



PROJECT REPORT

“Analyze the Healthcare cost and Utilization in Wisconsin hospitals”



MADE BY: AVINASH SHARMA

COURSE: R PROGRAMMING

CONTENT

SL NO	TITLE	PAGE NO
1.	BUSINESS SCENARIO	2
2.	EXPECTED OUTCOME	3
3.	CODE	4-5
4.	FUNCTIONS IMPLEMENTED	6
5.	ANALYSIS	7-

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.

Attributes Description:

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

EXPECTED OUTCOME

- 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.

CODE

```
h_c<-read.csv("C:/Users/bhati/Desktop/SimpliLearn/Data Science with R/projects to  
solve/Healthcare/Healthcare/HospitalCosts.csv")
```

```
summary(h_c)
```

```
hist(h_c$AGE, main = "Histogram for age frequency", xlab = "Age Group", ylab = "Frequency of  
Patients", prob= TRUE, col = "red")
```

```
lines(density(h_c$AGE))
```

```
summary(as.factor(h_c$AGE))
```

```
x <- aggregate(TOTCHG~AGE,FUN = sum,data = h_c)
```

```
x
```

```
max(x)
```

```
which.max(summary(as.factor(h_c$APRDRG)))
```

```
diagnosiscost <- aggregate(TOTCHG ~ APRDRG, FUN = sum, data = h_c)
```

```
diagnosiscost
```

```
diagnosiscost[which.max(diagnosiscost$TOTCHG),]
```

```
summary(as.factor(h_c$RACE))
```

```
head(h_c)
```

```
h_c<-na.omit(h_c)
```

```
h_c$RACE<-as.factor(h_c$RACE)
```

```
mod<- aov(TOTCHG ~ RACE, data = h_c)
```

```
mod
```

```
summary(mod)
```

```
summary(h_c$RACE)
```

```
model1 <- lm(TOTCHG ~ AGE + FEMALE, data = h_c)
h_c$FEMALE<-as.factor(h_c$FEMALE)
model1 <- lm(TOTCHG ~ AGE + FEMALE, data = h_c)
summary(model1)
summary(h_c$FEMALE)
head(h_c)

h_c$RACE<-as.factor(h_c$RACE)
model2 <- lm(TOTCHG ~ AGE + FEMALE + RACE, data = h_c)
summary(model2)

model3 <- lm(TOTCHG ~ ., data = h_c)
summary(model3)
```

FUNCTIONS IMPLEMENTED

- **read.csv ():** Use the read.csv () function to import data in CSV format. This function has a number of arguments, but the only essential argument is file, which specifies the location and filename.
- **summary ():** Function is a generic function used to produce result summaries of the results of various model fitting functions. It includes:

**Min. value, 1st Qu. Value, Median value, Mean value, 3rd quartile value,
Max value**

- **hist ():** A histogram represents the frequencies of values of a variable bucketed into ranges. Histogram is similar to bar chart but the difference is it groups the values into continuous ranges. Each bar in histogram represents the height of the number of values present in that range.

R creates histogram using **hist()** function. This function takes a vector as an input and uses some more parameters to plot histograms.

