Project 7

R Code

```
library(readxl)

HospData <-read_xlsx("E:/Simplilearn/Project/Project 7/hospitalcosts.xlsx")

View(HospData)

summary(HospData)

hist(HospData$AGE)

summary(as.factor(HospData$AGE))

aggregate(TOTCHG ~ AGE, FUN = sum, data = HospData)

max(aggregate(TOTCHG ~ AGE, FUN = sum, data = HospData))

which.max(summary(as.factor(HospData$APRDRG)))

CostForDiagnosis<- aggregate(TOTCHG~APRDRG, FUN = sum, data = HospData)

CostForDiagnosis

CostForDiagnosis [which.max(CostForDiagnosis$TOTCHG),]

summary(as.factor(HospData$RACE))

head(HospData)

HospData<-(na.omit(HospData))

HospData

PatientRace<-as.factor(HospData$RACE)

Mod1<- aov(TOTCHG~RACE, data = HospData)

Mod1

summary(Mod1)

summary(HospData)

Mod2<-lm(TOTCHG~AGE+FEMALE, data=HospData)

Mod2
```

Fem<-as.factor(HospData$FEMALE)

summary(Mod2)

summary(Fem)

head(HospData)

Mod3<-lm(TOTCHG~AGE+FEMALE+RACE, data = HospData)

Mod3

summary(Mod3)

Mod4<-lm(TOTCHG ~ ., data=HospData)

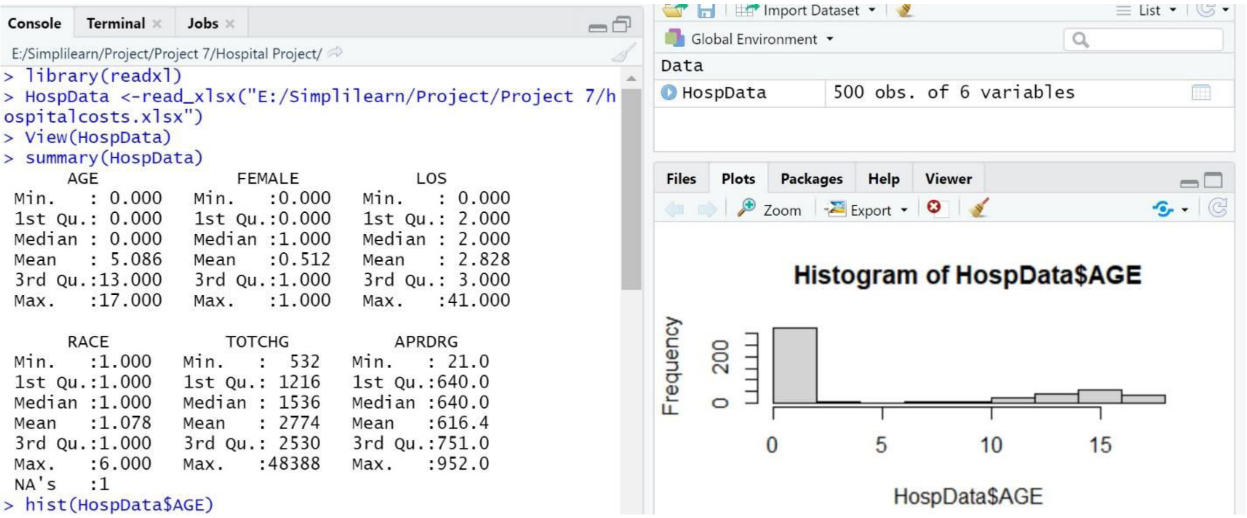
Mod4

summary(Mod4)

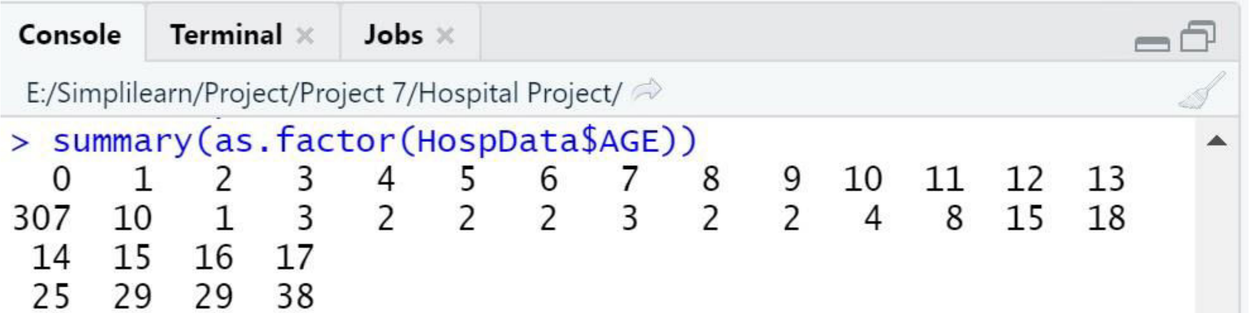
Analysis done:

QUESTION 1:

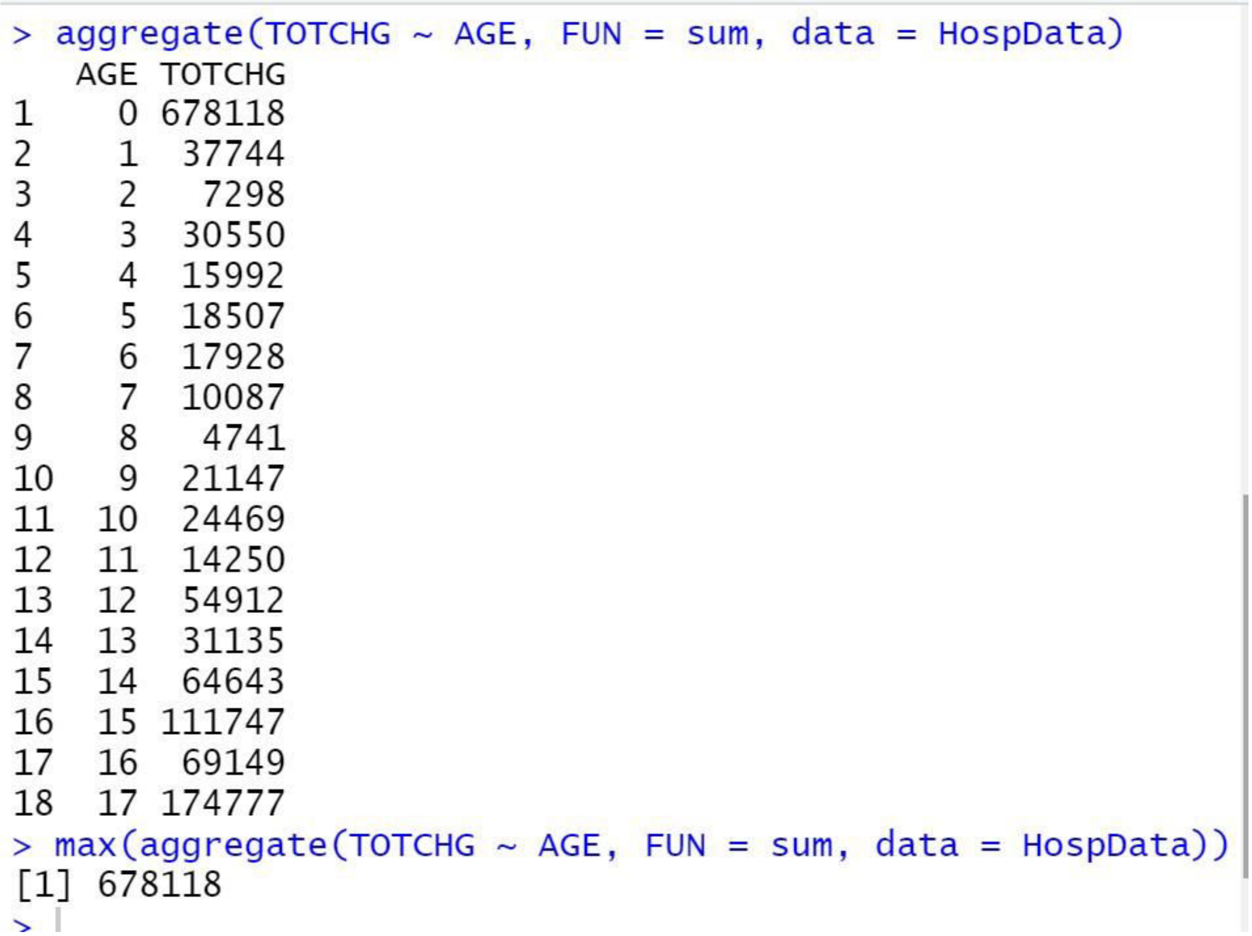
To record the patient statistics, the agency wants to find the age category of people who frequently visit the hospital and has the maximum expenditure.



```
> library(readxl)
> HospData <-read_xlsx("E:/Simplilearn/Project/Project 7/h
ospitalcosts.xlsx")
> View(HospData)
> summary(HospData)
      AGE             FEMALE            LOS        
 Min.   : 0.000   Min.   :0.000   Min.   : 0.000  
 1st Qu.: 0.000   1st Qu.:0.000   1st Qu.: 2.000  
 Median : 0.000   Median :1.000   Median : 2.000  
 Mean   : 5.086   Mean   :0.512   Mean   : 2.828  
 3rd Qu.:13.000   3rd Qu.:1.000   3rd Qu.: 3.000  
 Max.   :17.000   Max.   :1.000   Max.   :41.000  
                                                  
      RACE           TOTCHG          APRDRG     
 Min.   :1.000   Min.   :  532   Min.   : 21.0  
 1st Qu.:1.000   1st Qu.: 1216   1st Qu.:640.0  
 Median :1.000   Median : 1536   Median :640.0  
 Mean   :1.078   Mean   : 2774   Mean   :616.4  
 3rd Qu.:1.000   3rd Qu.: 2530   3rd Qu.:751.0  
 Max.   :6.000   Max.   :48388   Max.   :952.0  
 NA's   :1                                      
> hist(HospData$AGE)
```



```
> summary(as.factor(HospData$AGE))
  0   1   2   3   4   5   6   7   8   9  10  11  12  13 
307  10   1   3   2   2   2   3   2   2   4   8  15  18 
 14  15  16  17 
 25  29  29  38 
```



