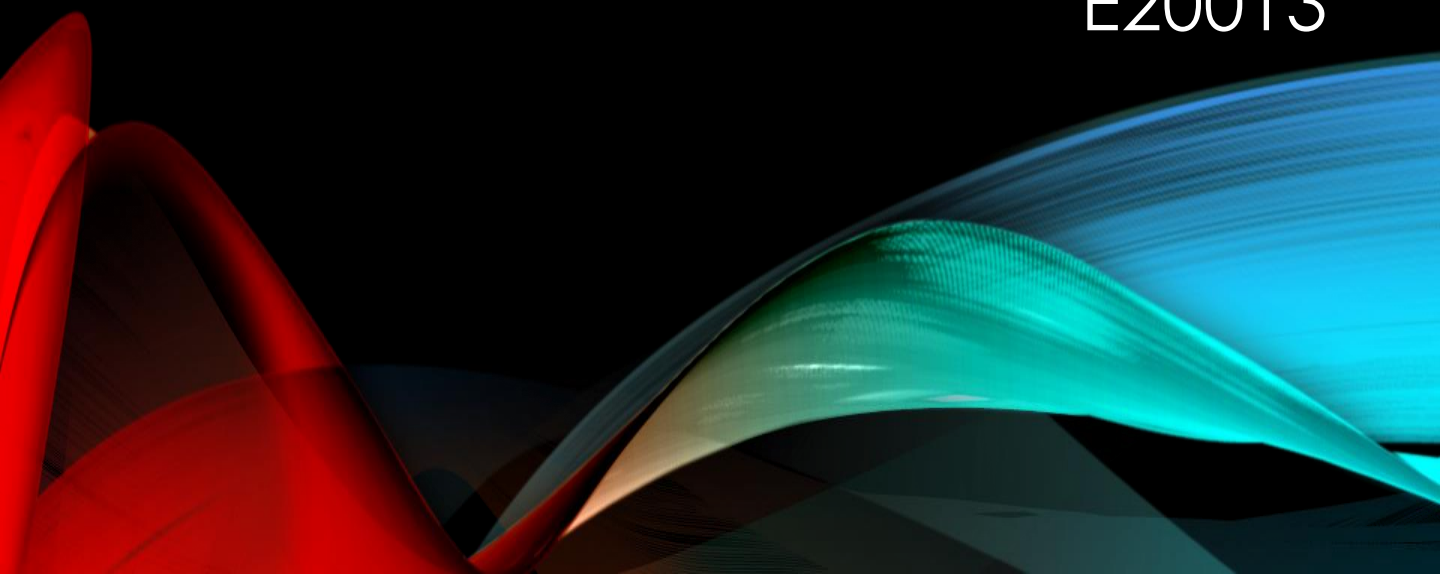
KING COUNTY HOUSE PRICE PREDICTION MODEL

James Ajeeth J
E20013



PROBLEM STATEMENT

King County is located in the U.S. state of Washington. The population was 2,149,970 in a 2016 census estimate. King is the most populous county in Washington, and the 13th-most populous in the United States. The county seat is Seattle, which is the state's largest city. King County is one of the three Washington counties that is included in the Seattle-Tacoma-Bellevue metropolitan statistical area. About two-thirds of King County's population lives in the city's suburbs. As of 2011, King County was the 86th highest-income county in the United States. This paper addresses the factors concerning the "house sale prices" in King County sold between May 2014 and May 2015.

Prices of real estate properties are sophisticatedly linked with country's economy. Despite this, we do not have accurate measures of housing prices based on the vast amount of data available. Therefore, the goal is to use machine learning to predict the selling prices of houses based on many economic factors.

Use Case

House Buyers vs House Sellers:

Client House buyer: This client wants to find their next dream home with a reasonable price tag. They have their locations of interest ready. Now, they want to know if the house price matches the house value. With this study, they can understand which features (ex. Number of bathrooms, location, etc.) influence the final price of the house. If everything matches, they can ensure that they are getting a fair price.

