King County House Prices Prediction Model

Instructor- Ebrahim Nasrabadi

Submitted by- Sayali Walke

SUMMARY

Overview:

This dataset contains house sale prices for King County including homes sold between May 2014 and May 2015.

Purpose:

The goal is to predict the sales price for each house based on the given features.

Objectives:

1] Data Analysis
2] Data Visualization
3] Data Preprocessing
4] Feature Selection
5] Model Implementation
6] Future Improvements
7] Learning Outcomes

Overview of Dataset

This dataset has 18 Features.

Id to denote the house

Date on which the house was sold

Column Name	Description
id	a notation for a house
date	Date house was sold
price	Price is prediction target
bedrooms	Number of Bedrooms/House
bathrooms	Number of bathrooms/House
sqft_living	square footage of the home
sqft_lot	square footage of the lot
floors	Total floors (levels) in house
waterfront	House which has a view to a waterfront
view	Has been viewed
condition	How good the condition is (Overall)
grade	overall grade given to the housing unit, based on King County grading system
sqft_above	square footage of house apart from basement
sqft_basement	square footage of the basement
yr_built	Built Year
yr_renovated	Year when house was renovated
Zipcode	zip
Lat	Latitude coordinate
Long	Longitude coordinate
sqft_living15	Living room area in 2015
sqft_lot15	lotSize area in 2015

Data Analysis and Data Cleaning

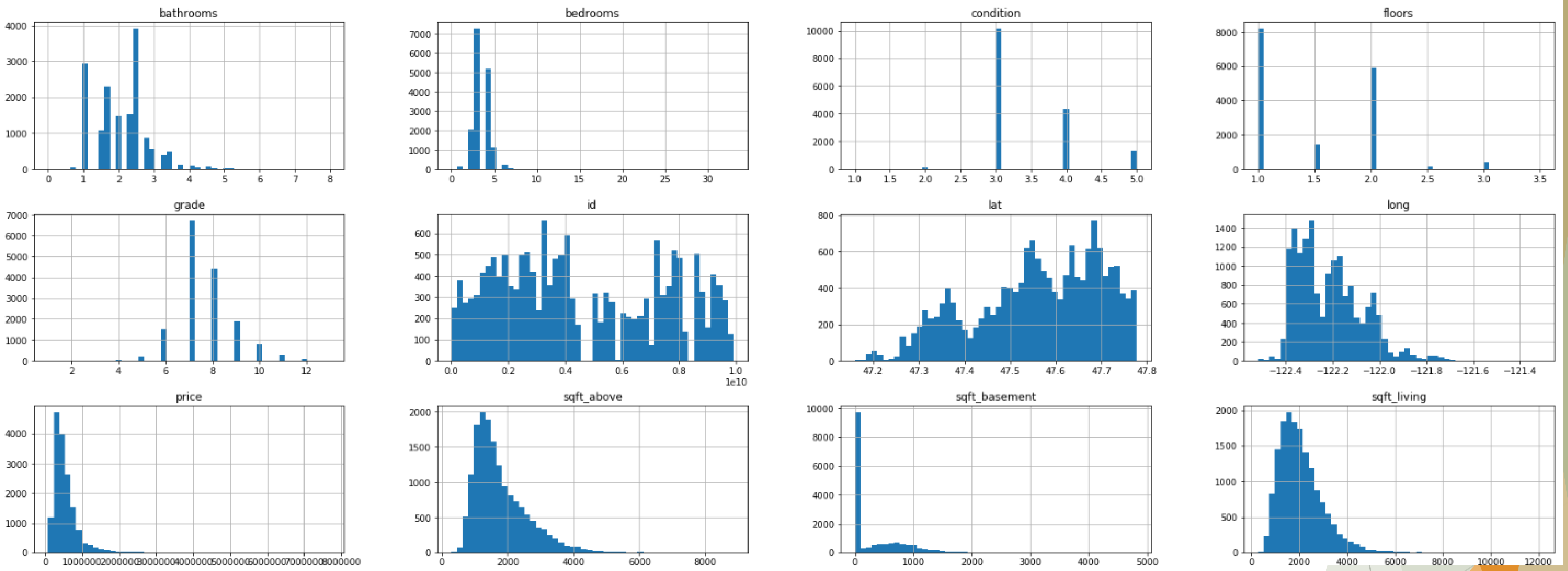
Simple checkpoints:

