Data Science with R: Project

This document contains Project 9 – Healthcare Cost Analysis

Project by: Timothy Bumagat

Mentor: Rajib Layek

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.

Domain: Healthcare

Dataset Description:

Here is a detailed description of the given dataset:

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

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.

Expectation /Goals:

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

Answers:

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.

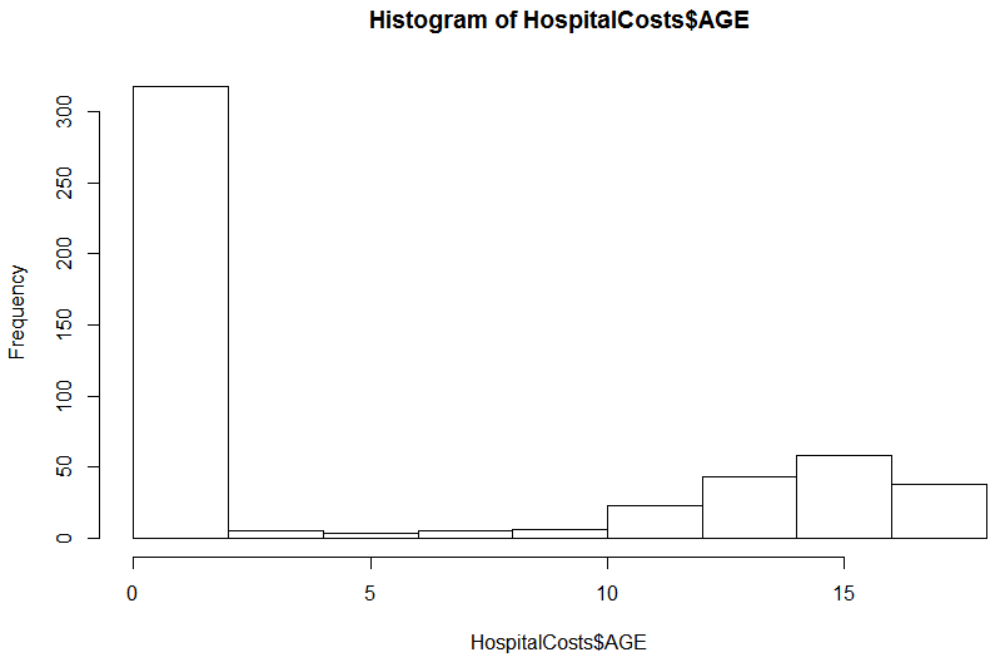
Code:

```
hist(HospitalCosts$AGE)

table(HospitalCosts$AGE)

HospitalCosts %>% group_by(AGE) %>%  summarize(count = n(), totalexp = sum(TOTCHG)) %>% arrange(desc(count,totalexp))
```

Output Screenshot:



```
> table(HospitalCosts$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
```

```
> HospitalCosts %>% group_by(AGE) %>%  summarize(count = n(), totalexp = sum(TOTCHG)) %>% arrange(desc(count,totalexp))
# A tibble: 18 x 3
     AGE count totalexp
   <int> <int>    <int>
 1     0   307   678118
 2    17    38   174777
 3    15    29   111747
 4    16    29    69149
 5    14    25    64643
 6    13    18    31135
 7    12    15    54912
 8     1    10    37744
 9    11     8    14250
10    10     4    24469
11     3     3    30550
12     7     3    10087
13     4     2    15992
14     5     2    18507
15     6     2    17928
16     8     2     4741
17     9     2    21147
18     2     1     7298
```

Analysis:

People aged 0 to 2 are the most prone to visit the hospital as 318 out of the 500 sample of people recorded were in this age bracket. Additionally, the age with the most expenditure is in the age 0 as expenditure in that age is 678118 out of the total from the sample of 1387194.

