Prediction of Housing Sales Price in King County, USA

General Information

Dataset : kc_house_data from Kaggle Datasets
Langauge : R
Predictive Modelling : Linear Regression

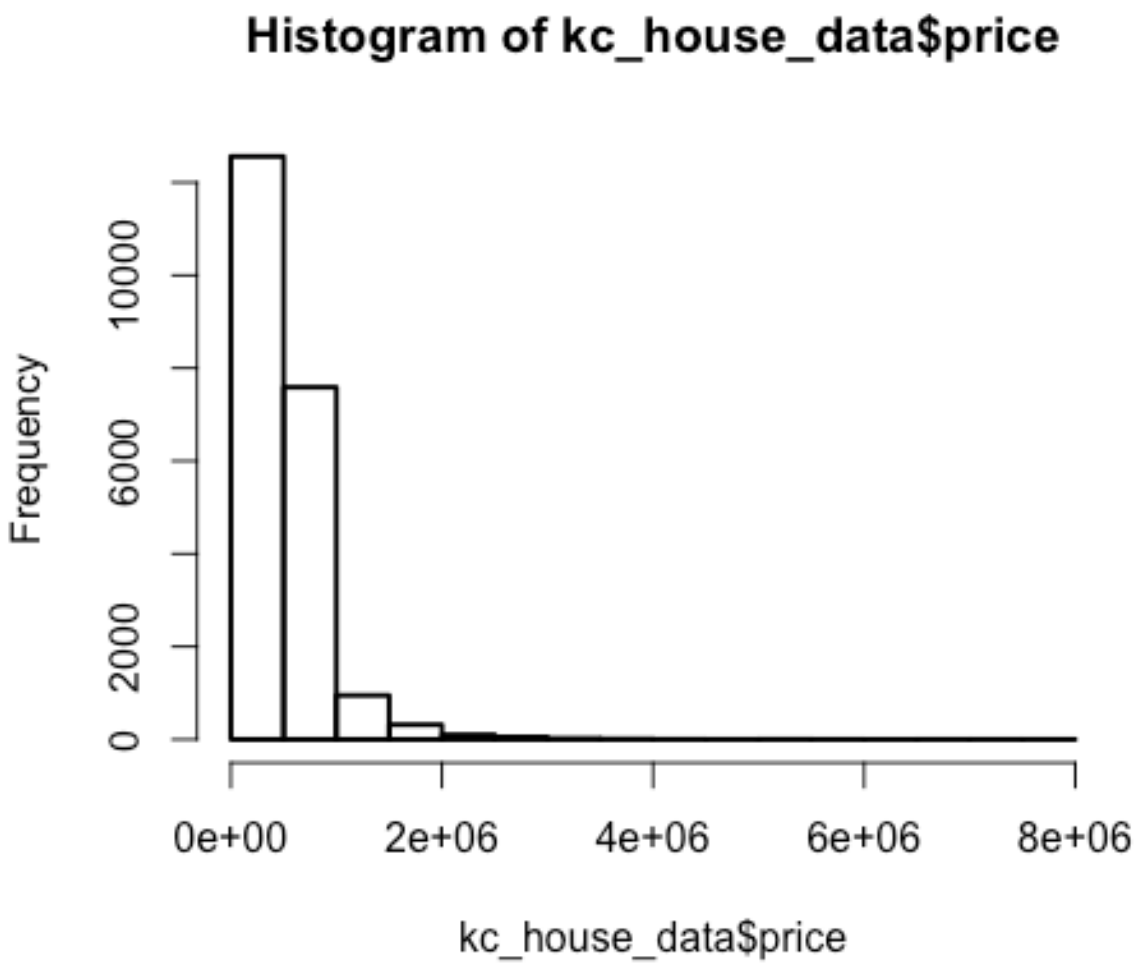
Problem Statement : Predicting House Sales Prices in King County, US using past housing sales data.

