ANALYSE THE HEALTHCARE COST AND UTILIZATION IN WISCONSIN HOSPITAL

R Studio Tool is used for solutions.

Question: A nationwide survey of hospital costs conducted by the US Agency for Healthcare consists of hospital records of inpatient samples. The given data is restricted to the city of Wisconsin and relates to patients in the age group 0-17 years. The agency wants to analyse the data to research on the research the healthcare costs and their utilization.

CODE

```
## To know working directory path
getwd()
```

OUTPUT

```
> ## To know working directory path
> getwd()
[1] "/home/radsrinivasan_gmail_com"
```

CODE

```
# Loading a HospitalCosts.csv file
# Note the sep argument usually needs "," because .csv files are mostly comma separated
HealthCare <- read.csv("/home/radsrinivasan_gmail_com/R/HospitalCosts.csv", sep=",", header=TRUE)
#Loading HospitalCosts.csv file data into variable HealthCare
```

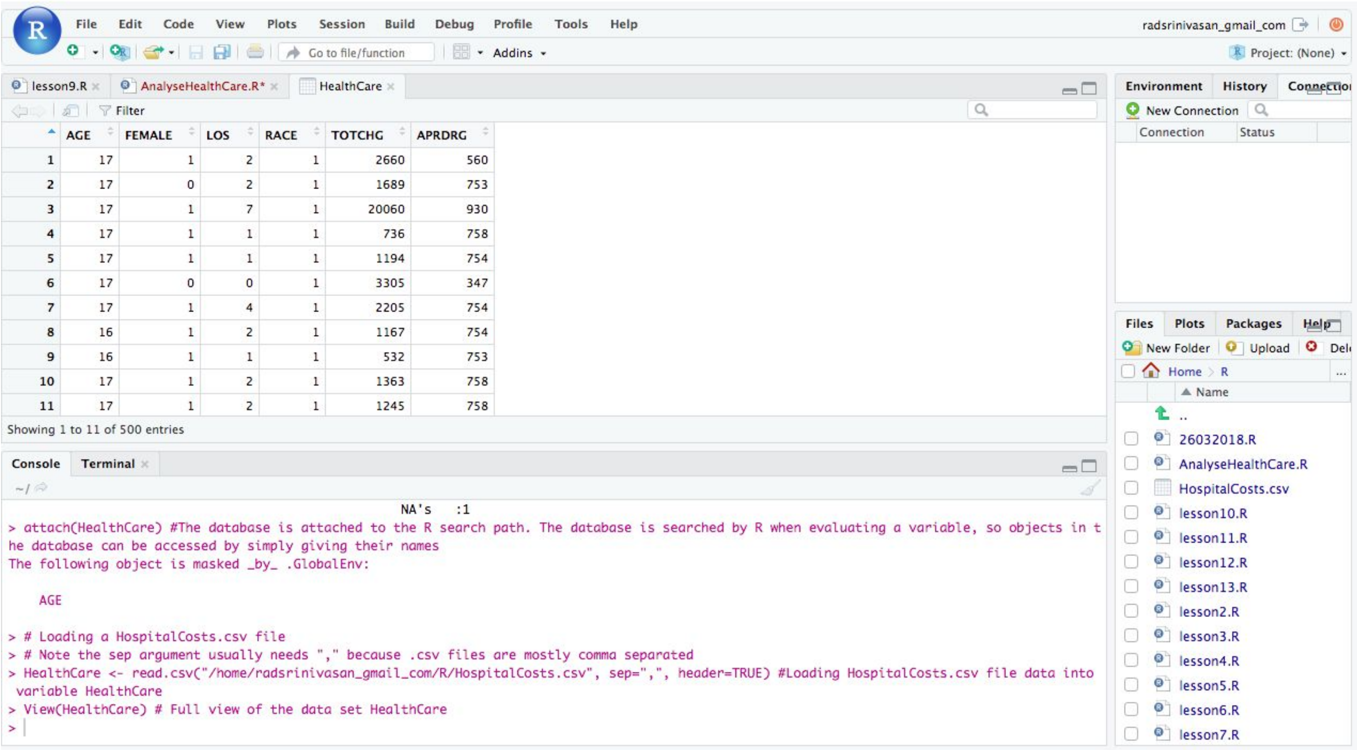
OUTPUT

```
> # Loading a HospitalCosts.csv file
> # Note the sep argument usually needs "," because .csv files are mostly comma separated
> HealthCare <- read.csv("/home/radsrinivasan_gmail_com/R/HospitalCosts.csv", sep=",", header=TRUE)
#Loading HospitalCosts.csv file data into variable HealthCare
```

