Healthcare Cost Analysis

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

Task 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.

```
# Import excel
library(readxl)
hospital_costs <- read_excel("1555054100_hospitalcosts.xlsx")
# check structure of the data
str(hospital_costs)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    500 obs. of  6 variables:
##  $ AGE   : num  17 17 17 17 17 17 17 16 16 17 ...
##  $ FEMALE: num  1 0 1 1 1 0 1 1 1 1 ...
##  $ LOS   : num  2 2 7 1 1 0 4 2 1 2 ...
##  $ RACE  : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ TOTCHG: num  2660 1689 20060 736 1194 ...
##  $ APRDRG: num  560 753 930 758 754 347 754 754 753 758 ...
```

```
# column names
names(hospital_costs)
```

```
## [1] "AGE"    "FEMALE" "LOS"    "RACE"   "TOTCHG" "APRDRG"
```

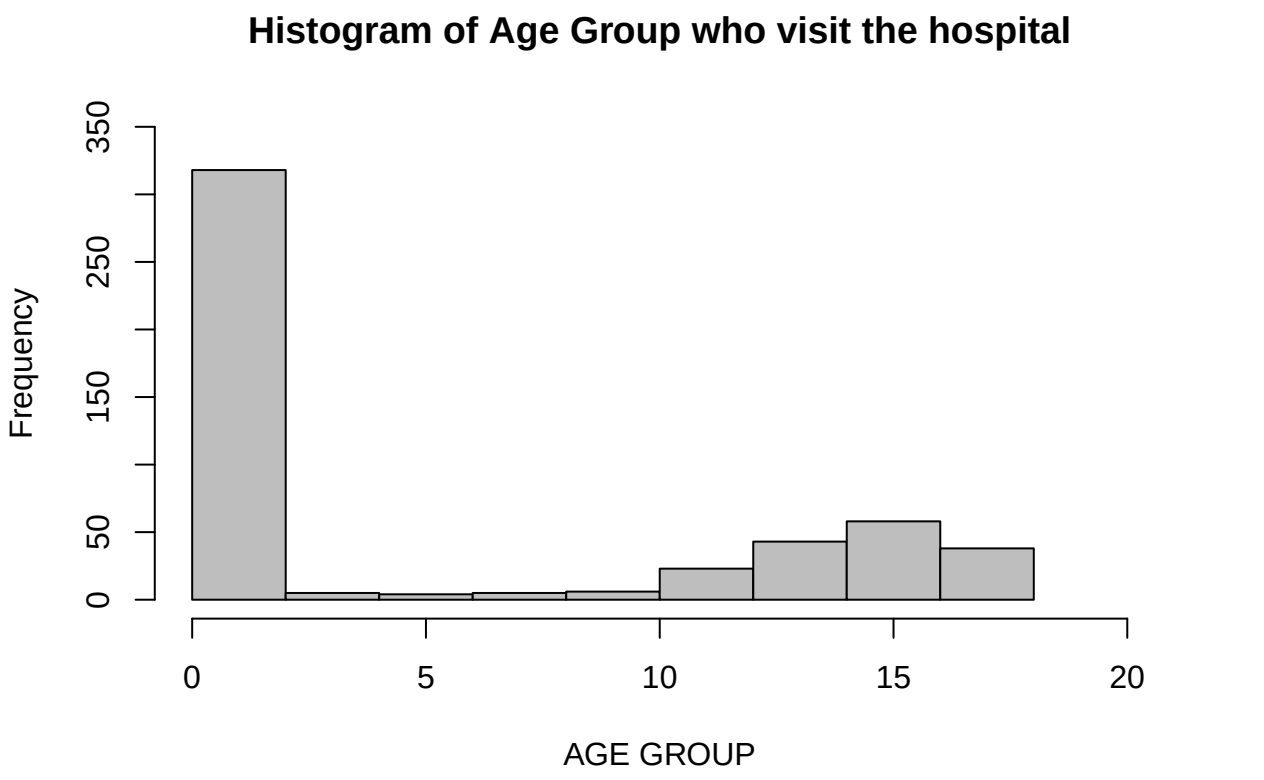
```
summary(hospital_costs)  # summary of the data
```

```
##       AGE             FEMALE           LOS              RACE
##  Min.   : 0.000   Min.   :0.000   Min.   : 0.000   Min.   :1.000
##  1st Qu.: 0.000   1st Qu.:0.000   1st Qu.: 2.000   1st Qu.:1.000
##  Median : 0.000   Median :1.000   Median : 2.000   Median :1.000
##  Mean   : 5.086   Mean   :0.512   Mean   : 2.828   Mean   :1.078
##  3rd Qu.:13.000   3rd Qu.:1.000   3rd Qu.: 3.000   3rd Qu.:1.000
##  Max.   :17.000   Max.   :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
##
```

