Project Report
House Price Prediction

Rohit K. Umredkar
Praxis Business School – rohit.umredkar@praxis.ac.in

Chapter 1

Introduction

Housing prices are an important reflection for real estate groups and housing price ranges are of great interest for both buyers and sellers. In this assignment house prices will be predicted given explanatory variables that cover many aspects of residential houses. As continuous house prices, they will be predicted with various regression techniques including Simple Linear Regression, Multiple Linear Regression and Decision Tree Regressor etc. The goal of this project is to create a regression model that can accurately estimate the price of the house with given features.

Problem Statement

Real Estate online portal like 99acres, commonfloor, magicbricks where properties are listed for sell/buy purposes. But there are lot of inconsistence in term of house prices, like same type of houses but different prices. This leads to diminish trust and decrease in transparency between buyer/seller and real estate groups. Again, there is no way for buyer/seller to know house prices when they listed their house details on portal.

Objective

- To build trust and increase transparency between buyer/seller and real estate groups, as proper & justified house prices can help.
- To help buyer/seller to know house prices when they list their house details on portal.

Chapter 2

Exploratory Data Analysis

Getting the Data & Previous Preprocess

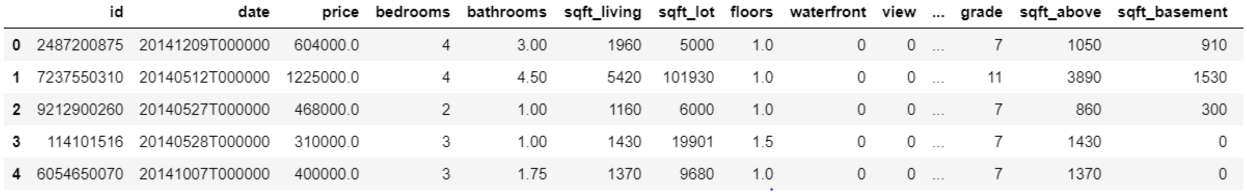
The dataset used in this project comes from the Kaggle Repository. This data was collected in 2014-2015. The dataset is split into training, validation and Test dataset, in training dataset each of the 9761 entries represents aggregate information about 19 features of homes from various suburbs located in Washington U.S. Further analysis performed on Training data.

The features can be summarized as follows:

- id – it identifies each observation in the dataset.
- date – This is the date when observation is entered in the dataset.
- price – This is the price of house in US$.
- bedrooms – No of bedrooms available in house.
- bathrooms - No of bathrooms available in house.
- sqft_living – Carpet area of house.
- sqft_lot – Courtyard area of house.
- floors – No of floor available in house.

- waterfront – whether house have waterfront or not (if yes then 1 else 0).
- view – How much good view house have. (min 0 to max 4)
- condition – how is the condition of house. (min 1 to max 5)
- grade – Rating of house. (min 1 to max 13)
- sqft_above – Carpet area of house except basement area.
- sqft_basement – Basement area of house.
- yr_built – Year when house is built.
- yr_renovated – Year when house is renovated.
- Zipcode – This shows locality of house.
- lat – This is exact geo location of house.
- long - This is exact geo location of house.

This is an overview of the original dataset, with its original features:

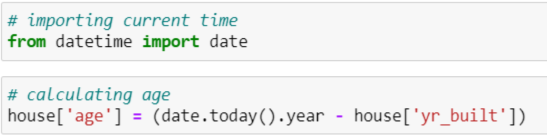


	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sqft_above	sqft_basement
0	2487200875	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	0	...	7	1050	910
1	7237550310	20140512T000000	1225000.0	4	4.50	5420	101930	1.0	0	0	...	11	3890	1530
2	9212900260	20140527T000000	468000.0	2	1.00	1160	6000	1.0	0	0	...	7	860	300
3	114101516	20140528T000000	310000.0	3	1.00	1430	19901	1.5	0	0	...	7	1430	0
4	6054650070	20141007T000000	400000.0	3	1.75	1370	9680	1.0	0	0	...	7	1370	0

For the purpose of the project: the dataset has been preprocessed as follows:

- The unnecessary features have been excluded for the project are: 'id, 'date, 'lat' and 'long'. The remaining features are essential for this project.
- No Missing value is observed in this dataset.
- The yr_built variable is transformed into age of house, and yr_renovated into renovated_age.

