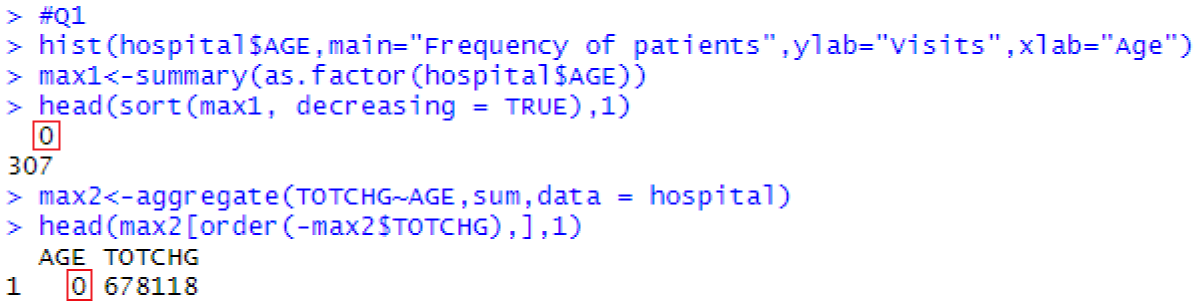
HEALTHCARE COST ANALYSYS

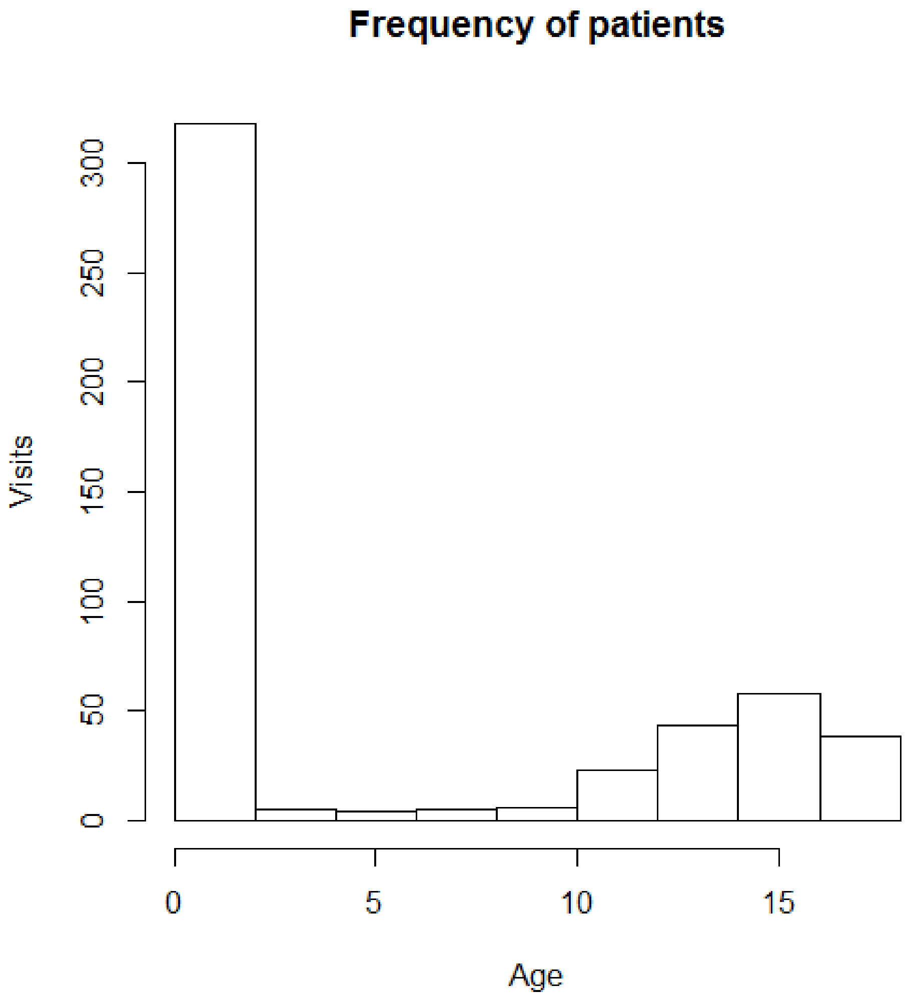
Domain: Healthcare

Solution:

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.



