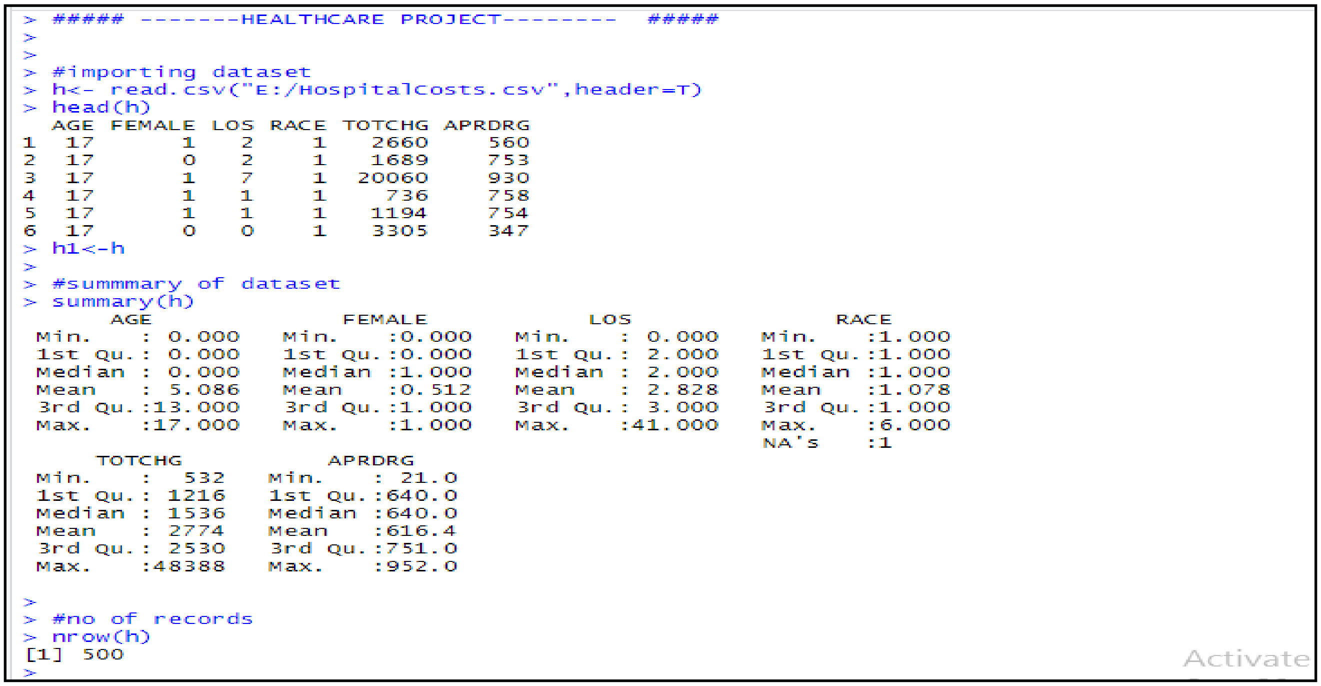
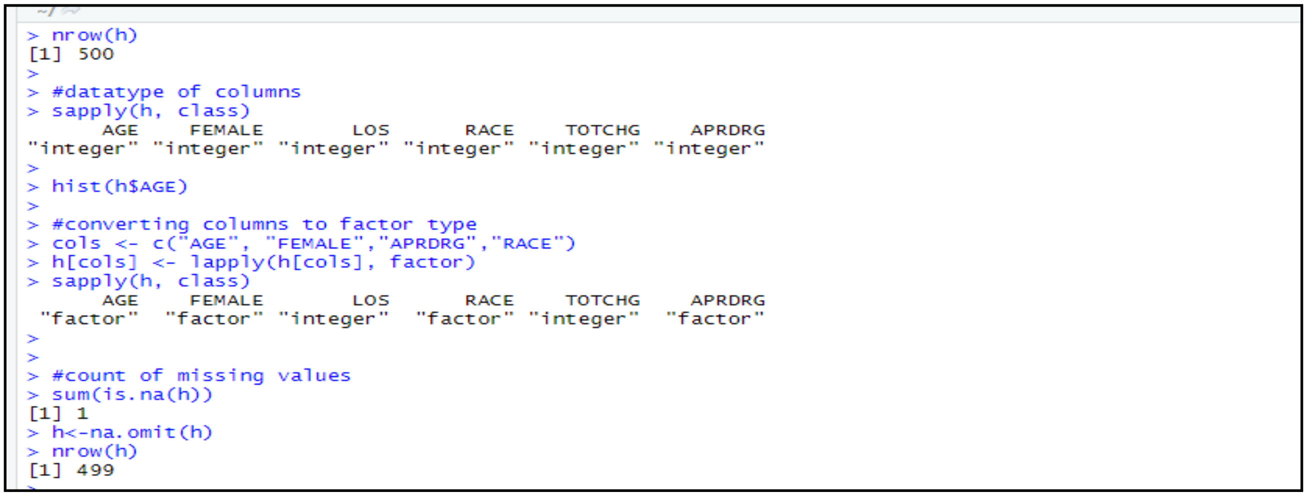