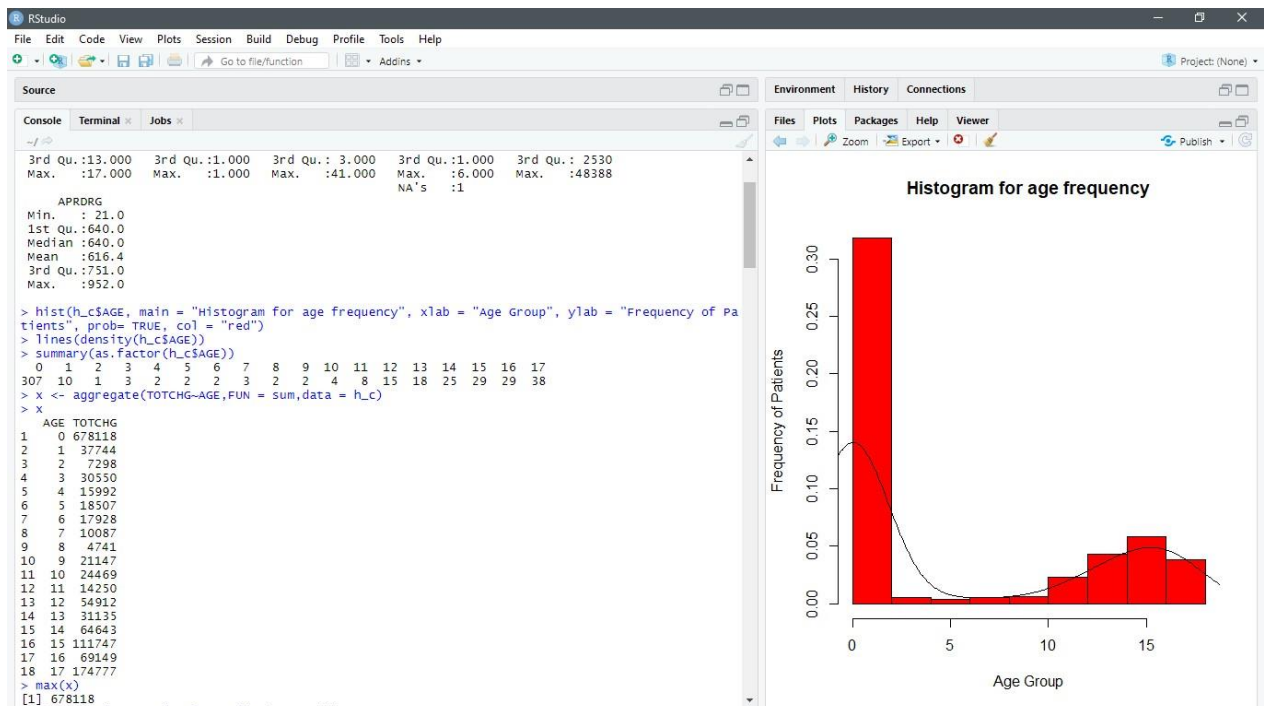
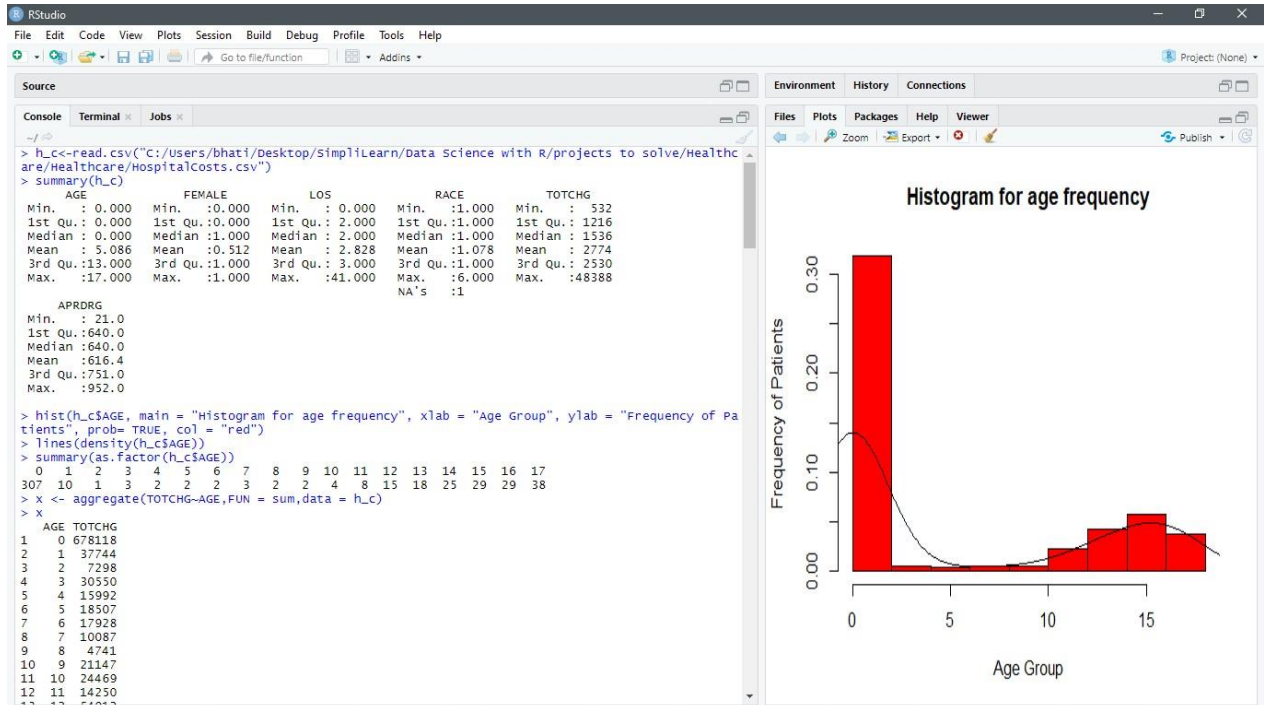
- **lines():** A generic function taking coordinates given in various ways and joining the corresponding points with line segments.
- **aggregate():** Aggregate is a function in base R which can, as the name suggests, aggregate the inputted data.frame d.f by applying a function specified by the FUN parameter to each column of sub-data.frame defined by the by input parameter.
- **max():** Function in R computes the maximum value of a vector or data frame.
- **which.max ():** Which.max returns the position of the element with the maximal value in a vector.

