House Prices Prediction with Data Science

AGENDA

1 Introduction

2 Exploratory Data Analysis

3 Feature Engineering

4 Model & Evaluation

1 Introduction

Business Objectives

1. *To help make home search easier*
2. *Apply machine learning predictive model*
3. *Deliver reliable predicted house values*

Data Overview

1. *House prices in King County, WA*
2. *Median household income in King County*
3. *The dataset contains 21 variables*

Data source:

- https://www.kaggle.com/harlfoxem/housesalesprediction/data
- https://censusreporter.org/data/table/?table=B19013&geo_ids=05000US5 3033,860|05000US53033&primary_geo_id=05000US53033

2 Exploratory Data Analysis

House Prices Distribution

Look into the distribution of house prices using distribution plot in Python Seaborn data visualization library

1. *Prices deviate from normal distribution*
2. *Skewed to the right*
3. *Potential outliers with prices over $6M*

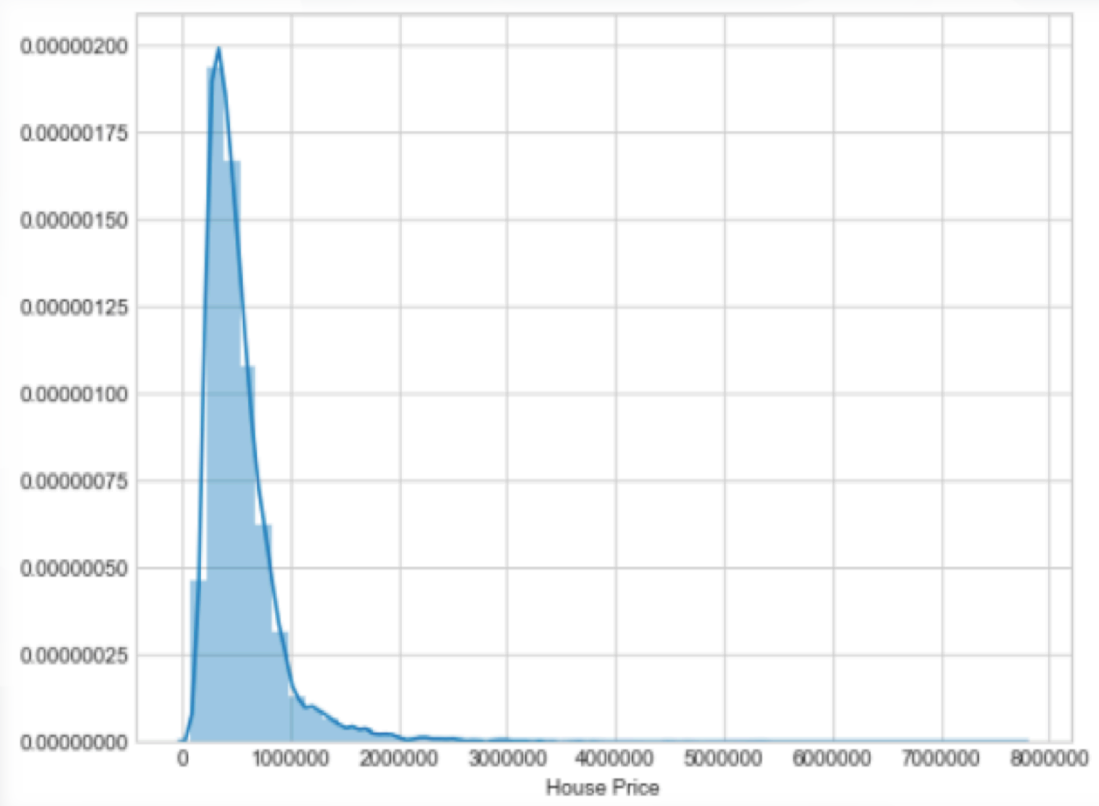


Figure1 Seaborn Distribution Plot

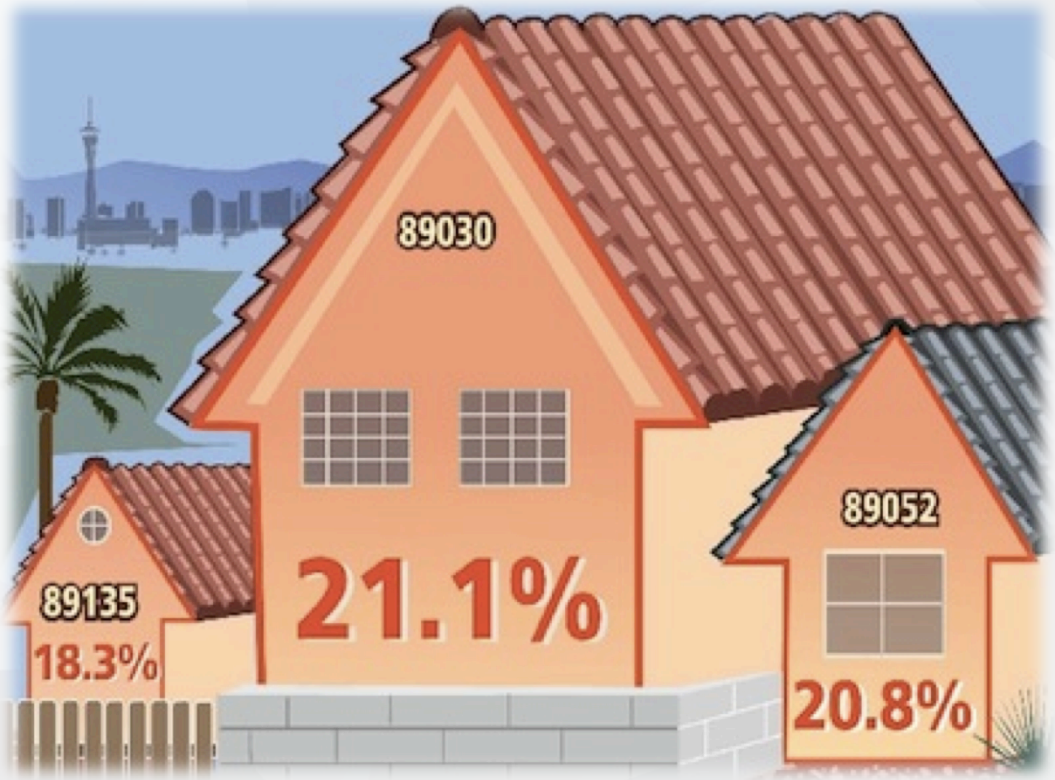
Does Zip-code Reflect on House Price?

Location usually plays an important role when it comes to house price.

1. *Visualize the distribution of housing price based on zip code in boxplot (Figure2)*

2. *Based on the boxplot, zip codes do give us information of house prices (Figure2)*

3. *Zip code is a good indicator when doing home search in certain areas*



Does Zip-code Reflect House Price?