- Missing values- checked for missing values in dataset
- Unique values- checked for unique values of house id

Statistical Analysis:

- Average price of house sold in King County is $538926.
- Very few houses which have some features and price appear far from others like 33 bedrooms or price $7700000
- There will always be some outliers as some luxury house prices in this dataset. I have addressed this problem in next step to reduce the effect of outliers
- The sqft_living column has maximum value of 12050 sqft, which is 3 standard deviations above mean.
- Avg no. of 3 bedrooms and 2 bathrooms per house were sold in King County
- Avg area of house is 2071 Sqft with one house having 12050 Sqft. Area
- The dataset contains houses built from year 1900 to 2015

Data Visualization



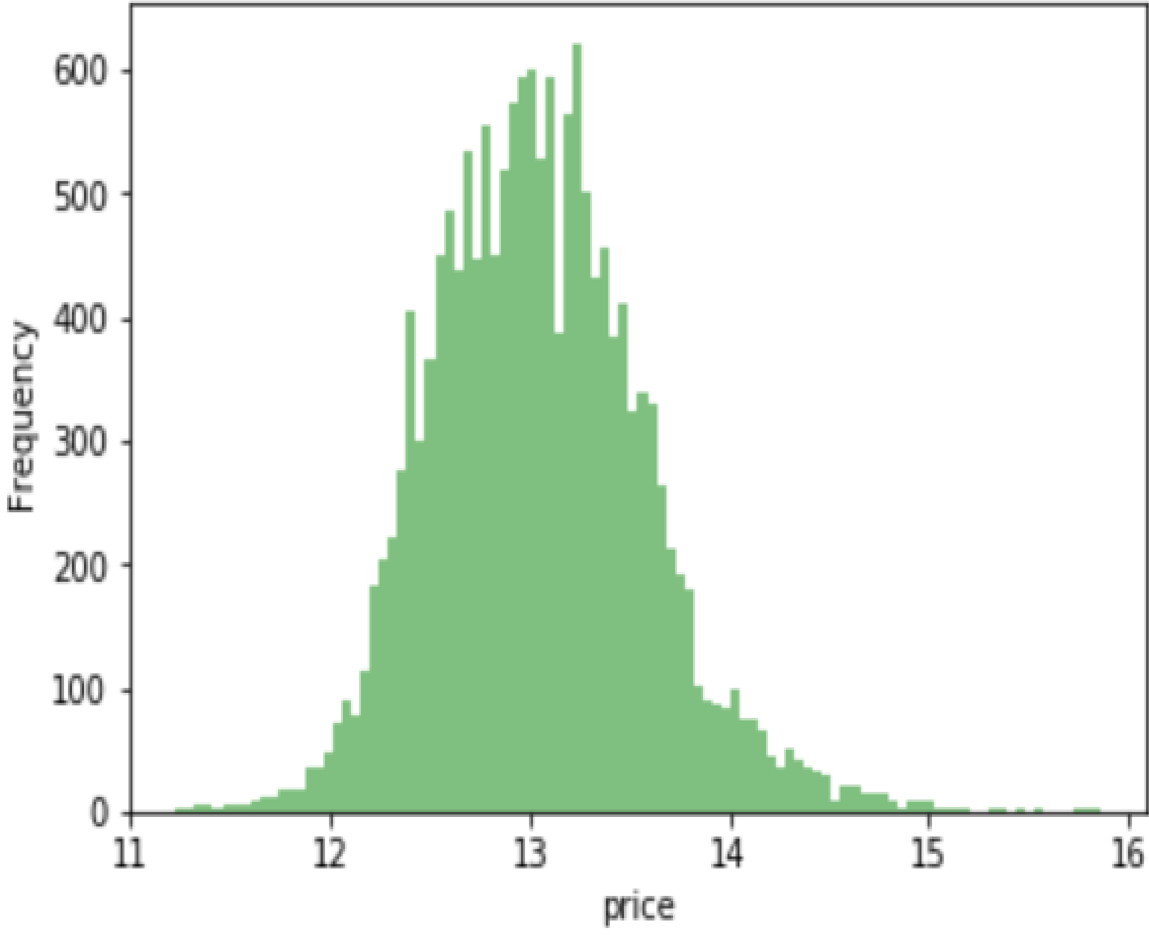
Observations:

- Most of the houses in the dataset have 3 bedrooms and have only 1 floor
- Most of the houses have more than average condition(3)
- For most of the houses overall grade is 7
- The histograms of Price is right-skewed.

Data Preprocessing

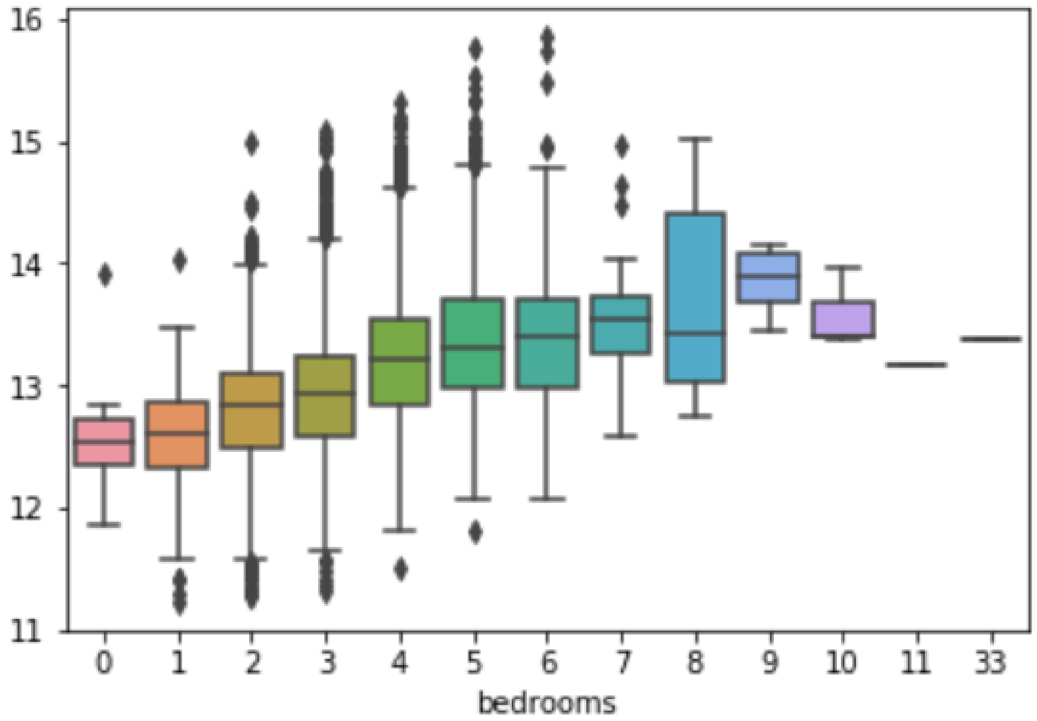
Log transformation:

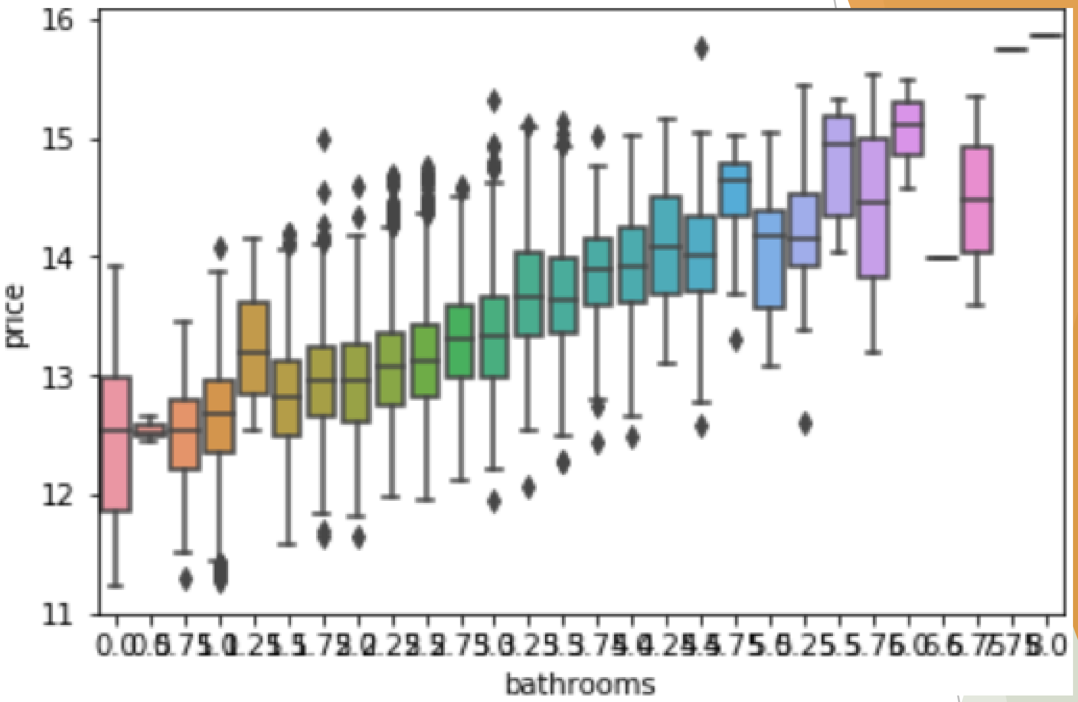
- According to visualization in previous slide histogram of price of the houses was right-skewed
- If the data is skewed, then the mean may not provide a good estimate for the center of the data and represent where most of the data fall
- The price of house varies from 75K to 7.7M
- By normalizing price variable I made value o this numeric column common scale, without distorting differences in the ranges of values.



Log transformed Histogram of price

Box plots





Outliers:

- From the statistical analysis I found that the features sqft_living, sqft_lot, bedrooms, bathrooms have maximum value above two standard deviations that means, these features have outliers
- From above box plots we can say the outliers for price variable correspond to outliers in these features and outliers in these features corresponds to outliers in grade, condition
- So after considering these dependencies I have decided to keep these outliers

Data Preprocessing

Computed Columns:

While traversing through dataset I found that some of the columns are not much significant. So, In order to provide us with better understanding of the data, I computed new columns.

- **Age of house = year_sold (extracted from date column) – yr_built**
- **Is_renovated (0 or 1) = 0 if yr_renovated is not present**
  1 if yr_renovated is present

Encoding high cardinality features:

- High-cardinality refers to columns with values that are very uncommon or unique
- For this dataset feature zip code with many distinct value could be very predictive as it could be telling prices of house according to area
- I have encoded zip code into dummy numerical value and later investigated its relationship with price.

