Prepared by Indu Voleti

Analyze the Healthcare Cost

1. 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 healthcare costs and their utilization.

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

Dataset downloaded from the URL mentioned below:

https://lms.simplilearn.com/courses/2716/Data-Science-with-R/assessment

2. Goals:

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 the 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 hospital costs.

3. R-Code:

```
hospitalcost <- read.csv("C:/Users/manish/Desktop/Simplilearn/Healthcare/HospitalCosts.csv")
view(hospitalcost)
head(hospitalcost)
summary(hospitalcost)
hist(hospitalcost$AGE)
summary(as.factor(hospitalcost$AGE))
aggregate(TOTCHG ~AGE, FUN = sum, data = hospitalcost)
max(aggregate(TOTCHG ~AGE, FUN = sum, data = hospitalcost))
diagnosiscost<-aggregate(TOTCHG ~APRDRG, FUN = sum, data = hospitalcost)
summary(diagnosiscost)
diagnosiscost[which.max(diagnosiscost$TOTCHG),]
summary(as.factor(hospitalcost$RACE))
hospitalcost<-na.omit(hospitalcost)
head(hospitalcost)
summary(as.factor(hospitalcost$RACE))
modelannova<-aov(TOTCHG~RACE, data = hospitalcost)
summary(modelannova)
model1<-lm(TOTCHG~AGE+FEMALE,data = hospitalcost)
summary(model1)
model2<-lm(LOS~AGE+FEMALE+RACE, data = hospitalcost)
summary(model2)
model3 <- lm(TOTCHG ~., data = hospitalcost)
summary(model3)
```

4. Analysis:

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.

Solutions:

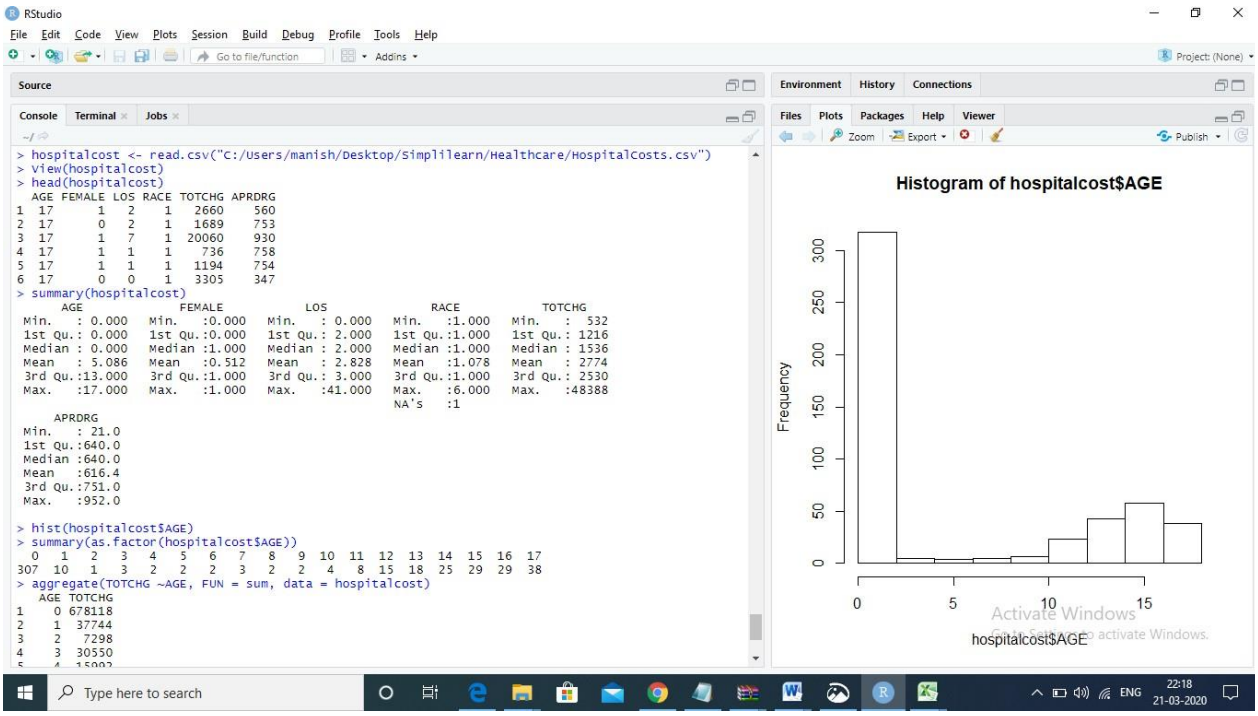
A histogram is graphical analysis and being used to find the number of occurrences of each age category that has the highest frequency of hospital visit.

