```
library(corrplot) # for corrplot
library(caTools) # for splitting dataset
library(ggplot2)
library(MLmetrics)
library(readxl)

dataset = read_excel("hospitalcosts.xlsx")
View(dataset)

##Removing NA Values
numberOfNA = sum(is.na(dataset))
numberOfNA

if(numberOfNA > 0){
  dataset = dataset[complete.cases(dataset),]
}
sum(is.na(dataset))

##Que 1
table(dataset$AGE)
age_totchg <- aggregate(x= dataset$TOTCHG, by = list(dataset$AGE), FUN=sum)
colnames(age_totchg) = c('Age', 'Total_Charge')
age_totchg
age_totchg[which.max(age_totchg$Total_Charge),] #People with Age = 0 have Max
Expenditure

##Que 2

age_aprdrg <- aggregate(x=dataset$APRDRG, by= list(dataset$AGE), FUN=sum)
colnames(age_aprdrg) = c('Age', 'Total_Aprdrg')
age_aprdrg
age_aprdrg[which.max(age_aprdrg$Total_Aprdrg),]

##Finding Correlation

dim(dataset)
corr = cor(dataset[, 1:6])
View(corr)

###Hospitalization costs no where seems to be coorelated to each other

aggregate(x=dataset$TOTCHG, by=list(dataset$AGE, dataset$RACE), FUN=sum)

##Age, Gender and TOTCHG

age_gender <- aggregate(x=dataset$TOTCHG, by=list(dataset$AGE, dataset$FEMALE),
FUN=mean)
colnames(age_gender) <- c('Age','Gender','Mean')
age_gender
class(age_gender)

###AGE GENDER and RACE

unique(dataset$AGE)
```

```
unique(dataset$FEMALE)
unique(dataset$RACE)


#Visualizing Data

par(mfrow=c(1,5))

boxplot(dataset$LOS ~  , data=dataset)$out

boxplot(dataset$LOS ~ dataset$AGE , data=dataset)$out

boxplot(dataset$LOS ~ dataset$FEMALE , data=dataset)$out

boxplot(dataset$LOS ~ dataset$RACE , data=dataset)$out

boxplot(dataset$LOS ~ dataset$TOTCHG, data=dataset)

boxplot(dataset$LOS ~ dataset$APRDRG, data=dataset)

##Model without dummy variables

model_1 = lm(formula= LOS ~ AGE+FEMALE+RACE, data = dataset)
summary(model_1)

##Model with dummy variables

##Creating Dummy Variables

age_fact = as.factor(dataset$AGE)
dummy_age = data.frame(model.matrix(~age_fact))[,-1]

gender_fact = as.factor(dataset$FEMALE)
dummy_gender = data.frame(model.matrix(~gender_fact))[,-1]

race_fact = as.factor(dataset$RACE)
dummy_race = data.frame(model.matrix(~race_fact))[,-1]

##Dropping Unwanted Columns


dataset_demo = subset(dataset , select = -c(AGE,RACE,FEMALE,APRDRG, TOTCHG))
View(dataset_demo)
dataset_demo = cbind(dataset_demo,dummy_gender, dummy_race, dummy_age)
dim(dataset_demo)

View(dataset_demo)

model = lm(formula = LOS ~  ., data=dataset_demo)
summary(model)

##As the correlation between the LOS and AGE,GENDER, RACE is very low the model
will not br able to predict length of Stay with these Factors
View(dataset)
```

```
model = lm(formula = TOTCHG ~ LOS,APRDRG, data = dataset)

summary(model)

#A better model can be built using TOTCHG against LOS and APRDRG
```