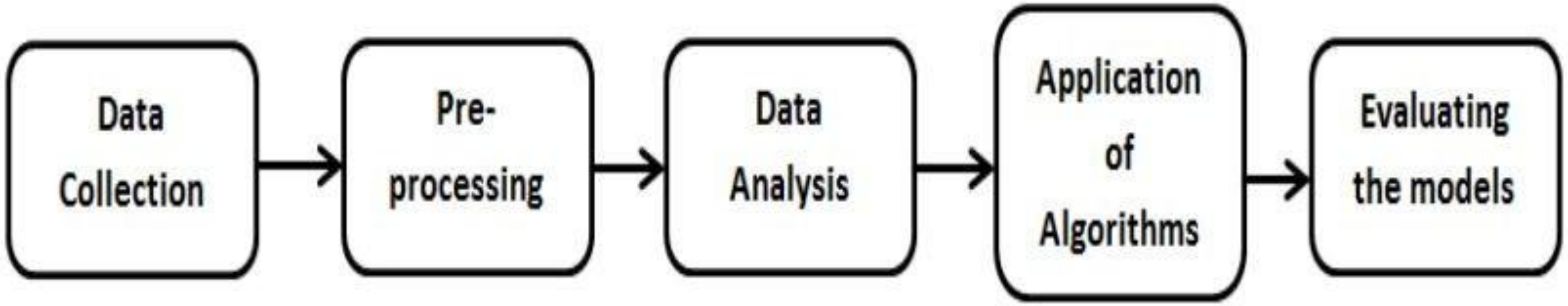
Dwelling Price Appraisal based on Physical, Economic and Social Indicators using Regression methods

JASKIRAT SINGH

Housing Price Prediction of King County, USA using Regression Algorithms

- The project may assist insurance companies to price policies, help contractors to price new houses and estimate demands, or even assess disasters by the government.
- **About The Dataset**
- This dataset contains house sale prices for King County, Washington DC which includes Seattle. It includes homes sold between May 2014 and May 2015.

Methodology



Data Set Overview

Feature Name	Description	Type
Date	Date on which the dwelling was sold	String
Price	Price of the dwelling which we have to predict so this is our target variable	Integer
bedrooms	Number of bedrooms per dwelling	Integer
bathrooms	Number of bathrooms per dwelling	Float
sqft_living	Square Footage of the dwelling	Integer
sqft_lot	Square footage of the lot	Integer
floors	Total floors (levels) in dwelling	Float
waterfront	dwelling which has a view to a waterfront	Integer
view	How many times the dwelling has been viewed	Integer
condition	How good is the condition (Overall)	Integer
grade	Grade of the dwelling	Integer
sqft_above	Square footage of the dwelling apart from basement	Integer
sqft_basement	Square footage of the basement	Integer
yr_built	Built year	Integer
yr_rennovated	Year when dwelling was renovated	Integer
zipcode	Zip	Integer
lat	Latitude coordinate	Float
long	Longitude coordinate	Float
sqft_living15	Living room area in 2015 (implies some renovation)	Integer
sqft_lot15	Lot size area in 2015 (implies some renovations)	Integer

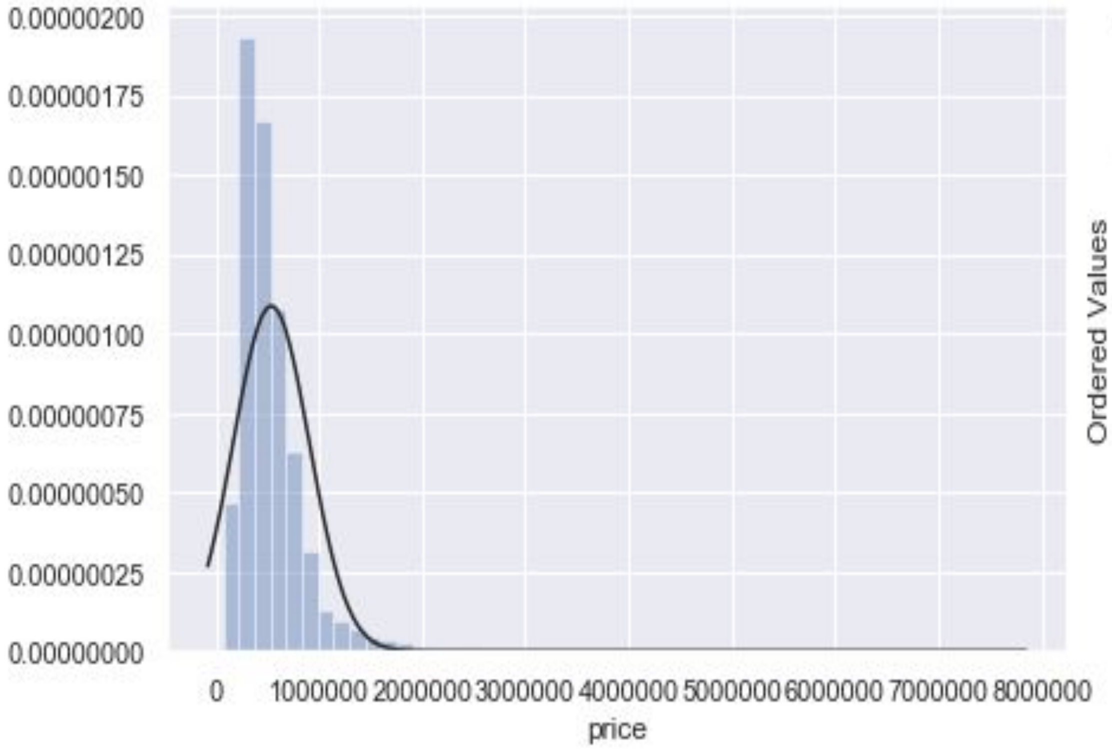
Data Pre-Processing

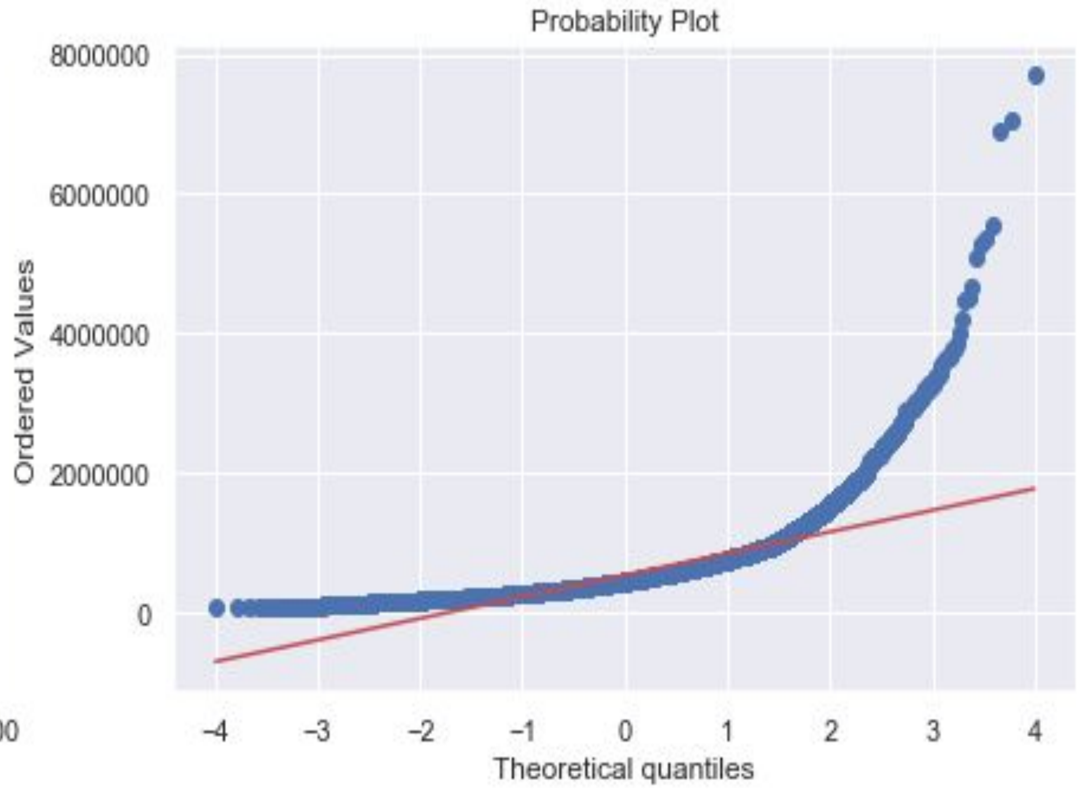
- Data Cleaning
- Statistical Analysis
- Feature Construction
- Identifying Outliers
- Data Conversion
- Collinearity Problem
- Data Visualization