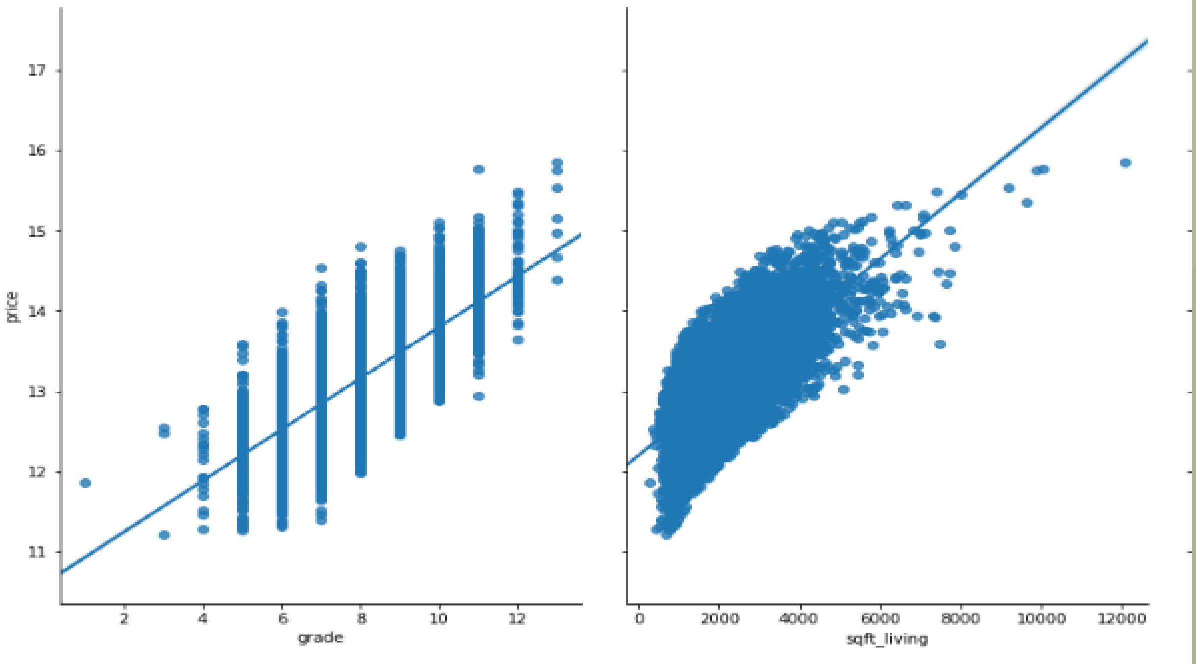
Feature Selection

1] Filter Method using correlation:

- Correlation refers to some statistical relationships involving dependencies between variables
- For better understanding of relationships between features, I have used heatmap
- But as we are more interested in finding out relationship between features and response,

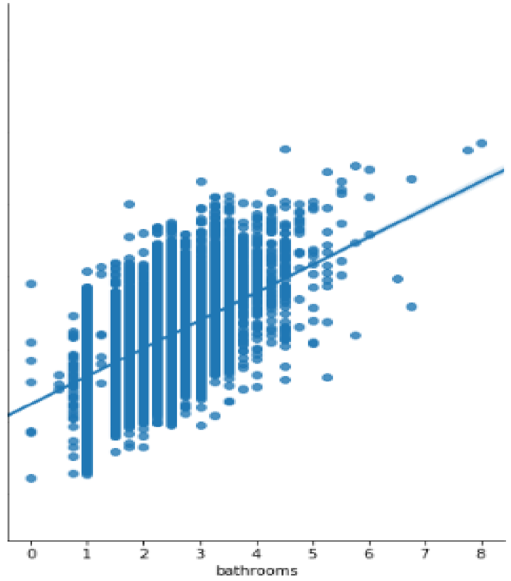
  I have calculated correlation coefficients of all the features with respect to price

These are the pair plots of grade and sqft_living with respect to price. The price of houses is more if the grade and sqft area is more. The correlation coefficient >0.5 and visualization makes it clear that price of houses is highly dependent on grade and Sqft area.