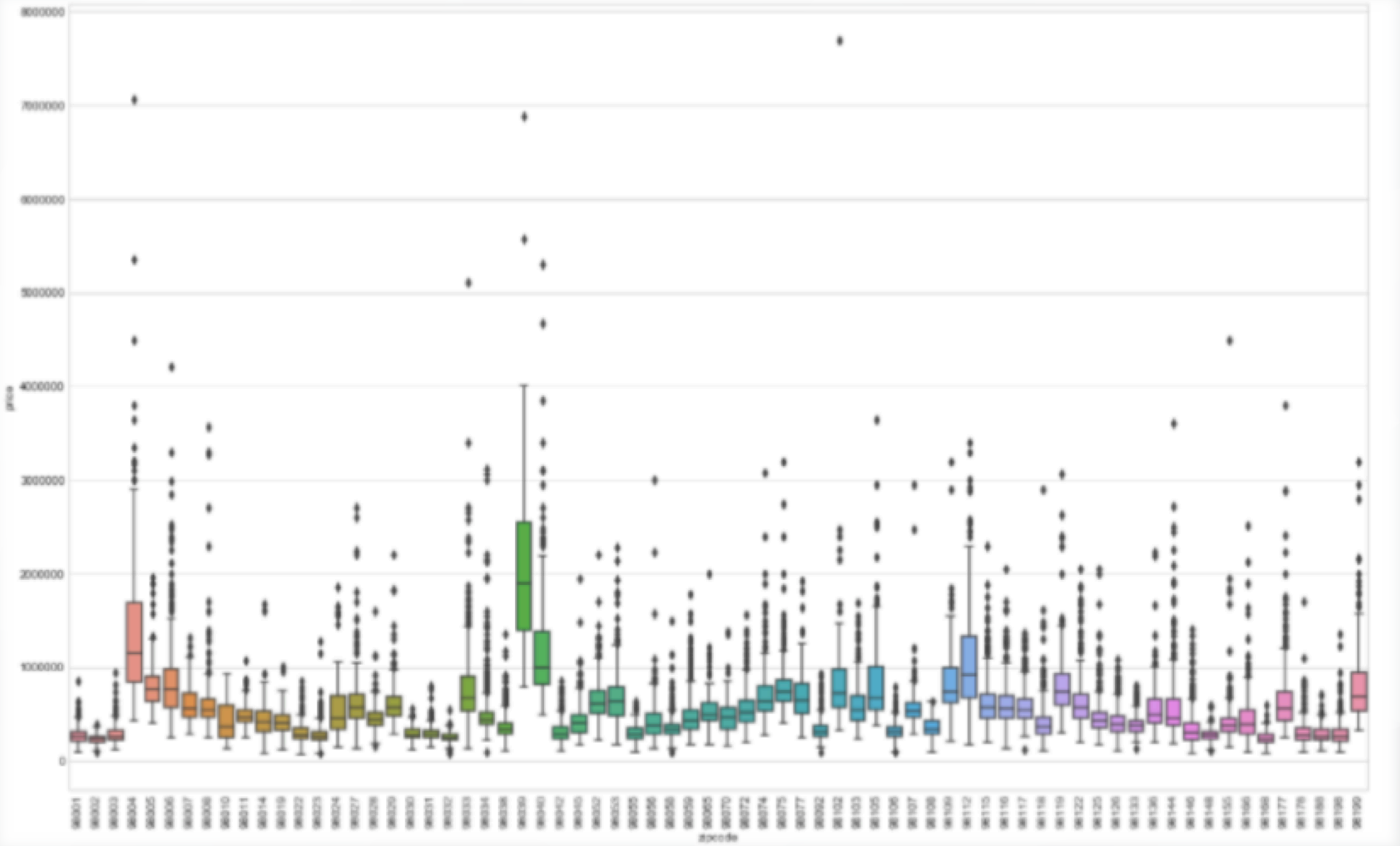
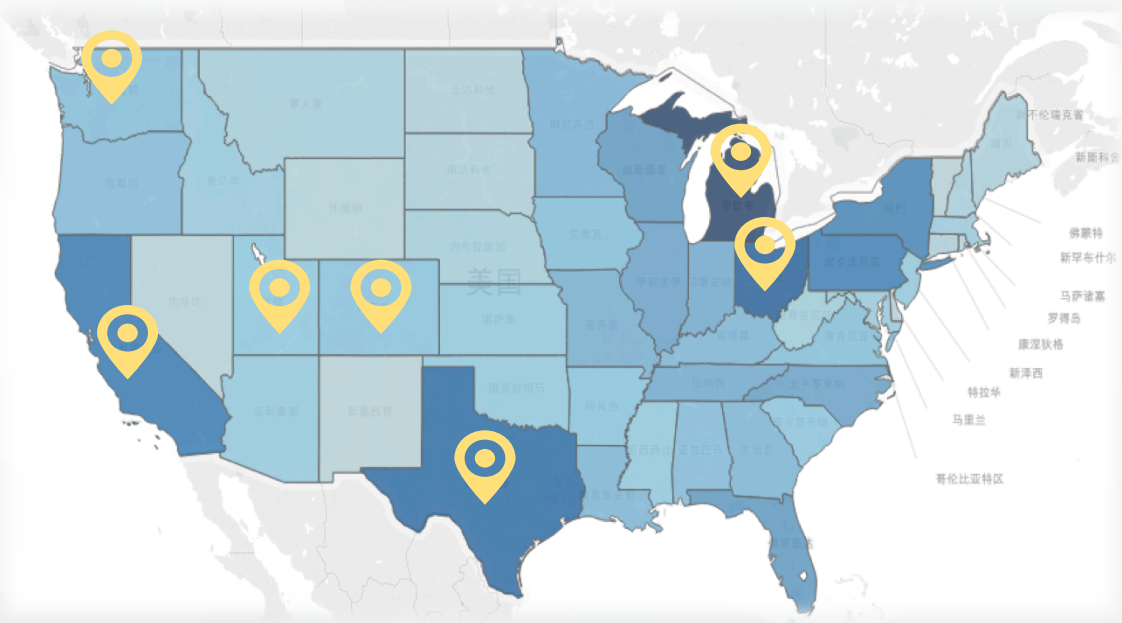


Figure2 Boxplot Zip-code vs House price

Where are houses located in?