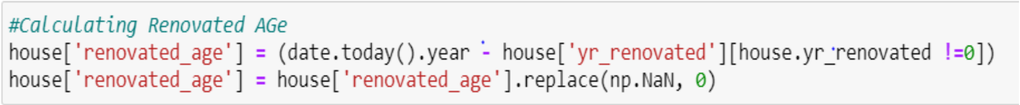
  This is explained as below:
  For age of house



```
# importing current time
from datetime import date
```

```
# calculating age
house['age'] = (date.today().year - house['yr_built'])
```

For renovated age of house



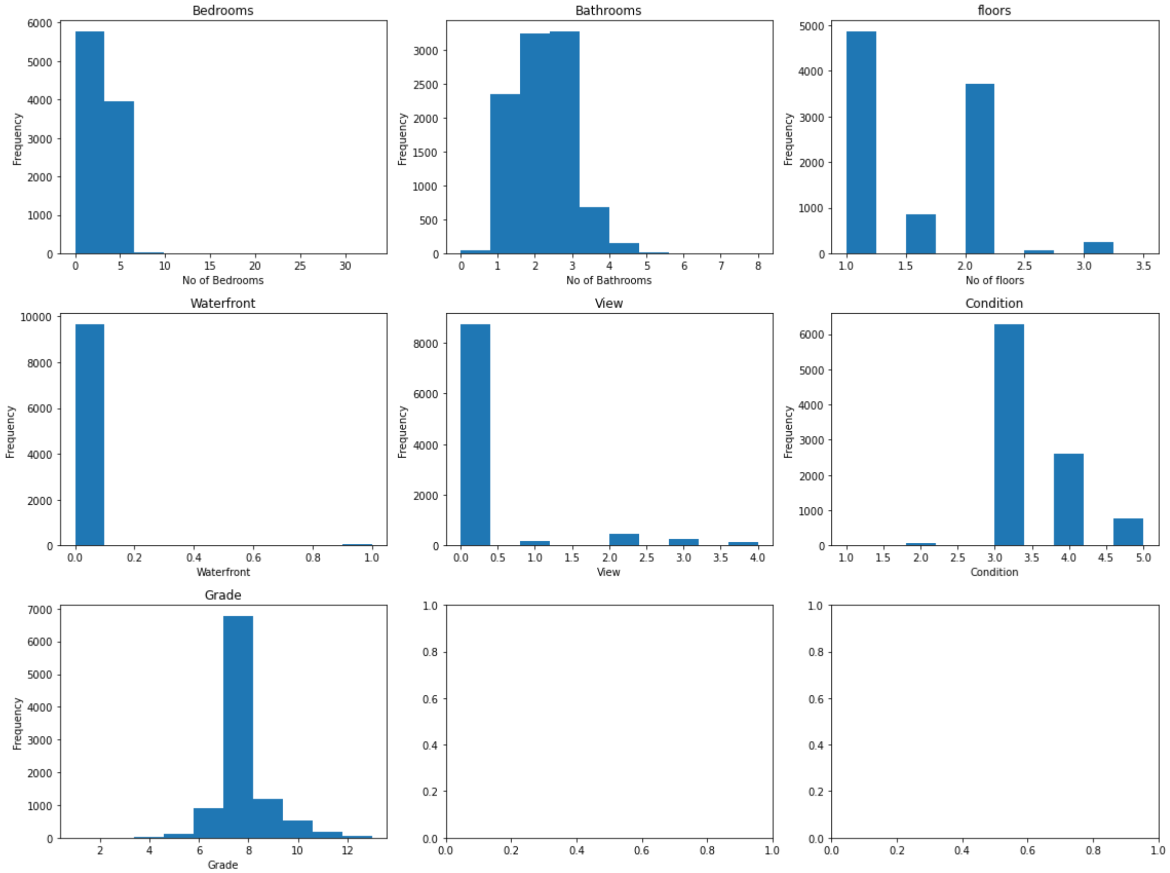
```
#Calculating Renovated AGe
house['renovated_age'] = (date.today().year - house['yr_renovated'][house.yr_renovated !=0])
house['renovated_age'] = house['renovated_age'].replace(np.NaN, 0)
```

Univariate Analysis

I segregate variable into discrete and continuous numerical for ease in understanding insights.

