> # Read the .csv file to a dataframe

> housing <- read.csv("USA\_Housing.csv")

>

> dim(housing)

[1] 5000 7

> names(housing)

[1] "Avg..Area.Income" "Avg..Area.House.Age"

[3] "Avg..Area.Number.of.Rooms" "Avg..Area.Number.of.Bedrooms"

[5] "Area.Population" "Price"

[7] "Address"

> #renaming the variables

> names(housing)<-c("income","age","rooms","beds","population","price","address")

> #checkng for any missing values

> sum(is.na(housing))

[1] 0

> #spiltting the dataframe to 2 dataframes, one to train and second to test

> h\_train<-housing[1:2500,1:6]

> h\_test<-housing[2501:5000,1:6]

> dim(h\_train)

[1] 2500 6

> names(h\_train)

[1] "income" "age" "rooms" "beds" "population"

[6] "price"

> dim(h\_test)

[1] 2500 6

> names(h\_test)

[1] "income" "age" "rooms" "beds" "population"

[6] "price"

> #plotting the histograms of each variable

> library(ggplot2)

> par(mfrow=c(3,2))

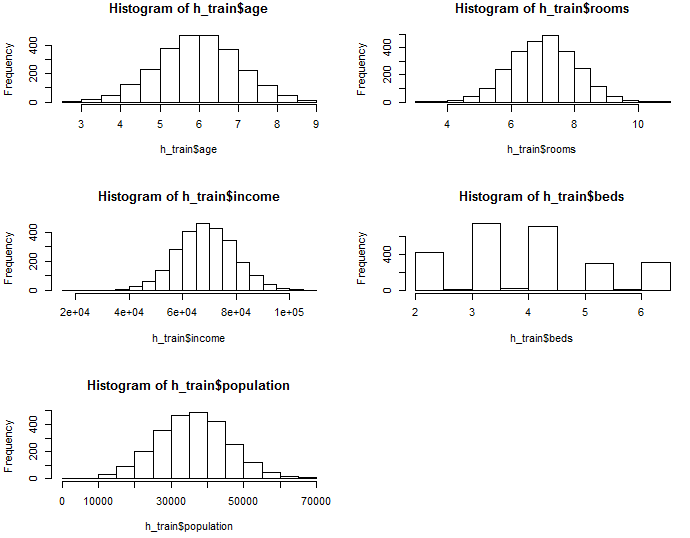
> hist(h\_train$age)

> hist(h\_train$rooms)

> hist(h\_train$income)

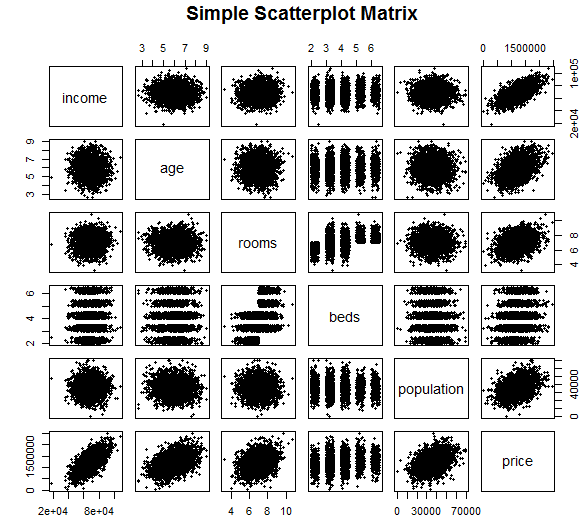
> hist(h\_train$beds)

> hist(h\_train$population)



> #Basic Scatterplot Matrix

> pairs(h\_train,pch=20,main="Simple Scatterplot Matrix")



> #Linear regression model building

> model=lm(price~.,data=h\_train)

> summary(model)

Call:

lm(formula = price ~ ., data = h\_train)

Residuals:

Min 1Q Median 3Q Max

-329225 -71105 -666 70155 313972

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.617e+06 2.413e+04 -108.440 <2e-16 \*\*\*

income 2.174e+01 1.899e-01 114.452 <2e-16 \*\*\*

age 1.644e+05 2.017e+03 81.525 <2e-16 \*\*\*

rooms 1.180e+05 2.289e+03 51.577 <2e-16 \*\*\*

beds 5.295e+02 1.850e+03 0.286 0.775

population 1.517e+01 2.072e-01 73.226 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101500 on 2494 degrees of freedom

Multiple R-squared: 0.9183, Adjusted R-squared: 0.9182

F-statistic: 5609 on 5 and 2494 DF, p-value: < 2.2e-16

> #since beds is not significant in determining, we remove it from the model

> model\_new=lm(price~income+age+rooms+population,data=h\_train)

> summary(model\_new)

Call:

lm(formula = price ~ income + age + rooms + population, data = h\_train)

Residuals:

