Prepare a prediction model for profit of 50\_startups data.Do transformations for getting better predictions of profit and make a table containing R^2 value for each prepared model.

R&D Spend -- Research and devolop spend in the past few years

Administration -- spend on administration in the past few years

Marketing Spend -- spend on Marketing in the past few years

State -- states from which data is collected

Profit -- profit of each state in the past few years

Here I am using R programming language to solve this problem

**#Install ‘psych’ package to perform different performance metrics**

install.packages("psych")

**#invoke the library**

library(psych)

**#upload the csv file**

Startups\_data <- read.csv(file.choose())

View(Startups\_data)

|  | **R.D.Spend** | **Administration** | **Marketing.Spend** | **State** | **Profit** |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| **1** | 165349.20 | 136897.80 | 471784.10 | 0 | 192261.83 |
| **2** | 162597.70 | 151377.59 | 443898.53 | 1 | 191792.06 |
| **3** | 153441.51 | 101145.55 | 407934.54 | 2 | 191050.39 |
| **4** | 144372.41 | 118671.85 | 383199.62 | 0 | 182901.99 |
| **5** | 142107.34 | 91391.77 | 366168.42 | 2 | 166187.94 |
| **6** | 131876.90 | 99814.71 | 362861.36 | 0 | 156991.12 |
| **7** | 134615.46 | 147198.87 | 127716.82 | 1 | 156122.51 |
| **8** | 130298.13 | 145530.06 | 323876.68 | 2 | 155752.60 |
| **9** | 120542.52 | 148718.95 | 311613.29 | 0 | 152211.77 |
| **10** | 123334.88 | 108679.17 | 304981.62 | 1 | 149759.96 |
| **11** | 101913.08 | 110594.11 | 229160.95 | 2 | 146121.95 |
| **12** | 100671.96 | 91790.61 | 249744.55 | 1 | 144259.40 |
| **13** | 93863.75 | 127320.38 | 249839.44 | 2 | 141585.52 |
| **14** | 91992.39 | 135495.07 | 252664.93 | 1 | 134307.35 |
| **15** | 119943.24 | 156547.42 | 256512.92 | 2 | 132602.65 |
| **16** | 114523.61 | 122616.84 | 261776.23 | 0 | 129917.04 |
| **17** | 78013.11 | 121597.55 | 264346.06 | 1 | 126992.93 |
| **18** | 94657.16 | 145077.58 | 282574.31 | 0 | 125370.37 |
| **19** | 91749.16 | 114175.79 | 294919.57 | 2 | 124266.90 |
| **20** | 86419.70 | 153514.11 | 0.00 | 0 | 122776.86 |
| **21** | 76253.86 | 113867.30 | 298664.47 | 1 | 118474.03 |
| **22** | 78389.47 | 153773.43 | 299737.29 | 0 | 111313.02 |
| **23** | 73994.56 | 122782.75 | 303319.26 | 2 | 110352.25 |
| **24** | 67532.53 | 105751.03 | 304768.73 | 2 | 108733.99 |
| **25** | 77044.01 | 99281.34 | 140574.81 | 0 | 108552.04 |
| **26** | 64664.71 | 139553.16 | 137962.62 | 1 | 107404.34 |
| **27** | 75328.87 | 144135.98 | 134050.07 | 2 | 105733.54 |
| **28** | 72107.60 | 127864.55 | 353183.81 | 0 | 105008.31 |
| **29** | 66051.52 | 182645.56 | 118148.20 | 2 | 103282.38 |
| **30** | 65605.48 | 153032.06 | 107138.38 | 0 | 101004.64 |
| **31** | 61994.48 | 115641.28 | 91131.24 | 2 | 99937.59 |
| **32** | 61136.38 | 152701.92 | 88218.23 | 0 | 97483.56 |
| **33** | 63408.86 | 129219.61 | 46085.25 | 1 | 97427.84 |
| **34** | 55493.95 | 103057.49 | 214634.81 | 2 | 96778.92 |
| **35** | 46426.07 | 157693.92 | 210797.67 | 1 | 96712.80 |
| **36** | 46014.02 | 85047.44 | 205517.64 | 0 | 96479.51 |
| **37** | 28663.76 | 127056.21 | 201126.82 | 2 | 90708.19 |
| **38** | 44069.95 | 51283.14 | 197029.42 | 1 | 89949.14 |
| **39** | 20229.59 | 65947.93 | 185265.10 | 0 | 81229.06 |
| **40** | 38558.51 | 82982.09 | 174999.30 | 1 | 81005.76 |
| **41** | 28754.33 | 118546.05 | 172795.67 | 1 | 78239.91 |
| **42** | 27892.92 | 84710.77 | 164470.71 | 2 | 77798.83 |
| **43** | 23640.93 | 96189.63 | 148001.11 | 1 | 71498.49 |
| **44** | 15505.73 | 127382.30 | 35534.17 | 0 | 69758.98 |
| **45** | 22177.74 | 154806.14 | 28334.72 | 1 | 65200.33 |
| **46** | 1000.23 | 124153.04 | 1903.93 | 0 | 64926.08 |
| **47** | 1315.46 | 115816.21 | 297114.46 | 2 | 49490.75 |
| **48** | 0.00 | 135426.92 | 0.00 | 1 | 42559.73 |
| **49** | 542.05 | 51743.15 | 0.00 | 0 | 35673.41 |
| **50** | 0.00 | 116983.80 | 45173.06 | 1 | 14681.40 |

Showing 1 to 15 of 50 entries, 5 total columns

**#get column name of dataframe startup\_data**

colnames(Startups\_data)

"**R.D.Spend" "Administration" "Marketing.Spend" "State" "Profit"**

**# get the structure of startup\_data**

str(Startups\_data)

|  |
| --- |
| 'data.frame': 50 obs. of 5 variables:  $ R.D.Spend : num 165349 162598 153442 144372 142107 ...  $ Administration : num 136898 151378 101146 118672 91392 ...  $ Marketing.Spend: num 471784 443899 407935 383200 366168 ...  $ State : num 0 1 2 0 2 0 1 2 0 1 ...  $ Profit : num 192262 191792 191050 182902 166188 ... |
|  |
| |  | | --- | | > | |

**#here we have one categoriacal variable 'state', se we convert it into a numerical variable**

install.packages("plyr")

library(plyr)

Startups\_data$State<-revalue(Startups\_data$State,c("New York"="0", "California"="1", "Florida"="2"))

View(Startups\_data)

**#view first six rows of a data drame**

head(Startups\_data) **# here we can see that data of state variable changed**

R.D.Spend Administration Marketing.Spend State Profit

1 165349.2 136897.80 471784.1 0 192261.8

2 162597.7 151377.59 443898.5 1 191792.1

3 153441.5 101145.55 407934.5 2 191050.4

4 144372.4 118671.85 383199.6 0 182902.0

5 142107.3 91391.77 366168.4 2 166187.9

6 131876.9 99814.71 362861.4 0 156991.1

**# here state variable is in character data types, we have to convert it into a numerical form**

Startups\_data$State = as.numeric(Startups\_data$State)

str(Startups\_data)

data.frame': 50 obs. of 5 variables:

$ R.D.Spend : num 165349 162598 153442 144372 142107 ...

$ Administration : num 136898 151378 101146 118672 91392 ...

$ Marketing.Spend: num 471784 443899 407935 383200 366168 ...

$ State : num 0 1 2 0 2 0 1 2 0 1 ...

$ Profit : num 192262 191792 191050 182902 166188 ...

**#check whether the data set contain null values**

sum(is.na(Startups\_data)) # data set contain zero null values

0

**# summary of dataset**

summary(Startups\_data)

R.D.Spend Administration Marketing.Spend State Profit

Min. : 0 Min. : 51283 Min. : 0 Min. :0.00 Min. : 14681

1st Qu.: 39936 1st Qu.:103731 1st Qu.:129300 1st Qu.:0.00 1st Qu.: 90139

Median : 73051 Median :122700 Median :212716 Median :1.00 Median :107978

Mean : 73722 Mean :121345 Mean :211025 Mean :0.98 Mean :112013

3rd Qu.:101603 3rd Qu.:144842 3rd Qu.:299469 3rd Qu.:2.00 3rd Qu.:139766

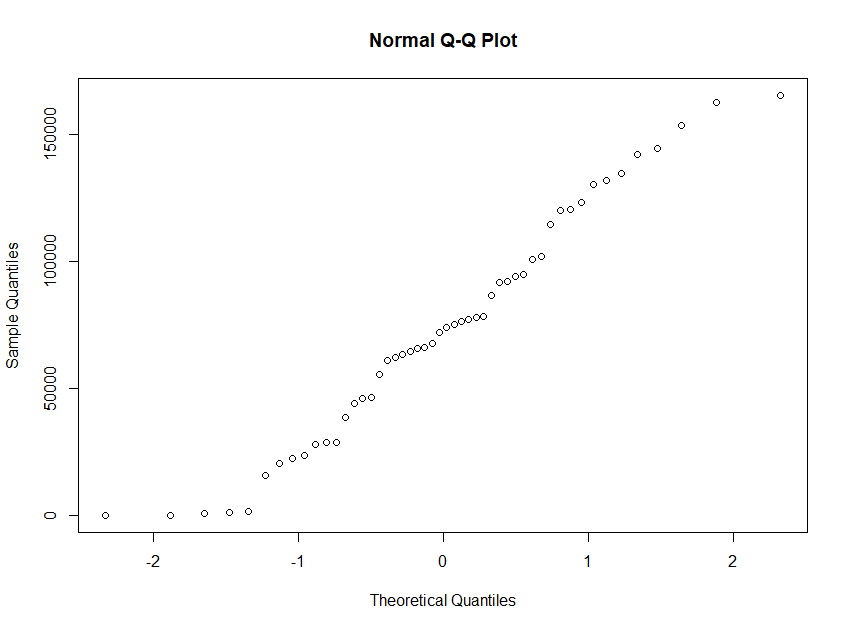
Max. :165349 Max. :182646 Max. :471784 Max. :2.00 Max. :192262

attach(Startups\_data)