Visualize the locations of houses on a map along with house prices

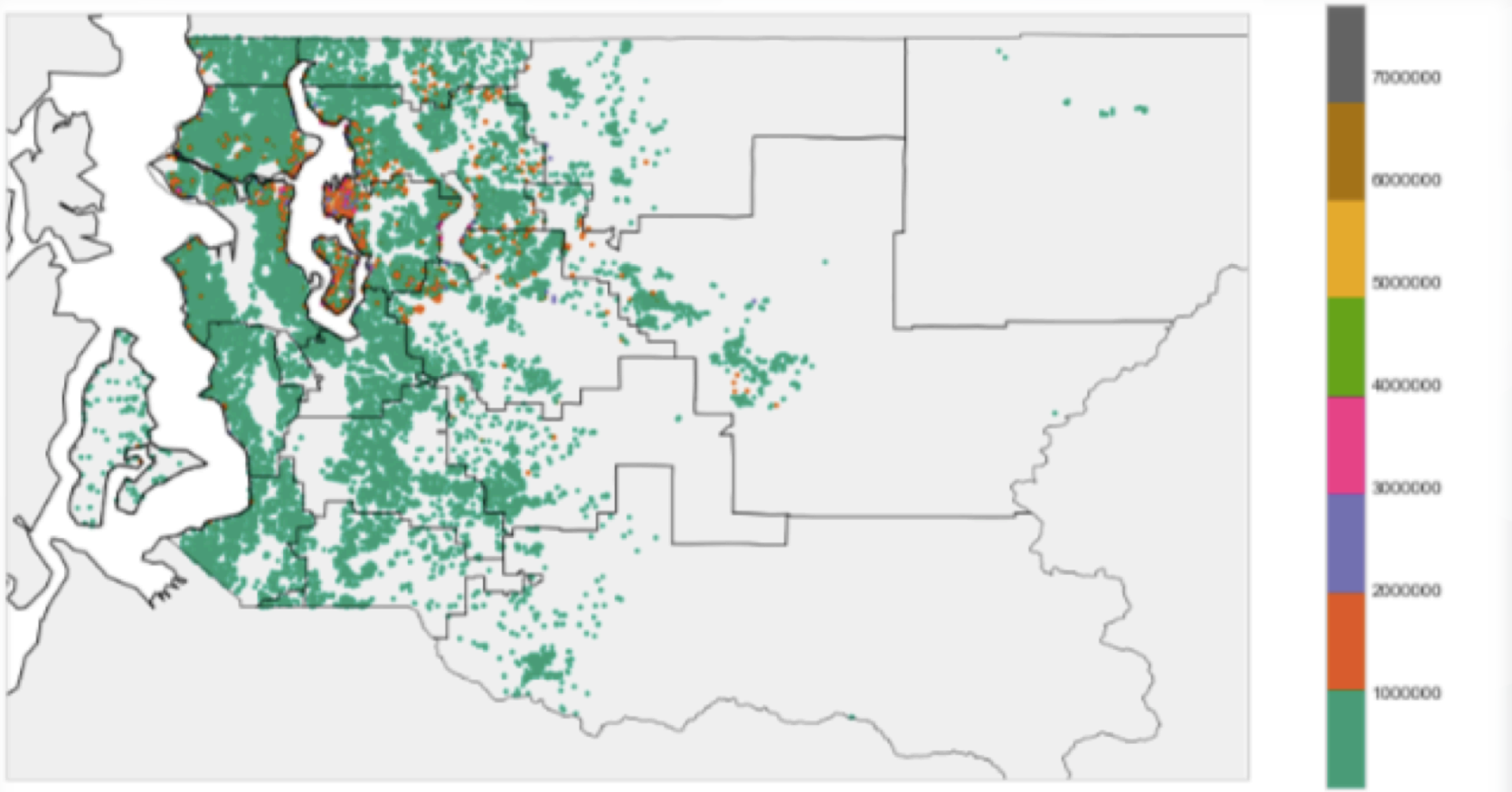
1. *With map, consumers can search houses easily in certain areas with price range*
2. *Create a map of King County neighborhoods using Python Basemap (Figure3)*
3. *Each dot is a house with a price, and the bar at the right is price range (Figure3)*



Where are houses located in?