CODE

```
View(HealthCare) # Full view of the data set HealthCare
```

OUTPUT



CODE

head(HealthCare) # View first 6 records of HealthCare

OUTPUT

```
> head(HealthCare) # View first 6 records of HealthCare
  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
```

CODE

str(HealthCare) # Describes structure of HealthCare and its variables

OUTPUT

```
> str(HealthCare) # Describes structure of HealthCare and its variables
'data.frame':	500 obs. of  6 variables:
 $ AGE   : int  17 17 17 17 17 17 17 16 16 17 ...
 $ FEMALE: int  1 0 1 1 1 0 1 1 1 1 ...
 $ LOS   : int  2 2 7 1 1 0 4 2 1 2 ...
 $ RACE  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ TOTCHG: int  2660 1689 20060 736 1194 3305 2205 1167 532 1363 ...
 $ APRDRG: int  560 753 930 758 754 347 754 754 753 758 ...
```

CODE

```
attach(HealthCare) #The database is attached to the R search path. The database is searched by R when evaluating a variable, so objects in the database can be accessed by simply giving their names
```

OUTPUT

```
> attach(HealthCare) #The database is attached to the R search path. The database is searched by R when evaluating a variable, so objects in the database can be accessed by simply giving their names
The following object is masked _by_ .GlobalEnv:

    AGE

The following objects are masked from HealthCare (pos = 3):

    AGE, APRDRG, FEMALE, LOS, RACE, TOTCHG
```

CODE

```
par(mfrow=c(2,3)) #Creates a matrix of 2 rows by 3 columns plots
```

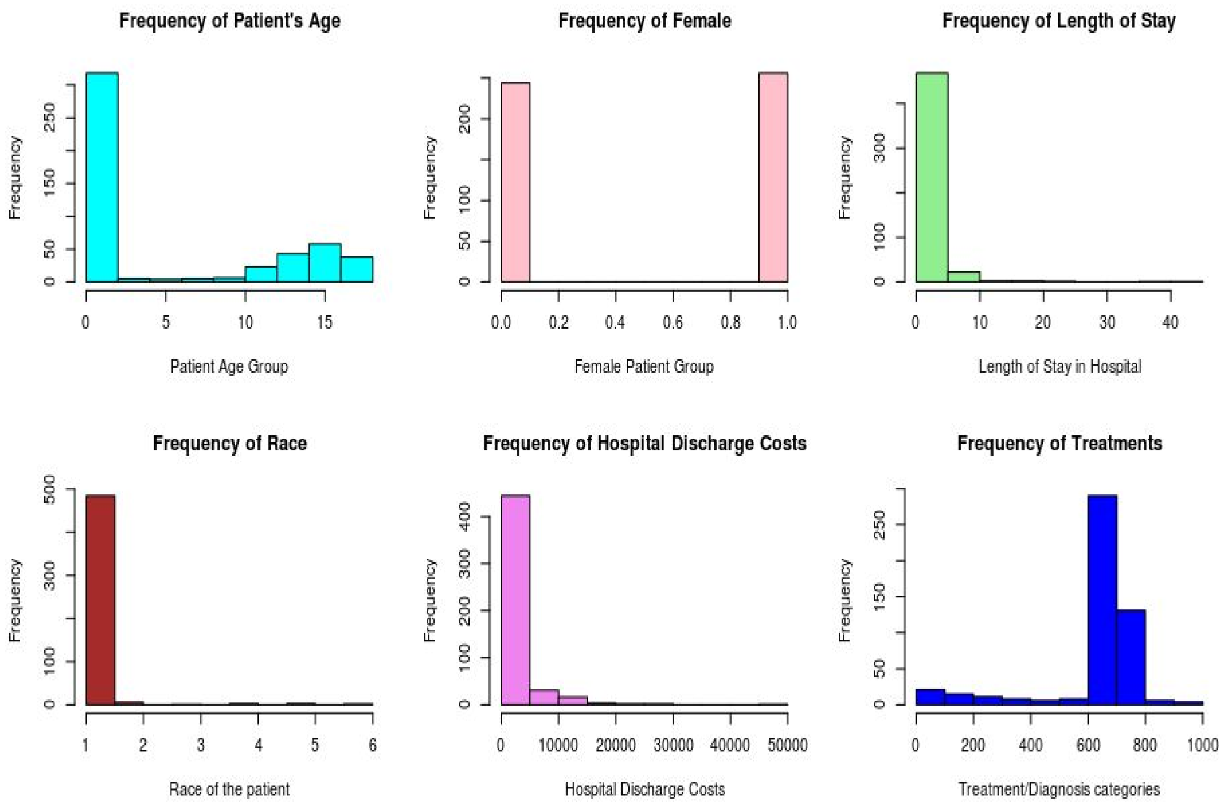
OUTPUT

```
> par(mfrow=c(2,3)) #Creates a matrix of 2 rows by 3 columns plots
```