The as.factor() is used to make sure that the categories are not treated as numbers.

Result:

Based on output, it is observed that frequent visit to hospital & maximum expenditure is by infant of 0 age 678118

Output:



Cont.

```
> hist(hospitalcost$AGE)
> summary(as.factor(hospitalcost$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 = hospitalcost)
   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 = hospitalcost))
[1] 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.

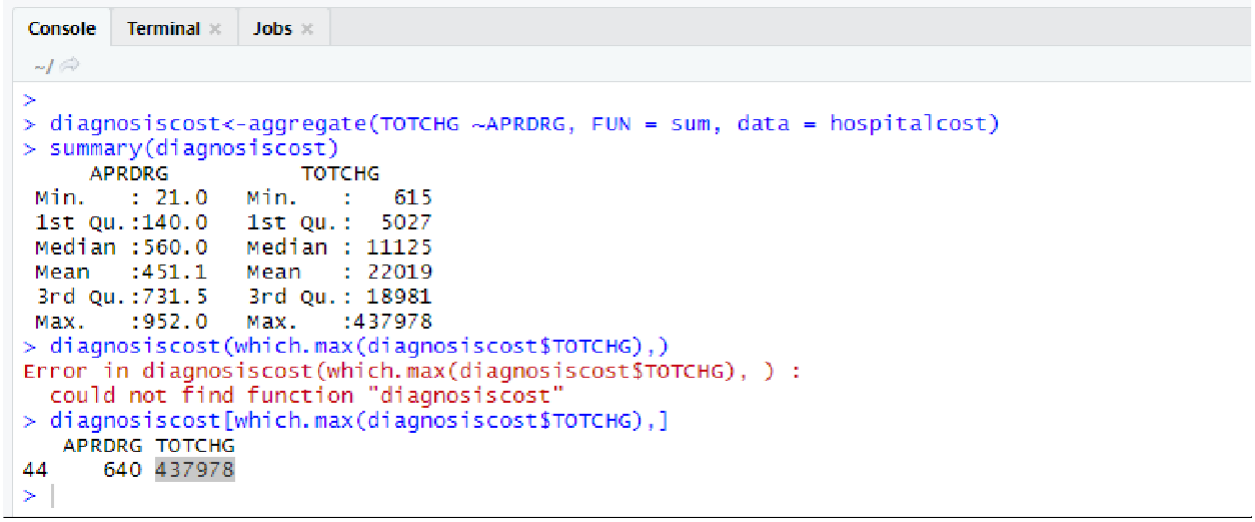
Solutions:

Aggregate ()-used to create a data frame and displaying the result on the basis of given formula which.max() is being used to identify the maximum charges against diagnosis-related group

Result:

As per below results, it is observed that APRDRG 640 has maximum hospitalization & highest total hospitalization cost 437978

Output:



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.

Solutions:

Annova test is being used to analyse the Race wise cost occurred.
Ho: The race of the patient is related to the hospitalization costs.
H1: No relation

Result:

p-value observed to be very high 68% which means we can take risk and reject the null hypothesis
Hence, there is no relation between the race of patient and the hospital cost.

Output:

```
Console   Terminal ×   Jobs ×
~/
summary"
> summary(as.factor(hospitalcost$RACE))
   1    2    3    4    5    6 NA's
 484    6    1    3    3    2    1
>
>
>
> summary(as.factor(hospitalcost$RACE))
   1    2    3    4    5    6 NA's
 484    6    1    3    3    2    1
> hospitalcost<-na.omit(hospitalcost)
> head(hospitalcost)
  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
> summary(as.factor(hospitalcost$RACE))
  1   2   3   4   5   6
484   6   1   3   3   2
> modelannova<-aov(TOTCHG~RACE)
Error in eval(predvars, data, env) : object 'TOTCHG' not found
> modelannova<-aov(TOTCHG~RACE, data = hospitalcost)
> summary(modelannova)
             Df    Sum Sq  Mean Sq F value Pr(>F)
RACE          1 2.488e+06  2488459   0.164  0.686
Residuals   497 7.540e+09 15170268
>
```

4. 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.

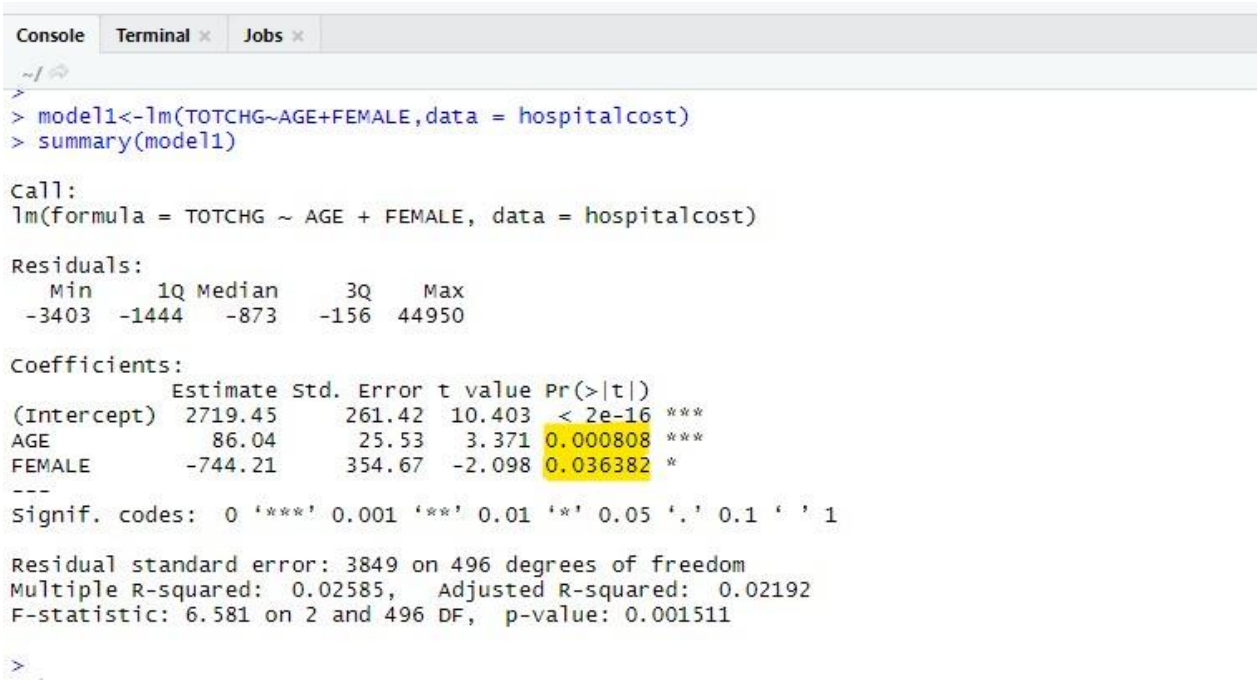
Solutions:

Linear Regression model to be used here.

Result:

As per below output, Age is important factor in hospital cost as seen by the significance levels and p-values
Also, p-values for gender are less, which means it is also having impact on cost and same with intercept

Output:



```
Console   Terminal   Jobs
~/
>
> model1<-lm(TOTCHG~AGE+FEMALE,data = hospitalcost)
> summary(model1)

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

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

>
```

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.

Solutions:

Linear Regression model to be used here.

Result:

p-values is observed to be very high which signifies that there is no linear relationship between given variables. Hence, we cannot predict length of stay of patients based on age, gender & race.

Output:

Console Terminal Jobs

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

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

Residuals:
   Min     1Q Median     3Q    Max
 -3.22  -1.22  -0.85   0.15  37.78

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.94377    0.39318   7.487 3.25e-13 ***
AGE         -0.03960    0.02231  -1.775   0.0766 .
FEMALE       0.37011    0.31024   1.193   0.2334
RACE        -0.09408    0.29312  -0.321   0.7484
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.363 on 495 degrees of freedom
Multiple R-squared:  0.007898,  Adjusted R-squared:  0.001886
F-statistic: 1.314 on 3 and 495 DF,  p-value: 0.2692

>
```

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

Solutions:

Linear Regression model to be used here.

Result:

It is observed that age & length of stay affects hospital cost.
APRDRG also affects hospital cost.

Output:

Console Terminal Jobs

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

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

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
Multiple R-squared:  0.5536,    Adjusted R-squared:  0.5491
F-statistic: 122.3 on 5 and 493 DF,  p-value: < 2.2e-16
```