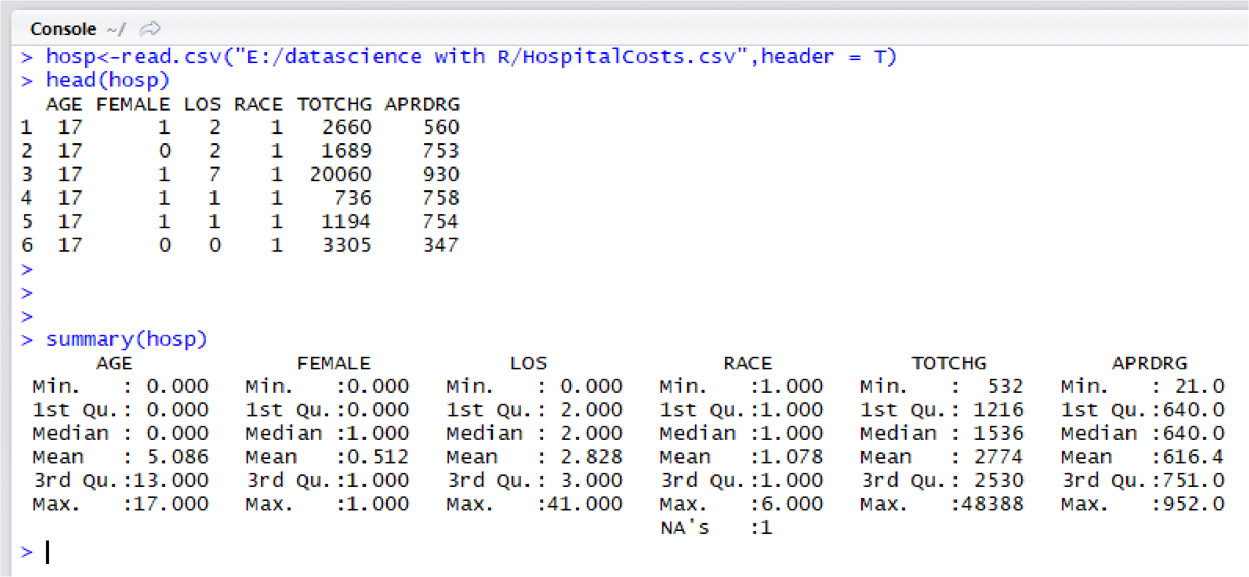
Analyze the Healthcare cost and Utilization in Wisconsin hospitals

Loading the dataset:

Code:

```
hosp<-read.csv("E:/datascience with R/HospitalCosts.csv",header = T)

head(hosp)

summary(hosp)
```

output:



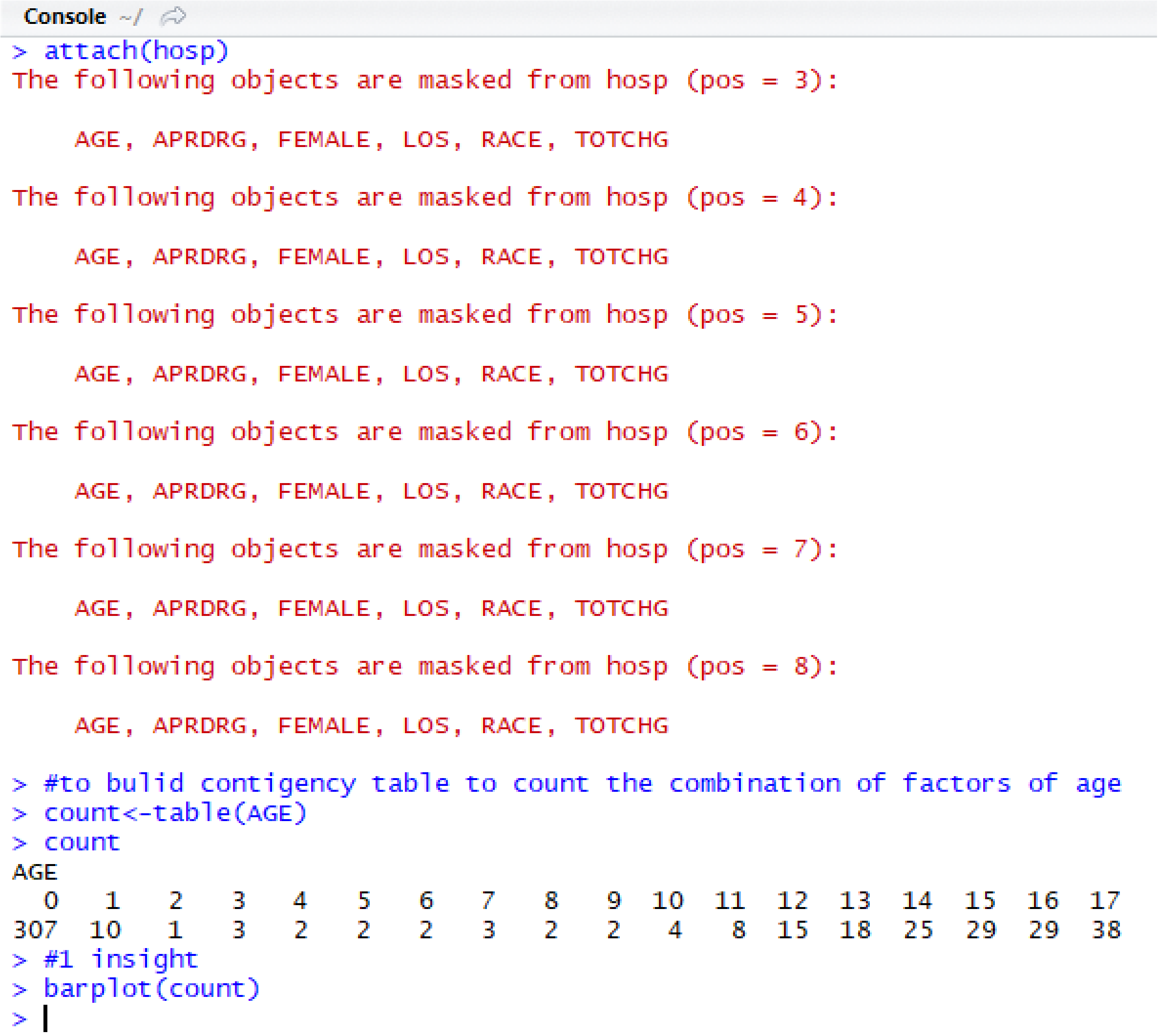
```
> hosp<-read.csv("E:/datascience with R/HospitalCosts.csv",header = T)
> head(hosp)
  AGE FEMALE LOS RACE TOTCHG APRDRG
1  17      1   2    1   2660    560
2  17      0   2    1   1689    753
3  17      1   7    1  20060    930
4  17      1   1    1    736    758
5  17      1   1    1   1194    754
6  17      0   0    1   3305    347
>
>
>
> summary(hosp)
      AGE              FEMALE           LOS              RACE           TOTCHG          APRDRG     
 Min.   : 0.000   Min.   :0.000   Min.   : 0.000   Min.   :1.000   Min.   :  532   Min.   : 21.0  
 1st Qu.: 0.000   1st Qu.:0.000   1st Qu.: 2.000   1st Qu.:1.000   1st Qu.: 1216   1st Qu.:640.0  
 Median : 0.000   Median :1.000   Median : 2.000   Median :1.000   Median : 1536   Median :640.0  
 Mean   : 5.086   Mean   :0.512   Mean   : 2.828   Mean   :1.078   Mean   : 2774   Mean   :616.4  
 3rd Qu.:13.000   3rd Qu.:1.000   3rd Qu.: 3.000   3rd Qu.:1.000   3rd Qu.: 2530   3rd Qu.:751.0  
 Max.   :17.000   Max.   :1.000   Max.   :41.000   Max.   :6.000   Max.   :48388   Max.   :952.0  
                                                   NA's   :1                                      
> |
```