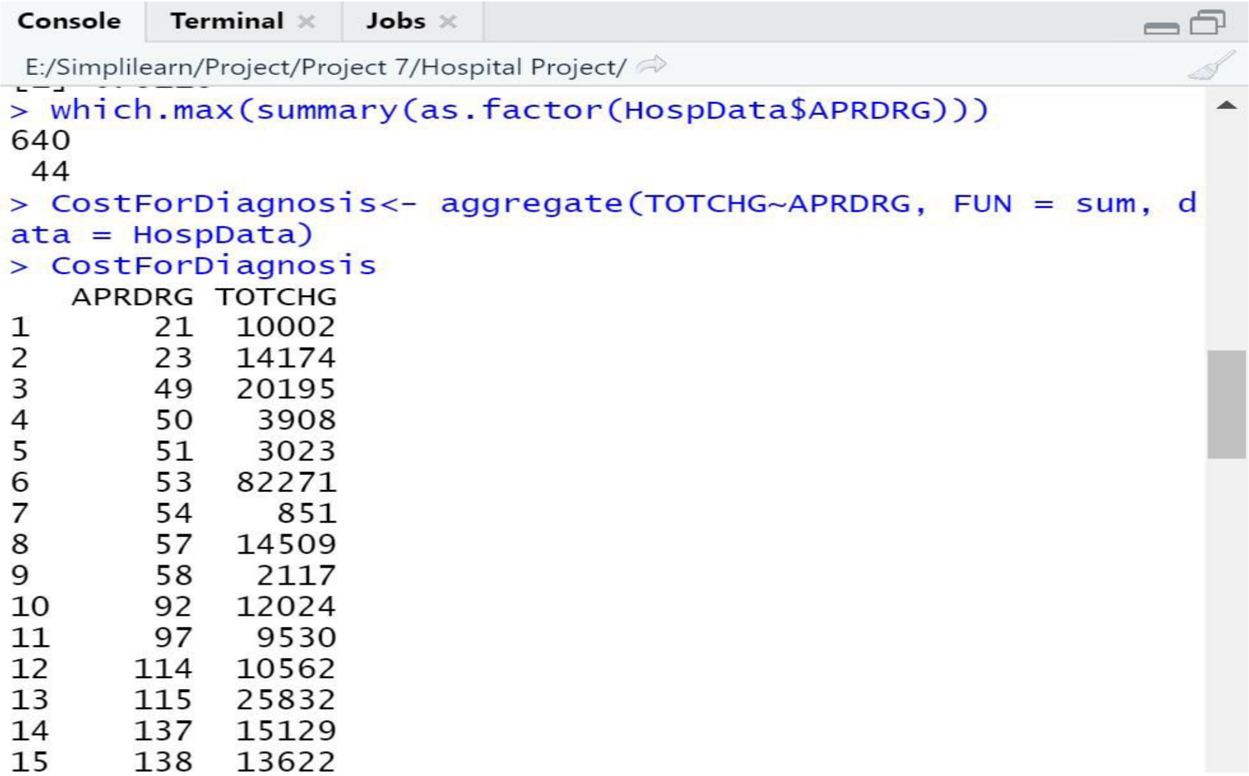
```
> aggregate(TOTCHG ~ AGE, FUN = sum, data = HospData)
   AGE TOTCHG
1    0 678118
2    1  37744
3    2   7298
4    3  30550
5    4  15992
6    5  18507
7    6  17928
8    7  10087
9    8   4741
10   9  21147
11  10  24469
12  11  14250
13  12  54912
14  13  31135
15  14  64643
16  15 111747
17  16  69149
18  17 174777
> max(aggregate(TOTCHG ~ AGE, FUN = sum, data = HospData))
[1] 678118
> 
```

ANSWER 1:

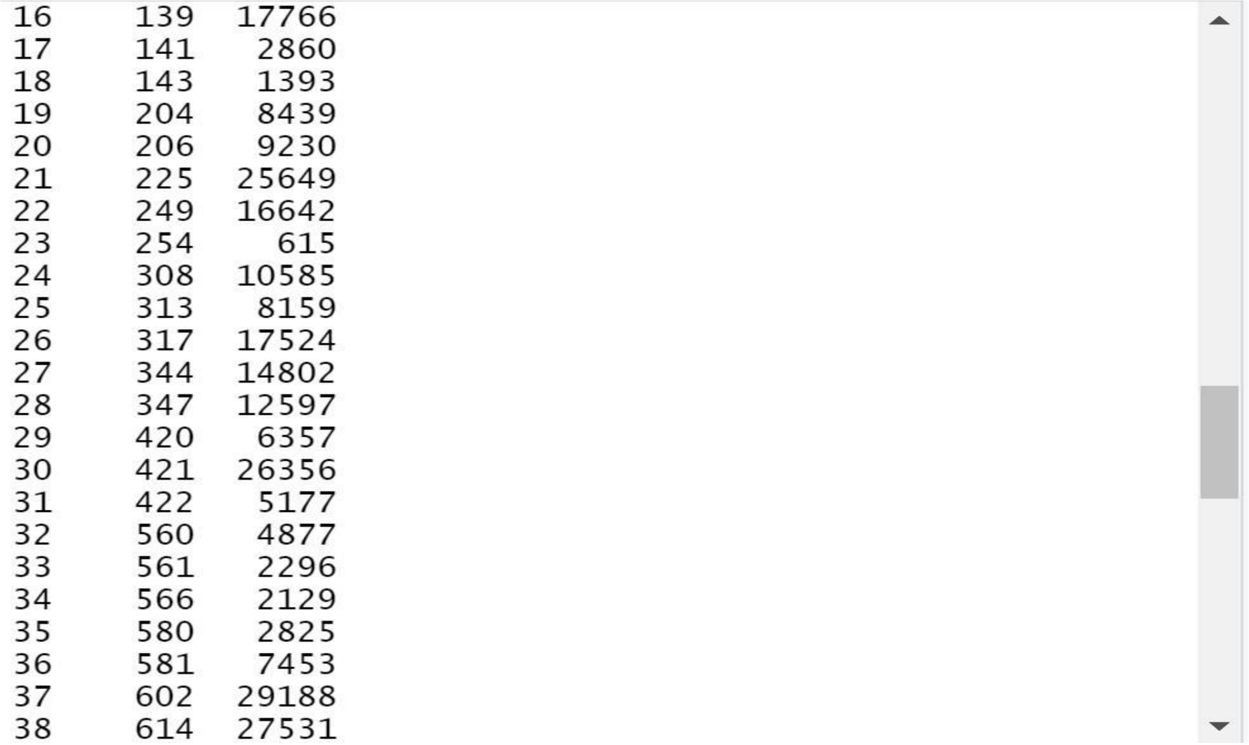
Frequent Visit Age Category is **0-1 year-old age group** and they have the maximum expenditure of **678118**