The value of that element can be found with max (...).

- **head():** head() returns the first 6 rows in keeping with the current data.frame convention in R.
- **as.factor():** This function converts a variable into a factor, but preserves variable and value label attributes.
- **Aov ():** aov() is used to summarize the analysis of variance model. The output includes the columns F value and Pr (>F) corresponding to the p-value of the test.
- **na.omit ():** The na. omit R function removes all incomplete cases of a data object (typically of a data frame, matrix or vector).
- **lm ():** It is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although aov may provide a more convenient interface for these).

ANALYSIS

- 1) Record the patient statistics, the agency wants to find the age category of people who frequent the hospital and has the maximum expenditure.

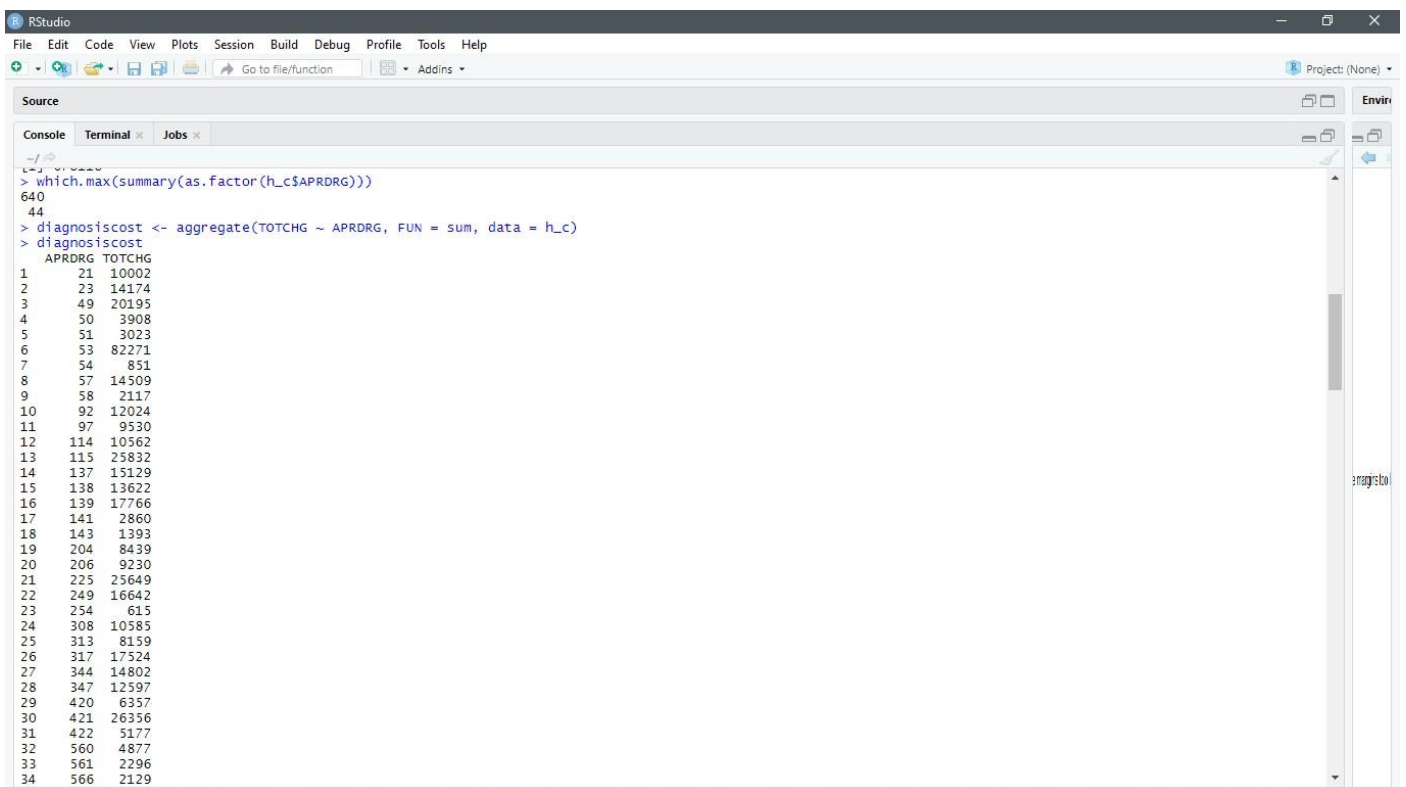


- After executing histogram, we can analyse that age group with most frequency of patients is: 0-1yrs.
- Using `as.factor()` function, allows us to get the count of patients belonging to different age groups.
- Using `aggregate`, we can find out the total charges by each age group. To achieve the total amount of all the people of that age group we provide (`fun = sum`) in `aggregate()` function.
- Hence,