Data Cleaning

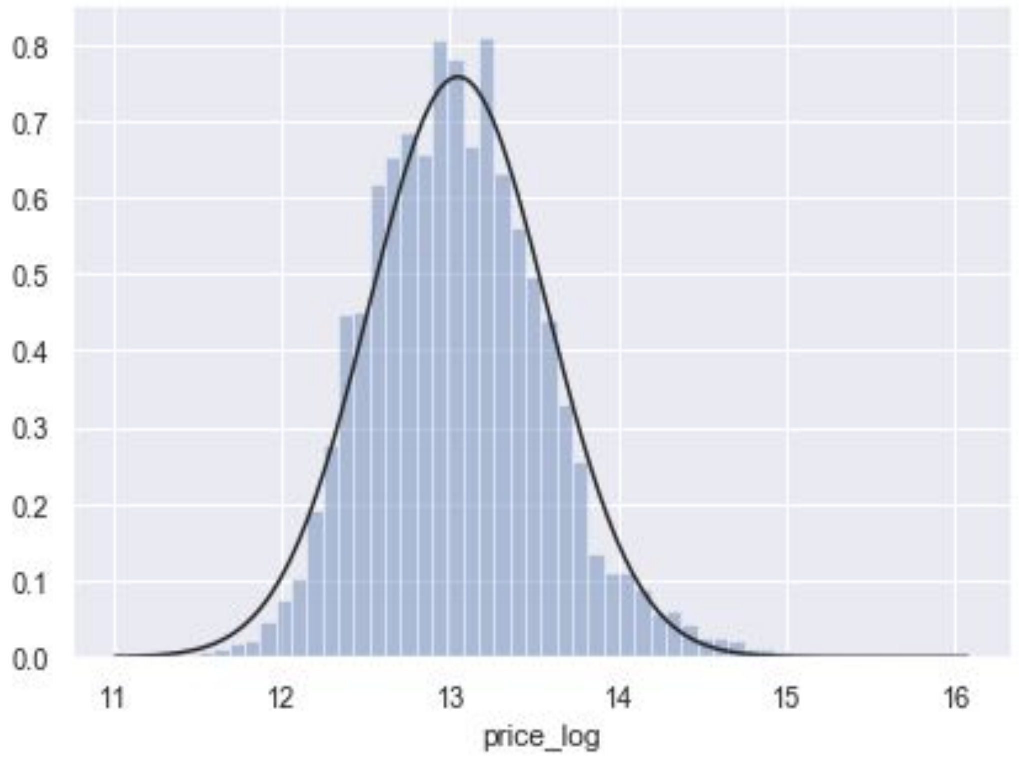
```
Data columns (total 21 columns):
id               21613 non-null int64
date             21613 non-null object
price            21613 non-null int64
bedrooms         21613 non-null int64
bathrooms        21613 non-null float64
sqft_living      21613 non-null int64
sqft_lot         21613 non-null int64
floors           21613 non-null float64
waterfront       21613 non-null int64
view             21613 non-null int64
condition        21613 non-null int64
grade            21613 non-null int64
sqft_above       21613 non-null int64
sqft_basement    21613 non-null int64
yr_built         21613 non-null int64
yr_renovated     21613 non-null int64
zipcode          21613 non-null int64
lat              21613 non-null float64
long             21613 non-null float64
sqft_living15    21613 non-null int64
sqft_lot15       21613 non-null int64
dtypes: float64(4), int64(16), object(1)
```