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:

library(dplyr) #640 has the highest count with the highest total expenditure of 437978

HospitalCosts %>% group_by(APRDRG) %>% summarize(count = n(), totalexp = sum(TOTCHG)) %>% arrange(desc(count,totalexp))

Output Screenshot:

```
> library(dplyr) #640 has the highest count with the highest total expenditure of 437978
> HospitalCosts %>% group_by(APRDRG) %>%  summarize(count = n(), totalexp = sum(TOTCHG)) %>% arrange(desc(count,totalexp))
# A tibble: 63 x 3
   APRDRG count totalexp
    <int> <int>    <int>
 1    640   267   437978
 2    754    37    59150
 3    753    36    79542
 4    758    20    34953
 5    751    14    21666
 6    755    13    11168
 7     53    10    82271
 8    249     6    16642
 9    626     6    23289
10    139     5    17766
# ... with 53 more rows
> |
```

Analysis:

640 has the highest frequency of patients of diagnosis related group (with 267 counts) and with an expenditure of 437978, with a combination of expenditure and patients, it ranks the highest compared to other groups.

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.

Code:

```
race <- as.factor(HospitalCosts$RACE)
summary(race)

HospitalCostsna <- na.omit(HospitalCosts)
modelannova <- aov(TOTCHG~race, data=HospitalCosts)
summary(modelannova)
```

Output Screenshot:

```
> race <- as.factor(HospitalCosts$RACE)
> summary(race)
   1    2    3    4    5    6 NA's
 484    6    1    3    3    2    1
> HospitalCostsna <- na.omit(HospitalCosts)
> modelannova <- aov(TOTCHG~race, data=HospitalCosts)
> summary(modelannova) #done
             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
1 observation deleted due to missingness
```

Analysis:

It is observed that the most common race among the patients is race 1 as it has the most frequency of patients in this classification with a frequency of 484 compared with other races. According to the model annova implemented, with a p-value of 0.943, there is no strong relationship and hence can be classified as insignificant. We can therefore conclude that the race of the patient has no relation with the hospitalization cost.

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:

```
lm_4 <- lm(TOTCHG~AGE+FEMALE, data=HospitalCosts)
```

summary(lm_4)

Output Screenshot:

```
> lm_4 <- lm(TOTCHG~AGE+FEMALE, data=HospitalCosts)
> summary(lm_4)

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

Residuals:
   Min     1Q Median     3Q    Max
 -3406  -1443   -869   -152  44951

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2718.63     261.14  10.411  < 2e-16 ***
AGE            86.28      25.48   3.387 0.000763 ***
FEMALE       -748.19     353.83  -2.115 0.034967 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3845 on 497 degrees of freedom
Multiple R-squared:  0.0261,    Adjusted R-squared:  0.02218
F-statistic:  6.66 on 2 and 497 DF,  p-value: 0.001399
```

Analysis:

According to the linear model implemented, with a p-value of 0.000763 with age and 0.034967 with gender, both age and gender have a strong relationship with the hospital costs, with more than 95% significance with gender and more than 99% significance with age.

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:

lm_5 <- lm(LOS~AGE+FEMALE+RACE, data=HospitalCosts)

summary(lm_5)

Output Screenshot:

```
> lm_5 <- lm(LOS~AGE+FEMALE+RACE, data=HospitalCosts)
> summary(lm_5)

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

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
  (1 observation deleted due to missingness)
Multiple R-squared:  0.007898,  Adjusted R-squared:  0.001886
F-statistic: 1.314 on 3 and 495 DF,  p-value: 0.2692
```

Analysis:

According to the linear model implemented, with a p-value of 0.0766 with age, 0.2334 with gender, and 0.7484 with race, it shows that there is a weak and insignificant relationship between the length of stay and the age, gender, and race. Therefore, age, gender, and race are not a good and reliable predictor of a patient's length of stay.

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

Code:

```
lm_6 <- lm(TOTCHG ~ .,data = HospitalCosts)
summary(lm2)
```

Output Screenshot:

```
> lm_6 <- lm(TOTCHG ~ .,data = HospitalCosts)
> summary(lm2)

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

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
  (1 observation deleted due to missingness)
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:

To find out if there are any variable that affects the hospital costs, the researcher has implemented a linear model in order to find out if there are any significant relationship with the hospital costs. According to the linear model, with a p-value of 7.02e-14 with age, 2e-16 with length of stay, and 2e-16 with All Patient Refined Diagnosis Related Groups. All these factors have more than 99% significance and hence, they show a strong relationship with the hospital costs and can therefore be assumed that these 3 factors affect hospital costs. On the other hand, gender and race with a p-value of 0.115 and 0.352, do not show a strong relationship with hospital costs and therefore these factors are assumed not to affect hospital costs.