Report – Project Healthcare

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.

I used a histogram to visually identify the age range more frequently hospitalised. Summary and aggregate functions will return the descriptive statistics in term of average and total expenditure

Code:

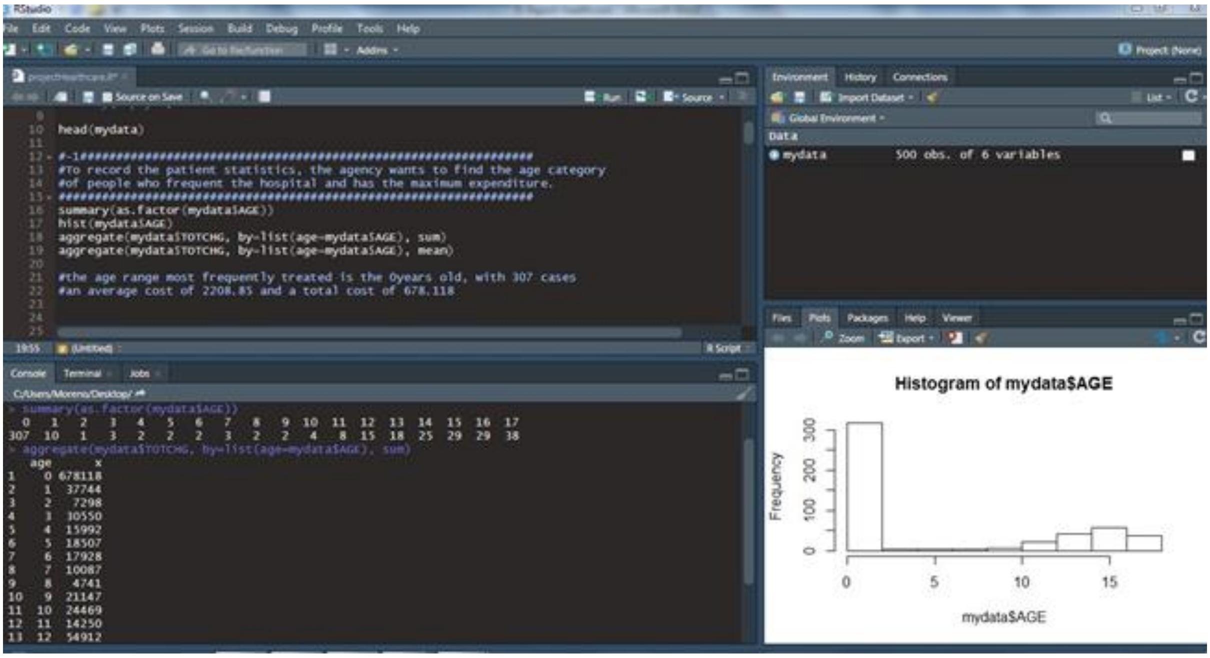
```
setwd('C:/Users/Ilaria/Dropbox/Rcourse')

mydata= read.csv('HospitalCosts.csv')

summary(as.factor(mydata$AGE))

hist(mydata$AGE)

aggregate(mydata$TOTCHG,  by=list(age=mydata$AGE),  sum)

aggregate(mydata$TOTCHG,  by=list(age=mydata$AGE),  mean)
```

Results:

As displayed in the histogram, the age range most frequently treated is the 0-1 years old.

From the summary, we can see that 307 cases belong to that age range. The aggregate function shows that they had an average cost of 2.208,85 and a total cost of 678.118

Output:



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Addins
Project: (None)
Source on Save
Run
Source
head(mydata)
#-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.
###
summary(as.factor(mydata$AGE))
hist(mydata$AGE)
aggregate(mydata$TOTCHG, by=list(age=mydata$AGE), sum)
aggregate(mydata$TOTCHG, by=list(age=mydata$AGE), mean)
#the age range most frequently treated is the 0years old, with 307 cases
#an average cost of 2208.85 and a total cost of 678.118
R Script
Console
Terminal
Jobs
> summary(as.factor(mydata$AGE))
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
307 10 1 3 2 2 2 3 2 2 4 8 15 18 25 29 29 38
> aggregate(mydata$TOTCHG, by=list(age=mydata$AGE), sum)
age x
1 0 678118
2 1 37744
3 2 7298
4 3 30550
5 4 15992
6 5 18507
7 6 17928
8 7 10087
9 8 4741
10 9 21147
11 10 24469
12 11 14250
13 12 54912
Environment History Connections
Import Dataset
Global Environment
List
Data
mydata 500 obs. of 6 variables
Files Plots Packages Help Viewer
Zoom
Export
Histogram of mydata$AGE
Frequency
0
100
200
300
0
5
10
15
mydata$AGE