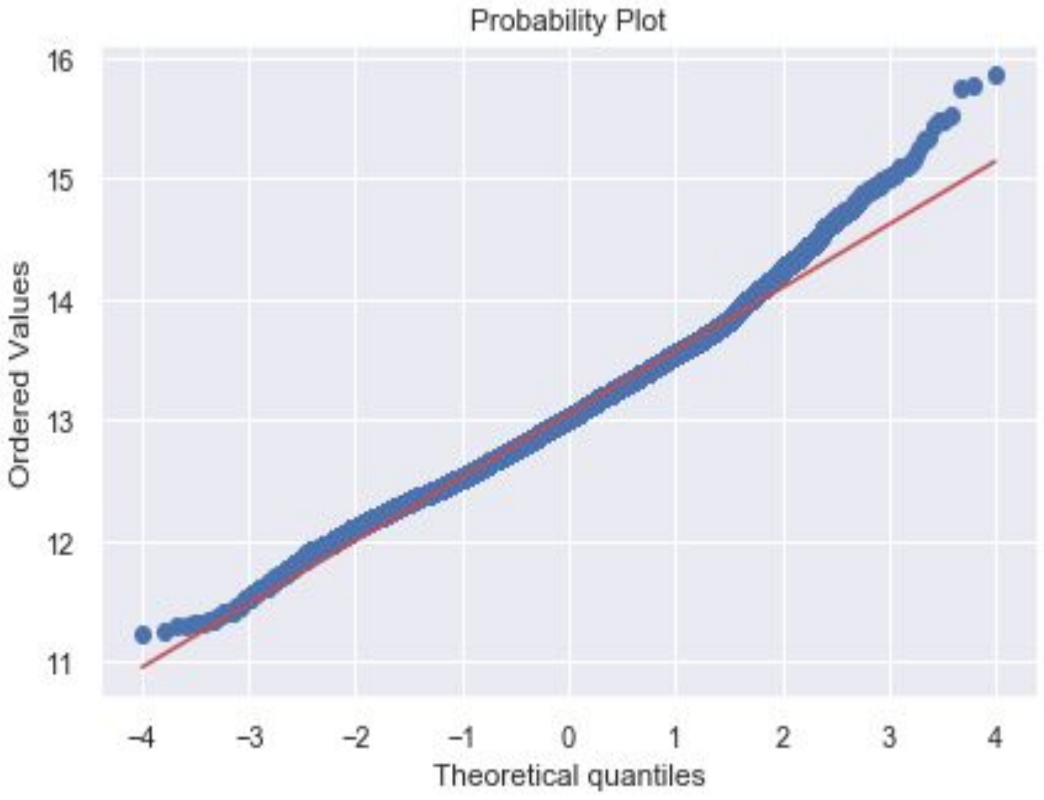
Statistical Analysis of Price Feature



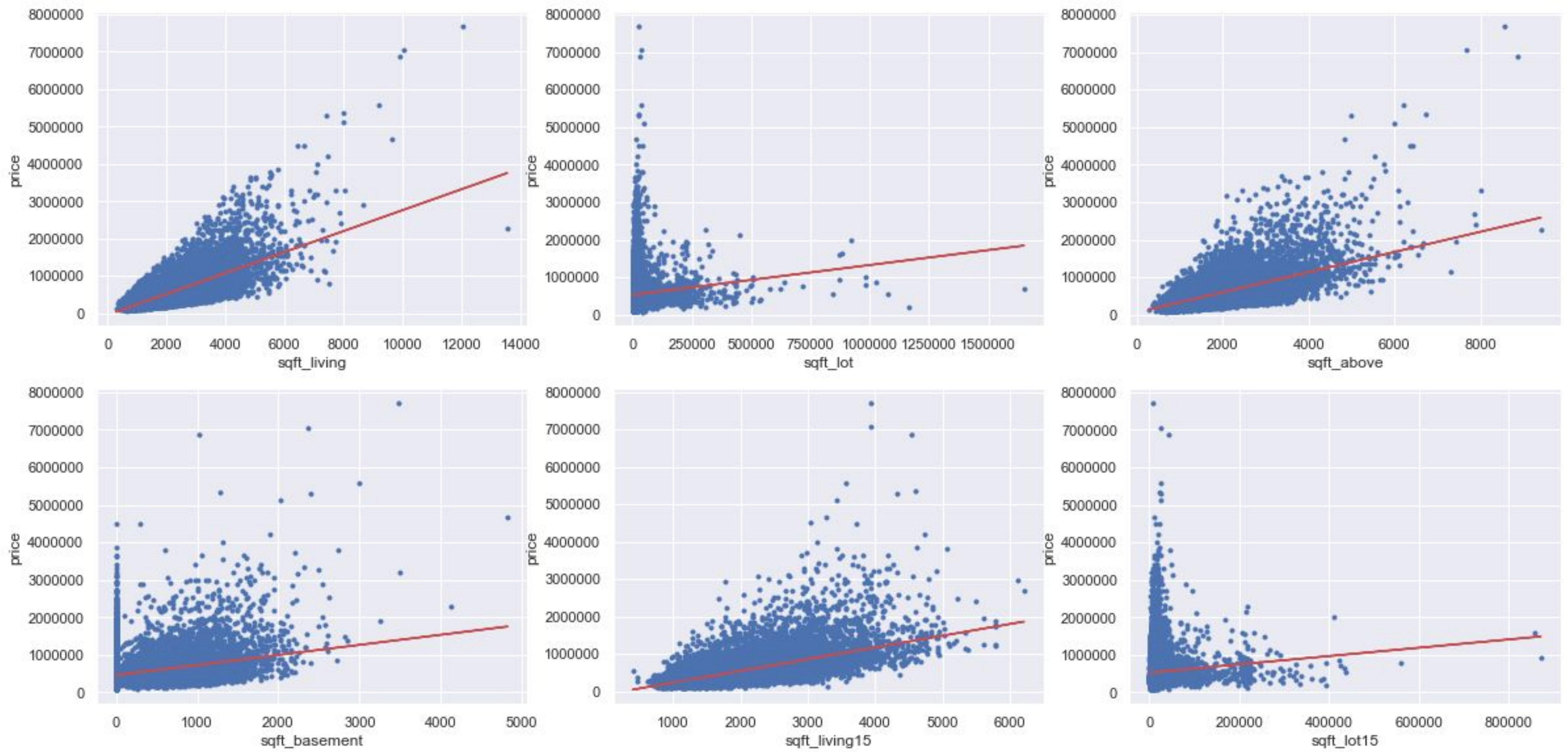


Logarithmic Transformation of Price

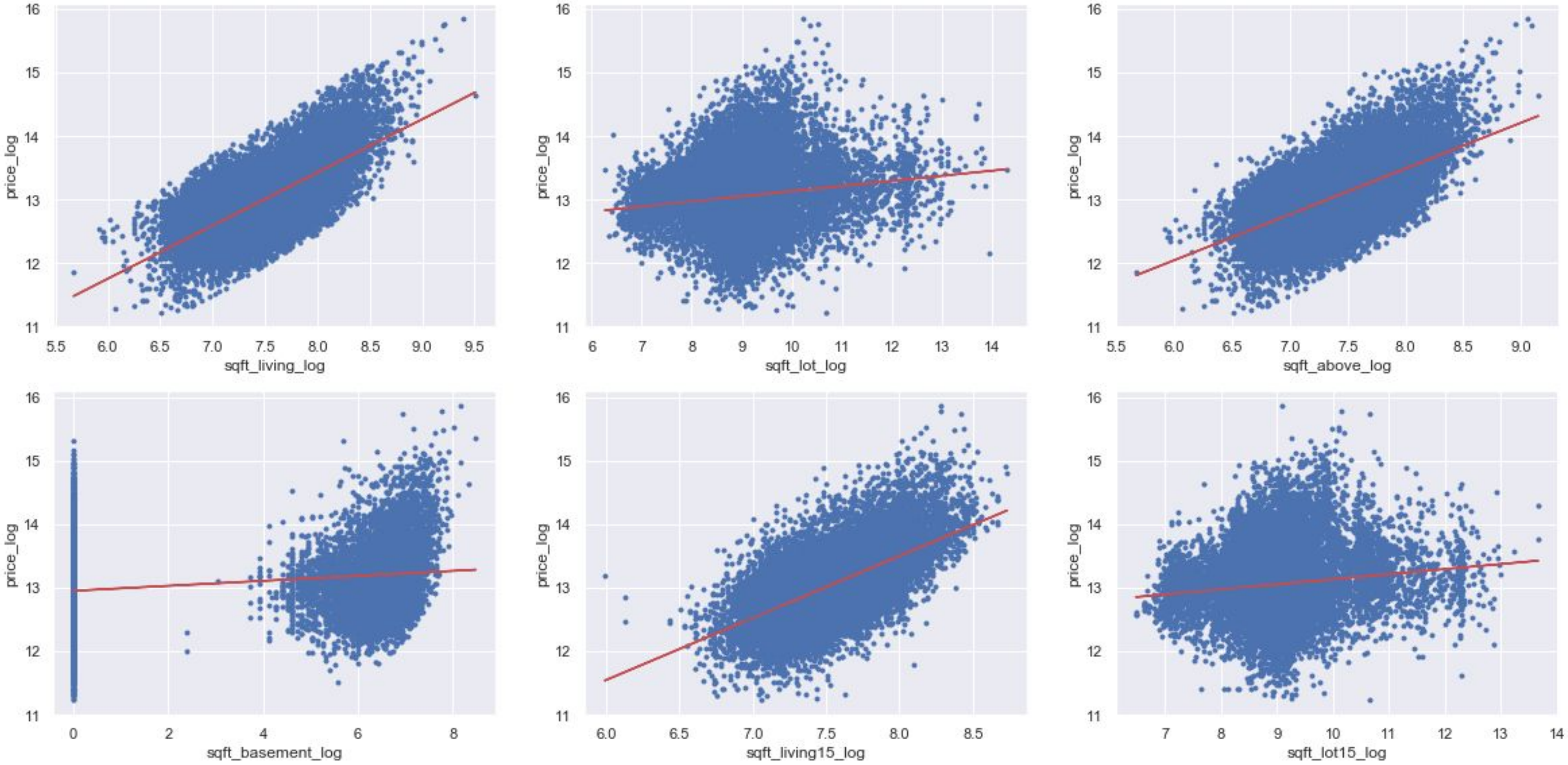




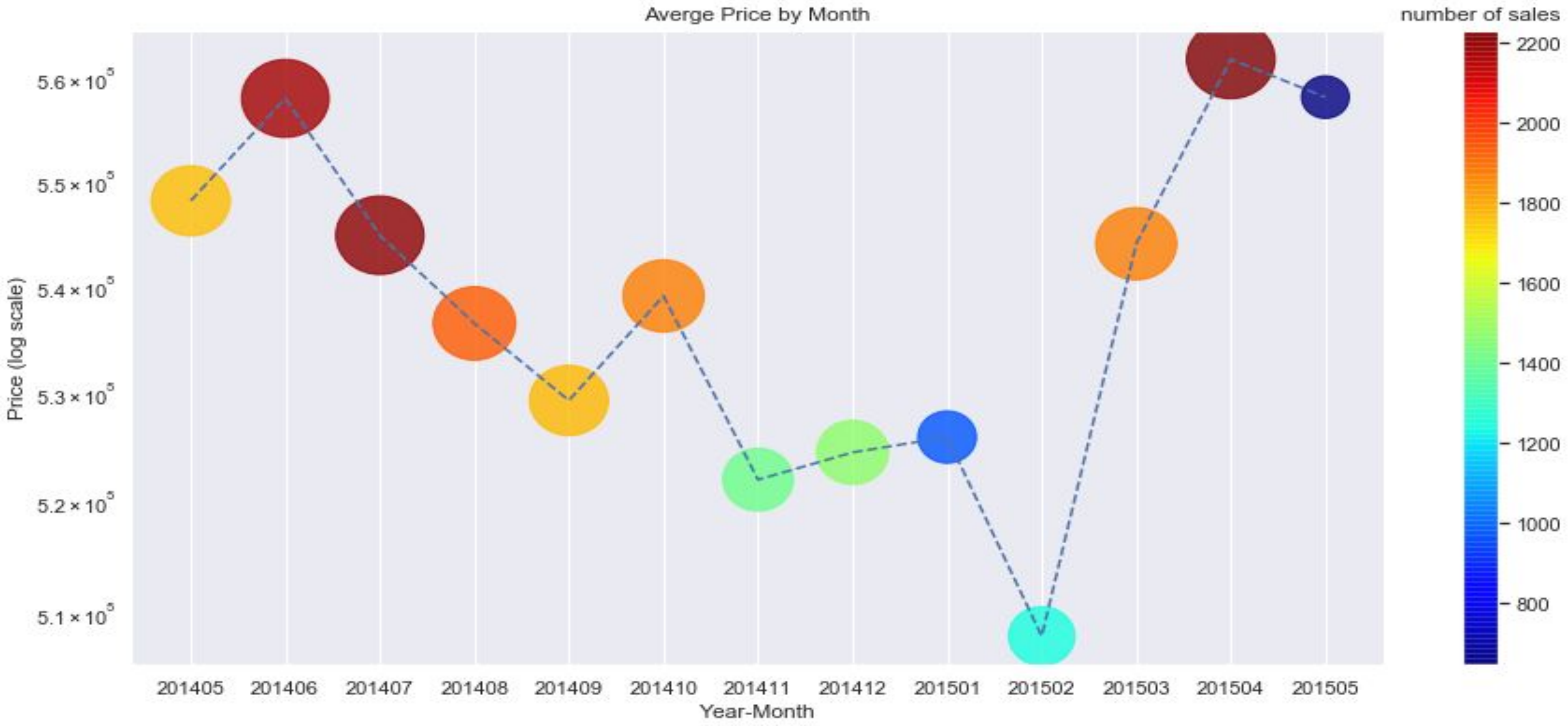
Too Much skewness in numerical Features



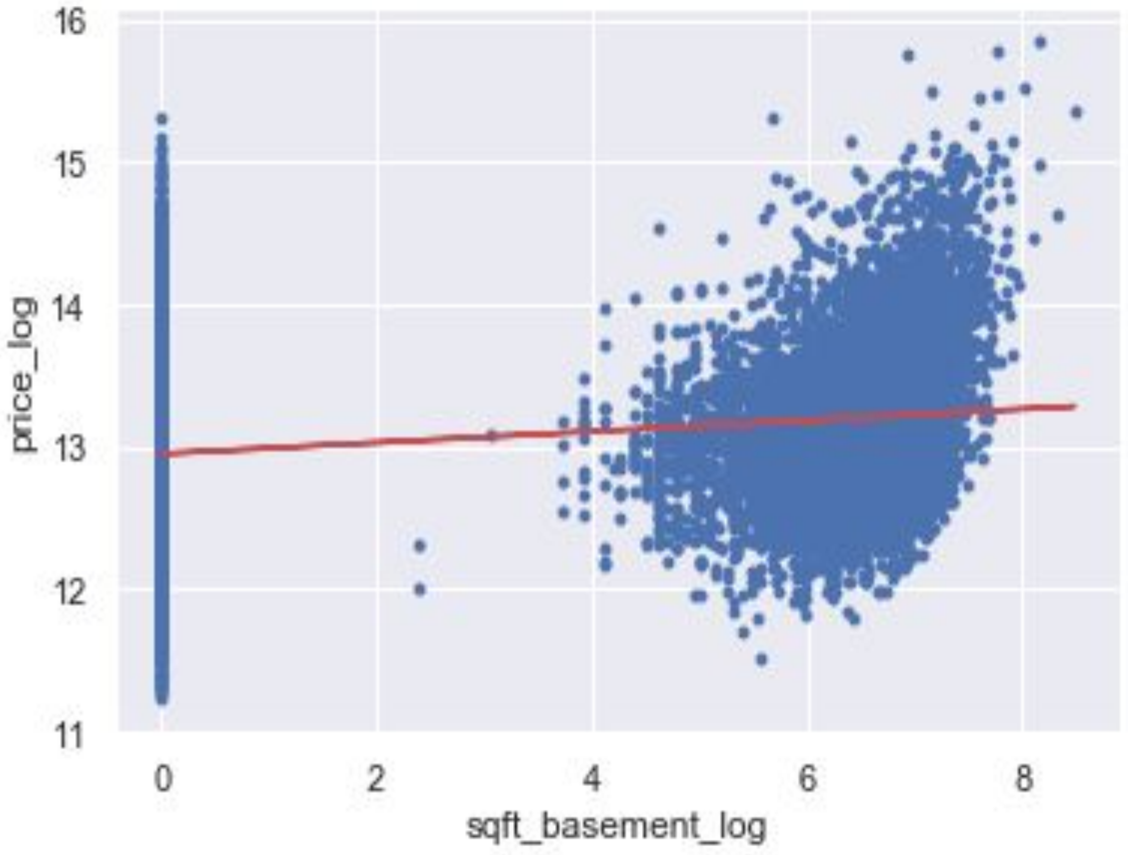
After Log Transformation of numerical features

