Analyze the Healthcare cost
AND
Utilization in Wisconsin hospitals

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.

Sol. We can analyse the age category who frequent the hospital by finding the maximum frequency is of which age. Also we can visualize the same by drawing a histogram having age on x-axis and frequency on y-axis.

Code:

```
hops = read.csv("HospitalCosts.csv")

which.max(summary(as.factor(hospitalData$AGE)))

hist(hospitalData$AGE,main="Age/Frequency Graph",xlab="Age",ylab="Frequency",ylim=c(0,350),xlim=c(0,25),col = "green",border = "blue")

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

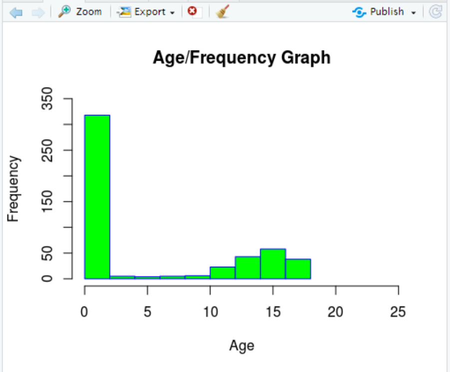
Result:

From the output we can see that zero age babies are hospitalized maximum i.e. 307. The same we can visualize on the histogram there is much difference between infants and other age group's hospitalization.

In case expenditure also infants have the maximum cost with a lot of difference compare to the next ones i.e. 15 & 17 age group.

Output:

```
> hospitalData = read.csv("HospitalCosts.csv")
> which.max(summary(as.factor(hospitalData$AGE)))
0
1
> hist(hospitalData$AGE,main="Age/Frequency Graph",xlab="Age",ylab="Frequency",ylim=c(0,350),xlim=c(0,25),
col = "green",border = "blue")
> aggregate(TOTCHG ~ AGE, FUN = sum, data = hospitalData)
   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
> 
```



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.

Sol. We can analyze the groups on histogram for seeing which group has max. hospitalization

and then by using aggregate function we can see the expenditure group.

Code:

```
summary(as.factor(hospitalData$APRDRG))

aggregate(TOTCHG ~ APRDRG, FUN = sum, data = hospitalData)
```

Result:

From the summary we can see that group 640 has maximum hospitalization with 267 entries and also has the highest total hospitalization cost i.e. 437978.

Output:

```
63     952   4833
> summary(as.factor(hospitalData$APRDRG))
 21  23  49  50  51  53  54  57  58  92  97 114 115 137 138 139 141 143 204 206 225 249 254 308 313 317 344 347 420 421 422 560 561 
  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   2   1 
566 580 581 602 614 626 633 634 636 639 640 710 720 723 740 750 751 753 754 755 756 758 760 776 811 812 863 911 930 952 
  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   1 
> aggregate(TOTCHG ~ APRDRG, FUN = sum, data = hospitalData)
   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
```

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.

Sol. As race of the patient is a categorical independent variable and cost is dependent variable we can perform anova test for seeing the relationship between them.

We can have our hypothesis as follow:

H0: Race of patient is not dependent on cost.

H1: Race of patient is dependent on cost.

Code:

summary(as.factor(hospitalData$RACE))

hospitalData= na.omit(hospitalData)

res=aov(TOTCHG ~ RACE, data = hospitalData)

summary(res)

Result:

From the anova test ran above we have our p-value(0.686) much higher than alpha(0.05), results in accepting null hypothesis(H0). It concludes that both are unrelated to each other. But from the summary we can analyze that race 1 has 484 entries out of 499 which is difficult for any test to understand the relationship between all race group present.

Output:

```
Console   Terminal ×
~/training/
>
> summary(as.factor(hospitalData$RACE))
  1   2   3   4   5   6
484   6   1   3   3   2
> hospitalData= na.omit(hospitalData)
> res=aov(TOTCHG ~ RACE, data = hospitalData)
> summary(res)
             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
>
```

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.

Sol: To see severity of the hospital costs by age and gender we can ran two-way anova test as follows:

Dependent variable: TOTCHG

Independent variable: AGE, FEMALE

Code:

res1 = aov(TOTCHG ~ AGE + FEMALE, data = hospitalData)

summary(res1)

Result:

From the test ran above we can see that both age and gender has impact on total charge in which gender has the highest impact as the p value indicates.

Output:

```
~/training/
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> res1 = aov(TOTCHG ~ AGE + FEMALE, data = hospitalData)
> summary(res1)
             Df    Sum Sq   Mean Sq F value  Pr(>F)
AGE           1 1.297e+08 129749266   8.759 0.00323 **
FEMALE        1 6.522e+07  65219972   4.403 0.03638 *
Residuals   496 7.347e+09  14812787
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

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.

Sol: We can predict the length of stay from age, gender and race by linear regression model.

Dependent variable: LOS

Independent variables: AGE, FEMALE, RACE

Code:

hospitalData$RACE=as.factor(hospitalData$RACE)

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

trainingData=lm(LOS~RACE+FEMALE, hospitalData)

summary(trainingData)

Result:

From the summary of linear model we can see that RACE and FEMALE are not significant as they have very high p-value. Hence we cant predict length of stay on the basis of RACE and GENDER

Output:

```
Console    Terminal ×
~/training/

> hospitalData$RACE=as.factor(hospitalData$RACE)
> hospitalData$FEMALE=as.factor(hospitalData$FEMALE)
> trainingData=lm(LOS~RACE+FEMALE, hospitalData)
> summary(trainingData)

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

Residuals:
   Min     1Q Median     3Q    Max 
-2.950 -0.950 -0.726  0.274 38.050 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.7258     0.2195  12.417   <2e-16 ***
RACE2        -0.6337     1.3906  -0.456    0.649    
RACE3         1.0504     3.3895   0.310    0.757    
RACE4         0.4584     1.9596   0.234    0.815    
RACE5        -0.7258     1.9653  -0.369    0.712    
RACE6        -0.8377     2.3969  -0.349    0.727    
FEMALE1       0.2238     0.3045   0.735    0.463    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.383 on 492 degrees of freedom
Multiple R-squared:  0.00256,	Adjusted R-squared:  -0.009604 
F-statistic: 0.2105 on 6 and 492 DF,  p-value: 0.9735

> |
```

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

Sol: To see complete analysis we will run a linear model which shows dependency of each factor on hospital costs.

Dependent variable: TOTCHG

Independent variables: All other variables

Code:

costData=lm(TOTCHG~., hospitalData)

summary(costData)

Result:

From the linear regression model ran above we can see that age, length of stay and diagnosis related group are affecting total charge. Rest all factors are insignificant with respect to charge of hospitalization.

Output:

Console Terminal ×

~/training/

```
F-statistic: 0.2105 on 6 and 492 DF,  p-value: 0.9735

> costData=lm(TOTCHG~., hospitalData)
> summary(costData)

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

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    
LOS          742.9637    35.0464  21.199  < 2e-16 ***
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    
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
```