Analyze the Healthcare cost and Utilization in Wisconsin hospitals

Business Analytic Foundation with R Tools- Question

Abstract

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.

Presented by: Ankita Agarwal

Problem Statement:

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.

Detailed description of the given dataset:

AGE: Age of the patient discharged
FEMALE: 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

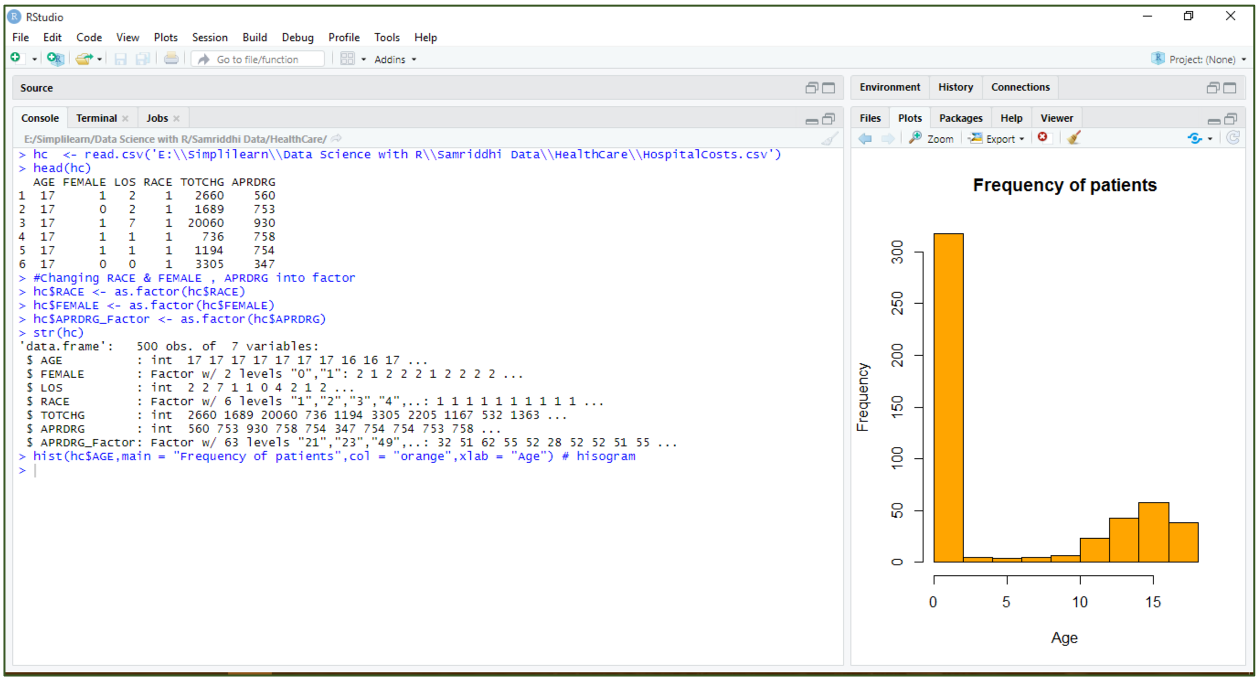
To Analyze:

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.

Analysis and Interpretations:

