R – Project: “Healthcare Cost “

Course: “Data Science with R”

Business Scenario:

Description:Analyse the Healthcare Cost and Utilization in Wisconsin hospital

Expectation /Goals:

The goals of this project are:

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 analyse if the race of the patient is related to the hospitalization costs.

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

R-Code File:



R-Code with Output screenshots and Analysis Details:

#Title: "Healthcare Cost"

#Author: "SANJAY TALLOLLI"

#---------------------------------------------------------------------------

getwd()

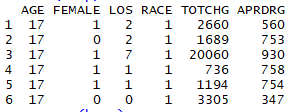
setwd("I:\\SIMPLILEARN COURSES LIVE 2018\\DATASCIENCE WITH R\\COURSE MATERIALS\\Project\\Projects for Submission\\Healthcare\\Healthcare")

# Importing data sets

hosp <- read.csv ("I:\\SIMPLILEARN COURSES LIVE 2018\\DATASCIENCE WITH R\\COURSE MATERIALS\\Project\\Projects for Submission\\Healthcare\\Healthcare\\ HospitalCosts.csv", header = T)

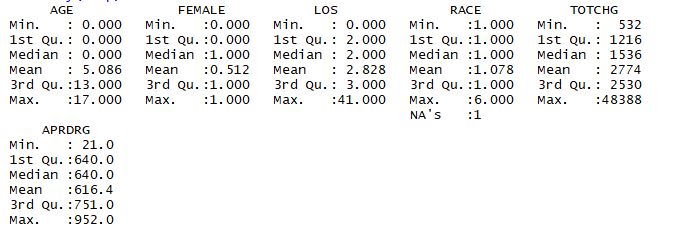
head(hosp) # Analysis: 500 observations and 6 Variables

# **OUTPUT**



summary(hosp)

**#OUTPUT:**



# Goal/Expectation (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

attach(hosp)

hist(AGE)

**#OUTPUT**



# To see the value of category of infants

ag <- as.factor(AGE)

summary(ag)

**# OUTPUT:**



# Age category of 0 seems to be frequently using the hospital

tapply(TOTCHG,AGE,sum)

**#OUTPUT:**



which.max(tapply(TOTCHG,AGE,sum))

**#OUTPUT:-**



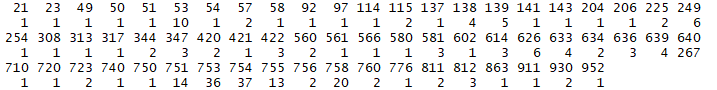
# Analysis (1): Max expenditure also by infant of 0 age =678118, 15=111747 17=174777

# Goal/Expectation (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

diagg <- as.factor(APRDRG)

summary(diagg)

**#OUTPUT:**



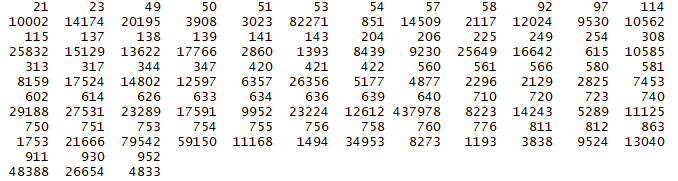
which.max(summary(diagg))

**#OUTPUT:**



tapply(TOTCHG,diagg,sum)

**#OUTPUT:**



which.max(tapply(TOTCHG,diagg,sum))

**#OUTPUT:**



max(tapply(TOTCHG,diagg,sum))

**#OUTPUT:**



#Analysis (2): From the results we can see that the category 640 has the maximum entries of hospitalization and also has the highest total hospitalization cost (437978).

# Goal/Expectation (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

# h0: The race of the patient is related to the hospitalization costs.

# ha: No relation

rc <- as.factor(RACE)

summary(rc)

**#OUTPUT:**



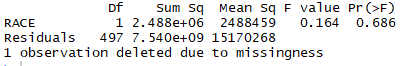
# Now to omit na values from data set

hospna <- na.omit(hosp)

modelannova <- aov(TOTCHG~RACE)

summary(modelannova)

**#OUTPUT:**



# Analysis (3): p-value comes out to be very high 68% this means we can take risk and reject the null hypothesis this means there is no relation between the race of patient and the hospital cost.

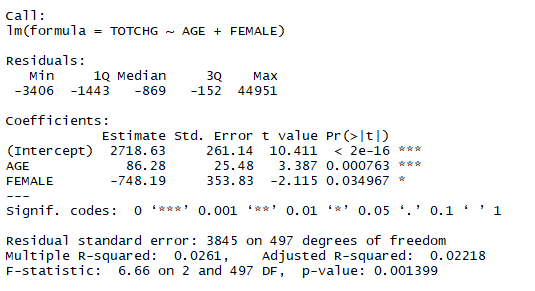
# Goal/Expectation (4) : To properly utilize the costs, the agency has to analyse

the severity of the hospital costs by age and gender for proper allocation of resources

model1 <- lm(TOTCHG~AGE+FEMALE)

summary(model1)

**#OUTPUT:**

****

# Analysis (4): p-value for age is very less this means it is an important factor in the hospital costs as seen by the significance levels and p-values. gender has also less p value means it is also having the impact on cost and same with intercept

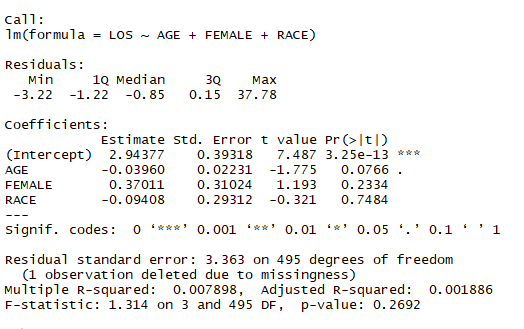
# Goal/Expectation (5): Since the length of stay is the crucial factor for inpatients,

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

model2 <- lm(LOS~AGE+FEMALE+RACE)

summary(model2)

**#OUTPUT:**



# Analysis (5): Except for the intercept.

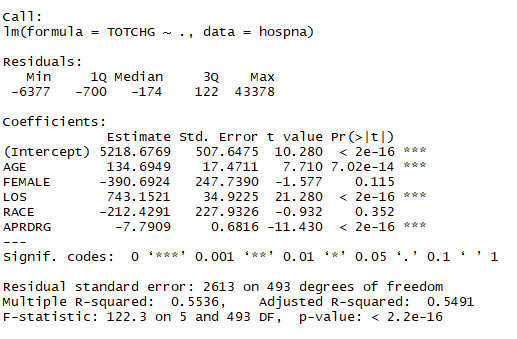
# The very high p-value signifies that there is no linear relationship between the given variables. That is, with just the age, gender, and race, it is not possible to predict the los of a patient

# Goal/Expectation (6) : To perform a complete analysis, the agency wants to find the variable that mainly affects the hospital costs

model3 <- lm(TOTCHG~ .,data=hospna)

summary(model3)

**#OUTPUT:**



# Analysis (6): APRDRG also affect

# We can see that age and length of stay affect the total hospital cost.

========================== END ===============================