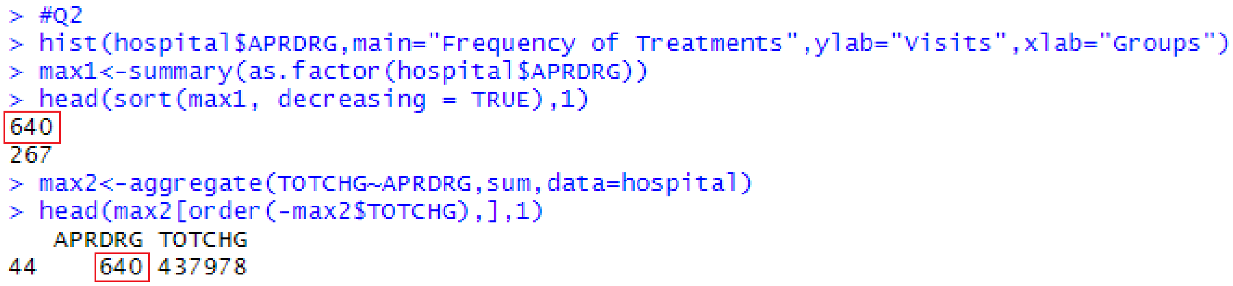
```
> #Q1
> hist(hospital$AGE,main="Frequency of patients",ylab="Visits",xlab="Age")
> max1<-summary(as.factor(hospital$AGE))
> head(sort(max1, decreasing = TRUE),1)
  0
307
> max2<-aggregate(TOTCHG~AGE,sum,data = hospital)
> head(max2[order(-max2$TOTCHG),],1)
  AGE TOTCHG
1   0 678118
```



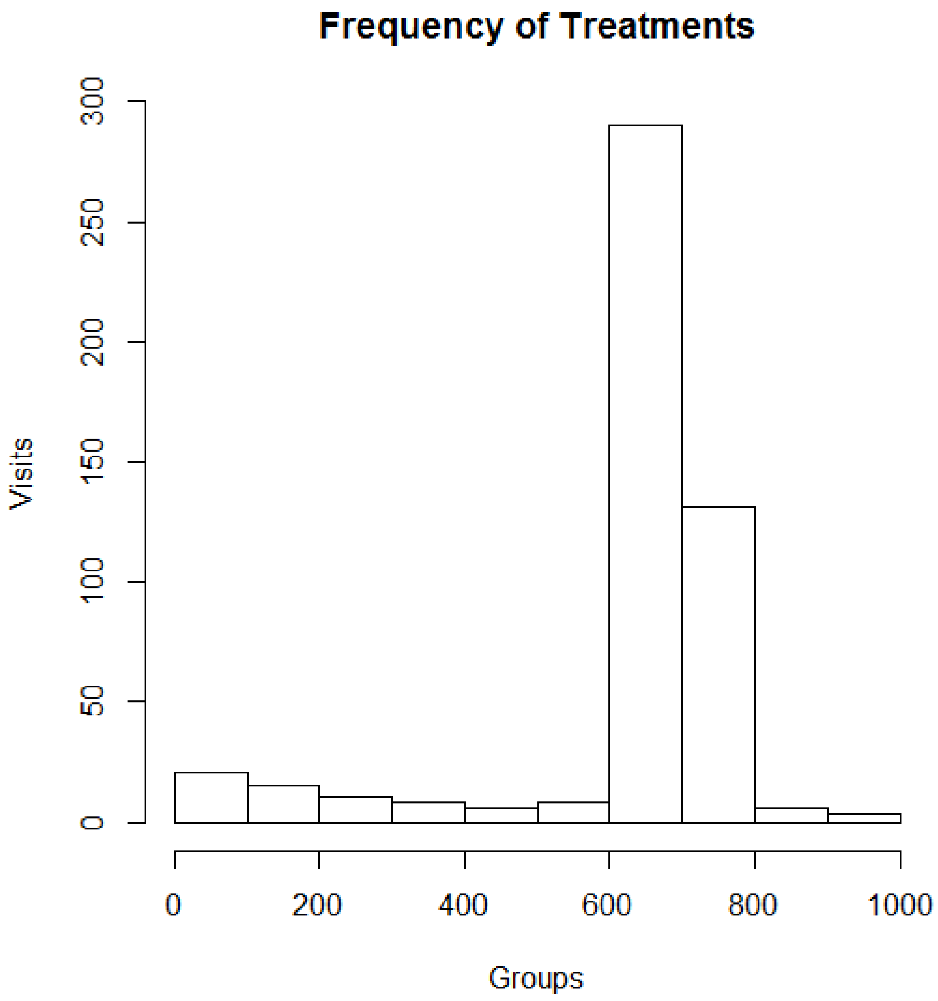
From the histogram above we can see that children below age 5 have the highest frequency of visit.