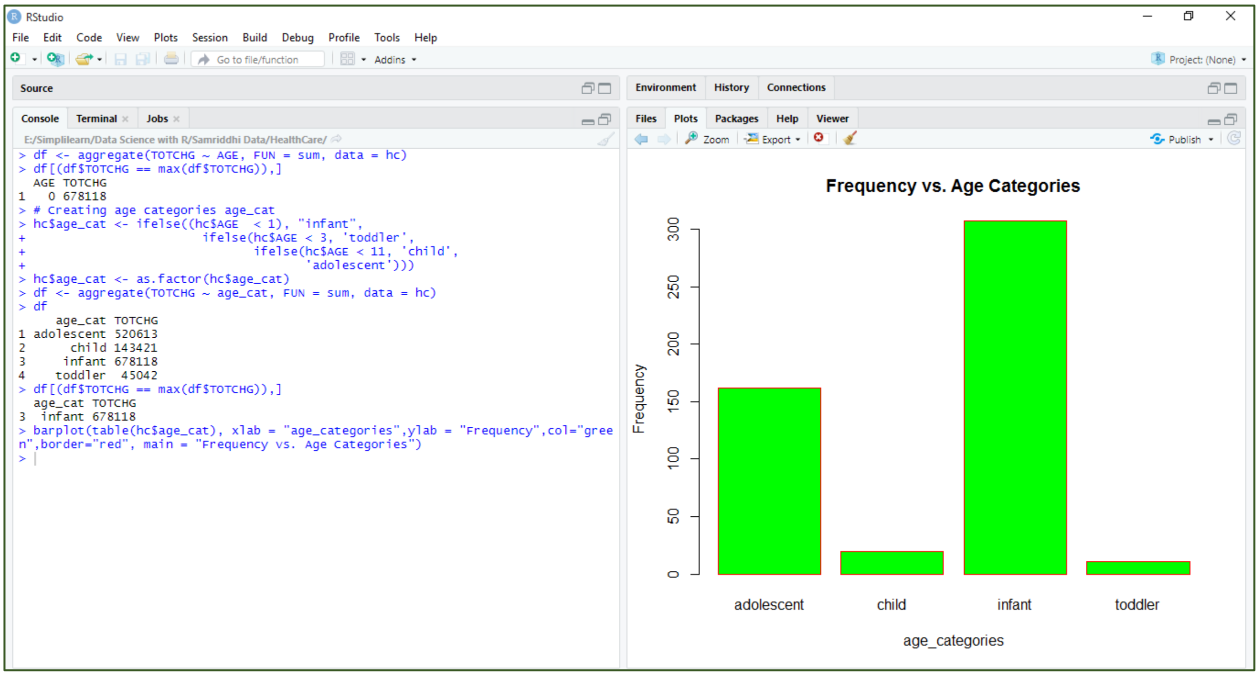
1. **To record the patient statistics, the agency wants to find the age category of people who frequent the hospital and has the maximum expenditure.**

a. To find the age category that has the highest frequency of hospital visit, a histogram can be plotted that would display the number of occurrences of each age category.



```
> hc  <- read.csv('E:\\Simplilearn\\Data Science with R\\Samriddhi Data\\HealthCare\\HospitalCosts.csv')
> head(hc)
  AGE FEMALE LOS RACE TOTCHG APRDRG
1  17      1   2    1   2660    560
2  17      0   2    1   1689    753
3  17      1   7    1  20060    930
4  17      1   1    1    736    758
5  17      1   1    1   1194    754
6  17      0   0    1   3305    347
> #Changing RACE & FEMALE , APRDRG into factor
> hc$RACE <- as.factor(hc$RACE)
> hc$FEMALE <- as.factor(hc$FEMALE)
> hc$APRDRG_Factor <- as.factor(hc$APRDRG)
> str(hc)
'data.frame':   500 obs. of  7 variables:
 $ AGE          : int  17 17 17 17 17 17 17 16 16 17 ...
 $ FEMALE       : Factor w/ 2 levels "0","1": 2 1 2 2 2 1 2 2 2 2 ...
 $ LOS          : int  2 2 7 1 1 0 4 2 1 2 ...
 $ RACE         : Factor w/ 6 levels "1","2","3","4",..: 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 ...
 $ APRDRG_Factor: Factor w/ 63 levels "21","23","49",..: 32 51 62 55 52 28 52 52 51 55 ...
> hist(hc$AGE,main = "Frequency of patients",col = "orange",xlab = "Age") # hisogram
> 
```

b. To find the age category with the maximum expenditure, we use the aggregate function to add the values of total expenditure according to the values of age.



