```r
# Healthcare Cost Analysis:

# importing data

library(dplyr)
library(readxl)
library(ggplot2)
setwd("E:/simplilearn/Data Science with R Programming/Assessment/Project 7")
getwd()
healthCost <- read_xlsx("hospitalcosts.xlsx")
View(healthCost)
str(healthCost)

# The age category of people who frequently visit the hospital and has maximum expenditure

summary(healthCost)
hist(healthCost$AGE, col = 'dark green')

ggplot(healthCost) +
  geom_histogram(aes(x = AGE), colour = 'black', fill = 'blue', binwidth = 2) +
  scale_x_continuous(name = 'Age of patients', breaks = seq(0, 20, 2)) +
  scale_y_continuous(name = 'Counts of Age', breaks = seq(0, 350, 50)) +
  ggtitle('Histogram of age')

summary(as.factor(healthCost$AGE))
aggregate(TOTCHG ~ AGE, FUN = sum, data = healthCost)
max(aggregate(TOTCHG ~ AGE, FUN = sum, data = healthCost))

# Insight 1: Based on above analysis, we can see that age wise hospital visit and expenditure
# Insight 2: 0-1 yrs age group has the maximum visits of 307
# Insight 3: 0-1 yrs age group has the maximum expenditure of 678118

# The diagnosis-related group that has maximum hospitalization and expenditure

which.max(summary(as.factor(healthCost$APRDRG)))
diagnosisCost <- aggregate(TOTCHG ~ APRDRG, FUN = sum, data = healthCost)
diagnosisCost
max(diagnosisCost)
diagnosisCost[which.max(diagnosisCost$TOTCHG),]

# Insight 4: Based on the output, we can see that the expenditure is based on the diagnosis and treatments
# Insight 5: 640 diagnosis-related group has maximum hospitalization and expenditure

# Analyze if the race of the patient is related to the hospitalization costs

summary(as.factor(healthCost$RACE))
head(healthCost)
healthCost <- na.omit(healthCost)
summary(as.factor(healthCost$RACE))

healthCost$RACE <- as.factor(healthCost$RACE)
anova <- aov(TOTCHG ~ RACE, data = healthCost)
anova
summary(anova)

# Insight 6: The data has 484 patient of race 1 out of the 500 entries where is very skewed
# Insight 7: The residual value is very high specifying that there is no relation between the race of patient and hospital costs

# Analyze the severity of the hospital costs by age ang gender for the proper allocation of resources

healthCost$FEMALE <- as.factor(healthCost$FEMALE)
costlm <- lm(TOTCHG ~ AGE + FEMALE, data = healthCost)
summary(costlm)
summary(healthCost$FEMALE)
str(healthCost)

# Insight 8: Age is a very important factor that affects the hospital costs as seen by the p-value
# Insight 9: Gender also has an impact to the hospital costs
# Insight 10: Based the negative coeffitient we can conclude that females incur lesser costs than males.

# Factors affect hospital costs

costlm1 <- lm(TOTCHG ~ AGE + FEMALE + RACE, data = healthCost)
summary(costlm1)

# Insight 11: The p-value high which signifies that there is no linear relationship between the given variables
# Insight 12: We can't predict the length of stay based on age, gender, and race

costlm2 <- lm(TOTCHG ~ ., data = healthCost)

# Insight 13: Based on the output we can see that the age and length of stay affects the hospital costs.
# INsight 14: With an increase of one day stay, the hospital cost will increase by 742
```