```
# loading libraries
library(readxl)   # for reading excel files
library(dplyr)    # for data manipulation
library(corrplot) # for correlation-plot
library(caTools) # for splitting dataset
library(ggplot2) # for data visualization
library(MLmetrics) # for machine learning metrics
library(caret)  # for feature importance
library(lattice)

# removing objects in environment
rm(list = ls())

# loading dataset
mydf <- read_xlsx('hospitalcosts.xlsx')
View(mydf)

# EDA
dim(mydf)
summary(mydf)
str(mydf)

# checking and handling NA
sapply(mydf, function(x) sum(is.na(x))) # count of NA for each feature
which(is.na(mydf$RACE)) # row number for NA in RACE
mydf <- na.omit(mydf)

# creating GENDER column
mydf$GENDER <- ifelse(mydf$FEMALE == 1, 'Female', 'Male')
```

```
mydf$GENDER <- as.factor(mydf$GENDER)

mydf$GENDER


# creating age class interval column i.e 0-5, 5-10, 10-15, 15-20

range(mydf$AGE)

bins <- seq(0,20,by=5)

bins

mydf$AGE_GROUP <- cut(mydf$AGE, bins, include.lowest = T)

summary(mydf)

View(mydf)


# question-1

group_age <- group_by(mydf,AGE_GROUP)

age_res <- summarise(group_age, Total_Expenditure = n())

age_res[which.max(age_res$Total_Expenditure),]

ggplot(age_res, aes(x = AGE_GROUP, y = Total_Expenditure)) + geom_bar(stat =
'identity', color = 'blue', fill = 'white')



# question-2

group_diagonsis <- aggregate(mydf$TOTCHG,

                        by = list(mydf$APRDRG),

                        FUN = sum)

names(group_diagonsis)[1] <- 'APRDRG'

names(group_diagonsis)[2] <- 'Total_Expenditure'

group_diagonsis[which.max(group_diagonsis$Total_Expenditure),]


# question-3

group_race <- aggregate(mydf$TOTCHG,
            by = list(mydf$RACE),
            FUN = sum)
names(group_race)[1] <- 'RACE'
names(group_race)[2] <- 'Total_Expenditure'
group_race
anova_model <- aov(mydf$TOTCHG~mydf$RACE, data = mydf)
summary(anova_model)
barplot(group_race$Total_Expenditure,
    border = 'dark blue',
    main = 'Race - Total_Expenditure',
    xlab = 'Race',
    ylab= 'Total_Expenditure',
    names.arg = c(1:6))

# question-4
group_age_gender <- aggregate(mydf$TOTCHG,
            by = list(mydf$AGE, mydf$GENDER),
            FUN = sum)
names(group_age_gender)[1] <- 'AGE'
names(group_age_gender)[2] <- 'GENDER'
names(group_age_gender)[3] <- 'Total_Expenditure'
group_age_gender
female <- group_age_gender[group_age_gender$GENDER == 'Female', ]
female
barplot(female$Total_Expenditure,
    col = heat.colors(12),
    log = "y",
    main = 'Female Chart',
    names.arg = female$AGE,
        xlab = 'AGE',
        ylab = 'Total_Expenditure')
male <- group_age_gender[group_age_gender$GENDER == 'Male',]
male
barplot(male$Total_Expenditure,
        border = 'dark blue',
        main = 'Male Chart',
        col = rainbow(20),
        names.arg = male$AGE,
        xlab = 'AGE',
        ylab = 'Total_Expenditure')

# question - 5
finaldf <- mydf[,c(1:4)]  # selecting features required to perform multiple regression
View(finaldf)
categcol <- finaldf[,c(1,2,4)] # selecting categorical features to create dummy variables


fact<- data.frame(sapply(categcol, function(x) factor(x)))
str(fact)
names(fact)
# creating dummy variables for factor attributes
dummies<- data.frame(sapply(fact,
                        function(x) data.frame(model.matrix(~x,data = fact))[,-1]))
View(dummies)
names(dummies)

finaldf <- cbind(finaldf[,3], dummies)
View(finaldf)
```

```
regressor <- lm(LOS~.,data = finaldf)
summary(regressor)


# question-6
imp <- varImp(regressor)
barplot(imp$Overall,names.arg = row.names(imp))
cor <- cor(mydf[,1:6])
corrplot(cor,
        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))
```