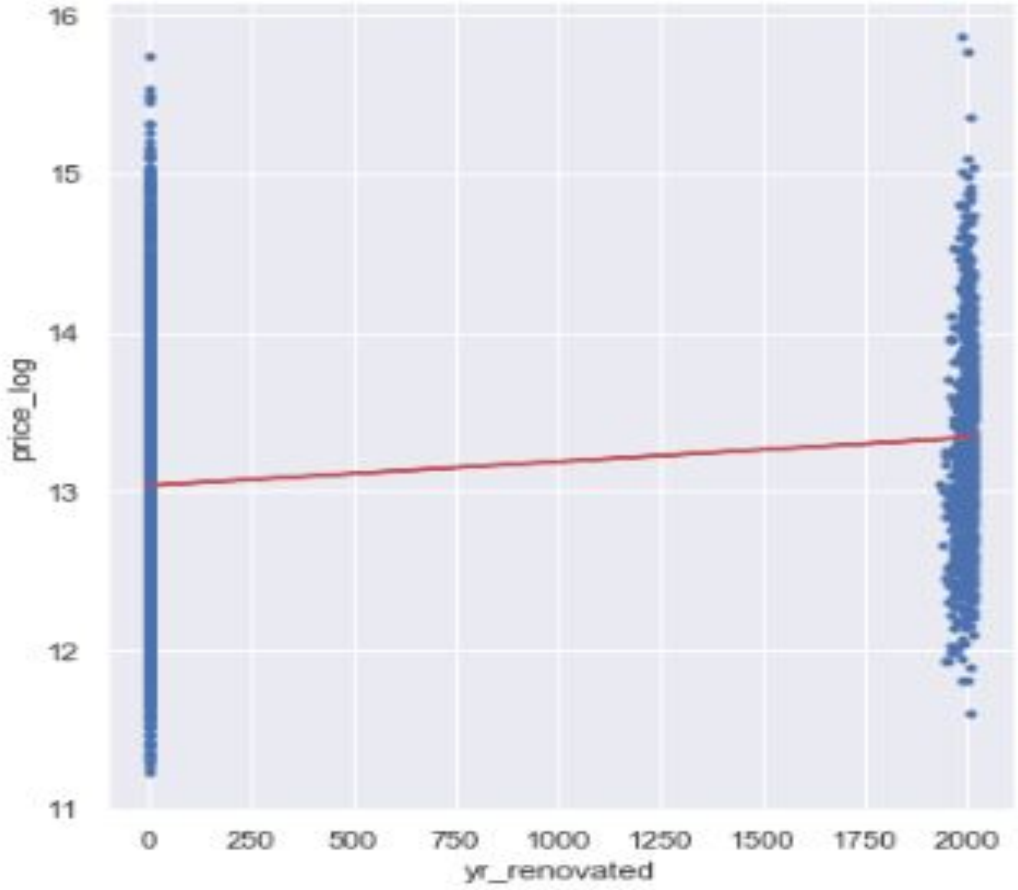


Feature Construction- Seasonality of house Price

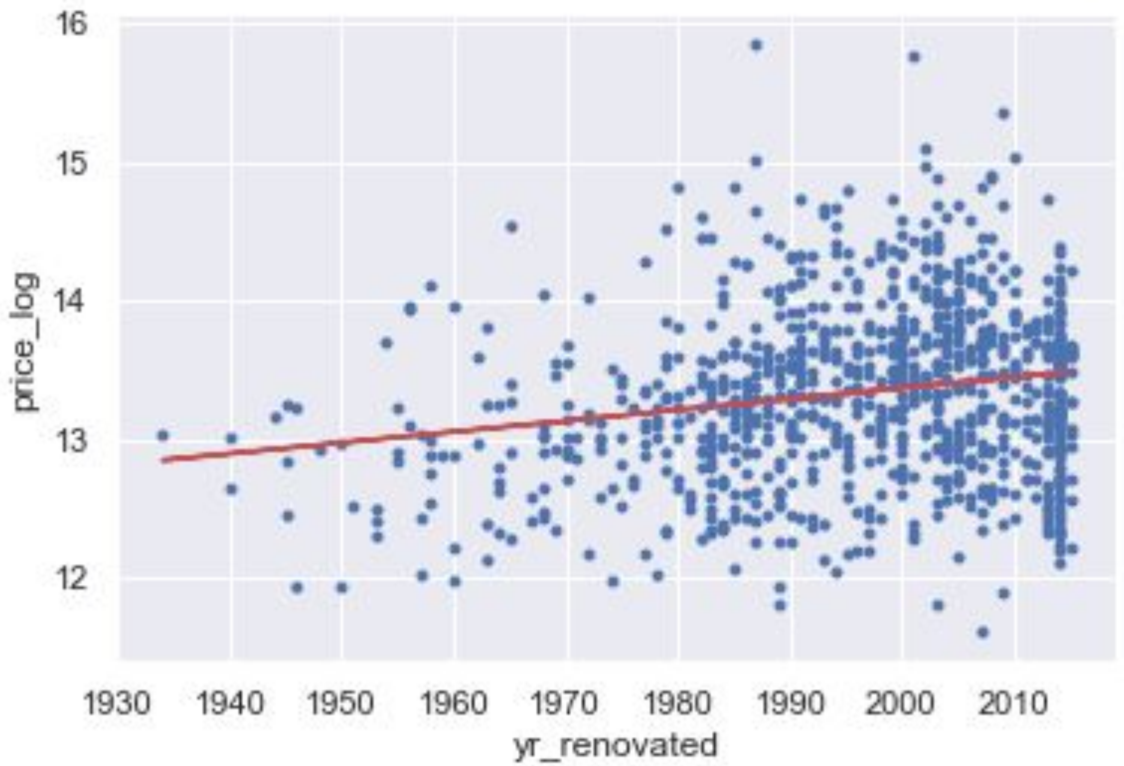


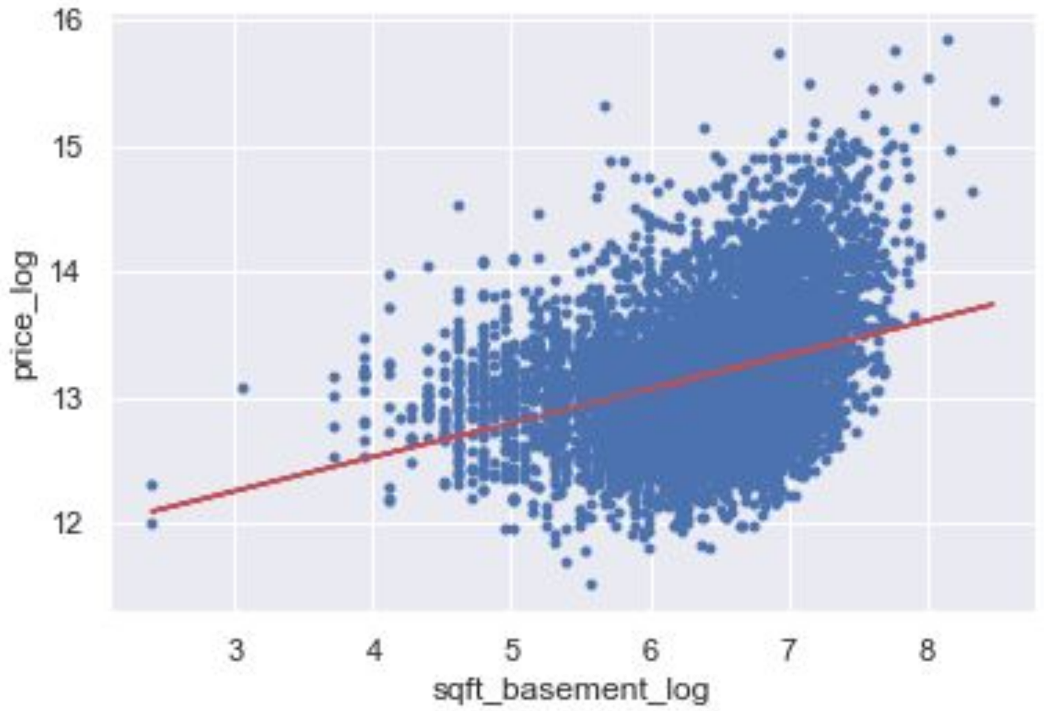
Feature extraction from basement area and year renovation attribute