CODE

```
#Identifies the Patient's age group in the histogram chart
hist(HealthCare$AGE, col="cyan", main = "Frequency of Patient's Age", xlab = "Patient Age Group")
#Identifies Female patient group in the histogram chart
hist(HealthCare$FEMALE, col="pink", main = "Frequency of Female", xlab = "Female Patient Group")
#Identifies the length of stay in the Hospital (in days) in the histogram chart
hist(HealthCare$LOS,col="lightgreen", main = "Frequency of Length of Stay", xlab = "Length of Stay in Hospital")
#Identifies the Race of Patient's in the histogram chart
hist(HealthCare$RACE, col="brown", main = "Frequency of Race", xlab = "Race of the patient")
#Identifies the Hospital Discharge Cost in the histogram chart
hist(HealthCare$TOTCHG, col="violet", main = "Frequency of Hospital Discharge Costs", xlab = "Hospital Discharge Costs")
#Identifies the Diagnosis/Treatments categories in the histogram chart
hist(HealthCare$APRDRG, col="blue", main = "Frequency of Treatments", xlab = "Treatment/Diagnosis categories")
```

OUTPUT



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

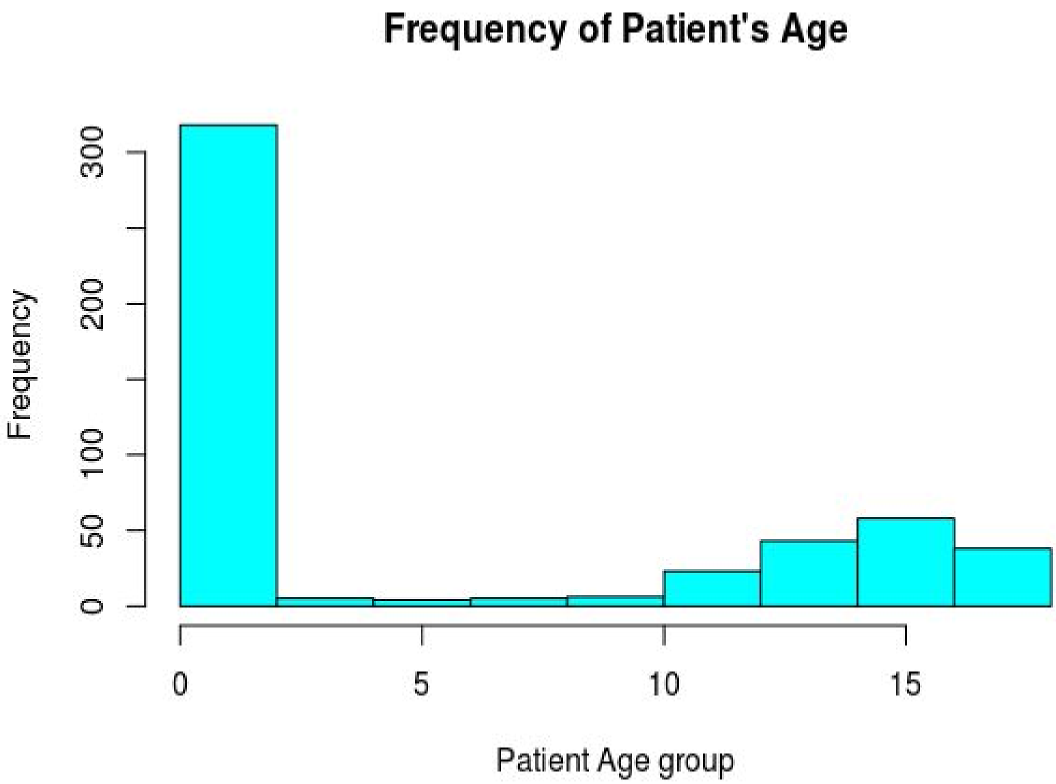
   Histogram chart is used to find age category with max frequency of hospital visits. To get an overview of all age categories we use histogram for frequency analysis

