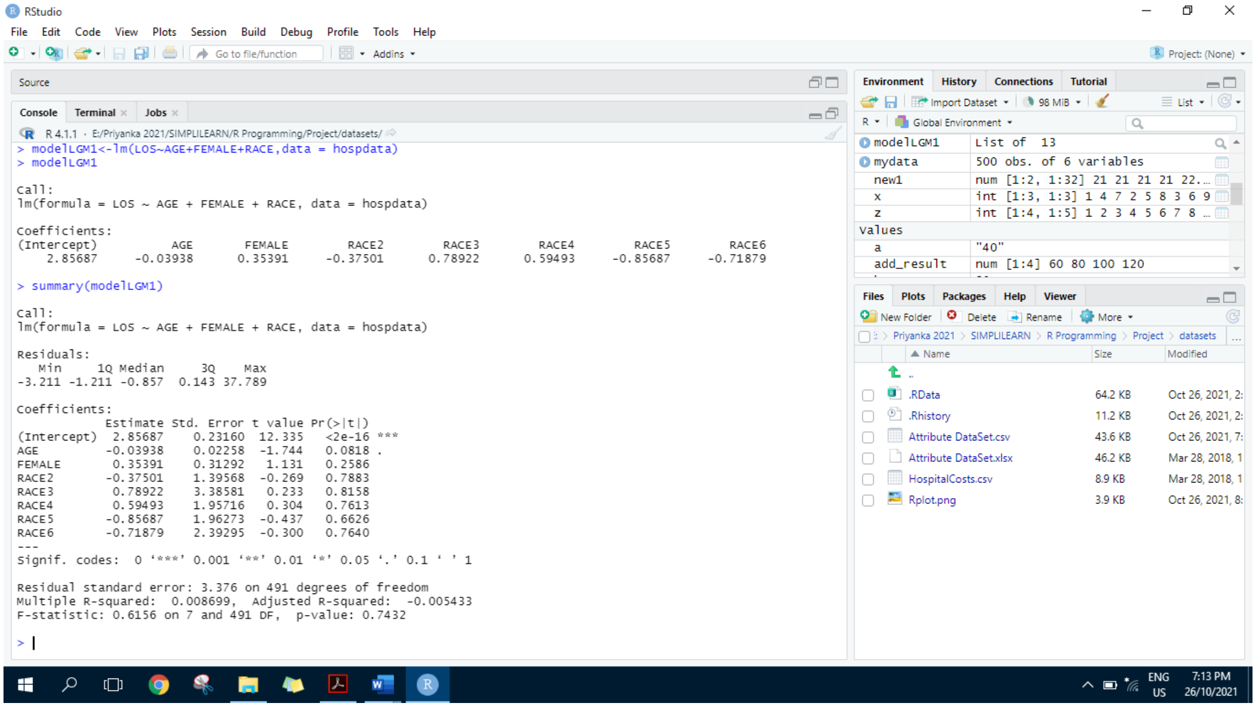
| | Name | Size | Modified |
|---|---|---|---|
| | .. | | |
| □ | .RData | 64.2 KB | Oct 26, 2021, 2: |
| □ | .Rhistory | 11.2 KB | Oct 26, 2021, 2: |
| □ | Attribute DataSet.csv | 43.6 KB | Oct 26, 2021, 7: |
| □ | Attribute DataSet.xlsx | 46.2 KB | Mar 28, 2018, 1 |
| □ | HospitalCosts.csv | 8.9 KB | Mar 28, 2018, 1 |
| □ | Rplot.png | 3.9 KB | Oct 26, 2021, 8: |

Analysis:

As we can see from the results that p value for age is very less this means it is an important factor in the hospital costs as seen by the significance levels and p-values

gender has also less p value means it is also having the impact on cost and same with intercept