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

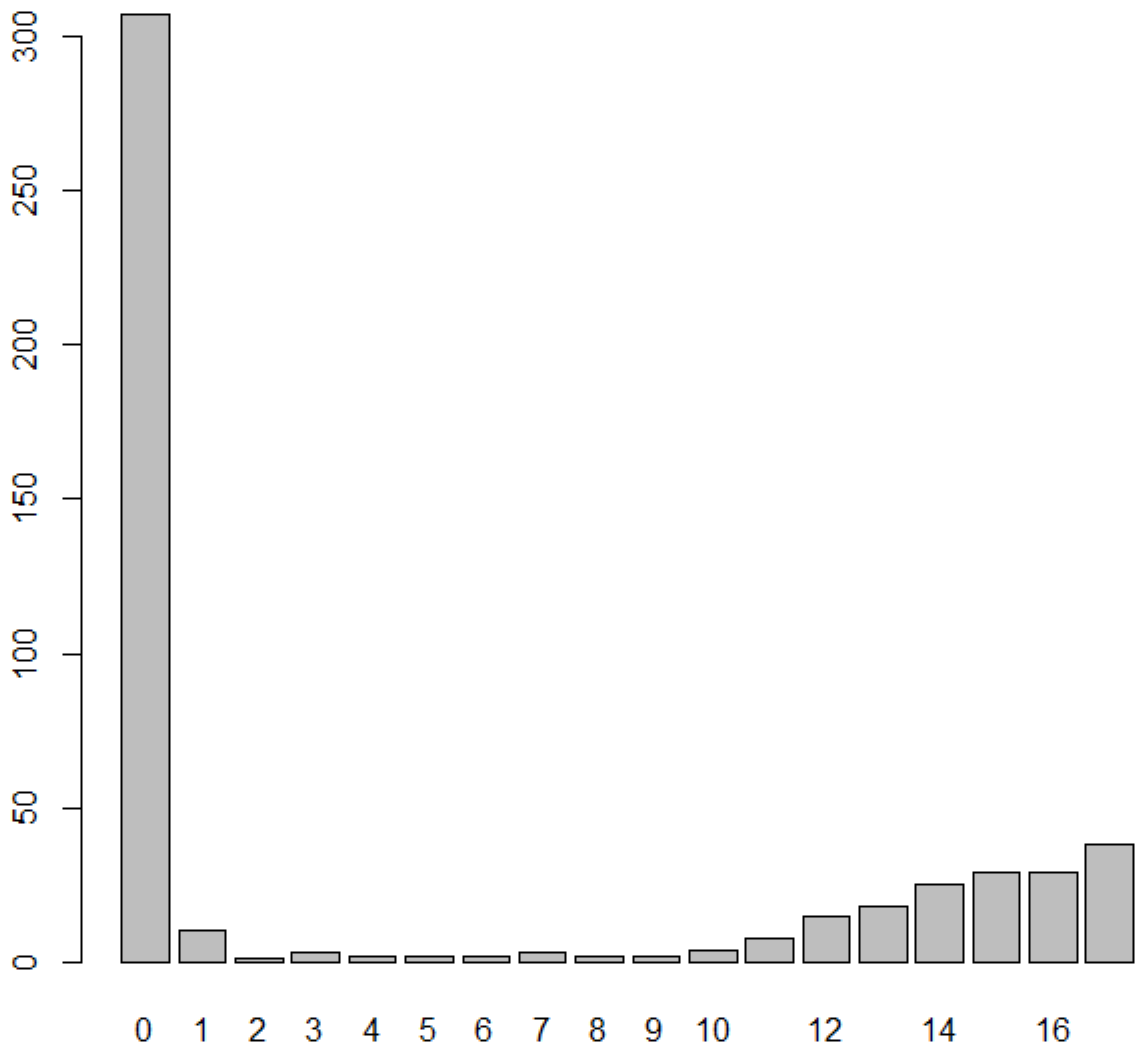
Code:

```
attach(hosp)

#to bulid contigency table to count the combination of factors of age

count<-table(AGE)

count

#1 insight

barplot(count)
```

Output:



```
> attach(hosp)
The following objects are masked from hosp (pos = 3):

    AGE, APRDRG, FEMALE, LOS, RACE, TOTCHG

The following objects are masked from hosp (pos = 4):

    AGE, APRDRG, FEMALE, LOS, RACE, TOTCHG

The following objects are masked from hosp (pos = 5):

    AGE, APRDRG, FEMALE, LOS, RACE, TOTCHG

The following objects are masked from hosp (pos = 6):

    AGE, APRDRG, FEMALE, LOS, RACE, TOTCHG

The following objects are masked from hosp (pos = 7):

    AGE, APRDRG, FEMALE, LOS, RACE, TOTCHG

The following objects are masked from hosp (pos = 8):

    AGE, APRDRG, FEMALE, LOS, RACE, TOTCHG

> #to bulid contigency table to count the combination of factors of age
> count<-table(AGE)
> count
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 
> #1 insight
> barplot(count)
> 
```



->#from the above barplot,we find that 0-1 agecategory people visits the hospital frequently.

