

Health care Cost Analysis

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

Solution 1.

The `as.factor()` is called to make sure that the categories are not treated as numbers. The package “ggplot2” is used to display the histogram.

Code:

```
library(readxl) # To read an excel file.
```

```
Hospital <- read_excel("D:/Simplilearn/Project Data Sets/7/Hospital.xlsx")
```

```
# Location of excel file.
```

```
View(Hospital) # To view the inputted dataset.
```

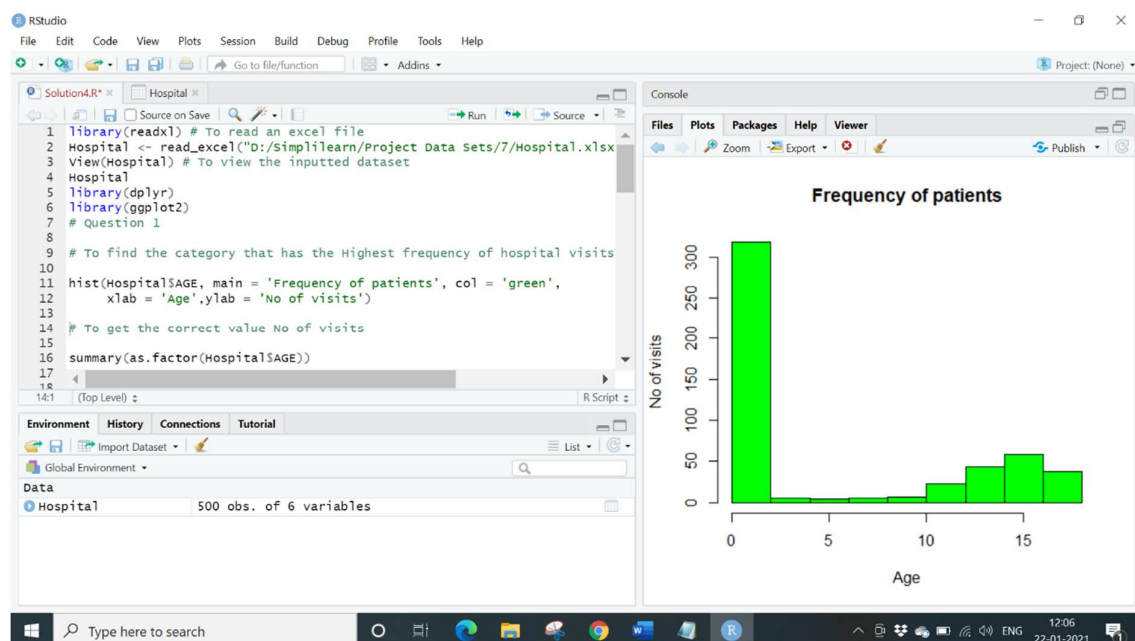
```
Hospital # To get the tibble 6*6 in the console.
```

```
library(dplyr)
```

```
library(ggplot2)
```

