REPORT ON
HOUSE PRICE PREDICTION MODEL
USING LINEAR REGRESSION TECHNIQUE

DONE BY

- RAJ KUMAR T (E20023)

REPORT ON KC_HOUSE PRICE PREDICTION

Problem Statement:

The corporation wants to evaluate the property taxes (of about 10% of market price) for every individual land owner based on their location and area of the property. They have a collected database for some houses with all the different features along with prices. They wanted to predict the closest market prices for the rest of the houses.

Data Description:

- Dataset was split into training, test and validation test as given below:

File name	**Type of Data**	**No. of Rows**
wk3_kc_house_train_data.csv	Training	9761
wk3_kc_house_valid_data.csv	Validation	9635
wk3_kc_house_test_data.csv	Test Data	9761

- The training dataset has 9761 rows and 21 columns.
- The independent variables for predicting price available in the dataset are listed below as per their scales:

Nominal	**Ordinal**	**Ratio**
Id	Bedrooms	sqft_living
Date	bathrooms	sqft_lot
zipcode	floors	sqft_above
Lat	condition	sqft_basement
Long	grade	sqft_living15
yr_built	view	sqft_lot15
yr_renovated	Waterfront	

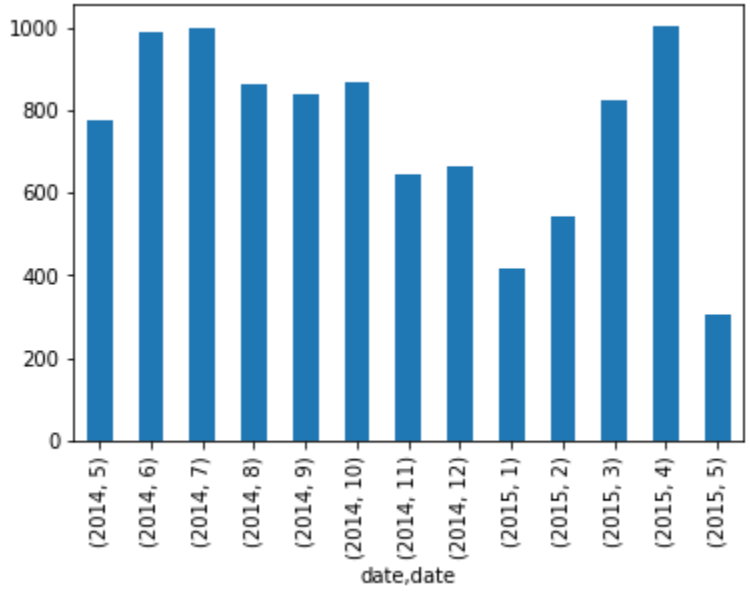
- There are no missing values in the dataset.
- For the purpose of readability, the date column has been converted to date format.
- The dataset has recordings of the sales of different houses in King's County from the time period of May 2014 to May 2015.

Exploratory Data Analysis(EDA) :

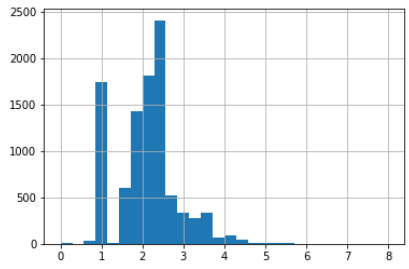
Univariate Analysis on possibly influencing variables:

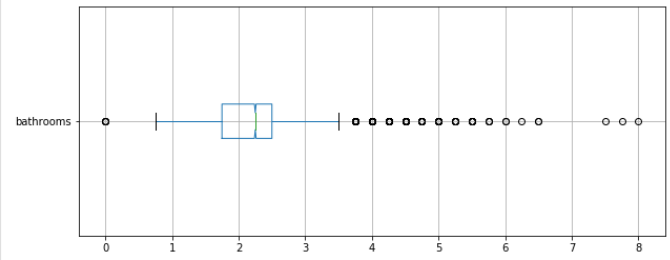
ID-(Identification no. for houses): There are 41 repetition of id's within the training dataset. On examination, the price with date has been updated. Hence the id with the last date provides us the final record. Hence those repetitive values have been deleted.