1b)

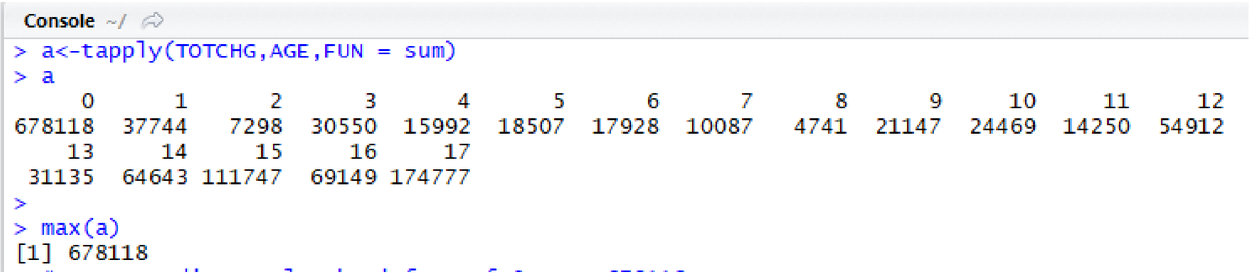
#to find the infant category has maximum hospital costs

Code:

a<-tapply(TOTCHG,AGE,FUN = sum)

a

max(a)



```
> a<-tapply(TOTCHG,AGE,FUN = sum)
> a
     0      1      2      3      4      5      6      7      8      9     10     11     12 
678118  37744   7298  30550  15992  18507  17928  10087   4741  21147  24469  14250  54912 
    13     14     15     16     17 
 31135  64643 111747  69149 174777 
> 
> max(a)
[1] 678118
```

#max expenditure also by infant of 0 age =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.

Code:

APRDRG1<-as.factor(APRDRG)

summary(APRDRG1)

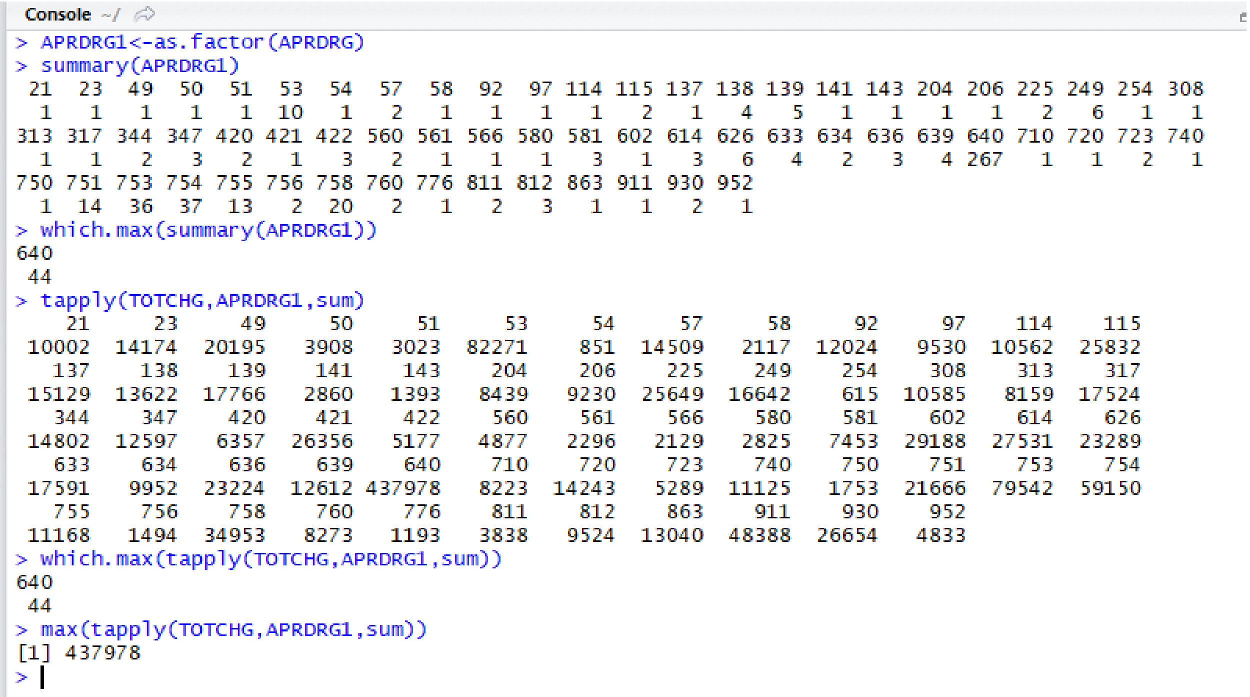
which.max(summary(APRDRG1))

tapply(TOTCHG,APRDRG1,sum)

which.max(tapply(TOTCHG,APRDRG1,sum))

max(tapply(TOTCHG,APRDRG1,sum))