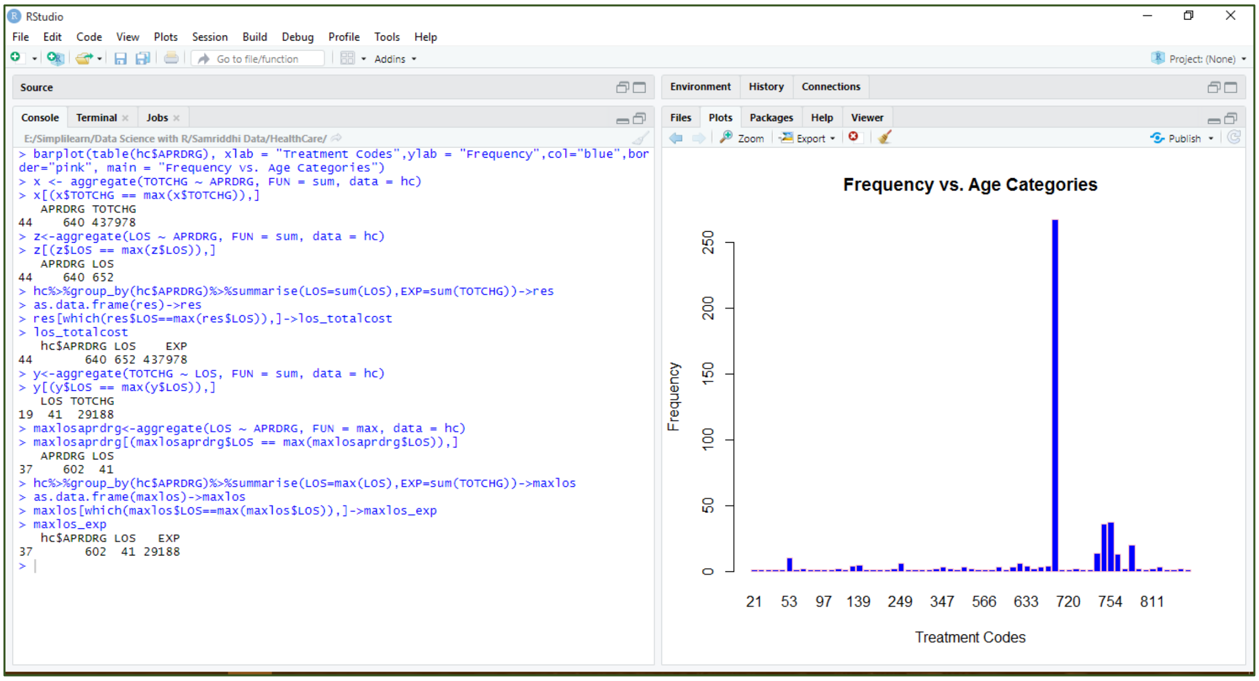
```
> df <- aggregate(TOTCHG ~ AGE, FUN = sum, data = hc)
> df[(df$TOTCHG == max(df$TOTCHG)),]
  AGE TOTCHG
1   0 678118
> # Creating age categories age_cat
> hc$age_cat <- ifelse((hc$AGE  < 1), "infant",
+                      ifelse(hc$AGE < 3, 'toddler',
+                             ifelse(hc$AGE < 11, 'child',
+                                    'adolescent')))
> hc$age_cat <- as.factor(hc$age_cat)
> df <- aggregate(TOTCHG ~ age_cat, FUN = sum, data = hc)
> df
     age_cat TOTCHG
1 adolescent 520613
2      child 143421
3     infant 678118
4    toddler  45042
> df[(df$TOTCHG == max(df$TOTCHG)),]
  age_cat TOTCHG
3  infant 678118
> barplot(table(hc$age_cat), xlab = "age_categories",ylab = "Frequency",col="gree
n",border="red", main = "Frequency vs. Age Categories")
> 
```

Interpretation:

By observing the above table, it is very clear that the people of age '0' frequent the hospital most often and has the most expenditure (678118). By converting age into age categories, we can observe that infants (0-1 years) are more prone to visit the hospital, fetching in most expenditure. Following the infants, adolescents between the ages of 11 to 17 years have high hospitalization costs.

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.

