**Project: HealthCare Cost Analysis :**

setwd("C:/Users/Workstation/OneDrive/Desktop/Data Science\_Material")

getwd()

library(readxl)

## *Reading Excel from Local*

Hospital\_Data <-read\_xlsx("Hospital\_Cost\_Analysis.xlsx")

View(Hospital\_Data)

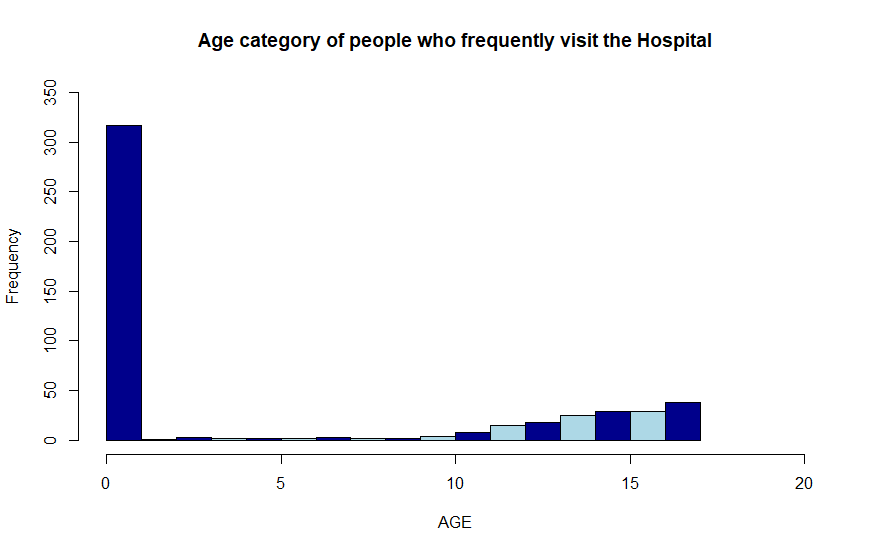
summary(Hospital\_Data)

**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.**

hist(Hospital\_Data$AGE, col= c('darkblue','lightblue'),breaks = 17, main= 'Age category of people who frequently visit the Hospital',xlab='AGE', border = 'black',xlim = c(0,20), ylim = c(0,350))

class(Hospital\_Data$AGE)

AGE <- as.factor(Hospital\_Data$AGE)

summary(AGE)

*#Here, We analyze the maximum expenditure categorized by Age using Aggregate function*

age\_category\_freqently\_visits <- aggregate(TOTCHG ~ AGE, FUN = sum, data = Hospital\_Data)

max(aggregate(Hospital\_Data$TOTCHG ~ Hospital\_Data$AGE, FUN = sum))

age\_category\_freqently\_visits[which.max(age\_category\_freqently\_visits$TOTCHG),]

Console Output : AGE TOTCHG

1. 0 678118

Conclusion: Age Category : '0-1' (Infants) frequently visit the hospital with maximum expenditure of '678118', as Histogram shows frequency of Infants (0-1) is >300.

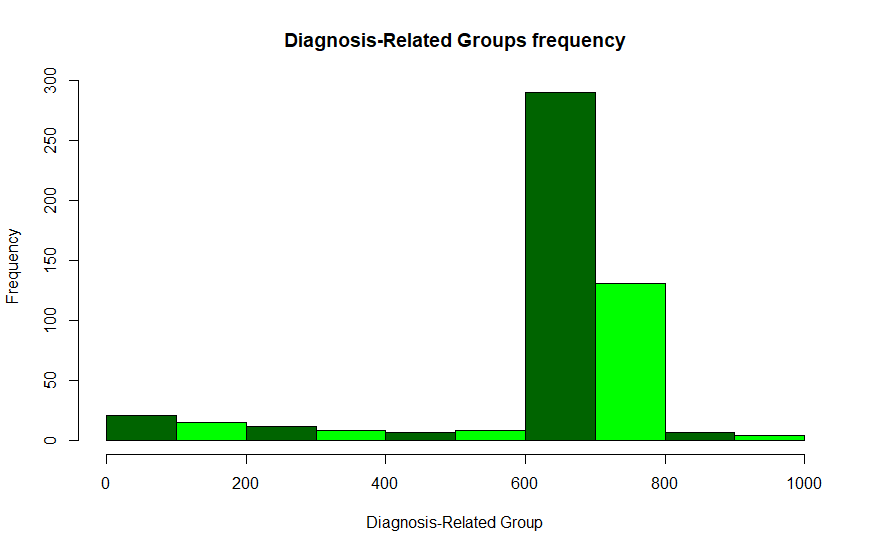
**Task 2:** In order of severity of the diagnosis and treatments and to find out the expensive treatments, the agency wants to find the diagnosis-related group that has maximum hospitalization and expenditure.