As we see, the age category of people who frequently visit the hospital is 0-1.

```
# Histogram
hist(as.numeric(hospital_costs$AGE), main="Histogram of Age Group who visit the hospital",
     xlab="AGE GROUP", border="black", col=c("gray"), xlim=c(0,20), ylim=c(0,350))
```




```
# Summarize expenditure based on age group
expenditure = aggregate(TOTCHG ~ AGE, FUN=sum, data=hospital_costs)
expenditure
```

```
##    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
```

The age category with the maximum expenditure is the 0-1 with 678118.

Task 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.

```
# Convert APRDRG column to factor
hospital_costs$APRDRG = as.factor(hospital_costs$APRDRG)
```

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

```
##    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
```

```
# find the diagnosis-related group that has maximum hospitalization and expenditure.
diagnosis[which.max(diagnosis$TOTCHG), ]
```

```
##    APRDRG TOTCHG
## 44    640 437978
```

Task 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.

```
# Convert RACE column to factor
hospital_costs$RACE = as.factor(hospital_costs$RACE)
# check for null values
sapply(hospital_costs, function(x) sum(is.na(x)))
```

```
##    AGE FEMALE    LOS   RACE TOTCHG APRDRG
##      0      0      0      1      0      0
```

As we see there is one null value that needs to be removed.

```
# use na.omit() to remove the null value
hospital_costs = na.omit(hospital_costs)
# check again if the null value exists
sapply(hospital_costs, function(x) sum(is.na(x)))
```

```
##    AGE FEMALE    LOS   RACE TOTCHG APRDRG
##      0      0      0      0      0      0
```

```r
summary(hospital_costs$RACE)
```

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

As we see, 484 patients belongs to age group 0-1.

Task 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.

```r
# Use Linear Regression Model
model <- lm(formula = TOTCHG ~ AGE + FEMALE, data = hospital_costs)
summary(model)
```

```
##
## Call:
## lm(formula = TOTCHG ~ AGE + FEMALE, data = hospital_costs)
##
## 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 ***
## AGE            86.04      25.53   3.371 0.000808 ***
## FEMALE       -744.21     354.67  -2.098 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
```

```r
# Convert FEMALE column to factor
hospital_costs$FEMALE = as.factor(hospital_costs$FEMALE)
summary(hospital_costs$FEMALE)
```

```
##   0   1
## 244 255
```

CONCLUSION: The severity of the hospital costs by age is very important as seen by the high p-value and the statistical significance (*** next to it). We see that there is similar distribution of genders. Based on the negative coefficient we conclude that females spend less hospital costs than males.

Task 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.

```r
model_2 <- lm(formula = LOS ~ AGE + FEMALE + RACE, data = hospital_costs)
summary(model_2)
```

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

CONCLUSION : The p-value is more than 0.05 for both age,gender and race which signifies that there is no relationship between these variables. As a result, we can't predict the length of stay for inpatients.

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

```r
hospital_costs$APRDRG = as.numeric(hospital_costs$APRDRG)
hospital_costs$RACE = as.numeric(hospital_costs$RACE)
hospital_costs$FEMALE = as.numeric(hospital_costs$FEMALE)
model3<-lm(TOTCHG~.,data=hospital_costs)
summary(model3)
```

```
##
## Call:
## lm(formula = TOTCHG ~ ., data = hospital_costs)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6572   -633   -182    123  43351
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6179.93     595.27  10.382  < 2e-16 ***
## AGE           142.22      17.45   8.148 3.06e-15 ***
## FEMALE       -413.53     245.86  -1.682   0.0932 .
```

```
## LOS           732.01      34.76  21.059  < 2e-16 ***
## RACE         -201.23     226.76  -0.887   0.3753
## APRDRG       -125.51      10.73 -11.700  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2600 on 493 degrees of freedom
## Multiple R-squared:  0.558,  Adjusted R-squared:  0.5535
## F-statistic: 124.5 on 5 and 493 DF,  p-value: < 2.2e-16
```

CONCLUSION : We see that the variables AGE and LOS affect the hospital costs. Also we can see that for the increasement of day stay by one, the hospital costs will increase by 732.