Data Understanding and Data Sanitisation :

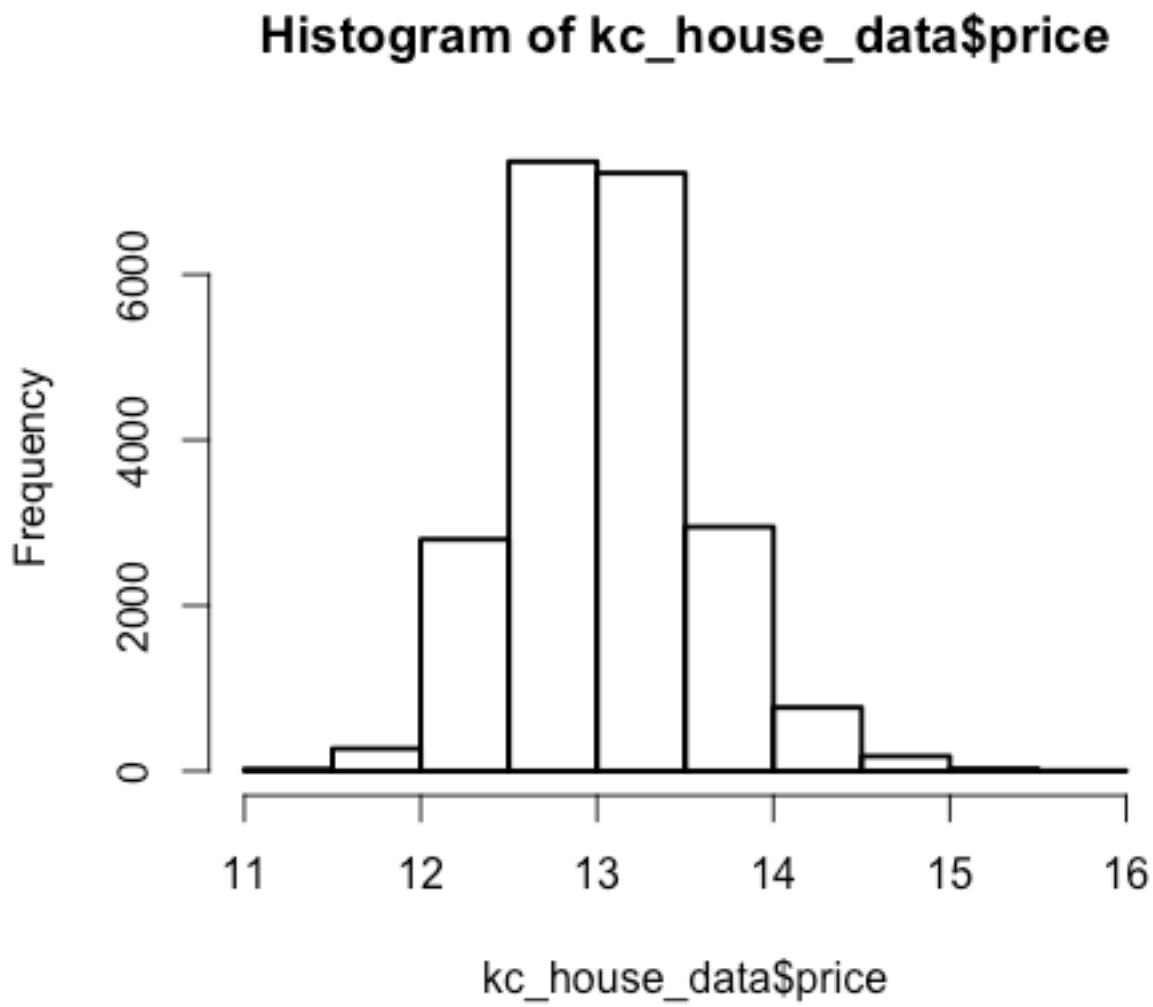
kc_house_data dataset contains 21613 observations of 21 Variables
No NA values found in Dataset

Histogram plot of the Dataset reveals that the dataset is highly skewed, hence log transformation of price variable is done to eliminate the skewness in data.

