DIC

PROJECT 1 : PROBLEM 4

LINEAR REGRESSION USING REAL DIRECT DATA

NAME : ANIKET THIGALE
UB NO : 50168090

PROBLEM DESCRIPTION

Write a R script to clean data, perform EDA and analyse the RealDirect data set to find some insights and make recommendations.

DATA

We use the Real Direct data used in problem 3. The data is present in the XLS format. There are 21 variables like borough,neighborhood, block, lot etc. per record.Use perl to import the data to R.

CLEANING AND ADDING FEATURES TO DATA

1. Created a new variable, sale_price_n, to remove the $ from sale.price
2. Make land.square.feet, year.built and gross.square.feet numeric
3. Make sale.date as date in R
4. Remove outlier data for eg, data with sale.price as 0, gross.square.feet=0, where neighborhood is not mentioned, no zipcodes, year.built is 0 etc. to finally get a full dataset.

BUILDING A MULTIPLE LINEAR REGRESSION MODEL

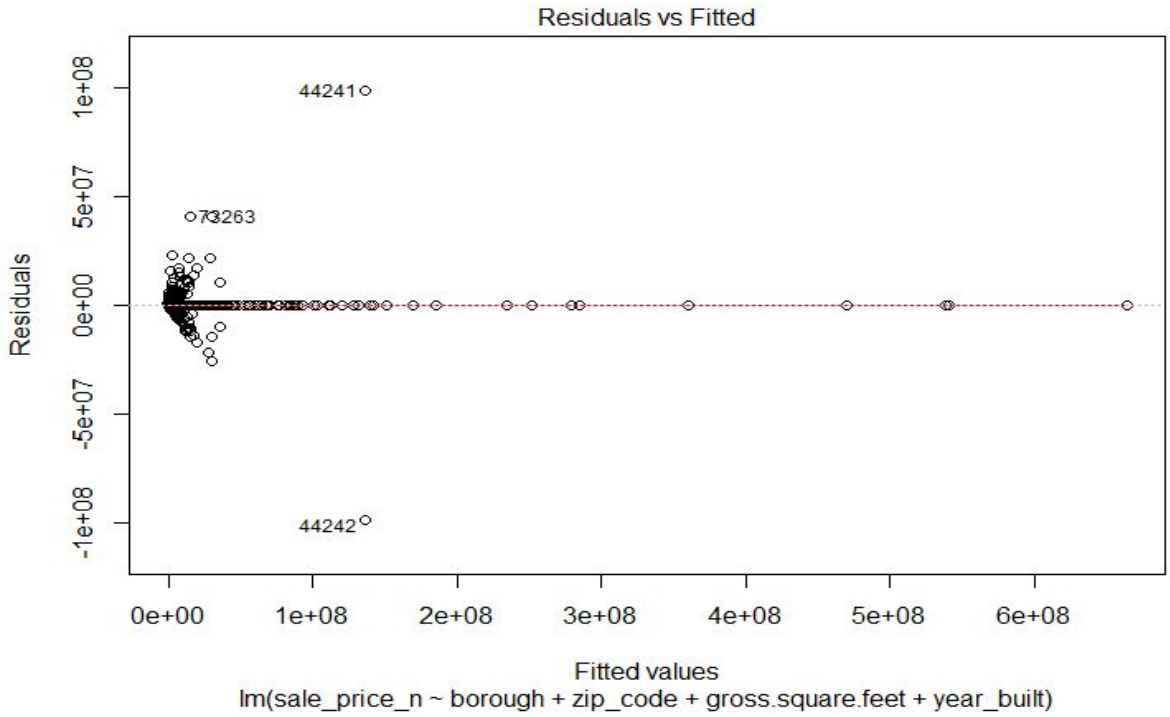
First pick a random subset for the model to train on.Then use the lm package available for R to build a prediction model for predicting sale prices based on :

