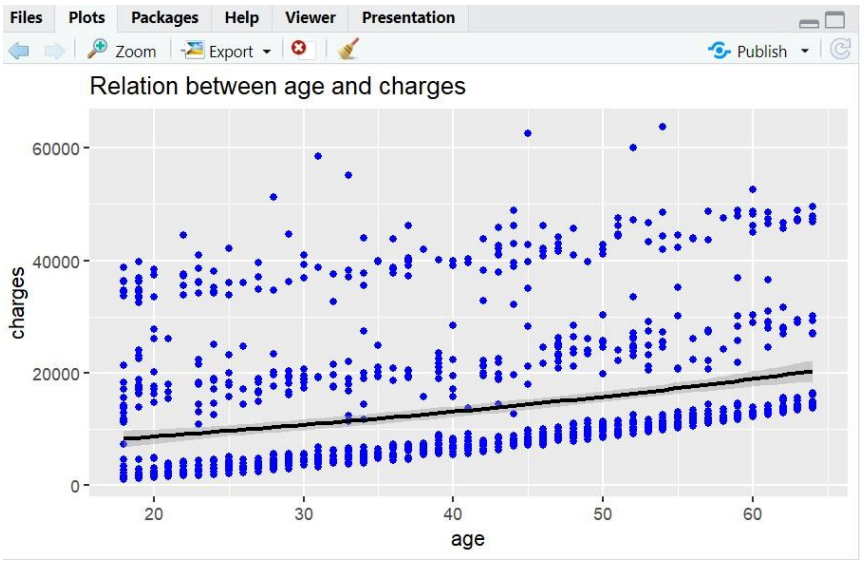
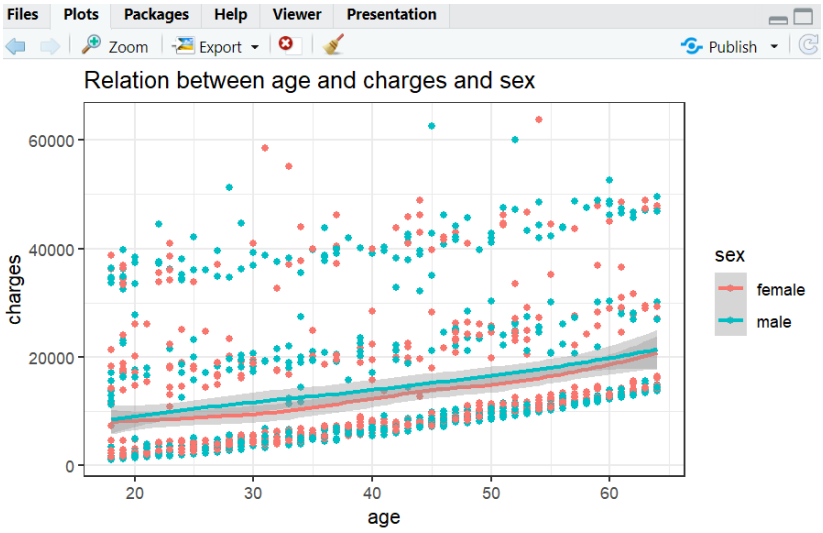
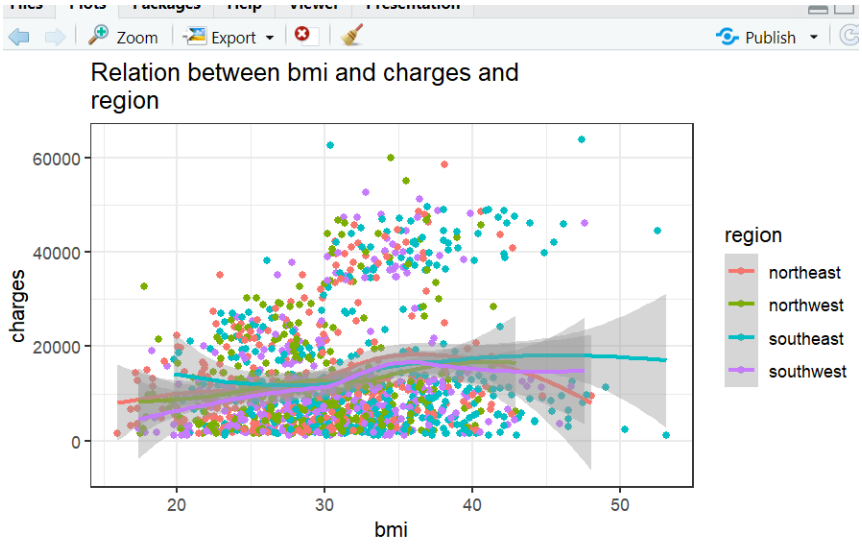
```
install.packages("tidyverse")
require("tidyverse")
mydata<-read.csv("C:/Users/91995/Favorites/Desktop/insurance.csv")
mydata
head(mydata)
summary(mydata)

library(ggplot2)
#Relation between age and charges

a <-ggplot(mydata, aes(age, charges))+
  geom_point(col="blue")+
  geom_smooth(col="black")+
  labs(x="age",y="charges",title="Relation between age and charges")
a
```



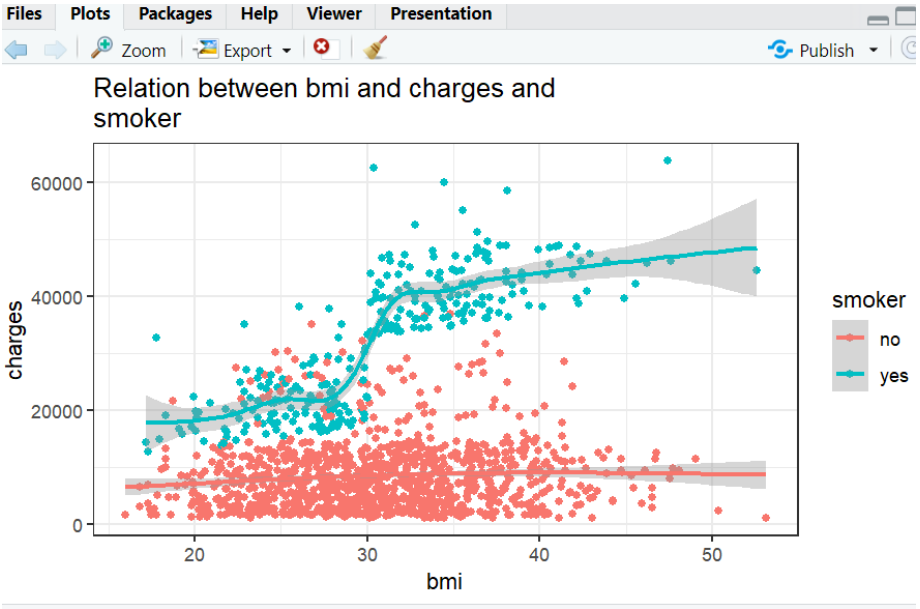

```
#Relation between age and charges and sex

b <- ggplot(mydata,aes(age,charges,color=sex))+
 geom_point()+
 geom_smooth()+
 labs(x="age",y="charges",title="Relation between age and charges and
sex")+
 theme_bw()
b
```



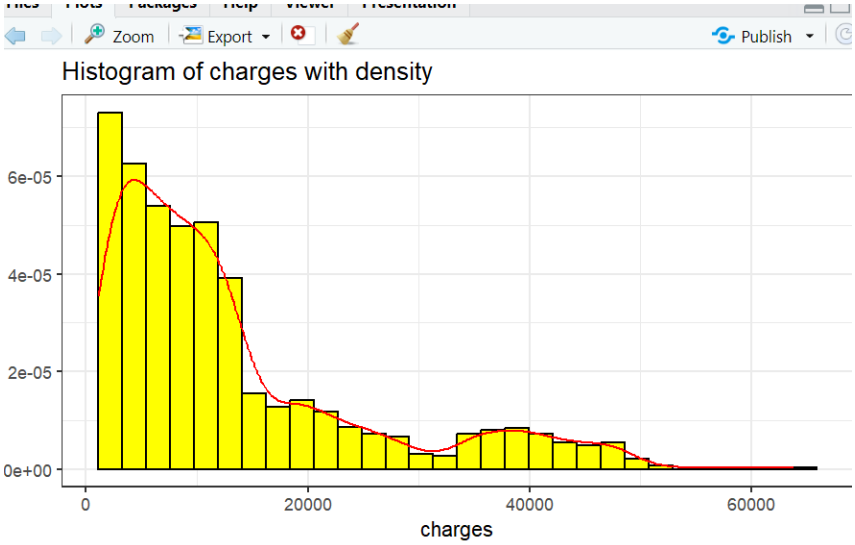

```
#Relation between age and charges and region

ca <- ggplot(mydata,aes(bmi,charges,color=region))+
 geom_point()+
 geom_smooth()+
 labs(x="bmi",y="charges",title="Relation between bmi and charges and
region")+
  theme_bw()
ca
```



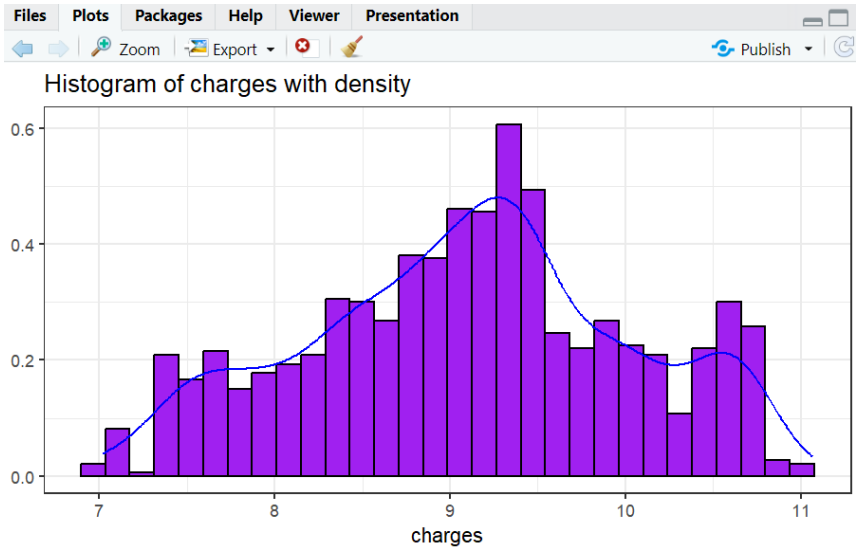