To find the diagnosis related group that has maximum hospitalization and expenditure, we use the aggregate function or we can use group by to summarize the characteristics of the variables LOS and TOTCRG with respect to APRDRG.



```
> barplot(table(hc$APRDRG), xlab = "Treatment Codes",ylab = "Frequency",col="blue",bor
der="pink", main = "Frequency vs. Age Categories")
> x <- aggregate(TOTCHG ~ APRDRG, FUN = sum, data = hc)
> x[(x$TOTCHG == max(x$TOTCHG)),]
   APRDRG TOTCHG
44    640 437978
> z<-aggregate(LOS ~ APRDRG, FUN = sum, data = hc)
> z[(z$LOS == max(z$LOS)),]
   APRDRG LOS
44    640 652
> hc%>%group_by(hc$APRDRG)%>%summarise(LOS=sum(LOS),EXP=sum(TOTCHG))->res
> as.data.frame(res)->res
> res[which(res$LOS==max(res$LOS)),]->los_totalcost
> los_totalcost
   hc$APRDRG LOS    EXP
44       640 652 437978
> y<-aggregate(TOTCHG ~ LOS, FUN = sum, data = hc)
> y[(y$LOS == max(y$LOS)),]
   LOS TOTCHG
19  41  29188
> maxlosaprdrg<-aggregate(LOS ~ APRDRG, FUN = max, data = hc)
> maxlosaprdrg[(maxlosaprdrg$LOS == max(maxlosaprdrg$LOS)),]
   APRDRG LOS
37    602  41
> hc%>%group_by(hc$APRDRG)%>%summarise(LOS=max(LOS),EXP=sum(TOTCHG))->maxlos
> as.data.frame(maxlos)->maxlos
> maxlos[which(maxlos$LOS==max(maxlos$LOS)),]->maxlos_exp
> maxlos_exp
   hc$APRDRG LOS   EXP
37       602  41 29188
> |
```

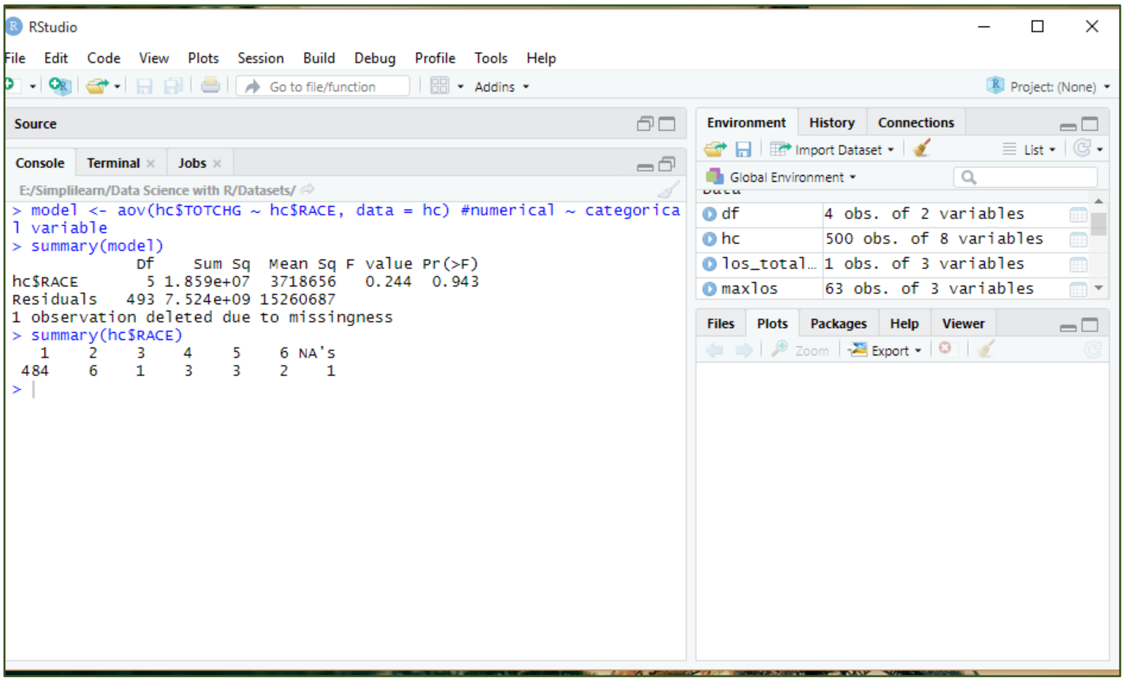
Interpretation:

As seen in the graph, maximum people are diagnosed with group 640, inferring that the most expensive treatment by far is for diagnosis group 640 with a total charge of 437978 and length of stay is summed up to 652 days. However, the longest length of stay is 41 days costing 29188 for diagnosis group 602.