Client House seller: Think of the average house-flipper. This client wants to take advantage of the features that influence a house price the most. They typically want to buy a house at a low price and invest on the features that will give the highest return. For example, buying a house at a good location and with small footage. The client will invest on making rooms at a small cost to get a large return

DATA DICTIONARY

Features	Definition
Id	Unique ID for each home sold
Date	Date of the home sale
Price	Price of each home sold
Bedrooms	Number of bedrooms
Bathrooms	Number of bathrooms
Sqft_living	Square footage of the apartments interior living space
Sqft_lot	Square footage of the land space
Floors	Number of floors
Watefront	A dummy variable for whether the apartment was overlooking the waterfront or not
View	An index from 0 to 4 of how good the view of the property was
Condition	An index from 1 to 5 on the condition of the apartment
Grade	An index from 1 to 13
Sqft_above	The square footage of the interior housing space that is above ground level
Sqft_basement	The square footage of the interior housing space that is below ground level
Yr_built	The year the house was initially built
Yr_renovated	The year of the house's last renovation
Zip Code	What zip code area the house is in
Lat	Latitude
Long	Longitude
Sqft_living15	The square footage of interior housing living space for the nearest 15 neighbors
Sqft_lot15	The square footage of the land lots of the nearest 15 neighbors

DATA PRE-PROCESSING

- **Missing Values Detection:** Missing data pattern is used to identify the missing data in the dataset. From the given training data it can be observed that the data does not consist of any missing data for any of the variables.
- New columns which are essential for analysis, are created by applying new formulae on existing variables. This provides us with a better understanding of the data and improve the accuracy of our model.

The columns transformed are listed below

1. Age: New column Data Type: Numeric Continuous

Calculation: Extracted Year from the date field - Yr_Built

2.Waterfront: Changed Data Type from Numeric continuous to Numeric Nominal

3. **View**: Changed Data Type from Numeric Continuous to Numeric Nominal

4. **Condition**: Changed Data Type from Numeric Continuous to Numeric Nominal

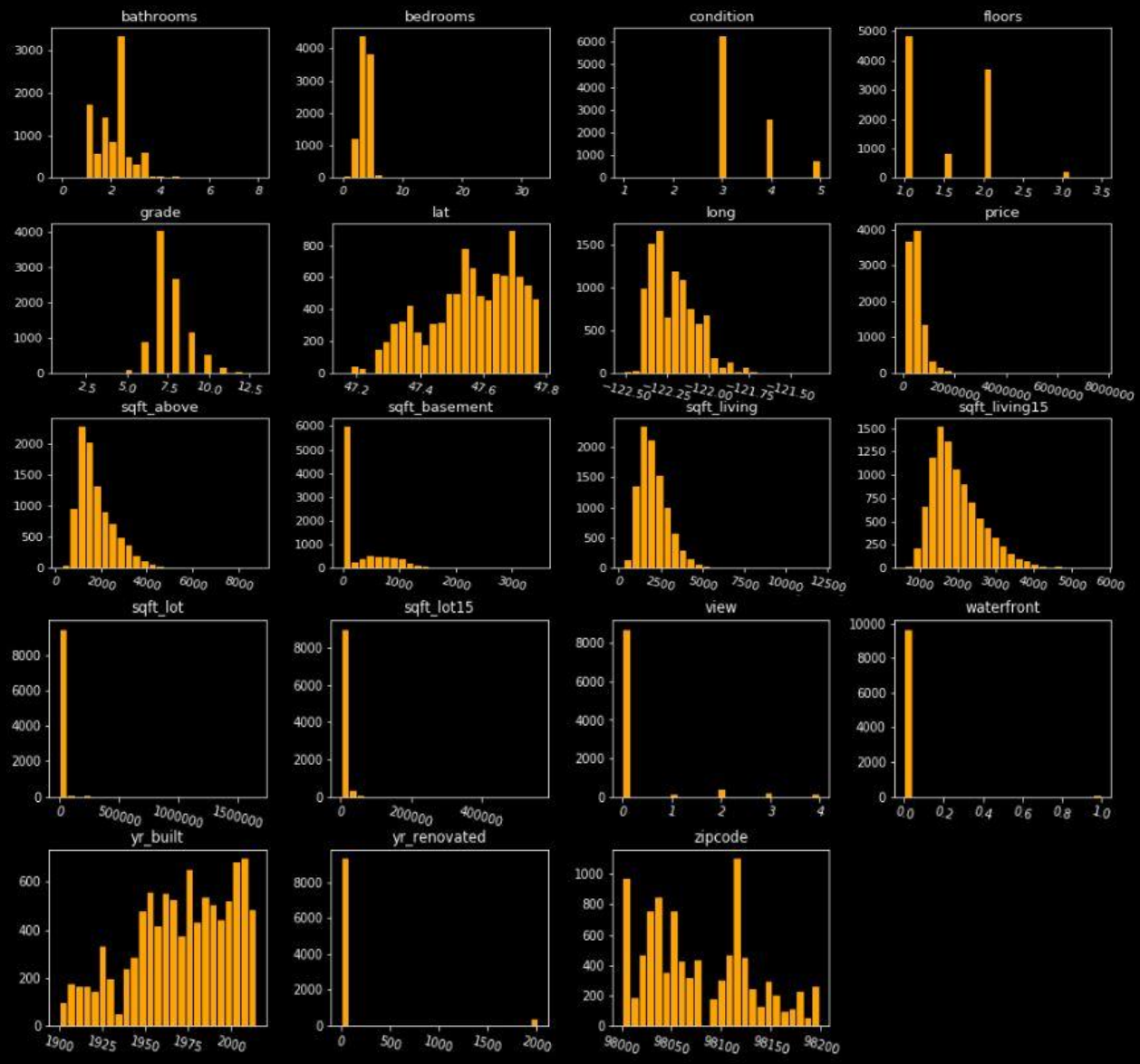
5. **Grade**: Changed Data Type from Numeric Continuous to Numeric Nominal

