HELTHCARE COST ANALYSIS

-SIDDHANT ARYA

1. DESCRIPTION

1.1. Background and Objective

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 healthcare costs and their utilization.

1.2. Domain

The domain of the project is healthcare.

1.3. Dataset Description

Given below is the detailed description of the dataset used.

Attribute	Description
Age	Age of the patient discharged
Female	A 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

1.4. Analysis To Be Performed

1. To record the patient statistics, the agency wants to find the age category of people who frequently visit the hospital and has the maximum 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.
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.
4. To properly utilize the costs, the agency has to analyze the severity of the hospital costs by age and gender for the proper allocation of resources.
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.
6. To perform a complete analysis, the agency wants to find the variable that mainly affects hospital costs.

2. ANALYSIS

Reading the given dataset:

```
hosp <- read.csv("HospitalCosts.csv", stringsAsFactors = TRUE)
```

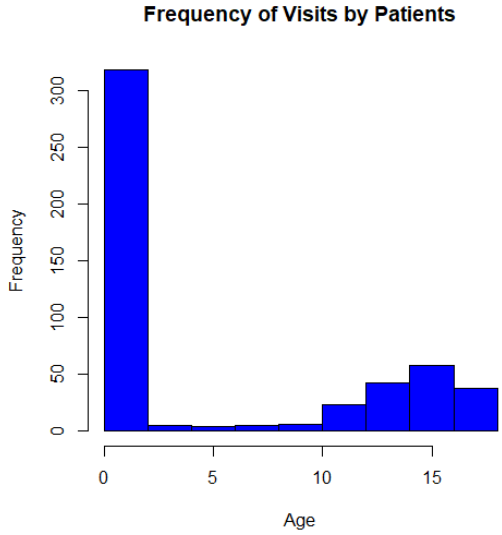
For the convenience of analysis, attribute **FEMALE** (being a dummy variable) and **AGE** (for analysis purposes) has been converted from **int** to **Factor** levels being "1" and "0":

```
hosp$FEMALE <- sapply(hosp$FEMALE, factor)
hosp$AGE <- sapply(hosp$AGE, factor)
```

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

In order to find the people who visit the hospital frequently a histogram can be plot for the age groups and the data that will be used for the frequency will be the attribute **AGE**:

```
hist(hosp$AGE, main = "Frequency of Visits by Patients", xlab = "Age", col = "blue")
```



From the histogram it is evident that a large number of visits (>300) to the hospital are made by the children, mostly the infants. However, for a better clarity :

```
library(dplyr)
hosp%>%desc(count(AGE))
```

So as can be seen from the table as well that the greatest number of occurrences is for age 0 (infants). Therefore,

```
> hosp%>%count(AGE)
   AGE   n
1   17  38
2   16  29
3   15  29
4   14  25
5   13  18
6   12  15
7   11   8
8   10   4
9    7   3
10   6   2
11   3   3
12   2   1
13   1  10
14   0 307
15   5   2
16   4   2
17   8   2
18   9   2
```

OBSERVATION 1: From the above two results we can conclude that the most frequent visits is from the infant age group followed by the people of age 17.

Now in order to find the maximum expenditure the total expenditure of all people in the age group can be used and the maximum of that will give the desired result. In order to add the expenditure for each age, the aggregate function is used:

```
temp <- aggregate(hosp$TOTCHG ~ hosp$AGE, FUN = sum, na.action = na.omit)
colnames(temp) <- c("Age","Expenditure")
temp
temp[temp$Expenditure == max(temp$Expenditure),]
```

So, as we can see from the output of the above code that the maximum expenditure is also done in the infant age category.

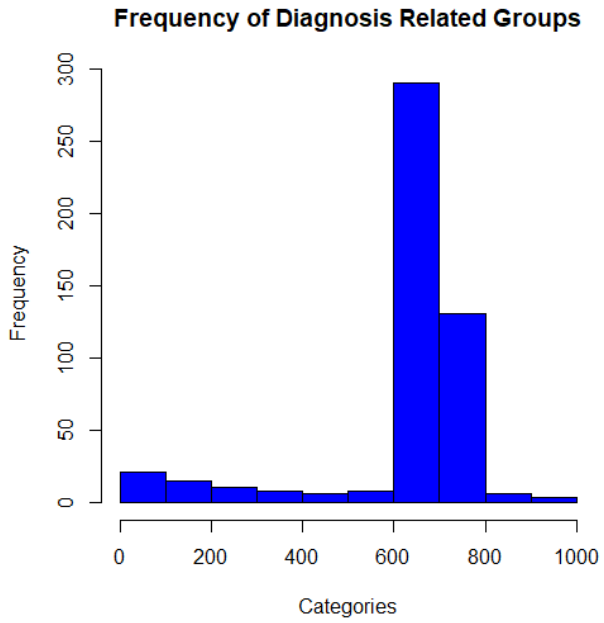
OBSERVATION 2: We can thus conclude that along with the maximum hospital visits the age group of infants also have the maximum expenditure.

```
> temp
   Age Expenditure
1   17      174777
2   16       69149
3   15      111747
4   14       64643
5   13       31135
6   12       54912
7   11       14250
8   10       24469
9    7       10087
10   6       17928
11   3       30550
12   2        7298
13   1       37744
14   0      678118
15   5       18507
16   4       15992
17   8        4741
18   9       21147
> temp[temp$Expenditure==max(temp$Expenditure),]
   Age Expenditure
14   0      678118
>
```

Analysis 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.