CODE

```
#Identifies the age group in the histogram chart
hist(HealthCare$AGE, breaks=7,col="cyan", main = "Frequency of Patient's Age", xlab = "Patient Age group")
```

OUTPUT

```
> #Identifies the age group in the histogram chart
> hist(HealthCare$AGE, breaks=7,col="cyan", main = "Frequency of Patient's Age", xlab = "Patient Age group")
```



Factor function is used to convert AGE column to numeric which will be used in summary function

CODE

AGE <- as.factor(HealthCare$AGE) #Converts it into factor and stores in a variable AGE

OUTPUT

> AGE <- as.factor(HealthCare$AGE) #Converts it into factor and stores in a variable AGE

CODE

summary(AGE) # Provides descriptive statistics about the AGE data set

OUTPUT

> summary(AGE) # Provides descriptive statistics about the AGE data set

```
  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
```

Conclusion: From the above Age summary results we can infer exact numeric output of infant category. It has the maximum hospital visits that is 307. Age ‘0’ patient’s has maximum visits followed by age group 17 and 15-16 ages.

CODE

```
which.max(summary(AGE)) #Generates the max index of the AGE category dataframe
```

OUTPUT

```
which.max(summary(AGE)) #Generates the max index of the AGE category dataframe
0
1
```

CODE

```
#Splits the data into subsets, computes summary statistics for each,
#and returns the result is reformatted into a dataframe containing variable ageGroup
?aggregate
#aggregate() is used to add the expenditure from each age
ageGroup <- aggregate(TOTCHG~AGE, FUN = sum, data=HealthCare)
```

OUTPUT

```
> #Splits the data into subsets, computes summary statistics for each,
> #and returns the result is reformatted into a dataframe containing variable ageGroup
> ?aggregate
```

Aggregate function is used to find sum up expenditure from each age category and then maximum function is used to find highest costs.

CODE

```
#aggregate() is used to add the expenditure from each age
ageGroup <- aggregate(TOTCHG~AGE, FUN = sum, data=HealthCare)
```

OUTPUT

```
> #aggregate() is used to add the expenditure from each age
> ageGroup <- aggregate(TOTCHG~AGE, FUN = sum, data=HealthCare)
```

CODE

```
ageGroup #review dataset
```

OUTPUT

```
  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
```

CODE

```
#max() is used to find highest costs
max_expenditure <- max(aggregate(TOTCHG~AGE, FUN = sum, data=HealthCare))
max_expenditure #review dataset
```

OUTPUT

```
> #max() is used to find highest costs
> max_expenditure <- max(aggregate(TOTCHG~AGE, FUN = sum, data=HealthCare))
> max_expenditure #review dataset
[1] 678118
```

Conclusion: We can infer from the output that the infants have maximum hospital expenditure followed by the Age groups of 17 & 15-16. Number of Hospital visits are proportional to hospital expenditure.