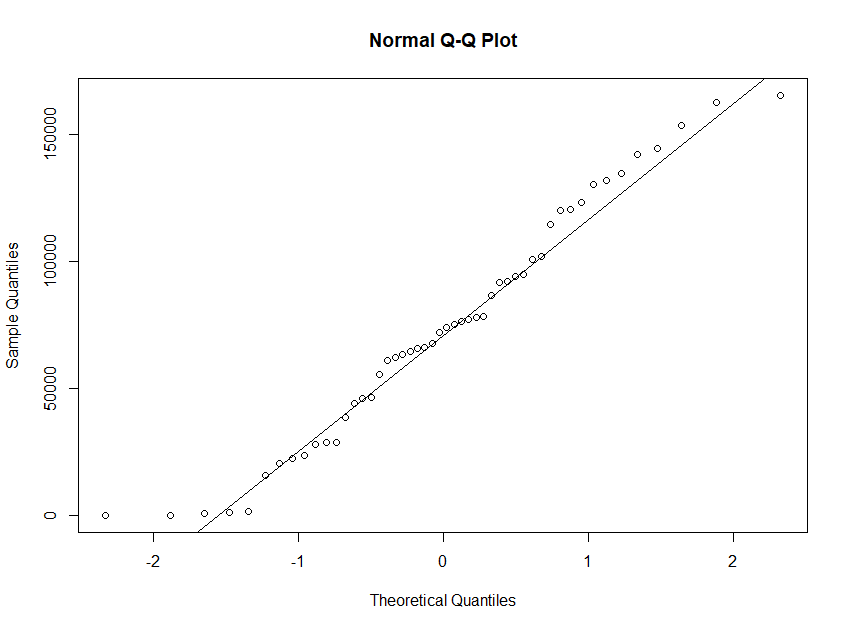
# check the data is normally distributed or not

# using qqplot and shapiro test

qqnorm(R.D.Spend)



qqline(R.D.Spend)



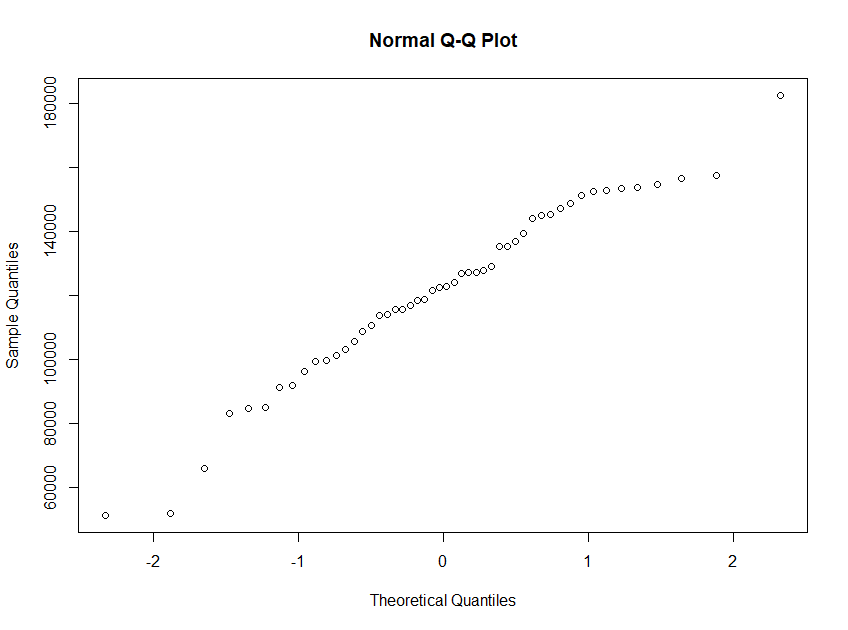
shapiro.test(R.D.Spend) **# pvalue=0.1801 , if the p value>0.05, we can say that data is normally distributed**

Shapiro-Wilk normality test

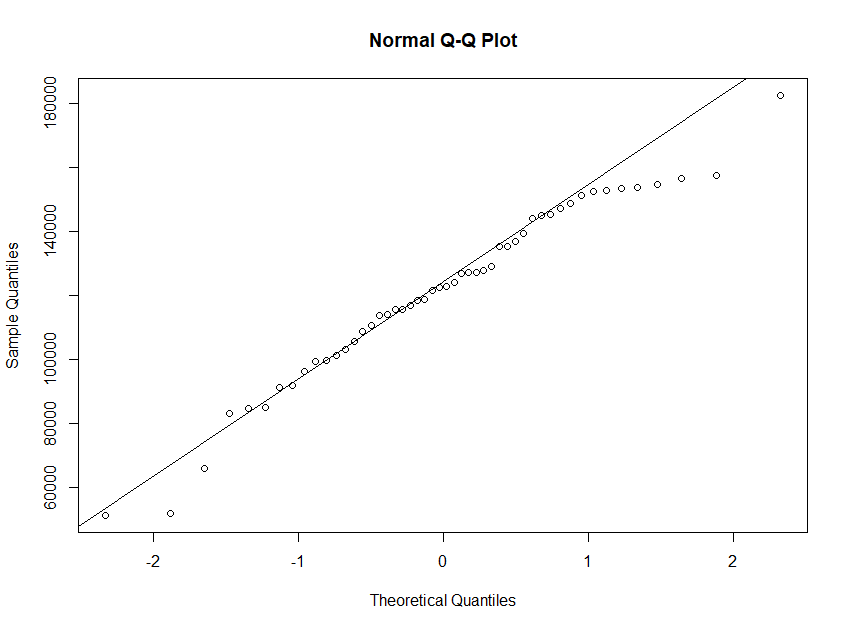
data: R.D.Spend

W = 0.96734, p-value = 0.1801

qqnorm(Administration)



qqline(Administration)



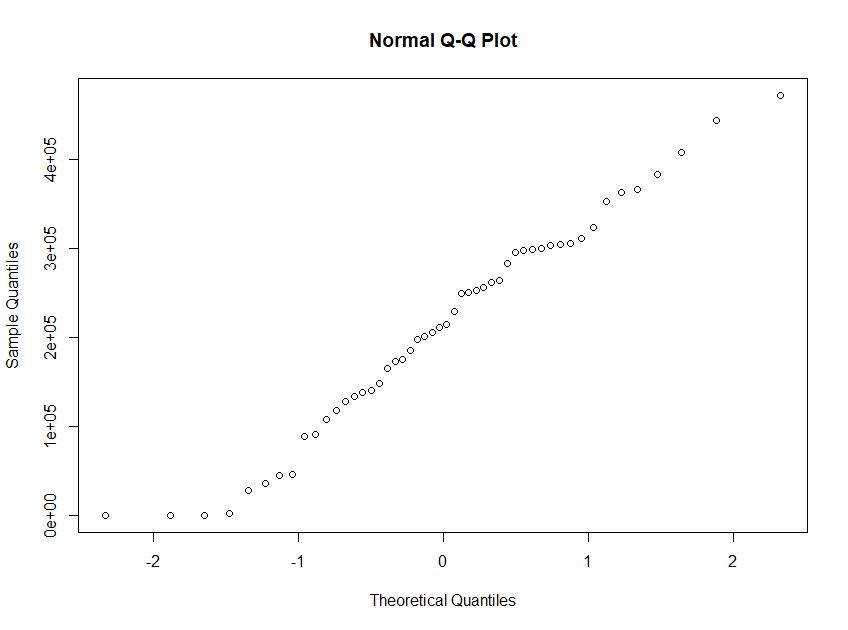
shapiro.test(Administration) **# pvalue=0.2366, if the p value>0.05, we can say that data is normally distributed**