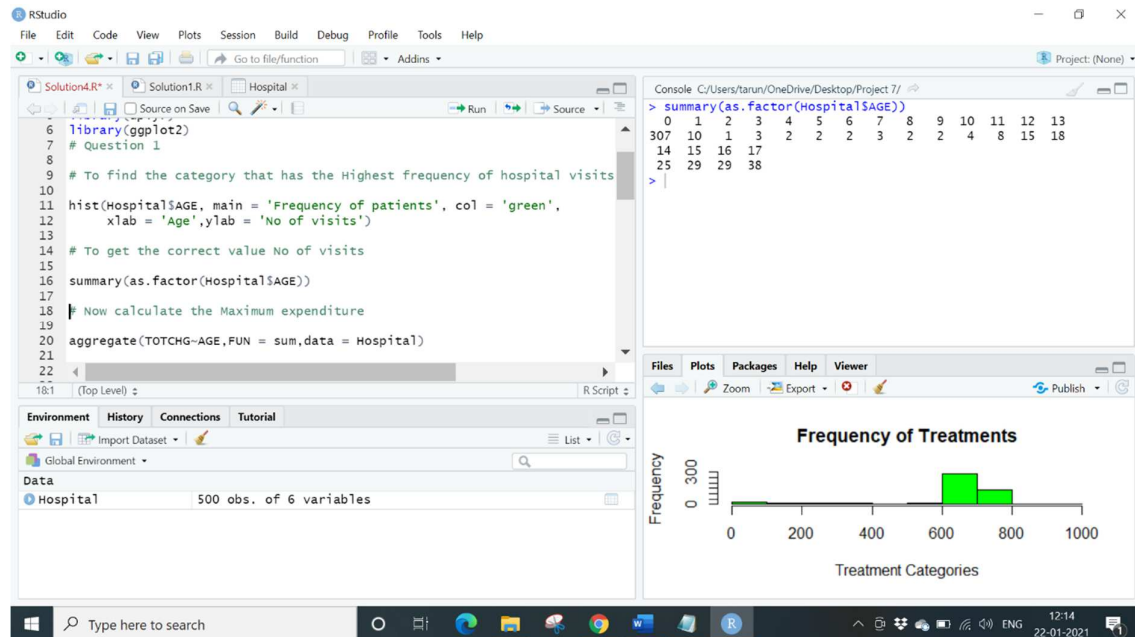
```
# To find the category that has the Highest frequency of hospital visits
```

```
hist(Hospital$AGE, main = "Frequency of patients", col = "green", xlab = "Age",  
ylab = "No of visits")
```



To get the correct value No of visits

summary(as.factor(Hospital\$AGE))



Conclusion:

From the graph that is displayed, we can see that (0) infants have the maximum frequency of hospital visit.

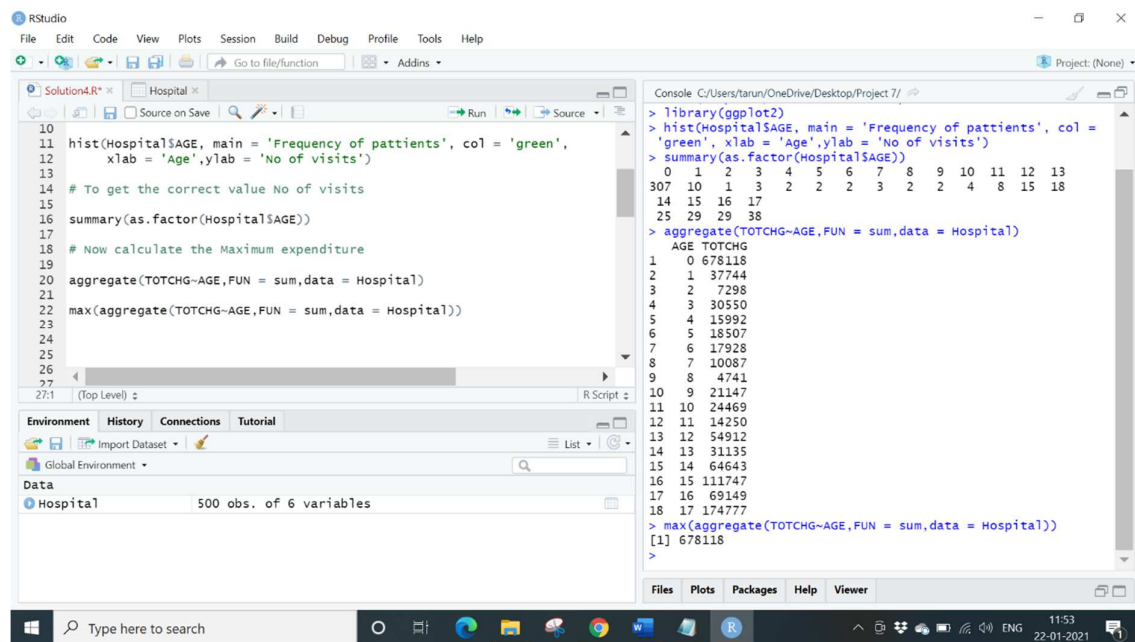
Now to calculate maximum expenditure:

Aggregate function is used to add the expenditure from each age and then max function used to find highest costs.

