Analyze the Healthcare cost and Utilization in Wisconsin hospitals

Business Analytic Foundation with R Tools- Solutions

HEALTHCARE COST ANALYSIS

Business Objective:
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.

Data : Hospital.csv
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

Question 1 : People with maximum expenditure age group
To record the patient statistics, the agency wants to find the age category of people who frequent the hospital and has the maximum expenditure.

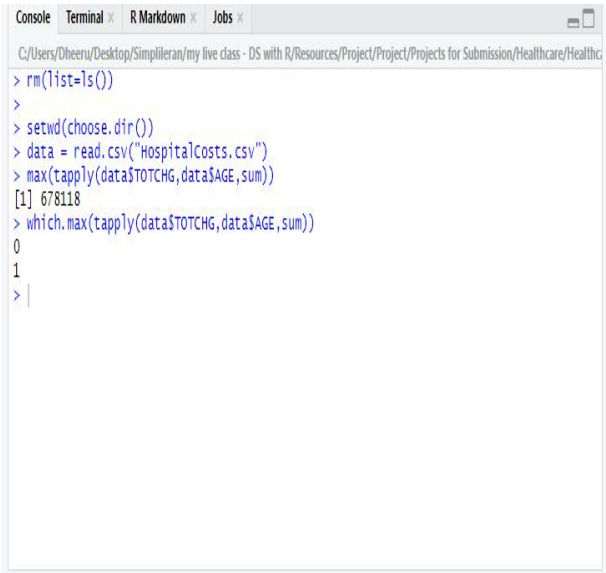
CODE

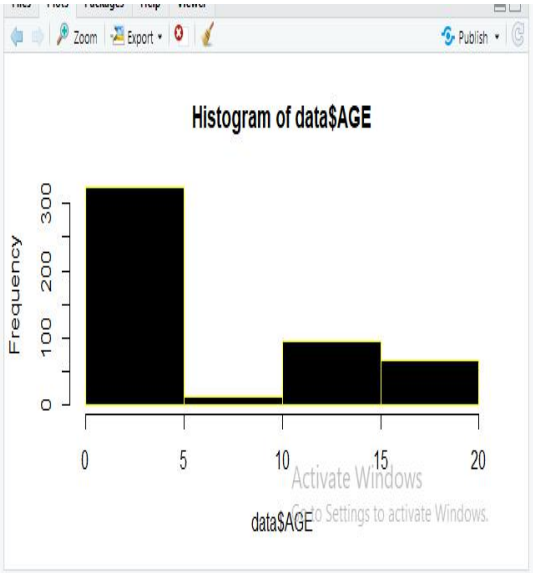
```
rm(list=ls())

setwd(choose.dir())

data = read.csv("HospitalCosts.csv")

View(data)

head(data)

str(data)

data$FEMALE = as.factor(data$FEMALE)
data$RACE=as.factor(data$RACE)
data$APRDRG=as.factor(data$APRDRG)

data$age_bins <- ifelse((data$AGE < 1), "infant",
ifelse(data$AGE < 3, 'toddler',
ifelse(data$AGE < 11, 'child',
'adolescent')))

str(data)

hist(data$AGE,breaks = 3,col = "black",border = "yellow")


max(tapply(data$TOTCHG,data$AGE,sum))
  the maximum charge is found out to be 678118

which.max(tapply(data$TOTCHG,data$AGE,sum))
```

RESULT

1) From Here we can arrive at the conclusion that children between the age group of (0-5) are the most frequent patients in the hospital. the next age group with high but comparatively lower hospital visits than (0-5) are patients of age group(10-15)

2) This shows that the maximum cost of 678118 is predominantly for children of age 0-1(new born babies or 0-12months).



