Business Scenario:

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 analyze the data to research on the healthcare costs and their utilization.

Here is a detailed description of the given dataset:

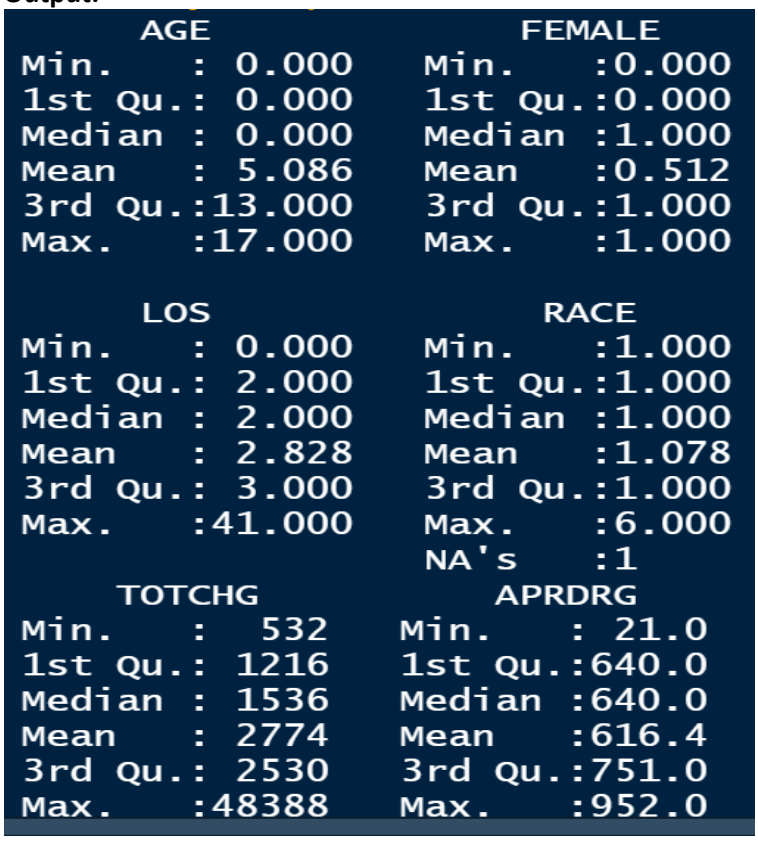
- AGE : Age of the patient discharged
- FEMALE : Binary variable that indicates if the patient is female
- LOS : Length of stay, in days
- RACE : Race of the patient (specified numerically)
- TOTCHG : Hospital discharge costs APRDRG : All Patient Refined Diagnosis Related Groups

```
# Importing Dataset
library(readxl)
hospdata = read_excel("hospitalcosts.xlsx",sheet="HospitalCosts")
View(hospdata)
```

	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
7	17	1	4	1	2205	754
8	16	1	2	1	1167	754
9	16	1	1	1	532	753
10	17	1	2	1	1363	758
11	17	1	2	1	1245	758
12	15	0	2	1	1656	753

Summary(hospdata)

Output:



```
      AGE              FEMALE
Min.   : 0.000   Min.   :0.000
1st Qu.: 0.000   1st Qu.:0.000
Median : 0.000   Median :1.000
Mean   : 5.086   Mean   :0.512
3rd Qu.:13.000   3rd Qu.:1.000
Max.   :17.000   Max.   :1.000

      LOS              RACE
Min.   : 0.000   Min.   :1.000
1st Qu.: 2.000   1st Qu.:1.000
Median : 2.000   Median :1.000
Mean   : 2.828   Mean   :1.078
3rd Qu.: 3.000   3rd Qu.: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
```

Goal 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.

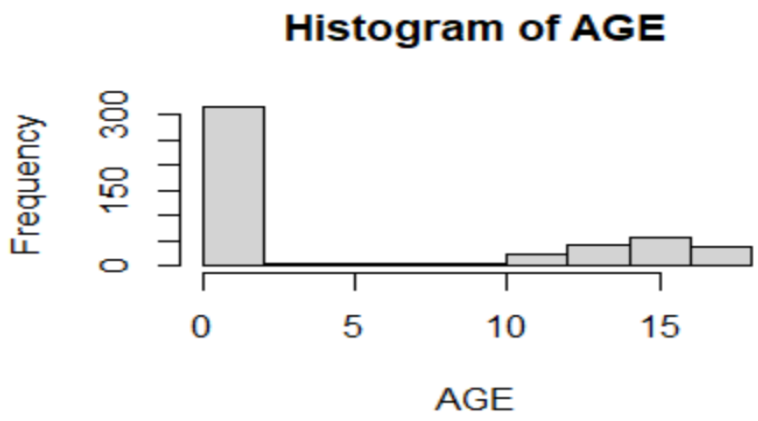
#To find the age category of people who frequent the hospital

Code:

attach(hospdata)

hist(AGE)

Output:



#To see the value for age group 0-1

table(AGE)

```
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
```

So the age category of people who frequent the hospital is: 0-1 years (307)

#To find the age category of people who has maximum expenditure

Code:

tapply(TOTCHG,AGE,sum)

Output:

```
     0      1      2      3      4      5      6      7
678118  37744   7298  30550  15992  18507  17928  10087
     8      9     10     11     12     13     14     15
  4741  21147  24469  14250  54912  31135  64643 111747
    16     17
 69149 174777
```

Analysis:

So the maximum expenditure is for the age group 0-1: 678118

Goal 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.