Code:

```
aggregate (TOTCHG~AGE, FUN=sum, data = Hospital)
```

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



Conclusion:

So again, result is age group 0 (infant) for maximum expenditure.

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.

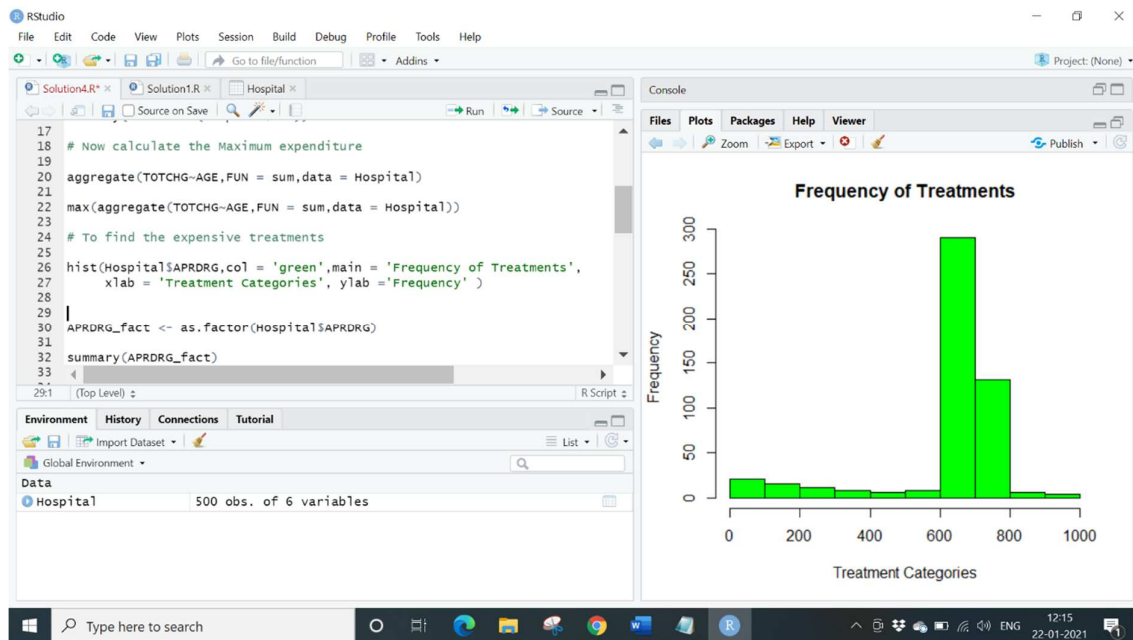
Solution 2.

Code:

```

hist (Hospital$APRDRG, col = 'green', main = 'Frequency of Treatments',
      xlab = 'Treatment Categories', ylab = 'Frequency')

```



The `as. factor ()` is called to make sure that the categories are not treated as numbers.

Code:

```
APRDRG_fact<-as. factor (Hospital$APDRG)
```

```
summary (APRDRG_fact)
```

```
which.max (summary (APRDRG_fact))
```

```
df <-aggregate (TOTCHG~APRDRG, FUN = sum, data=hospital)
```

```
df
```

```
df[which.max(df$TOTCHG),]
```

```

29 # The as.factor() is called to make sure that the categories
30 # are not treated as numbers.
31 APRDRG_factor <- as.factor(Hospital$APRDRG)
32
33 summary(APRDRG_factor)
34
35 which.max(summary(APRDRG_factor))
36
37 df<-aggregate(TOTCHG~APRDRG,FUN = sum,data = Hospital)
38
39 df
40
41 df[which.max(df$TOTCHG),]
42
43 Hospital <- na.omit(Hospital)
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99