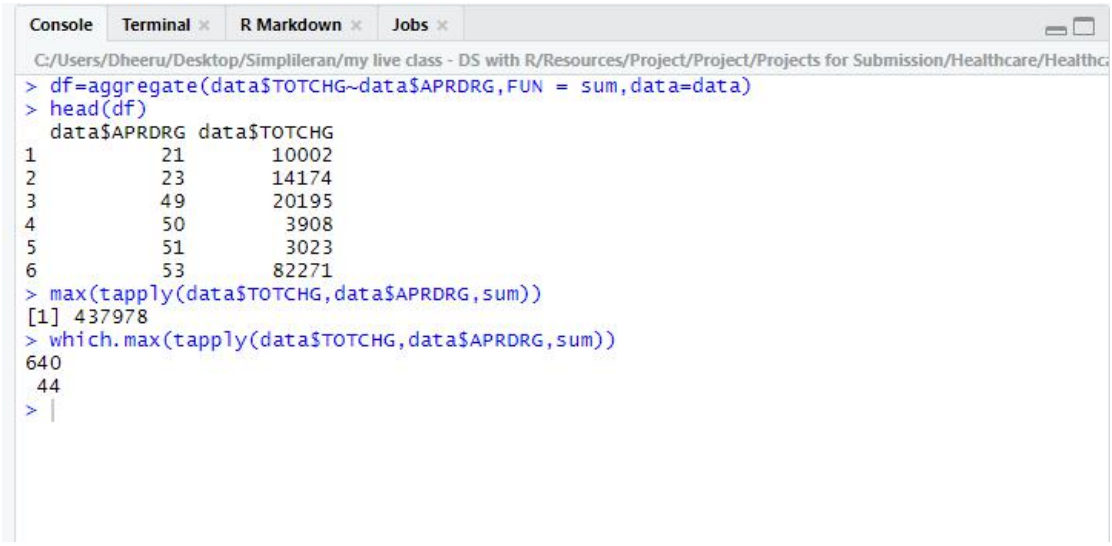


Question 2:
In order of severity of the diagnosis and treatments and to find out the expensive treatments, the agency wants to find the diagnosis related group that has maximum hospitalization and expenditure.

CODE

```
df=aggregate(data$TOTCHG~data$APRDRG,FUN = sum,data=data)
df

max(tapply(data$TOTCHG,data$APRDRG,sum))
437978

which.max(tapply(data$TOTCHG,data$APRDRG,sum))
640
```

RESULT:
based on this we found that group 640 had the maximum expenditure of 437978



Question 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.

Hypothesis Testing
H0:Race had no imapct on cost
H1:Race had impact on cost
Here we are trying to see whether cost(numeric) is affected by race(categorical)
In such a case we use anova for testing hypothesis

CODE

```
colSums(is.na(data))

data=na.omit(data)

colSums(is.na(data))

model = aov(data$TOTCHG~data$RACE,data = data)

summary(model)

alpha = 0.05
```

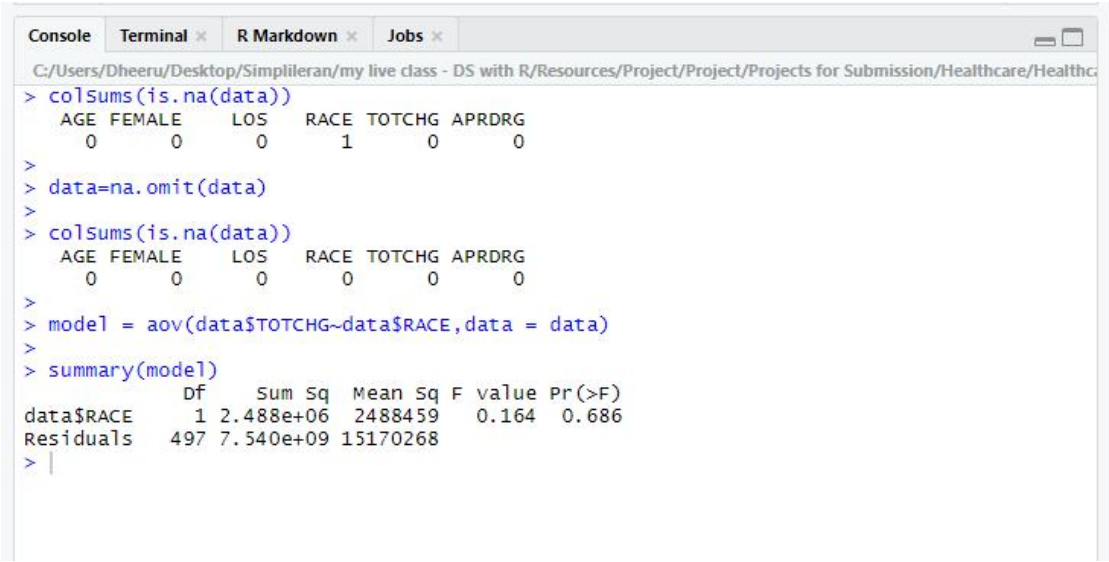
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data$RACE	5	1.859e+07	3718656	0.244	0.943
Residuals	493	7.524e+09	15260687		

