> library(boot)

> library(car)

> library(QuantPsyc)

> library(lmtest)

> library(sandwich)

> library(vars)

> library(nortest)

> library(MASS)

> data <- read.csv("Data.csv")

> head(data)

Taxi\_dist Market\_dist Hospital\_dist Carpet\_area Builtup\_area Parking\_type City\_type

1 9796 5250 10703 1659 1961 Open CAT B

2 8294 8186 12694 1461 1752 Not Provided CAT B

3 11001 14399 16991 1340 1609 Not Provided CAT A

4 8301 11188 12289 1451 1748 Covered CAT B

5 10510 12629 13921 1770 2111 Not Provided CAT B

6 6665 5142 9972 1442 1733 Open CAT B

Rainfall Price\_house

1 530 6649000

2 210 3982000

3 720 5401000

4 620 5373000

5 450 4662000

6 760 4526000

> summary(data)

Taxi\_dist Market\_dist Hospital\_dist Carpet\_area Builtup\_area

Min. : 146 Min. : 1666 Min. : 3227 Min. : 775 Min. : 932

1st Qu.: 6476 1st Qu.: 9354 1st Qu.:11302 1st Qu.: 1318 1st Qu.: 1583

Median : 8230 Median :11161 Median :13163 Median : 1480 Median : 1774

Mean : 8230 Mean :11019 Mean :13072 Mean : 1512 Mean : 1795

3rd Qu.: 9937 3rd Qu.:12670 3rd Qu.:14817 3rd Qu.: 1655 3rd Qu.: 1982

Max. :20662 Max. :20945 Max. :23294 Max. :24300 Max. :12730

NA's :13 NA's :13 NA's :1 NA's :8 NA's :15

Parking\_type City\_type Rainfall Price\_house

Length:932 Length:932 Min. :-110.0 Min. : 30000

Class :character Class :character 1st Qu.: 600.0 1st Qu.: 4658000

Mode :character Mode :character Median : 780.0 Median : 5866000

Mean : 785.6 Mean : 6084695

3rd Qu.: 970.0 3rd Qu.: 7187250

Max. :1560.0 Max. :150000000

> sapply(data, function(x) sum(is.na(x)))

Taxi\_dist Market\_dist Hospital\_dist Carpet\_area Builtup\_area Parking\_type

13 13 1 8 15 0

City\_type Rainfall Price\_house

0 0 0

> data2 <- na.omit(data)

> data <- data2

> sapply(data, function(x)sum(is.na(x)))

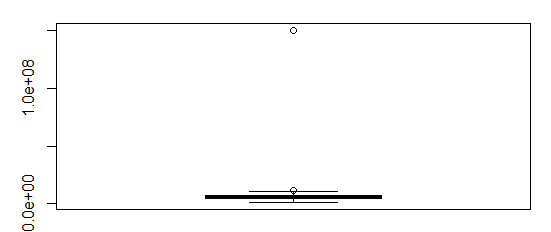
Taxi\_dist Market\_dist Hospital\_dist Carpet\_area Builtup\_area Parking\_type

0 0 0 0 0 0

City\_type Rainfall Price\_house

0 0 0

> boxplot(data$Price\_house)



> quantile(data$Price\_house, c(0,0.05,0.1,0.25,0.5,0.75,0.90,0.95,0.99,0.995,1))

0% 5% 10% 25% 50% 75% 90% 95%

1492000 3340150 3778300 4643000 5860500 7195750 8207300 8795850

99% 99.5% 100%

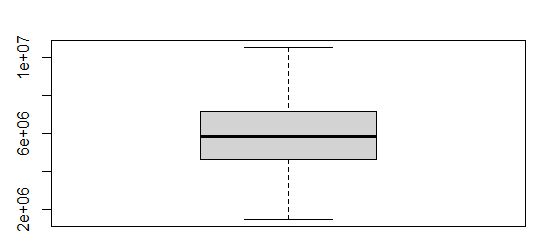
9948270 10207235 150000000

> #removing the outliers

> data2 <- data[data$Price\_house <11207235, ]

> data <- data2

> boxplot(data$Price\_house)



> head(data)

Taxi\_dist Market\_dist Hospital\_dist Carpet\_area Builtup\_area Parking\_type City\_type

1 9796 5250 10703 1659 1961 Open CAT B

2 8294 8186 12694 1461 1752 Not Provided CAT B

3 11001 14399 16991 1340 1609 Not Provided CAT A

4 8301 11188 12289 1451 1748 Covered CAT B

5 10510 12629 13921 1770 2111 Not Provided CAT B

6 6665 5142 9972 1442 1733 Open CAT B

Rainfall Price\_house

1 530 6649000

2 210 3982000

3 720 5401000

4 620 5373000

5 450 4662000

6 760 4526000

> fit<- lm(Price\_house ~ Taxi\_dist + Market\_dist + Hospital\_dist + Carpet\_area + Builtup\_area + Parking\_type + City\_type + Rainfall, data=data)

> summary(fit)

Call:

lm(formula = Price\_house ~ Taxi\_dist + Market\_dist + Hospital\_dist +

Carpet\_area + Builtup\_area + Parking\_type + City\_type + Rainfall,

data = data)

Residuals:

