HEALTHCARE COST ANALYSIS

MADE BY

SIDDHI KASLIWAL

ANALYZE THE HEALTHCARE COST AND UTILIZATION IN WISCONSIN HOSPITALS

DESCRIPTION

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

Analysis to be done:

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.

DECLARATION:

In order to make this project more informative and effective , I did some research and came up with certain findings, which helped me learn and understand better and at the same time trying my best to make this analysis more knowledgeable and interesting.

To begin with the project, let us understand the dataset first, then we've been asked to analyse the data of Wisconsin Hospitals supported some attributes and the way they're affecting the entire costs involved in a treatment Length of stay of the patients and we will investigate the practices followed in Wisconsin Hospital supported the race.

The tool we'll be using for the analysis is the **rStudio.**

QUESTION1:

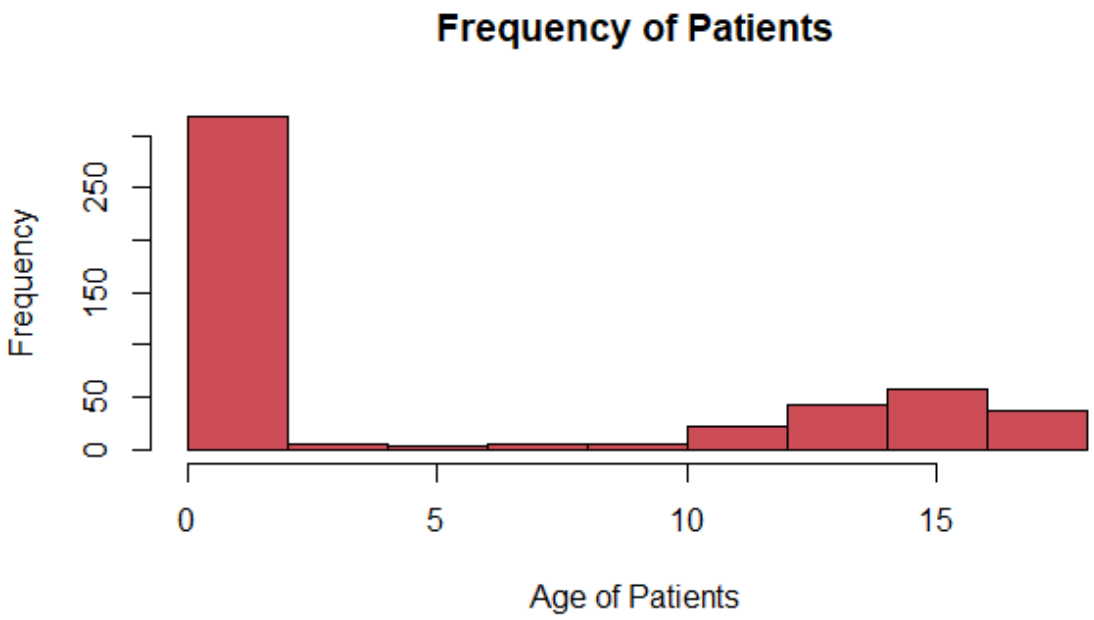
To record the patient statistics, the agency wants to find the age category of people who frequently visit the hospital and has the maximum expenditure.

Solution:

Here we have to find two things ,the category with the maximum frequency of hospital visits for this we use data visualization to get an overview of all the categories,in this case we will use a histogram for frequency analysis.

Code:

```
>hist(hist(Healthcare$AGE, main = "Frequency of Patients", col="#cc4d56", xlab
="Age of Patients"))
```



After that we will factor function the make the “AGE” column numerical which will be later used in summary function.

```
>attach(Healthcare)
>AGE <- as.factor(AGE)
>AGE_Dataframe <- data.frame(summary(AGE))
>head(AGE_Dataframe)
 >summary.AGE.
```

```
0    307
1     10
2      1
3      3
4      2
5      2
```

```
AGE_Table <- table(Healthcare$AGE)

AGE_Table

  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 results we conclude that infant category has the maximum hospital visits (above 300). The summary of Age gives us the exact numerical output showing that Age 0 patients have the maximum visits followed by Ages 15-17.

Now the other thing is that ,we have to find which age group has highest total cost, so for this we will use aggregate function to add the expenditure from each age and then the max function to find highest cost.

Code:

```
AGE_Aggregated <- aggregate(x= Healthcare$TOTCHG, by=
list(Healthcare$AGE),FUN=sum)

AGE_Aggregated

   Group.1      x
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

> max(AGE_Aggregated)

[1] 678118

Conclusion:

Thus, we can conclude that the infants also have the maximum hospital costs followed by Age groups 15 to 17, additionally we can say confidently that number of hospital visits are proportional to hospital costs.

