

**Project 1:**

**Healthcare Cost Analysis**



**Contents**

**1. Objectives of Case Study …………………… 3**

**2. Data Description …………………… 3**

**3.Exploratory Data …………………… 3**

**4.To record the patient statistics, the agency wants to find the age category of people who frequent the hospital and has the maximum expenditure. …………………… 5**

**5. In order to determine the 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. …………………… 7**

**6. 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. …………………… 8**

**7.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. …………………… 9**

**8.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. …………………… 9**

**9.To perform a complete analysis, the agency wants to find the variable that mainly affects hospital costs. …………………… 11**

**Objective Of Case Study :**

In this project we are going to analyse the hospital data and we will give result to the customer , by using histogram and with other R function we will give results for frequent visitors , Which Diagnosis yielding more cost , which race of people spent , length of stay based on age, gender, and race. variable effects the hospital costs .

**Data Description :**

Healthcare : Name of The Data frame

AGE : Age of the patient discharged

FEMALE : 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

**Exploratory Data :**

First we need to import the dataset .

To Import Data we use syntact

**Healthcare=read.csv(choose.files(),header=T)**

**1) Structure Of Data :**

Str is a compact way to display the structure of an R object

**str(Healthcare)**

$ AGE : num 17 17 17 17 17 17 17 16 16 17 ...

$ FEMALE: num 1 0 1 1 1 0 1 1 1 1 ...

$ LOS : num 2 2 7 1 1 0 4 2 1 2 ...

$ RACE : num 1 1 1 1 1 1 1 1 1 1 ...

$ TOTCHG: num 2660 1689 20060 736 1194 ...

$ APRDRG: num 560 753 930 758 754 347 754 754 753 758 ...

**2) Dimensions of Data :**

to return Dimension of Dataset

**dim(Healthcare)**

> dim(Healthcare)

[1] 500 6

it says the there are 500 observation 8 columns

**3) Name :**

it will give names of an object.

**names(Healthcare)**

"AGE" "FEMALE" "LOS" "RACE" "TOTCHG" "APRDRG"

**4) head :**

Returns the first 6 rows of data frame

**head(Healthcare)**

1 17 1 2 1 2660 560

2 17 0 2 1 1689 753

3 17 1 7 1 20060 930

4 17 1 1 1 736 758

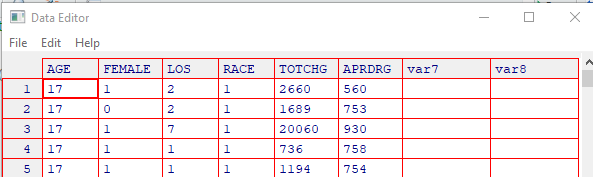
5 17 1 1 1 1194 754

6 17 0 0 1 3305 347

**5) fix :**

edited on the fly, For modifications of data set

**fix(Healthcare)**

****

**6) Listing Of Single variable values**

**Healthcare$AGE**

This lists the age values of age variable

> Healthcare$AGE

[1] 17 17 17 17 17 17 17 16 16 17 17 15 15 15 15 15 14 14 15 14 13 17 12 15 13 12 11 15 11 10 11 7

**7) Summary Of Data**

It is multipurpose function where it shows the number of objects includes the datasets variables , the variables in summary of the data are mean,max,min and medians

**> summary(Healthcare)**

AGE FEMALE LOS RACE TOTCHG APRDRG

Min. : 0.000 Min. :0.000 Min. : 0.000 Min. :1.000 Min. : 532 Min. : 21.0

1st Qu.: 0.000 1st Qu.:0.000 1st Qu.: 2.000 1st Qu.:1.000 1st Qu.: 1216 1st Qu.:640.0

Median : 0.000 Median :1.000 Median : 2.000 Median :1.000 Median : 1536 Median :640.0

Mean : 5.086 Mean :0.512 Mean : 2.828 Mean :1.078 Mean : 2774 Mean :616.4

3rd Qu.:13.000 3rd Qu.:1.000 3rd Qu.: 3.000 3rd Qu.:1.000 3rd Qu.: 2530 3rd Qu.:751.0

Max. :17.000 Max. :1.000 Max. :41.000 Max. :6.000 Max. :48388 Max. :952.0 NA's :1

Summary For Age VAriable

> summary(as.factor(Healthcare$AGE))

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

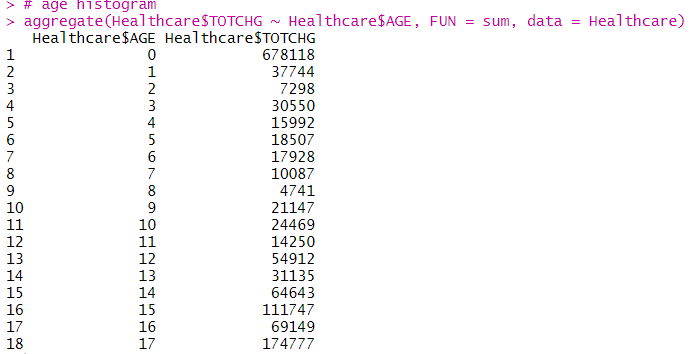
307 10 1 3 2 2 2 3 2 2 4 8 15 18 25 29 29 38

