Healthcare Cost Analysis

```
setwd("C:/Users/SUDESHNA/Desktop/Rstudio_Projects")
print(getwd())
```

```
## [1] "C:/Users/SUDESHNA/Desktop/Rstudio_Projects"
```

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

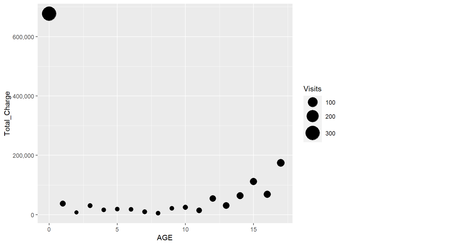
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(scales)
rm(list=ls())
hops<-read.csv("HospitalCostDataset.csv")
hops<-na.omit(hops)
```

1.The agency wants to find the age category of people who frequently visit the hospital and has the maximum expenditure.

```
hops$RACE <- as.factor(hops$RACE)
hops$FEMALE <- as.factor(hops$FEMALE)
Visits <- summary(as.factor(hops$AGE))
df <- summarise(group_by(hops,AGE), Total_Charge = sum(TOTCHG))

print(ggplot(data = df, aes(x = AGE, y = Total_Charge)) +
   geom_point(aes(size = Visits))
   + scale_y_continuous(label = scales::comma)+
   scale_size(range = c(2.5, 9.5)))
```



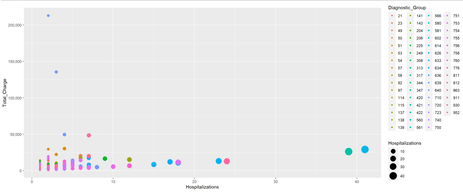

```
print(arrange(df,desc(Total_Charge))[1,])
```

```
## # A tibble: 1 x 2
##     AGE Total_Charge
##   <int>        <int>
## 1     0       676962
```

2.To find out the expensive treatments, the agency wants to find the diagnosis-related group that has maximum hospitalization and expenditure.

```
hop<-hops[hops$LOS > 0,]
hop$APRDRG <- as.factor(hop$APRDRG)


df1 <- aggregate(hop$TOTCHG ~ hop$APRDRG + hop$LOS, data = hop, FUN = sum, na.rm = TRUE)
colnames(df1)[1] <- "Diagnostic_Group"
colnames(df1)[2] <- "Hospitalizations"
colnames(df1)[3] <- "Total_Charge"
print(ggplot(data = df1, aes(x = Hospitalizations , y = Total_Charge ))
      + geom_point(aes(
        size = Hospitalizations,
        color = Diagnostic_Group))
      + scale_y_continuous(label = scales::comma)
      + scale_size(range = c(1.75,8)))
```




```
df2<-df1[df1$`Total_Charge`== max(df1$`Total_Charge`),]
print(df2)
```

```
##    Diagnostic_Group Hospitalizations Total_Charge
## 42              640                2       212832
```

```
df3<-df1[df1$`Hospitalizations`== max(df1$`Hospitalizations`),]
print(df3)
```

```
##     Diagnostic_Group Hospitalizations Total_Charge
## 116              602                41        29188
```

3. To determine if the race of the patient is related to the hospitalization costs.

Ho/Null-Hypothesis : Race of patient is not related to Total Cost

Then, to verify if the races made an impact on the costs, perform an ANOVA with the following variables:

ANOVA dependent variable: TOTCHG Categorical/grouping variable: RACE Missing values: 1 NA value, use na.omit to remove the NA value numerical/int ~ categorical variable, dependent variable ~ independent variable

```
model <- aov(TOTCHG ~ RACE, data = hops)
alpha = 0.05
Pvalue = summary(model)[[1]][,"Pr(>F)"][1]
print(Pvalue)
```

```
## [1] 0.9428886
```

```
print(Pvalue < alpha) #if P-value < alpha is true we reject the null hypothesis
```

```
## [1] FALSE
```

Here we do not reject null hypothesis therefore, we can say that there is no relation between the race of patient and the hospital cost.

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.

Let, H0/Null-Hypothesis be that there is no relation between the hospital costs by age and gender

```
av <- aov ( TOTCHG ~ AGE + FEMALE, data = hops)
Pval_Age = summary(av)[[1]][,"Pr(>F)"][1]
Pval_Gender = summary(av)[[1]],"Pr(>F)"][2]
print(Pval_Age)
```

```
## [1] 0.003227653
```

```
print(Pval_Gender)
```

```
## [1] 0.03638199
```

```
print(Pval_Age < alpha && Pval_Gender < alpha) #if P-value < alpha is true we reject the null hypothesis
```

```
## [1] TRUE
```

Therefore we can conclude that there is severity of the hospital costs by age and gender, model is statistically significant.

5.The agency wants to find if the length of stay can be predicted from age, gender, and race.

let the H0/null hypothesis be the LOS cannot be predicted from age, gender & race

```
aov_model <- aov ( LOS ~ AGE + FEMALE + RACE, data = hops)
Pval_model_age = summary(aov_model)[[1]][,"Pr(>F)"][1]
Pval_model_gender = summary(aov_model)[[1]][,"Pr(>F)"][2]
Pval_model_race = summary(aov_model)[[1]][,"Pr(>F)"][3]
print(Pval_model_age)
```

```
## [1] 0.125039
```

```
print(Pval_model_gender)
```

```
## [1] 0.2293082
```

```
print(Pval_model_race)
```

```
## [1] 0.9921102
```

```
print(Pval_model_age < alpha && Pval_model_gender < alpha && Pval_model_race < alpha)  #if P-value < alpha is true we reject the null hypothesis
```

```
## [1] FALSE
```

Therefore we don't reject the null hypothesis. In conclusion we can say that LOS cannot be predicted from age, gender & race.

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

let the null hypothesis be the TOTCHG cannot be predicted from any of the parameters.

```
av_model <- aov ( TOTCHG ~ ., data = hops)
P_val_model = summary(av_model)[[1]]["Pr(>F)"]
print(P_val_model)
```

```
##              Pr(>F)
## AGE          2e-05 ***
## FEMALE     0.00218 **
## LOS        < 2e-16 ***
## RACE       0.85885
## APRDRG     < 2e-16 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

THEREFORE, WE CAN SAY APART FROM RACE ALL THE OTHER VARIABLES AFFECT THE TOTCHG, SINCE P-VALUE OF AGE,GENDER,LOS,APRDRG ARE LESS THAN APLHA=0.05 THEREFORE WE REJECT NULL HYPOTHESIS