```
#Relation between age and charges and smoker
sc<-ggplot(mydata,aes(bmi,charges,color=smoker))+
  geom_point()+
  geom_smooth()+
  labs(x="bmi",y="charges",title="Relation between bmi and charges and
smoker")+
  theme_bw()
sc
```



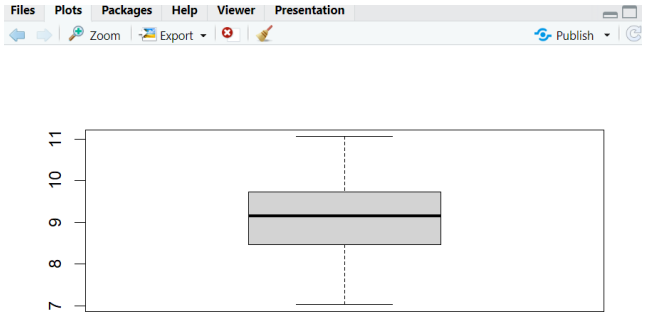

```
#Histogram of charges with density
dc <-ggplot(mydata,aes(charges))+
  geom_histogram(aes(y=after_stat(density)),bins=
30,col="black",fill="yellow")+
  geom_density(col="red")+
  labs(x="charges",y=NULL,title="Histogram of charges with density")+
  theme_bw()
dc
```




```
#Histogram of charges with density with log since its not normally
distributed#
cdl <-ggplot(mydata,aes(log(charges)))+
  geom_histogram(aes(y=stat(density)),bins=
30,col="black",fill="purple")+
  geom_density(col="blue")+
  labs(x="charges",y=NULL,title="Histogram of charges with density")+
   theme_bw()
cdl
```



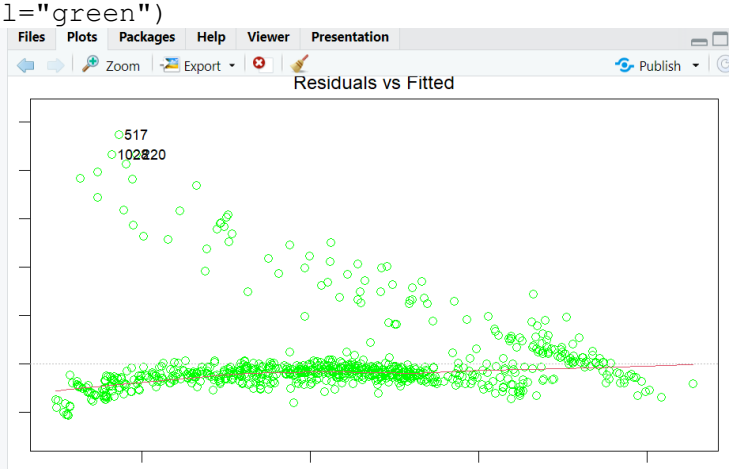

```
#Boxplot for log of charges
bp <- boxplot(log(mydata$charges))
bp
```




```
#hence we observe no outliers

set.seed(9654)
install.packages("caTools")
library(caTools)