Discrete variables are like bedrooms, bathrooms, floors, waterfront, view, condition and grade are plotted using matplotlib library subplot function.

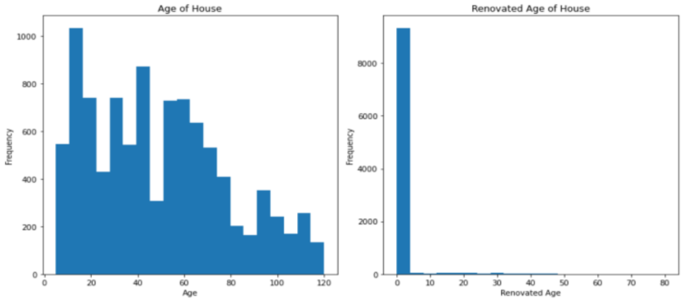
The graph is shown below:



All variables have skewed distribution with long right tail except grade and condition. I observed house have 33 bedrooms with 1.75 bathrooms, this must be anomalies and need to be corrected.

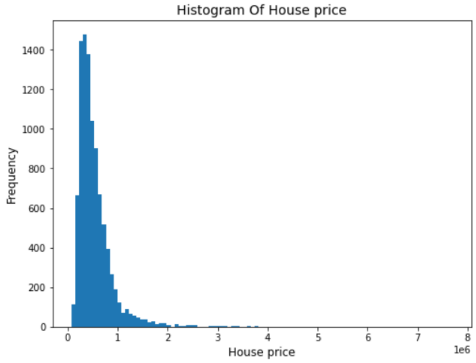
Continuous variables are like age, renovated age, price, sqft_living, sqft_lot, sqft_basement, sqft_above plotted as follows.

Age and renovated age of house



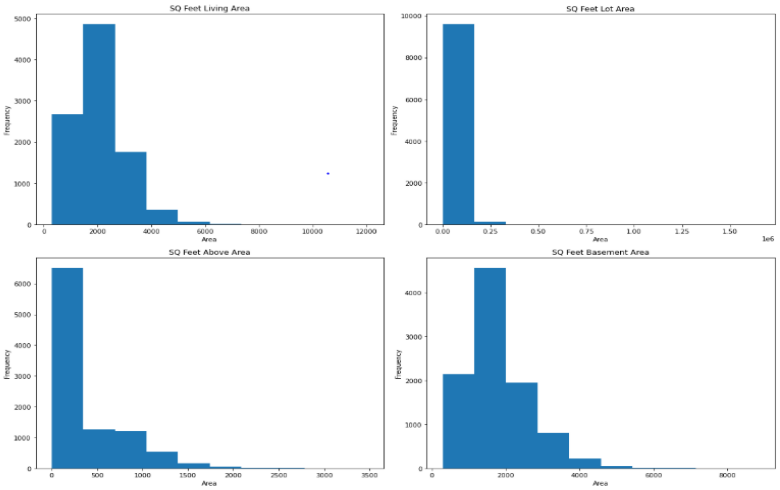
Age of house have skewed distribution with right tail and most of the house (9340 out of 9671) is not renovated.

Price



Price have highly skewed distribution with long right tail.