Shapiro-Wilk normality test

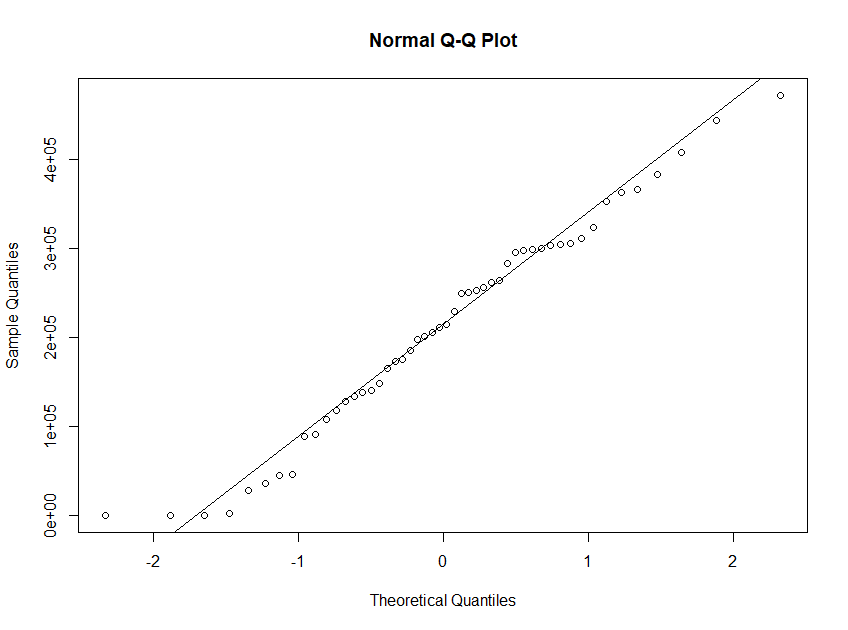
data: Administration

W = 0.97024, p-value = 0.2366

qqnorm(Marketing.Spend)



qqline(Marketing.Spend)

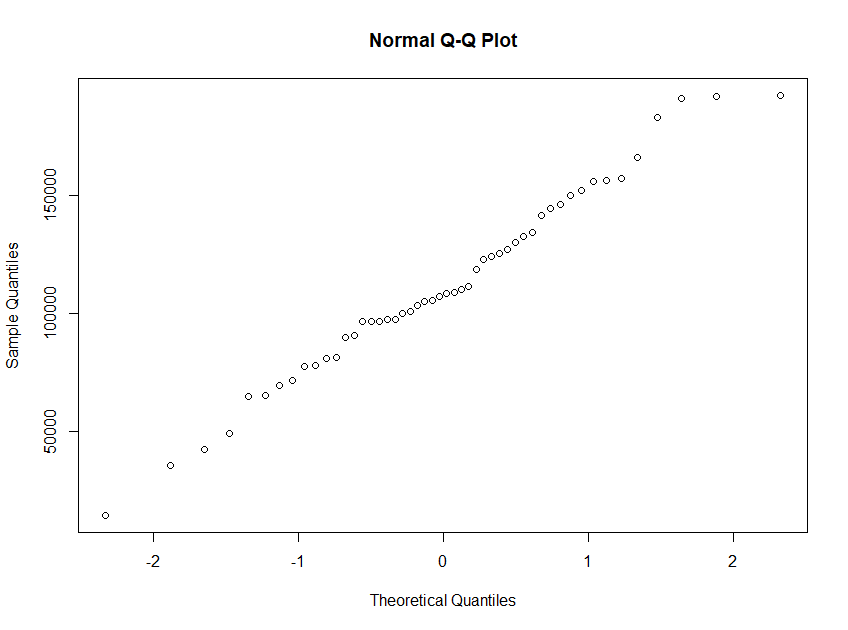


shapiro.test(Marketing.Spend) #p value=0.3 so data is normally distributed

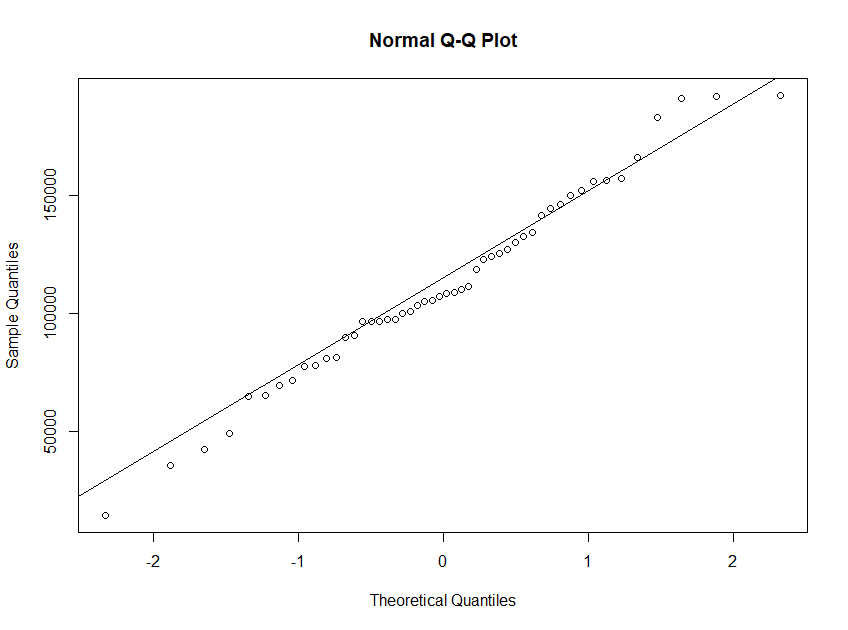
data: Marketing.Spend

W = 0.97437, p-value = 0.3451

qqnorm(Profit)



qqline(Profit)



shapiro.test(Profit) **#p value=0.7, so data is normally distributed**

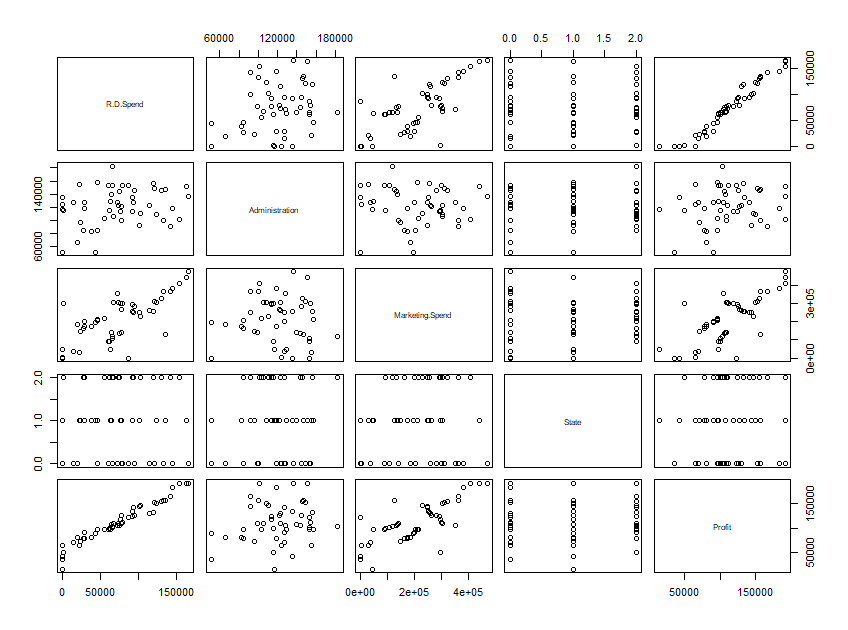
Shapiro-Wilk normality test

data: Profit

W = 0.98488, p-value = 0.7666

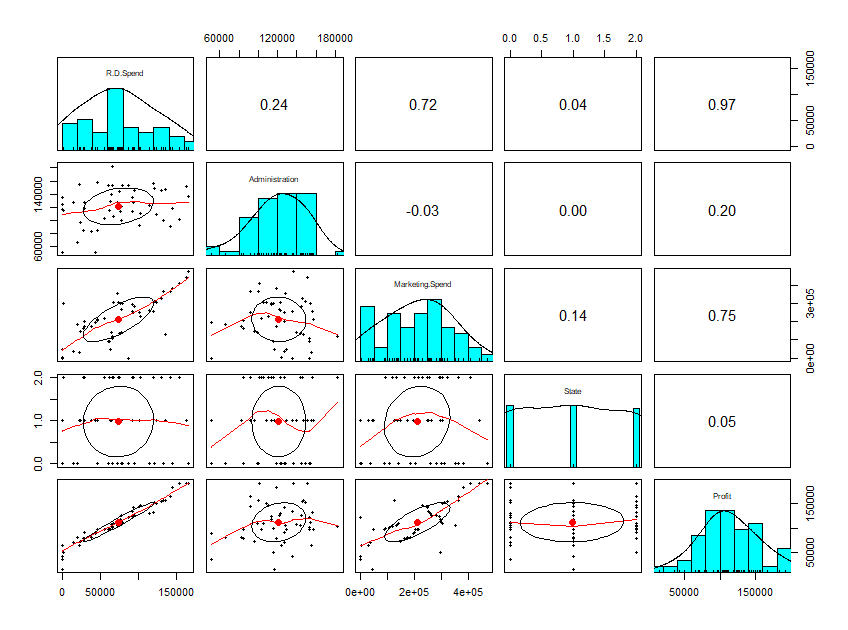
**# Explore the data**

pairs(Startups\_data) **# Scatter plot for all pairs of variables**



**#using pairs.panels to show the correlation between variables**

pairs.panels(Startups\_data)



**# check correlation between target variable profit and independant variable using cor()**

cor(Profit,R.D.Spend) **#get 0.9729, highly correlated, so this variable important to get better model**