Output:



```
> APRDRG1<-as.factor(APRDRG)
> summary(APRDRG1)
 21  23  49  50  51  53  54  57  58  92  97 114 115 137 138 139 141 143 204 206 225 249 254 308 
  1   1   1   1   1  10   1   2   1   1   1   1   2   1   4   5   1   1   1   1   2   6   1   1 
313 317 344 347 420 421 422 560 561 566 580 581 602 614 626 633 634 636 639 640 710 720 723 740 
  1   1   2   3   2   1   3   2   1   1   1   3   1   3   6   4   2   3   4 267   1   1   2   1 
750 751 753 754 755 756 758 760 776 811 812 863 911 930 952 
  1  14  36  37  13   2  20   2   1   2   3   1   1   2   1 
> which.max(summary(APRDRG1))
640 
 44 
> tapply(TOTCHG,APRDRG1,sum)
    21     23     49     50     51     53     54     57     58     92     97    114    115 
 10002  14174  20195   3908   3023  82271    851  14509   2117  12024   9530  10562  25832 
   137    138    139    141    143    204    206    225    249    254    308    313    317 
 15129  13622  17766   2860   1393   8439   9230  25649  16642    615  10585   8159  17524 
   344    347    420    421    422    560    561    566    580    581    602    614    626 
 14802  12597   6357  26356   5177   4877   2296   2129   2825   7453  29188  27531  23289 
   633    634    636    639    640    710    720    723    740    750    751    753    754 
 17591   9952  23224  12612 437978   8223  14243   5289  11125   1753  21666  79542  59150 
   755    756    758    760    776    811    812    863    911    930    952 
 11168   1494  34953   8273   1193   3838   9524  13040  48388  26654   4833 
> which.max(tapply(TOTCHG,APRDRG1,sum))
640 
 44 
> max(tapply(TOTCHG,APRDRG1,sum))
[1] 437978
> |
```

#From the results we can see that the category 640 has the maximum entries of hospitalization

and also has the highest total hospitalization cost (437978).

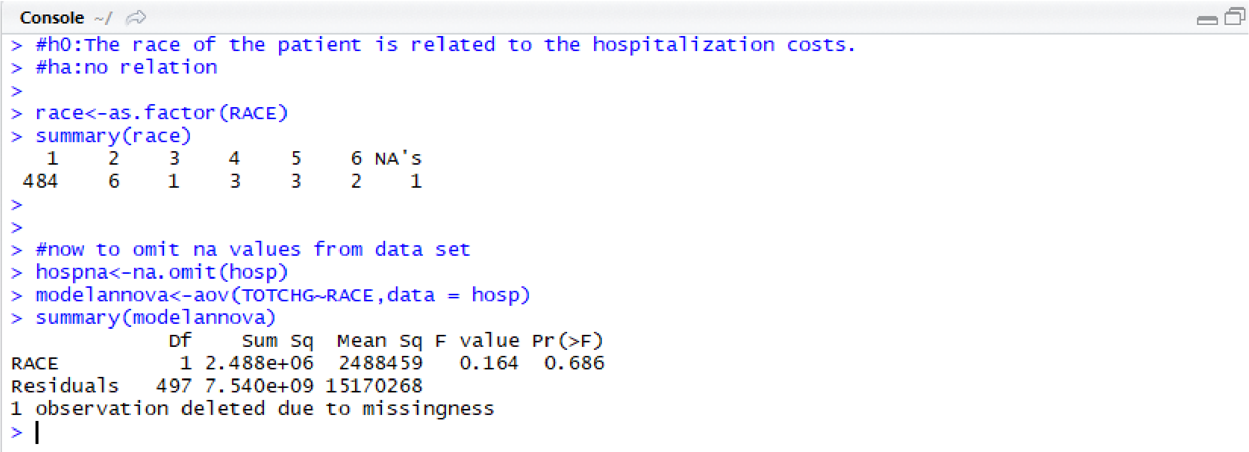
3.To make sure that there is no malpractice, the agency needs to analyze if the race of the patient is related to the hospitalization costs.

Code:

#h0:The race of the patient is related to the hospitalization costs.

```
#ha:no relation


race<-as.factor(RACE)

summary(race)



#now to omit na values from data set

hospna<-na.omit(hosp)

modelannova<-aov(TOTCHG~RACE,data = hosp)

summary(modelannova)
```

Output:



```
> #h0:The race of the patient is related to the hospitalization costs.
> #ha:no relation
> 
> race<-as.factor(RACE)
> summary(race)
   1    2    3    4    5    6 NA's 
 484    6    1    3    3    2    1 
> 
> 
> #now to omit na values from data set
> hospna<-na.omit(hosp)
> modelannova<-aov(TOTCHG~RACE,data = hosp)
> 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               
1 observation deleted due to missingness
> |
```

#pvalue comes out to be very high 68% this means we can take risk and reject the null hypothesis

#this means there is no relation between the race of patient and the hospital cost.