Age group with max visits: 0-1 yrs

Maximum total charges: 678118

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.



```

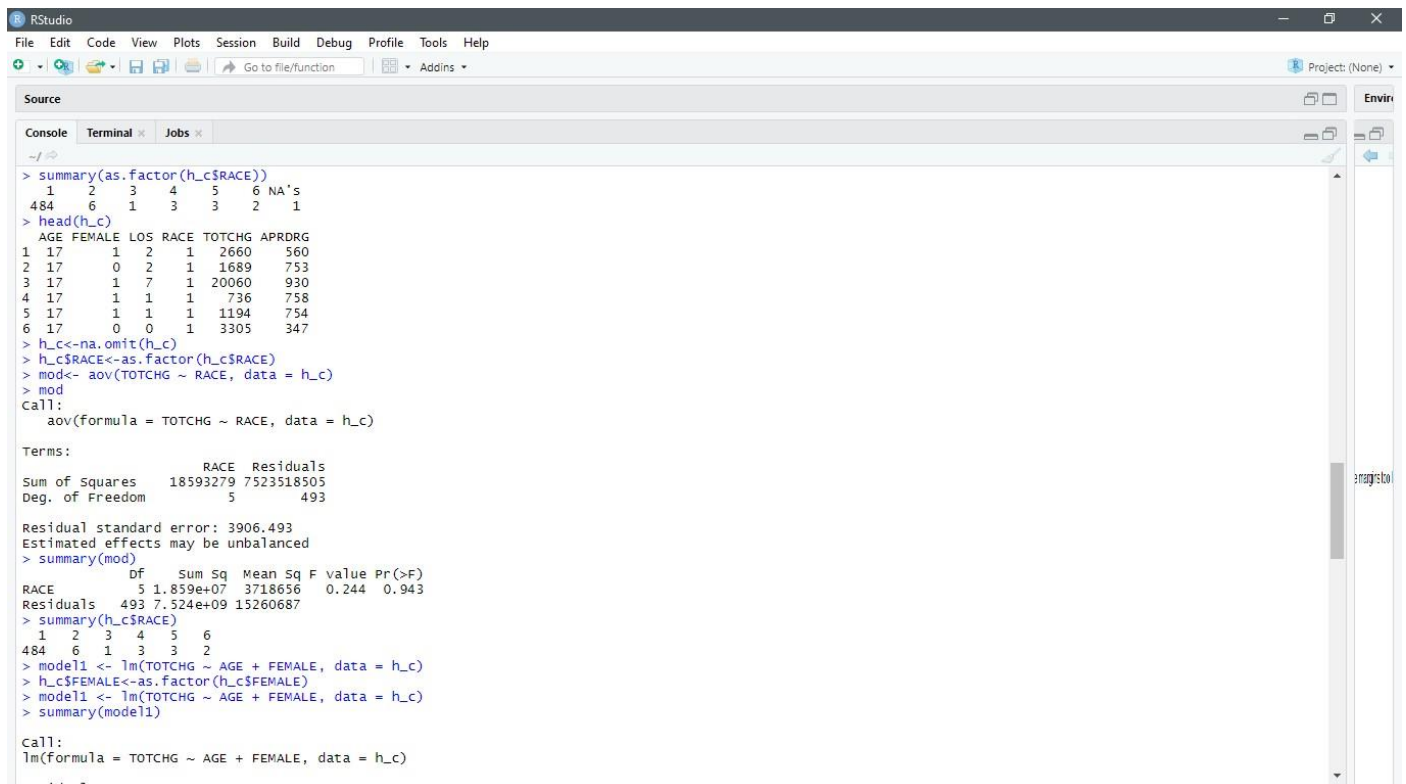
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Project: (None)
Source
Console Terminal Jobs
> which.max(summary(as.factor(h_c$APDRG)))
640
44
> diagnosis$cost <- aggregate(TOTCHG ~ APRDRG, FUN = sum, data = h_c)
> diagnosis$cost
  APRDRG TOTCHG
1      21  10002
2      23  14174
3      49  20195
4      50   3908
5      51   3023
6      53  82271
7      54    851
8      57  14509
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  16642
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

```

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Source
Console Terminal Jobs
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
> diagnosis$cost[which.max(diagnosis$cost$TOTCHG),]
APRDRG TOTCHG
640 437978
> summarise(factor(h_c$PAC))
```

- Using `which.max()` we get the elements with the biggest value in the vector. Also to get the total maximum expenditure from all the diagnosis groups we use `aggregate()` function.
- So, diagnosis group with max value : 640
- Max expenditure from group 640 is : 437978

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.



```
> summary(as.factor(h_c$RACE))
 1    2    3    4    5    6 NA's 
484    6    1    3    3    2     1 
> head(h_c)
  AGE FEMALE LOS RACE TOTCHG APRDRG
1  17      1  2    1  2660    560
2  17      0  2    1  1689    753
3  17      1  7    1  20060   930
4  17      1  1    1   736    758
5  17      1  1    1  1194    754
6  17      0  0    1  3305    347
> h_c<-na.omit(h_c)
> h_c$RACE<-as.factor(h_c$RACE)
> mod<- aov(TOTCHG ~ RACE, data = h_c)
> mod
Call:
aov(formula = TOTCHG ~ RACE, data = h_c)