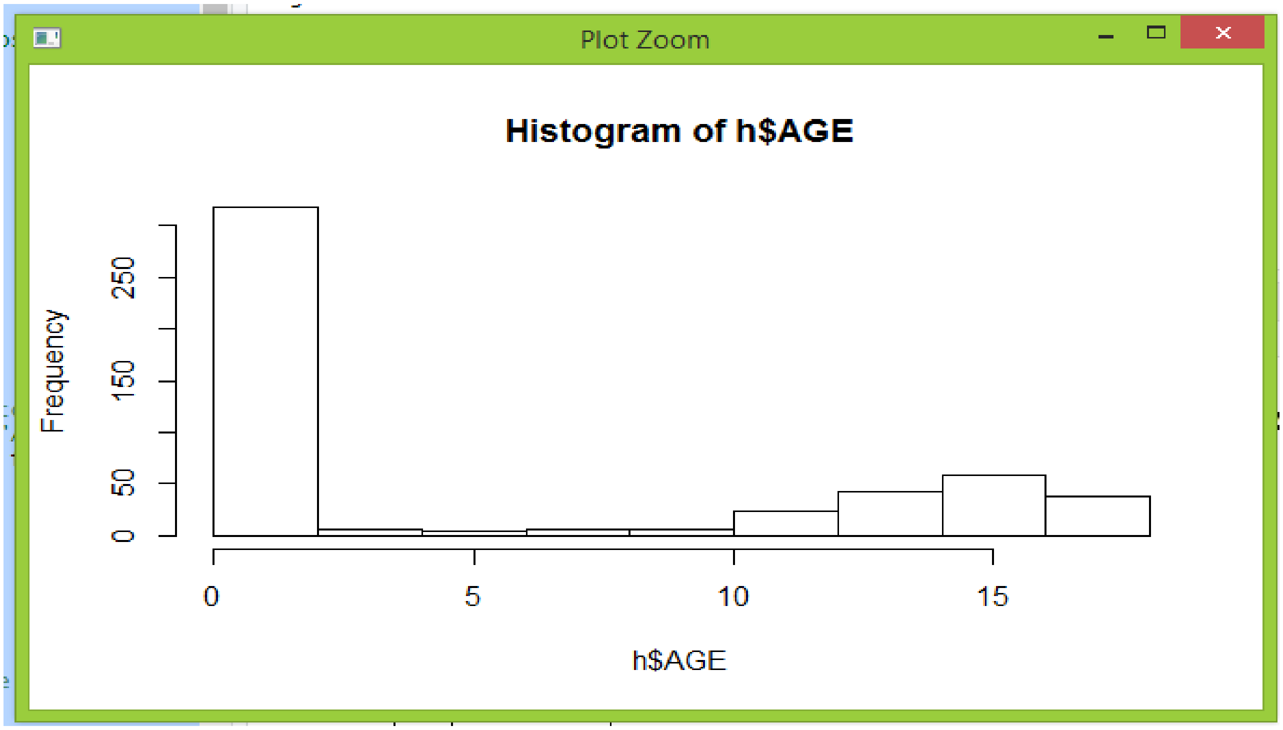
Exploring Data