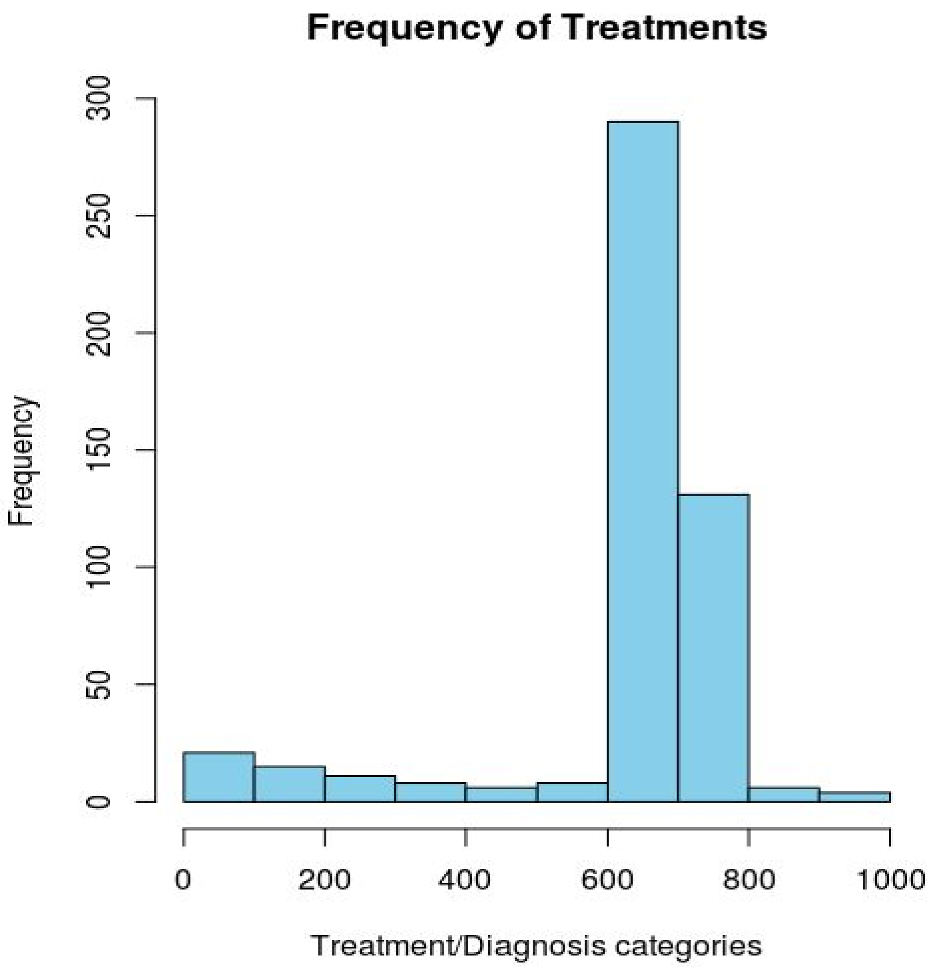
2. In order of severity of the diagnosis and treatments and to find out the expensive treatments, the agency wants to find the diagnosis related group that has maximum hospitalization and expenditure.

   To visualise diagnosis and treatments based on the categories of frequenty using histograms

CODE

```
#Identifies the Diagnosis and Treatments in the histogram chart
hist(HealthCare$APRDRG, col="skyblue", main = "Frequency of the treatments", xlab = "Patient Diagnosis Related Groups")
```

OUTPUT



Factor function is used to convert APRDRG column to numeric which will be used in summary function along with which.max() to generate maximum index of the category dataframe.

CODE

APRDRG <- as.factor(HealthCare$APRDRG) #Converts it into factor and stores in a variable APRDRG
summary(APRDRG) # Provides descriptive statistics about the APRDRG data set

OUTPUT

```
> APRDRG <- as.factor(HealthCare$APRDRG) #Converts it into factor and stores in a variable APRDRG

> summary(APRDRG) # Provides descriptive statistics about the APRDRG data set
 21  23  49  50  51  53  54  57  58  92  97 114 115 137 138 139 141 143 204 206 225 249 254 308 313 317
344 347 420 421 422
  1   1   1   1   1  10   1   2   1   1   1   1   2   1   4   5   1   1   1   1   2   6   1   1   1   1   2   3   2   1   3
560 561 566 580 581 602 614 626 633 634 636 639 640 710 720 723 740 750 751 753 754 755 756 758 760
776 811 812 863 911 930
  2   1   1   1   3   1   3   6   4   2   3   4 267   1   1   2   1   1  14  36  37  13   2  20   2   1   2   3   1   1   2
952
  1
```

CODE

which.max(summary(APRDRG)) #Generates the max index of the category dataframe

OUTPUT

```
> which.max(summary(APRDRG)) #Generates the max index of the category dataframe
640
 44
```

Aggregate function is used to find sum up expenditure for treatment/diagnosis categories and then maximum function is used to find highest costs.