0.9729005

cor(Profit,Marketing.Spend) **# get 0.74, better correlation**

0.7477657

cor(Profit,Administration) **# get 0.20, less correlation, the variable administration is not making any impact on model**

0.2007166

**# create first linear model using**

model1<-lm(Profit~R.D.Spend+Administration+State+Marketing.Spend,data=Startups\_data )

summary(model1)

|  |
| --- |
| Call:  lm(formula = Profit ~ R.D.Spend + Administration + State + Marketing.Spend,  data = Startups\_data)  Residuals:  Min 1Q Median 3Q Max  -33546 -4709 109 6638 17374  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 5.004e+04 6.731e+03 7.434 2.32e-09 \*\*\*  R.D.Spend 8.061e-01 4.588e-02 17.569 < 2e-16 \*\*\*  Administration -2.699e-02 5.164e-02 -0.523 0.604  State 1.185e+02 1.649e+03 0.072 0.943  Marketing.Spend 2.703e-02 1.687e-02 1.602 0.116  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 9334 on 45 degrees of freedom  Multiple R-squared: 0.9508, Adjusted R-squared: 0.9464  F-statistic: 217.2 on 4 and 45 DF, p-value: < 2.2e-16 |
|  |
| |  | | --- | | > | |

**#Here only one variable R.D.Spend is significant, so we would individually check other variable significant or not**

model1\_Adm<-lm(Profit~Administration,data=Startups\_data)

summary(model1\_Adm)

|  |
| --- |
| Call:  lm(formula = Profit ~ Administration, data = Startups\_data)  Residuals:  Min 1Q Median 3Q Max  -96072 -23426 -3564 25438 84870  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 7.697e+04 2.532e+04 3.040 0.00382 \*\*  Administration 2.887e-01 2.034e-01 1.419 0.16222  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 39900 on 48 degrees of freedom  Multiple R-squared: 0.04029, Adjusted R-squared: 0.02029  F-statistic: 2.015 on 1 and 48 DF, p-value: 0.1622 |
|  |
| |  | | --- | | > | |

model1\_markng<-lm(Profit~Marketing.Spend,data=Startups\_data)

summary(model1\_markng)

|  |
| --- |
| Call:  lm(formula = Profit ~ Marketing.Spend, data = Startups\_data)  Residuals:  Min 1Q Median 3Q Max  -83739 -18802 4925 15879 64642  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 6.000e+04 7.685e+03 7.808 4.29e-10 \*\*\*  Marketing.Spend 2.465e-01 3.159e-02 7.803 4.38e-10 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 27040 on 48 degrees of freedom  Multiple R-squared: 0.5592, Adjusted R-squared: 0.55  F-statistic: 60.88 on 1 and 48 DF, p-value: 4.381e-10 |
|  |
| |  | | --- | | > # we have checked individually , to get the marketing.spend variable is significant | |

**# check partial correlation**

**###Partial Correlation matrix**

install.packages("corpcor")

library(corpcor)

cor2pcor(cor(Startups\_data))

|  |
| --- |
| [,1] [,2] [,3] [,4] [,5]  [1,] 1.00000000 0.21023496 0.04571253 -0.04612757 0.93421673  [2,] 0.21023496 1.00000000 -0.28547598 0.04680553 -0.07766241  [3,] 0.04571253 -0.28547598 1.00000000 0.15926762 0.23232585  [4,] -0.04612757 0.04680553 0.15926762 1.00000000 0.01071012  [5,] 0.93421673 -0.07766241 0.23232585 0.01071012 1.00000000 |
|  |
| |  | | --- | | > | |

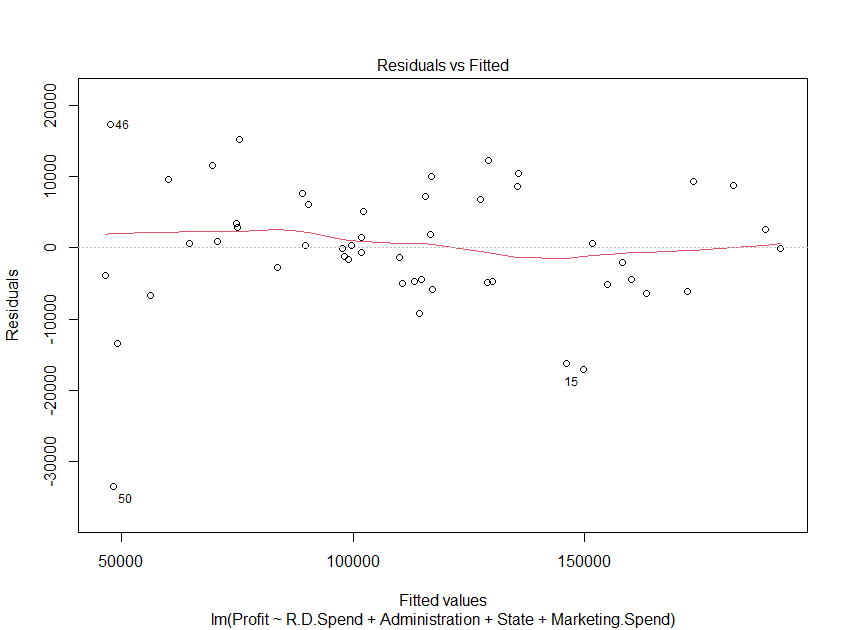
**# use some diagnostic plot to detect outliers**

install.packages("car")

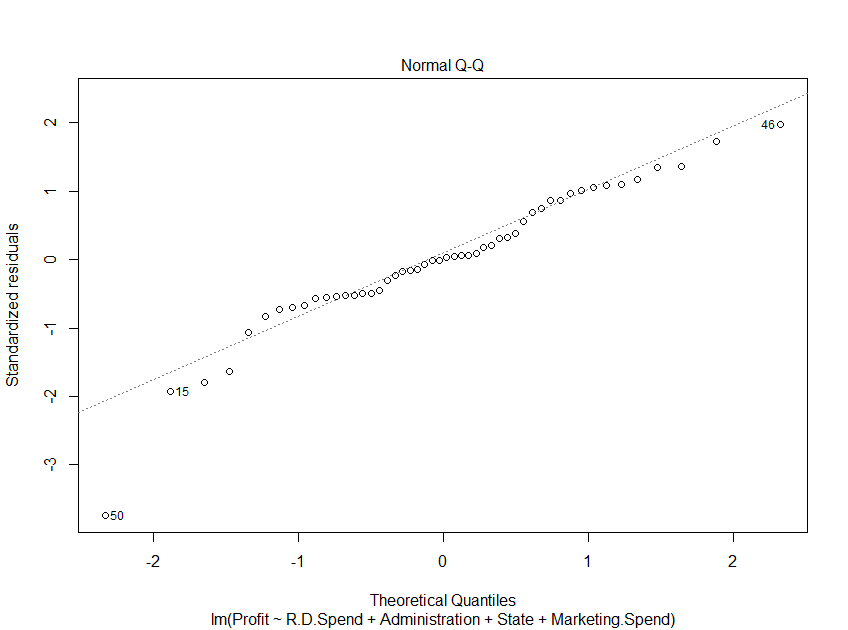
library(car)

plot(model1) # **Residual Plots, QQ-Plos, Std. Residuals vs Fitted, Cook's distance**