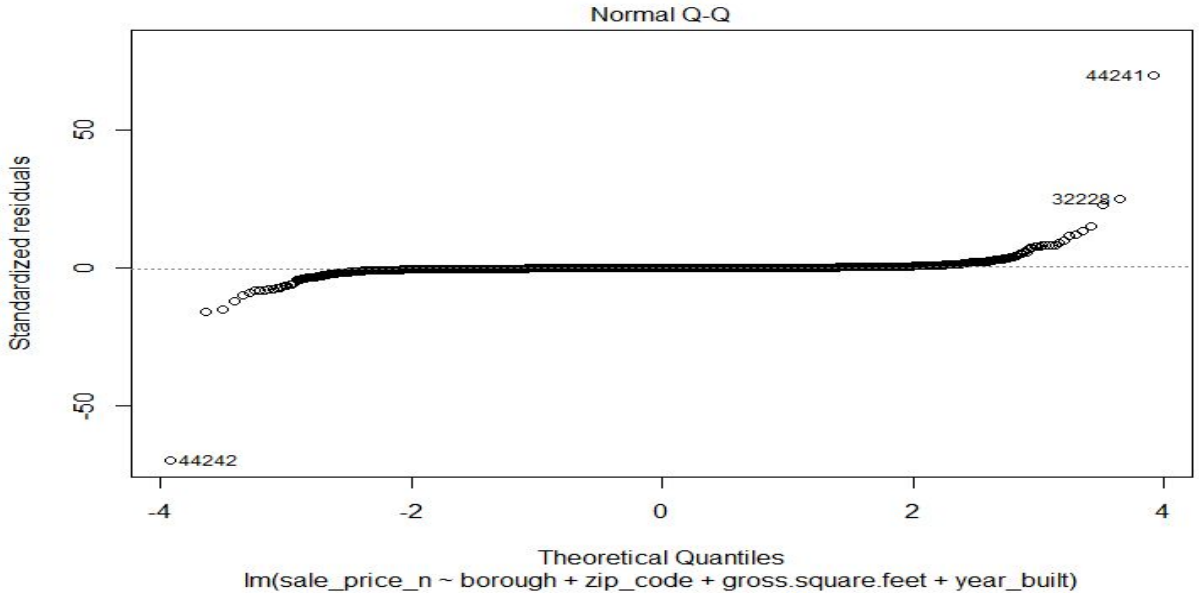
year_built, zipcode, borough, gross.square.feet

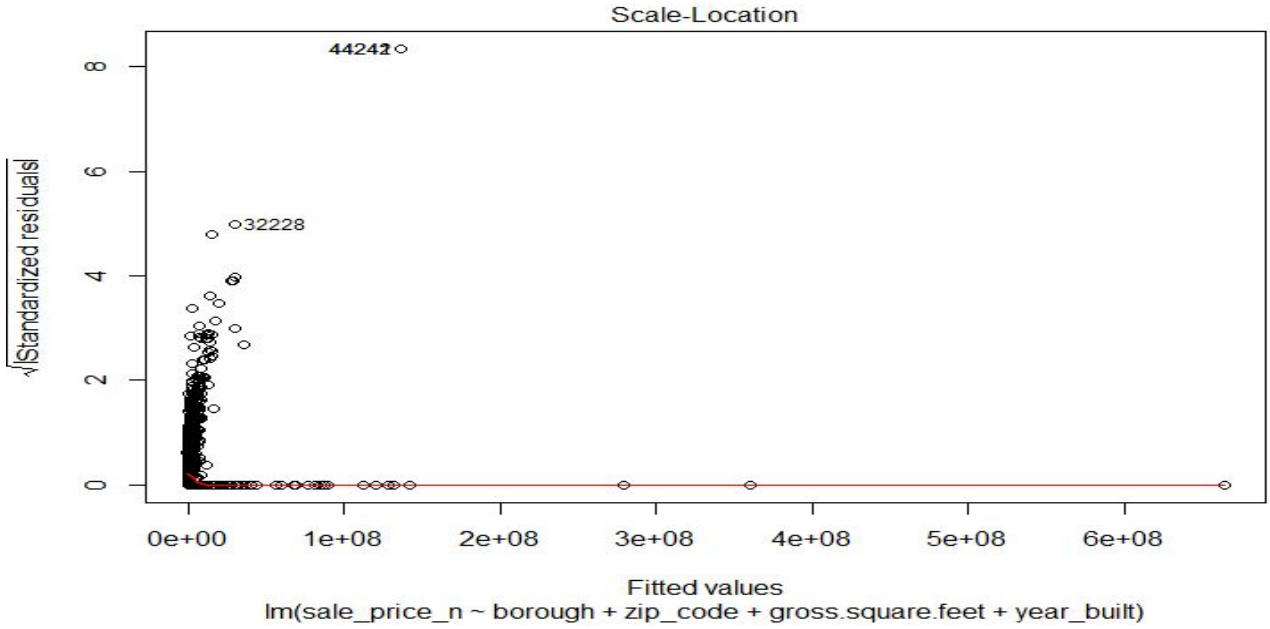
Then we can use the predict() function to predict sale prices