split=sample.split(mydata , SplitRatio = 0.7 , group=NULL)
data=subset(mydata,split==TRUE)
data2<-subset(mydata,split==FALSE)

model
=lm(log(charges)~age+sex+bmi+children+smoker+region+age:smoker+bmi:smoker
+

I(log(age))+I(log(bmi))+age:children+age:region+smoker:children,data
          = data)
Model
```

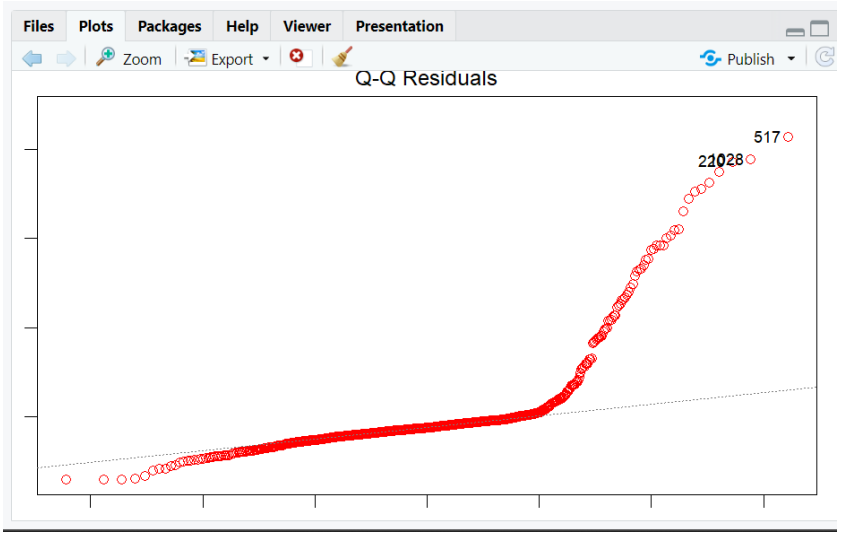
```
Call:
lm(formula = log(charges) ~ age + sex + bmi + children + smoker +
    region + age:smoker + bmi:smoker + I(log(age)) + I(log(bmi)) +
    age:children + age:region + smoker:children, data = data)

Coefficients:
        (Intercept)                 age              sexmale
           3.471818            0.030618           -0.107822
                bmi            children           smokeryes
          -0.042331            0.273847            1.291908
    regionnorthwest     regionsoutheast     regionsouthwest
          -0.190114           -0.323188           -0.408022
       I(log(age))         I(log(bmi))       age:smokeryes
           0.331727            1.232816           -0.032214
     bmi:smokeryes       age:children age:regionnorthwest