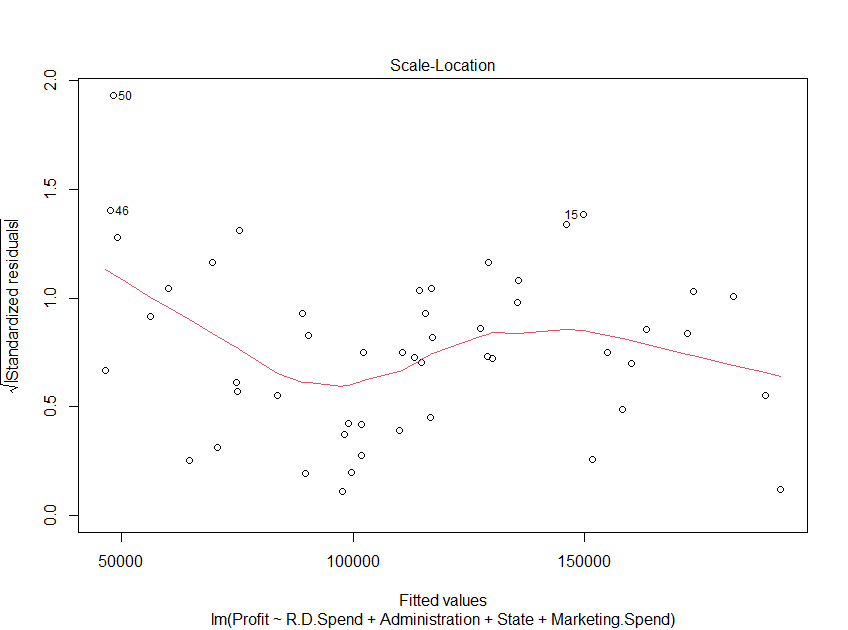
Hit <Return> to see next plot: 1



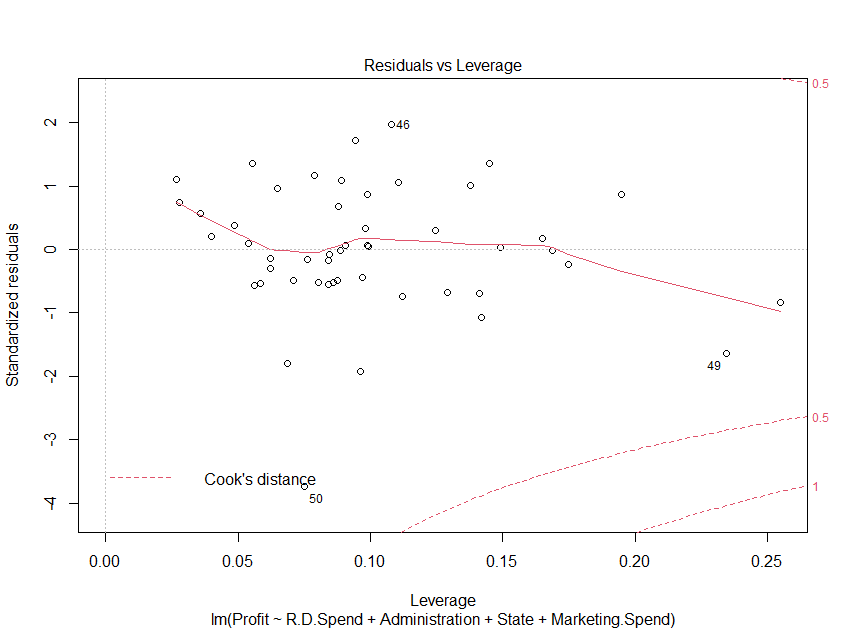
Hit <Return> to see next plot:2



Hit <Return> to see next plot:3



Hit <Return> to see next plot:4

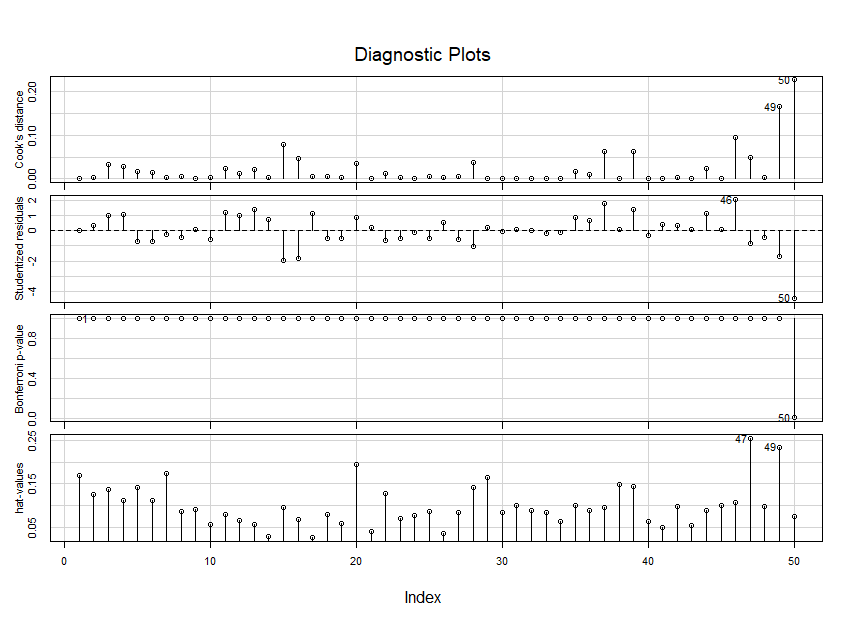


**# Deletion Diagnostics for identifying influential variable**

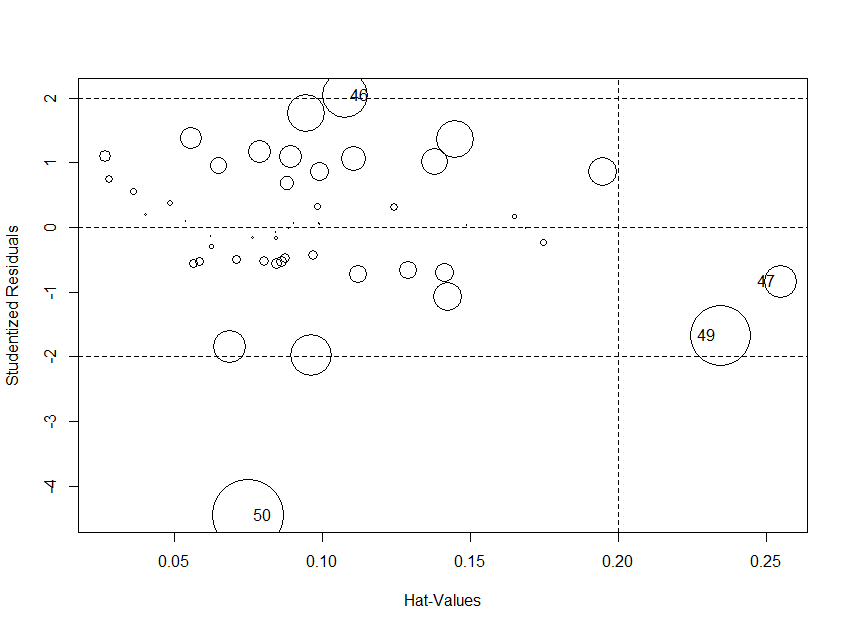
influence.measures(model1)

|  |
| --- |
| Influence measures of  lm(formula = Profit ~ R.D.Spend + Administration + State + Marketing.Spend, data = Startups\_data) :  dfb.1\_ dfb.R.D. dfb.Admn dfb.Stat dfb.Mr.S dffit cov.r cook.d hat inf  1 0.001574 -0.000645 -0.001016 0.003175 -0.002811 -0.00622 1.346 7.90e-06 0.1689 \*  2 -0.060079 0.017425 0.041360 -0.009609 0.045475 0.11345 1.265 2.63e-03 0.1244  3 0.052392 0.200965 -0.175296 0.181314 -0.006444 0.40567 1.156 3.29e-02 0.1378  4 0.028798 0.112377 -0.049377 -0.212845 0.081390 0.37429 1.109 2.79e-02 0.1106  5 -0.085164 -0.155233 0.162906 -0.131957 0.042204 -0.28271 1.233 1.62e-02 0.1411  6 -0.092251 -0.082775 0.105510 0.143119 -0.038351 -0.25997 1.186 1.37e-02 0.1122  7 -0.007800 -0.093273 0.005235 -0.013611 0.086696 -0.10880 1.347 2.42e-03 0.1747 \*  8 0.069750 -0.045951 -0.041655 -0.085851 -0.002665 -0.15077 1.194 4.62e-03 0.0874  9 -0.006817 0.001131 0.008829 -0.012981 0.005807 0.02089 1.230 8.92e-05 0.0906  10 -0.045528 -0.082464 0.064409 -0.001863 0.021617 -0.13704 1.144 3.81e-03 0.0563  11 0.087169 0.189817 -0.136130 0.234561 -0.147865 0.34341 1.041 2.34e-02 0.0787  12 0.174700 0.142179 -0.191767 0.012767 -0.081572 0.25290 1.079 1.28e-02 0.0648  13 -0.050063 0.072505 0.011828 0.250095 -0.034559 0.33290 0.961 2.17e-02 0.0556  14 -0.038622 0.002144 0.050050 -0.002615 0.023299 0.12477 1.082 3.15e-03 0.0278  15 0.322346 -0.225143 -0.259494 -0.377243 0.122979 -0.64853 0.806 7.90e-02 0.0964  16 -0.113810 -0.212039 0.064880 0.322262 0.055808 -0.50066 0.827 4.76e-02 0.0685  17 -0.008516 -0.057920 0.025843 -0.011592 0.089256 0.18154 1.004 6.56e-03 0.0266  18 0.048342 0.034392 -0.075917 0.103256 -0.065790 -0.15392 1.180 4.82e-03 0.0805  19 0.003098 -0.004062 0.019100 -0.090940 -0.023085 -0.13242 1.151 3.56e-03 0.0583  20 0.061189 0.253440 0.040641 -0.106873 -0.329370 0.42230 1.278 3.59e-02 0.1946  21 0.000742 -0.018588 0.000656 -0.004006 0.027623 0.04085 1.160 3.41e-04 0.0401  22 0.116992 0.133051 -0.164071 0.147634 -0.163489 -0.25761 1.221 1.34e-02 0.1290  23 0.037995 0.055942 -0.026153 -0.077591 -0.070455 -0.13555 1.172 3.74e-03 0.0708  24 -0.000141 0.016768 0.004666 -0.023961 -0.021216 -0.04328 1.208 3.83e-04 0.0763  25 -0.119982 -0.078080 0.087830 0.079317 0.081704 -0.16066 1.187 5.25e-03 0.0859  26 -0.011507 0.009566 0.043491 0.009289 -0.038301 0.10774 1.120 2.36e-03 0.0359  27 0.032526 -0.047992 -0.041327 -0.115150 0.078706 -0.16849 1.180 5.77e-03 0.0843  28 0.093996 0.270852 -0.147632 0.252778 -0.350724 -0.43636 1.146 3.80e-02 0.1421  29 -0.046513 -0.004489 0.055602 0.035819 -0.011977 0.07707 1.335 1.21e-03 0.1649 \*  30 0.002785 -0.001341 -0.011130 0.012250 0.005870 -0.02310 1.221 1.09e-04 0.0844  31 0.004038 0.005524 -0.003444 0.008550 -0.008589 0.01270 1.242 3.30e-05 0.0991  32 0.000299 -0.000258 -0.001634 0.001809 0.001105 -0.00358 1.228 2.62e-06 0.0887  33 -0.017753 -0.029435 0.005618 -0.008534 0.045792 -0.05392 1.218 5.94e-04 0.0843  34 -0.009266 0.005255 0.010305 -0.024750 -0.001802 -0.03534 1.191 2.55e-04 0.0623  35 -0.167034 -0.192657 0.230403 -0.017555 0.149088 0.28688 1.141 1.66e-02 0.0991  36 0.146735 -0.048638 -0.101790 -0.125118 0.042023 0.21080 1.165 9.00e-03 0.0880  37 -0.137164 -0.374319 0.191758 0.295404 0.240557 0.56956 0.878 6.20e-02 0.0944  38 0.013985 0.000957 -0.013616 0.000436 -0.001478 0.01549 1.315 4.91e-05 0.1490  39 0.420831 -0.186484 -0.311370 -0.262203 0.125908 0.56089 1.064 6.17e-02 0.1449  40 -0.062266 0.012650 0.051179 -0.002051 0.000379 -0.07707 1.181 1.21e-03 0.0624  41 0.006436 -0.062087 0.018471 -0.002541 0.035517 0.08439 1.158 1.45e-03 0.0487  42 0.054216 -0.021920 -0.050766 0.061131 -0.002350 0.10597 1.226 2.29e-03 0.0982  43 0.013632 -0.010359 -0.007813 0.000389 0.002783 0.02276 1.181 1.06e-04 0.0539  44 0.097656 -0.087355 0.060461 -0.172487 -0.069870 0.34121 1.075 2.32e-02 0.0892  45 -0.004386 -0.005027 0.012104 0.001633 -0.005448 0.02119 1.241 9.18e-05 0.0990  46 0.235235 -0.204640 0.094171 -0.317826 -0.158680 0.70883 0.799 9.39e-02 0.1079  47 0.115584 0.429519 -0.142888 -0.115941 -0.373296 -0.48886 1.388 4.81e-02 0.2549 \*  48 -0.011763 0.047993 -0.045816 -0.012380 0.038581 -0.14454 1.212 4.25e-03 0.0969  49 -0.893571 -0.116114 0.683871 0.245870 0.384139 -0.92708 1.073 1.65e-01 0.2345  50 -0.386898 0.581546 -0.102670 -0.082337 0.154587 -1.26784 0.189 2.27e-01 0.0751 \* |
|  |
| |  | | --- | | > | |