class(Hospital\_Data$APRDRG)

as.factor(Hospital\_Data$APRDRG)

summary(as.factor(Hospital\_Data$APRDRG))

max(summary(as.factor(Hospital\_Data$APRDRG)))

hist(Hospital\_Data$APRDRG, main = 'Diagnosis-Related Groups frequency', xlab = 'Diagnosis-Related Group', col= 'green')

Diagnosis\_Cost <- aggregate(TOTCHG ~ APRDRG, FUN = sum, data = Hospital\_Data)

max(Diagnosis\_Cost <- aggregate(TOTCHG ~ APRDRG, FUN = sum, data = Hospital\_Data))

Diagnosis\_Cost[which.max(Diagnosis\_Cost$TOTCHG),]

Console Output : 437978

CONCLUSION: - From the above output we conclude that 640 has the maximum hospitalization (267 out of 500) and also has the highest expenditure cost(437978).

Task 3 : To make sure that there is no malpractice, the agency needs to analyse if the race of the patient is related to the hospitalization costs.

summary(as.factor(Hospital\_Data$RACE))

anyNA(Hospital\_Data)

Hospital\_Data <- na.omit(Hospital\_Data)

anyNA(Hospital\_Data)

*# As can be seen 484 patients out of 499 falls under group 1, showing that the number of observations for 1 category is way higher than others - hence data is skewed. This will only affect the results from linear regression or ANOVA analysis*

cor(Hospital\_Data$TOTCHG, Hospital\_Data$RACE)

Console Output : -0.0181643

*# correlation value is too less indicating that there is no relation*

Race\_Influence <- lm(Hospital\_Data$TOTCHG ~ Hospital\_Data$RACE)

summary(Race\_Influence)

Console Output : Residuals: Min 1Q Median 3Q Max

-2256 -1560 -1227 -258 45600

Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept) 2925.7 405.0 7.224 1.92e-12 \*\*\*

Hospital\_Data$RACE -137.3 339.1 -0.405 0.686

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3895 on 497 degrees of freedom

Multiple R-squared: 0.0003299, Adjusted R-squared: -0.001681

F-statistic: 0.164 on 1 and 497 DF, p-value: 0.6856

Conclusion: #R- Squared Value is 0.0003299, very less than 0.5

#pValue is 0.69 it is much higher than 0.5

#We can say that race does not affect the hospitalization costs

*## Analyzing Using ANOVA*

Race\_Influence\_AOV <- aov(TOTCHG ~ RACE, data = Hospital\_Data)

Race\_Influence\_AOV

summary(Race\_Influence\_AOV)

Console 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

Conclusion : The residual variance (deviation from original) (of all other variables) is very high. This implies that there is extraordinarily little influence from RACE on hospitalization costs

#As can be seen, the degree of freedom (Df) for RACE is 1 and that of residuals is 497 observations

#The F-Value, the test statistic is 0.16 which is much less than 0.5 showing that RACE does not affect the hospitalization cost.

#The Pr(>F), the p value of 0.69 is high confirming that RACE does not affect hospitalization cost.

Task 4 : To properly utilize the costs, the agency has to analyse the severity of the hospital costs by age and gender for the proper allocation of resources.

summary(Hospital\_Data$FEMALE)

*#As can be seen here, there is equal distribution of male and female in the group*

Age\_Gender\_Influence\_Cost <- lm(TOTCHG ~ AGE + FEMALE, data = Hospital\_Data)

summary(Age\_Gender\_Influence\_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 : Since the pValues of AGE is much lesser than 0.05, the ideal statistical significance level, and it also has three stars (\*\*\*) next to it, it means AGE has the most statistical significance

#Similarly, gender is also less than 0.05.

#Hence, we can conclude that the model is statistically significant

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.

Age\_Gender\_Race\_Inf\_lModel <- lm(LOS ~ AGE + FEMALE + RACE, data = Hospital\_Data)

Age\_Gender\_Race\_Inf\_lModel

summary(Age\_Gender\_Race\_Inf\_lModel)

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 : R - Squared value = 0.007898 (<0.05)

#The p-value is higher than 0.05 for age, gender and race, indicating there is no linear relationship between these variables and length of stay.

#Hence, age, gender and race cannot be used to predict the length of stay of inpatients.