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

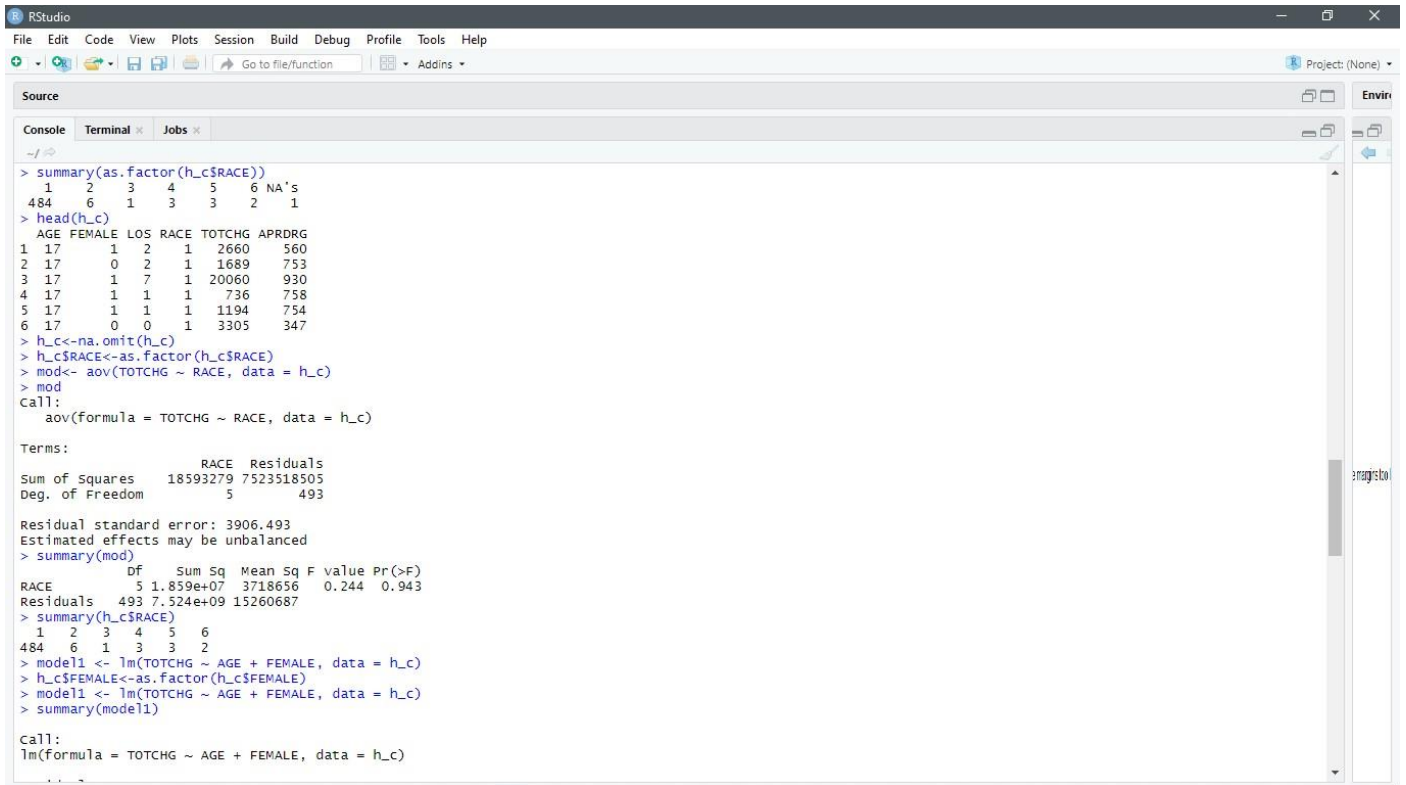
Residual standard error: 3906.493
Estimated effects may be unbalanced
> summary(mod)
          Df Sum Sq Mean Sq F value Pr(>F)
RACE      5 1.859e+07 3718656  0.244  0.943
Residuals 493 7.524e+09 15260687
> summary(h_c$RACE)
 1    2    3    4    5    6 
484    6    1    3    3    2 
> model1 <- lm(TOTCHG ~ AGE + FEMALE, data = h_c)
> h_c$FEMALE<-as.factor(h_c$FEMALE)
> model1 <- lm(TOTCHG ~ AGE + FEMALE, data = h_c)
> summary(model1)

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

- The Residual Value (deviation of the observed value) is very high specifying that there is no relation between the race of patient and the hospital cost.
- From the summary we can also see 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.