           0.053483           -0.003601            0.003560
age:regionsoutheast age:regionsouthwest  children:smokeryes
           0.005655            0.006962           -0.122567
```

```
mydata =mydata[c(-431,-398,-103),]
mydata =mydata[c(-354,-1132,-514),]
mydata =mydata[c(-521,-219,-4),]
mydata =mydata[c(-1020,-338,-1032),]
mydata =mydata[c(-1178,-525,-1094),]
mydata =mydata[c(-1317,-545,-304),]
mydata =mydata[c(-1199,-1147,-461),]
mydata =mydata[c(-421,-945,-1177),]

set.seed(9625)
split=sample.split(mydata,SplitRatio = 0.7)
data=subset(mydata,split==TRUE)
data2=subset(mydata,split==FALSE)

model=lm(log(charges)~age+sex+bmi+children+smoker+region+age:smoker+bmi:s
moker+

I(log(age))+I(log(bmi))+age:children+age:region+smoker:children,data =
data)
model
summary(model)

#1 residual vs fitted plot

par(mar=c(1,1,1,1))
plot(model,1,col="green")
```

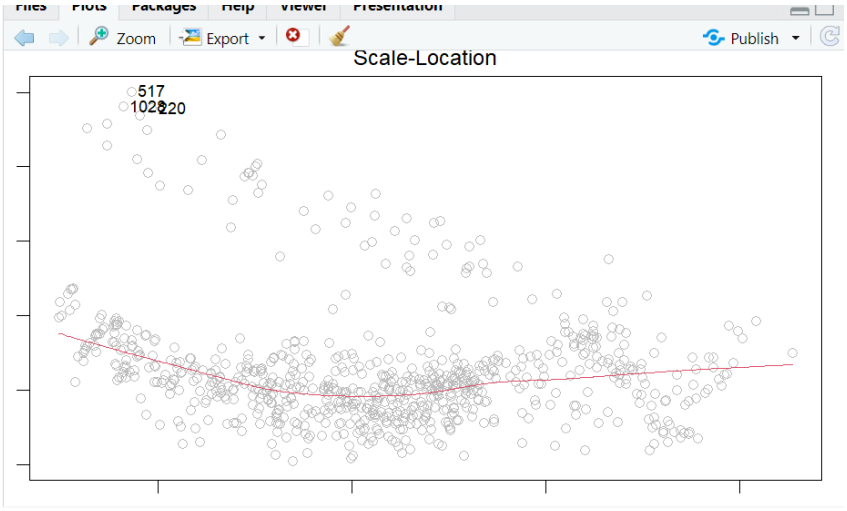



```
#2 Q-Q plot
```

```
plot(model,2,col="red")
```




```
#3 scale location graph
plot(model,3,col="grey")
```




```
anova(model)

newdata =data2[,-7]