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.

To analyze if there is a relation between Race and Total Cost, we need to convert the Race variable to factor (done in the beginning of the project). We now do an ANOVA test on RACE and TOTCHG to verify the impact of Race on Total Costs.

Defining Hypothesis:
Ho: The race has no an impact on the costs.
Ha: The race has an impact on the costs.



E:/Simplilearn/Data Science with R/Datasets/

```
> model <- aov(hc$TOTCHG ~ hc$RACE, data = hc) #numerical ~ categorica
l variable
> summary(model)
             Df    Sum Sq  Mean Sq F value Pr(>F)
hc$RACE       5 1.859e+07  3718656   0.244  0.943
Residuals   493 7.524e+09 15260687
1 observation deleted due to missingness
> summary(hc$RACE)
   1    2    3    4    5    6 NA's
 484    6    1    3    3    2    1
> |
```

df	4 obs. of 2 variables
hc	500 obs. of 8 variables
los_total…	1 obs. of 3 variables
maxlos	63 obs. of 3 variables

Interpretation:

As we know, when p-value < alpha, p-value is less than alpha; we reject the null hypothesis. We take alpha value as 0.05 at 95% confidence level.

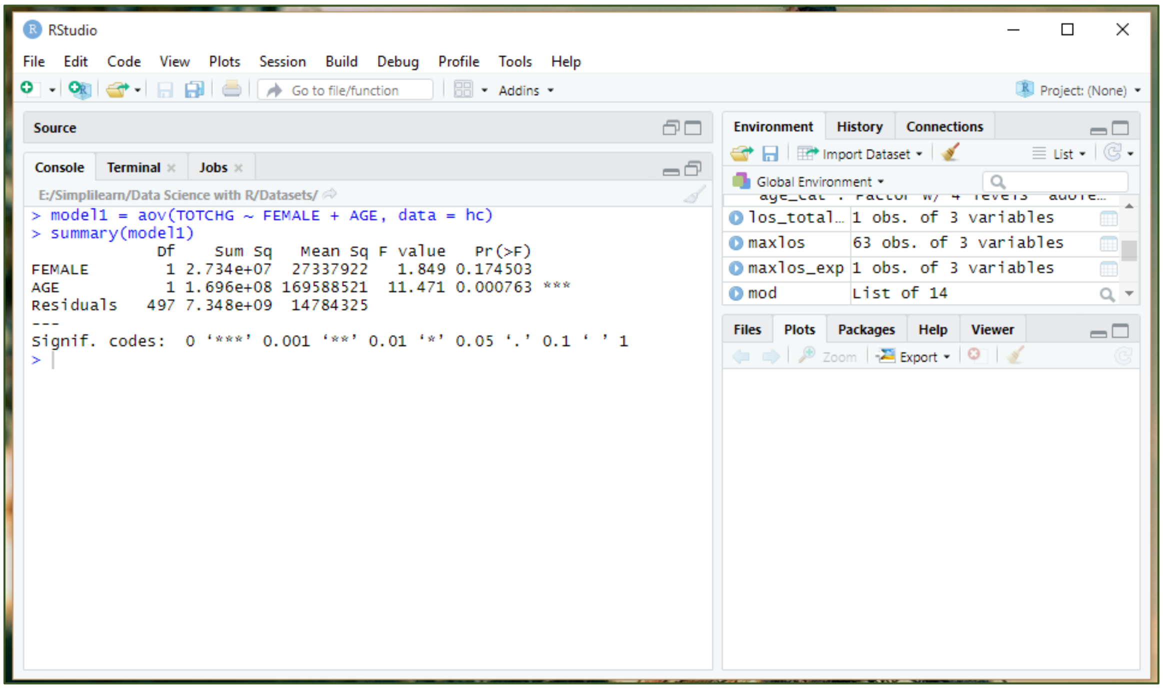
Here, p-value = 0.943. which means that the p-value is significantly high, emphasizing that there is no relation between the race of patient and the hospital cost. Also, from the summary we notice that, the data has 484 patients of Race 1 out of the 500 entries. This will affect the results of ANOVA as well, since the number of observations is very much skewed.

