Source Code:

```
hosp <- read.csv("hospital.csv")
head(hosp)
summary(hosp)
attach(hosp)

#1
#Histogram to represent the age group that frequently visit the hospital.
hist(AGE, col = "Blue")
# The category of infants(0) has the highest vist to the hospital.
#To see the value of category of infants.
high<-as.factor(AGE)
summary(high)
#there are 307 cases in the category 0. which means infants have a highest frequency to visit the hospital.
#age category of 0 seems to be  frequently using the hospital.
tapply(TOTCHG,AGE,sum)
which.max(tapply(TOTCHG,AGE,sum))
#max expenditure also by infant of 0 age =678118, 15=111747 17=174777

#2
Expnd<-as.factor(APRDRG)
summary(Expnd)
which.max(summary(Expnd))
tapply(TOTCHG,Expnd,sum)
which.max(tapply(TOTCHG,Expnd,sum))
```

```
max(tapply(TOTCHG,Expnd,sum))

#From the results we can see that the category 640 has the maximum entries of hospitalization

#and also has the highest total hospitalization cost (437978).


#3

#To find out the relationship between the race of the patient and the hospitalization costs. We perform a ANOVA test based on the following assumptions.

#Ho: there is a relationsip between the the race and the cost.  H1:No relation

linear<-as.factor(RACE)

summary(linear)

hospna<-na.omit(hosp)

modelannova<-aov(TOTCHG~RACE)

summary(modelannova)

#Pvalue comes out to be very high 68% this means we can take risk and reject the null hypothesis

#This means  there is no relation between the race of patient and the hospital cost.


#4

#To analyse the severity of hospital cost by age and gender , we use the  Linear Regression analysis.

linear1<-lm(TOTCHG~AGE+FEMALE)

summary(linear1)

#Pvalue 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

#To see if we can predict the Length of stay based on the age, gender and race we perform an Linear Regression between them.

linear2<-lm(LOS~AGE+FEMALE+RACE)

summary(linear2)

#The higher 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 LOS of a patient.


#6

#To perform a complete analysis of the main factors that affect the hospital cost another Linear regression analysis is performed.

linear3<-lm(TOTCHG~ .,data=hospna)

summary(linear3)

#We can see that age and length of stay (LOS) and the APRDRG are the major factors  affecting the total hospital cost.
```