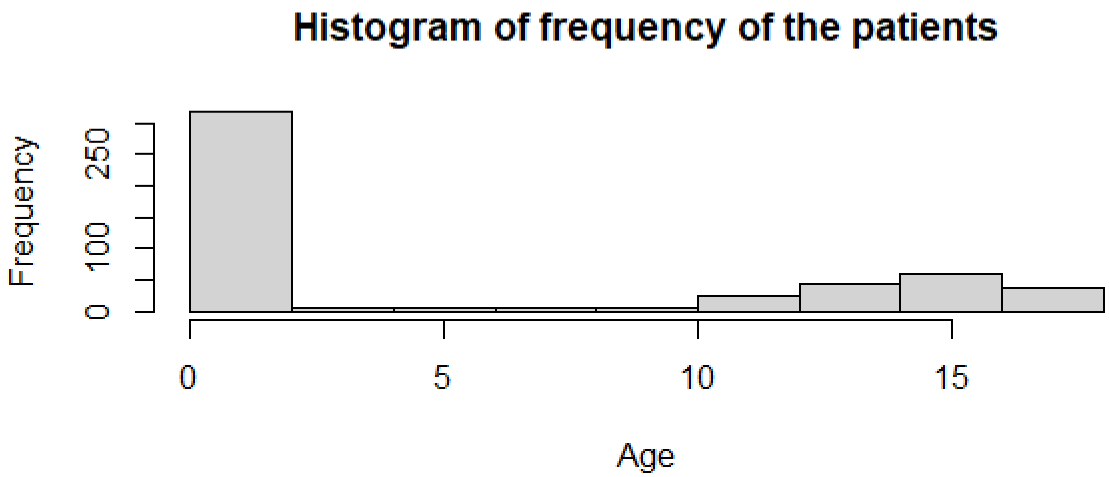
Project Writeup: HEALTHCARE ANALYSIS

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 category of maximum frequency of hospital visit, I have used Histogram.

   Below is the histogram that I obtained in RStudio.

   To find expenditure the all the ages I have used aggregate function. Then I took the max function to find which age has maximum hospital cost.

   Conclusion: from the output obtained I observed that infants have maximum visits to hospital.

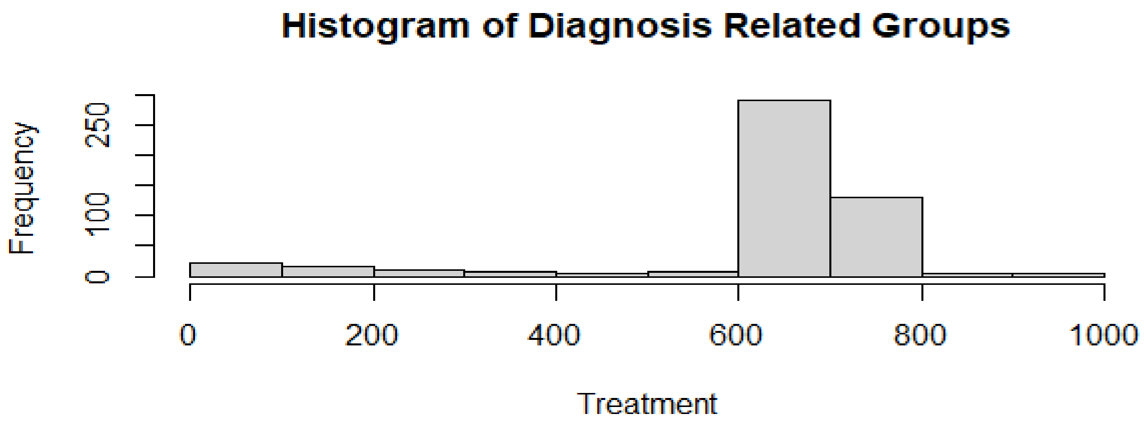
   Also the infants have maximum expenditure or hospital cost.

   So I infer that number of hospital visits is proportional to the hospital cost.

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.

   First I have visualized the frequency of patients using histogram.

   Below is the histogram that I obtained in RStudio.



Then I took the summary function followed by which.max to find maximum index of the category.

Then I used aggregate function as used in first question above.

Conclusion: I found out that category 640 has maximum hospitalization cost. Also this category has maximum hospital cost.

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.

   Firstly I removed the NA values using na.omit. Then I used anova function between TOTCHG and AGE function.

   I got following results:

```
Df    Sum Sq  Mean Sq F value Pr(>F)
RACE          1 2.488e+06  2488459   0.164  0.686
Residuals   497 7.540e+09 15170268
```

   Conclusion: I got P value of 68.6% which is greater than 5%. So null hypothesis is rejected.

   Hence I conclude there is no relation between race of the patient and 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.

   Here I took linear regression between the TOTCHG and dependent variables AGE and FEMALE.

   I got the following results:

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

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
```

Conclusion: P values of both AGE and FEMALE are less than 5%.

But the AGE is more significant than FEMALE due as it can be seen from the significant code(AGE has *** and FEMALE has * significant code).

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.

Here I used linear regression between length of stay(LOS) and dependent variables AGE, FEMALE and RACE.

I got the following results:

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

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
```

Conclusion: I observed that for all the dependent variables then P value is more than 5%.

So I conclude that length of stay cannot 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.

Here I used linear regression between TOTCHG and all the dependent variables.(AGE, FEMALE, LOS, RACE and APRDRG).

I got the following results:

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


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
```

Conclusion: From the P values and the significant codes I conclude that AGE, RACE and APRDRG affect more the hospital costs than other variables. .i.e. They are more significant than other variables.