Figure3 King County Neighborhood Map along with Price Bar

- Most of the houses have higher price in Greater Seattle area.
- Some houses in suburban areas also have higher prices.



3 Feature Engineering

Correlation Heatmap

Correlation heat-map presents the correlation between each variable

Check Multi-Collinearity

- Two variables have > 90% correlation
- Multi-Collinearity makes it hard to assess the effect of independent variables on dependent variable
- The heatmap shows no correlation between two variables > 90%, thus no multi-collinearity issue

Normality Test

Check the normality of variables before applying models in order to reduce the chances of getting misleading results

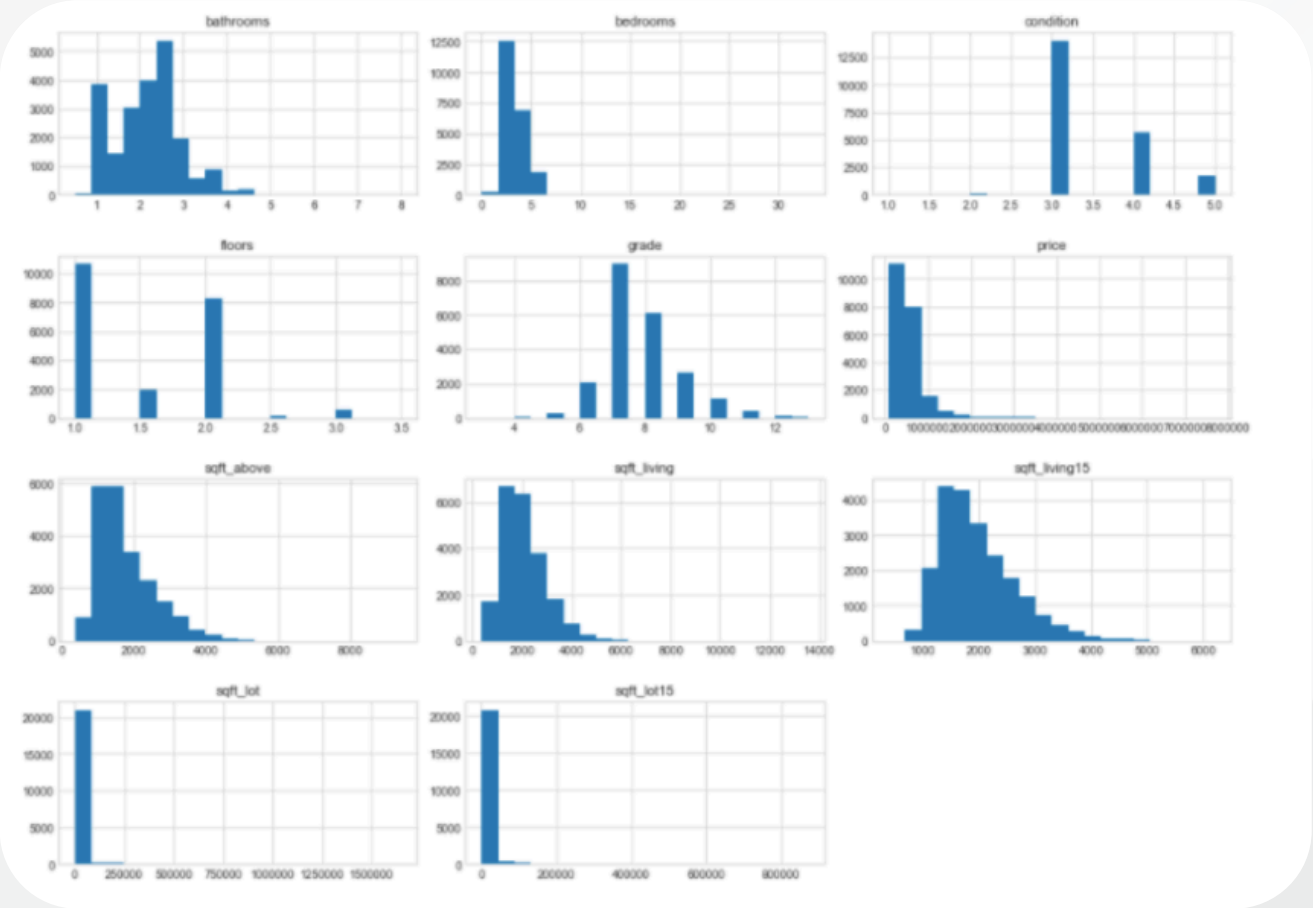


Figure 4 Histogram

- Plot out histograms to look into the distribution of each variable
- Most of the variables are skewed to the right