influenceIndexPlot(model1) **# Index Plots of the influence measures**



influencePlot(model1) **# A user friendly representation of the above**



|  |
| --- |
| StudRes Hat CookD  46 2.0386686 0.10785323 0.09390326  47 -0.8357182 0.25494143 0.04811938  49 -1.6749535 0.23451409 0.16526587  50 -4.4496953 0.07508785 0.22675297 |
|  |
| |  | | --- | | > | |

**#records 49,50 considered as outliers, so remove these records from dataset and create new model**

**#create second model**

model2<-lm(Profit~R.D.Spend+Administration+State+Marketing.Spend,data=Startups\_data[-c(49,50),])

summary(model2)

|  |
| --- |
| Call:  lm(formula = Profit ~ R.D.Spend + Administration + State + Marketing.Spend,  data = Startups\_data[-c(49, 50), ])  Residuals:  Min 1Q Median 3Q Max  -16479 -4844 -1976 5856 13844  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 5.932e+04 6.114e+03 9.702 2.14e-12 \*\*\*  R.D.Spend 7.889e-01 3.690e-02 21.378 < 2e-16 \*\*\*  Administration -6.345e-02 4.442e-02 -1.428 0.160  State -2.347e+02 1.327e+03 -0.177 0.861  Marketing.Spend 1.721e-02 1.380e-02 1.247 0.219  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 7432 on 43 degrees of freedom  Multiple R-squared: 0.9627, Adjusted R-squared: 0.9592  F-statistic: 277.5 on 4 and 43 DF, p-value: < 2.2e-16 |
|  |
| |  | | --- | | > | |

**#check multicollinearity present in a data set**

**# here using Variance Inflation Factors(VIF) technique helps to identify the multicollinearity**

vif(model2) **# VIF is > 10 => collinearity**

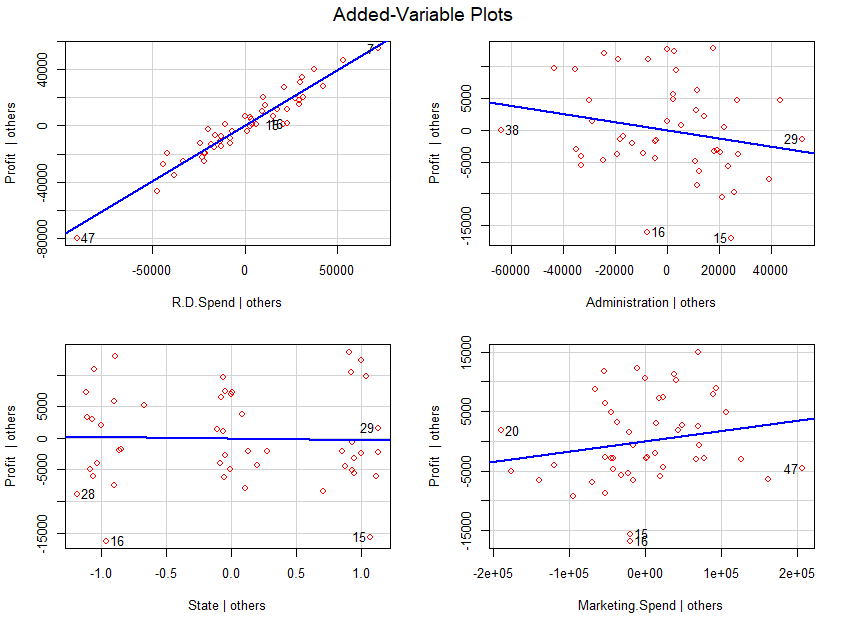
|  |
| --- |
| R.D.Spend Administration State Marketing.Spend  2.268684 1.196206 1.020948 2.267071 |
|  |
| |  | | --- | | > | |

**# here we get VIF values of all variable less than 10, so we can say that there is no multicollinearity present in our data set**

**# another method to check the multicollinearity use added variable plot(AV plots)**

**#### Added Variable Plots ######**

avPlots(model2, id.n=5, id.cex=100, col="red")



**#The Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. In statistics, AIC is used to compare different possible models and determine which one is the best fit for the data.**

install.packages("MASS")

library("MASS")

stepAIC(model1) # backward

|  |
| --- |
| Start: AIC=918.87  Profit ~ R.D.Spend + Administration + State + Marketing.Spend  Df Sum of Sq RSS AIC  - State 1 4.4975e+05 3.9209e+09 916.88  - Administration 1 2.3789e+07 3.9442e+09 917.17  <none> 3.9204e+09 918.87  - Marketing.Spend 1 2.2368e+08 4.1441e+09 919.65  - R.D.Spend 1 2.6891e+10 3.0811e+10 1019.96  Step: AIC=916.88  Profit ~ R.D.Spend + Administration + Marketing.Spend  Df Sum of Sq RSS AIC  - Administration 1 2.3539e+07 3.9444e+09 915.18  <none> 3.9209e+09 916.88  - Marketing.Spend 1 2.3349e+08 4.1543e+09 917.77  - R.D.Spend 1 2.7147e+10 3.1068e+10 1018.37  Step: AIC=915.18  Profit ~ R.D.Spend + Marketing.Spend  Df Sum of Sq RSS AIC  <none> 3.9444e+09 915.18  - Marketing.Spend 1 3.1165e+08 4.2560e+09 916.98  - R.D.Spend 1 3.1149e+10 3.5094e+10 1022.46  Call:  lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = Startups\_data)  Coefficients:  (Intercept) R.D.Spend Marketing.Spend  4.698e+04 7.966e-01 2.991e-02 |
|  |
| |  | | --- | | > | |

**# Lower the AIC (Akaike Information Criterion) value better is the model. AIC is used only if you build**

**# multiple models.**

**#Profit ~ R.D.Spend + Marketing.Spend, this combination got lowest AIC value, so choose**

**#this combination for prediction**

**# create next model**

model3<-lm( Profit ~ R.D.Spend + Marketing.Spend, data = Startups\_data[-c(49,50), ])

summary(model3)

|  |
| --- |
| Call:  lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = Startups\_data[-c(49,  50), ])  Residuals:  Min 1Q Median 3Q Max  -17147 -4243 -1595 4888 12845  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 5.127e+04 2.376e+03 21.574 <2e-16 \*\*\*  R.D.Spend 7.694e-01 3.399e-02 22.636 <2e-16 \*\*\*  Marketing.Spend 2.415e-02 1.271e-02 1.899 0.0639 .  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 7437 on 45 degrees of freedom  Multiple R-squared: 0.9609, Adjusted R-squared: 0.9592  F-statistic: 553.1 on 2 and 45 DF, p-value: < 2.2e-16 |
|  |
| |  | | --- | | > | |