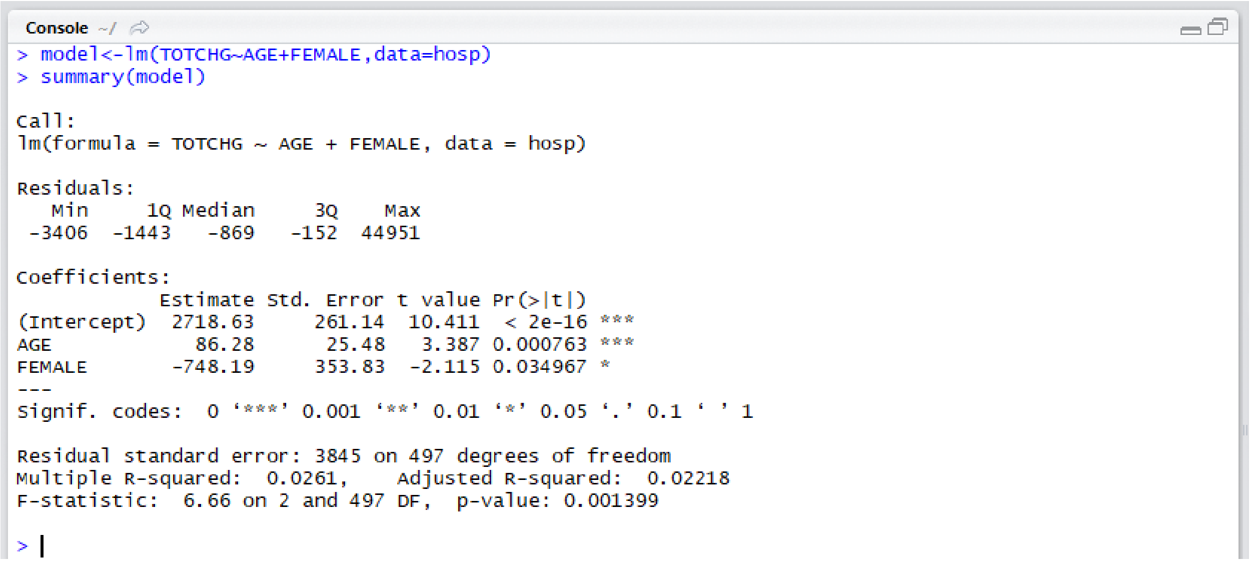
4.To properly utilize the costs, the agency has to analyze the severity of the hospital costs by age and gender for proper allocation of resources.

Code:

model<-lm(TOTCHG~AGE+FEMALE,data=hosp)

summary(model)

Output:



```
> model<-lm(TOTCHG~AGE+FEMALE,data=hosp)
> summary(model)

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

Residuals:
   Min     1Q Median     3Q    Max 
 -3406  -1443   -869   -152  44951 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2718.63     261.14  10.411  < 2e-16 ***
AGE            86.28      25.48   3.387 0.000763 ***
FEMALE       -748.19     353.83  -2.115 0.034967 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3845 on 497 degrees of freedom
Multiple R-squared:  0.0261,	Adjusted R-squared:  0.02218 
F-statistic:  6.66 on 2 and 497 DF,  p-value: 0.001399

> |
```

#pvalue for age is very less this means it is a important factor in the hospital costs as seen by the significance levels and p-values

#gender has also less p value means it is also having the impact on cost and same with intercept.