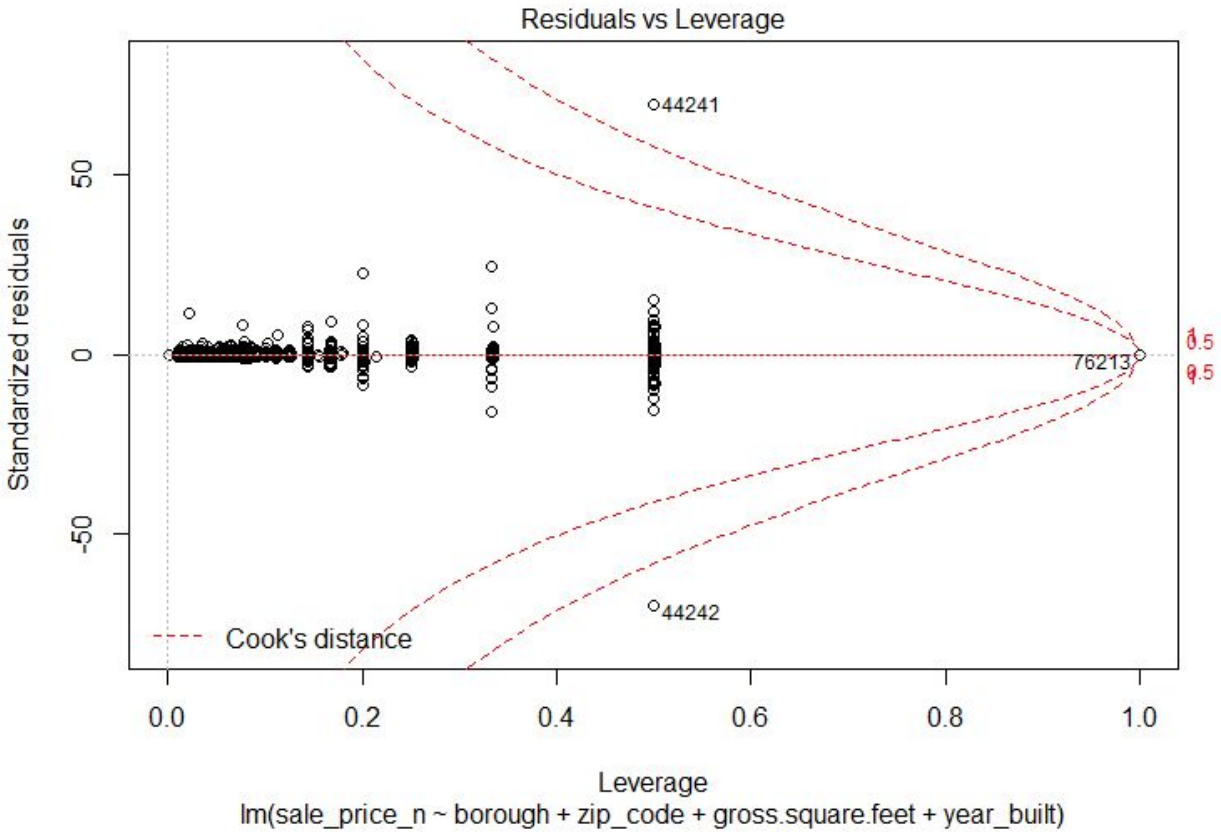
```
> predict(threePredictorModel, bk_testdata, interval="predict")
            fit       lwr     upr
77143  351787.79 -3656733 4360309
84931  945000.00 -4617709 6507709
14459  596038.42 -3372504 4564581
75514 1500000.00 -4062709 7062709
34372 3697159.21 -1120288 8514607
3170   652438.59 -3300286 4605163
69855  347480.68 -4470049 5165010
75185  352422.68 -3618856 4323701
43424   68326.56 -3868492 4005146
43800   68326.56 -3868492 4005146
>
```

PLOTS




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(sale_price_n ~ borough + zip_code + gross.square.feet + year_built)
44241
32228
44242


Scale-Location
√|Standardized residuals|
Fitted values
lm(sale_price_n ~ borough + zip_code + gross.square.feet + year_built)
32228



CONCLUSION

Thus, we have analyzed and identified interesting facts in the Real Direct Dataset