```

```

> xlab = 'Treatment Categories', ylab = 'Frequency' )
> APRDRG_factor <- as.factor(Hospital$APRDRG)
> summary(APRDRG_factor)
21 23 49 50 51 53 54 57 58 92 97 114 115 137
1 1 1 1 1 10 1 2 1 1 1 1 2 1
138 139 141 143 204 206 225 249 254 308 313 317 344 347
4 5 1 1 1 1 2 6 1 1 1 1 2 3
420 421 422 560 561 566 580 581 602 614 626 633 634 636
2 1 3 2 1 1 1 3 1 3 6 4 2 3
639 640 710 720 723 740 750 751 753 754 755 756 758 760
4 267 1 1 2 1 1 14 36 37 13 2 20 2
776 811 812 863 911 930 952
1 2 3 1 1 2 1
> which.max(summary(APRDRG_factor))
640
44
> df<-aggregate(TOTCHG~APRDRG, FUN = sum,data = Hospital)
> df
  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

```

```

29 # The as.factor() is called to make sure that the categories
30 # are not treated as numbers.
31 APRDRG_factor <- as.factor(Hospital$APRDRG)
32
33 summary(APRDRG_factor)
34
35 which.max(summary(APRDRG_factor))
36
37 df<-aggregate(TOTCHG~APRDRG,FUN = sum,data = Hospital)
38
39 df
40
41 df[which.max(df$TOTCHG),]
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99

```

```

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
> df[which.max(df$TOTCHG),]
  APRDRG TOTCHG
44     640 437978

```

Conclusion: So, the category 640 has the maximum hospitalizations along with this it also has the highest hospitalization cost.

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.

Solution 3.

Remove the “NA” values from our database, then by using `as.factor()` the Race variable to generate a summary to verify whether race made an impact on the hospital costs we will use ANOVA function with TOTCHG as dependent variable and RACE as grouping variable.

Code:

```
Hospital <- na.omit(hospital) #first we remove “NA” values
```

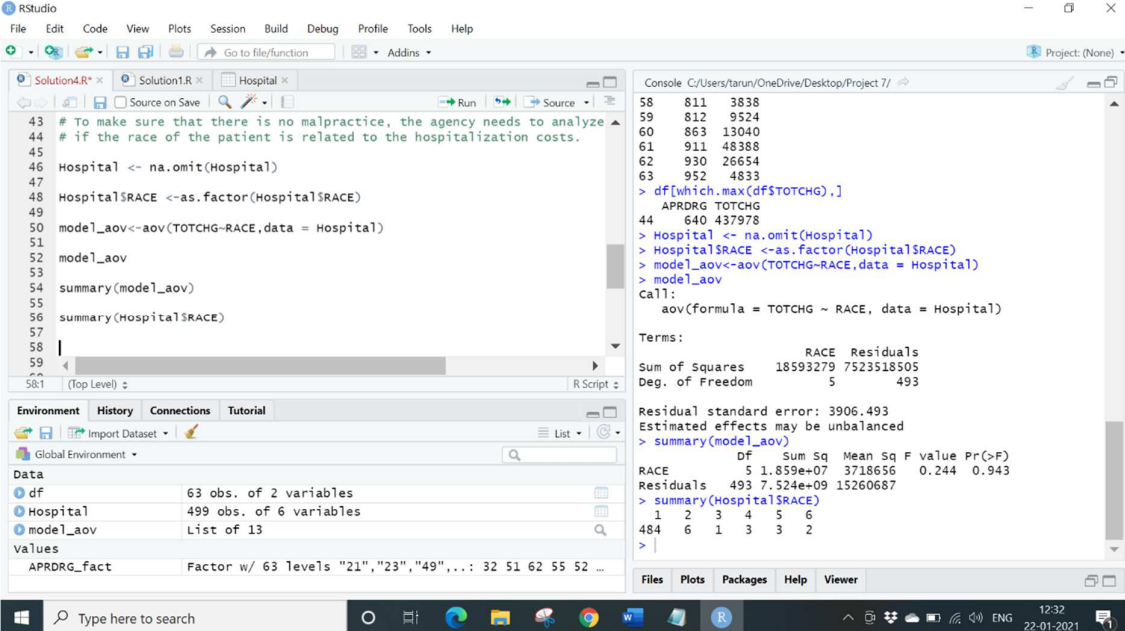
```
Hospital$RACE <- as.factor(Hospital$RACE)
```

```
model_aov <- aov(TOTCHG ~ RACE, data = Hospital)
```

```
model_aov #ANOVA RESULTS
```

```
summary(model_aov)
```

```
summary(Hospital$RACE)
```