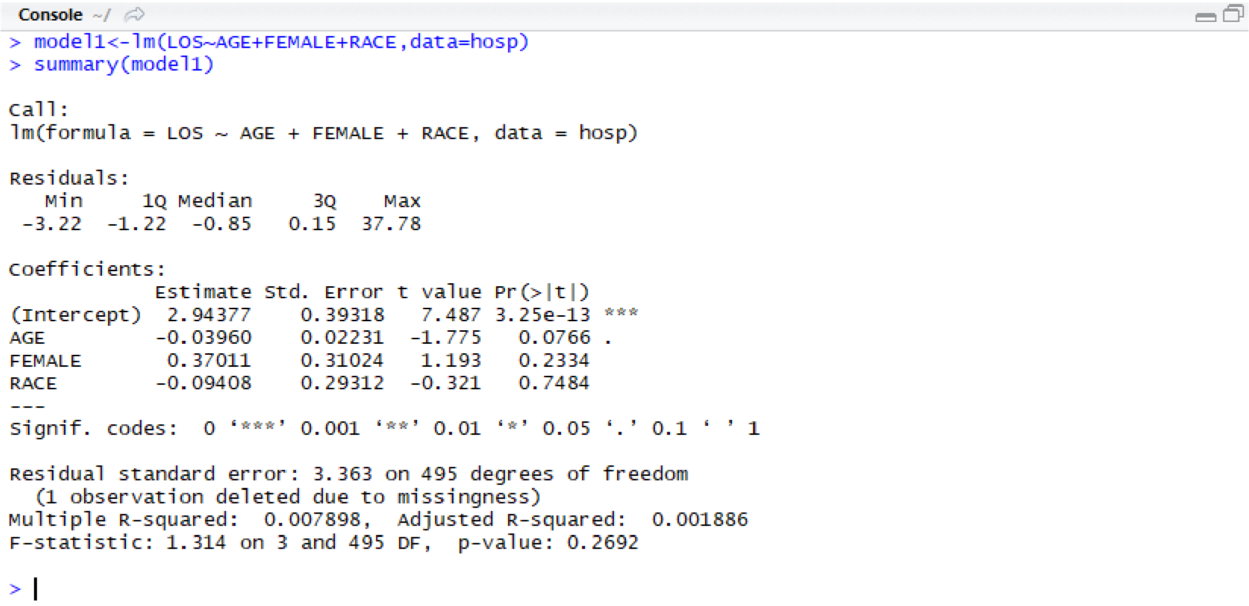
5. Since the length of stay is the crucial factor for inpatients, the agency wants to find if the length of stay can be predicted from age, gender, and race.

Code:

model1<-lm(LOS~AGE+FEMALE+RACE,data=hosp)

summary(model1)

Output:



```
> model1<-lm(LOS~AGE+FEMALE+RACE,data=hosp)
> summary(model1)

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

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
  (1 observation deleted due to missingness)
Multiple R-squared:  0.007898,	Adjusted R-squared:  0.001886 
F-statistic: 1.314 on 3 and 495 DF,  p-value: 0.2692

> |
```

#except for the intercept.

#The very high p-value signifies that there is no linear relationship between the given variables.

#That is, with just the age, gender, and race, it is not possible to predict the los of a patient

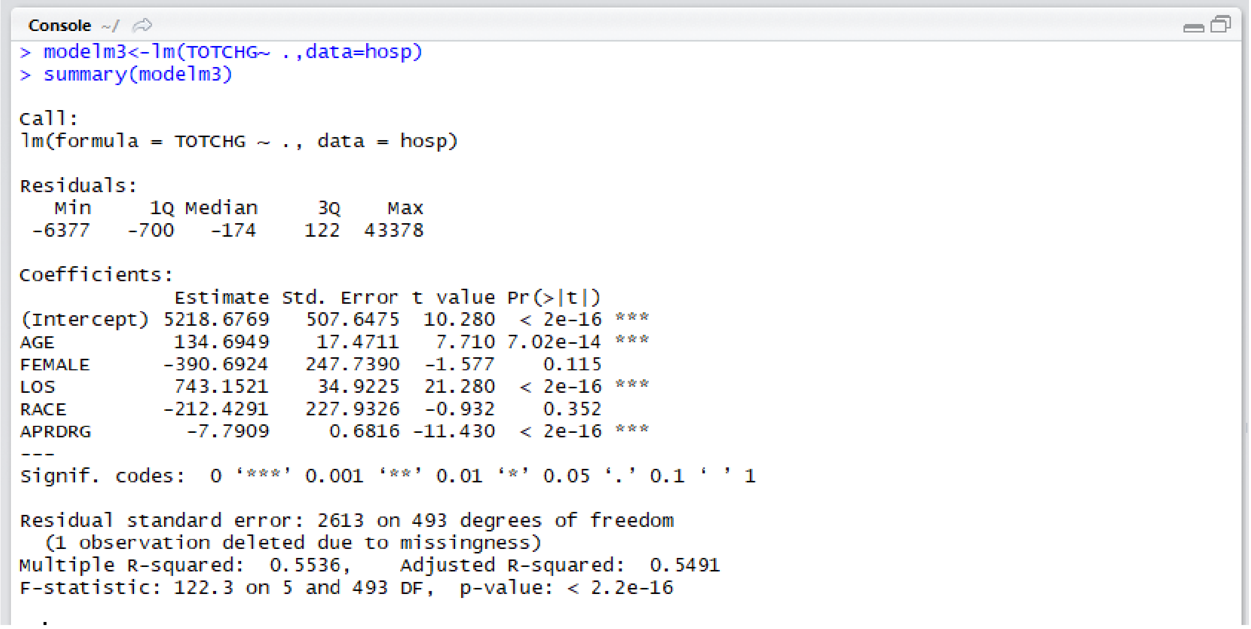
6. To perform a complete analysis, the agency wants to find the variable that mainly affects the hospital costs