Date (Date of Sale of houses): We are trying to find if any pattern exists between months. But there seems to be no trend with respect to date.

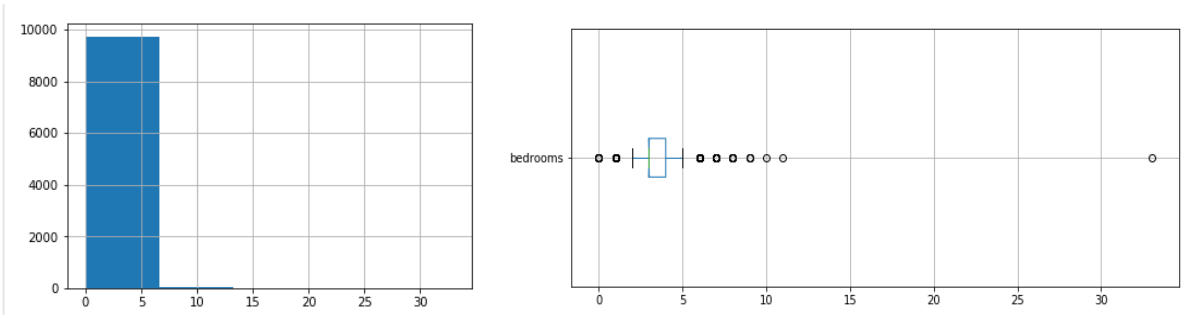


Bathrooms (No of Bathrooms): There are 28 unique values of bathrooms and the distribution is skewed towards right side. There are around 252 outliers on the upper limit.

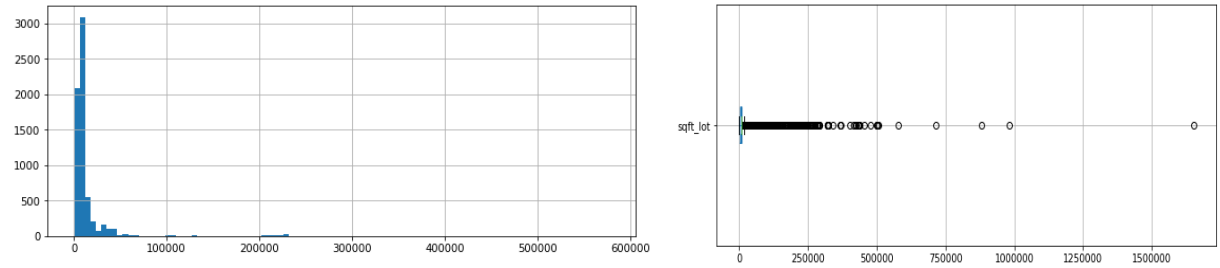




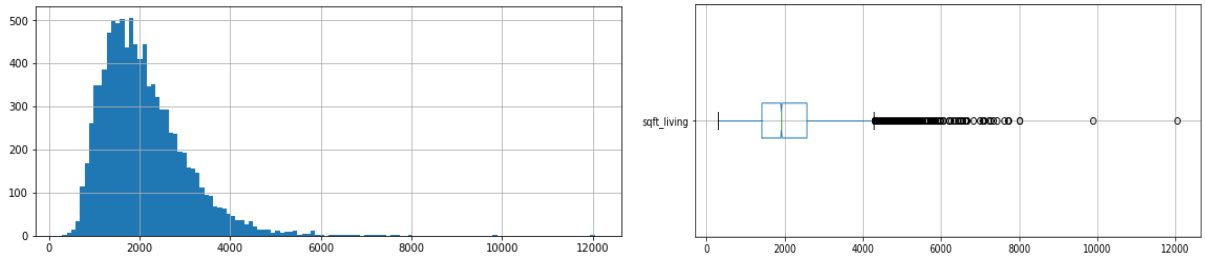
Bedrooms (No of Bedrooms): There are 13 unique values of bedrooms and the distribution is skewed towards right side. There seems to be anomaly with 33 bedrooms which is a recording error. Hence the record is removed. There are 94 lower and 145 upper outliers.