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

To answer this question I calculated:

a) the diagnosis with higher frequency of hospitalization and higher total cost, with its total and average length of stay;

b) the diagnosis group with the most expensive treatment on average

c) the diagnosis with the longer length of stay and its costs

Codes:

a)

```
hist(mydata$APRDRG)

freq.dia= summary(as.factor(mydata$APRDRG))

max(freq.dia)

sort(freq.dia, decreasing= TRUE)

mydataAPRDRG=as.factor(mydata$APRDRG)

totcost= aggregate(mydata$TOTCHG, by=list(diagnosis=mydata$APRDRG), sum)

totcost[order(totcost$x , decreasing= TRUE),]
```

b)

```
mydataAPRDRG=as.factor(mydata$APRDRG)

meancost= aggregate(mydata$TOTCHG, by=list(diagnosis=mydata$APRDRG), mean)

meancost[order(meancost$x, decreasing=TRUE),]
```

c)

hospitalization= aggregate(mydata$LOS, by=list(diagnosis=mydata$APRDRG),mean)

hospitalization[order(hospitalization$x, decreasing = TRUE),]

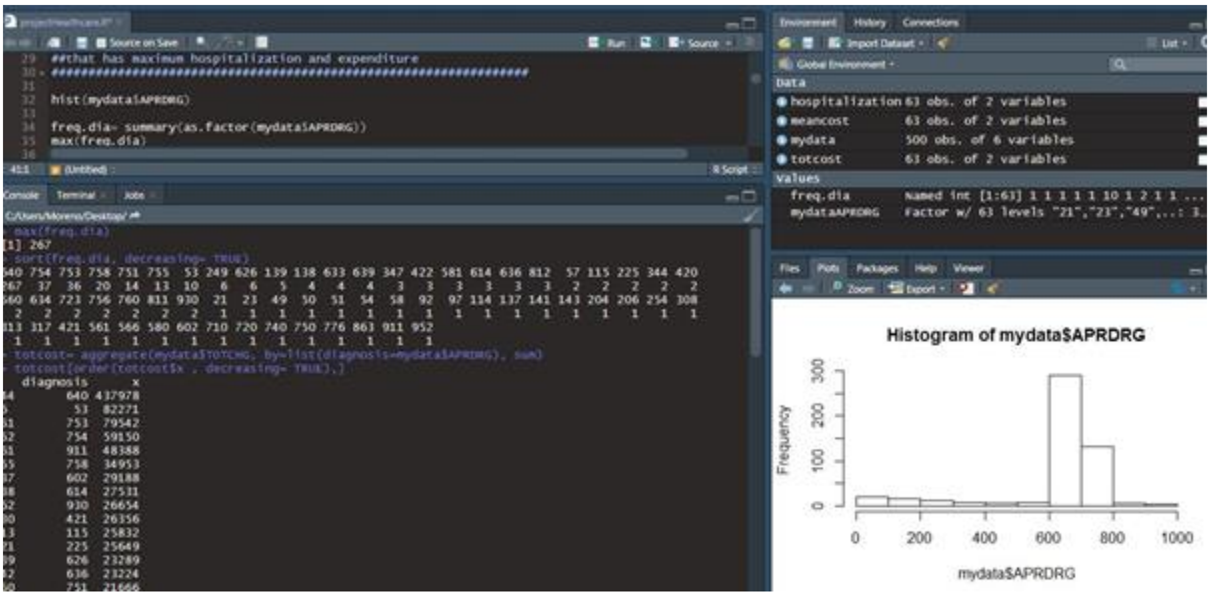
Results:

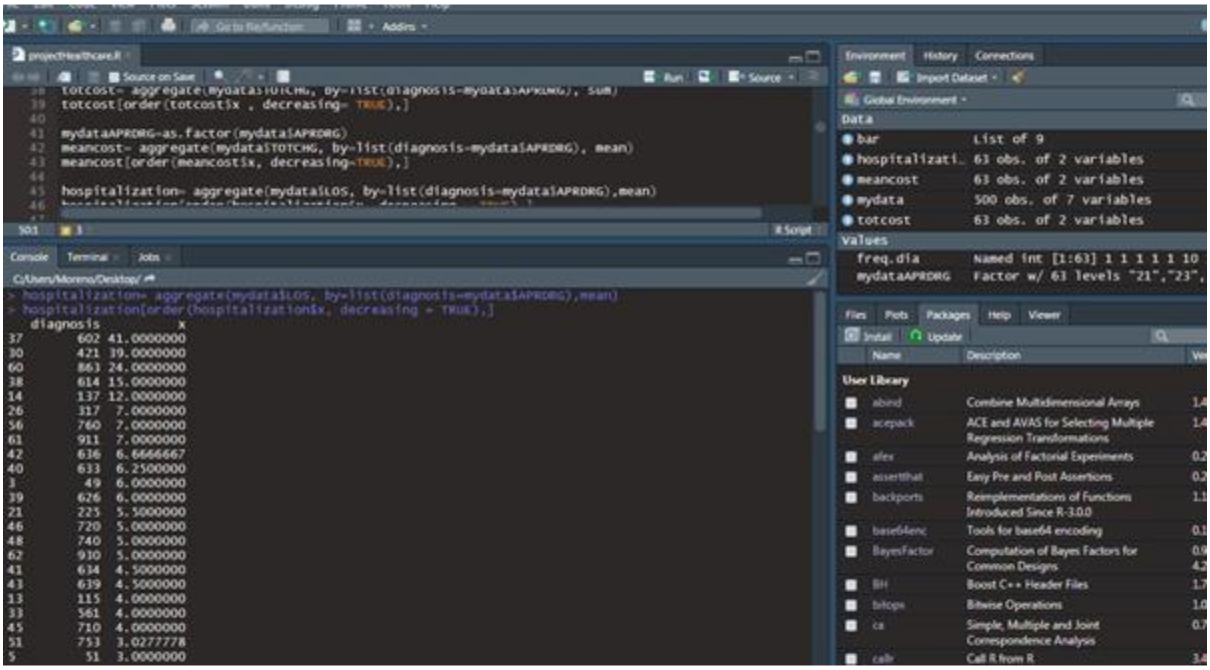
a) The diagnosis group which is more frequently hospitalized is the 640, with 267 cases. This has an impact on the hospital costs as this category represent the one with the higher total cost, of 437.978. The ‘order’ function allows listing the diagnosis groups from the one with the most total expenditure to the least.

b) Despite the group 640 has the highest impact on expenses, on average is does not represent the most expensive treatment, with an average cost of 164.037. The diagnosis group with the most expensive treatment on average is the group 911, with an average cost of 48.388.

c) On average, the 602 group is the one with longer permanence, 41 days on average. Only one case is reported with a cost of 29.188

Output:





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.

To answer this question, I visually explored the data by using a barlot, and summarized them to obtain descriptive statistics. I used SummarySE which returns also sd and frequency. I also calculated the total costs through aggregate function. I run a linear regression model to check whether race was statistically impacting hospitalization costs. As NAs were present, I removed those data points from the analysis

Codes:

```
str(mydata)

mydata$RACE=as.factor(mydata$RACE)  #transform the variable



racecost= summarySE(mydata,'TOTCHG','RACE')

racecost = racecost[-which(is.na(racecost$RACE)), ] #remove NAs

print(racecost) #calculate the average cost
```

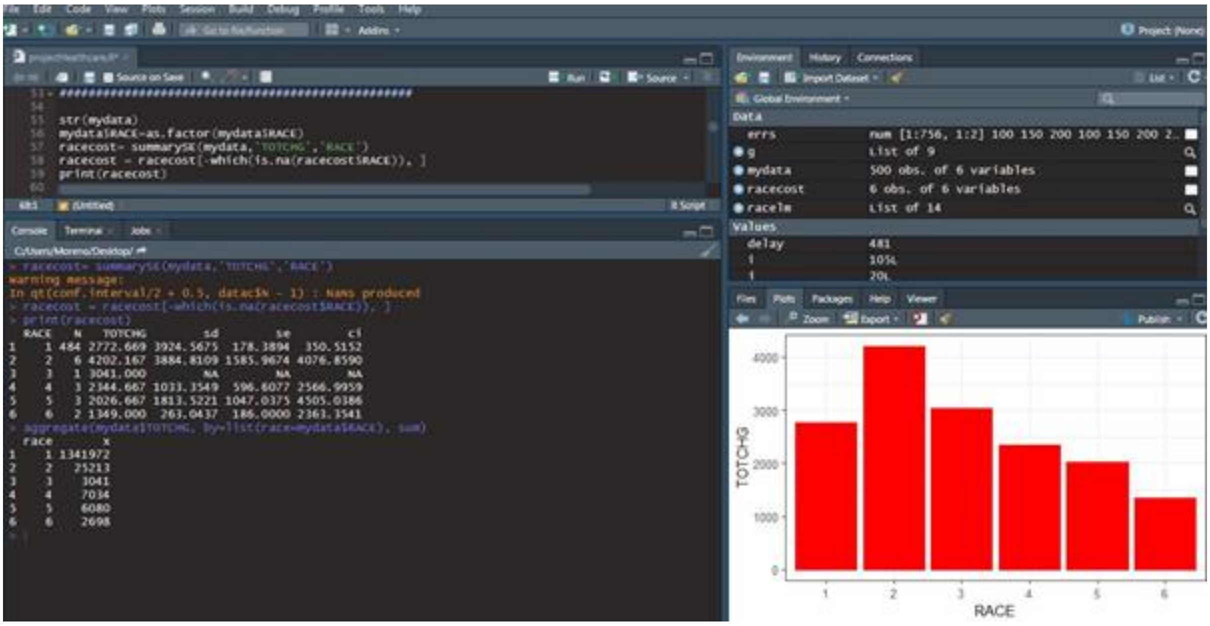
```
g = ggplot(racecost, aes(RACE, TOTCHG))

g + geom_col(colour = "red", fill = "red") +

  theme_bw(base_size = 12)


aggregate(mydata$TOTCHG, by=list(race=mydata$RACE), sum) #calculate the total cost


racelm = lm(TOTCHG~RACE, mydata, na.action=na.omit) #inferential

summary(racelm)
```