Code:

summary(as.factor(APRDRG))

output:

```
> summary(as.factor(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
  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
347 420 421 422 560 561 566 580 581 602 614 626 633 634 636 639 640 710 720 723 740 750 751 753 754 755 756
  3   2   1   3   2   1   1   1   3   1   3   6   4   2   3   4 267   1   1   2   1   1  14  36  37  13   2
758 760 776 811 812 863 911 930 952
 20   2   1   2   3   1   1   2   1
```

#to get the diagnostic realted cost

Code:

tapply(TOTCHG,as.factor(APRDRG),sum)

Output:

```
   21     23     49     50     51     53     54     57     58     92     97    114    115    137    138
10002  14174  20195   3908   3023  82271    851  14509   2117  12024   9530  10562  25832  15129  13622
  139    141    143    204    206    225    249    254    308    313    317    344    347    420    421
17766   2860   1393   8439   9230  25649  16642    615  10585   8159  17524  14802  12597   6357  26356
  422    560    561    566    580    581    602    614    626    633    634    636    639    640    710
 5177   4877   2296   2129   2825   7453  29188  27531  23289  17591   9952  23224  12612 437978   8223
  720    723    740    750    751    753    754    755    756    758    760    776    811    812    863
14243   5289  11125   1753  21666  79542  59150  11168   1494  34953   8273   1193   3838   9524  13040
  911    930    952
48388  26654   4833
```

To get maximum cost

which.max(tapply(TOTCHG,as.factor(APRDRG),sum))

640

max(tapply(TOTCHG,as.factor(APRDRG),sum))

[1] 437978

Analysis:

So here 640 diag. related group has max cost: 437978

Goal 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.

#h0: the race of patient is related to the hospitalization cost
#ha: No relation between race and cost

Code:

```
summary(as.factor(RACE))
  1   2   3   4   5   6 NA's
484   6   1   3   3   2   1

hspdt=na.omit(hospdata)
summary(as.factor(hspdt$RACE))

 1  2  3  4  5  6
484  6  1  3  3  2
```

#Applying ANNOVA

Code:

```
anv<- aov(TOTCHG~RACE,data=hspdt)

summary(anv)
```

```
Call:
   aov(formula = TOTCHG ~ RACE, data = hspdt)

Terms:
                    RACE  Residuals
Sum of Squares     2488459 7539623326
Deg. of Freedom          1        497

Residual standard error: 3894.903
Estimated effects may be unbalanced
```

Analysis:

- **here the p-value is .68(high), so we can reject the null hypothesis**
- **So we conclude there is no relation between race of the patient and the hospital cost**

Goal 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.

Code:

```
#Applying regression modeling

md1<- lm(formula = TOTCHG~AGE+FEMALE, data = hspdt)
summary(md1)
```

Output:

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

Residuals:
   Min     1Q Median     3Q    Max
 -3403  -1444   -873   -156  44950

Coefficients:
            Estimate Std. Error t value
(Intercept)  2719.45     261.42  10.403
AGE            86.04      25.53   3.371
FEMALE       -744.21     354.67  -2.098
            Pr(>|t|)
(Intercept)  < 2e-16 ***
AGE         0.000808 ***
FEMALE      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
```

Analysis:

- **Since the p-value for age is lesser than 0.05 and has 3*, it has most statistical significance**
- **Also gender has less p-value.**
- **So we can conclude that the model is statistically significance**

Goal 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.

Code:

```
md2<-lm(formula=LOS~AGE+FEMALE+RACE, data = hspdt)

summary(md2)
```

Output:

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

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
Multiple R-squared:  0.007898,  Adjusted R-squared:  0.001886
F-statistic: 1.314 on 3 and 495 DF,  p-value: 0.2692
```

Analysis:

- **Here we can see the p-value for age, gender and race is higher, so it is statistically insignificant**
- **Hence age, gender and race can’t be used to predict length of stay.**

Goal 6.

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

Code:

```
md3<- lm(formula=TOTCHG~ ., data = hspdt)

summary(md3)
```

Output:

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

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
Multiple R-squared:  0.5536,    Adjusted R-squared:  0.5491
F-statistic: 122.3 on 5 and 493 DF,  p-value: < 2.2e-16
```

Analysis:

- **Here we can see Age, LOS and APRDRG have 3 stars, so they are the once with statistical significance.**
- **we can see that age and length of stay affect the total hospital cost.**