QUESTION 2:

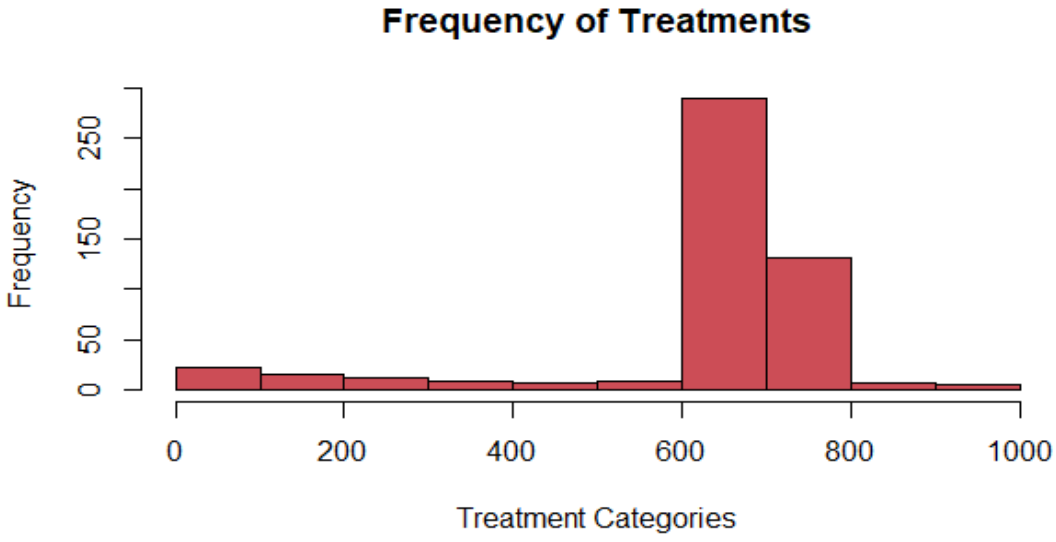
In order of severity of the diagnosis and treatments and to find out the expensive treatments, the agency wants to find the diagnosis-related group that has maximum hospitalization and expenditure.

Solution:

Here first we visualize the categories based on their frequency using histograms.

Code:

```
hist(Healthcare$APRDRG,col = "#cc4d56",main = "Frequency of Treatments",xlab =
"Treatment Categories")
```



Now we'll confirm that category column("APRDRG") is numerical then generates a summary along with the which.max to generate the max index of the category data frame, this will be followed by aggregate function used in an identical way as above.

.

Code:

```
APRDRG <- as.factor(Healthcare$APRDRG)

APRDRG_Dataframe <- data.frame(summary(APRDRG))

(APRDRG_Dataframe)
```

summary.APRDRG.

21	1
23	1
49	1
50	1
51	1
53	10
54	1
57	2
58	1
92	1
97	1
114	1
115	2
137	1
138	4
139	5
141	1
143	1
204	1
206	1
225	2
249	6
254	1
308	1

313	1
317	1
344	2
347	3
420	2
421	1
422	3
560	2
561	1
566	1
580	1
581	3
602	1
614	3
626	6
633	4
634	2
636	3
639	4
640	267
710	1
720	1
723	2
740	1
750	1
751	14
753	36
754	37
755	13
756	2
758	20
760	2
776	1

```
811             2
812             3
863             1
911             1
930             2
952             1
> which.max(summary(APRDRG))
640
 44
APRDG_Aggregated <- aggregate(TOTCHG ~ APRDRG, FUN = sum, data =
Healthcare)
APRDG_Aggregated
   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
> APRDG_Aggregated[which.max(APRDG_Aggregated$TOTCHG),]
   APRDRG TOTCHG
44    640 437978
```

Conclusion:

Hence can conclude that category 640 has the maximum hospitalizations by a huge number (267 out of 500), along with this it also has the highest hospitalization cost.

QUESTION 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.

Solution:

Here we will first remove the “NA” values from our database, then factorize the Race variable to generate a summary, additionally to verify whether race made an impact on the hospital costs we will use ANOVA function with TOTCHG as dependent variable and RACE as grouping variable.