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 perform this analysis, we use Linear Regression model.



```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal Jobs
> summary(as.factor(h_c$RACE))
 1    2    3    4    5    6 NA's
484    6    1    3    3    2    1
> head(h_c)
  AGE FEMALE LOS RACE TOTCHG APRDRG
1  17      1  2    1  2660    560
2  17      0  2    1  1689    753
3  17      1  7    1 20060    930
4  17      1  1    1   736    758
5  17      1  1    1  1194    754
6  17      0  0    1  3305    347
> h_c<-na.omit(h_c)
> h_c$RACE<-as.factor(h_c$RACE)
> mod<- aov(TOTCHG ~ RACE, data = h_c)
> mod
Call:
aov(formula = TOTCHG ~ RACE, data = h_c)

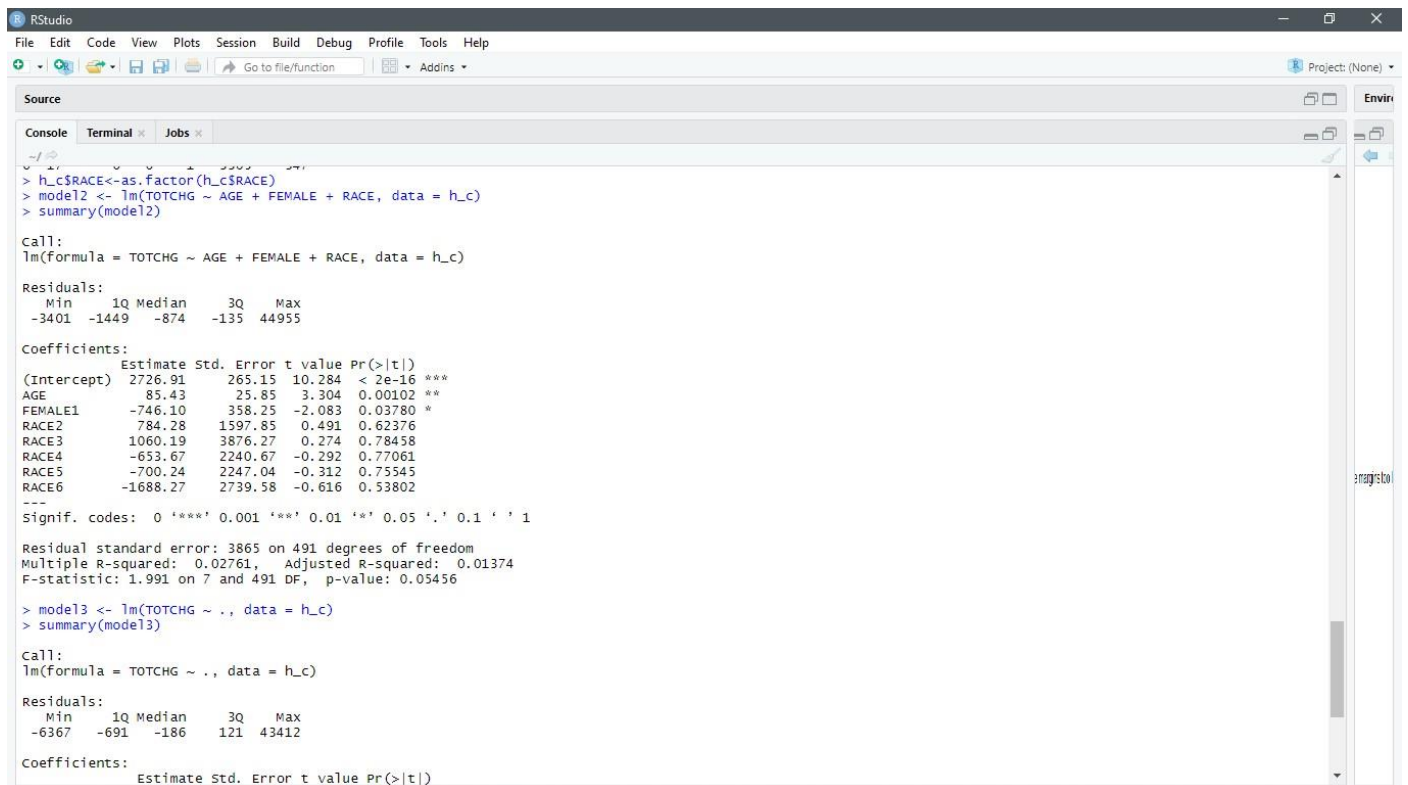
Terms:
RACE
Sum of Squares    18593279 7523518505
Deg. of Freedom      5      493

Residual standard error: 3906.493
Estimated effects may be unbalanced
> summary(mod)
              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(h_c$RACE)
 1    2    3    4    5    6
484    6    1    3    3    2
> model1 <- lm(TOTCHG ~ AGE + FEMALE, data = h_c)
> h_c$FEMALE<-as.factor(h_c$FEMALE)
> model1 <- lm(TOTCHG ~ AGE + FEMALE, data = h_c)
> summary(model1)

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

- Age is a very important factor in the hospital costs as seen by the significance levels and p-values.
- The gender also seems to have an impact.
- There is an equal number of male and female patients.
- Based the negative coefficient we can conclude that females incur lesser cost than males.

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.