The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains the R code for removing NA values, converting RACE to a factor, fitting an ANOVA model, and generating summaries.
- Console:** Displays the output of the code, including the ANOVA table and the summary of the RACE factor.
- Environment Pane:** Shows the objects created in the global environment: `df` (63 obs. of 2 variables), `Hospital` (499 obs. of 6 variables), and `model_aov` (List of 13).

Console Output:

```
58 811 3838
59 812 9524
60 863 13040
61 911 48388
62 930 26654
63 952 4833
> df[which.max(df$TOTCHG),]
  APRDRG TOTCHG
44    640 437978
> Hospital <- na.omit(Hospital)
> Hospital$RACE <- as.factor(Hospital$RACE)
> model_aov <- aov(TOTCHG ~ RACE, data = Hospital)
> model_aov
Call:
aov(formula = TOTCHG ~ RACE, data = Hospital)

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

Residual standard error: 3906.493
Estimated effects may be unbalanced
> summary(model_aov)
              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(Hospital$RACE)
  1  2  3  4  5  6
484 6  1  3  3  2
>
```

Environment Pane:

Object	Description
<code>df</code>	63 obs. of 2 variables
<code>Hospital</code>	499 obs. of 6 variables
<code>model_aov</code>	List of 13

Conclusion: The result shows that there is no relationship between race and hospital costs, thereby accepting the Null hypothesis.

Question 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.

Solution 4.

To analyze the costs we will use linear regression with TOTCHG(Cost) and independent variable along with AGE and Female as dependent variables

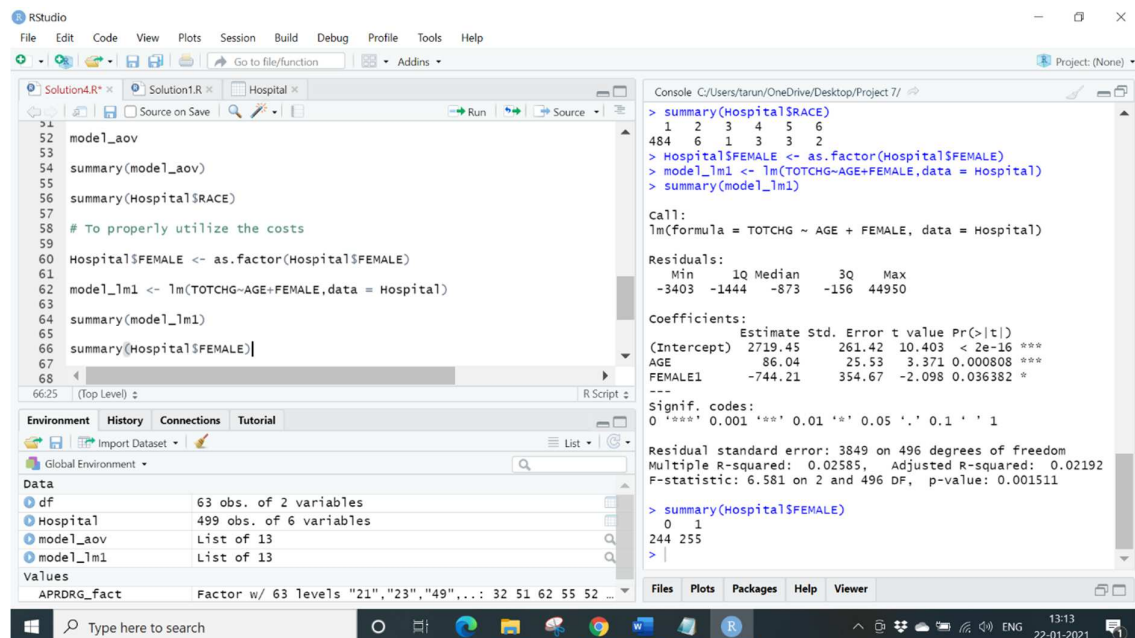
Code:

```
Hospital$FEMALE <- as.factor(Hospital$FEMALE)
```

```
model_lm1 <- lm(TOTCHG~AGE+FEMALE, data =Hospital) #calling Regression function
```

```
summary(model_lm1)
```

```
summary(hosp$FEMALE) #comapring genders
```