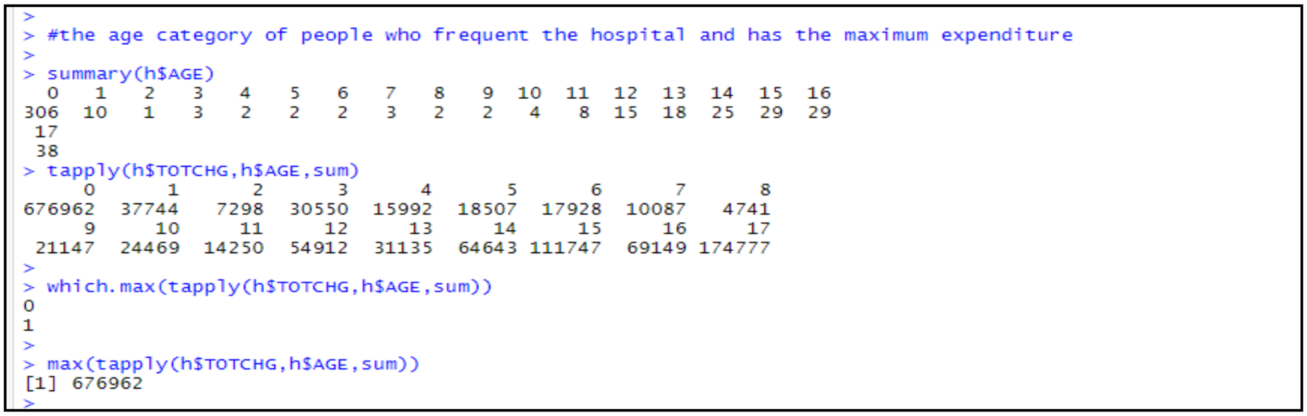
```
> #####  -------HEALTHCARE PROJECT--------   #####
>
>
> #importing dataset
> h<- read.csv("E:/Hospitalcosts.csv",header=T)
> head(h)
  AGE FEMALE LOS RACE TOTCHG APRDRG
1  17      1   2    1   2660    560
2  17      0   2    1   1689    753
3  17      1   7    1  20060    930
4  17      1   1    1    736    758
5  17      1   1    1   1194    754
6  17      0   0    1   3305    347
> h1<-h
>
> #summmary of dataset
> summary(h)
      AGE             FEMALE          LOS              RACE      
 Min.   : 0.000   Min.   :0.000   Min.   : 0.000   Min.   :1.000  
 1st Qu.: 0.000   1st Qu.:0.000   1st Qu.: 2.000   1st Qu.:1.000  
 Median : 0.000   Median :1.000   Median : 2.000   Median :1.000  
 Mean   : 5.086   Mean   :0.512   Mean   : 2.828   Mean   :1.078  
 3rd Qu.:13.000   3rd Qu.:1.000   3rd Qu.: 3.000   3rd Qu.:1.000  
 Max.   :17.000   Max.   :1.000   Max.   :41.000   Max.   :6.000  
                                                   NA's   :1      
     TOTCHG          APRDRG     
 Min.   :  532   Min.   : 21.0  
 1st Qu.: 1216   1st Qu.:640.0  
 Median : 1536   Median :640.0  
 Mean   : 2774   Mean   :616.4  
 3rd Qu.: 2530   3rd Qu.:751.0  
 Max.   :48388   Max.   :952.0  

>
> #no of records
> nrow(h)
[1] 500
>
```



```
> nrow(h)
[1] 500
>
> #datatype of columns
> sapply(h, class)
      AGE    FEMALE       LOS      RACE    TOTCHG    APRDRG 
"integer" "integer" "integer" "integer" "integer" "integer" 
>
> hist(h$AGE)
>
> #converting columns to factor type
> cols <- c("AGE", "FEMALE","APRDRG","RACE")
> h[cols] <- lapply(h[cols], factor)
> sapply(h, class)
      AGE    FEMALE       LOS      RACE    TOTCHG    APRDRG 
 "factor"  "factor" "integer"  "factor" "integer"  "factor" 
>
>
> #count of missing values
> sum(is.na(h))
[1] 1
> h<-na.omit(h)
> nrow(h)
[1] 499
```