**# to improve the R-squared values, here I am using different transformation techniques**

**#logarithmic transformation**

model4<-lm( Profit ~ log(R.D.Spend) + log(Marketing.Spend), data = Startups\_data[-c(49,50), ])

summary(model4)

|  |
| --- |
| Call:  lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = Startups\_data[-c(49,  50), ])  Residuals:  Min 1Q Median 3Q Max  -17147 -4243 -1595 4888 12845  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 5.127e+04 2.376e+03 21.574 <2e-16 \*\*\*  R.D.Spend 7.694e-01 3.399e-02 22.636 <2e-16 \*\*\*  Marketing.Spend 2.415e-02 1.271e-02 1.899 0.0639 .  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 7437 on 45 degrees of freedom  Multiple R-squared: 0.9609, Adjusted R-squared: 0.9592  F-statistic: 553.1 on 2 and 45 DF, p-value: < 2.2e-16 |
|  |
| |  | | --- | | > | |

**#exponential transformation**

model5<-lm( log(Profit) ~ R.D.Spend + Marketing.Spend, data = Startups\_data[-c(49,50), ])

summary(model5)

|  |
| --- |
| Call:  lm(formula = log(Profit) ~ R.D.Spend + Marketing.Spend, data = Startups\_data[-c(49,  50), ])  Residuals:  Min 1Q Median 3Q Max  -0.37520 -0.03278 0.00667 0.05383 0.15858  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 1.103e+01 3.089e-02 357.165 <2e-16 \*\*\*  R.D.Spend 7.207e-06 4.419e-07 16.308 <2e-16 \*\*\*  Marketing.Spend 8.151e-08 1.653e-07 0.493 0.624  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.09669 on 45 degrees of freedom  Multiple R-squared: 0.9221, Adjusted R-squared: 0.9187  F-statistic: 266.5 on 2 and 45 DF, p-value: < 2.2e-16 |
|  |
| |  | | --- | | > | |

#square root transformation

model6<-lm( sqrt(Profit) ~ R.D.Spend + Marketing.Spend, data = Startups\_data[-c(49,50), ])

summary(model6)

|  |
| --- |
| Call:  lm(formula = sqrt(Profit) ~ R.D.Spend + Marketing.Spend, data = Startups\_data[-c(49,  50), ])  Residuals:  Min 1Q Median 3Q Max  -34.643 -5.725 -0.881 8.303 21.943  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 2.409e+02 3.837e+00 62.795 <2e-16 \*\*\*  R.D.Spend 1.164e-03 5.489e-05 21.211 <2e-16 \*\*\*  Marketing.Spend 2.447e-05 2.053e-05 1.192 0.239  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 12.01 on 45 degrees of freedom  Multiple R-squared: 0.9541, Adjusted R-squared: 0.952  F-statistic: 467.4 on 2 and 45 DF, p-value: < 2.2e-16 |
|  |
| |  | | --- | | > | |

**# polynomial model transformation**

model7<-lm( Profit ~ R.D.Spend+I(R.D.Spend^2)+I(R.D.Spend^3) + Marketing.Spend+I(Marketing.Spend^2)+I(Marketing.Spend^3), data = Startups\_data[-c(49,50), ])

summary(model7)

|  |
| --- |
| Call:  lm(formula = Profit ~ R.D.Spend + I(R.D.Spend^2) + I(R.D.Spend^3) +  Marketing.Spend + I(Marketing.Spend^2) + I(Marketing.Spend^3),  data = Startups\_data[-c(49, 50), ])  Residuals:  Min 1Q Median 3Q Max  -15952 -4066 -1781 3793 14807  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 5.069e+04 4.254e+03 11.915 6.75e-15 \*\*\*  R.D.Spend 1.270e+00 3.308e-01 3.839 0.000419 \*\*\*  I(R.D.Spend^2) -9.570e-06 5.490e-06 -1.743 0.088806 .  I(R.D.Spend^3) 4.654e-11 2.521e-11 1.846 0.072084 .  Marketing.Spend -9.277e-02 1.044e-01 -0.889 0.379383  I(Marketing.Spend^2) 7.500e-07 6.165e-07 1.217 0.230722  I(Marketing.Spend^3) -1.343e-12 1.032e-12 -1.302 0.200249  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 7349 on 41 degrees of freedom  Multiple R-squared: 0.9652, Adjusted R-squared: 0.9601  F-statistic: 189.7 on 6 and 41 DF, p-value: < 2.2e-16 |
|  |
| |  | | --- | | > | |

**#quadratic model**

model8<-lm( Profit ~ R.D.Spend+I(R.D.Spend^2) + Marketing.Spend+I(Marketing.Spend^2), data = Startups\_data[-c(49,50), ])

summary(model8)

|  |
| --- |
| Call:  lm(formula = Profit ~ R.D.Spend + I(R.D.Spend^2) + Marketing.Spend +  I(Marketing.Spend^2), data = Startups\_data[-c(49, 50), ])  Residuals:  Min 1Q Median 3Q Max  -16990.4 -4031.9 -749.6 5261.9 13097.6  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 5.417e+04 3.628e+03 14.928 < 2e-16 \*\*\*  R.D.Spend 6.834e-01 1.104e-01 6.192 1.92e-07 \*\*\*  I(R.D.Spend^2) 5.628e-07 7.613e-07 0.739 0.464  Marketing.Spend 2.002e-02 3.934e-02 0.509 0.613  I(Marketing.Spend^2) 3.527e-09 1.022e-07 0.035 0.973  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 7500 on 43 degrees of freedom  Multiple R-squared: 0.962, Adjusted R-squared: 0.9585  F-statistic: 272.2 on 4 and 43 DF, p-value: < 2.2e-16 |
|  |
| |  | | --- | | > | |

# now we choose best model for prediction

|  |  |  |
| --- | --- | --- |
| MODEL | Multiple R-squared | Adjusted R-squared |
| Model1 | 0.9508 | 0.9464 |
| Model2 | :0.9627 | 0.9592 |
| Model3 | 0.9614 | 0.9596 |
| Model4 | 0.9609 | 0.9592 |
| Model5 | 0.9221 | 0.9187 |
| Model6 | 0.9541 | 0.952 |
| Model7 | 0.9652 | 0.9601 |
| Model8 | 0.962 | 0.9585 |

**# here we can see that model7 has higher adjusted R-squared value. so we can choose model as our final model 7 for predict the profit**

final\_model<-model7

summary(final\_model)

|  |
| --- |
| Call:  lm(formula = Profit ~ R.D.Spend + I(R.D.Spend^2) + I(R.D.Spend^3) +  Marketing.Spend + I(Marketing.Spend^2) + I(Marketing.Spend^3),  data = Startups\_data[-c(49, 50), ])  Residuals:  Min 1Q Median 3Q Max  -15952 -4066 -1781 3793 14807  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 5.069e+04 4.254e+03 11.915 6.75e-15 \*\*\*  R.D.Spend 1.270e+00 3.308e-01 3.839 0.000419 \*\*\*  I(R.D.Spend^2) -9.570e-06 5.490e-06 -1.743 0.088806 .  I(R.D.Spend^3) 4.654e-11 2.521e-11 1.846 0.072084 .  Marketing.Spend -9.277e-02 1.044e-01 -0.889 0.379383  I(Marketing.Spend^2) 7.500e-07 6.165e-07 1.217 0.230722  I(Marketing.Spend^3) -1.343e-12 1.032e-12 -1.302 0.200249  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 7349 on 41 degrees of freedom  Multiple R-squared: 0.9652, Adjusted R-squared: 0.9601  F-statistic: 189.7 on 6 and 41 DF, p-value: < 2.2e-16 |
|  |
| |  | | --- | | > | |