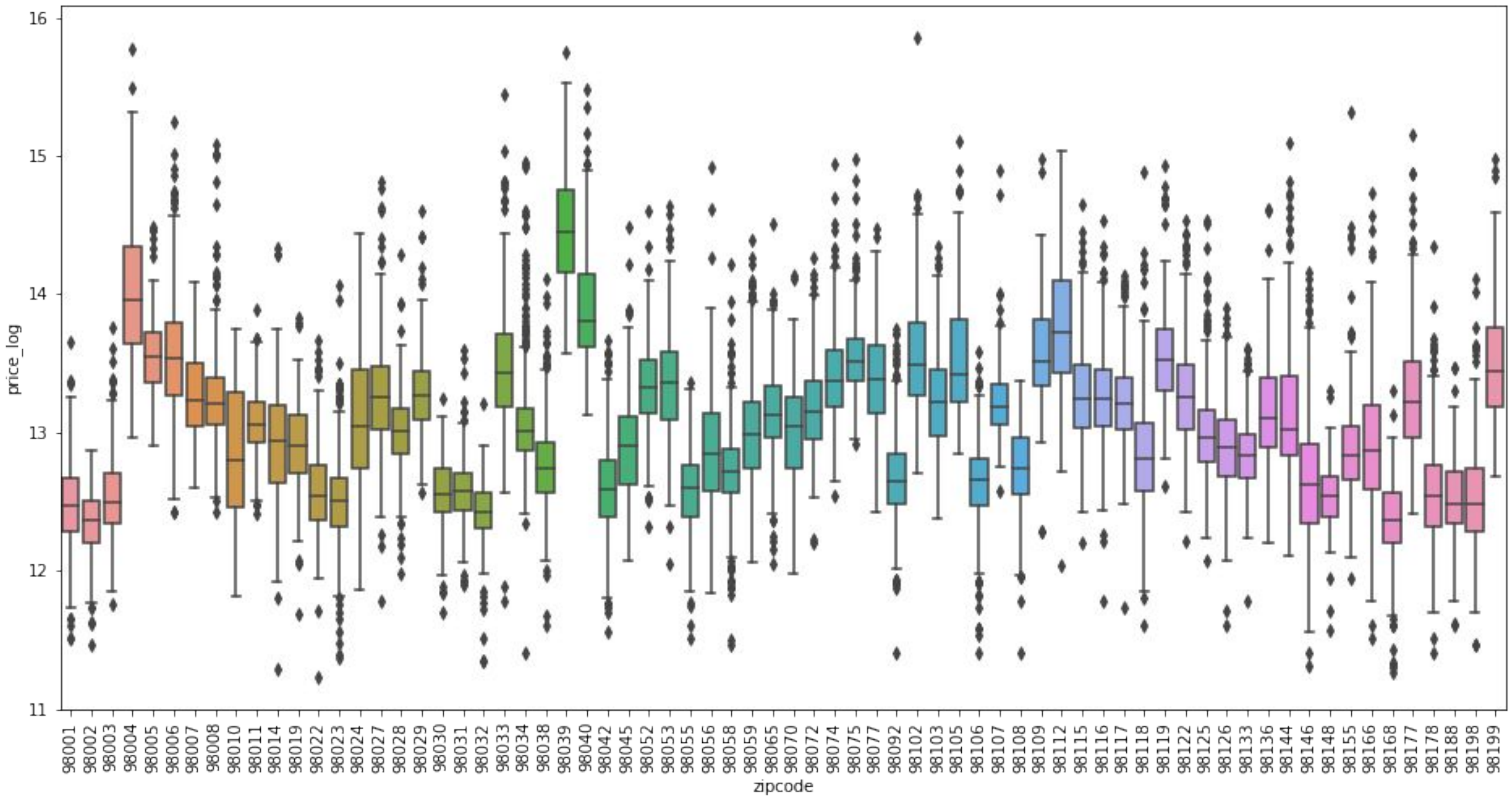
After Feature Extraction of basement area and year of renovation