**Analysis**

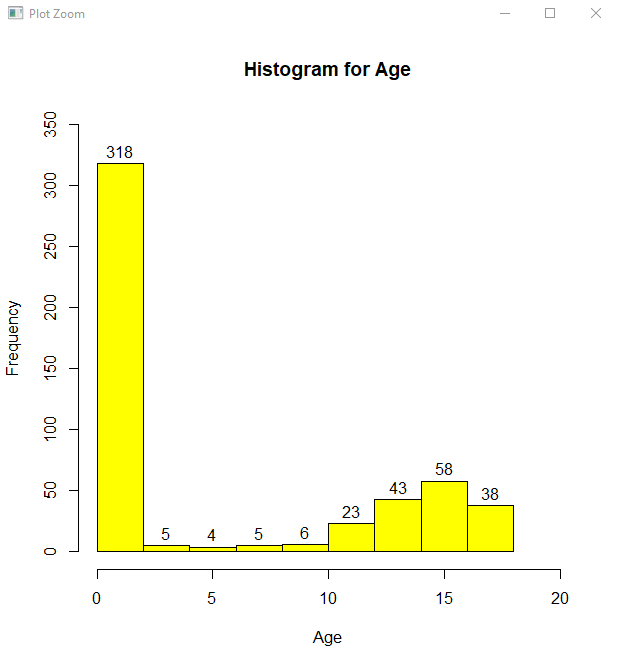
**1 ) To record the patient statistics, the agency wants to find the age category of people who frequent the hospital and has the maximum expenditure.**

**aggregate(Healthcare$TOTCHG ~ Healthcare$AGE, FUN = sum, data = Healthcare)**

aggregates the inputted data. frame d.f by applying a function specified by the FUN parameter to each column of sub-data. frames defined by the by input parameter

****

**hist(Healthcare$AGE, main = "Histogram for Age", labels = T,xlim=c(0,20),ylim=c(0,350), breaks = 10, xlab = "Age", col = "yellow", )**

****

**Based On OutPut we can see Age wise hospital visits and expenses**

max(summary(as.factor(Healthcare$AGE)))

> summary(as.factor(Healthcare$AGE))

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

307 10 1 3 2 2 2 3 2 2 4 8 15 18 25 29 29 38

Maximum hospital visits 0-1 age group 307

**max(aggregate(Healthcare$TOTCHG ~ Healthcare$AGE, FUN = sum, data = Healthcare))**

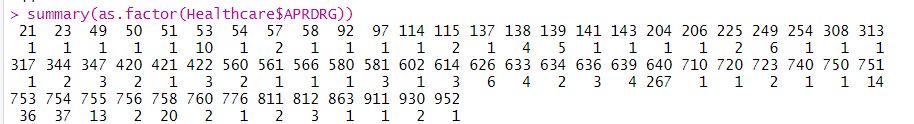
> max(aggregate(Healthcare$TOTCHG ~ Healthcare$AGE, FUN = sum, data = Healthcare))

[1] 678118

Maximum expenditure 0-1 age group 678118

**2. In order to determine the 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.**

**summary(as.factor(Healthcare$APRDRG))**

****

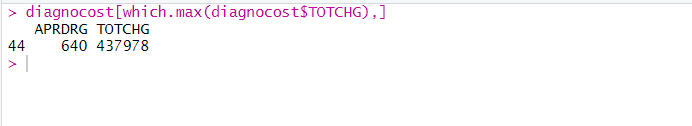
**which.max(summary(as.factor(Healthcare$APRDRG)))**



**diagnocost <- aggregate(TOTCHG ~ APRDRG, FUN = sum, data = Healthcare)**



**diagnocost[which.max(diagnocost$TOTCHG),]**

****

Diagnosis group which is having more Hospitalization is 640 and For That expenditure is 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.**

**We can achieve Race wise hospitalization cost by ANOVA**

**head(Healthcare)**

**Healthcare<-na.omit(Healthcare)**

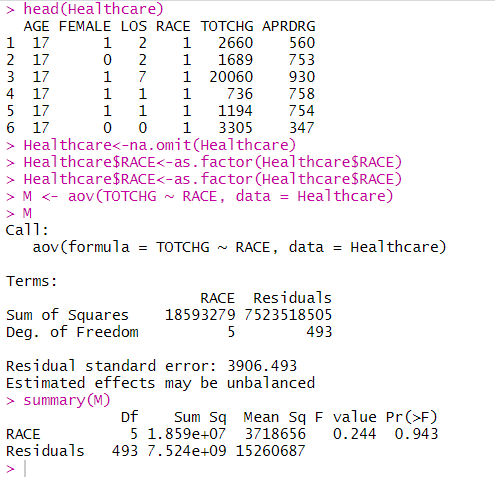
**Healthcare$RACE<-as.factor(Healthcare$RACE)**