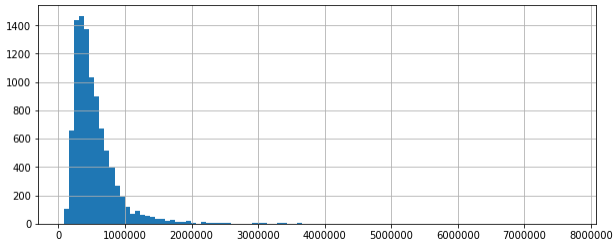
Sqft_lot (Sqft area of the lot): This variable is skewed to the right as well. There are 1085 upper outliers.

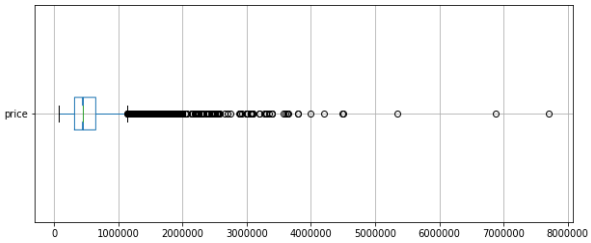


Sqft_living (Sqft area of the living space): This variable is skewed to the right as well. There are 240 upper outliers.



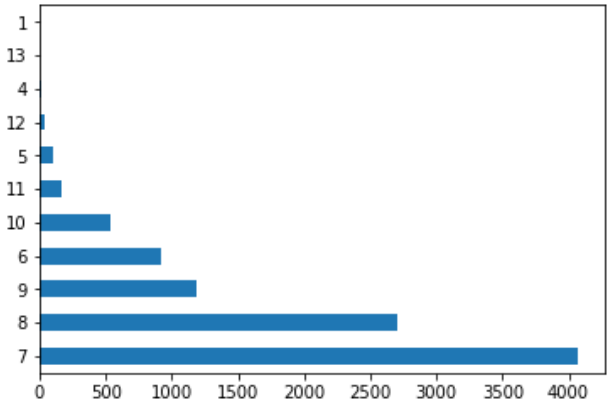
Price (Price of the plots): Price seems to be skewed as well to the right.





All these continuous variables seem to be skewed to the right defining that there are certain plots with bigger areas thereby resulting in greater price. Hence even if individually examined and these points appear to be outlier, they seems to be linearly increasing with one another.

Grade (Grade of the property): There are 11 different grades available and has been classified on their basis.



Building_Age (Computed from yr_built till year 2014): This variable was computed to find if any relationship exist between price and building_age.