QUESTION 2:

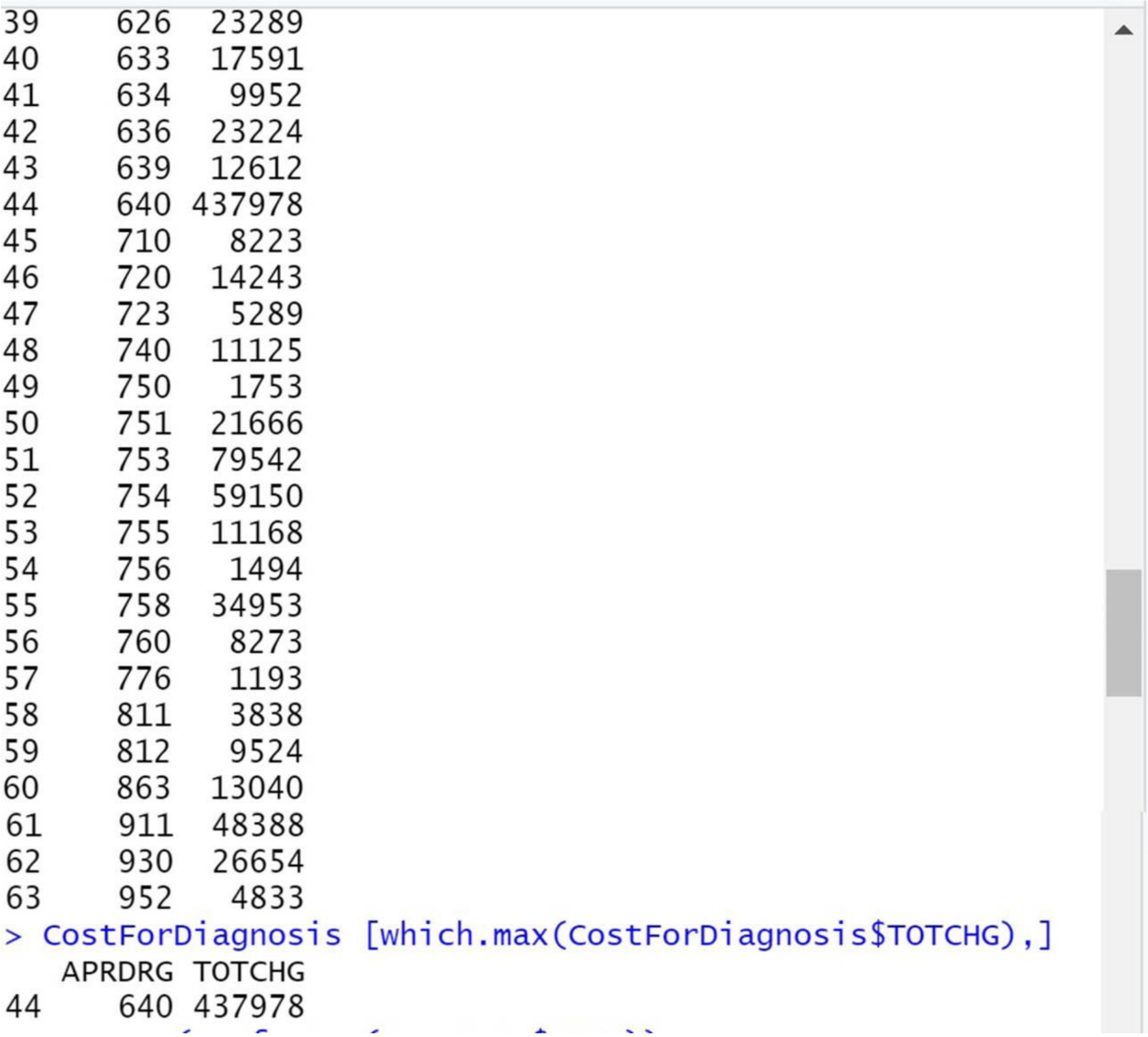
In order of severity of the diagnosis and treatments and to find out the expensive treatments, the agency wants to find the diagnosis-related group that has maximum hospitalization and expenditure.



```
> which.max(summary(as.factor(HospData$APRDRG)))
640
 44
> CostForDiagnosis<- aggregate(TOTCHG~APRDRG, FUN = sum, d
ata = HospData)
> CostForDiagnosis
   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
35    580   2825
36    581   7453
37    602  29188
38    614  27531
```



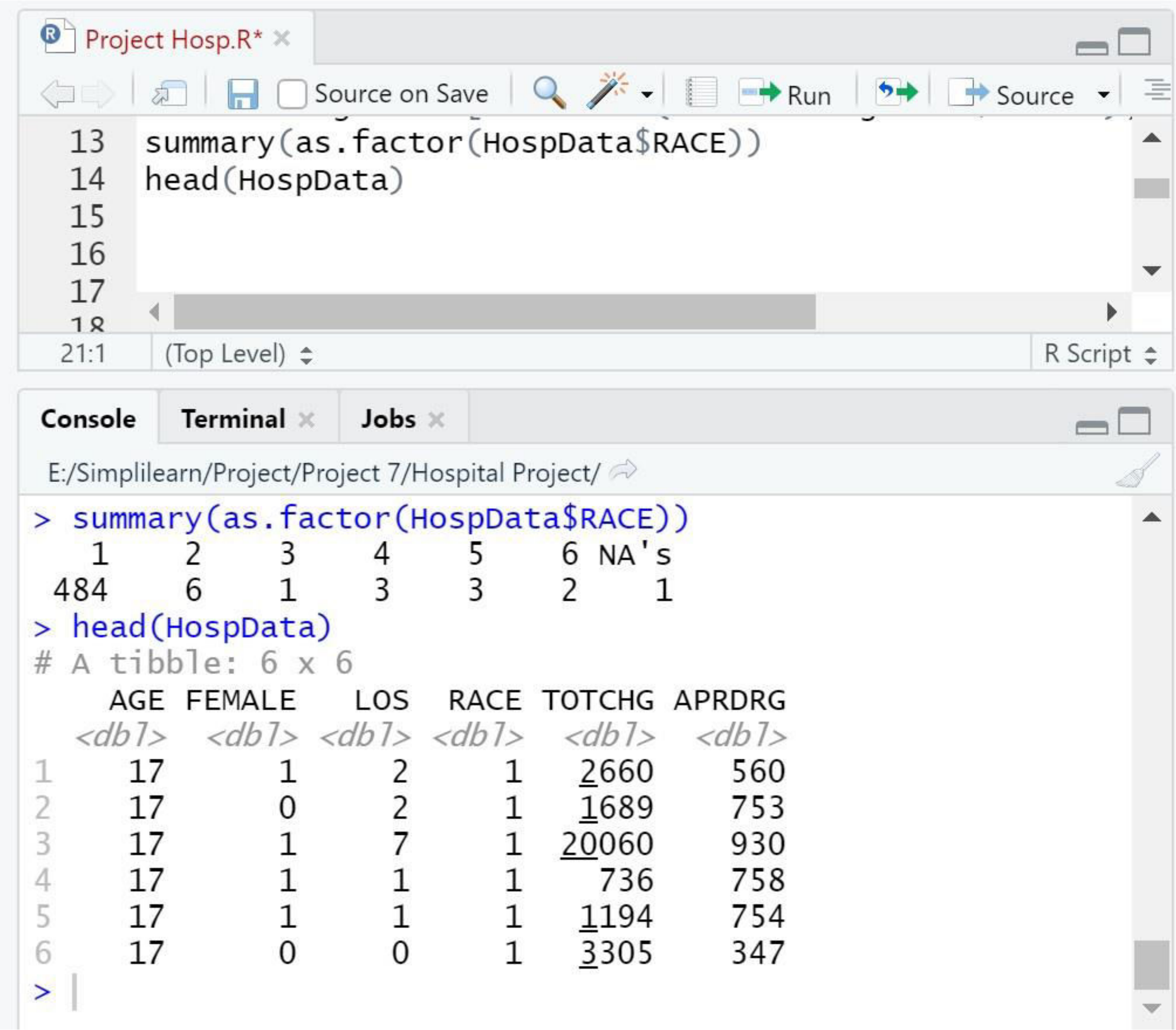
```
39     626  23289
40     633  17591
41     634   9952
42     636  23224
43     639  12612
44     640 437978
45     710   8223
46     720  14243
47     723   5289
48     740  11125
49     750   1753
50     751  21666
51     753  79542
52     754  59150
53     755  11168
54     756   1494
55     758  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
> CostForDiagnosis [which.max(CostForDiagnosis$TOTCHG),]
   APRDRG TOTCHG
44    640 437978
```