Normality Test

Transform data to make data normally distributed if data is not normal distribution

	skewness
sqft_lot	13.057033
sqft_lot15	9.505720
price	4.025608
bedrooms	2.002336
sqft_living	1.472613
sqft_above	1.446937
sqft_living15	1.106906
condition	1.035838
grade	0.785404
floors	0.614549
bathrooms	0.519449

Figure 5 Skewness

Skewness

- Use skewness to check normality
- Skewness close to 0 is normal distribution
- Skewness > 0 means skewed to the right
- Based on Figure 5, most of the variables are skewed to the right with skewness > 0

Box-Cox Transformation

Apply box-cox transformation to get transformed data

Before box-cox transformation

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sqft_above
0	7129300520	2014-10-13	221900.0	3	1.00	1180	5650	1.0	0	0	...	7	1180
1	6414100192	2014-12-09	538000.0	3	2.25	2570	7242	2.0	0	0	...	7	2170
2	5631500400	2015-02-25	180000.0	2	1.00	770	10000	1.0	0	0	...	6	770
3	2487200875	2014-12-09	604000.0	4	3.00	1960	5000	1.0	0	0	...	7	1050
4	1954400510	2015-02-18	510000.0	3	2.00	1680	8080	1.0	0	0	...	8	1680

5 rows x 21 columns

Box-Cox Transformation

Apply box-cox transformation to get transformed data

After box-cox transformation

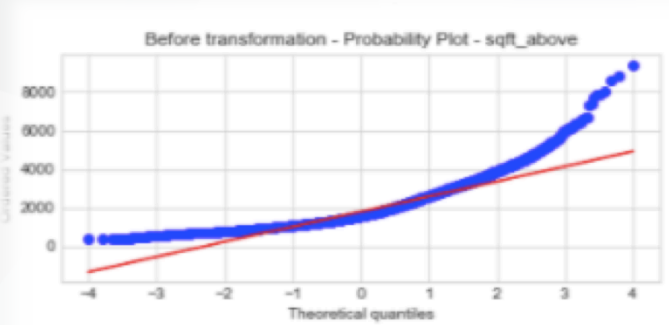
	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sq
0	7129300520	2014-10-13	35.583909	1.540963	0.730463	12.597323	17.696152	0.730463	0	0	...	2.440268	12
1	6414100192	2014-12-09	41.586543	1.540963	1.289269	14.981646	18.620287	1.194318	0	0	...	2.440268	14
2	5631500400	2015-02-25	34.278249	1.194318	0.730463	11.403697	19.874209	0.730463	0	0	...	2.259674	11
3	2487200875	2014-12-09	42.431400	1.820334	1.540963	14.119786	17.253669	0.730463	0	0	...	2.440268	12
4	1954400510	2015-02-18	41.201236	1.540963	1.194318	13.644922	19.038978	0.730463	0	0	...	2.602594	13