- While exploring the data it is found that few instances of the data were inconsistent and was not logical. We choose to either make the values consistent by recoding or exclude those instances from the data.

1.One observation with 33 bedrooms in 1620 Square feet with 1.75 bathrooms, it is better to exclude the observations with bedrooms greater than 10

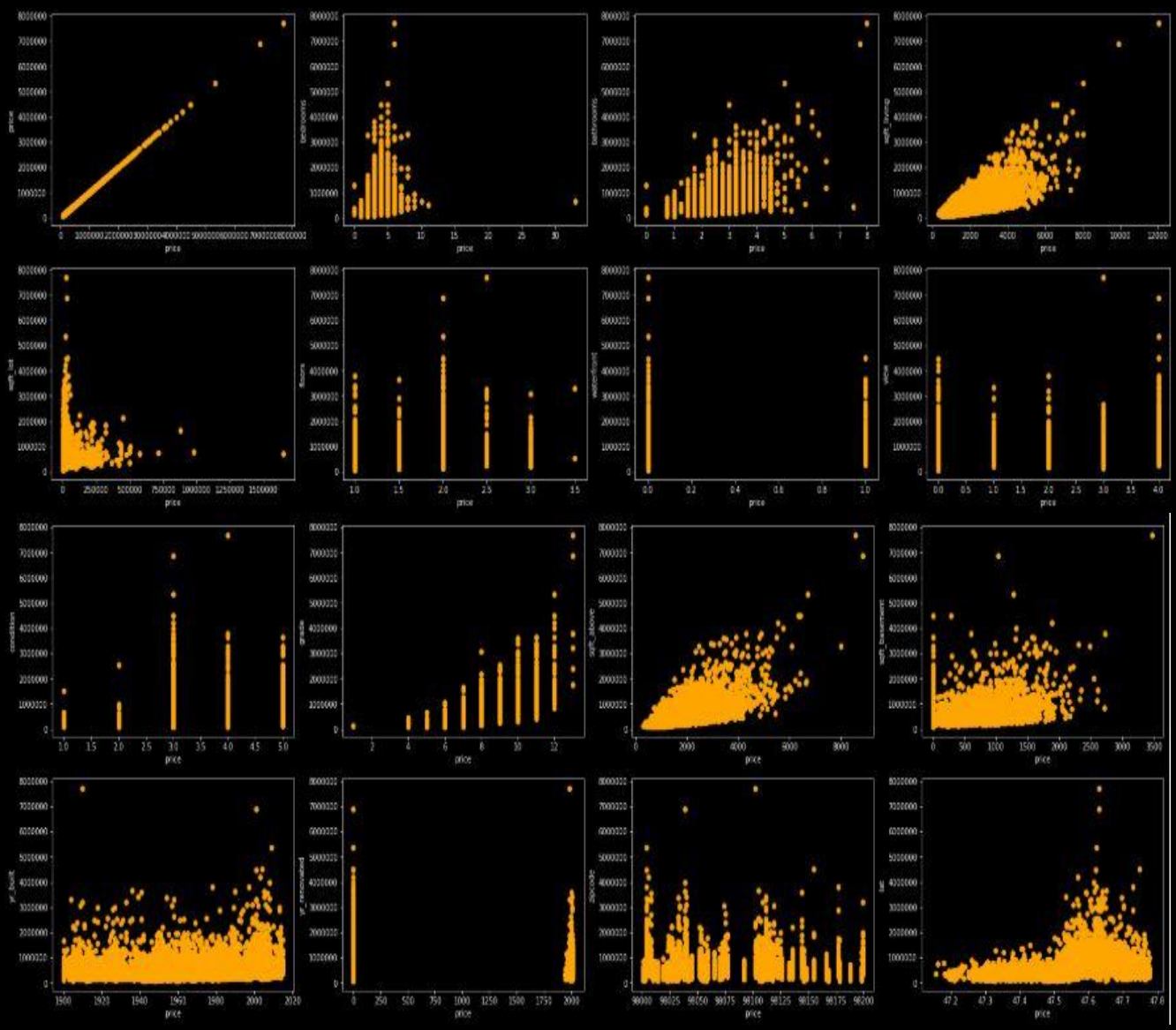
2. Four observations with 0 bathrooms. Since it is not conventional to have houses without bathrooms, it is better to exclude these observations.