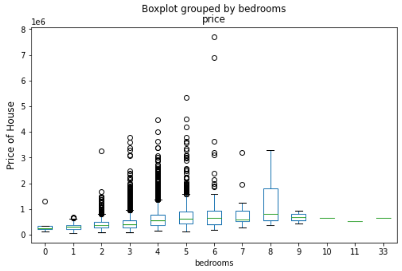
Area of House

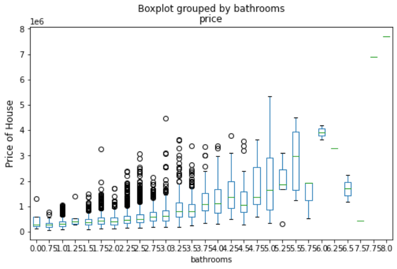


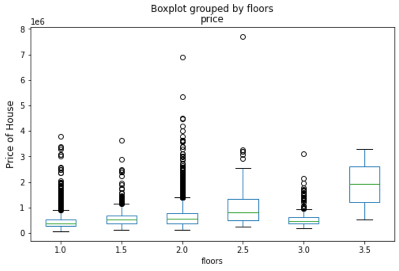
Square foot living, lot, above and basement areas have highly skewed distribution with long right tail as observed in above figure.

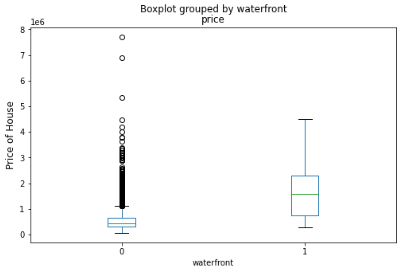
Bivariate Analysis

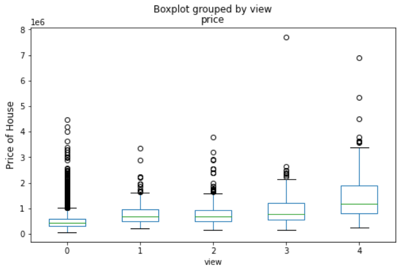
Boxplot for bedrooms, bathrooms, floors, waterfront, view, condition and grade

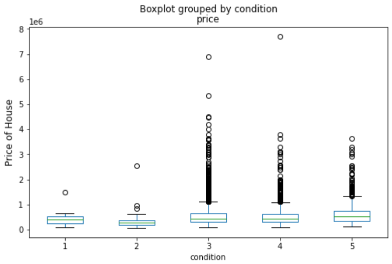


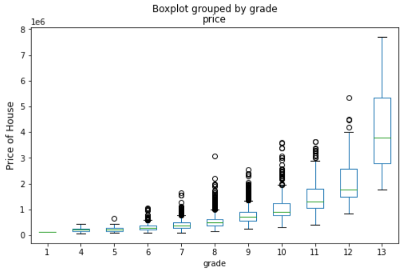






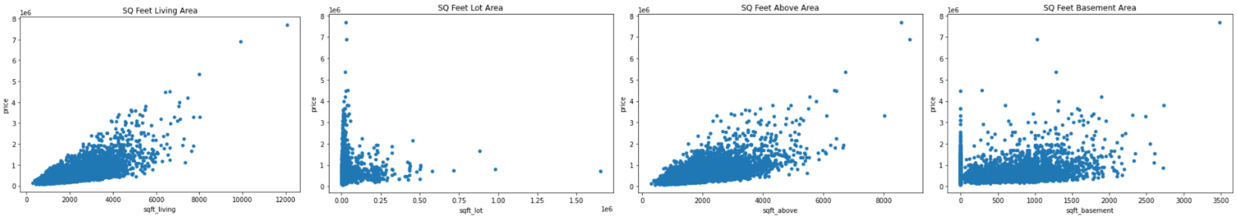






From above side by side box plot we can say that variables like bedrooms, floors, views, condition do not have significant effect on house price, whereas variable like waterfront and bathrooms have some effect on house price, again house price is exponentially increasing with increase in house grade.

Scatter plot of House area



From above scatter we can say that house price is gradually increasing with increase in living, above and basement area except sqft_lot area, and from data we can also say that house living area is divided into sqft_above and sqft_basement.