p=0.943

RESULT:

As p>alpha we accept the null hypothesis

We can say that race had no impact on cost



Question 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.

FEMALE - Categorical

Cost - Continuous

age_bins - Categorical

Define Hypothesis

H0:Age and Gender have no imapct on cost

H1:Age and Gender have impact on cost

CODE:

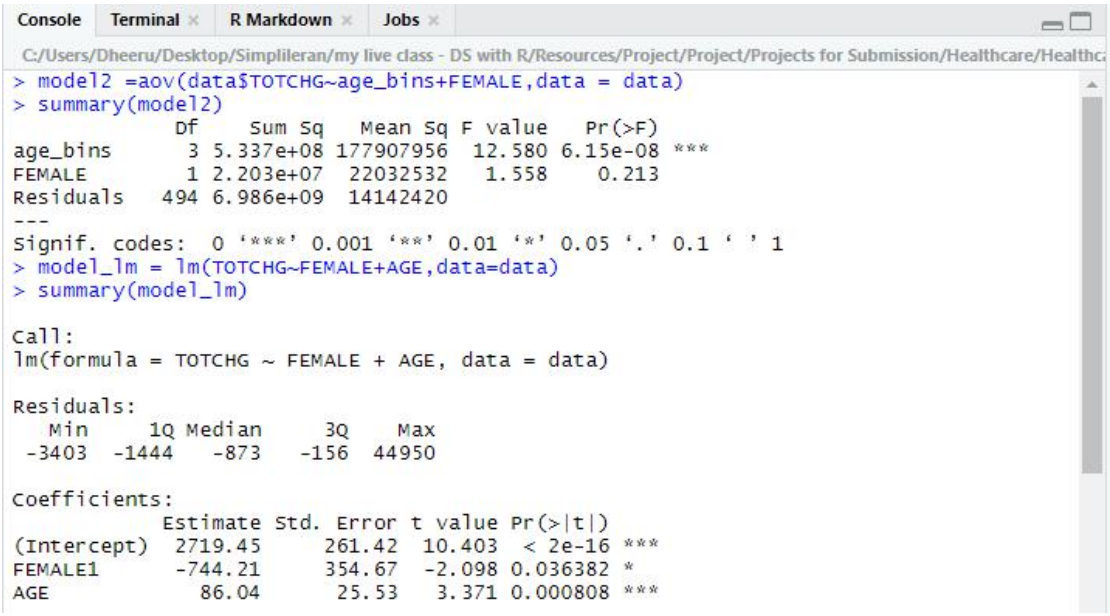
```
model2 =aov(data$TOTCHG~age_bins+FEMALE,data = data)
summary(model2)


model_lm = lm(TOTCHG~FEMALE+AGE,data=data)
summary(model_lm)

Male
TOTCGH_M=2719.45+(86.04*0)+(-744.21*0)
TOTCGH_M

Female
TOTCGH_F=2719.45+(86.04*0)+(-744.21*1)
TOTCGH_F
```