In conclusion, there is not enough data to verify if the race of patient is related to the hospitalization cost.

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 analyze if there is a relation between Gender + Age and Total Cost, we need to convert both FEMALE and AGE variables to factors (done in the beginning of the project). Then, we do an ANOVA test with the following variables: FEMALE + AGE and TOTCHG to verify its impact on Total Costs.

Defining Hypothesis
Ho: The gender and age have no an impact on the total costs.
Ha: The gender and age have an impact on the total costs.



```
> model1 = aov(TOTCHG ~ FEMALE + AGE, data = hc)
> summary(model1)
             Df    Sum Sq   Mean Sq F value   Pr(>F)    
FEMALE        1 2.734e+07  27337922   1.849 0.174503    
AGE           1 1.696e+08 169588521  11.471 0.000763 ***
Residuals   497 7.348e+09  14784325                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

```
  age_cat : Factor w/ 4 levels "above…
los_total… 1 obs. of 3 variables
maxlos     63 obs. of 3 variables
maxlos_exp 1 obs. of 3 variables
mod        List of 14
```

Interpretation:

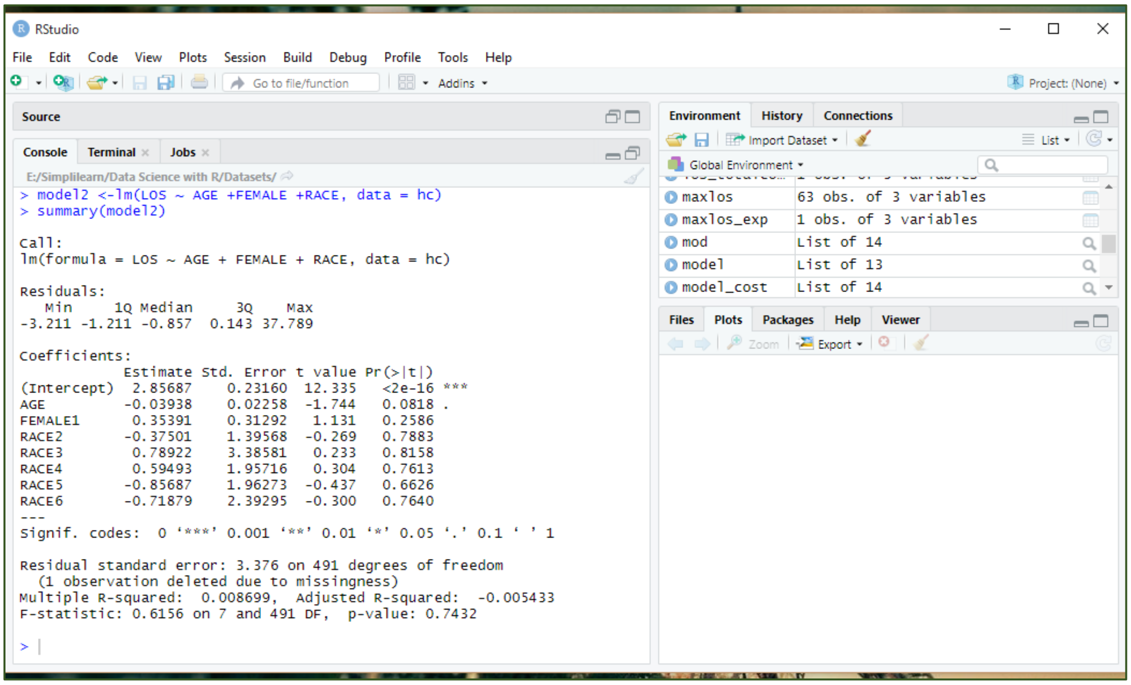
From the above ANOVA model, we observe that p-value of AGE (0.000763) is much lower than 0.05, which suggests that Age is quite significant. Hence, we retain the Null Hypothesis (Ho). However, p-value of FEMALE (0.174503) is higher than 0.05, which suggests that gender has very less impact on the costs. Hence, we cannot retain the Null Hypothesis (Ho).

To conclude, Age does impact the Total Cost of hospitalization whereas Gender does not have much severity towards the Total Cost.

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.

To predict if LOS is based on Gender, Age and Race, we use multiple linear regression model to analyze as we have more than one independent variable.

Defining Hypothesis:
Ho: There exists no relation between age, gender and race with LOS
Ha: There exists a relation between age, gender and race with LOS



E:/Simplilearn/Data Science with R/Datasets/