Min 1Q Median 3Q Max

-329567 -71208 -898 70481 313419

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.617e+06 2.412e+04 -108.51 <2e-16 \*\*\*

income 2.174e+01 1.898e-01 114.53 <2e-16 \*\*\*

age 1.644e+05 2.015e+03 81.60 <2e-16 \*\*\*

rooms 1.183e+05 2.025e+03 58.43 <2e-16 \*\*\*

population 1.517e+01 2.071e-01 73.26 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101500 on 2495 degrees of freedom

Multiple R-squared: 0.9183, Adjusted R-squared: 0.9182

F-statistic: 7014 on 4 and 2495 DF, p-value: < 2.2e-16

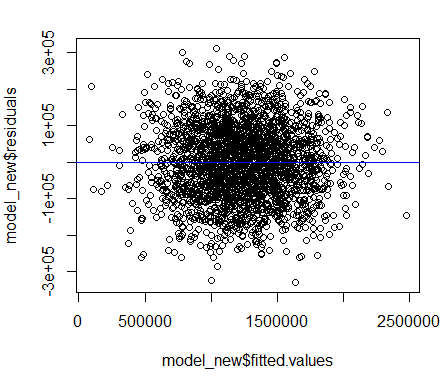
> #Adjusted R square is 0.91, so the model build is 91% accurate.

> par(mfrow=c(1,1))

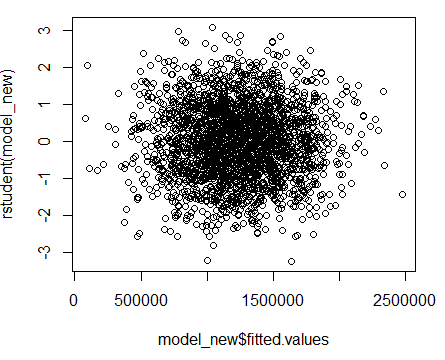
> #plotting the graph of fitted values vs residual for the new model

> plot(model\_new$fitted.values,model\_new$residuals)

> abline(h=0,col="blue")



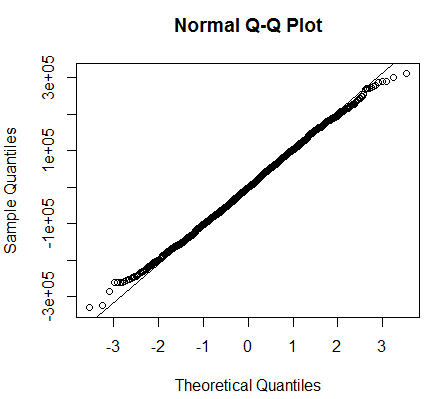
> plot(model\_new$fitted.values,rstudent(model\_new))



> #Checking the qqplot

> qqnorm(model\_new$residuals)

> qqline(model\_new$residuals)

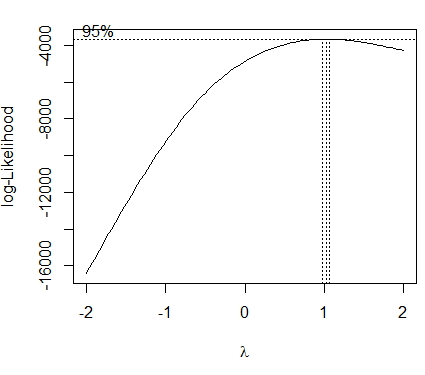


> #Doing boxcox to know the lambda value

> library(car)

> require(MASS)

> boxcox(model\_new)



> #checking for any mutlicollinearity

> vif(model\_new)

income age rooms population

1.000751 1.001193 1.000616 1.000762

> #Confindence interval for the build model

> confint(model\_new)

2.5 % 97.5 %

(Intercept) -2.664378e+06 -2.569786e+06

income 2.136538e+01 2.210972e+01

age 1.604827e+05 1.683861e+05

rooms 1.143707e+05 1.223132e+05

population 1.476576e+01 1.557791e+01

> #Predicting the values of price for the test dataframe with 99% confidence and this 91% accurate with the actual results.

> head(predict(model\_new,h\_test,interval=c("pred"),level=0.99, type="response"))

fit lwr upr

2501 1689585.1 1427727.8 1951442.3

2502 1440330.7 1178640.1 1702021.3

2503 838716.9 577031.3 1100402.5

2504 1697472.8 1435425.5 1959520.1

2505 709906.9 448049.3 971764.6

2506 1805245.5 1543078.3 2067412.8

> head(h\_test)

income age rooms beds population price

2501 74691.95 5.492571 7.937276 5.29 55403.09 1658381.0

2502 68217.68 6.373444 6.586337 3.32 49240.84 1312093.4

2503 55306.14 6.125857 6.410318 4.06 32142.95 647982.7

2504 69012.77 6.981767 9.841095 5.09 33069.76 1492786.6

2505 60846.62 4.072355 7.666556 3.16 28172.07 627733.7

2506 93838.10 4.267184 8.544650 5.50 44138.01 1827500.9

>

>