UNIVARIANT ANALYSIS - HISTOGRAM



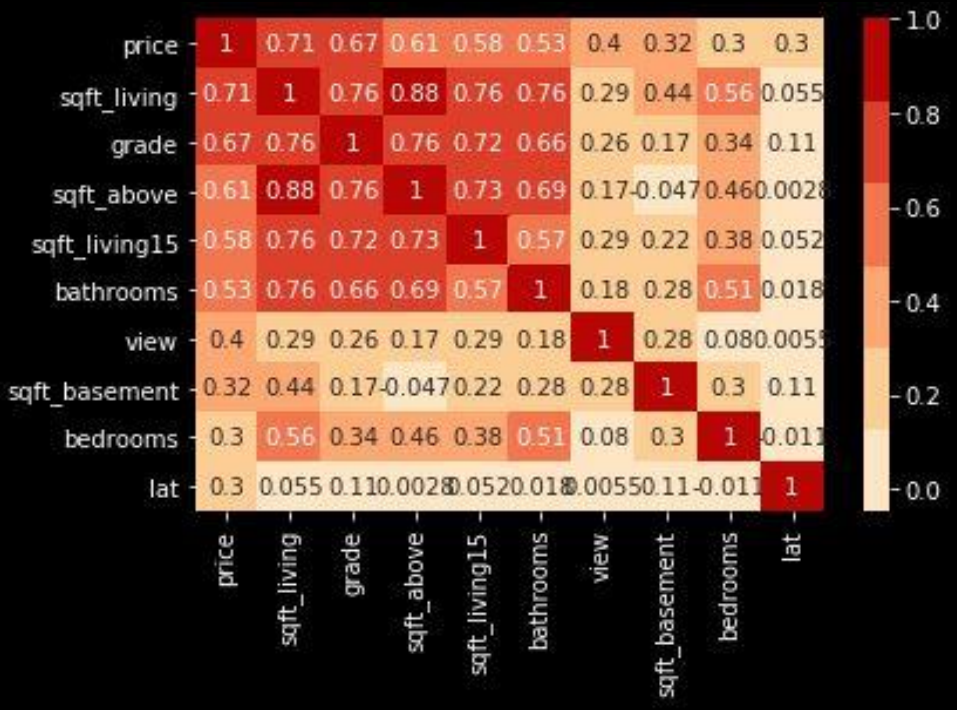
- It can inferred that price, sqft_above, sqft_lot, sqft_living, sqft_basement, sqft_living15, sqft_lot15 features are all right skewed so it is necessary to log transform those feature to make the prediction more accurate.
- Since sqft_living = sqft_above + sqft_basement, it is not necessary to take log transformation on those features as it is not necessary to include them in fitting the model