In order to get a clear view first the APRDRG column can be visualized using a histogram:

```
hist(hosp$APRDRG, main = "Frequency of Diagnosis Related Groups", xlab =
"Categories", col = "blue")
```



From the above histogram it can be seen that the maximum hospitalization is in a group between 600-700. To find out the category with maximum hospitalizations:

```
hosp$APRDRG <- as.factor(hosp$APRDRG)
summary(hosp$APRDRG)
which.max(summary(hosp$APRDRG))
```

```
> summary(hosp$APRDRG)
 21  23  49  50  51  53  54  57  58  92  97 114 115 137 138 139 141 143 204 206 225 249 254 308 313 317 344 347 420 421 422 560 561 566
  1   1   1   1   1  10   1   2   1   1   1   1   2   1   4   5   1   1   1   1   2   6   1   1   1   1   2   3   2   1   3   2   1   1
580 581 602 614 626 633 634 636 639 640 710 720 723 740 750 751 753 754 755 756 758 760 776 811 812 863 911 930 952
  1   3   1   3   6   4   2   3   4 267   1   1   2   1   1  14  36  37  13   2  20   2   1   2   3   1   1   2   1
> which.max(summary(hosp$APRDRG))
640
 44
>
```

Therefore, treatment category 640 has maximum number of hospitalizations.

Now coming to the category with maximum expenditure:

```
temp2 <- aggregate(hosp$TOTCHG ~ hosp$APRDRG, FUN = sum, na.action = na.omit)
colnames(temp2) <- c("Category","Expenditure")
temp2[temp2$Expenditure==max(temp2$Expenditure),]
```

```
> temp2[temp2$Expenditure==max(temp2$Expenditure),]
   Category Expenditure
44      640      437978
```

Therefore, the category with maximum expenditure is 640.

Observation: The category with maximum number of hospitalizations is "Category 640", out of a total of 500 hospitalizations (total entries in dataset) 267 were for "Category 640". Moreover, "Category 640" is also the category that had the maximum expenditure.

Analysis 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.

Considering the following hypothesis for testing:

H_0 : There is no relation between Race and Hospitalization cost.

H_1 : There is a relation between Race and Hospitalization cost.

In order to perform the analysis that whether race made an impact on hospital costs, ANOVA test can be used with **TOTCHG** as dependent variable and **RACE** as grouping variable.

```
hosp <- na.omit(hosp)   #Removing NA values for ANOVA to work (Because
received: 1 observation deleted due to missingness)
hosp$RACE <- as.factor(hosp$RACE)
anova <- aov(hosp$TOTCHG ~ hosp$RACE)
anova
summary(anova)
```

```
> anova
Call:
   aov(formula = hosp$TOTCHG ~ hosp$RACE)

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

Residual standard error: 3906.493
Estimated effects may be unbalanced
> summary(anova)
             Df    Sum Sq  Mean Sq F value Pr(>F)
hosp$RACE     5 1.859e+07  3718656   0.244  0.943
Residuals   493 7.524e+09 15260687
>
```

1. F value is quite low, therefore, variation among different races is much lesser than variation within each race.
2. P value is a lot more than the standard 5%, therefore, we can accept the Null hypothesis i.e., there is no relationship between race and hospital costs.
3. However, having a look at summary of Race we get an element of skewness:

```
summary(hosp$RACE)
```

```
> summary(hosp$RACE)
  1   2   3   4   5   6
484   6   1   3   3   2
>
```

What we see is that we have 484 entries for Race category 1 and for the rest 5 we have only 15 entries, which signals that there might be a high level of skewness in the result of ANOVA test.

OBSERVATION: Although, the P value being much greater than 0.05 results in the acceptance of the NULL hypothesis, however, as the data available for a single race is too large hence there is high level of skewness. If there would have been more entries for other races then it would have been easy to come to conclusion. Therefore, enough data is not present to verify the hypothesis.

Analysis 4: To properly utilize the costs, the agency has to analyze the severity of the hospital costs by age and gender for the proper allocation of resources.

To analyze the severity the best bet is to use linear regression here. We can take TOTCHG as the independent variable and, AGE and FEMALE can be taken as dependent variable. In the initial lines of code FEMALE was already set as a factor type attribute.

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

```
> summary(lr_model)

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

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 ***
FEMALE1      -744.21     354.67  -2.098 0.036382 *
AGE            86.04      25.53   3.371 0.000808 ***
---
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
```

Also, the count of number of male and female:

```
summary(hosp$FEMALE)
```

```
> summary(hosp$FEMALE)
  0   1
244 255
>
```

Therefore, the number of male and female in the dataset is almost equal.

OBSERVATION: The level of significance of AGE is higher as compared to the Gender, therefore, AGE has a greater impact on cost. Moreover, as the number of males and females are same, therefore, we can say that females spend lesser than male patients (Negative intercept of FEMALE1 -- 1 specifies a female and 0 specifies a male). However, the accuracy can’t be guaranteed as the value of Adjusted R-squared is too low.

Analysis 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.

This can also be analysed using linear regression (multi variate). Here we will take LOS as dependent variable and age, gender, and race as the independent variable.

```
lr_model2 <- lm(formula = LOS ~ AGE + FEMALE + RACE, data = hosp)
summary(lr_model2)
```

```
> summary(lr_model2)

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

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
```

OBSERVATION: As the P-values for the independent variables are quite high thus a low level of significance, hence, we can say that the length of stay can’t be predicted from age, gender, and race of the patient.

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

In this analysis too we can use the multi variable linear regression model to predict the variable that affects the hospital cost the most. Here the TOTCHG will be used as a dependent variable and all other variables will be considered as independent variables.

```
lr_model3 <- lm(formula = TOTCHG ~ ., data=hosp)
summary(lr_model3)
```

```
> summary(lr_model3)

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

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
```

OBSERVATION: Here we can see that the level of significance is very high for age, length of stay and the category of the illness. Moreover, the accuracy of the estimate also comes out to be 54.62% as the Adjusted R-squared value for the estimate is 0.5462, which seems quite good.