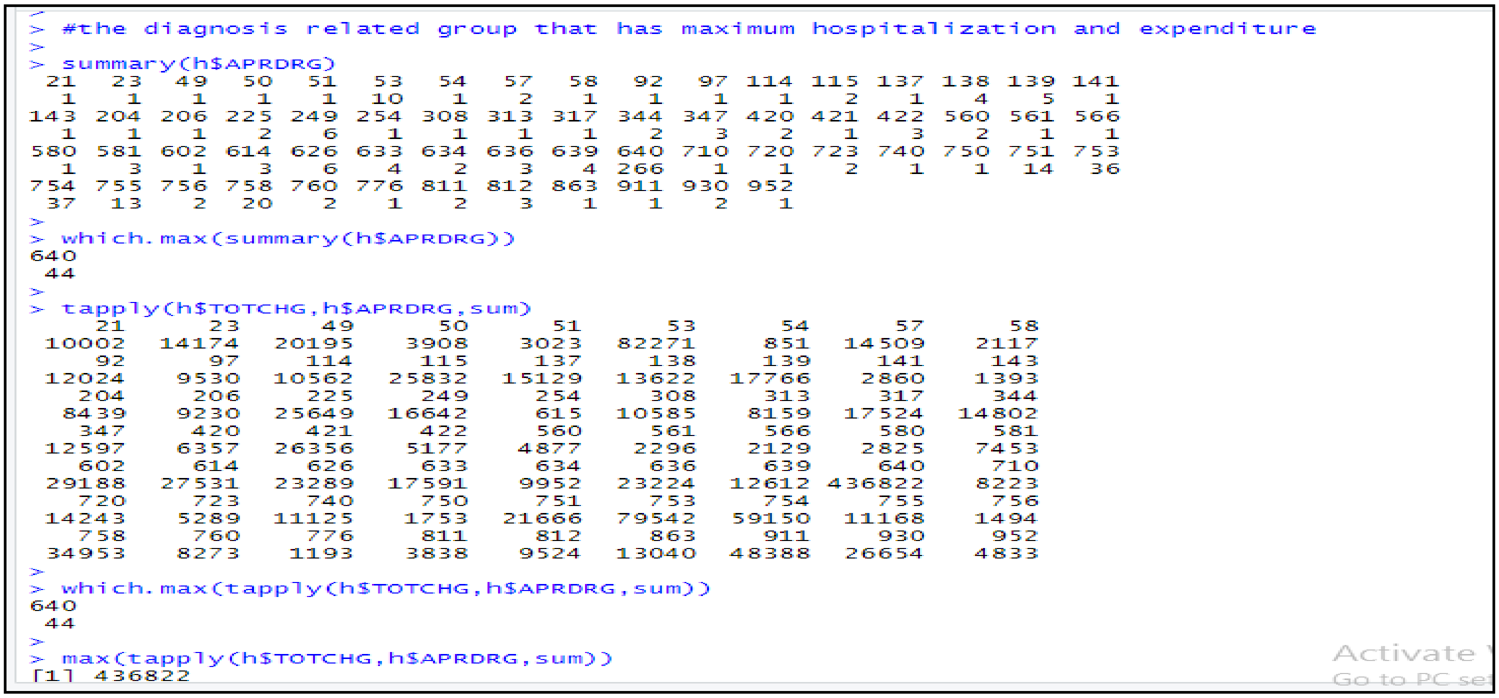
– To record the patient statistics, the agency wants to find the age category of people who frequent the hospital and has the maximum expenditure.



```
>
> #the age category of people who frequent the hospital and has the maximum expenditure
>
> summary(h$AGE)
  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16 
306  10   1   3   2   2   2   3   2   2   4   8  15  18  25  29  29 
 17 
 38 
> tapply(h$TOTCHG,h$AGE,sum)
     0      1      2      3      4      5      6      7      8 
676962  37744   7298  30550  15992  18507  17928  10087   4741 
     9     10     11     12     13     14     15     16     17 
 21147  24469  14250  54912  31135  64643 111747  69149 174777 
>
> which.max(tapply(h$TOTCHG,h$AGE,sum))
0 
1 
>
> max(tapply(h$TOTCHG,h$AGE,sum))
[1] 676962
>
```

