My code:

library(dplyr)

setwd("/Users/rushikeshkhankar/Desktop/R")

getwd()

#Loading Dataset

data <- read.csv("~/Desktop/R/Project/Projects for Submission/Healthcare/Healthcare/HospitalCosts.csv")

data

#Descriptive Analysis

View(data)

str(data)

summary(data)

#Data Cleaning

#Handling Missing Values

data <- na.omit(data)

##Q1. To record the patient statistics, the agency wants to find the age category

#of people who frequent the hospital and has the maximum expenditure.

hist(data$AGE, main = "Frequency of Patients", breaks = 8, col = "darkorange")

attach(data)

AGE = as.factor(AGE)

summary(AGE)

max\_value = max(aggregate(TOTCHG~AGE, FUN = sum,data = data))

max\_value

#Q.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.

hist(data$APRDRG, main = "Diagnosis Related Group", breaks = 8, col = "darkorange")

APRDRG=as.factor(APRDRG)

summary(APRDRG)

df=aggregate(TOTCHG~APRDRG,FUN = sum, data = data)

df[which.max(df$TOTCHG),]

#Q.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.

aovt <-aov(TOTCHG~factor(RACE), data = data)

summary(aovt)

#Q.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.

model1=lm(TOTCHG~AGE+factor(FEMALE),data = data)

summary(model1)

#Q.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+factor(RACE), data = data)

summary(model2)

#Q.6 To perform a complete analysis, the agency wants to find the variable that

#mainly affects the hospital costs.

model3 = lm(TOTCHG~AGE+FEMALE+LOS+RACE+APRDRG,data = data)

summary(model3)