**Healthcare$RACE<-as.factor(Healthcare$RACE)**

**M <- aov(TOTCHG ~ RACE, data = Healthcare)**

**M**

**summary(M)**

****

**Result InterPretation**

**p value comes out to be very high 93% this means we can take risk and reject the null hypothesis . This means there is no relation between the race of patients 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.**

This can be done by Linear Regression

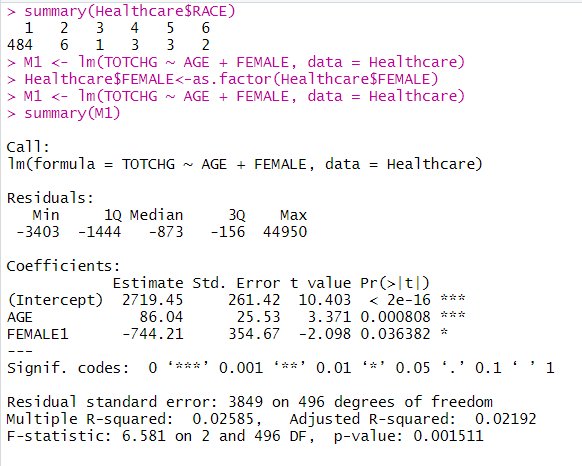
**summary(Healthcare$RACE)**

**M1 <- lm(TOTCHG ~ AGE + FEMALE, data = Healthcare)**

**Healthcare$FEMALE<-as.factor(Healthcare$FEMALE)**

**M1 <- lm(TOTCHG ~ AGE + FEMALE, data = Healthcare)**

**summary(M1)**

****

**Result Interpretation**

**p value for age is very less this means it is a 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**

**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.**

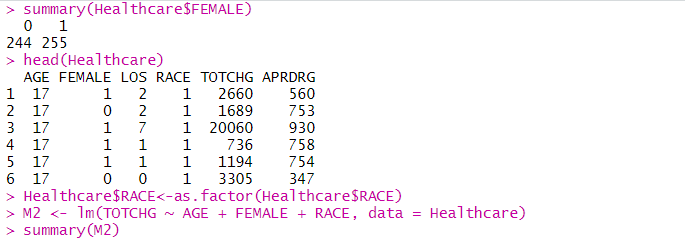
**summary(Healthcare$FEMALE)**

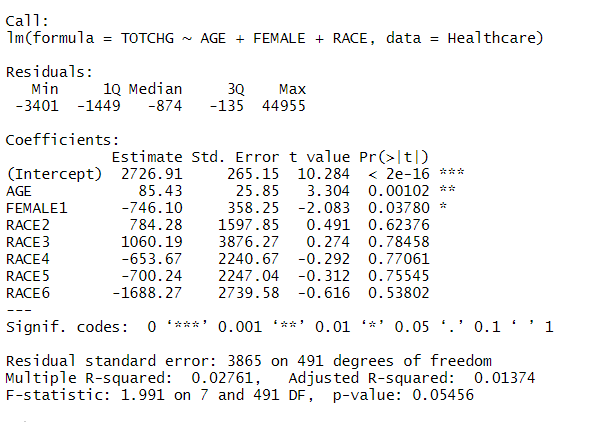
**head(Healthcare)**

**Healthcare$RACE<-as.factor(Healthcare$RACE)**

**M2 <- lm(TOTCHG ~ AGE + FEMALE + RACE, data = Healthcare)**

**summary(M2)**

****

****

**Result Interpretation**

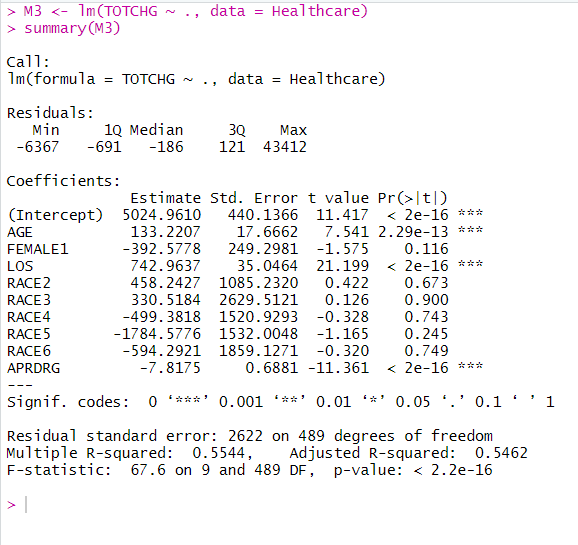
**The very high p-value signifies that there is no linear relationship between the given variables.**

**That is, with just the age, gender, and race, it is not possible to predict the loss of a patient**

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

**M3 <- lm(TOTCHG ~ ., data = Healthcare)**

**summary(M3)**

****

**Result Interpretation**

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

**------------------END-----------------**