ANSWER 2:

The output shows the diagnosis-related group that has the maximum hospitalization and expenditure Group-640 with 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.



```
summary(as.factor(HospData$RACE))
head(HospData)
```

```
> summary(as.factor(HospData$RACE))
   1     2     3     4     5     6  NA's 
 484     6     1     3     3     2     1 
> head(HospData)
# A tibble: 6 x 6
    AGE FEMALE   LOS  RACE TOTCHG APRDRG
  <dbl>  <dbl> <dbl> <dbl>  <dbl>  <dbl>
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
> 
```

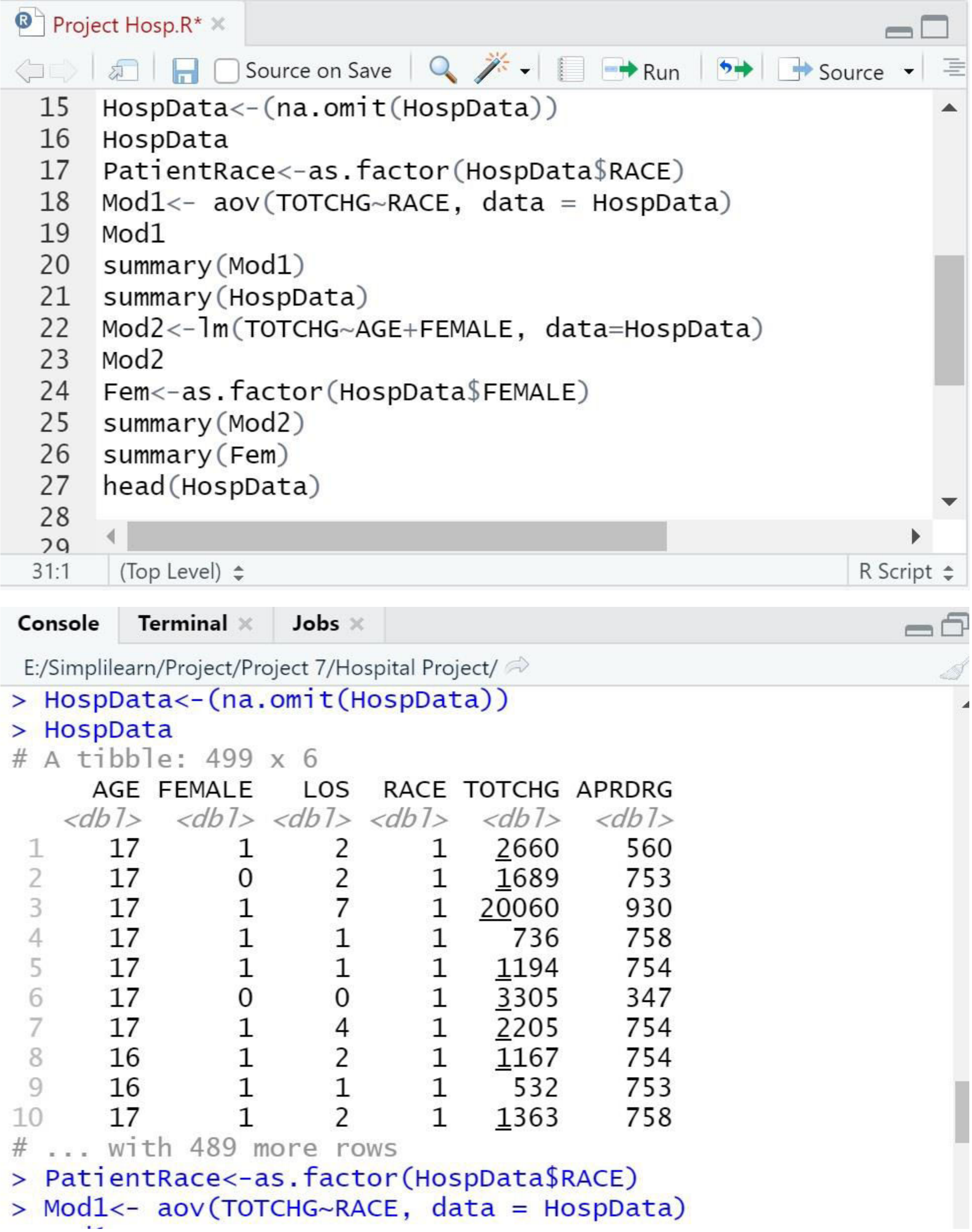
ANSWER 3:

As the **Race 1** patient group contributes to **484 out of 500 records**, this will affect the ANOVA results.

The above output shows that the Residual Value is very high specifying that **there is no relationship between the race and the hospital cost of the patient.**