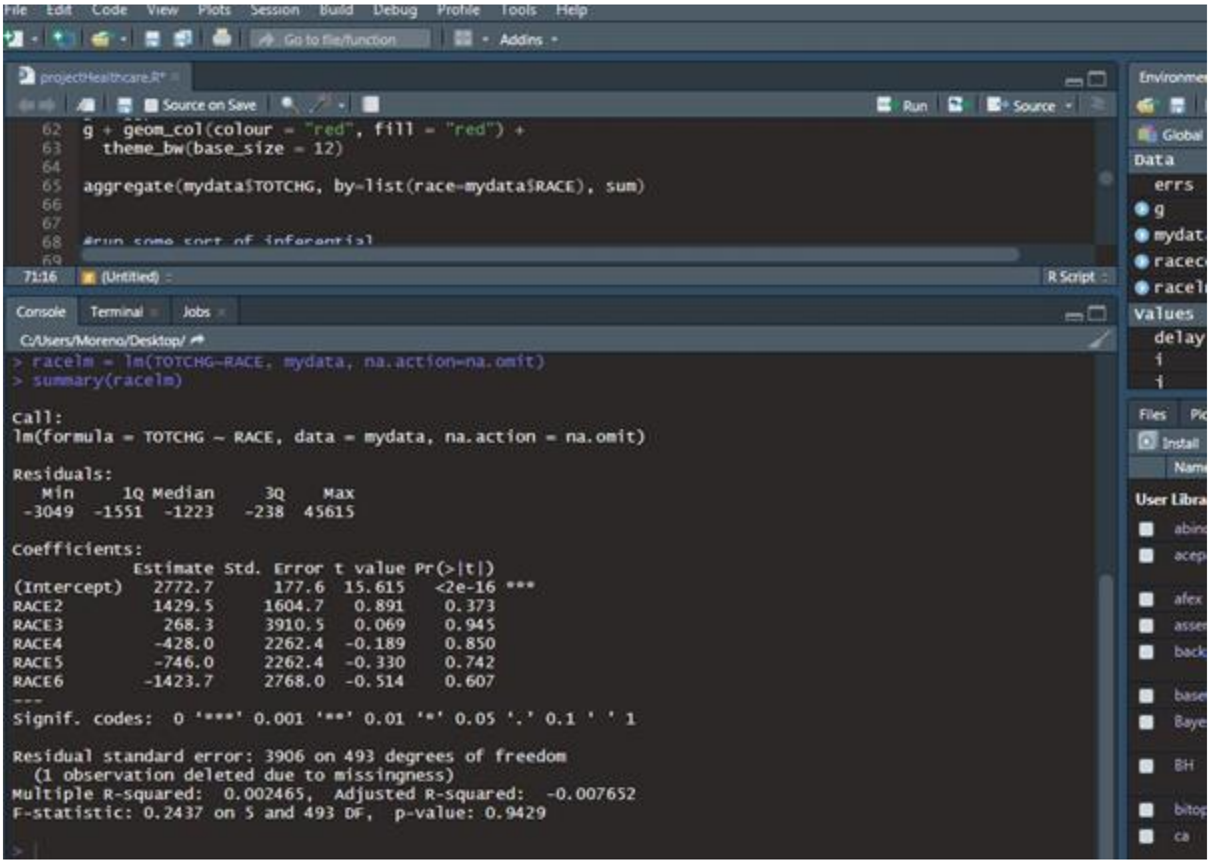
Results:

As shown in the graph, the race with the higher average cost is the race2, with an expenditure of 4202,2 and 6 cases. Race1 has the higher total cost of 1.341.972 and 484 cases (on average 2.772,7).