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

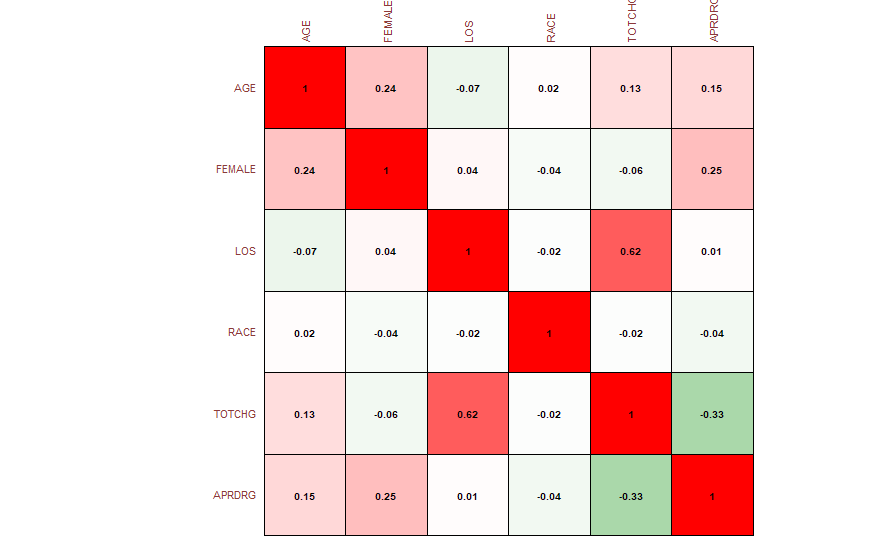
library(corrplot)

library(car)

library(MLmetrics)

factors\_Inf\_cost <- cor(Hospital\_Data[,1:6])

View(factors\_Inf\_cost)

cor\_plot <- corrplot(factors\_Inf\_cost, method = "color", outline = T, cl.pos = 'n', rect.col = "black", tl.col = "indianred4", addCoef.col = "black", number.digits = 2, number.cex = 0.60, tl.cex = 0.7, cl.cex = 1, col = colorRampPalette(c("green4","white","red"))(100))

*#As it is apparent from the coefficient values, Age, Length of stay (LOS) and patient refined diagnosis related groups(APRDRK)*

*#Also, RACE and Female is the least significant. build a model after removing RACE and LENGTH*

Hospital\_Data <- subset(Hospital\_Data, select = -c(RACE, FEMALE))

names(Hospital\_Data)

comp\_analysis = lm(formula = TOTCHG ~ . , data = Hospital\_Data)

summary(comp\_analysis)

Residuals: Min 1Q Median 3Q Max

-6603 -719 -169 124 43350

Coefficients : Estimate Std. Error t value Pr(>|t|)

(Intercept) 4960.1705 433.6579 11.44 < 2e-16 \*\*\*

AGE 128.5519 17.0946 7.52 2.59e-13 \*\*\*

LOS 740.8057 34.9161 21.22 < 2e-16 \*\*\*

APRDRG -8.0055 0.6643 -12.05 < 2e-16 \*\*\*

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2617 on 495 degrees of freedom

Multiple R-squared: 0.5506, Adjusted R-squared: 0.5479

F-statistic: 202.2 on 3 and 495 DF, p-value: < 2.2e-16

*#Since APRDRG has -ve t-value, dropping it.*

comp\_analysis\_1 <- lm(formula = TOTCHG ~ AGE + LOS, data = Hospital\_Data)

summary(comp\_analysis\_1)

Residuals: Min 1Q Median 3Q Max

-4783 -1103 -458 -133 41382

Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept) 200.66 203.48 0.986 0.325

AGE 97.96 19.21 5.101 4.83e-07 \*\*\*

LOS 734.27 39.66 18.512 < 2e-16 \*\*\*

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2973 on 496 degrees of freedom

Multiple R-squared: 0.4188, Adjusted R-squared: 0.4164

F-statistic: 178.7 on 2 and 496 DF, p-value: < 2.2e-16

*#Removing Race and gender doesn’t change the R2 value. It doesn’t impact cost*

*#Removing APRDRG increases the standard error. Hence “comp\_analysis” model holds good.*

***#*** *AGE, LOS AND APRDRG that mainly affects hospital costs.*

**Analysis Conclusion:**

* As is evident in the multiple models above, health care costs is dependent on age, length of stay and the diagnosis type.
* Healthcare cost is the most for patients in the 0-1 yrs age group category
* Maximum expenditure for 0-1 yr is 678118
* Length of Stay increases the hospital cost
* All Patient Refined Diagnosis Related Groups also affects healthcare costs
* 640 diagnosis related group had a max cost of 437978
* Race or Female(gender) doesn’t have that much impact on hospital cost