#### Analyze the Healthcare cost and Utilization in Wisconsin hospitals ####

```r
getwd() ## to see which is our current working directory

#### Reading the data from the csv file ####

Hospital <- read.csv("HospitalCosts.csv")

View(Hospital) ## first look at the dataset for further analysis

str(Hospital) ## structure of the dataset

dim(Hospital) ## get dimension of the dataset
```

## 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

```r
library(tidyverse) ## loading required package for visualization

## Let's create a histogram which shows the counts of age category which is frequent in the hospital

Hospital_Histogram <- ggplot(data = Hospital, mapping = aes(x = AGE)) +
  geom_histogram(color = "black", fill = "green") +
  ggtitle("Frequency of Hospital visits by age")

Hospital_Histogram

age <- as.factor(Hospital$AGE) # convering age into factor for further analysis

summary(age)
```

## from histogram and summary we can see that 0 or newborns has maximum visits 307 in the hospital

## Now we will use the aggregate function to see which age category has maximum expenditure

```r
cost <- aggregate(TOTCHG~AGE, FUN = sum, data = Hospital)

cost

cost[which.max(cost$TOTCHG),] ## to get which is the maximum
```

## Also from above analysis we found that age 0 or newborns have maximum expenditure

##2 find the diagnosis related group that has maximum hospitalization and expenditure

```r
## Here we will also create a histogram to see which diagnosis related group has maximum hospitalization

aprdrg_histogram <- ggplot(data = Hospital, mapping = aes(x = APRDRG)) +
  geom_histogram(color = "black", fill = "blue") +
  ggtitle("Frequency of Diagnosis Related Groups")

aprdrg_histogram
```

```
aprdrg <- as.factor(Hospital$APRDRG)

summary(aprdrg)

which.max(summary(aprdrg))
```

## So from above analysis the diagnosis related group 640 has maximum hospitalization compared to other

## Let's check which diagnosis related group has maximum expenditure using the same aggregate function

```
Treatment_cost <- aggregate(TOTCHG~APRDRG, FUN = sum, data = Hospital)

Treatment_cost

Treatment_cost[which.max(Treatment_cost$TOTCHG),]
```

## Here we found that diagnosis related group 640 has maximum hospitalization as well as maximum expenditure of 437978

## 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

```
## Lets check if there is any missing values in data ##

sum(is.na(Hospital))

View(Hospital[is.na(Hospital)])

## So we found that there is only one missing value and also we can use following function to remove any missing data ##

Hospital <- na.omit(Hospital)

dim(Hospital)

Hospital$RACE <- as.factor(Hospital$RACE)
```

## We will use ANOVA Table to check if RACE of the patient is related to the hospitalization costs.

```
race_anova <- aov(TOTCHG~RACE, data = Hospital)

race_anova

summary(race_anova)

summary(Hospital$RACE)
```

## As the P value is very high so there is no relationship between the race and hospital cost.

## Also from summary of race we found that RACE 1 has max number of records = 484 out of 500 so maybe there is no sufficient data of other RACES to see if there is any effect of RACE on hospital costs.

### 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.

```
## Here we will use basic linear regression function lm to analyze if AGE and GENDER has any effect
on cost

Hospital$FEMALE <- as.factor(Hospital$FEMALE)

Severity_model <- lm(formula = TOTCHG~AGE+FEMALE, data = Hospital)

summary(Severity_model)

summary(Hospital$FEMALE)
```

## By looking at the summary of the model we found that age has more impact than gender on the hospital cost as per the p-values. Also the summary of Gender shows there are 244 Male and 245 Female which are almost equal

## Let's explore further using ANOVA function to analyze

```
Severity_anova <- aov(TOTCHG~AGE+FEMALE, data = Hospital)

Severity_anova

summary(Severity_anova)
```

## Here also the ANOVA table shows that AGE has more impact than GENDER on the hospital costs

### 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.

```
## We will use the same linear regression method to analyze if the length of stay can be predicted
from age, gender, and race

Hospital$RACE <- as.factor(Hospital$RACE)

LOS_model <- lm(formula = LOS~AGE+FEMALE+RACE, data = Hospital)

summary(LOS_model)
```

## from above analysis we found that the p-values are high for AGE, GENDER and RACE so we can say that there is no relationship found that can predict the length of stay of a patient based on AGE, GENDER and RACE.

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

```
Complete_Analysis <- lm(formula = TOTCHG~., data = Hospital)

summary(Complete_Analysis)
```

## from above analysis model we can see that AGE and LOS has main effect on the hospital cost