CODE

```
treatmentGrp <- aggregate(TOTCHG~APRDRG, FUN = sum, data=HealthCare) #aggregate() is used to add the expenditure for treatment/diagnosis
treatmentGrp #review dataset
```

OUTPUT

```
> treatmentGrp <- aggregate(TOTCHG~APRDRG, FUN = sum, data=HealthCare) #aggregate() is used to add the expenditure for treatment/diagnosis
> treatmentGrp #review dataset
   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

Maximum function is used to find highest costs.

CODE

```
#max() is used to find highest costs
max_hospitalization_Expenditure <- max(aggregate(TOTCHG~APRDRG, FUN = sum, data=HealthCare))
max_hospitalization_Expenditure #review dataset
```

OUTPUT

```
> #max() is used to find highest costs
> max_hospitalization_Expenditure <- max(aggregate(TOTCHG~APRDRG, FUN = sum, data=HealthCare))
> max_hospitalization_Expenditure #review dataset
[1] 437978
```

CODE

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

OUTPUT

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

Conclusion: We can infer from the output that 640 has the maximum hospitalization by a huge number (267 out of 500) and has the highest hospitalisation cost.

CODE

```
treatmentGrp#review dataset
```

OUTPUT

```
> treatmentGrp#review dataset
   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
```

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.

1. We need to remove the 'NAs' value from the data
2. Factorise RACE variable to generate RACE summary
3. If RACE has impact on Hospital costs
4. ANOVA function with TOTCHG as dependent variable and RACE as independent variable

CODE

```
names(HealthCare) #for reading
```

OUTPUT

```
> names(HealthCare) #for reading
[1] "AGE"    "FEMALE" "LOS"    "RACE"   "TOTCHG" "APRDRG"
```

CODE

```
#Create confusion matrix and accuracy/error rates for this model
#NOTE - we need to remove rows that had NAs in any variable
HealthCare <- na.omit(HealthCare)
```

OUTPUT

```
> #Create confusion matrix and accuracy/error rates for this model
> #NOTE - we needed to remove rows that had NAs in any variable
> HealthCare <- na.omit(HealthCare)
```

CODE

```
#Apply One Way Anova function
# aov(dependent ~ independent, data)
av <- aov(TOTCHG ~ RACE, data=HealthCare)
```

OUTPUT

```
> #Apply One Way Anova function
> # aov(dependent ~ independent, data)
> av <- aov(TOTCHG ~ RACE, data=HealthCare)
```

CODE

```
av
```

OUTPUT

```
> av
Call:
   aov(formula = TOTCHG ~ RACE, data = HealthCare)

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

Residual standard error: 3906.493
Estimated effects may be unbalanced
```

CODE

```
#Print out the ANOVA table with the summary function.
summary(av)
#Displays Maximum Hospital Expenditure per race
summary(HealthCare$RACE)
```

OUTPUT

```
> #Print out the ANOVA table with the summary function.
> summary(av)
             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
> #Displays Maximum Hospital Expenditure per race
> summary(HealthCare$RACE)
  1   2   3   4   5   6 
484   6   1   3   3   2 
```

Conclusion: F value is very low. Accepting Null hypothesis based on the variation between hospital expenditure among different RACES which is lesser than the variation of hospital expenditure within each RACE. P value is high i.e., it depicts that there is no relationship between RACE and hospital expenses.
We have more data for RACE 1 in comparison to other races i.e., 484/500 as per the output which makes the observations skewed.
Data is not sufficient to verify whether RACE of a patient affects hospital expenditure.

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.

To analyse the severity of costs we will use linear regression with TOTCHG (expense) an independent variable with AGE and FEMALE (gender) as dependent variables.

CODE

```
#Using Linear Regression to analyse the severity of Hospital costs by age & gender
#Converts it into factor and stores Gender in a variable HealthCare$FEMALE
HealthCare$FEMALE <- as.factor(HealthCare$FEMALE)
```

OUTPUT