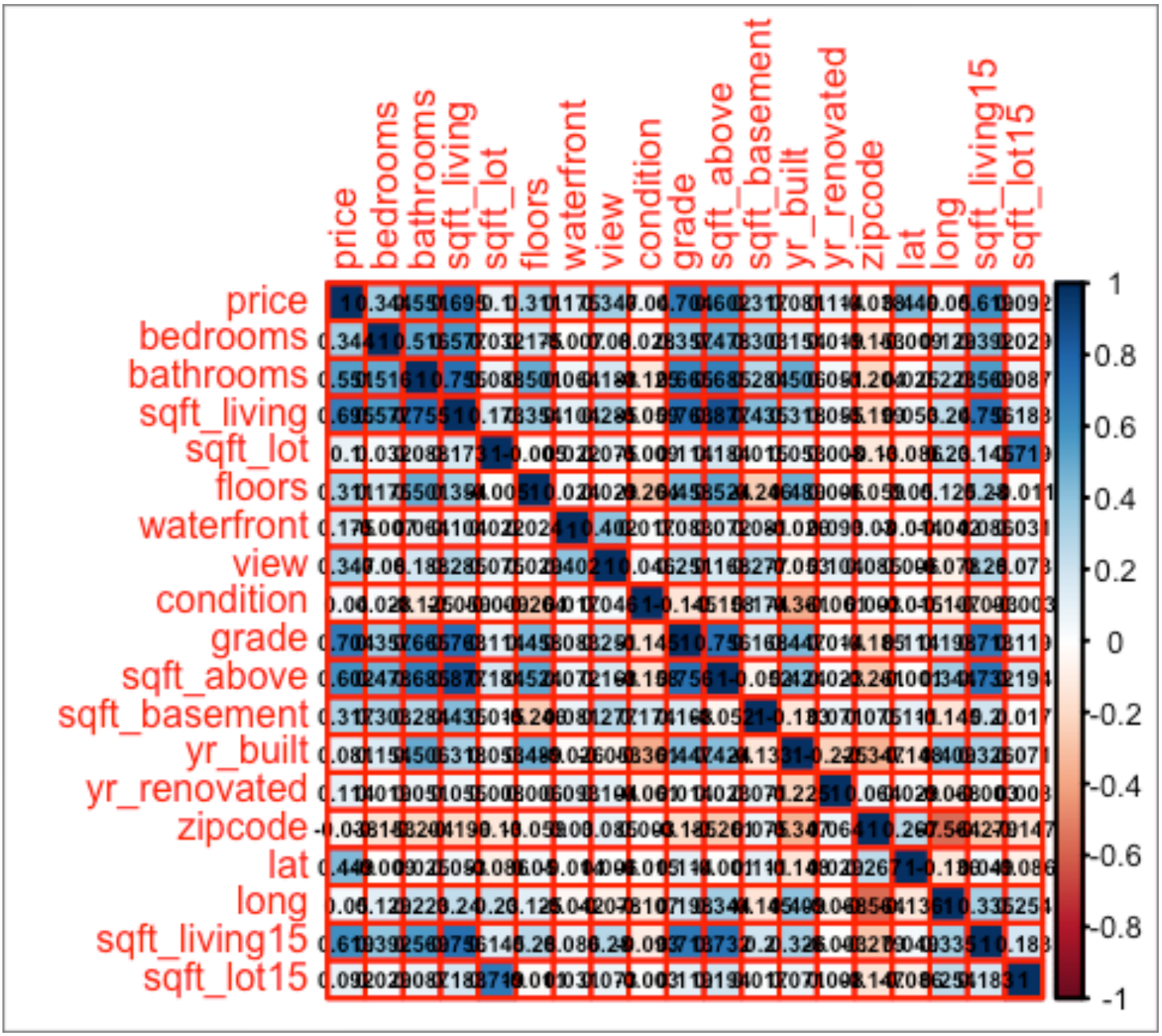
Pre Log Transform Histogram



Post Log Transformation Histogram



Correlation Matrix



Columns with less correlation value :