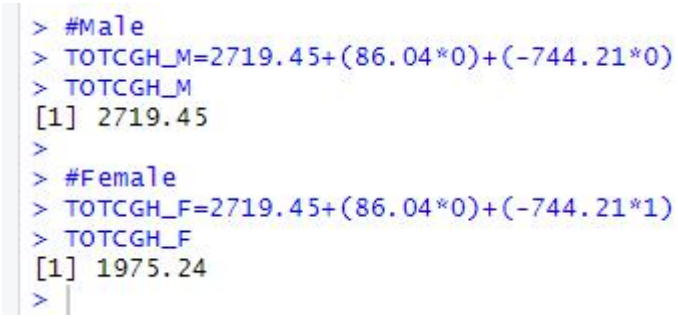
RESULT:

p_age=8.28e-07 there is impact of age on cost
p_gender=0.208 there is no imapct of gender on cost

we had to futher refine the relation using linear regression equation

we can see that cost from males is higher than that of female
for a given age group
Female =1 patient is female
=0 Patient is male





Question 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.

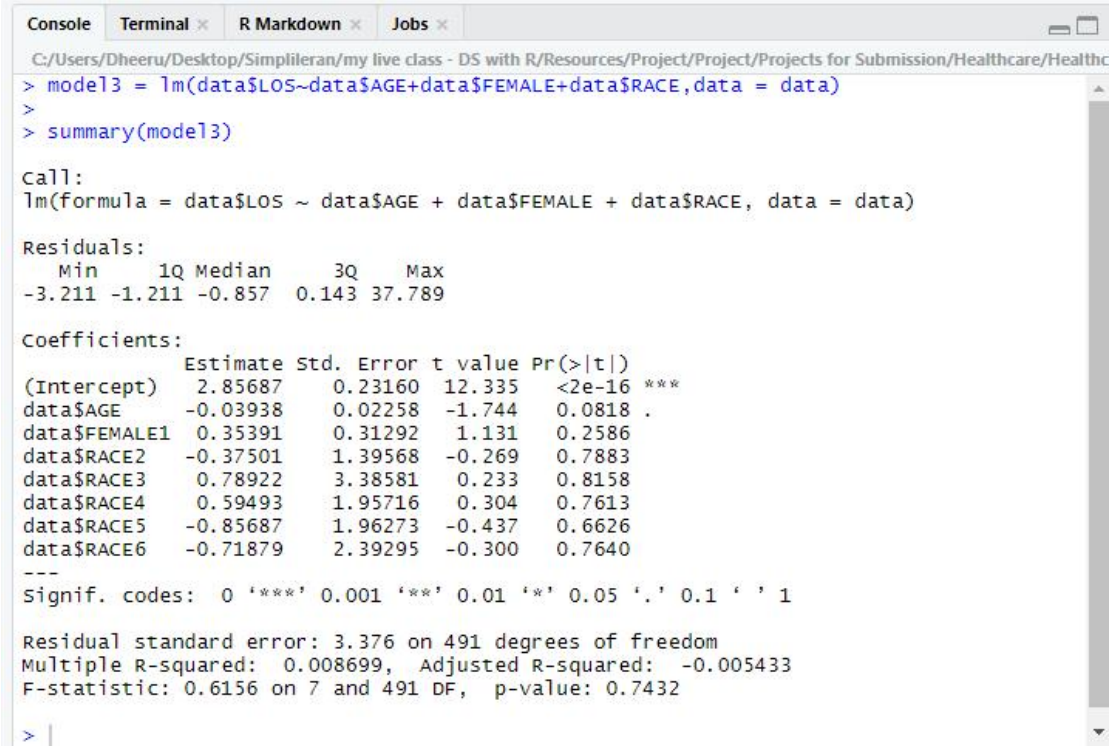
predicting length of stay LOS(numerical and dependent) and and three independent variables like AGE,RACE,FEMALE

CODE:

```
model3 = lm(data$LOS~data$AGE+data$FEMALE+data$RACE,data = data)

summary(model3)
```

RESULT:

We can see with very high intercept Significance we can say that these are not adequate features for predicting LOS and also the r-squared value is too low.