The comparison of max of summary() and aggregate() function leads us to the conclusion that children of age **0-1** frequent the hospital and has the max expenditure. So it can be interpreted that visit and cost are directly proportional

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.



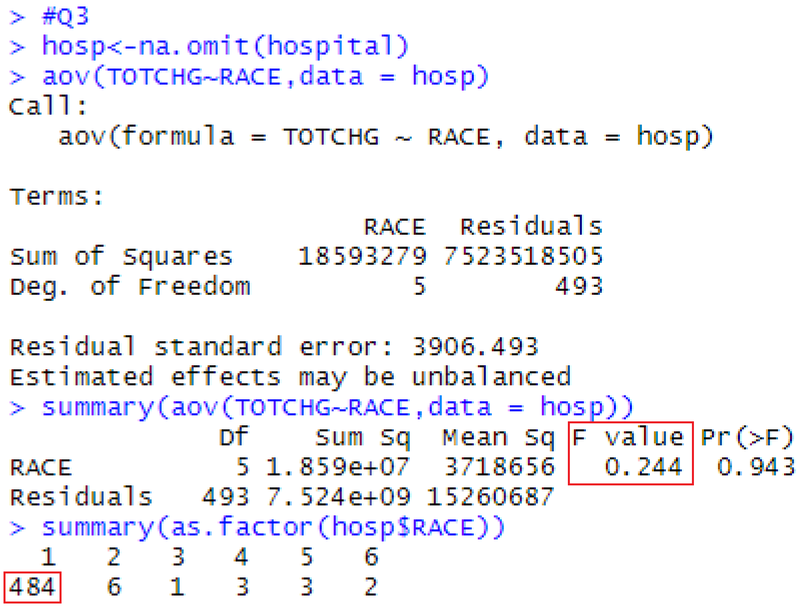
```
> #Q2
> hist(hospital$APRDRG,main="Frequency of Treatments",ylab="Visits",xlab="Groups")
> max1<-summary(as.factor(hospital$APRDRG))
> head(sort(max1, decreasing = TRUE),1)
640
267
> max2<-aggregate(TOTCHG~APRDRG,sum,data=hospital)
> head(max2[order(-max2$TOTCHG),],1)
   APRDRG TOTCHG
44    640 437978
```



From the histogram above we can see that the diagnostic-related group which falls between 600 - 800 has the highest frequency of visit.

The comparison of max of summary() and aggregate() function leads us to the conclusion that patients falling under diagnostic-related group **640** 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.