A linear model shows that race does not significantly affect hospitalization cost, although its accuracy is very low (R2 = -0.007652)

Outputs:


str(mydata)
mydata$RACE<-as.factor(mydata$RACE)
racecost<- summarySE(mydata,'TOTCHG','RACE')
racecost = racecost[-which(is.na(racecost$RACE)),]
print(racecost)
Warning message:
In qt(conf.interval/2 + 0.5, datac$N - 1) : NaNs produced
RACE N TOTCHG sd se ci
1 1 484 2772.669 3924.5675 178.3894 350.5152
2 2 6 4202.167 3884.8109 1585.9674 4076.8590
3 3 1 3041.000 NA NA NA
4 4 3 2344.667 1033.3549 596.6077 2566.9959
5 5 3 2026.667 1813.5221 1047.0375 4505.0386
6 6 2 1349.000 263.0437 186.0000 2363.3541
race x
1 1 1341972
2 2 25213
3 3 3041
4 4 7034
5 5 6080
6 6 2698
errs num [1:756, 1:2] 100 150 200 100 150 200 2...
g List of 9
mydata 500 obs. of 6 variables
racecost 6 obs. of 6 variables
racelm List of 14
delay 481
TOTCHG
RACE


g + geom_col(colour = "red", fill = "red") +
theme_bw(base_size = 12)
aggregate(mydata$TOTCHG, by=list(race=mydata$RACE), sum)
> racelm = lm(TOTCHG~RACE, mydata, na.action=na.omit)
> summary(racelm)
Call:
lm(formula = TOTCHG ~ RACE, data = mydata, na.action = na.omit)
Residuals:
Min 1Q Median 3Q Max
-3049 -1551 -1223 -238 45615
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 2772.7 177.6 15.615 <2e-16 ***
RACE2 1429.5 1604.7 0.891 0.373
RACE3 268.3 3910.5 0.069 0.945
RACE4 -428.0 2262.4 -0.189 0.850
RACE5 -746.0 2262.4 -0.330 0.742
RACE6 -1423.7 2768.0 -0.514 0.607

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 3906 on 493 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared: 0.002465, Adjusted R-squared: -0.007652
F-statistic: 0.2437 on 5 and 493 DF, p-value: 0.9429

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

I transformed age and gender into factor and used the functions SummarySE to calculate the average costs by age and gender, and the aggregate function to calculate the total costs by age and gender. Outputs are ordered in decreasing order by using the order function. A barplot shows how the average expenditure differ by gender.

Codes:

```
mydata$AGE=as.factor(mydata$AGE)

hosdemo= aggregate(mydata$TOTCHG, by=list(age=mydata$AGE),sum)#total cost by age

hosdemo[order(hosdemo$x, decreasing= TRUE),]


hosage= summarySE(mydata,'TOTCHG','AGE')#average cost by age

hosage[order(hosage$TOTCHG, decreasing = TRUE),]


mydata$FEMALE= as.factor(mydata$FEMALE)

bar <- ggplot(mydata, aes(x = FEMALE, y = TOTCHG))

bar+stat_summary(fun= mean, geom = "bar", fill = "lightgrey", colour = "black") +

  stat_summary(fun.data = mean_cl_normal, geom = 'errorbar', width=0.3)


gndcost= summarySE(mydata,'TOTCHG','FEMALE')

print(gndcost)


totgndcost=aggregate(mydata$TOTCHG, by=list(gender=mydata$FEMALE), sum)

print(totgndcost)
```

