 # Health Care Cost w/ Linear Regression

 # Understanding the data

 # Age:age of primary beneficiary

 # Sex: gender, [female, male]

 # BMI: Body mass index, providing an understanding of body, weights that are relatively

 # high or low relative to height, objective index of body weight (kg / m ^ 2) using the

 # ratio of height to weight, ideally 18.5 to 24.9

 # Children: number of children

 # Smoker: smoking, [yes, no]

 # Region: the beneficiary's residential area in the US, [northeast, southeast, southwest, northwest]

 # Charges: Individual medical costs billed by health insurance, $ #predicted value

 # load required libraries

```
library(ggplot2)
library(dplyr)
library(gridExtra)
library(psych)
library(corrplot)


# Load the dataset
d<-read.csv("D:/Rprograms/insurance.csv")
#print(d)


# Print head
print(head(d))


# Print tail
print(tail(d))


# To View the contents in the dataet
View(d)
```

```
# To print column names
print(colnames(d))


# Dimention of data
print(dim(d))


# Print Statistical summary
describe(d)


# Summary of the dataset
print(summary(d))


# Internal structure of R object
print(str(d))


# Display columns  and display some portions of the data
print(glimpse(d))
```

```
# To print unique columns
print(unique(d$age))
print(unique(d$bmi))
print(unique(d$charges))


# statistical values
print(is.na(d))
print(ncol(d))
print(nrow(d))
print(max(d$charges))
print(min(d$charges))
print(sort(d$charges))
print(which.max(d$charges))# Return the index of the first maximum value
print(which.min(d$charges))# Return the index of the first minimum value
print(mean(d$charges))
print(mean(d$charges,trim=0.10))
print(var(d$charges))
print(median(d$charges))
print(mad(d$charges))# mean absolute division
print(sd(d$charges))
print(range(d$charges))
```

```
print(quantile(d$charges))
print(IQR(d$charges))
print(t.test(d$charges))


    # Data visualisation

# Histogram of Numerical data
hist(d$age,breaks=15,col="green")


hist(d$bmi,breaks=15,col="cyan")
# BMI values are normally disributed.


hist(d$charges,breaks=15,col="pink")
# As we expected, the figure shows right skewed distribution


# To see the distribution of data
table(d$region)
table(d$age)
table(d$sex)
table(d$smoker)
```

```
table(d$children)
table(d$bmi)
table(d$charges)


# Barplot of Categorical data

barplot(table(d$children),col="brown1")
# majority of them having no children.

barplot(table(d$sex),col="blue1")
#  Here the graph shows,number of males are more than females.


barplot(table(d$smoker),col="cadetblue")
#  The number of  persons without smoke  are more than others.


barplot(table(d$region),col="aquamarine")
#  Shows,more number of persons are from southeast.


# Boxplot male and female with BMI values
sex_bmi<-ggplot(d,aes(x=sex,y=bmi))+geom_boxplot(fill="green3")
```

```
print(sex_bmi)
# BMI value is more for male than female


## Boxplot of male and female with charges
sex_chr<-
ggplot(d,aes(x=sex,y=charges))+geom_boxplot(fill="green3")
print(sex_chr)
# More charges are paid by male


# Boxplot of smoker and nonsmoker with BMI values
smok_bmi<-
ggplot(d,aes(x=smoker,y=bmi))+geom_boxplot(fill="brown")
print(smok_bmi)
# BMI value of smokers are more than without smokers


# Boxplot of age with region
age_reg<-
ggplot(d,aes(x=region,y=age))+geom_boxplot(fill="tomato")
print(age_reg)
# Here Maximum age from all regions are almost same
```

```
# geom_jitter with region and age
g1 <- ggplot(d, aes(region, age)) +