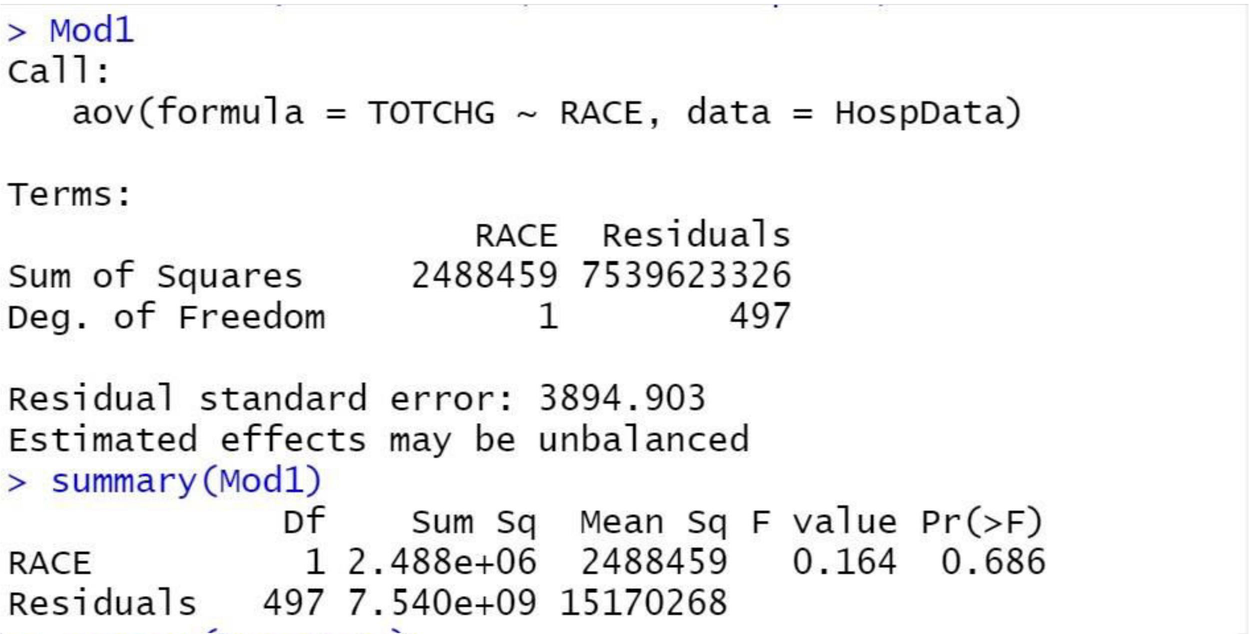
QUESTION 4:

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.

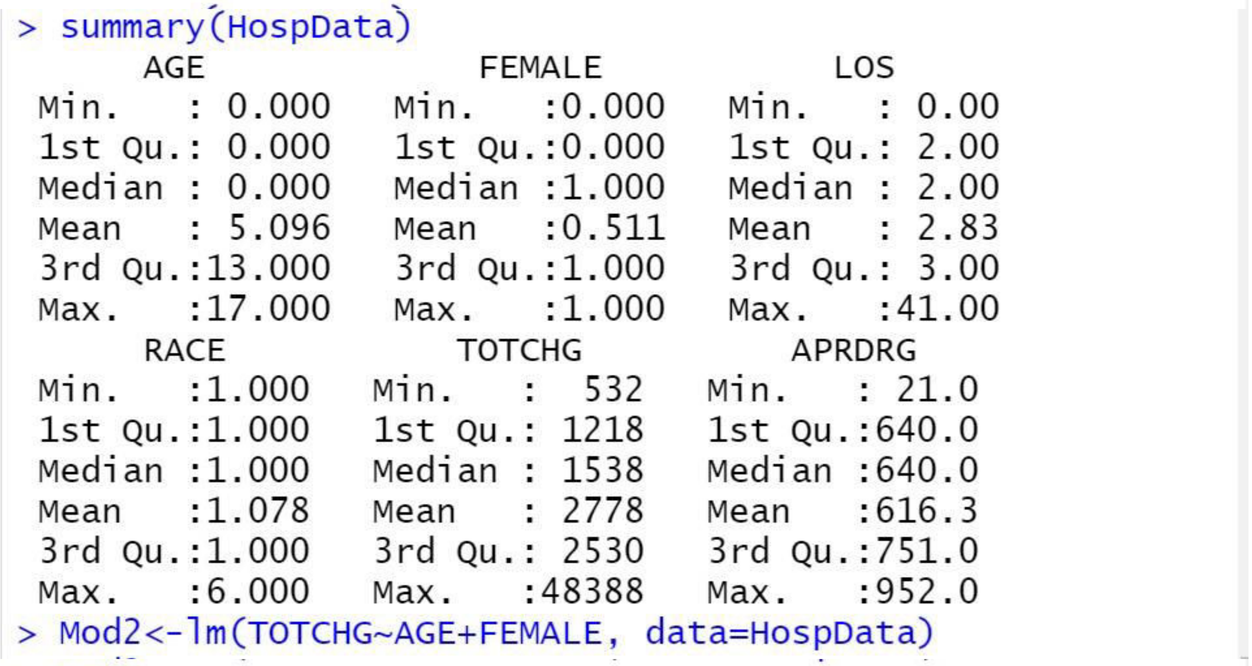


```r
HospData<-(na.omit(HospData))
HospData
PatientRace<-as.factor(HospData$RACE)
Mod1<- aov(TOTCHG~RACE, data = HospData)
Mod1
summary(Mod1)
summary(HospData)
Mod2<-lm(TOTCHG~AGE+FEMALE, data=HospData)
Mod2
Fem<-as.factor(HospData$FEMALE)
summary(Mod2)
summary(Fem)
head(HospData)
```

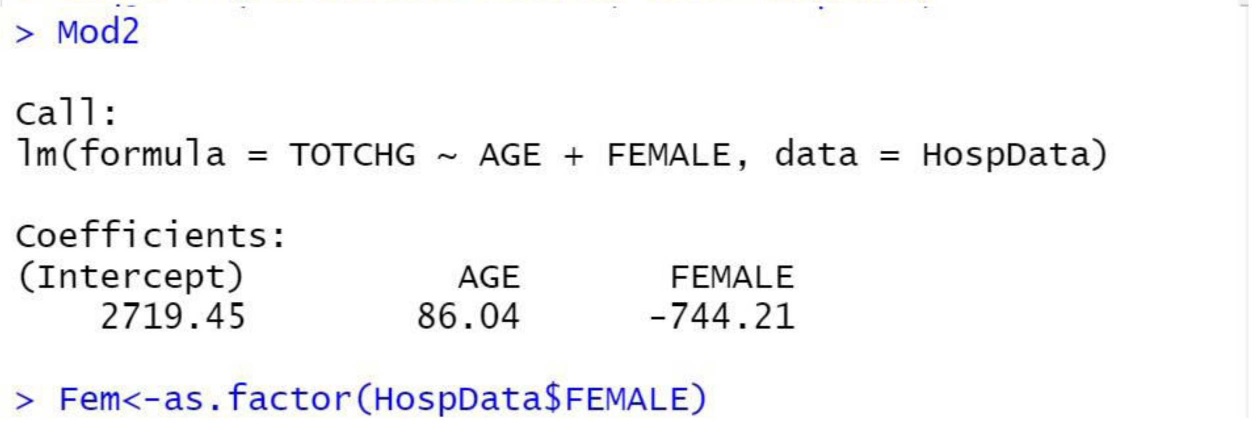
```
> HospData<-(na.omit(HospData))
> HospData
# A tibble: 499 x 6
     AGE FEMALE   LOS  RACE TOTCHG APRDRG
   <dbl>  <dbl> <dbl> <dbl>  <dbl>  <dbl>
 1    17      1     2     1   2660    560
 2    17      0     2     1   1689    753
 3    17      1     7     1  20060    930
 4    17      1     1     1    736    758
 5    17      1     1     1   1194    754
 6    17      0     0     1   3305    347
 7    17      1     4     1   2205    754
 8    16      1     2     1   1167    754
 9    16      1     1     1    532    753
10    17      1     2     1   1363    758
# ... with 489 more rows
> PatientRace<-as.factor(HospData$RACE)
> Mod1<- aov(TOTCHG~RACE, data = HospData)
```