age category of 0 seems to be frequently using the hospital with maximum expenditure 676962

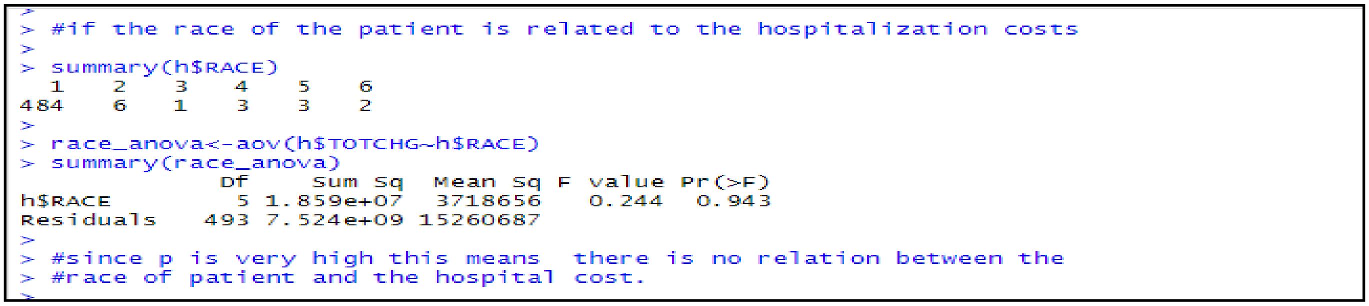
– In order of severity of the diagnosis and treatments and to find out the expensive treatments, the agency wants to find the diagnosis related group that has maximum hospitalization and expenditure.



```
>
> #the diagnosis related group that has maximum hospitalization and expenditure
>
> summary(h$APRDRG)
 21  23  49  50  51  53  54  57  58  92  97 114 115 137 138 139 141 
  1   1   1   1   1  10   1   2   1   1   1   1   2   1   4   5   1 
143 204 206 225 249 254 308 313 317 344 347 420 421 422 560 561 566 
  1   1   1   2   6   1   1   1   1   2   3   2   1   3   2   1   1 
580 581 602 614 626 633 634 636 639 640 710 720 723 740 750 751 753 
  1   3   1   3   6   4   2   3   4 266   1   1   2   1   1  14  36 
754 755 756 758 760 776 811 812 863 911 930 952 
 37  13   2  20   2   1   2   3   1   1   2   1 
>
> which.max(summary(h$APRDRG))
640 
 44 
>
> tapply(h$TOTCHG,h$APRDRG,sum)
    21     23     49     50     51     53     54     57     58 
 10002  14174  20195   3908   3023  82271    851  14509   2117 
    92     97    114    115    137    138    139    141    143 
 12024   9530  10562  25832  15129  13622  17766   2860   1393 
   204    206    225    249    254    308    313    317    344 
  8439   9230  25649  16642    615  10585   8159  17524  14802 
   347    420    421    422    560    561    566    580    581 
 12597   6357  26356   5177   4877   2296   2129   2825   7453 
   602    614    626    633    634    636    639    640    710 
 29188  27531  23289  17591   9952  23224  12612 436822   8223 
   720    723    740    750    751    753    754    755    756 
 14243   5289  11125   1753  21666  79542  59150  11168   1494 
   758    760    776    811    812    863    911    930    952 
 34953   8273   1193   3838   9524  13040  48388  26654   4833 
>
> which.max(tapply(h$TOTCHG,h$APRDRG,sum))
640 
 44 
>
> max(tapply(h$TOTCHG,h$APRDRG,sum))
[1] 436822
```

From the results we can see that the category 640 has the maximum entries of hospitalization and also has the highest total hospitalization cost of 436822.

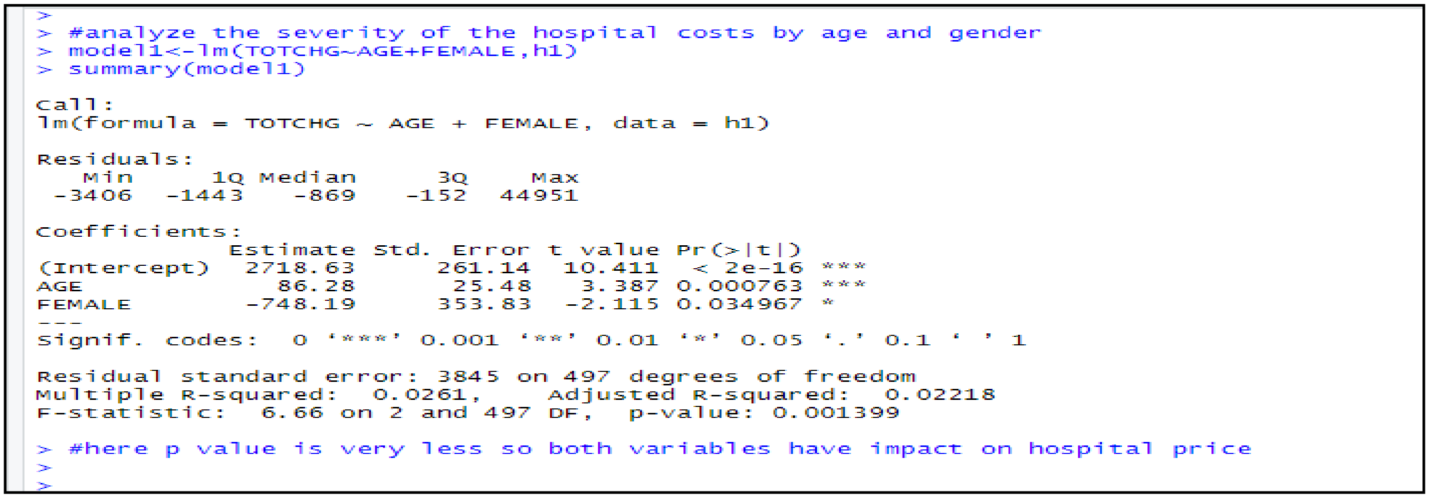
– To make sure that there is no malpractice, the agency needs to analyze if the race of the patient is related to the hospitalization costs.



```
>
> #if the race of the patient is related to the hospitalization costs
>
> summary(h$RACE)
  1   2   3   4   5   6 
484   6   1   3   3   2 
>
> race_anova<-aov(h$TOTCHG~h$RACE)
> summary(race_anova)
             Df    Sum Sq  Mean Sq F value Pr(>F)
h$RACE        5 1.859e+07  3718656   0.244  0.943
Residuals   493 7.524e+09 15260687               
>
> #since p is very high this means  there is no relation between the
> #race of patient and the hospital cost.
>
```