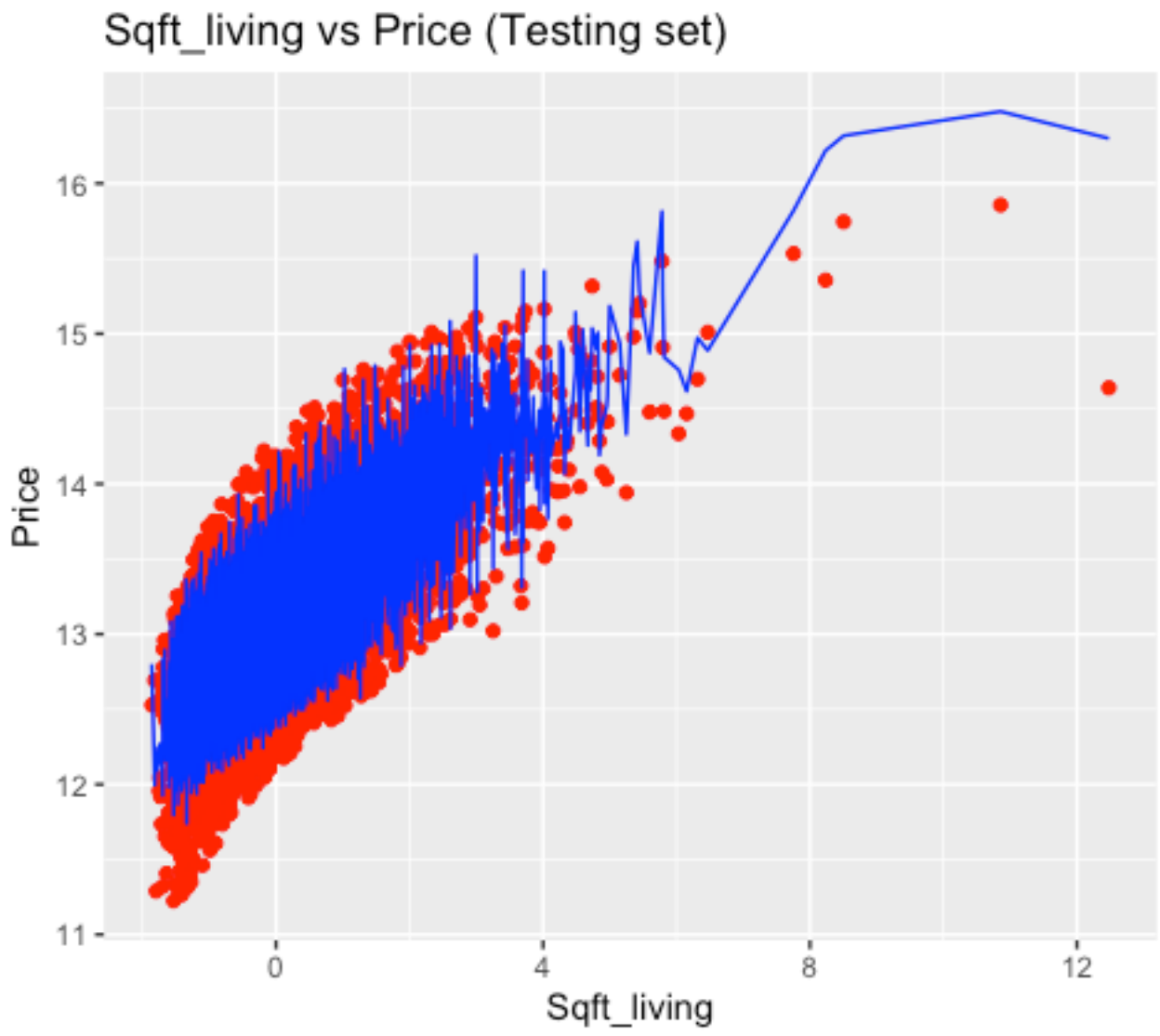
date
sqft_lot
sqft_lot15
yr_built
lat
long