The screenshot displays the RStudio interface with the following components:

- Script Editor:** Contains the R code for the analysis, including data preparation and model fitting.
- Console:** Shows the output of the R commands, including the summary of the linear model and the frequency table for the female variable.
- Environment:** Lists the objects in the global environment, including the data frame and the fitted model.

```
51 Hospital$FEMALE <- as.factor(Hospital$FEMALE)
52 model_lm1 <- lm(TOTCHG~AGE+FEMALE, data =Hospital)
53 summary(model_lm1)
54 summary(Hospital$FEMALE)
```

Console Output:

```
> summary(Hospital$FEMALE)
 1  2  3  4  5  6 
484 6  1  3  3  2 
> Hospital$FEMALE <- as.factor(Hospital$FEMALE)
> model_lm1 <- lm(TOTCHG~AGE+FEMALE,data = Hospital)
> summary(model_lm1)

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

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 ***
FEMALE1     -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(Hospital$FEMALE)
 0  1 
244 255
```

Environment:

Object	Class	Attributes
df	data.frame	63 obs. of 2 variables
Hospital	data.frame	499 obs. of 6 variables
model_lm1	lm	List of 13
model_lm1	lm	List of 13

Conclusion: There are equal number of Females and Males and on an average (based on the negative coefficient values) females makes lesser hospital costs than males.

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.

Solution 5.

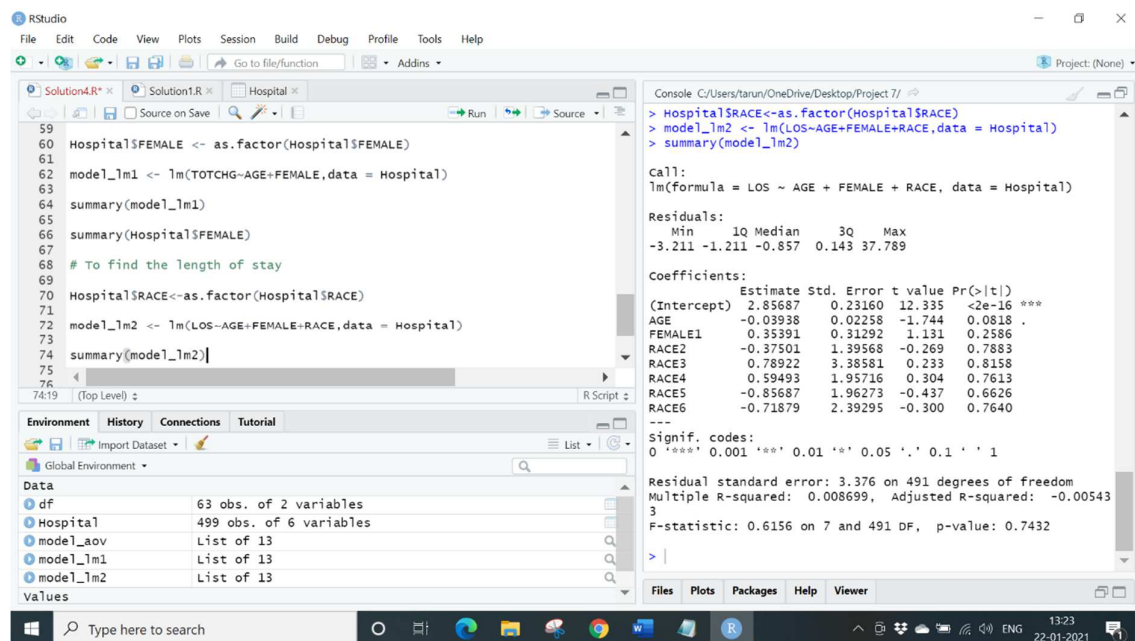
To find the length of stay we are using linear regression. Here length of stay is the dependent variable and age, gender and race are independent variables