since p is very high this means there is no relation between the race of patient and the hospital cost

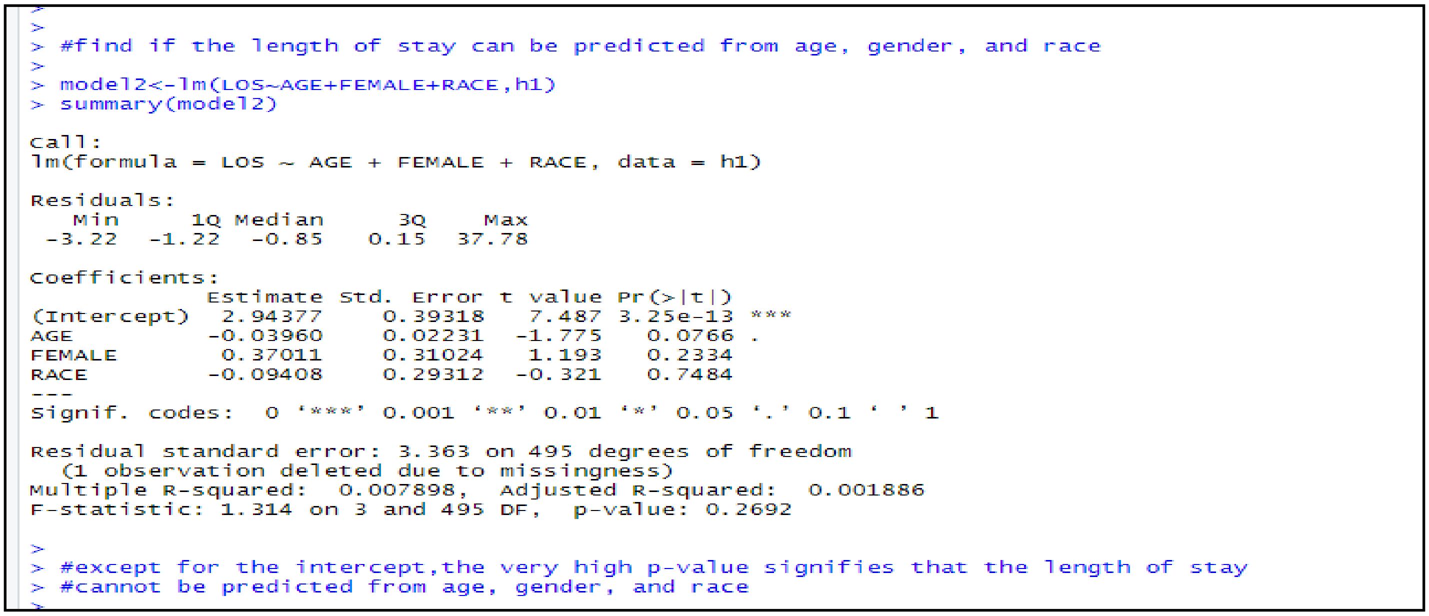
– To properly utilize the costs, the agency has to analyze the severity of the hospital costs by age and gender for proper allocation of resources.



```
>
> #analyze the severity of the hospital costs by age and gender
> model1<-lm(TOTCHG~AGE+FEMALE,h1)
> summary(model1)

Call:
lm(formula = TOTCHG ~ AGE + FEMALE, data = h1)

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

> #here p value is very less so both variables have impact on hospital price
>
>
```

here p value is very less so both variables have impact on hospital price

– 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.