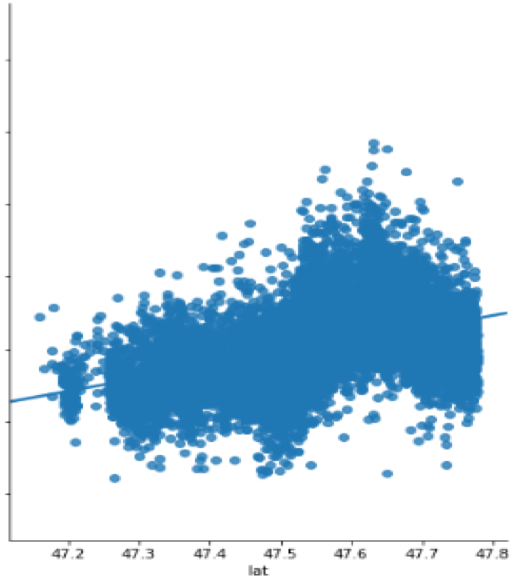


Grade= 0.705

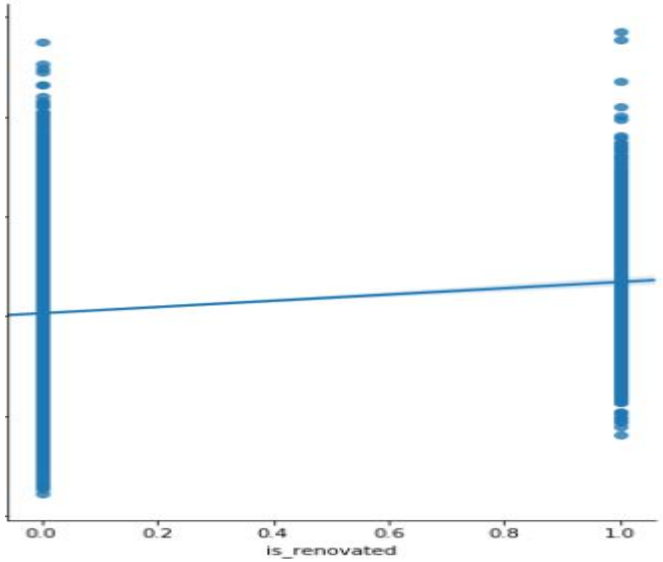
Sqft_living= 0.698



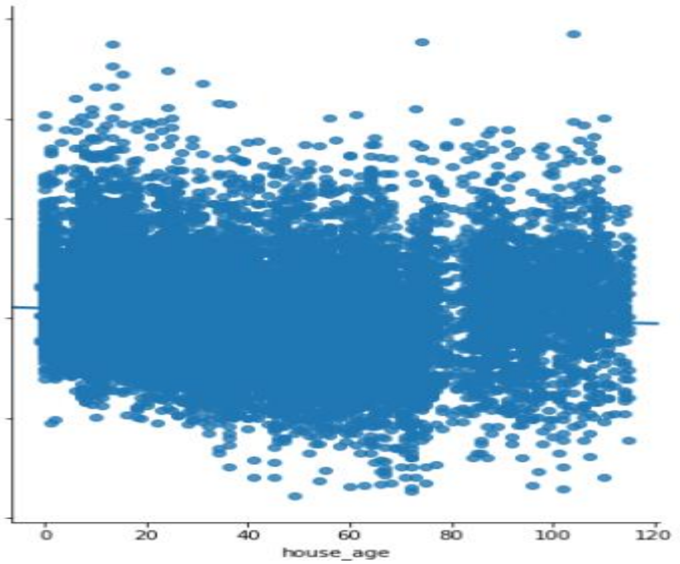
Bathrooms= 0.552



Lat=0.495

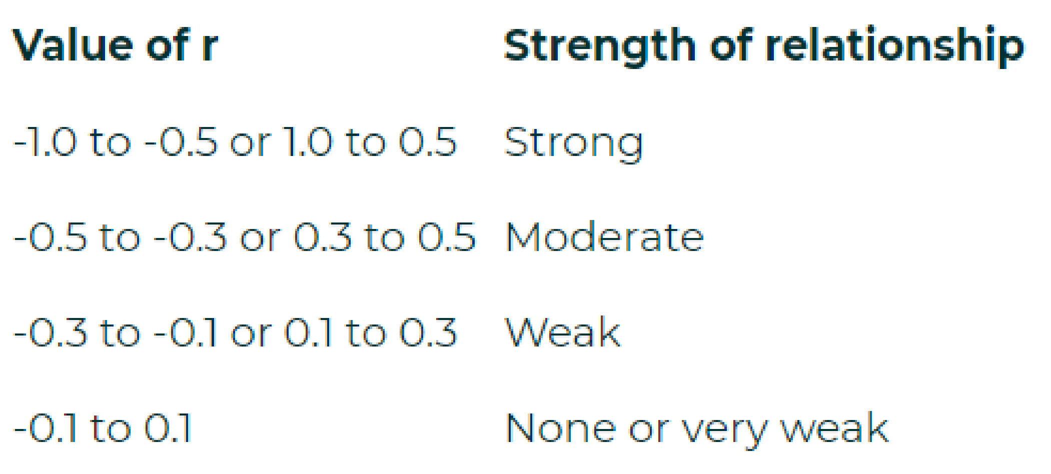


Is_renovated=0.125



House_age=-0.03

- From the above pair plot it is clear that as correlation coefficient decreases, the dependency of price on that feature also decreases.
- This table shows the interpretation of Correlation Coefficients.
- Using this I have selected different features for my model