```
> model2 <-lm(LOS ~ AGE +FEMALE +RACE, data = hc)
> summary(model2)

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

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 .
FEMALE1      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
  (1 observation deleted due to missingness)
Multiple R-squared:  0.008699,  Adjusted R-squared:  -0.005433
F-statistic: 0.6156 on 7 and 491 DF,  p-value: 0.7432

> |
```

maxlos	63 obs. of 3 variables
maxlos_exp	1 obs. of 3 variables
mod	List of 14
model	List of 13
model_cost	List of 14

Interpretation:

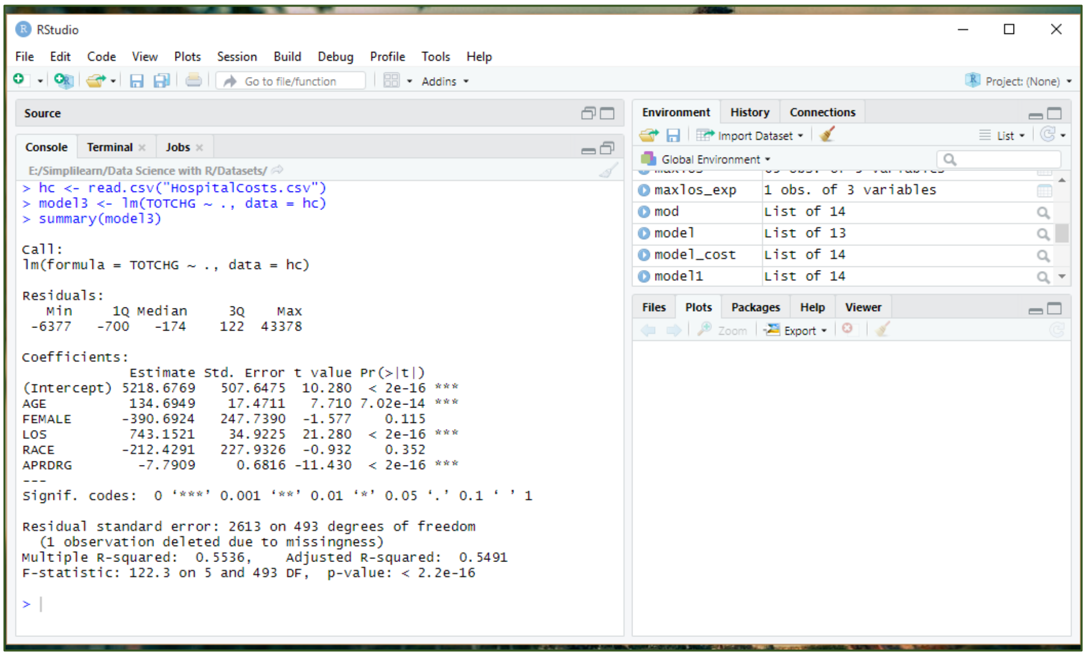
As we know, p-value < alpha, p-value is less than alpha; we reject the null hypothesis. We take alpha value as 0.05 at 95% confidence level.

We can clearly see that the significance codes are almost null for all the variables, except for the intercept. The p-value (0.7432) is quiet high which signifies that there is no linear relationship between the given variables. Hence, we cannot predict the length of stay of the patients with respect to the age, gender and race.

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

To find the variables that mainly affect the total costs, we construct a linear regression model with all the variables.

Defining Hypothesis:
Ho: There is no linear relationship between dependent and independent variables
Ha: There is a linear relationship between dependent and independent variables



```
> hc <- read.csv("HospitalCosts.csv")
> model3 <- lm(TOTCHG ~ ., data = hc)
> summary(model3)

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

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
LOS          743.1521    34.9225  21.280  < 2e-16 ***
RACE        -212.4291   227.9326  -0.932    0.352
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