Code:

modelm3<-lm(TOTCHG~ .,data=hosp)

summary(modelm3)

Output:



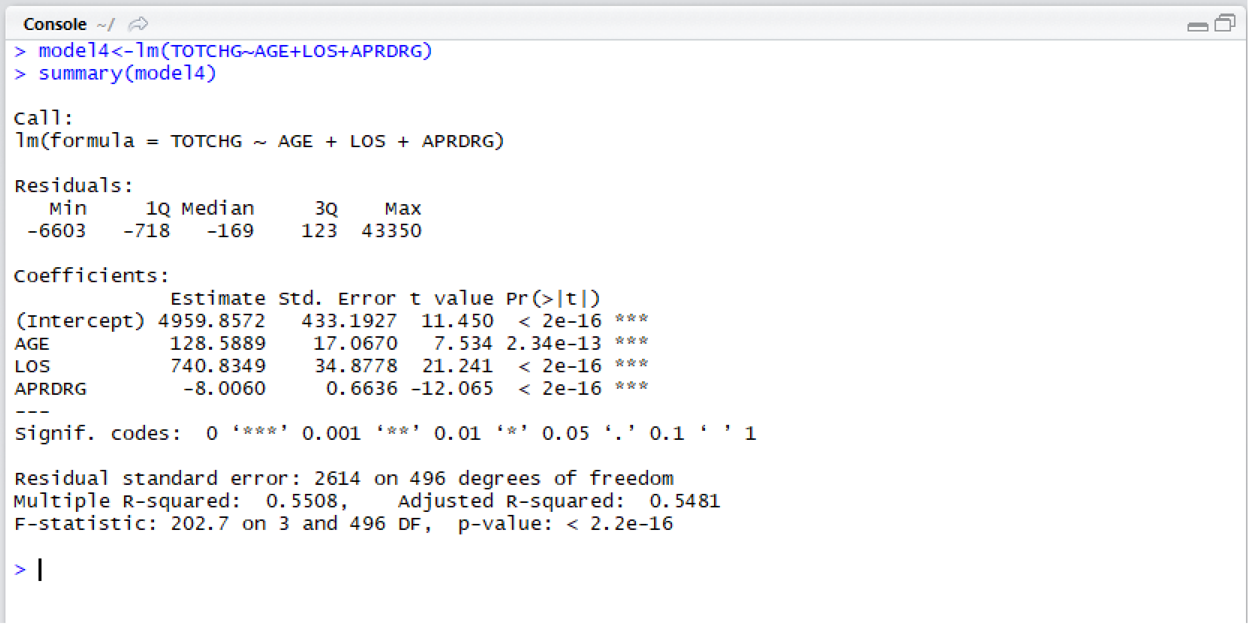
```
> modelm3<-lm(TOTCHG~ .,data=hosp)
> summary(modelm3)

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

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
```

#creating a model with only significant features

model4<-lm(TOTCHG~AGE+LOS+APRDRG)

summary(model4)

Output:



```
> model4<-lm(TOTCHG~AGE+LOS+APRDRG)
> summary(model4)

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

Residuals:
   Min     1Q Median     3Q    Max 
 -6603   -718   -169    123  43350 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4959.8572   433.1927  11.450  < 2e-16 ***
AGE          128.5889    17.0670   7.534 2.34e-13 ***
LOS          740.8349    34.8778  21.241  < 2e-16 ***
APRDRG        -8.0060     0.6636 -12.065  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2614 on 496 degrees of freedom
Multiple R-squared:  0.5508,	Adjusted R-squared:  0.5481 
F-statistic: 202.7 on 3 and 496 DF,  p-value: < 2.2e-16

> |
```

Conclusion:

#APRDRG also affect

#We can see that age and length of stay affect the total hospital cost