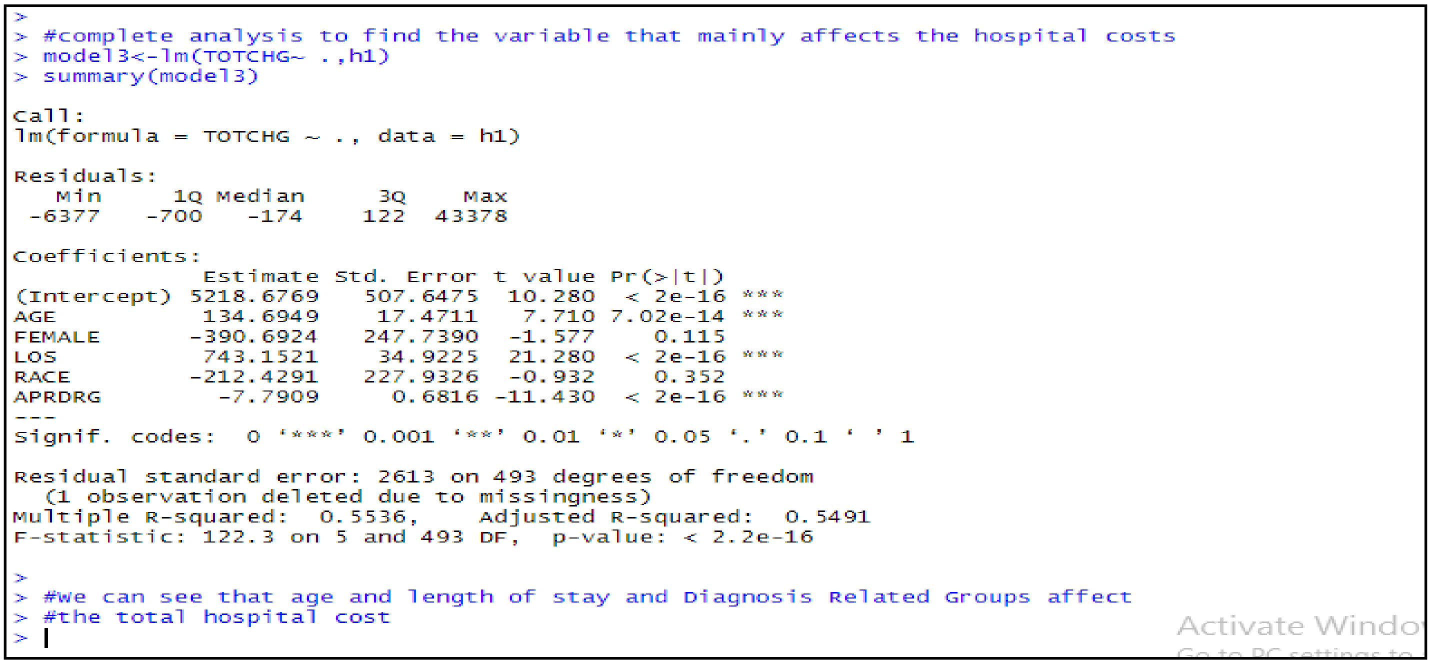


```
>
>
> #find if the length of stay can be predicted from age, gender, and race
>
> model2<-lm(LOS~AGE+FEMALE+RACE,h1)
> summary(model2)

Call:
lm(formula = LOS ~ AGE + FEMALE + RACE, data = h1)

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 the length of stay
> #cannot be predicted from age, gender, and race
>
```

except for the intercept , the very high p-value signifies that the length of stay

cannot be predicted from age, gender, and race

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



```
>
> #complete analysis to find the variable that mainly affects the hospital costs
> model3<-lm(TOTCHG~ .,h1)
> summary(model3)

Call:
lm(formula = TOTCHG ~ ., data = h1)

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

>
> #We can see that age and length of stay and Diagnosis Related Groups affect
> #the total hospital cost
> 
```

We can see that age and length of stay and Diagnosis Related Groups affect the total hospital cost