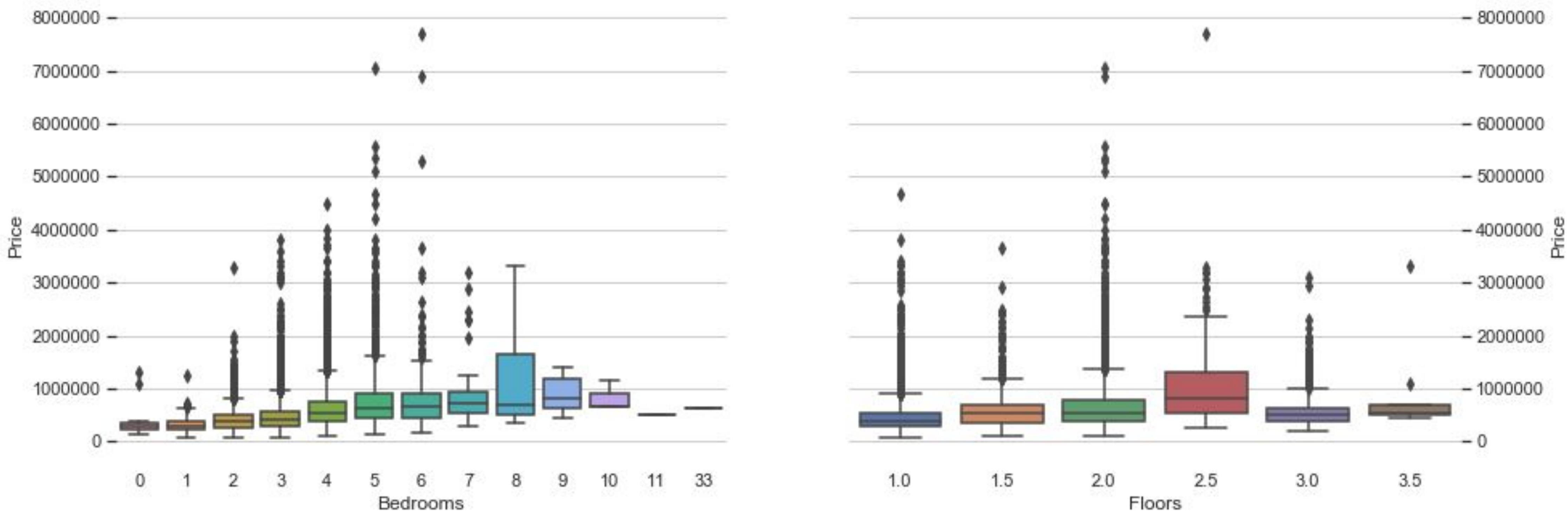


Identifying Outliers

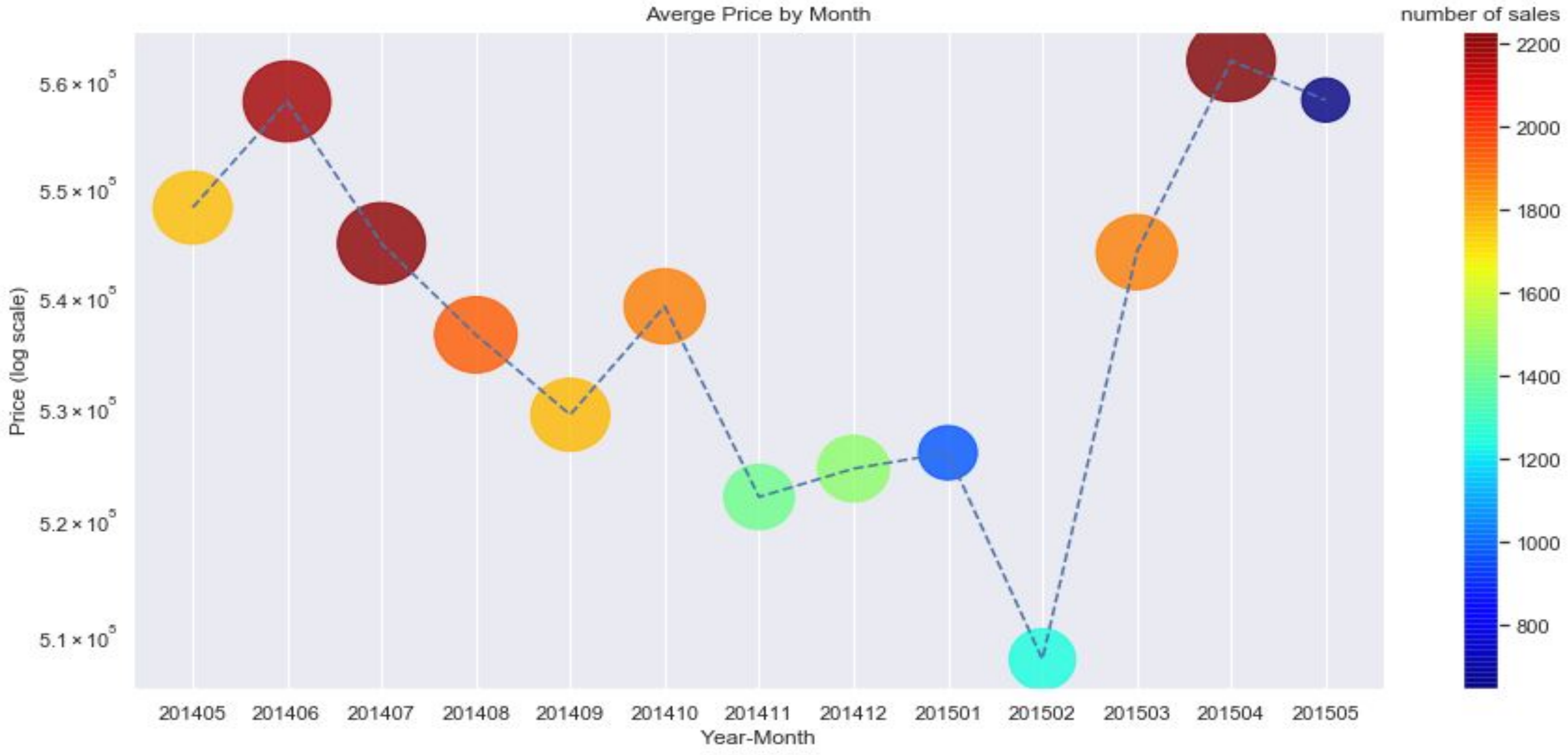
Data Conversion of zip code



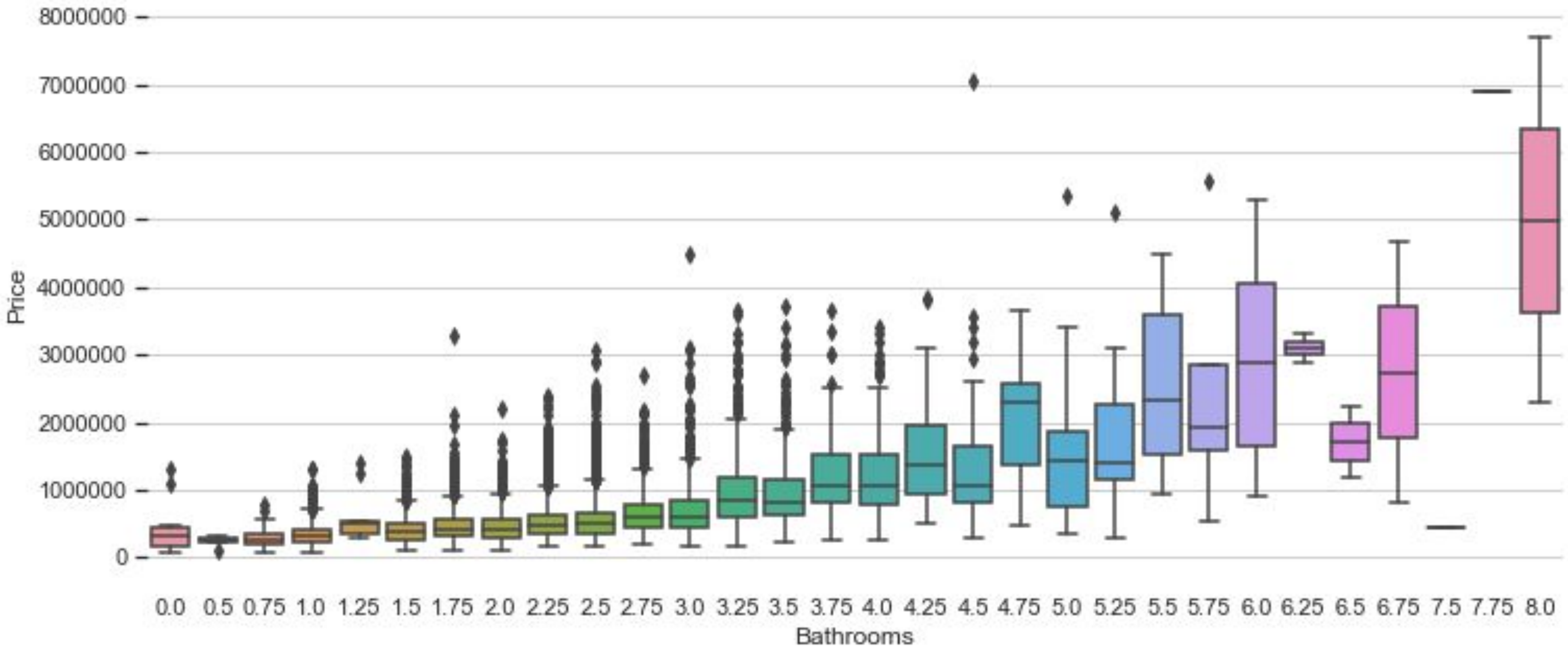
Data conversion of bedrooms and Floors



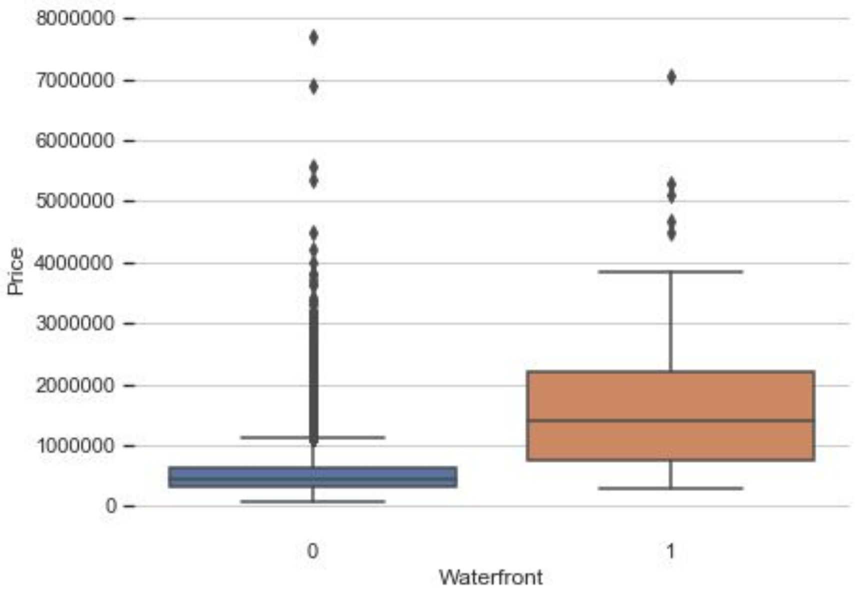
Data conversion of seasonality attribute