Code:

```
Healthcare_New <- na.omit(Healthcare)
colSums(is.na(Healthcare_New))
AGE FEMALE    LOS   RACE TOTCHG APRDRG
    0      0      0      0      0      0
> Healthcare_New$RACE <- as.factor(Healthcare_New$RACE)
> AOV_Model <- aov(TOTCHG ~ RACE, data = Healthcare_New)
> AOV_Model
Call:
```

```
  aov(formula = TOTCHG ~ RACE, data = Healthcare_New)

Terms:
                    RACE Residuals
Sum of Squares  1.9e+07   7.5e+09
Deg. of Freedom       5       493

Residual standard error: 3906
Estimated effects may be unbalanced
> summary(AOV_Model)
             Df   Sum Sq  Mean Sq F value Pr(>F)
RACE          5 1.86e+07  3718656    0.24   0.94
Residuals   493 7.52e+09 15260687
>summary(Healthcare_New$RACE)
  1   2   3   4   5   6
484   6   1   3   3   2
```

Conclusion:

F value is sort of low, which suggest that variation between hospital costs among different races is much smaller than the variation of hospital costs within each race, and P value being quite high shows that there's no relationship between race and hospital costs, thereby accepting the Null hypothesis. Additionally, we've more data for Race 1 as compared to other races (484 out of 500 patients) which make the observations skewed and thus all we can say is that there isn't enough data to verify whether race of a patient affects hospital costs.

QUESTION 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.

Solution:

Now to analyze the severity of costs we will use linear regression with TOTCHG(Cost) and independent variable along with AGE and Female as dependent variable.

Code:

```
> Healthcare_New$FEMALE <- as.factor(Healthcare_New$FEMALE)
> LM_Model <- lm(TOTCHG~AGE + FEMALE, data = Healthcare_New)
```

```
> summary(LM_Model)

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

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

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2719.4      261.4   10.40  < 2e-16 ***
AGE             86.0       25.5    3.37  0.00081 ***
FEMALE1       -744.2      354.7   -2.10  0.03638 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3850 on 496 degrees of freedom
Multiple R-squared:  0.0259,         Adjusted R-squared:  0.0219
F-statistic: 6.58 on 2 and 496 DF,  p-value: 0.00151
> summary(Healthcare_New$FEMALE)
  0   1
244 255
```

QUESTION 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.

Solution:

Using linear Regression, we will show whether length of stay depends on age, gender or race. Here we LOS is the dependent variable and age, gender and race are independent variables.

Code:

```
> LM_Model_2 <- lm(LOS ~ RACE + FEMALE + AGE, data = Healthcare_New)
```

```
> summary(LM_Model_2)

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

Residuals:
   Min     1Q Median     3Q    Max
 -3.21  -1.21  -0.86   0.14  37.79

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.8569     0.2316   12.34   <2e-16 ***
RACE2        -0.3750     1.3957   -0.27    0.788
RACE3         0.7892     3.3858    0.23    0.816
RACE4         0.5949     1.9572    0.30    0.761
RACE5        -0.8569     1.9627   -0.44    0.663
RACE6        -0.7188     2.3929   -0.30    0.764
FEMALE1       0.3539     0.3129    1.13    0.259
AGE          -0.0394     0.0226   -1.74    0.082 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.4 on 491 degrees of freedom
Multiple R-squared:  0.0087,        Adjusted R-squared:  -0.00543
F-statistic: 0.616 on 7 and 491 DF,  p-value: 0.743
```

Conclusion:

p-values for all independent variables are quite high thus signifying that there is no linear relationship between the given variables, finally concluding the fact that we can’t predict length of stay of a patient based on age, gender and race.

QUESTION 6:

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

Solution:

Using linear Regression, we can show which variable affects the hospital costs the most, thus TOTCHG becomes dependent variable and rest all variables are taken as independent.

Code:

```
> LM_Model_3 <- lm(TOTCHG ~ AGE + FEMALE + RACE + LOS + APRDRG, data = Healthcare_New)

> summary(LM_Model_3)


Call:

lm(formula = TOTCHG ~ AGE + FEMALE + RACE + LOS + APRDRG, data = Healthcare_New)


Residuals:

   Min     1Q Median     3Q    Max

 -6367   -691   -186    121  43412


Coefficients:

             Estimate Std. Error t value Pr(>|t|)

(Intercept)  5024.961    440.137   11.42  < 2e-16 ***

AGE           133.221     17.666    7.54  2.3e-13 ***

FEMALE1      -392.578    249.298   -1.57     0.12

RACE2         458.243   1085.232    0.42     0.67

RACE3         330.518   2629.512    0.13     0.90

RACE4        -499.382   1520.929   -0.33     0.74

RACE5       -1784.578   1532.005   -1.16     0.24

RACE6        -594.292   1859.127   -0.32     0.75

LOS           742.964     35.046   21.20  < 2e-16 ***

APRDRG         -7.818      0.688  -11.36  < 2e-16 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 2620 on 489 degrees of freedom

Multiple R-squared:  0.554,          Adjusted R-squared:  0.546
```

F-statistic: 67.6 on 9 and 489 DF, p-value: <2e-16

Conclusion :

Age and length of stay affect the entire hospital costs. Additionally, there is positive relationship between length of stay to the cost, so with an rise of 1 day there is an addition of a value of 742 to the cost.