**#prediction**

confint(final\_model,level = 0.95)

|  |
| --- |
| 2.5 % 97.5 %  (Intercept) 4.209487e+04 5.927763e+04  R.D.Spend 6.020989e-01 1.938402e+00  I(R.D.Spend^2) -2.065803e-05 1.517540e-06  I(R.D.Spend^3) -4.368395e-12 9.744907e-11  Marketing.Spend -3.036002e-01 1.180612e-01  I(Marketing.Spend^2) -4.950068e-07 1.994991e-06  I(Marketing.Spend^3) -3.426562e-12 7.404677e-13 |
|  |
| |  | | --- | | > | |

profit\_pred<-predict(final\_model,interval = "predict")

profit\_pred

|  |
| --- |
| fit lwr upr  1 191595.46 172604.35 210586.56  2 193401.93 176386.56 210417.30  3 184196.13 167948.30 200443.97  4 173656.73 157983.42 189330.03  5 172144.10 156276.46 188011.73  6 159424.43 143886.88 174961.97  7 159374.46 142116.14 176632.79  8 159669.34 144058.72 175279.96  9 149541.88 134120.27 164963.48  10 152457.10 136980.80 167933.40  11 131968.80 116482.99 147454.61  12 131745.97 116397.39 147094.55  13 126778.91 111521.67 142036.15  14 125558.04 110326.76 140789.33  15 148554.38 132924.73 164184.02  16 143563.06 128087.32 159038.80  17 116710.71 101483.29 131938.13  18 128015.42 112711.02 143319.82  19 126035.88 110661.59 141410.17  20 119024.59 101433.91 136615.27  21 115948.14 100463.37 131432.91  22 117278.04 101810.07 132746.02  23 114517.13 98951.90 130082.36  24 110526.87 94812.20 126241.54  25 111077.05 95395.61 126758.50  26 103342.40 87948.15 118736.64  27 109766.65 94092.23 125441.06  28 111588.98 94942.16 128235.79  29 103540.22 88030.65 119049.80  30 102990.21 87451.52 118528.90  31 100500.13 84986.54 116013.71  32 99939.98 84431.44 115448.52  33 101803.87 85966.93 117640.80  34 101018.11 85608.65 116427.58  35 94879.11 79297.41 110460.81  36 94360.40 78810.89 109909.91  37 81082.65 65426.93 96738.37  38 92626.70 77104.26 108149.15  39 72866.55 57174.11 88558.99  40 87640.50 72174.61 103106.38  41 79839.16 64316.93 95361.39  42 78736.03 63227.44 94244.63  43 74326.42 58761.04 89891.81  44 65845.21 50151.86 81538.55  45 72601.04 56618.99 88583.08  46 51773.34 34783.00 68763.69  47 55758.69 37924.19 73593.19  48 50686.25 33537.08 67835.41 |
|  |
| |  | | --- | | > | |

Final\_data <- cbind(Startups\_data$R.D.Spend,Startups\_data$Administration,

Startups\_data$Marketing.Spend,Startups\_data$State,Startups\_data$Profit,profit\_pred)

View(Final\_data)

|  | R.D.Spend | Administration | Marketing | State | Profit | **fit** | **lwr** | **upr** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |
| **1** | 165349.20 | 136897.80 | 471784.10 | 0 | 192261.83 | 191595.46 | 172604.35 | 210586.56 |
| **2** | 162597.70 | 151377.59 | 443898.53 | 1 | 191792.06 | 193401.93 | 176386.56 | 210417.30 |
| **3** | 153441.51 | 101145.55 | 407934.54 | 2 | 191050.39 | 184196.13 | 167948.30 | 200443.97 |
| **4** | 144372.41 | 118671.85 | 383199.62 | 0 | 182901.99 | 173656.73 | 157983.42 | 189330.03 |
| **5** | 142107.34 | 91391.77 | 366168.42 | 2 | 166187.94 | 172144.10 | 156276.46 | 188011.73 |
| **6** | 131876.90 | 99814.71 | 362861.36 | 0 | 156991.12 | 159424.43 | 143886.88 | 174961.97 |
| **7** | 134615.46 | 147198.87 | 127716.82 | 1 | 156122.51 | 159374.46 | 142116.14 | 176632.79 |
| **8** | 130298.13 | 145530.06 | 323876.68 | 2 | 155752.60 | 159669.34 | 144058.72 | 175279.96 |
| **9** | 120542.52 | 148718.95 | 311613.29 | 0 | 152211.77 | 149541.88 | 134120.27 | 164963.48 |
| **10** | 123334.88 | 108679.17 | 304981.62 | 1 | 149759.96 | 152457.10 | 136980.80 | 167933.40 |
| **11** | 101913.08 | 110594.11 | 229160.95 | 2 | 146121.95 | 131968.80 | 116482.99 | 147454.61 |
| **12** | 100671.96 | 91790.61 | 249744.55 | 1 | 144259.40 | 131745.97 | 116397.39 | 147094.55 |
| **13** | 93863.75 | 127320.38 | 249839.44 | 2 | 141585.52 | 126778.91 | 111521.67 | 142036.15 |
| **14** | 91992.39 | 135495.07 | 252664.93 | 1 | 134307.35 | 125558.04 | 110326.76 | 140789.33 |
| **15** | 119943.24 | 156547.42 | 256512.92 | 2 | 132602.65 | 148554.38 | 132924.73 | 164184.02 |
| **16** | 114523.61 | 122616.84 | 261776.23 | 0 | 129917.04 | 143563.06 | 128087.32 | 159038.80 |
| **17** | 78013.11 | 121597.55 | 264346.06 | 1 | 126992.93 | 116710.71 | 101483.29 | 131938.13 |
| **18** | 94657.16 | 145077.58 | 282574.31 | 0 | 125370.37 | 128015.42 | 112711.02 | 143319.82 |
| **19** | 91749.16 | 114175.79 | 294919.57 | 2 | 124266.90 | 126035.88 | 110661.59 | 141410.17 |
| **20** | 86419.70 | 153514.11 | 0.00 | 0 | 122776.86 | 119024.59 | 101433.91 | 136615.27 |
| **21** | 76253.86 | 113867.30 | 298664.47 | 1 | 118474.03 | 115948.14 | 100463.37 | 131432.91 |
| **22** | 78389.47 | 153773.43 | 299737.29 | 0 | 111313.02 | 117278.04 | 101810.07 | 132746.02 |
| **23** | 73994.56 | 122782.75 | 303319.26 | 2 | 110352.25 | 114517.13 | 98951.90 | 130082.36 |
| **24** | 67532.53 | 105751.03 | 304768.73 | 2 | 108733.99 | 110526.87 | 94812.20 | 126241.54 |
| **25** | 77044.01 | 99281.34 | 140574.81 | 0 | 108552.04 | 111077.05 | 95395.61 | 126758.50 |
| **26** | 64664.71 | 139553.16 | 137962.62 | 1 | 107404.34 | 103342.40 | 87948.15 | 118736.64 |
| **27** | 75328.87 | 144135.98 | 134050.07 | 2 | 105733.54 | 109766.65 | 94092.23 | 125441.06 |
| **28** | 72107.60 | 127864.55 | 353183.81 | 0 | 105008.31 | 111588.98 | 94942.16 | 128235.79 |
| **29** | 66051.52 | 182645.56 | 118148.20 | 2 | 103282.38 | 103540.22 | 88030.65 | 119049.80 |
| **30** | 65605.48 | 153032.06 | 107138.38 | 0 | 101004.64 | 102990.21 | 87451.52 | 118528.90 |
| **31** | 61994.48 | 115641.28 | 91131.24 | 2 | 99937.59 | 100500.13 | 84986.54 | 116013.71 |
| **32** | 61136.38 | 152701.92 | 88218.23 | 0 | 97483.56 | 99939.98 | 84431.44 | 115448.52 |
| **33** | 63408.86 | 129219.61 | 46085.25 | 1 | 97427.84 | 101803.87 | 85966.93 | 117640.80 |
| **34** | 55493.95 | 103057.49 | 214634.81 | 2 | 96778.92 | 101018.11 | 85608.65 | 116427.58 |
| **35** | 46426.07 | 157693.92 | 210797.67 | 1 | 96712.80 | 94879.11 | 79297.41 | 110460.81 |
| **36** | 46014.02 | 85047.44 | 205517.64 | 0 | 96479.51 | 94360.40 | 78810.89 | 109909.91 |
| **37** | 28663.76 | 127056.21 | 201126.82 | 2 | 90708.19 | 81082.65 | 65426.93 | 96738.37 |
| **38** | 44069.95 | 51283.14 | 197029.42 | 1 | 89949.14 | 92626.70 | 77104.26 | 108149.15 |
| **39** | 20229.59 | 65947.93 | 185265.10 | 0 | 81229.06 | 72866.55 | 57174.11 | 88558.99 |
| **40** | 38558.51 | 82982.09 | 174999.30 | 1 | 81005.76 | 87640.50 | 72174.61 | 103106.38 |
| **41** | 28754.33 | 118546.05 | 172795.67 | 1 | 78239.91 | 79839.16 | 64316.93 | 95361.39 |
| **42** | 27892.92 | 84710.77 | 164470.71 | 2 | 77798.83 | 78736.03 | 63227.44 | 94244.63 |
| **43** | 23640.93 | 96189.63 | 148001.11 | 1 | 71498.49 | 74326.42 | 58761.04 | 89891.81 |
| **44** | 15505.73 | 127382.30 | 35534.17 | 0 | 69758.98 | 65845.21 | 50151.86 | 81538.55 |
| **45** | 22177.74 | 154806.14 | 28334.72 | 1 | 65200.33 | 72601.04 | 56618.99 | 88583.08 |
| **46** | 1000.23 | 124153.04 | 1903.93 | 0 | 64926.08 | 51773.34 | 34783.00 | 68763.69 |
| **47** | 1315.46 | 115816.21 | 297114.46 | 2 | 49490.75 | 55758.69 | 37924.19 | 73593.19 |
| **48** | 0.00 | 135426.92 | 0.00 | 1 | 42559.73 | 50686.25 | 33537.08 | 67835.41 |

Showing 1 to 15 of 48 entries, 8 total columns