Code:

```
Hospital$RACE <- as.factor(Hospital$RACE)
```

```
model_lm2 <- lm(LOS~AGE+FEMALE+RACE, data = Hospital)
```

```
summary(model_lm2)
```



The screenshot shows the RStudio interface with the following content:

```
59  
60 Hospital$FEMALE <- as.factor(Hospital$FEMALE)  
61  
62 model_lm1 <- lm(TOTCHG~AGE+FEMALE,data = Hospital)  
63  
64 summary(model_lm1)  
65  
66 summary(Hospital$FEMALE)  
67  
68 # To find the length of stay  
69  
70 Hospital$RACE<-as.factor(Hospital$RACE)  
71  
72 model_lm2 <- lm(LOS~AGE+FEMALE+RACE,data = Hospital)  
73  
74 summary(model_lm2)
```

The console output for `summary(model_lm2)` is as follows:

```
call:  
lm(formula = LOS ~ AGE + FEMALE + RACE, data = Hospital)  
  
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 .  
FEMALE1      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:  -0.00543  
F-statistic: 0.6156 on 7 and 491 DF, p-value: 0.7432
```

The Environment pane shows the following objects:

Object	Details
df	63 obs. of 2 variables
Hospital	499 obs. of 6 variables
model_aov	List of 13
model_lm1	List of 13
model_lm2	List of 13

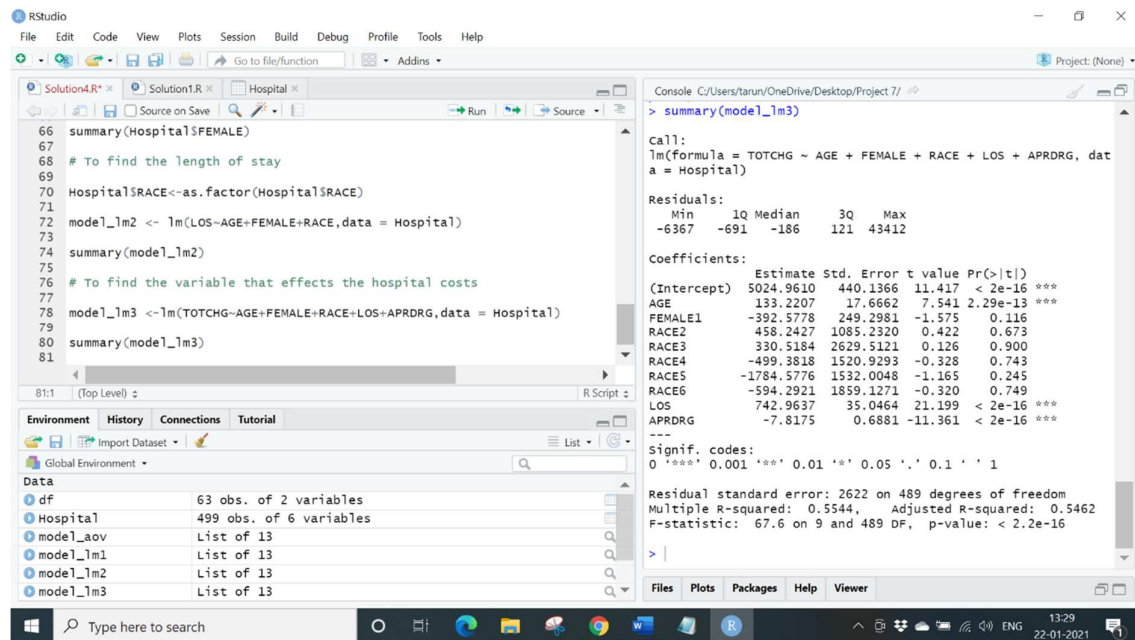
Conclusion: All independent variables are quite high thus signifying that there is no linear relationship between the given variables, so we can't predict length of stay of a patient based on age, gender and race.