5)



```
R 4.1.1 · E:/Priyanka 2021/SIMPLILEARN/R Programming/Project/datasets/
> modelLGM1<-lm(LOS~AGE+FEMALE+RACE,data = hospdata)
> modelLGM1

Call:
lm(formula = LOS ~ AGE + FEMALE + RACE, data = hospdata)

Coefficients:
(Intercept)          AGE       FEMALE        RACE2        RACE3        RACE4        RACE5        RACE6  
    2.85687     -0.03938      0.35391     -0.37501      0.78922      0.59493     -0.85687     -0.71879  

> summary(modelLGM1)

Call:
lm(formula = LOS ~ AGE + FEMALE + RACE, data = hospdata)

Residuals:
   Min     1Q Median     3Q    Max 
-3.211 -1.211 -0.857  0.143 37.789 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.85687    0.23160  12.335   <2e-16 ***
AGE         -0.03938    0.02258  -1.744   0.0818 .  
FEMALE       0.35391    0.31292   1.131   0.2586    
RACE2       -0.37501    1.39568  -0.269   0.7883    
RACE3        0.78922    3.38581   0.233   0.8158    
RACE4        0.59493    1.95716   0.304   0.7613    
RACE5       -0.85687    1.96273  -0.437   0.6626    
RACE6       -0.71879    2.39295  -0.300   0.7640    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.376 on 491 degrees of freedom
Multiple R-squared:  0.008699,	Adjusted R-squared:  -0.005433 
F-statistic: 0.6156 on 7 and 491 DF,  p-value: 0.7432

> |
```

```
modelLGM1     List of  13
mydata        500 obs. of 6 variables
  new1        num [1:2, 1:32] 21 21 21 21 22....
  x           int [1:3, 1:3] 1 4 7 2 5 8 3 6 9
  z           int [1:4, 1:5] 1 2 3 4 5 6 7 8 ...
Values
  a           "40"
  add_result  num [1:4] 60 80 100 120
```

| | Name | Size | Modified |
|---|---|---|---|
| | .. | | |
| □ | .RData | 64.2 KB | Oct 26, 2021, 2: |
| □ | .Rhistory | 11.2 KB | Oct 26, 2021, 2: |
| □ | Attribute DataSet.csv | 43.6 KB | Oct 26, 2021, 7: |
| □ | Attribute DataSet.xlsx | 46.2 KB | Mar 28, 2018, 1 |
| □ | HospitalCosts.csv | 8.9 KB | Mar 28, 2018, 1 |
| □ | Rplot.png | 3.9 KB | Oct 26, 2021, 8: |

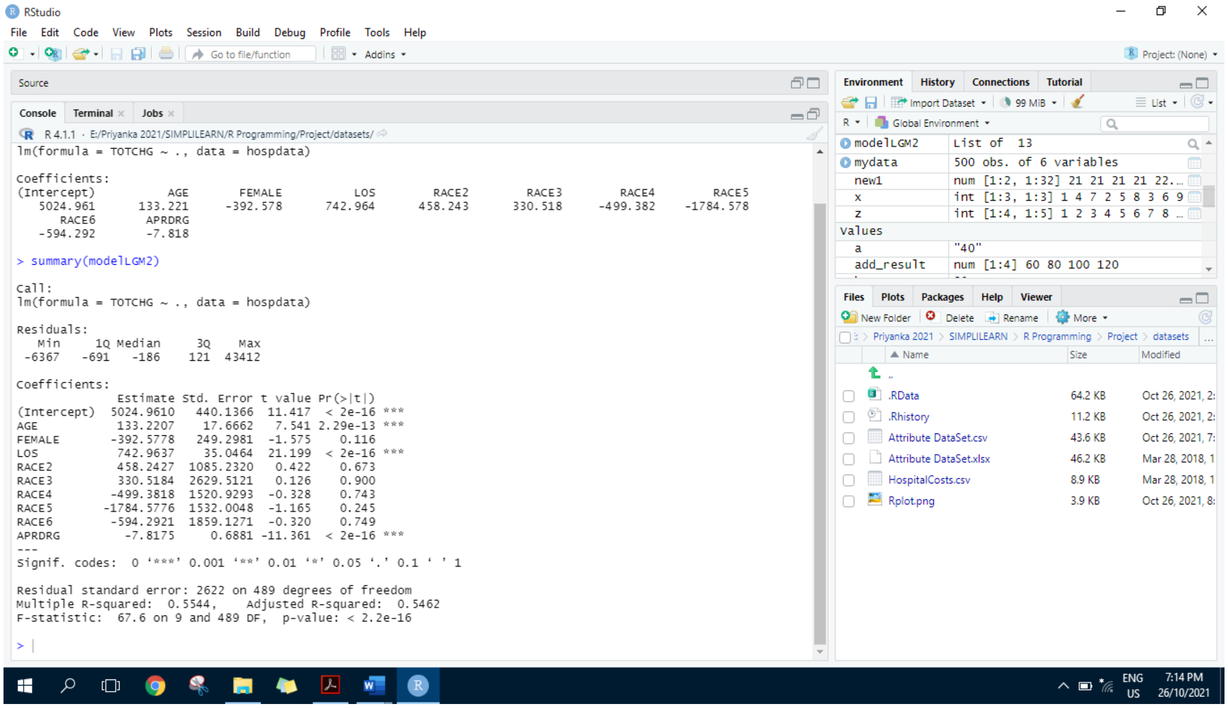
Analysis:

By observing the results of model, we can infer that except for the intercept p values are very high.

The very high p-value signifies that there is no linear relationship between the given variables.

That is, with just the age, gender, and race, it is not possible to predict the LOS (length of stay) of a patient

6)



```
lm(formula = TOTCHG ~ ., data = hospdata)

Coefficients:
(Intercept)          AGE       FEMALE          LOS        RACE2        RACE3        RACE4        RACE5
   5024.961      133.221     -392.578      742.964      458.243      330.518     -499.382    -1784.578
      RACE6       APRDRG
   -594.292       -7.818

> summary(modelLGM2)

Call:
lm(formula = TOTCHG ~ ., data = hospdata)

Residuals:
   Min     1Q Median     3Q    Max
 -6367   -691   -186    121  43412

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5024.9610   440.1366  11.417  < 2e-16 ***
AGE           133.2207    17.6662   7.541 2.29e-13 ***
FEMALE       -392.5778   249.2981  -1.575    0.116
LOS           742.9637    35.0464  21.199  < 2e-16 ***
RACE2         458.2427  1085.2320   0.422    0.673
RACE3         330.5184  2629.5121   0.126    0.900
RACE4        -499.3818  1520.9293  -0.328    0.743
RACE5       -1784.5776  1532.0048  -1.165    0.245
RACE6        -594.2921  1859.1271  -0.320    0.749
APRDRG         -7.8175     0.6881 -11.361  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2622 on 489 degrees of freedom
Multiple R-squared:  0.5544,    Adjusted R-squared:  0.5462
F-statistic:  67.6 on 9 and 489 DF,  p-value: < 2.2e-16

> |
```

| | |
|---|---|
| modelLGM2 | List of 13 |
| mydata | 500 obs. of 6 variables |
| new1 | num [1:2, 1:32] 21 21 21 21 22.… |
| x | int [1:3, 1:3] 1 4 7 2 5 8 3 6 9 |
| z | int [1:4, 1:5] 1 2 3 4 5 6 7 8 … |
| Values | |
| a | "40" |
| add_result | num [1:4] 60 80 100 120 |

| Name | Size | Modified |
|---|---|---|
| .. | | |
| .RData | 64.2 KB | Oct 26, 2021, 2: |
| .Rhistory | 11.2 KB | Oct 26, 2021, 2: |
| Attribute DataSet.csv | 43.6 KB | Oct 26, 2021, 7: |
| Attribute DataSet.xlsx | 46.2 KB | Mar 28, 2018, 1 |
| HospitalCosts.csv | 8.9 KB | Mar 28, 2018, 1 |
| Rplot.png | 3.9 KB | Oct 26, 2021, 8: |

<u>Analysis</u>

from the output we can say that Age and LOS(Length of stay) affects the hospital costs as higher length

of stay of the patients will result in higher hospital costs