PROJECT: Healthcare Cost Analysis

Business Scenerio:

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.

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

Analysis to be done:

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.

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.

Code:

#Importing file

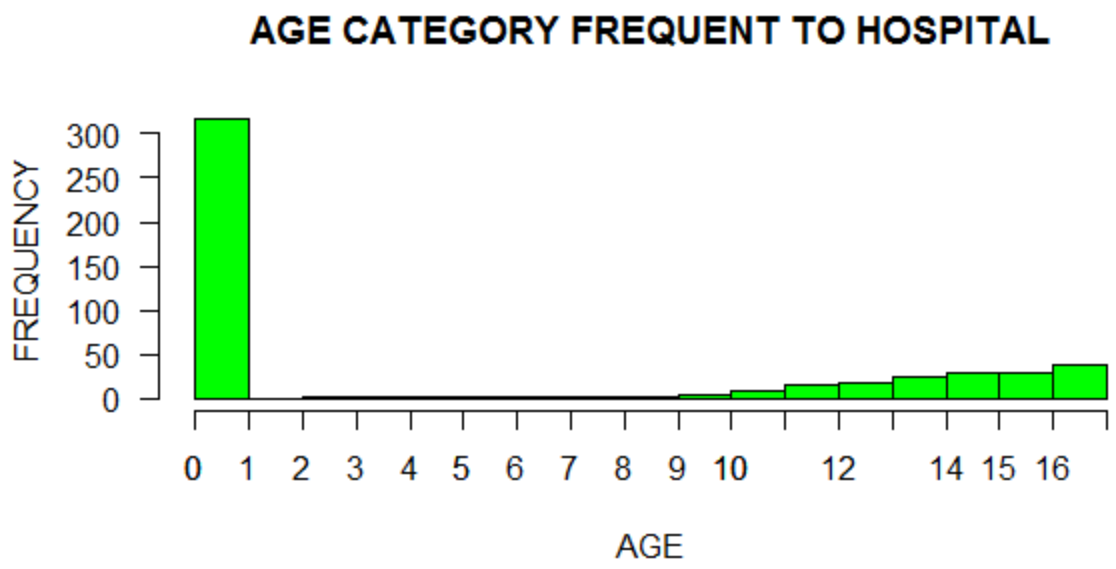
```
library(dplyr)

library(readxl)

data = readxl::read_excel(choose.files())

View(data)
```

#Dealing with missing values

```
anyNA(data)

data=na.exclude(data)
```

#Histogram of age Group Frequent to Hospital

```
hist(data$AGE, col = "green",border = "black", xlab = "AGE", ylab = "FREQUENCY",

    main = "AGE CATEGORY FREQUENT TO HOSPITAL",breaks = 17,las =1)

axis(side = 1,at = seq(0,17))
```



#Grouping charges by age

summary(as.factor(data$AGE))

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

306 10 1 3 2 2 2 3 2 2 4 8 15 18 25 29 29 38

total_charge_by_age = aggregate(data = data ,TOTCHG~AGE, FUN = sum)

#Table for Total Charge by Age

head(total_charge_by_age)