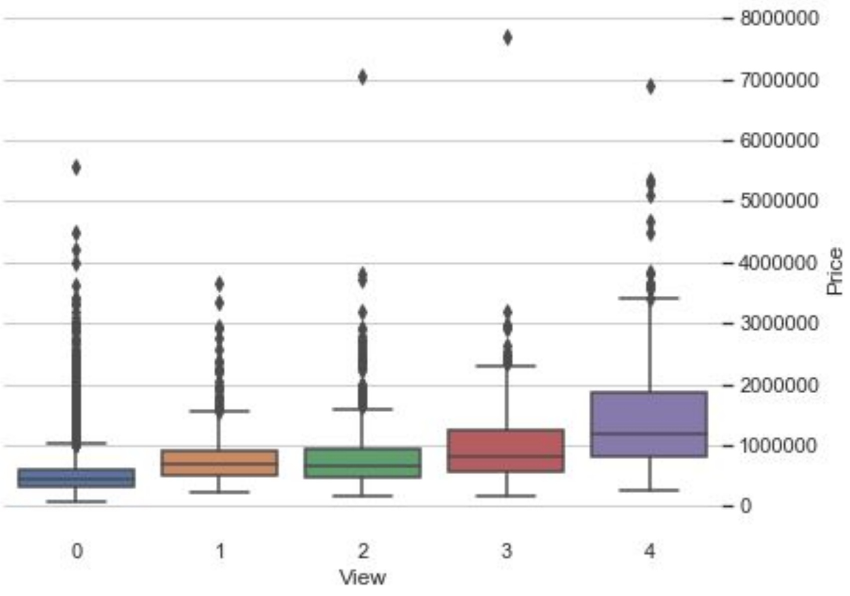


Bathrooms Feature

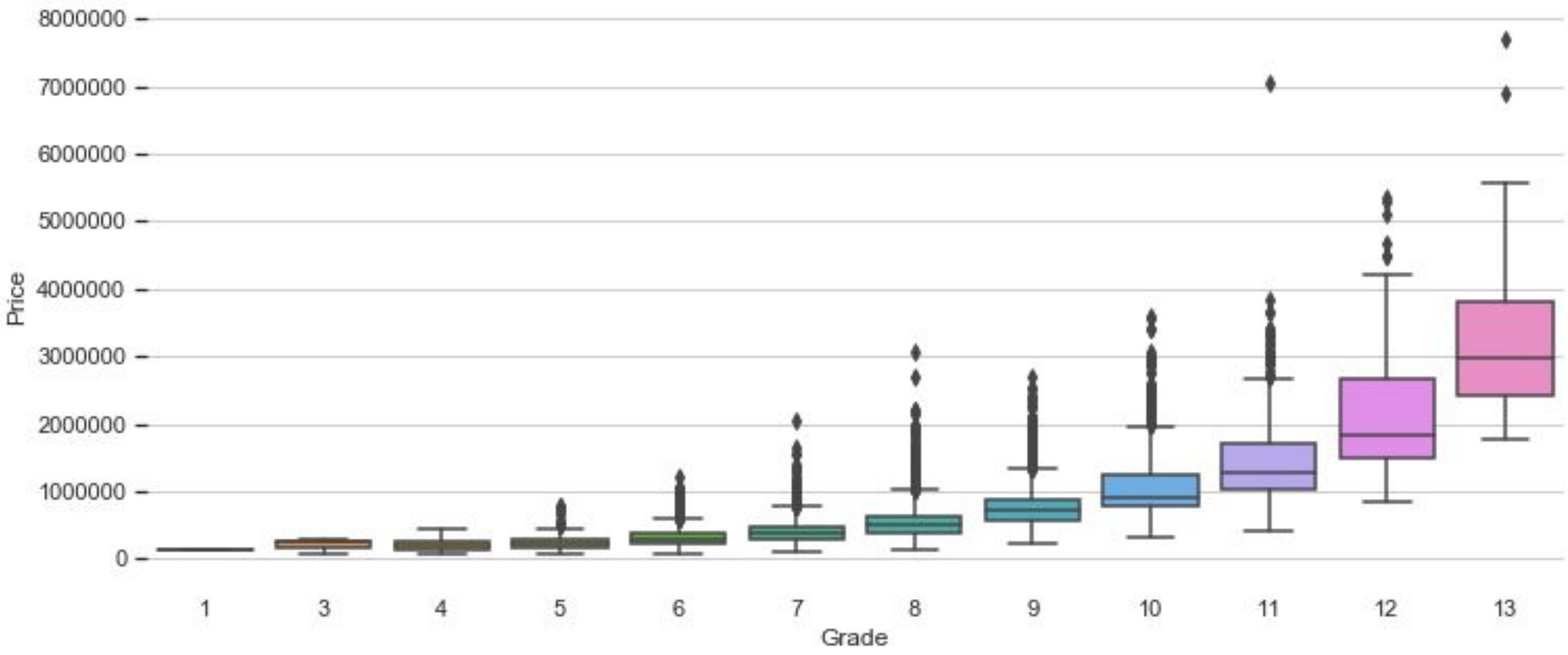


Water front and view feature

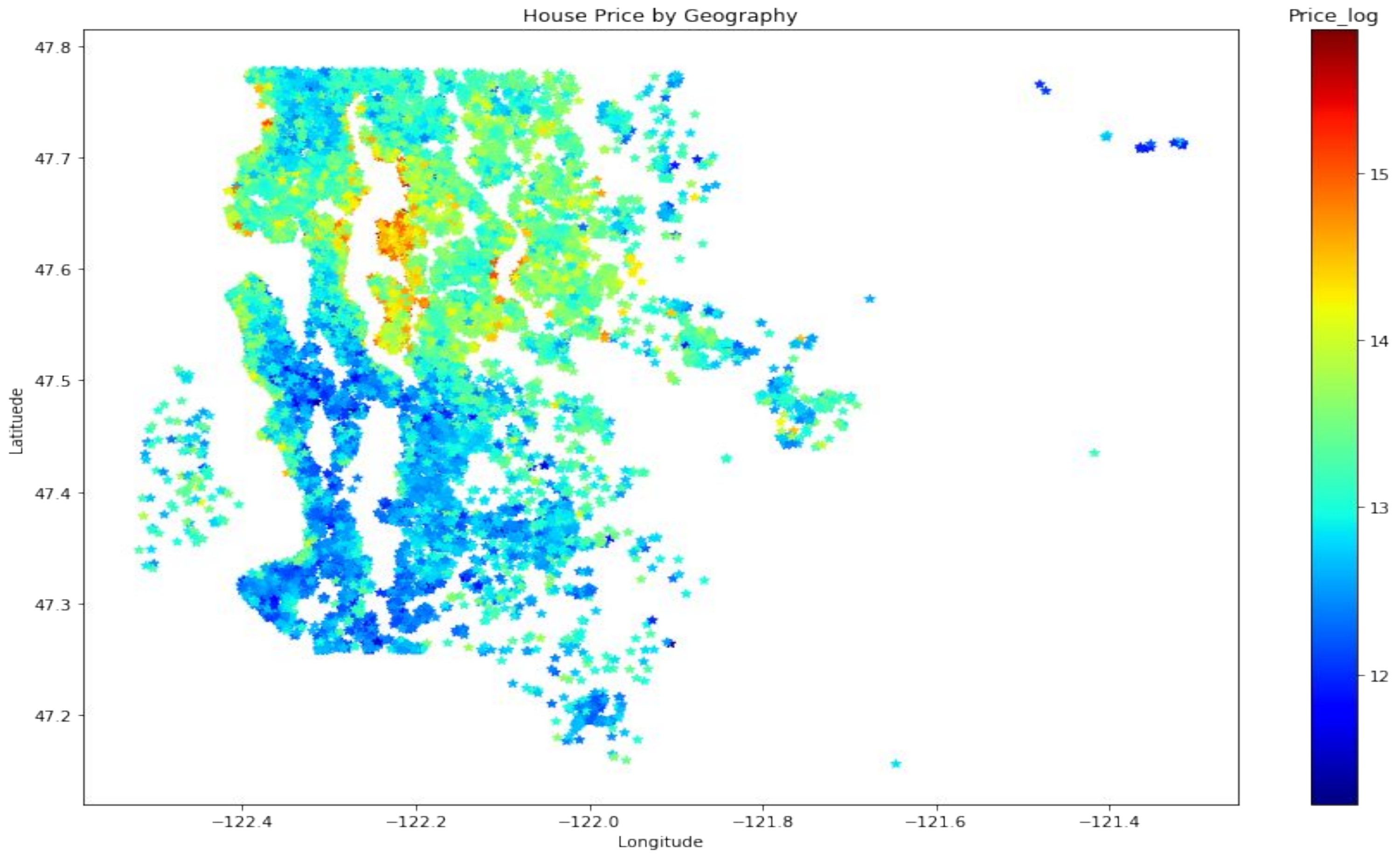




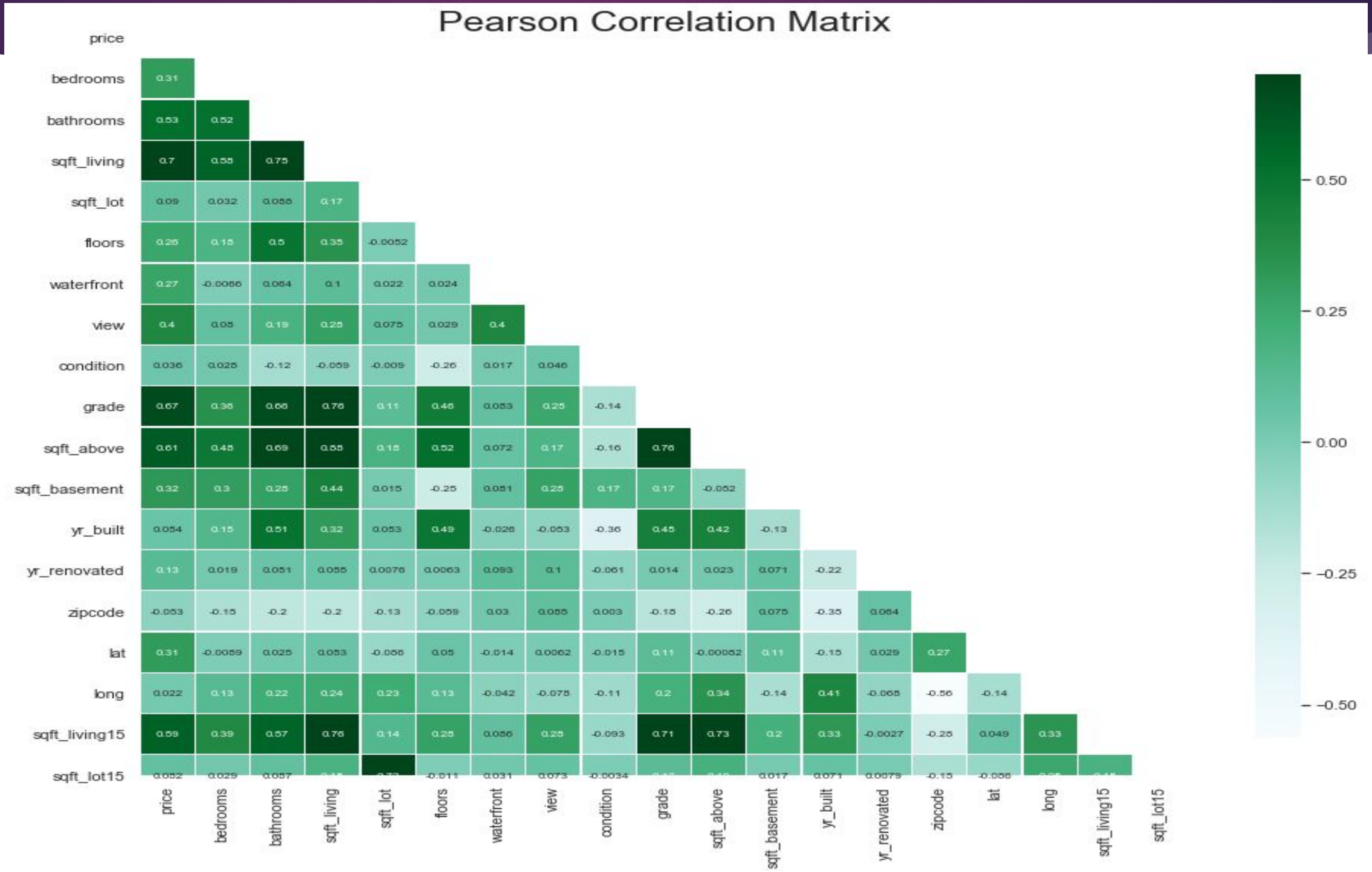
Grade Feature



Data Visualization of geography feature



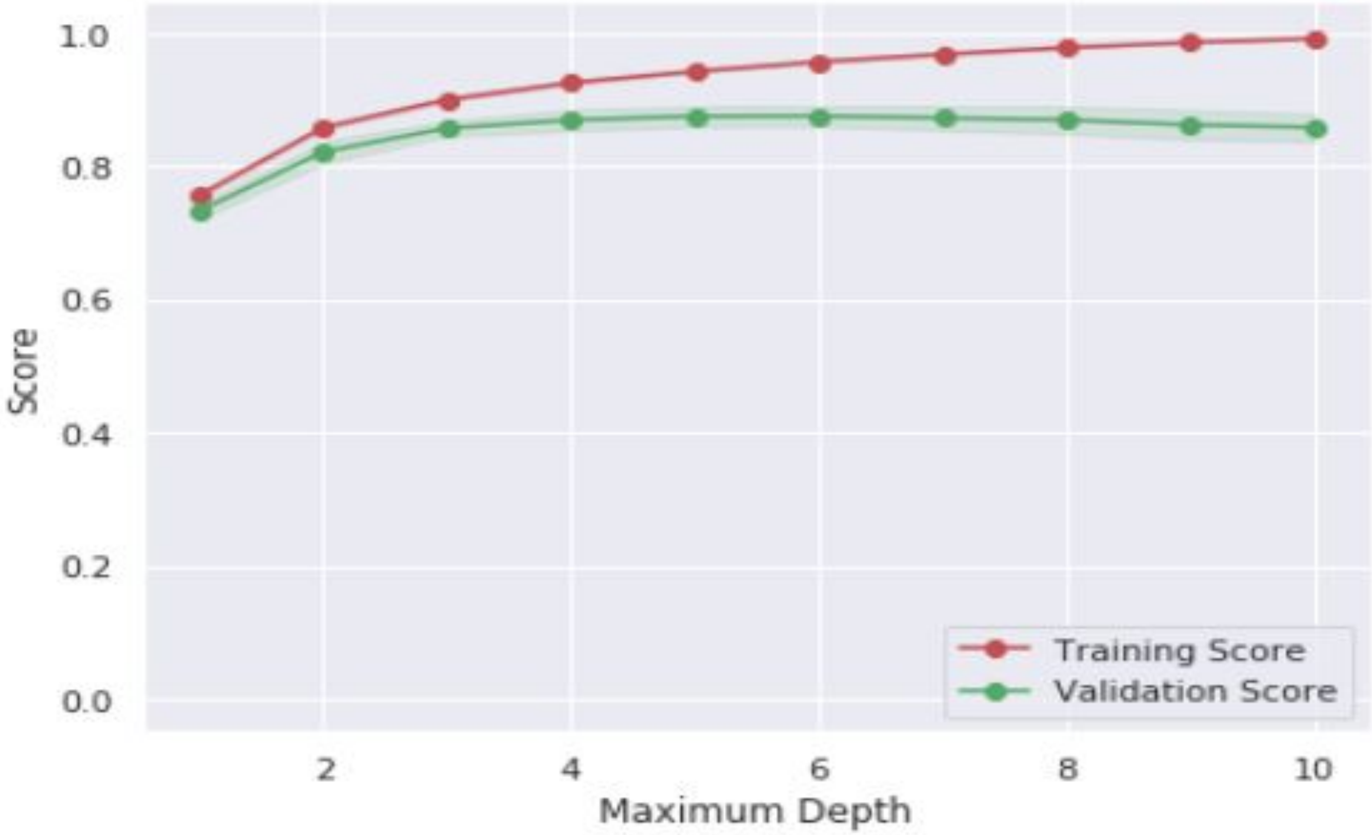
Correlation Among Features



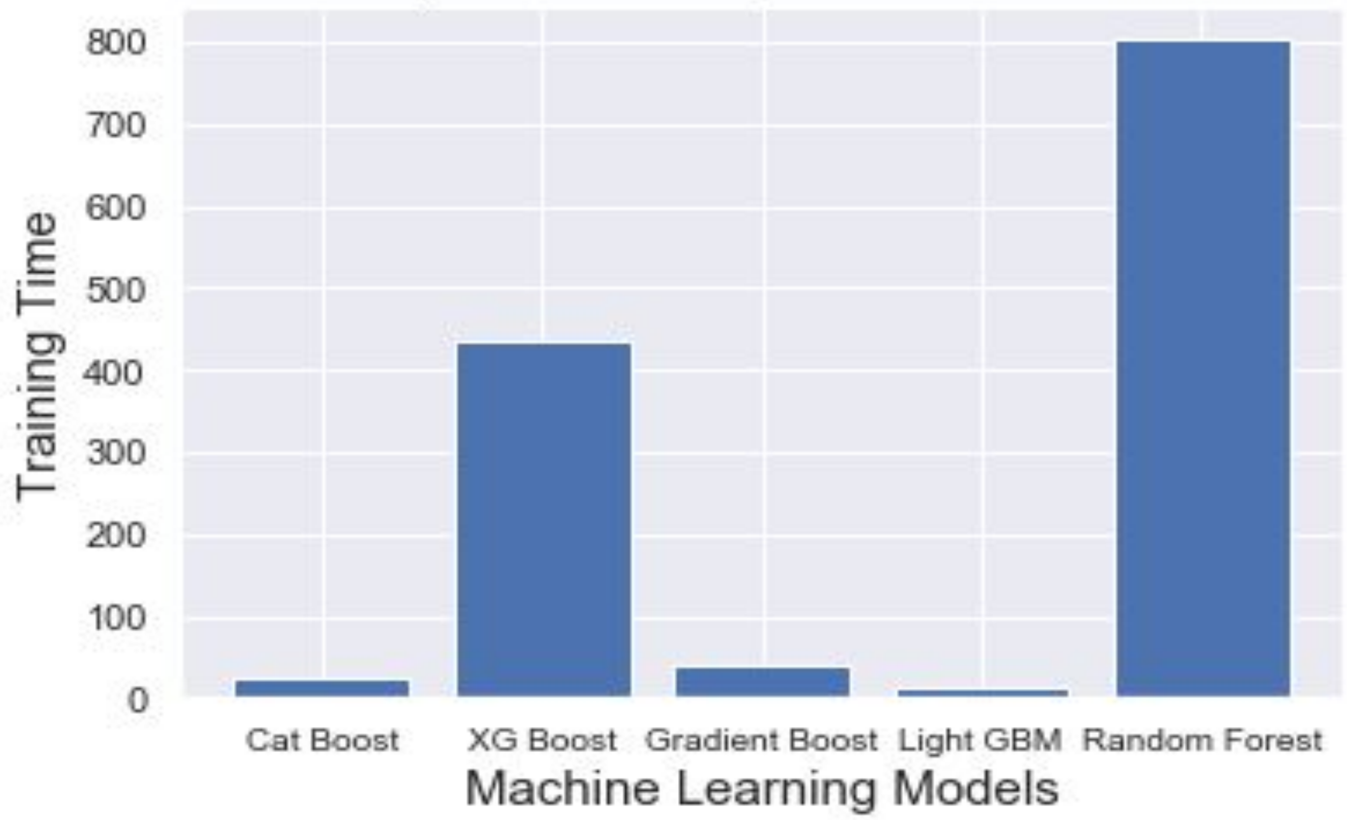
Models Evaluation based on manual tuning of hyper parameters

Model	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)	R-squared (train)	R-squared (test)	Adjusted R-squared (train)	Adjusted R-squared (test)	Explained Variance (train)	Explained Variance (test)	5-Fold Cross Validation
Cat Boost	0.1515	0.1073	0.9551	0.9150	0.9551	0.9146	0.9551	0.9150	0.9145
XGBoost	0.1545	0.1089	0.9858	0.9115	0.9858	0.9111	0.9858	0.9115	0.9107
Gradient Boost	0.1537	0.1101	0.9296	0.9125	0.9295	0.9121	0.9296	0.9125	0.9090
Light GBM	0.1577	0.1125	0.9883	0.9078	0.9883	0.9074	0.9883	0.9078	0.9067