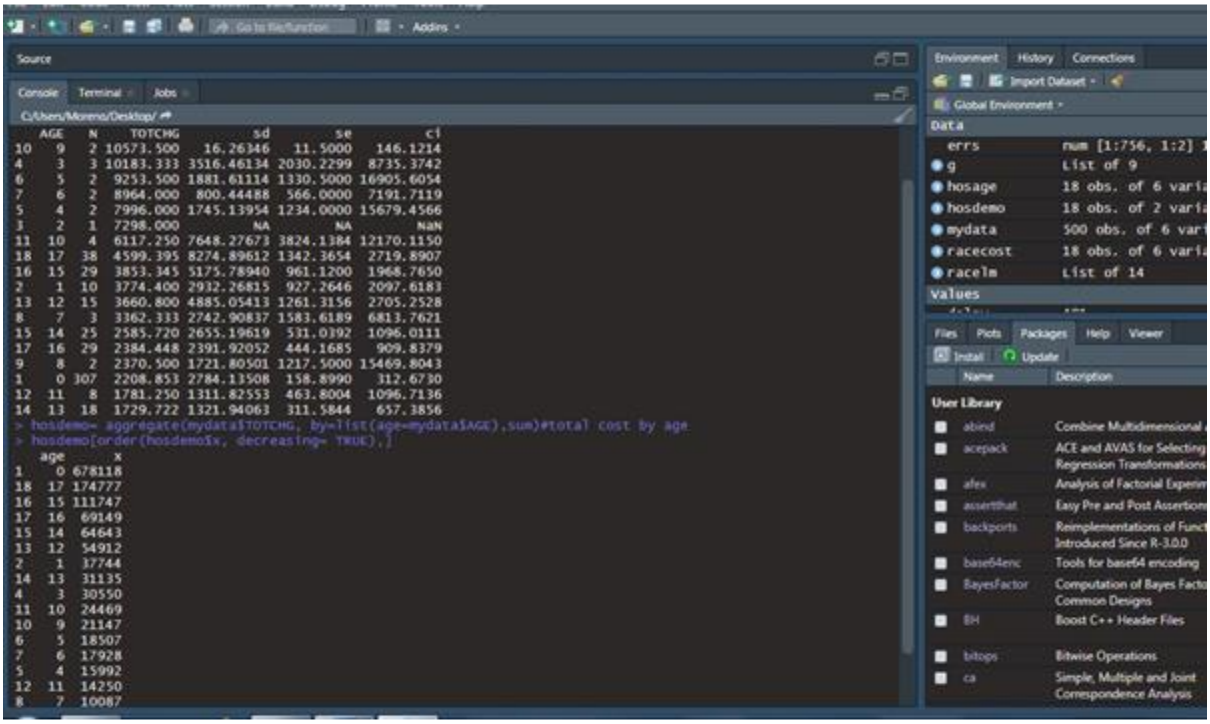
Results:

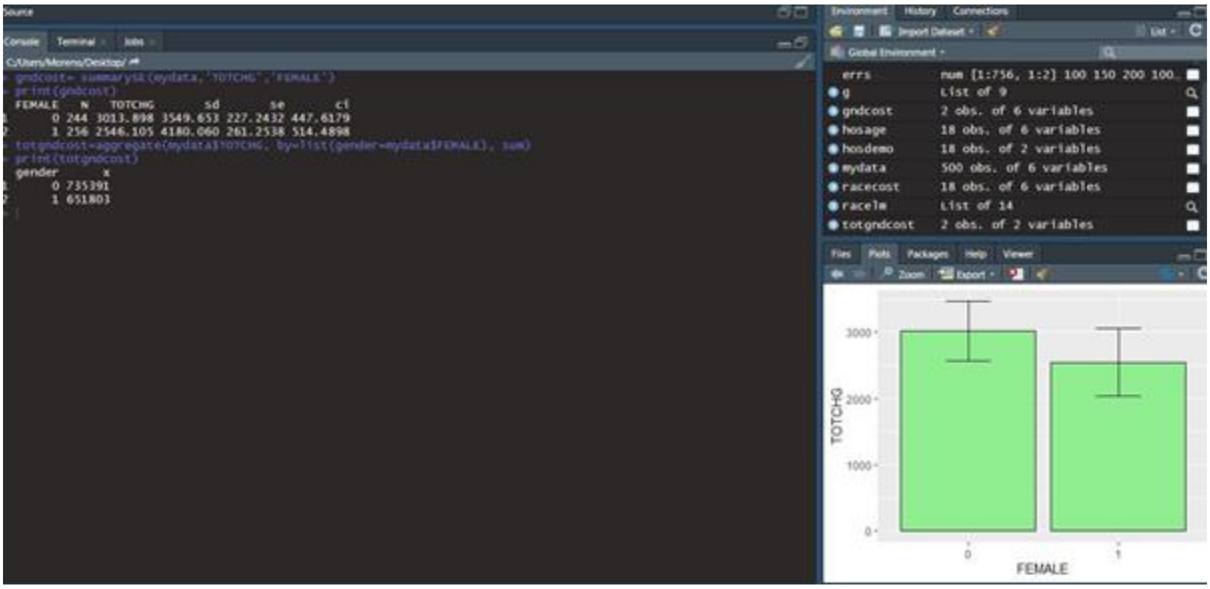
The age range with the higher cost on average is the age range 9-10, with 10573.500 and 2 cases.

The age range 0-1, as previously sees, is the one with higher hospitalization frequency and consequently it is the age range with the higher total cost.

Males have higher cost, on average (3013.898) and in total (735.391) than females.

Output:





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.

To answer this question I run a multiple linear model, with age, gender and race as predictor of length of stay.