##Output

	AGE	TOTCHG
1	0	676962
2	1	37744
3	2	7298
4	3	30550
5	4	15992
6	5	18507

#Maximum Charge for AGE group

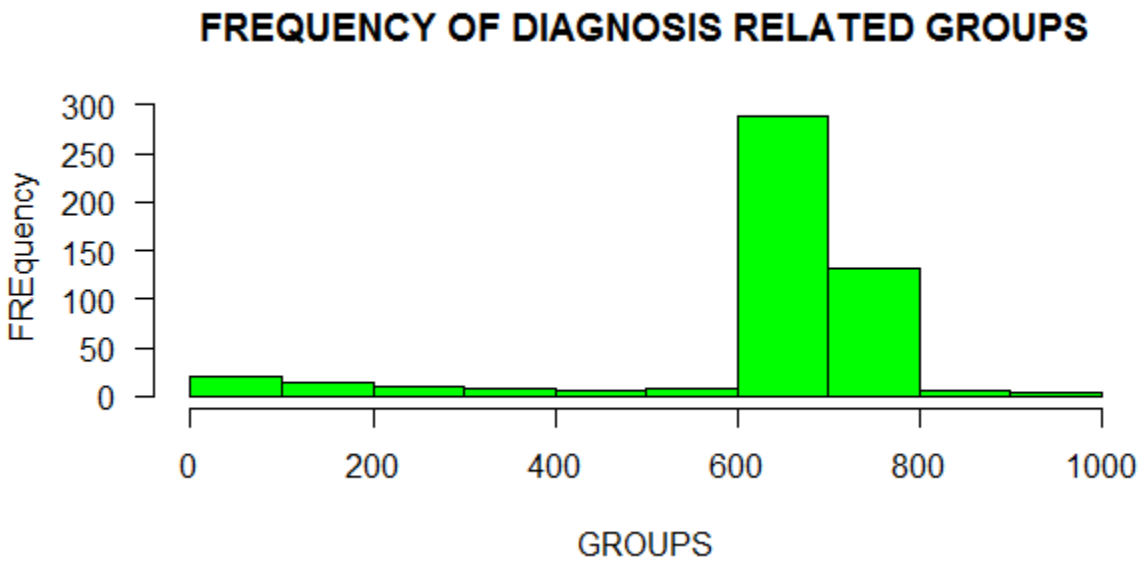
```
max(total_charge_by_age$TOTCHG)
```

#Output

676962

#Histogram of Diagnosis Related Group

```
hist(data$APRDRG,main = "FREQUENCY OF DIAGNOSIS RELATED GROUPS",xlab = "GROUPS",
    ylab = "FREquency",las=1, col = "green")
```



#Calculation of Maximum Diagnosed Group

```
diagnosis_groups = as.factor(data$APRDRG)

summary(diagnosis_groups)

max(summary(diagnosis_groups))

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

```
max(group_charges)
```

#Output

```
266
436822
```

#Relation of race to hospitalization charge

```
#H0: There is a relation between Race and Hospitalization costs
#H1: There is no relation
model_race = aov( data= data, TOTCHG~RACE)
summary(model_race)
```

#Output

```
             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
#p-value is high so we reject our NULL  Hypothesis.
```

#Relation of charges to age and gender

```
model_cost = lm(data = data,TOTCHG~AGE+FEMALE)
summary(model_cost)
```

#Output

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

FEMALE1 -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

#p-value is less than 0.05 our NULL Hypothesis is not rejected. Hence there is a relation between Charges and age and gender. As it can be seen AGE is significant variable.

#relation between LOS and AGE,GENDER and RACE

```
model_los = lm(data = data,LOS~AGE+ FEMALE +RACE)
summary(model_los)
```

#Output

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

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	.
FEMALE1	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

#p-value is greater than 0.05 hence NULL Hypothesis is not rejected. Hence on the basis of AGE, RACE and GENDER we cannot predict LOS.

#Factors affecting Total charges

```
model_cost1 = lm(data = data,TOTCHG~.)

summary(model_cost1)
```

#Output

```
Call:

lm(formula = TOTCHG ~ ., data = data)

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 ***

FEMALE1     -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

#From the Output it can be deduced that AGE, LOS and ARDRG are the signifaicant factors in predicting Hospital Charges .

ANALYSIS:

Our First conclusion was that infant category has the max hospital visits . The summary of Age gives us the exact numerical output showing that Age 0 patients(306) have the max visits followed by Ages 15-17. Hence maximum discharge cost collected was also from age group 0 which is 676962.

From the summary of diagnosis group function we conclude that category 640 has the maximum hospitalizations by a huge number (266 out of 500), along with this it also has the highest hospitalization cost .i.e 436822.

P-value for RACE and Charges was high(68%). It shows there was no relation between them.

And it was also seen that LOS was not affected by AGE, RACE and Gender.

AND it can be seen that Total Hospitalization Cost is affected by LOS, AGE and APRDRG as they are the significant variables.