# install required package to read xls file

# install.packages("readxl")

# load readxl library

library("readxl")

# dataname

setwd('F:\\my\_work\\simplilearn\\data-science\\data-science-with-r\\Hospital-Costs')

filename <- "hospitalcosts.xlsx"

# read dataset

dataset <- read\_excel(filename)

newdataset <- dataset

# Summary of dataset

summary(dataset)

nrows <- nrow(dataset)

# Get datatype of columns

sapply(dataset, class)

hist(dataset$AGE)

# Convert independent columns to factor type

cols <- c("AGE", "FEMALE", "APRDRG", "RACE")

dataset[cols] <- lapply(dataset[cols], factor)

sapply(dataset, class)

# dep\_cols <- c("LOS", "TOTCHG")

# dataset[dep\_cols] <- lapply(dataset[dep\_cols], as.integer)

# sapply(dataset, class)

# count missing values

sum(is.na(dataset))

dataset <- na.omit(dataset)

# Verify na is removed from all datasets, sum of na should be 0

sum(is.na(dataset))

nrow(dataset)

# 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.

# Get the age of frequently visited patient

summary(dataset$AGE)

# Get maximum expenditure based on age

tapply(dataset$TOTCHG, dataset$AGE, sum)

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

max(tapply(dataset$TOTCHG, dataset$AGE, sum))

# Age category of 0 seems to visit frequently with maximum expenditure.676962

# 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.

summary(dataset$APRDRG)

which.max(summary(dataset$APRDRG))

tapply(dataset$TOTCHG, dataset$APRDRG, sum)

which.max(tapply(dataset$TOTCHG, dataset$APRDRG, sum))

max(tapply(dataset$TOTCHG, dataset$APRDRG, sum))

# From the result we can see that the category 644 has maximum entries of

# hospitalization with expenditure 436822.

# 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.

summary(dataset$RACE)

race\_anova <- aov(dataset$TOTCHG~dataset$RACE)

summary(race\_anova)

# Here, we can see the value of P is very high, that mean there is not any

# relationship between RACE of the patient and hospitalization costs.

# 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.

model1 <- lm(formula = TOTCHG~AGE+FEMALE, data = newdataset)

summary(model1)

# Here p-value is very less, so both variable have impact on hospital price

# 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.

model1 <- lm(formula = LOS~AGE+FEMALE+RACE, data = newdataset)

summary(model1)

# Except for the intercept, p-value is very high that mean, age, female and

# race can not be factor for prediction of length of stay.

# 6. To perform a complete analysis,

# the agency wants to find the variable that mainly affects hospital costs.

model1 <- lm(formula = TOTCHG~., data = newdataset)

summary(model1)

# Here, we can see the large value of p for AGE, LOS and APRDRG, that mean,

# they are responsible for hospital costs.