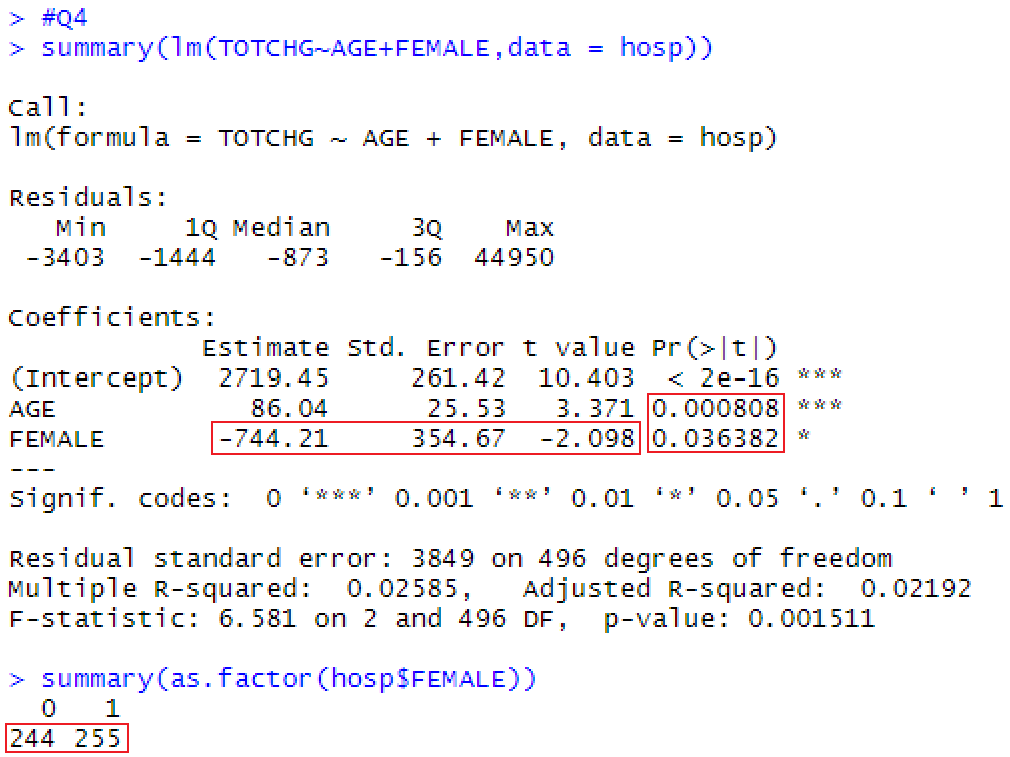


```
> #Q3
> hosp<-na.omit(hospital)
> aov(TOTCHG~RACE,data = hosp)
Call:
   aov(formula = TOTCHG ~ RACE, data = hosp)

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

Residual standard error: 3906.493
Estimated effects may be unbalanced
> summary(aov(TOTCHG~RACE,data = hosp))
             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(as.factor(hosp$RACE))
  1   2   3   4   5   6
484   6   1   3   3   2
```

The output of the ANOVA test shows a very low F value implying that variation with respect to race is very less, and a very high P value implying that **cost and race are independent**.

However as majority of the observations belongs to RACE – 1 this prediction may not be accurate.

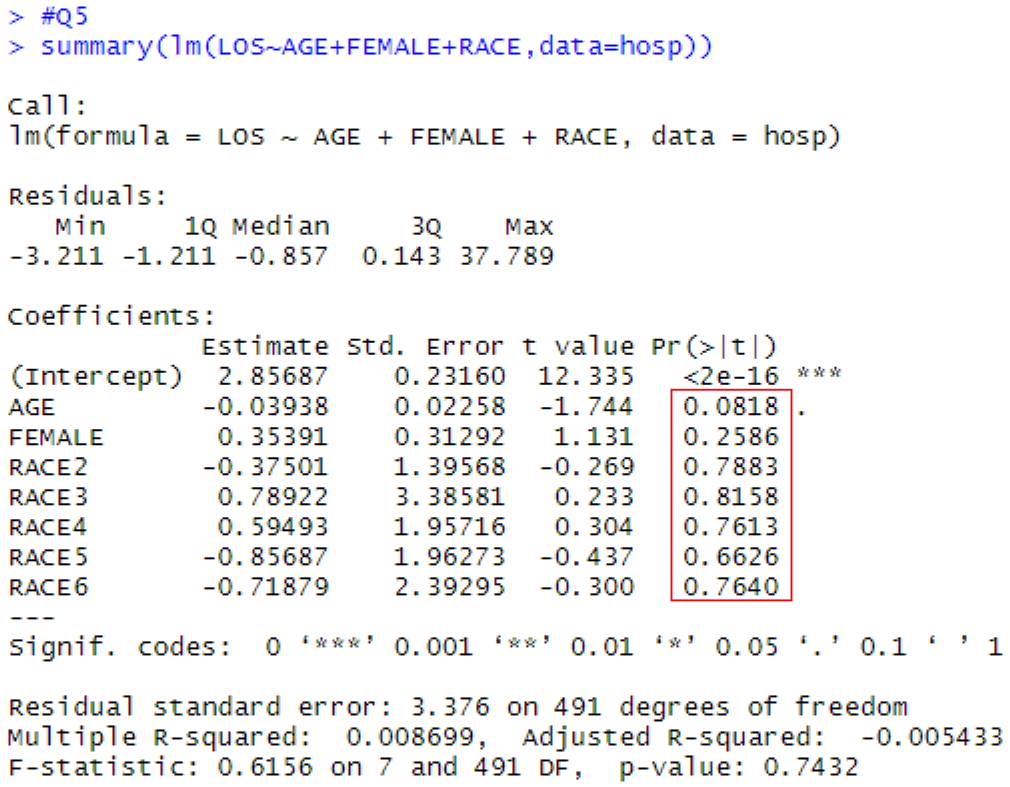
4. To properly utilize the costs, the agency has to analyze the severity of the hospital costs by age and gender for the proper allocation of resources.



```
> #Q4
> summary(lm(TOTCHG~AGE+FEMALE,data = hosp))

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

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

> summary(as.factor(hosp$FEMALE))
  0   1 
244 255 
```

If we compare the P values obtained from the Linear Regression Model we can say that age has more weightage than gender. We can also see that there are almost equal males and females, and the coefficient of female being negative means that the **cost for hospitalization for females is less than males**.