Multicollienarity Testing

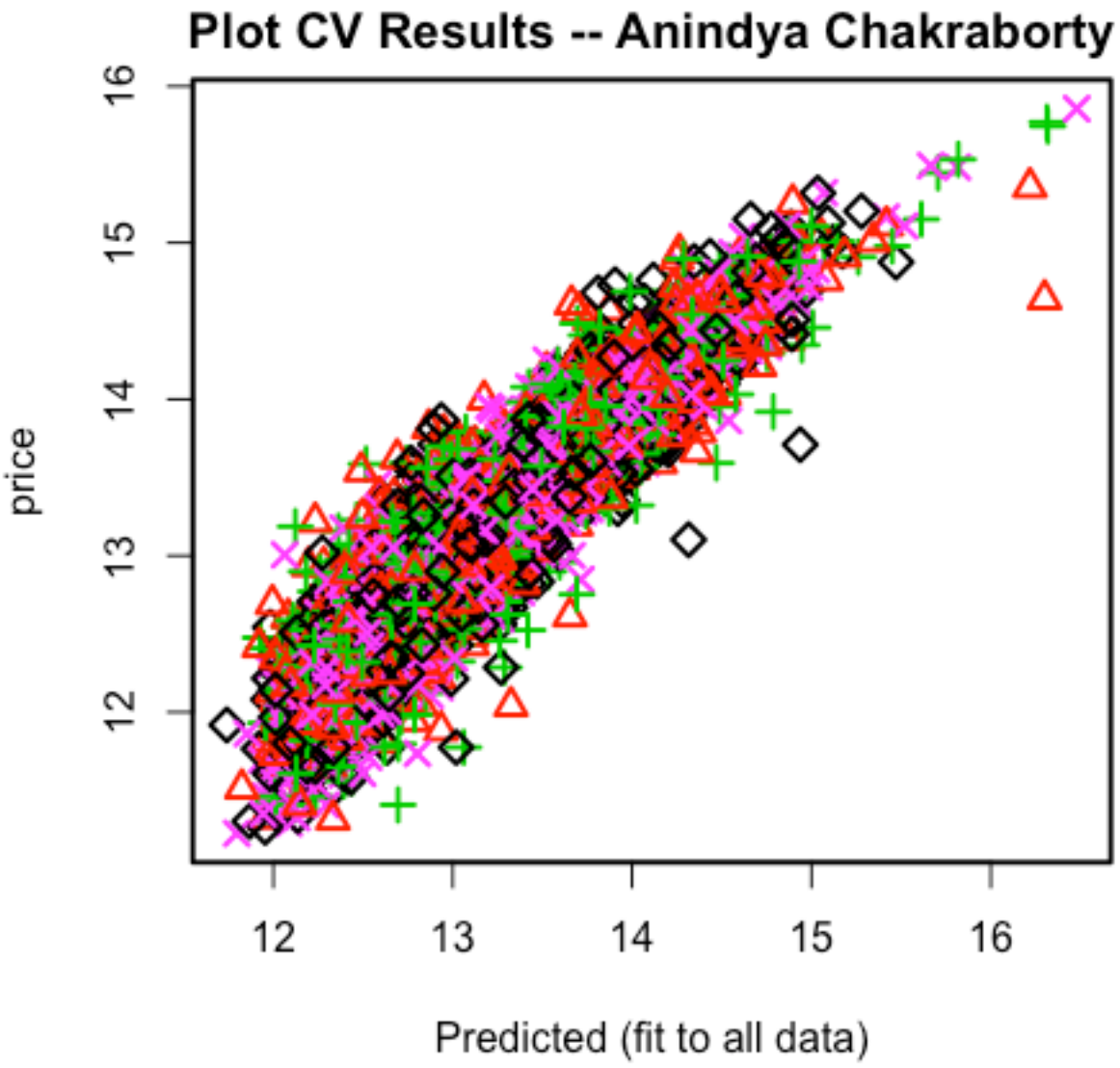
Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly linearly related to each other apart from being related to the predictive variable.
Multicollienarity is high for **sqft_above** and **sqft_living** wrt **price** variable. However correlation of **sqft_living** is more that **sqft_above**, and hence keeping **sqf_living** variable for model building. Rest all variable were found to be under VIF = 4, and hence have been considered for the analysis.

Predictive Modelling
Linear Regression was done on the dataset for predicting the housing sales prices.
Adjusted R squared value is 0.8726 ~ 87%
Root Mean Square Error (RMSE) : 0.1835

Model Fit on Testing Data :



K-Fold Validation :



Anindya Chakraborty
Electronics Engineering , BBDNITM (Lucknow)
PGDDS, IIIT Bangalore