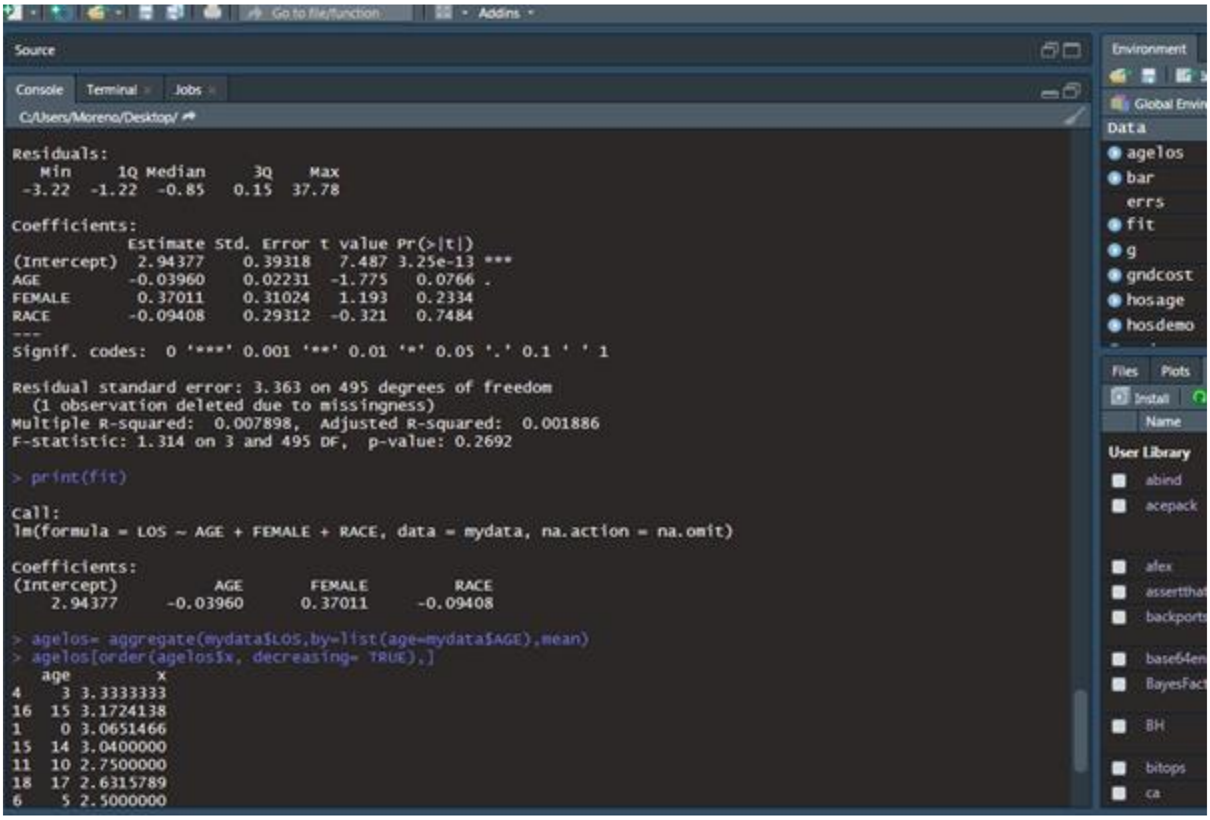
Codes:

```
fit <- lm(LOS ~ AGE + FEMALE + RACE, data=mydata, na.action = na.omit)

summary(fit) # show results

print(fit)


unique(mydata$AGE)
```

Results:

None of the variables seems to predict the length of stay. Age approaches significance level, with the age of 3 having the longer length of stay on average (3.3 days). The model accuracy is very low.

Output:



```
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

> print(fit)

Call:
lm(formula = LOS ~ AGE + FEMALE + RACE, data = mydata, na.action = na.omit)

Coefficients:
(Intercept)          AGE       FEMALE         RACE
    2.94377     -0.03960      0.37011     -0.09408

> agelos= aggregate(mydata$LOS,by=list(age=mydata$AGE),mean)
> agelos[order(agelos$x, decreasing= TRUE),]
   age         x
4    3 3.3333333
16  15 3.1724138
1    0 3.0651466
15  14 3.0400000
11  10 2.7500000
18  17 2.6315789
6    5 2.5000000
```

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

In this case, we can run a linear model including all the variable as predictors of the costs. I ran a separate model with diagnosis as a predictor in order to break down the diagnosis affecting the costs.