  geom_jitter(color = "gold", alpha = 0.5) +
  theme_light()+
  stat_summary(aes(x=region,y=age),fun=mean,color="blue")+
  stat_summary(aes(x=region,y=age),fun=median,color="red")
print(g1)
# Here all the region shows almost same mean and median value for age


#  geom_jitter with sex and charges
g2 <- ggplot(d, aes(sex, charges)) +
  geom_jitter(color = "green", alpha = 0.5) +
  theme_light()+
  stat_summary(aes(x=sex,y=charges),fun=mean,color="blue")+
  stat_summary(aes(x=sex,y=charges),fun=median,color="red")
print(g2)
#  Here, there is a small differnce in mean value of male and female w.r.t
#  charges
```

```
#  geom_jitter of sex and bmi
g3 <- ggplot(d, aes(sex, bmi)) +
  geom_jitter(color = "brown", alpha = 0.5) +
  theme_light()+
  stat_summary(aes(x=sex,y=bmi),fun=mean,color="blue")+
  stat_summary(aes(x=sex,y=bmi),fun=median,color="red")
print(g3)
# There is a small difference in mean and median value of male and female
# w.r.t bmi values

#  geom_jitter of age and charges
g4<-ggplot(d, aes(age, charges)) +
  geom_jitter(color = "violet", alpha = 0.5) +
  theme_light()+
  stat_summary(aes(x=age,y=charges),fun=mean,color="blue")+
  stat_summary(aes(x=age,y=charges),fun=median,color="red")
print(g4)
# Here,mean and median values are different w.r.t age and charges


#  geom_point with region and age
p1<-
ggplot(data=d)+geom_point(aes(x=region,y=age,color=region),alpha=.2)+
```

```
  theme_light()+
  stat_summary(aes(x=region,y=age),fun=mean,color="blue")+
  stat_summary(aes(x=region,y=age),fun=median,color="red")
print(p1)


#  geom_point with sex and charges
p2<-
ggplot(data=d)+geom_point(aes(x=sex,y=charges,color=region),alp
ha=.2)+
  theme_classic()+
  stat_summary(aes(x=sex,y=charges),fun=mean,color="blue")+
  stat_summary(aes(x=sex,y=charges),fun=median,color="red")
print(p2)


#  geom_point with sex and bmi
p3<-
ggplot(data=d)+geom_point(aes(x=sex,y=bmi,color=children),alpha
=.2)+
  theme_classic()+
  stat_summary(aes(x=sex,y=bmi),fun=mean,color="blue")+
  stat_summary(aes(x=sex,y=bmi),fun=median,color="red")
print(p3)
```

```
#  geom_point with age and charges
p4<-
ggplot(data=d)+geom_point(aes(x=age,y=charges,color=sex),alpha=
.2)+
  theme_classic()+
  stat_summary(aes(x=age,y=charges),fun=mean,color="blue")+
  stat_summary(aes(x=age,y=charges),fun=median,color="red")
print(p4)


#  Combination of geom_jitter and geom_point
print(grid.arrange(g1,p1,nrow=1))
print(grid.arrange(g2,p2,nrow=1))
print(grid.arrange(g3,p3,nrow=1))
print(grid.arrange(g4,p4,nrow=1))


#  To check the summary of charges
summary(d$charges)


#  To find the relation among variables.So we will use correlation
matrix
corr<-cor(d[c("age","bmi","children","charges")])
```

```
corrplot(corr,method="square",type="upper")


#  Scatterplot matrix
pairs(d[c("age","bmi","children","charges")],col="blue")


#  To add more information to scatterplot.
#  To enhance the plot,already load the package "psych"
pairs.panels(d[c("age","bmi","children","charges")])


#  To train a model on to the data
#  To fit the linear regression model to the data with R,we will use
#  the function lm()
model<-lm(charges~age+children+bmi+sex+region,data=d)
model<-lm(charges~.,data=d)


#  To build the model
model


#  To view more information about the model
```

summary(model)

 # In this analysis,applied linear regression.

 # As we can see, summary of a model showed us the significance of variable.