Random Forest Regression	0.1703	0.1221	0.9849	0.8925	0.9849	0.8921	0.9849	0.8925	0.8898

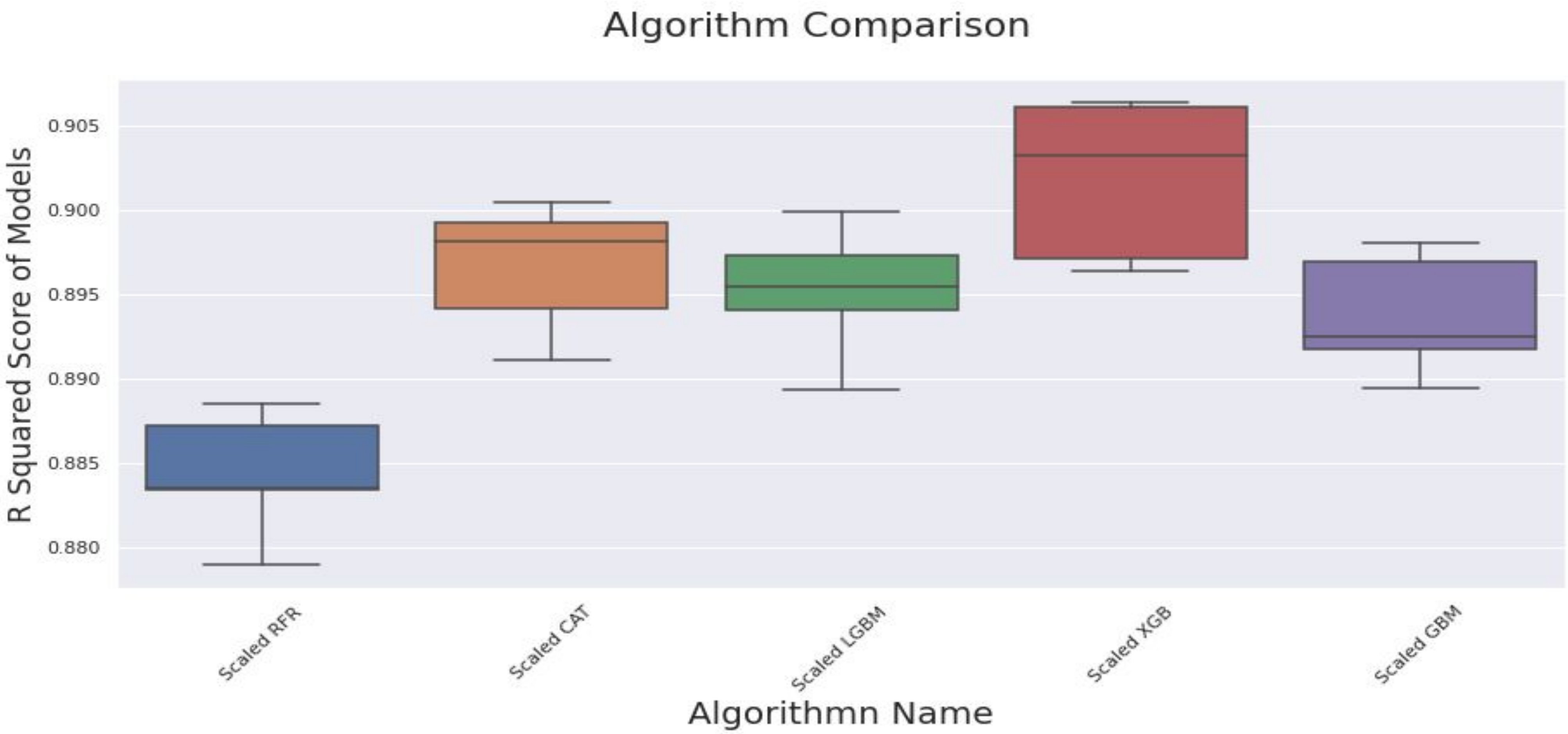
Why number of folds are set as 5 in cross validation score



Comparison of training time of all models



Models Evaluation based on automatic hyper parameter tuning



Future Enhancement of House Price Forecasting

- Considering more factors influencing the dwelling prices
- Adding Safety feature
- Using Deep Learning
- Using Principal component analysis
- Zip code feature engineering
- Using stacked model

Why does Organization's need this predictive model?

- possibly many real-estate firm's are interested in intelligent decision making regarding house price forecasting.
- The Organization's will use this data to help clients purchase properties at affordable price.
- Current process is good but manual and time consuming
- Organization's wants an edge over competition