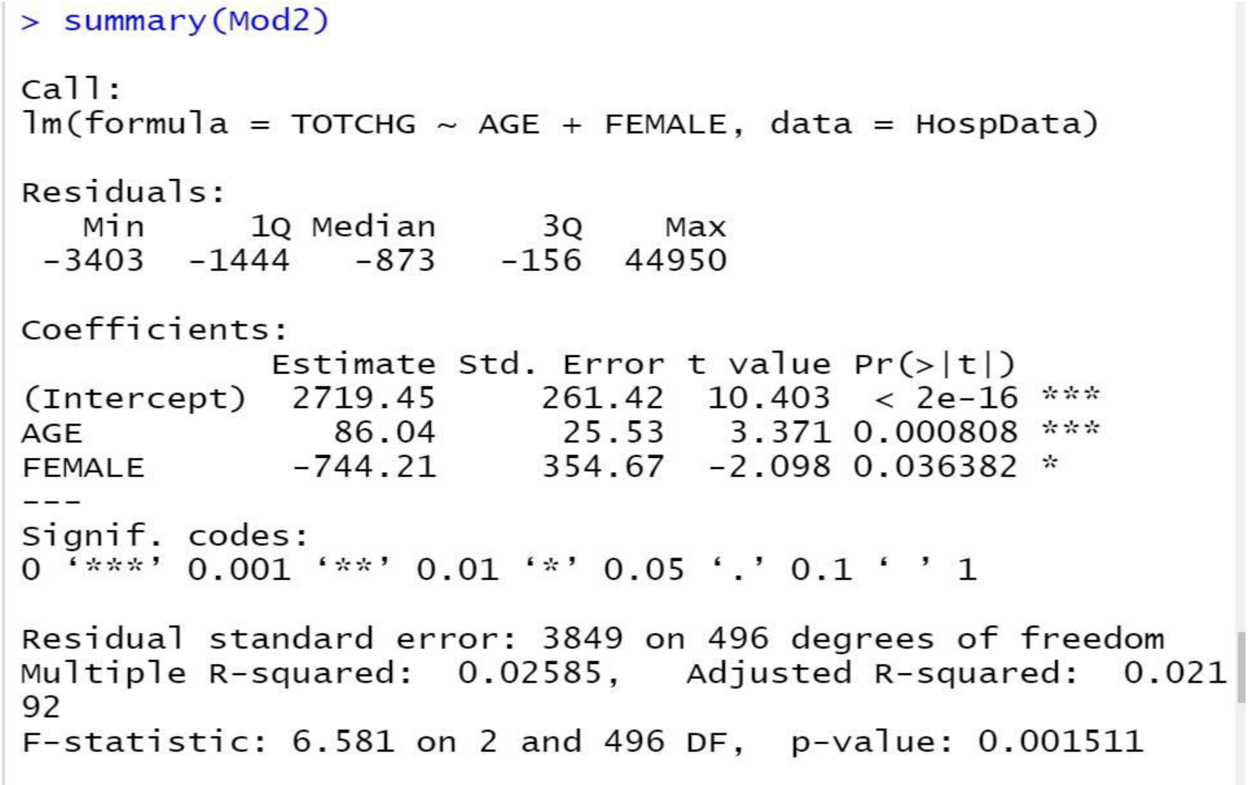
```
> Mod1
Call:
   aov(formula = TOTCHG ~ RACE, data = HospData)

Terms:
                   RACE  Residuals
Sum of Squares     2488459 7539623326
Deg. of Freedom          1        497

Residual standard error: 3894.903
Estimated effects may be unbalanced
> summary(Mod1)
             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
```



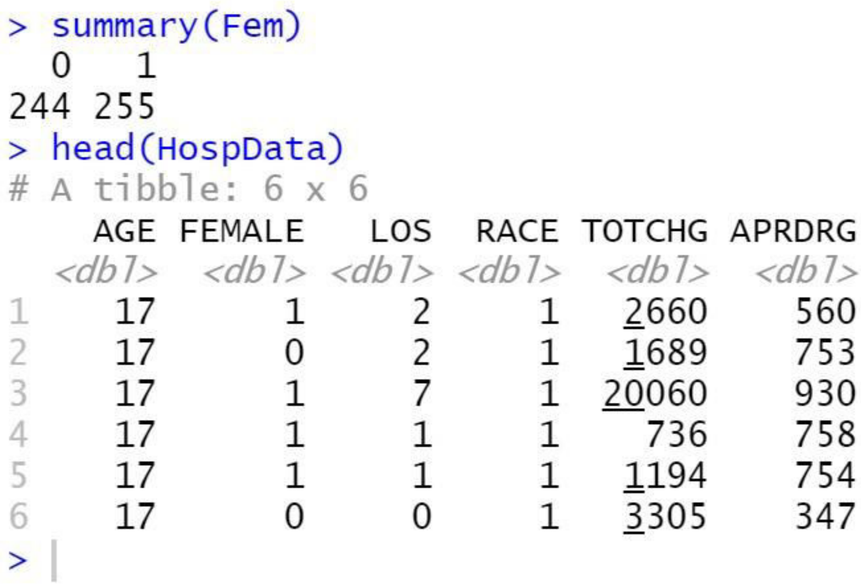
```
> summary(HospData)
      AGE             FEMALE            LOS
 Min.   : 0.000   Min.   :0.000   Min.   : 0.00
 1st Qu.: 0.000   1st Qu.:0.000   1st Qu.: 2.00
 Median : 0.000   Median :1.000   Median : 2.00
 Mean   : 5.096   Mean   :0.511   Mean   : 2.83
 3rd Qu.:13.000   3rd Qu.:1.000   3rd Qu.: 3.00
 Max.   :17.000   Max.   :1.000   Max.   :41.00
      RACE            TOTCHG           APRDRG
 Min.   :1.000   Min.   :  532   Min.   : 21.0
 1st Qu.:1.000   1st Qu.: 1218   1st Qu.:640.0
 Median :1.000   Median : 1538   Median :640.0
 Mean   :1.078   Mean   : 2778   Mean   :616.3
 3rd Qu.:1.000   3rd Qu.: 2530   3rd Qu.:751.0
 Max.   :6.000   Max.   :48388   Max.   :952.0
> Mod2<-lm(TOTCHG~AGE+FEMALE, data=HospData)
```



```
> Mod2

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

Coefficients:
(Intercept)          AGE       FEMALE
    2719.45        86.04      -744.21

> Fem<-as.factor(HospData$FEMALE)
```



```
> summary(Mod2)

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

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:  0.021
92 
F-statistic: 6.581 on 2 and 496 DF,  p-value: 0.001511
```



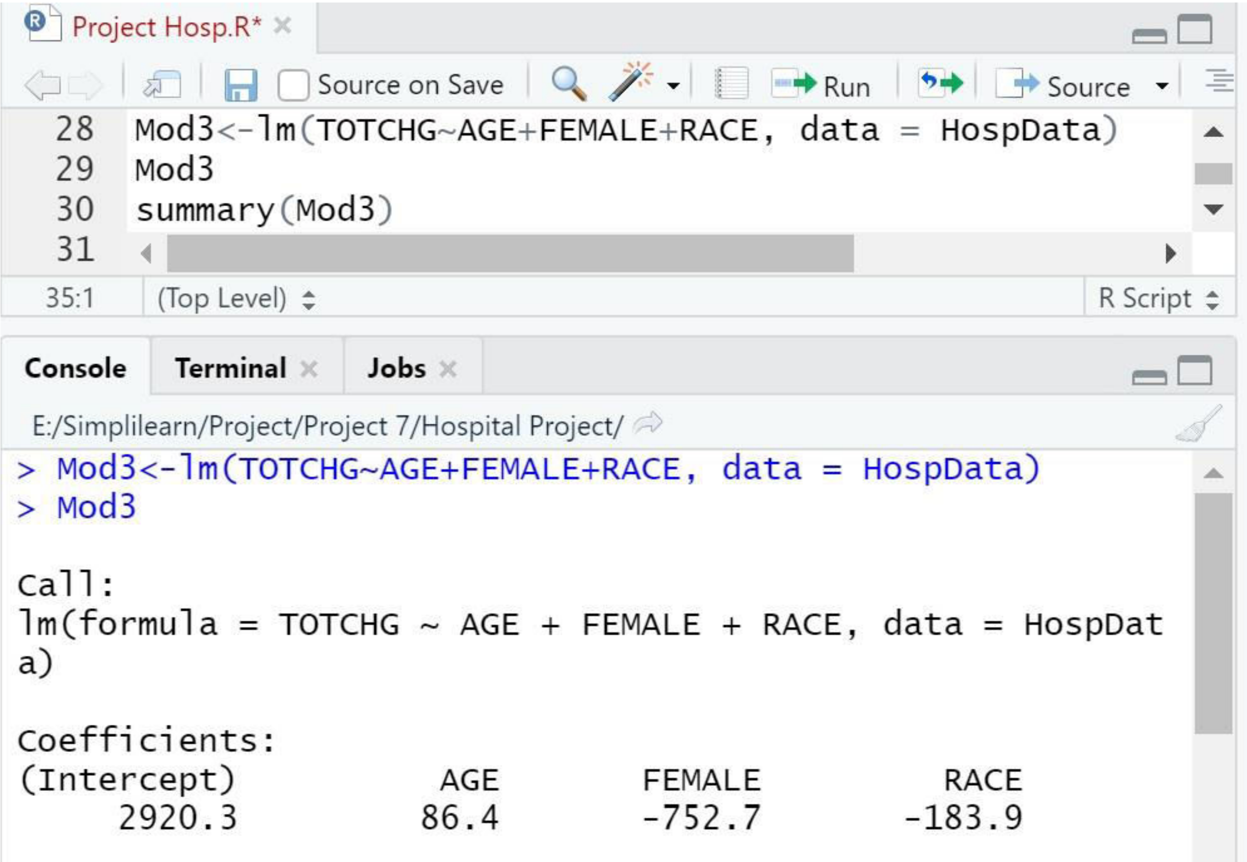
```
> summary(Fem)
  0   1 
244 255 
> head(HospData)
# A tibble: 6 x 6
    AGE FEMALE   LOS  RACE TOTCHG APRDRG
  <dbl>  <dbl> <dbl> <dbl>  <dbl>  <dbl>
1    17      1     2     1   2660    560
2    17      0     2     1   1689    753
3    17      1     7     1  20060    930
4    17      1     1     1    736    758
5    17      1     1     1   1194    754
6    17      0     0     1   3305    347
> 
```