Min 1Q Median 3Q Max

-3583969 -794789 -67950 764835 4396473

Coefficients:

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

(Intercept) 5.639e+06 3.652e+05 15.441 < 2e-16 \*\*\*

Taxi\_dist 3.065e+01 2.666e+01 1.150 0.2506

Market\_dist 1.337e+01 2.067e+01 0.647 0.5179

Hospital\_dist 4.891e+01 2.989e+01 1.636 0.1021

Carpet\_area 2.552e+02 3.453e+03 0.074 0.9411

Builtup\_area 4.226e+02 2.882e+03 0.147 0.8835

Parking\_typeNo Parking -6.462e+05 1.382e+05 -4.677 3.37e-06 \*\*\*

Parking\_typeNot Provided -4.974e+05 1.227e+05 -4.053 5.50e-05 \*\*\*

Parking\_typeOpen -2.661e+05 1.119e+05 -2.378 0.0176 \*

City\_typeCAT B -1.862e+06 9.550e+04 -19.499 < 2e-16 \*\*\*

City\_typeCAT C -2.881e+06 1.052e+05 -27.393 < 2e-16 \*\*\*

Rainfall -1.085e+02 1.532e+02 -0.709 0.4787

---

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

Residual standard error: 1215000 on 884 degrees of freedom

Multiple R-squared: 0.5019, Adjusted R-squared: 0.4957

F-statistic: 80.98 on 11 and 884 DF, p-value: < 2.2e-16

|  |
| --- |
| > #Multicollinearity Test  > #checking the VIF score; vif>2 means presence of multicollinearity  > vif(fit)  GVIF Df GVIF^(1/(2\*Df))  Taxi\_dist 2.754343 1 1.659621  Market\_dist 1.657598 1 1.287477  Hospital\_dist 3.574413 1 1.890612  Carpet\_area 450.904649 1 21.234516  Builtup\_area 450.967435 1 21.235994  Parking\_type 1.038044 3 1.006242  City\_type 1.024406 2 1.006046  Rainfall 1.014887 1 1.007416 |
|  |
| |  | | --- | | > | |

> data$pred <- fitted(fit)

> (sum((abs(data$Price\_house-data$pred))/data$Price\_house))/nrow(data)

[1] 0.1862668

> #Test for Auto-Correlation

> #Durbin Watson Test

> #if values are greater than 2, then auto-correlation exists

> dwt(fit)

lag Autocorrelation D-W Statistic p-value

1 0.03935394 1.920034 0.202

Alternative hypothesis: rho != 0

> #Test for Homoscedasticity

> #Breusch-Pagan Test

> #p-value should be greater than 0.05, so we accept null hypothesis, i.e test is homoscedasticity

> bptest(fit)

studentized Breusch-Pagan test

data: fit

BP = 10.69, df = 11, p-value = 0.4696

> resids <- fit$residuals

> ad.test(resids)

Anderson-Darling normality test

data: resids

A = 0.5671, p-value = 0.1415

> write.csv(data,"house\_prediction.csv")

Summary:

Based on the above Assumption Diagnostic Test, we generate the below insights of the linear model:

1. **Normality Test**: We start with Hypothesis as below

* Null Hypothesis: The errors are normally distributed. p-value should be more than 0.05.
* Alternative Hypothesis: The errors are not normally distributed. p-value is less than 0.05.

Based on the Anderson- Daring Test, we find the p-value is 0.1415. Since the p-value is high than 0.05, we accept the null hypothesis and reject the alternative hypothesis. Finally, we conclude that the errors are normally distributed. We have passed the Normality Test for this linear model.

1. **Homoscedasticity Test**: The assumption means that the variance around the regression line is same for all values of the predictor variable (X). For this test, we use Breusch-Pagan Test.

* Null Hypothesis: The error variances are all equal. p-value should be greater than 0.05.
* Alternative Hypothesis: The error variances are not equal. p-value should be less than 0.05

Based on the Breusch-Pagan Test, we find the p-value is 0.4696. Since the p-value is greater than 0.05, we accept the null hypothesis which says the variance is scattered similarly. We reject the alternative hypothesis and the regression model is homoscedasticity – the variances are similarly scattered.

1. **Multicollinearity Test**: Here, we check if the independent variables have relationship between them, i.e, the correlation between them. The correlation between independent or explanatory variables are called multicollinearity. We use VIF (Variance Inflation Factor) for this test.

* The value of VIF should be lower than 1.7.

Based on the test, we find the VIF score for the variables are less than 1.7. As such, we conclude that there is no multicollinearity on the linear model i.e, no correlation between independent variables in the model.

1. **Auto – Correlation Test**: The good model should not have autocorrelation in the data. Auto- Correlation occurs when the residuals are not independent from each other. For this, we use Durbin-Watson’s d test.

* Null Hypothesis: Residuals are not linearly auto correlated. Values 1.5 < d < 2.5 shows there is no auto correlation in data.
* Alternative Hypothesis: The residuals are linearly auto correlated. Values 1.5 > d > 2.5 shows there is linear auto correlations between data.

Based on the Durbin-Watson Test, we find that the value is 1.92, which is less than 2.5, as such there is no linear auto correlation between data.

Finally, we pass all the assumption diagnostic test.