> |
```

maxlos_exp	1 obs. of 3 variables
mod	List of 14
model	List of 13
model_cost	List of 14
model1	List of 14

Interpretation:

Since p-value (2.2e-16) < alpha, there is a linear relationship between the all variables.

Based on the output, the Age and LOS are statistically significant with a very low p-value and affects the total cost. Also, cost is directly proportional to the LOS i.e. higher the LOS, higher is the cost; increase in a day of stay, increase the cost by 743.

All Patient Refined Diagnosis Related Groups with a low p-value, is also affecting the costs. As seen in our prior analysis that APRDRG 640 has high expenditure and highest LOS. So, we can say that people with APRDRG 640 are more likely to be hospitalized and increase the hospitalization costs.

From the significance codes we can see RACE and Gender are clearly not affecting the hospital costs.

Programming Codes:

#Reading Hospital cost data

```
library(dplyr)
hc <- read.csv('E:\\Simplilearn\\Data Science with R\\Samriddhi Data\\HealthCare\\HospitalCosts.csv')
head(hc)

#Changing RACE & FEMALE, APRDRG into factor
hc$RACE <- as.factor(hc$RACE)
hc$FEMALE <- as.factor(hc$FEMALE)
hc$APRDRG_Factor <- as.factor(hc$APRDRG)
```

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

```
hist(hc$AGE,main = "Frequency of patients",col = "orange",xlab = "Age") # hisogram
summary(as.factor(hc$AGE)) # summary of age data

df <- aggregate(TOTCHG ~ AGE, FUN = sum, data = hc)
df[(df$TOTCHG == max(df$TOTCHG)),]

# Creating age categories age_cat
hc$age_cat <- ifelse((hc$AGE < 1), "infant",
        ifelse(hc$AGE < 3, 'toddler',
          ifelse(hc$AGE < 11, 'child',
             'adolescent')))
hc$age_cat <- as.factor(hc$age_cat)

df <- aggregate(TOTCHG ~ age_cat, FUN = sum, data = hc)
df[(df$TOTCHG == max(df$TOTCHG)),]

barplot(table(hc$age_cat), xlab = "age_categories",ylab = "Frequency",col="green",border="red", main =
"Frequency vs. Age Categories")
```

#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.

```
barplot(table(hc$APRDRG), xlab = "Treatment Codes",ylab = "Frequency",col="blue",border="pink", main =
"Frequency vs. Age Categories")

x <- aggregate(TOTCHG ~ APRDRG, FUN = sum, data = hc)
x[(x$TOTCHG == max(x$TOTCHG)),]

z<-aggregate(LOS ~ APRDRG, FUN = sum, data = hc)
z[(z$LOS == max(z$LOS)),]
```

```
#or using this

hc%>%group_by(hc$APRDRG)%>%summarise(LOS=sum(LOS),EXP=sum(TOTCHG))->res
as.data.frame(res)->res
res[which(res$LOS==max(res$LOS)),]->los_totalcost
los_totalcost

y<-aggregate(TOTCHG ~ LOS, FUN = sum, data = hc)
y[(y$LOS == max(y$LOS)),]

s<-aggregate(LOS ~ APRDRG, FUN = max, data = hc)
s[(s$LOS == max(s$LOS)),]

#or using this

hc%>%group_by(hc$APRDRG)%>%summarise(LOS=max(LOS),EXP=sum(TOTCHG))->maxlos
as.data.frame(maxlos)->maxlos
maxlos[which(maxlos$LOS==max(maxlos$LOS)),]->maxlos_exp
maxlos_exp
```

#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.

```
model <- aov(hc$TOTCHG ~ hc$RACE, data = hc) #numerical ~ categorical variable
summary(model)

summary(hc$RACE)
```

#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.

```
model1 = aov(TOTCHG ~ FEMALE + AGE, data = hc)
summary(model1)
```

#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.

```
model2 <-lm(LOS ~ AGE +FEMALE +RACE, data = hc)
summary(model2)
```

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

```
hc <- read.csv("HospitalCosts.csv")
model3 <- lm(TOTCHG ~ ., data = hc)
summary(model3)
```

--The End--