Value of r	Strength of relationship
-1.0 to -0.5 or 1.0 to 0.5	Strong
-0.5 to -0.3 or 0.3 to 0.5	Moderate
-0.3 to -0.1 or 0.1 to 0.3	Weak
-0.1 to 0.1	None or very weak

Feature Selection

- **Recursive feature elimination:**
- Method works by recursively removing attributes and building a model on those attributes that remain. It uses accuracy metric to rank the feature according to their importance.

- **Backward Feature elimination:**
- Method works by feeding all the possible features to the model at first. We check the performance of the model and then iteratively remove the worst performing features one by one till the overall performance of the model comes in acceptable range
- The performance metric used here to evaluate feature performance is pvalue. If the pvalue is above 0.05 then we remove the feature, else we keep it

Using these different techniques the various feature subsets were found and later used in model implementation.

Model Implementation

1. Simple Linear regression:

- Selected features whose correlation coefficient with price is more than 0.5.
- One of the assumptions of linear regression is that the independent variables need to be uncorrelated with each other so, I discarded those features which were highly correlated to each other
- R squared for this model was low because number of features were less.

2. Complex Linear regression:

- Selected features whose correlation coefficient with price is more than 0.2
- Out of 13 features found, selected best possible 11 features using Recursive elimination technique
- R squared for this model showed much improvement compared to previous model

3. Ridge regression:

- In ridge regression, the cost function is altered by adding a penalty equivalent to square of the magnitude of the coefficients.
- Ridge was implemented by using same set of features in previous model and R squared value increased slightly

4. Polynomial regression:

- The given equation represents how polynomial regression model calculates response

$$Y = \theta_0 + \theta_1 x + \theta_2 x^2$$

- This model was built using the same feature subset in previous model.
- Model gives highest R squared value i.e 77% and least RMSLE.