```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Project: (None)
Source
Console Terminal Jobs
~/...
> h_c$RACE<-as.factor(h_c$RACE)
> model2 <- lm(TOTCHG ~ AGE + FEMALE + RACE, data = h_c)
> summary(model2)

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

Residuals:
    Min       1Q   Median       3Q      Max
-3401  -1449   -874   -135  44955

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2726.91    265.15   10.284 < 2e-16 ***
AGE           85.43     25.85    3.304  0.00102 **
FEMALE1     -746.10    358.25   -2.083  0.03780 *
RACE2        784.28    1597.85    0.491  0.62376
RACE3       1060.19    3876.27    0.274  0.78458
RACE4       -653.67    2240.67   -0.292  0.77061
RACE5       -700.24    2247.04   -0.312  0.75545
RACE6      -1688.27    2739.58   -0.616  0.53802
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3865 on 491 degrees of freedom
Multiple R-squared:  0.02761, Adjusted R-squared:  0.01374
F-statistic: 1.991 on 7 and 491 DF, p-value: 0.05456

> model3 <- lm(TOTCHG ~ ., data = h_c)
> summary(model3)

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

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

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

- The significance codes are almost null for all the variables, except for the intercept.
- The p-value high which signifies that there is no linear relationship between the given variables.
- Hence we cannot predict the length of stay of the patients based on the age, gender, and race.

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

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Project: (None)
Source
Console Terminal Jobs
~/
RACE6 -1688.27 2739.58 -0.616 0.53802
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3865 on 491 degrees of freedom
Multiple R-squared: 0.02761, Adjusted R-squared: 0.01374
F-statistic: 1.991 on 7 and 491 DF, p-value: 0.05456

> model3 <- lm(TOTCHG ~ ., data = h_c)
> summary(model3)

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

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 ***
FEMALE1     -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
APDRG       -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

> ?read.csv()
> ?read.csv()
> |

```

- Based on the output we can see that the Age and Length of stay affects the total Hospital cost.
- Cost is directly proportional to the Length i.e. higher the Length of stay of patients will result to higher hospital cost.
- As per the output we can see that with an increase of 1 day stay, the hospital cost will increase by 742.

**THANK
YOU**