```
> #Using Linear Regression to analyse the severity of Hospital costs by age & gender
> #Converts it into factor and stores Gender in a variable HealthCare$FEMALE
> HealthCare$FEMALE <- as.factor(HealthCare$FEMALE)
```

CODE

```
#Linear regression model for predicting expenditure with AGE and gender
#TOTCHG (expenditure) is independent variable
#AGE and FEMALE are dependent variables
Severity_linRegression <- lm(TOTCHG ~ AGE + FEMALE, data = HealthCare)
```

OUTPUT

```
> #Linear regression model for predicting expenditure with AGE and gender
> #TOTCHG (expenditure) is independent variable
> #AGE and FEMALE are dependent variables
> Severity_linRegression <- lm(TOTCHG ~ AGE + FEMALE, data = HealthCare)
```

CODE

```
#Summary function
summary(Severity_linRegression)
```

OUTPUT

```
> #Summary function
> summary(Severity_linRegression)
```

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

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
```

CODE

```
#Summary function to compare Patient gender
summary(HealthCare$FEMALE)
```

OUTPUT

```
> #Summary function to compare Patient gender
> summary(HealthCare$FEMALE)
  0   1
244 255
```

Conclusion: Age has a high impact than female gender according to p-values and significant levels. There are equal number of male and female gender on an average based on the negative coefficient values from summary. Female gender incur lesser hospital expenses than that of the males.

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.

Linear regression is used to find whether length of stay is dependent on age, female gender and race. LOS is a dependent variable and Age, gender and race are independent variables.

CODE

```
> #Converts it into factor and stores RACE in a variable HealthCare$RACE
> HealthCare$RACE <- as.factor(HealthCare$RACE)
```

OUTPUT

```
> #Converts it into factor and stores RACE in a variable HealthCare$RACE
> HealthCare$RACE <- as.factor(HealthCare$RACE)
```

CODE

```
#Linear model for predicting length of stay with AGE, gender and Race
#LOS is dependent variable
#AGE, FEMALE and RACE are independent variables.
# Finding if Age, gender and race are affecting length of stay
patLos <- lm(LOS ~ AGE + FEMALE + RACE, data = HealthCare)
```

OUTPUT

```
> #Linear model for predicting length of stay with AGE, gender and Race
> #LOS is dependent variable
> #AGE, FEMALE and RACE are independent variables.
> # Finding if Age, gender and race are affecting length of stay
> patLos <- lm(LOS ~ AGE + FEMALE + RACE, data = HealthCare)
```

CODE

```
#Summary function Patient's length of stay
summary(patLos)
```

OUTPUT

```
> #Summary function Patient's length of stay
> summary(patLos)

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

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:  -0.005433
F-statistic: 0.6156 on 7 and 491 DF,  p-value: 0.7432
```

Conclusion: p-values for all independent variables are very high so there is no linear relationship between the given variables.
Therefore, we can conclude that we cannot predict length of stay of a patient based on age, gender and race.

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

Using Linear regression, we can find variables that affect the hospital expenses the most. TOTCHG is a dependent variable and age, female, race, APRDRG are independent variables.

CODE

```
#TOTCHG (expenditure) is dependent variable
#AGE, FEMALE, RACE, LOS and APRDRG are independent variables
# Finding if Age, gender, race, length of stay and treatment category are affecting expenditure
AffectsHospCost <- lm(TOTCHG ~ AGE + FEMALE + RACE + LOS + APRDRG, data = HealthCare)
```

OUTPUT

```
> #TOTCHG (expenditure) is dependent variable
> #AGE, FEMALE, RACE, LOS and APRDRG are independent variables
> # Finding if Age, gender, race, length of stay and treatment category are affecting expenditure
> AffectsHospCost <- lm(TOTCHG ~ AGE + FEMALE + RACE + LOS + APRDRG, data = HealthCare)
```

CODE

```
#Summary function
summary(AffectsHospCost)
```

OUTPUT

```
> #Summary function
> summary(AffectsHospCost)

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

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
```

Final conclusion: Age and length of stay affect the total hospital expenditure. Therefore, there is a positive relationship between length of stay to expense/cost, so with one additional day there is an additional cost of 742 to the hospital cost.