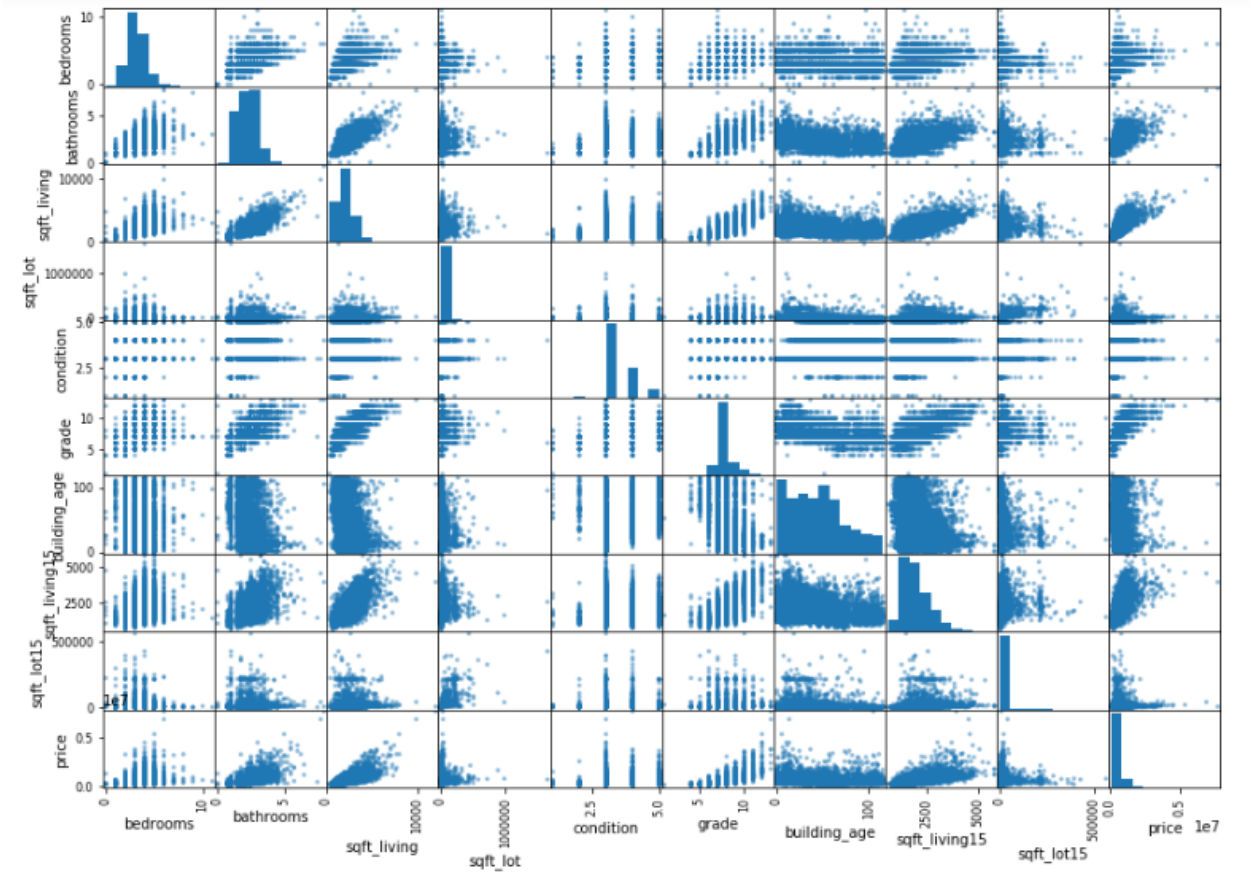
Zip-Code: It has 70 unique values and can be possibly used with one-hot encoding technique to find out if influences the price.

The other continuous variables like Sqft_living15, Sqft_lot15 can be ignored as their definition is unknown.

Observations & Alterations from Univariate Analysis: 1. Duplicate ID values were eliminated with the latest dates. 2.All continuous variables are skewed to the right. 3.Though most of the variables have considerable outliers in univariate analysis, they seem to be roughly in accordance with the behaviour of other variables. 4.Building Age was computed with Yr_built variable.

Multivariate Analysis:

Plotting a scatter matrix to find the linear relationship between the variables:



1.Only sqft_living, grade and bathrooms seems to have some linear relationship with price. We'll analyze it further.

2.Other than this, bathrooms and sqft_living has some linearity (This should be noted as it might result in multi-collinearity within these variables.)

Finding Correlation of variables with price:

1.Grade, Sqft_living and bathroom seem to have higher correlation with the price. The other variables seem to have lesser correlation but still can be input into the models to check if they increase the accuracy.

2.Sum of Sqft_above and sqft_basement results in sqft_living. Hence all the three cannot be use in a single model as it might lead to multi-collinearity.

Model Building:

The model is now built on training dataset and tested on validation dataset. Different models are tried to improve their efficiency and accuracy.

Model No:	Description	Predictor Variables	RMSE	R-Squared value(%)
1	High correlation variables	sqft_living, grade, bathrooms	242771	53.53
2	All continuous features	sqft_living, sqft_lot, sqft_above	255044	48.72
3	Including building_age in previous model	sqft_living, sqft_lot, sqft_above, building_age	247423	51.74
4	Combination of best variables	sqft_living, grade, view, waterfront	228554	58.82
5	Including zipcode through one hot encoding.	sqft_living, grade, view, waterfront, encoded zip-code	163028	79.02

Testing the best model:

The final model of selection (with attributes of sqft_living, grade, view, waterfront, encoded zip-code) is put through test data to find its accuracy. It obtains a result of

RMSE	R-squared value(%)
158323	80