ANSWER 4:

The above output indicates that **Age** and **Gender** is an important factor in the **Hospital Costs**, which is signified by the **p-values**. The data does have almost equal number of males and females and the **negative coefficient** suggests female patients tend to have less hospital costs than male patients.

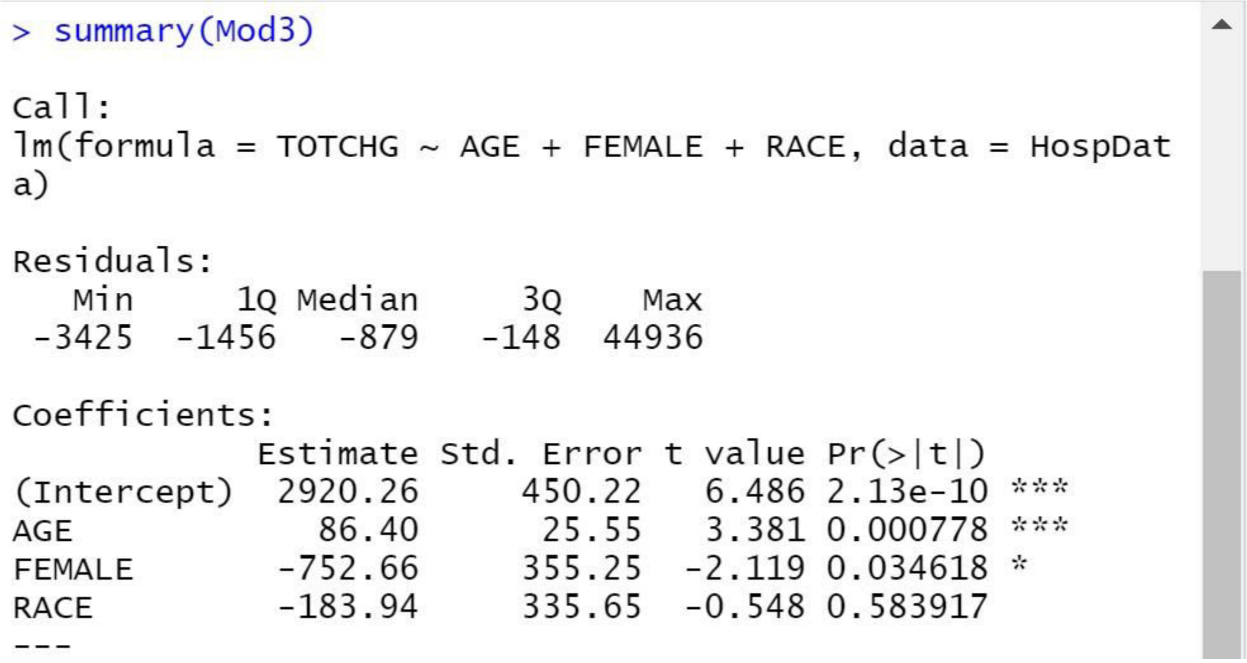
QUESTION 5:

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.

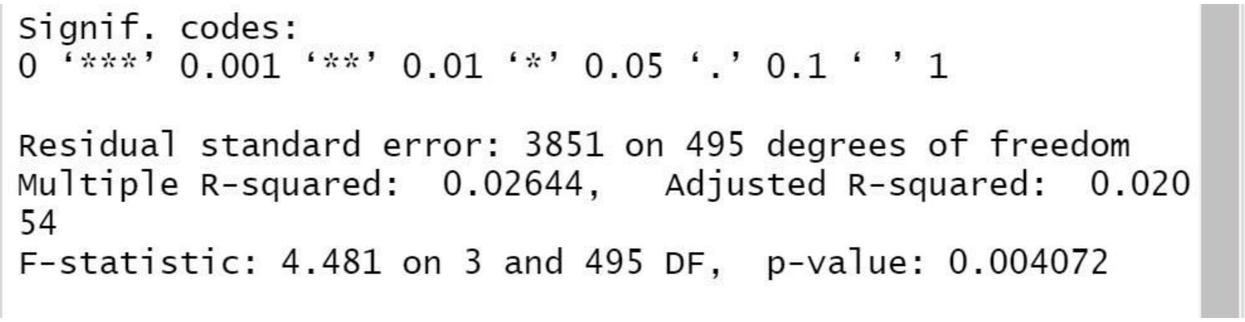


```
Mod3<-lm(TOTCHG~AGE+FEMALE+RACE, data = HospData)
Mod3
summary(Mod3)
```

```
> Mod3<-lm(TOTCHG~AGE+FEMALE+RACE, data = HospData)
> Mod3

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

Coefficients:
(Intercept)          AGE       FEMALE         RACE
     2920.3         86.4       -752.7       -183.9
```



```
> summary(Mod3)

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

Residuals:
   Min     1Q Median     3Q    Max 
 -3425  -1456   -879   -148  44936 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2920.26     450.22   6.486 2.13e-10 ***
AGE            86.40      25.55   3.381 0.000778 ***
FEMALE       -752.66     355.25  -2.119 0.034618 *  
RACE         -183.94     335.65  -0.548 0.583917    
---
```