BIVARIANT ANALYSIS – SCATTER PLOT



- Price increases with increase in Square Feet Living, Square Feet above, Number of Bathrooms and Number of Bedrooms.
- Price increase with increase in Grade and Condition.
- Houses with View 3 and 4 are likely to have a higher price compared to houses with View 0, 1, or 2.

CORRELATION MATRIX



- Correlation gives us the value of how strongly a single variable is linearly associated with another.
- Price has a high positive correlation with Square Feet Living and low positive correlation with Number of Bedrooms, Floors, Square Feet Basement
- Square Feet Living has high positive correlation with Square Feet Above, Square Feet Living15, Number of Bathrooms and Number of Bedrooms features which may explain the same variation in Price as Square Feet Living
- Price of an area increase depending upon the area. So Zip code will play a vital role in predicting the price of the house. Since Zip code is a categorical data, and it shows a negative correlation with the price of the house it is necessary to convert the categorical data into binary forms of column using one hot encoding method. This might increase the prediction of the model.

MODEL BUILDING

In this House price prediction experiment we chose Multiple linear Regression model to predict the house price that will help prospective homeowners, developers, investors and other participants of real estate market

Model 1:

Standard Least Squares: Variable selection is done on the basis of the relationships between house features and price revealed during univariant and bivariant analysis.

Features that are selected : *bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, sqft_above, sqft_basement, yr_built, yr_renovated, zipcode, age, sqft_living15, sqft_lot15*

	Mean Absolute Error	Root Mean Square Error	R Squared Value
Validation	72336.169	125404.843	0.877

Model 2:

Forward Selection

Stepwise model is to identify model with Maximum Validation R-Square. The objective is to compare the best model from forward selection.

Feature selection is done using forward selection method. Starting with one feature, then continue adding extra features and checking if model improves (improved R-squared)or not . If it doesn't, remove the particular feature.

Validation Scores:

	Features	Mean Absolute Error	Root Mean Square Error	R Squared Value
1	Sqft_living	173509.458	255529.738	0.485
2	Sqft_living, Zipcde	114010.674	186771.112	0.725
3	Sqft_living, Zipcde, View	107717.47	175090.842	0.758
4	Sqft_living, Zipcde, View, grade	104691.145	170607.345	0.771
5	Sqft_living, Zipcde, View, grade, Waterfront	102543.594	163060.215	0.79
6	Sqft_living, Zipcde, View, grade, Waterfront, age	100731.934	160796.852	0.796
7	Sqft_living, Zipcde, View, grade, Waterfront, age, condition	100212.358	160522.312	0.797
8	Log(Sqft_living), Zipcde, View, grade, Waterfront, age, condition	75523.716	126432.916	0.865
9	Log(Sqft_living), Zipcde, View, grade, Waterfront, age, condition, sqft_living15	74305.251	124955.018	0.87

MODEL SELECTION

Model 1:

The number of features used is high and it produces an accuracy of about 87.7 % with RMSE 125404.843

Model 2:

The number of features used is low and it produces an accuracy of about 87 % with RMSE 124955.018

Selected model: Model 2 → Low RMSE and high accuracy

Testing our model on test data:

Using the model on the test data,

MAE: 76645.219

RMSE: 130401.557

R2_Square: 0.87

Conclusion:

Even though the model 1 produces high accuracy, choosing the model with low RMSE and least number of variables results a best model. Therefore the suggested model is Model 2 (forward selection).