Codes:

fit2 <- lm(TOTCHG ~ ., data=mydata)

summary(fit2)

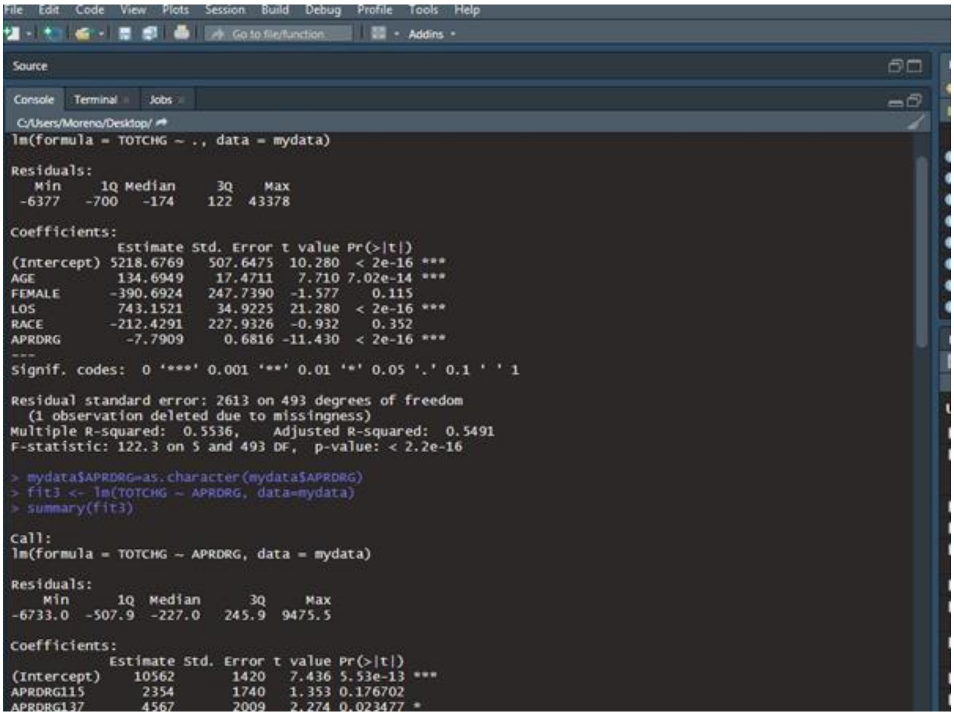
mydata$APRDRG=as.character(mydata$APRDRG)

fit3 <- lm(TOTCHG ~ APRDRG, data=mydata)

Results:

The model is 55% accurate. Age, length of stay and diagnosis significantly predict the hospitalization cost. Some diagnosis have a significantly higher impact on the costs.

Output:



Console Terminal Jobs

C:/Users/Moreno/Desktop/

```
APRDRG602     18626      2009   9.272  < 2e-16 ***
APRDRG614     -1385      1640  -0.844 0.398883
APRDRG626     -6680      1534  -4.354 1.66e-05 ***
APRDRG633     -6164      1588  -3.882 0.000120 ***
APRDRG634     -5586      1740  -3.211 0.001420 **
APRDRG636     -2821      1640  -1.720 0.086180 .
APRDRG639     -7409      1588  -4.665 4.10e-06 ***
APRDRG640     -8922      1423  -6.269 8.71e-10 ***
APRDRG710     -2339      2009  -1.164 0.244893
APRDRG720      3681      2009   1.832 0.067559 .
APRDRG723     -7918      1740  -4.551 6.92e-06 ***
APRDRG740       563      2009   0.280 0.779399
APRDRG750     -8809      2009  -4.385 1.45e-05 ***
APRDRG751     -9014      1470  -6.131 1.95e-09 ***
APRDRG753     -8352      1440  -5.800 1.27e-08 ***
APRDRG754     -8963      1439  -6.227 1.12e-09 ***
APRDRG755     -9703      1474  -6.583 1.33e-10 ***
APRDRG756     -9815      1740  -5.642 3.02e-08 ***
APRDRG758     -8814      1455  -6.056 3.01e-09 ***
APRDRG760     -6426      1740  -3.694 0.000249 ***
APRDRG776     -9369      2009  -4.664 4.13e-06 ***
APRDRG811     -8643      1740  -4.968 9.71e-07 ***
APRDRG812     -7387      1640  -4.504 8.56e-06 ***
APRDRG863      2478      2009   1.234 0.218012
APRDRG911     37826      2009  18.831  < 2e-16 ***
APRDRG92       1462      2009   0.728 0.467115
APRDRG930      2765      1740   1.589 0.112688
APRDRG952     -5729      2009  -2.852 0.004550 **
APRDRG97      -1032      2009  -0.514 0.607684
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1420 on 437 degrees of freedom
Multiple R-squared:  0.8831,	Adjusted R-squared:  0.8666
F-statistic: 53.27 on 62 and 437 DF,  p-value: < 2.2e-16

>
```