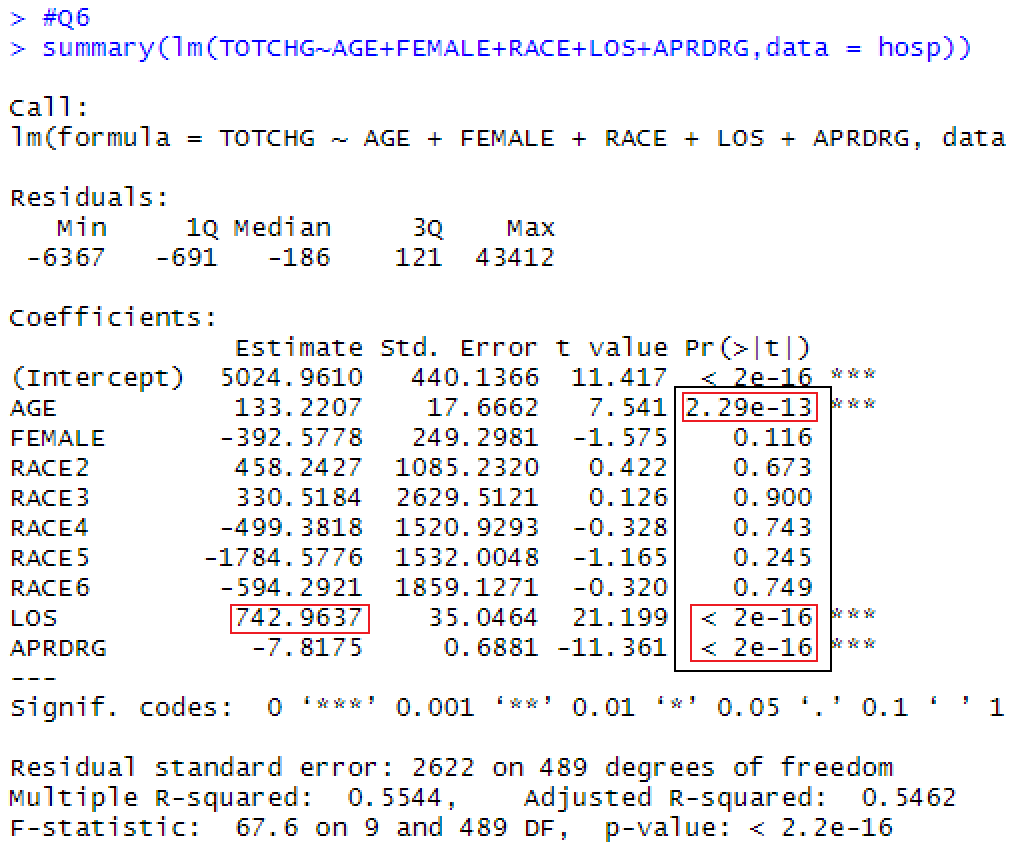
5. Since the length of stay is the crucial factor for inpatients, the agency wants to find if the length of stay can be predicted from age, gender, and race.



```
> #Q5
> summary(lm(LOS~AGE+FEMALE+RACE,data=hosp))

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

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
```

Since all the P values of the independent variables are high there exists no linear relationship among them, therefore we are **unable to predict length of stay from age, gender, and race**.

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



```
> #Q6
> summary(lm(TOTCHG~AGE+FEMALE+RACE+LOS+APRDRG,data = hosp))

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

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    
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    
LOS           742.9637    35.0464  21.199  < 2e-16 ***
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
```

By looking at the P values we can see that **age, length of stay and diagnostic-related group affect the hospital costs**, further we can see the positive coefficient of length of stay implying that each increase in LOS increases the TOTCHG by 742 in value.

****