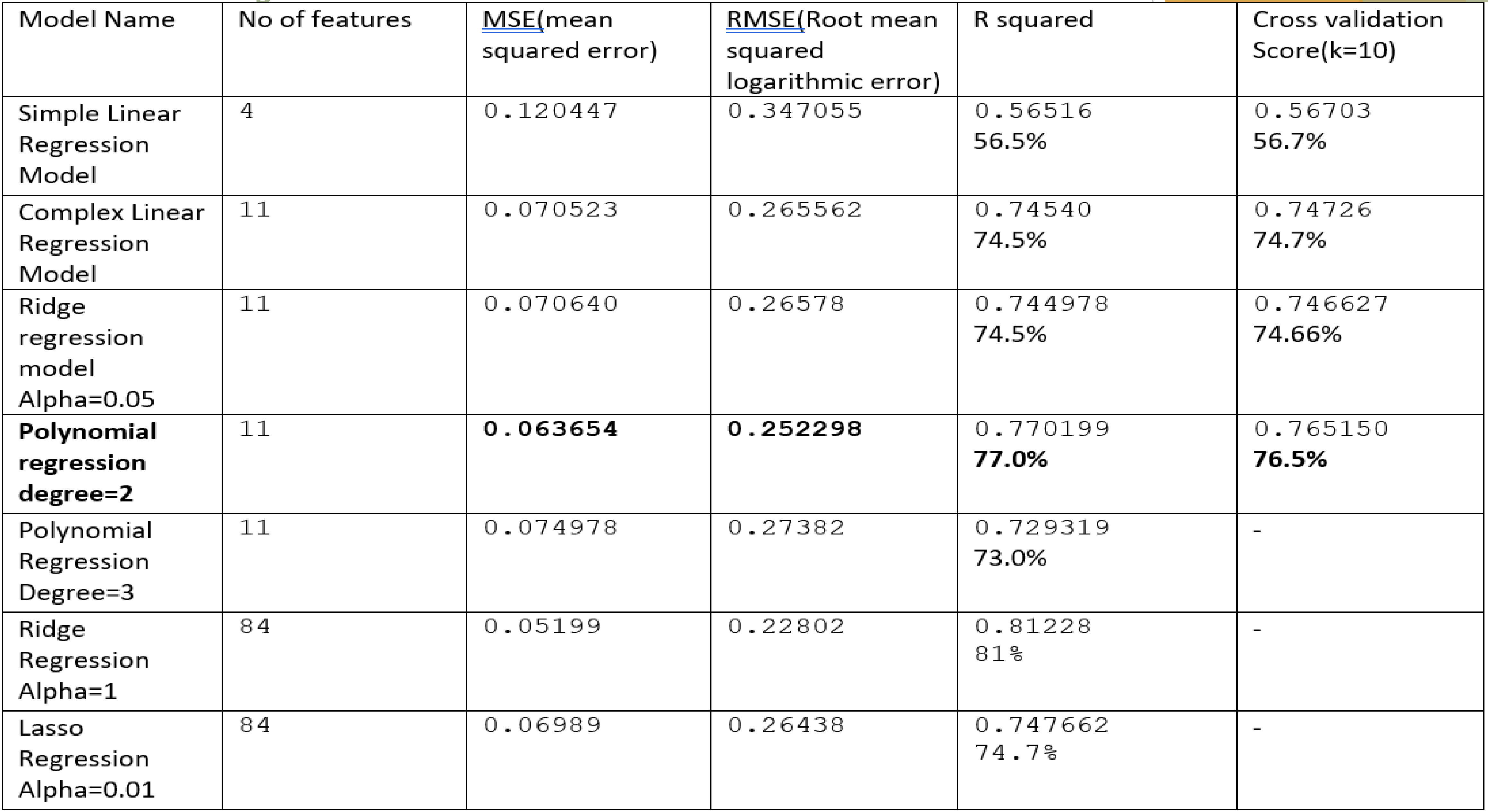
5. Ridge regression with more features:

- Selected all the features in datasets including the computed columns
- Using Backward elimination technique created the feature subset of 84 features
- Model gave R squared value i.e. 81% with alpha value 1
- But I have not selected this model because it has more features.

6. Lasso regression with more features:

- In Lasso model instead of taking the square of the coefficients, magnitudes are taken into account. This regularization leads to zero coefficients which reduces overfitting
- This model with 84 features and alpha value 0.01 gave R squared value 74.7%

Summary of models



Model Name	No of features	MSE(mean squared error)	RMSE(Root mean squared logarithmic error)	R squared	Cross validation Score(k=10)
Simple Linear Regression Model	4	0.120447	0.347055	0.56516 56.5%	0.56703 56.7%
Complex Linear Regression Model	11	0.070523	0.265562	0.74540 74.5%	0.74726 74.7%
Ridge regression model Alpha=0.05	11	0.070640	0.26578	0.744978 74.5%	0.746627 74.66%
Polynomial regression degree=2	11	**0.063654**	**0.252298**	0.770199 **77.0%**	0.765150 **76.5%**
Polynomial Regression Degree=3	11	0.074978	0.27382	0.729319 73.0%	-
Ridge Regression Alpha=1	84	0.05199	0.22802	0.81228 81%	-
Lasso Regression Alpha=0.01	84	0.06989	0.26438	0.747662 74.7%	-

Future Improvements

- More techniques can be used for data cleaning, preprocessing
- Feature selection process can be done more precisely
- Embedded method can be used to decide feature importance
- Explore different machine learning algorithms like Random Forest
- The model can be made more precise by using techniques Gradient boosting, XGBoost
- Evaluate model performance using different accuracy metrics

Learning Outcomes

- Dealing with high cardinality columns
- Computed useful columns
- Learned analyzing and processing large dataset
- Implemented and interpreted some good visualizations
- Learned different techniques of feature selection
- Implemented various machine learning algorithm and evaluated their performances

Thank you!