Chapter 3

Model Building

In this section of the project, we will develop the tools and techniques necessary for a model to make a prediction. Being able to make accurate evaluations of each model's performance using these tools and techniques helps to reinforce greatly the confidence in the prediction

Model 1

Decision Tree Regressor

My goal is to solve the regression problem where the target variable is the price and the independent variables are bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, age and zipcode as categorical variable. The model is then used to predict house prices with given features in validation dataset same as training dataset and is compared to the actual house prices of houses given in validation dataset and I got results as follows.

R2: 0.69

RSME: 198302.84

MAPE: 24.38 %

Accuracy: 75.62 %

Model 2

House Price Log Transformation

Then I used log transformation house price column, keeping rest features same as previous No significant improvement observed as shown below.

R2: 0.66

RSME: 205128.15

MAPE: 23.28 %

Accuracy: 76.72 %

Model 3

Linear Regression

Linear regression is a natural choice of baseline model for regression problems. So, I try to solve the following Problem given a processed list of features for a house, we would like to predict its house prices. Initially used features are are bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, age. The model is then used to predict house prices with given features in validation dataset same as training dataset and is compared to the actual house prices of houses given in validation dataset and No significant improvement observed as shown below.

R2: 0.657
RSME: 210105.99
MAPE: 29.44 %
Accuracy: 70.56 %

Model 4

Now I include zipcode in features as house prices also depend on locality of house by using dummies function in pandas as zipcode is categorical type and using it as a numerical variable make no sense. Again, few features has been transformed like log transformation for house price and sqft_living. Square transformation for bedrooms and bathrooms. Condition and sqft_lot has dropped as it has low co-relation coefficient, keeping rest of features same as previous, After predicting house price using features in validation dataset, I compared predicted price value with actual house prices in validation dataset and significant improvement has been observed as shown below.

R2: 0.875
RSME: 128050.22
MAPE: 14.35 %
Accuracy: 85.65 %

Model 5

We have achieved good RSME & MAPE in model 4 with 9 features but with large amount of features model may get over fitted, to overcome overfitting no of features need to be reduced. I proceed futher with only 5 features that are square of bathrooms, sqft_living_log, waterfront, grade, zipcode with target as price_log. I choose to include square of bathrooms instead bedrooms as it has higher co-relation coefficient than bedrooms with house price. The model is then used to predict house prices with features in validation dataset same as in training dataset and is compared to the actual house prices given in validation dataset, results as shown below.

R2: 0.863
RSME: 135328.366
MAPE: 15.06 %
Accuracy: 84.94 %

Testing

Even after dropping 4 features there is no significant dropped in R2, RSME, MAPE has been observed. Thus, it is highly recommended to used model 5 for testing. The model is then used to predict house prices with features in test dataset same as in training dataset and is compared to the actual house prices given in test dataset, results as shown below.

RSME: 137562.08
MAPE: 15.32 %
Accuracy: 84.68 %

As results RSME, MAPE is obtained almost similar for validation and testing for model 5 thus we can say that model 5 is stable and can predict unknown dataset having only 5 same features used in model 5 with higher accuracy.

Conclusion

The best performing model is Linear Regression with features log of square feet living area, square of bathrooms, waterfront, grade and zipcode as dummies with price_log as target varaible used in model 5 with RSME 137562.08 and MAPE 15.32 %.
According to my analysis, square feet living area, bathrooms, waterfront, grade and zipcode have the greatest statistical significance in predicting a house price.

References

PropTech for Proactive Pricing of Houses in Classified Advertisements in the Indian Real Estate Market by S Putatunda. *https://arxiv.org/ftp/arxiv/papers/1904/1904.05328.pdf*

Housing Price Prediction by An Nguyen, *https://pdfs.semanticscholar.org/782d/3fdf15f5ff99d5fb6acafb61ed8e1c60fab8.pdf*

https://towardsdatascience.com/machine-learning-project-predicting-boston-house-prices-with-regression-b4e47493633d

GitHub Link:

https://github.com/Roh1702/KC_House_Price_Prediction