newdata[c(1:6),]
pv=predict(model,newdata = newdata)
```

```
> anova(model)
Analysis of Variance Table

Response: log(charges)
                 Df  Sum Sq Mean Sq   F value    Pr(>F)
age               1 175.745 175.745 1170.7393 < 2.2e-16 ***
sex               1   0.556   0.556    3.7012  0.054761 .
bmi               1   8.995   8.995   59.9231 3.282e-14 ***
children          1  13.505  13.505   89.9643 < 2.2e-16 ***
smoker            1 308.077 308.077 2052.2717 < 2.2e-16 ***
region            3   3.088   1.029    6.8570  0.000147 ***
I(log(age))       1   3.533   3.533   23.5323 1.501e-06 ***
I(log(bmi))       1   0.089   0.089    0.5931  0.441488
age:smoker        1  24.208  24.208  161.2660 < 2.2e-16 ***
bmi:smoker        1  10.139  10.139   67.5434 9.351e-16 ***
age:children      1   2.475   2.475   16.4881 5.423e-05 ***
age:region        3   1.363   0.454    3.0258  0.028925 *
children:smoker   1   2.641   2.641   17.5928 3.072e-05 ***
Residuals       732 109.884   0.150
```

```
ff<-
ggplot(data2,aes(x=pv,y=log(data2$charges)))+geom_point()+geom_abline()
ff
```