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

Solution 6. To find the variable that mainly affects the hospital costs we use linear regression thus TOTCHG becomes dependent rest all becomes independent.

Code:

```
model_lm3 <- lm (TOTCHG~AGE+FEMALE+RACE+LOS+APRDRG, data = Hospital)
summary(model_lm3)
```



The screenshot shows the RStudio interface. The script editor on the left contains the following code:

```
66 summary(Hospital$FEMALE)
67
68 # To find the length of stay
69
70 Hospital$RACE<-as.factor(Hospital$RACE)
71
72 model_lm2 <- lm(LOS~AGE+FEMALE+RACE,data = Hospital)
73
74 summary(model_lm2)
75
76 # To find the variable that effects the hospital costs
77
78 model_lm3 <-lm(TOTCHG~AGE+FEMALE+RACE+LOS+APRDRG,data = Hospital)
79
80 summary(model_lm3)
81
```

The console on the right displays the output of the `summary(model_lm3)` command:

```
> summary(model_lm3)

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

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 ***
FEMALE1      -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
```

Conclusion: Age and Length of stay affect the total hospital costs.

Code:

```
library(readxl) # To read an excel file
Hospital <- read_excel ("D:/Simplilearn/Project Data Sets/7/Hospital.xlsx")
# Location of excel file

View (Hospital) # To view the inputted dataset

Hospital

library(dplyr)
library(ggplot2)

# Question 1

# To find the category that has the highest frequency of hospital visits and has
the maximum expenditure.

hist (Hospital$AGE, main = 'Frequency of patients', col = 'green',
      xlab = 'Age', ylab = 'No of visits')

# To get the correct value No of visits.

summary (as. factor (Hospital$AGE))

# Now calculate the Maximum expenditure.

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

# Question 2

# To find out the expensive treatments

hist (Hospital$APRDRG, col = 'green', main = 'Frequency of Treatments',
      xlab = 'Treatment Categories', ylab = 'Frequency')

# The as. factor () is called to make sure that the categories are not treated as
numbers.

APRDRG_fact <- as. factor (Hospital$APRDRG)

summary (APRDRG_fact)

which.max (summary (APRDRG_fact))

df <- aggregate (TOTCHG~APRDRG, FUN = sum, data = Hospital)
```

df

```
df[which.max(df$TOTCHG),]
```

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.

```
Hospital <- na.omit (Hospital)
```

```
Hospital$RACE <- as.factor (Hospital$RACE)
```

```
model_aov <- aov (TOTCHG~RACE, data = Hospital)
```

```
model_aov
```

```
summary(model_aov)
```

```
summary (Hospital$RACE)
```

Question 4

To properly utilize the costs

```
Hospital$FEMALE <- as.factor (Hospital$FEMALE)
```

```
model_lm1 <- lm (TOTCHG~AGE+FEMALE, data = Hospital)
```

```
summary(model_lm1)
```

```
summary (Hospital$FEMALE)
```

Question 5

To find the length of stay

```
Hospital$RACE <- as.factor (Hospital$RACE)
```

```
model_lm2 <- lm (LOS~AGE+FEMALE+RACE, data = Hospital)
```

```
summary(model_lm2)
```

Question 6

To find the variable that effects the hospital costs

```
model_lm3 <- lm (TOTCHG~AGE+FEMALE+RACE+LOS+APRDRG, data = Hospital)
```

```
summary(model_lm3)
```

(End of Project)