Question 6:

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

CODE:

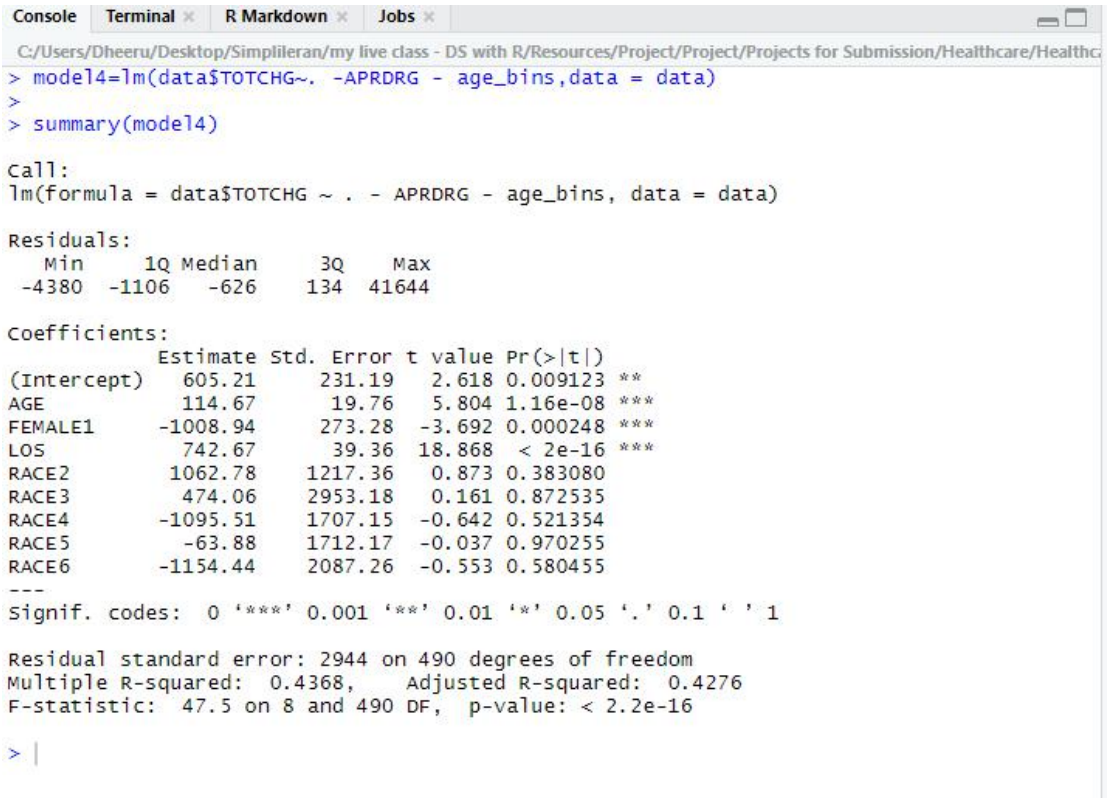
```
model4=lm(data$TOTCHG~. -APRDRG - age_bins,data = data)

summary(model4)


model5=lm(TOTCHG~AGE+LOS+APRDRG,data = data)
```

summary(model5)