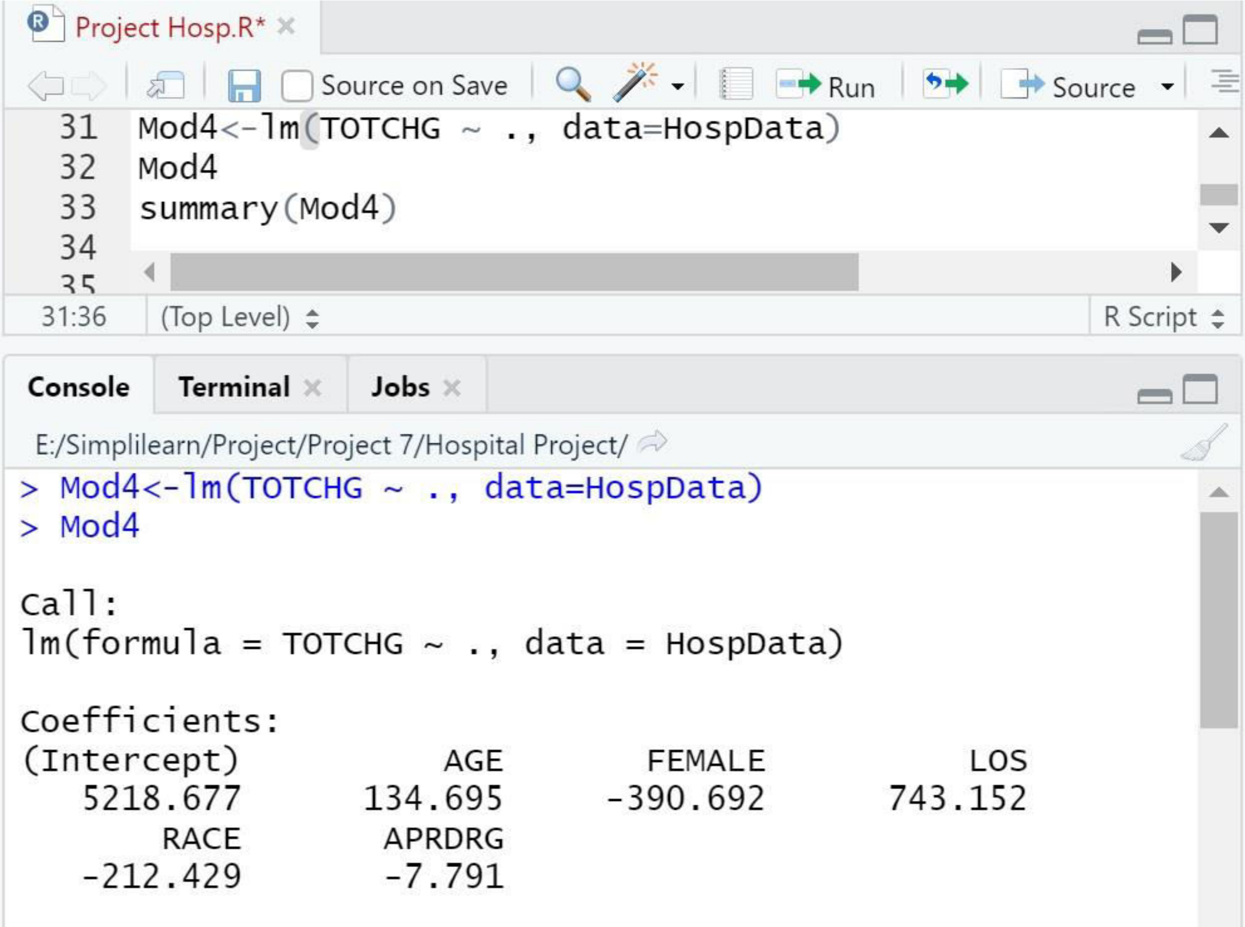
```
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3851 on 495 degrees of freedom
Multiple R-squared:  0.02644,	Adjusted R-squared:  0.020
54
F-statistic: 4.481 on 3 and 495 DF,  p-value: 0.004072
```

ANSWER 5:

The above output denotes that **Significance codes** are negligible except for the **Intercept** and the high **p-value** signifies that the **Length of Stay** is not related to **Age, Gender** and **Race.**

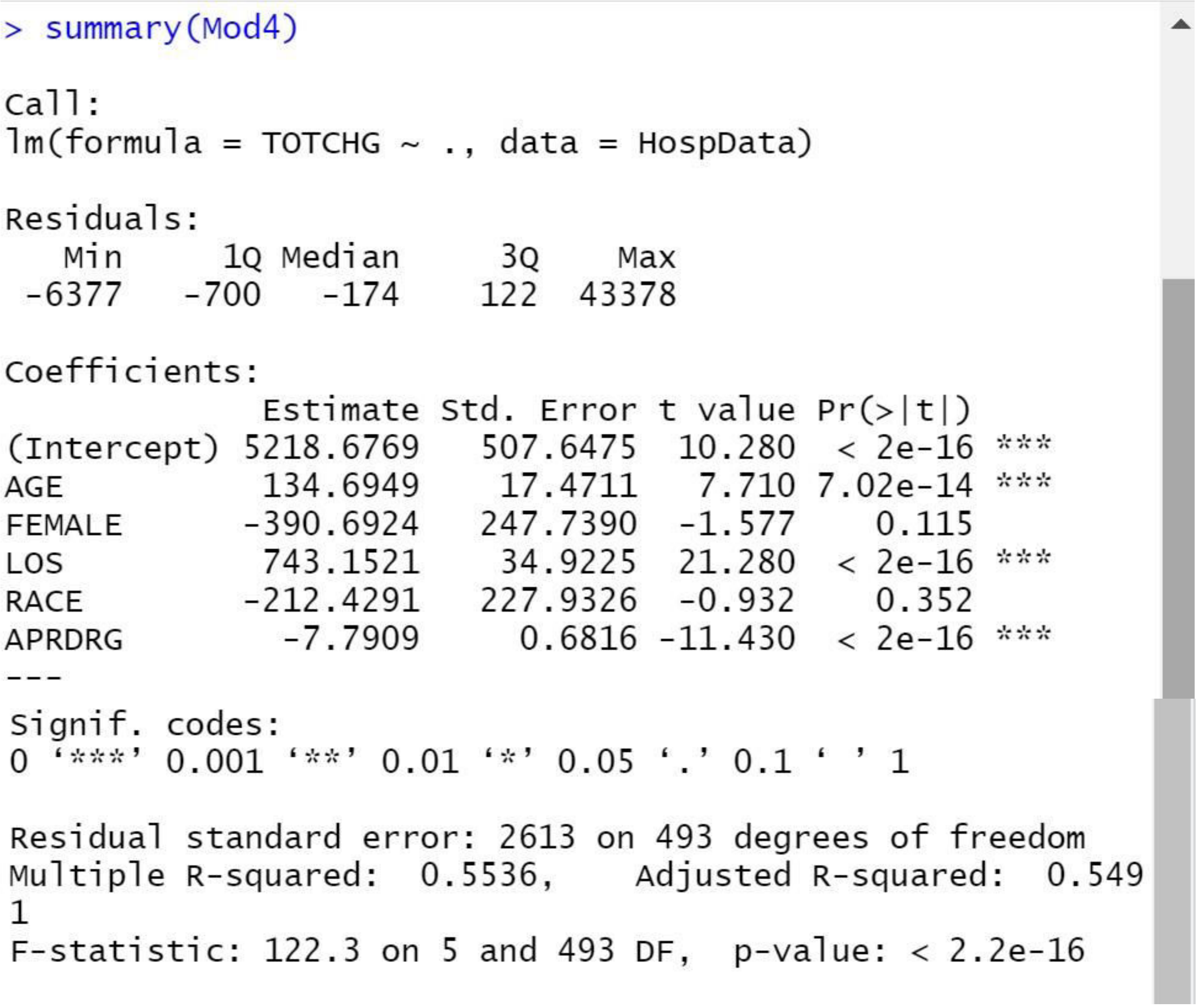
QUESTION 6:

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



```
Mod4<-lm(TOTCHG ~ ., data=HospData)
Mod4
summary(Mod4)
```

```
> Mod4<-lm(TOTCHG ~ ., data=HospData)
> Mod4

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

Coefficients:
(Intercept)          AGE       FEMALE          LOS
   5218.677      134.695     -390.692      743.152
       RACE       APRDRG
   -212.429       -7.791
```



```
> summary(Mod4)

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

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:  0.549
1 
F-statistic: 122.3 on 5 and 493 DF,  p-value: < 2.2e-16
```

ANSWER 6:

The output suggests that **Age** and **Length of Stay** factors contribute mainly to the **Hospital Cost.** The results suggest the **Hospital Cost** increases by around **743** for every day the patient stays.