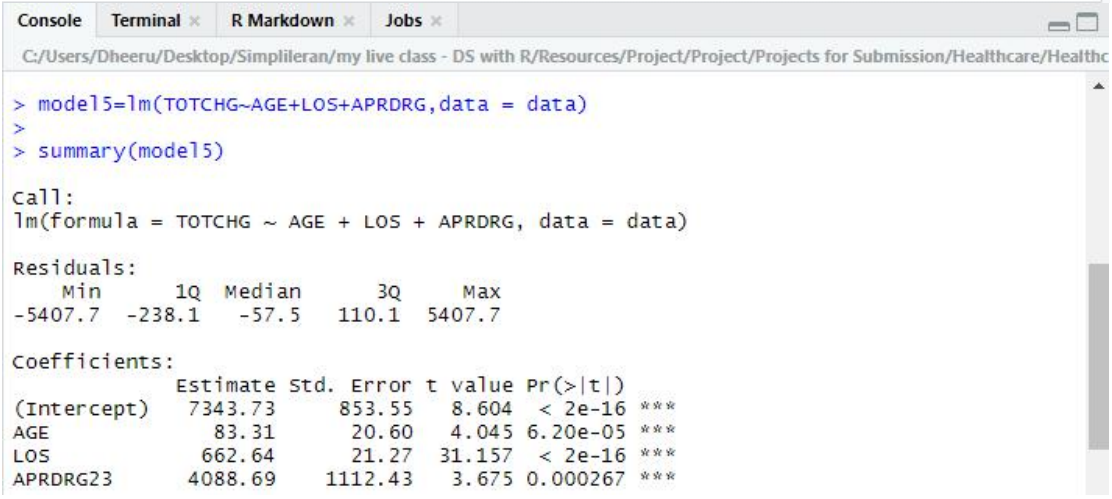
RESULT :

1) R-square and f-stats is low for model4.
2) We can see that AGE , APRDG and LOS have high significance so we can create a model using these three.
3) We have impoved our f-statistic and R-square way better than the earlier one

so we can use AGE , LOS and APRDRG are best suited to predict the total charge for a patient



```
Console  Terminal  R Markdown  Jobs
C:/Users/Dheeru/Desktop/Simplileran/my live class - DS with R/Resources/Project/Project/Projects for Submission/Healthcare/Healthc
> model4=lm(data$TOTCHG~. -APRDRG - age_bins,data = data)
>
> summary(model4)

Call:
lm(formula = data$TOTCHG ~ . - APRDRG - age_bins, data = data)

Residuals:
   Min     1Q Median     3Q    Max
 -4380  -1106   -626    134  41644

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   605.21     231.19   2.618 0.009123 **
AGE           114.67      19.76   5.804 1.16e-08 ***
FEMALE1     -1008.94     273.28  -3.692 0.000248 ***
LOS           742.67      39.36  18.868  < 2e-16 ***
RACE2        1062.78    1217.36   0.873 0.383080
RACE3         474.06    2953.18   0.161 0.872535
RACE4       -1095.51    1707.15  -0.642 0.521354
RACE5         -63.88    1712.17  -0.037 0.970255
RACE6       -1154.44    2087.26  -0.553 0.580455
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2944 on 490 degrees of freedom
Multiple R-squared:  0.4368,    Adjusted R-squared:  0.4276
F-statistic:  47.5 on 8 and 490 DF,  p-value: < 2.2e-16

>
```



```
Console  Terminal  R Markdown  Jobs
C:/Users/Dheeru/Desktop/Simplileran/my live class - DS with R/Resources/Project/Project/Projects for Submission/Healthc
> model5=lm(TOTCHG~AGE+LOS+APRDRG,data = data)
>
> summary(model5)

Call:
lm(formula = TOTCHG ~ AGE + LOS + APRDRG, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5407.7  -238.1   -57.5   110.1  5407.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   7343.73     853.55   8.604  < 2e-16 ***
AGE             83.31      20.60   4.045 6.20e-05 ***
LOS            662.64      21.27  31.157  < 2e-16 ***
APRDRG23      4088.69    1112.43   3.675 0.000267 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 786.5 on 434 degrees of freedom
Multiple R-squared:  0.9644,    Adjusted R-squared:  0.9592
F-statistic: 183.7 on 64 and 434 DF,  p-value: < 2.2e-16
```