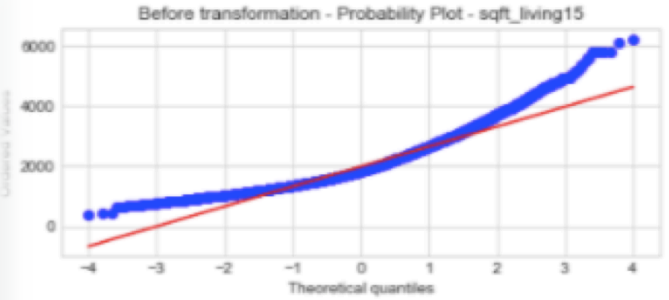
5 rows × 21 columns

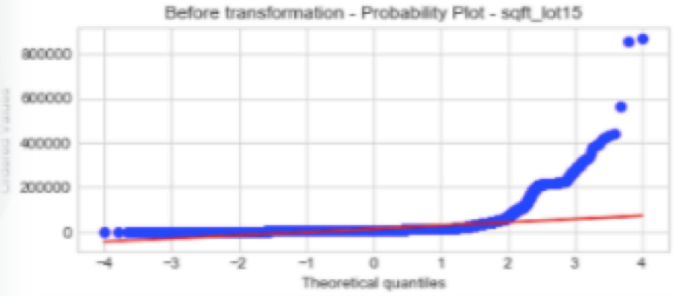
Probability Plot

Use probability plot to check if the transformed data are normally distributed

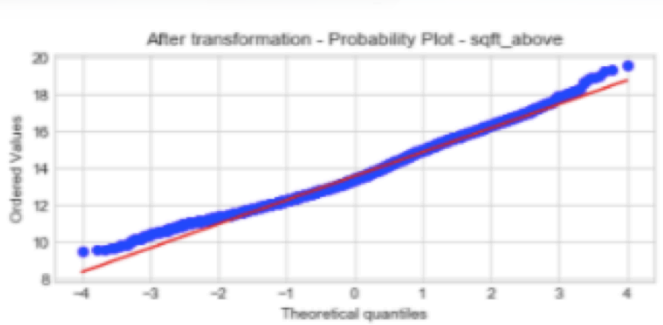
Before box-cox transformation

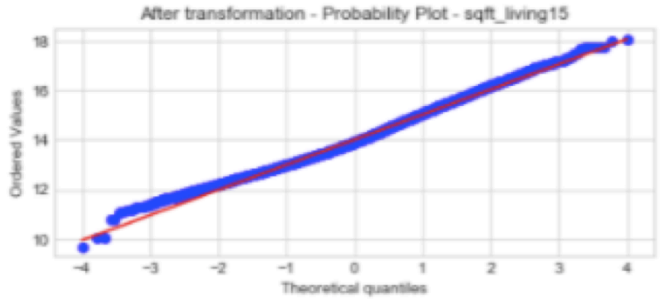


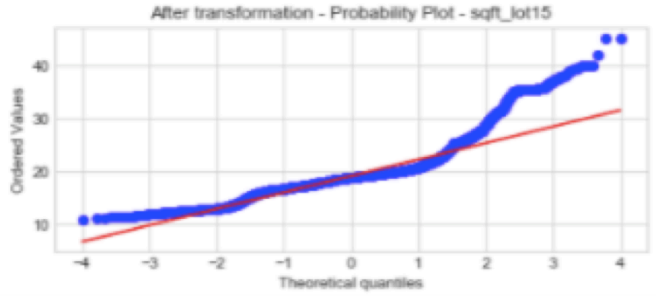




After box-cox transformation







- Variables after transformation (right hand side) are closer to normal distribution even.


4
Model & Evaluation

Predictive Models

1. Ridge Regression
2. Random Forest Regressor
3. Gradient Boosting with Grid Search

01 Ridge Regression

What is Ridge Regression

- Linear least squares with L2 regularization
- L2 regularization is a technique used by Machine Learning training algorithm to reduce model over-fitting.

L2 Regularization

- Pros
  - L2 regularization can be used in many types of training algorithms.
  - Big penalty on large weights
- Cons
  - If there are irrelevant features in the input, L2 regularization will give them small weights rather than 0

02 Random Forest Regressor

What is Random Forest

- Random Forest makes the decision tree building process use different predictors to split at different times.

Random Forest

- Pros
  - Use averaging to improve the predictive accuracy and reduce model over-fitting.

  - Boostrap samples when building trees.

- Cons
  - Sometimes can be computationally expensive

04 Gradient Boosting with Grid Search

What is Grid Search

- Grid search is used for regularization hyper parameter selection.

- Using grid search k-fold cross validation to reduce model over-fitting and find the optimal hyper parameters.

Gradient Boosting with Grid Search

- Pros
  - Reduce variance (over-fitting) and thus higher accuracy

- Cons
  - Computationally expensive and takes more time to run the model

05 Model Training & Test Score

Gradient Boosting with Grid Search

```
Test data score     :  0.721614432648
Training data score:  0.758956656955
```

- Gradient Boosting with Grid Search generates better prediction result with hyper-parameter selection and improve the overall model.
- Test score increased to 72% from 62%

Random Forest Regressor

```
Test data score     :  0.664314490415
Training data score :  0.677593177162
```

- Random Forest regressor improved the model from 62% to 66% without grid search.

Ridge Regression

```
Test data score     :  0.629247618078
Training data score :  0.618334226155
```

- Although Ridge Regression has L2 regularization, without grid search tuning hyper-parameter we can still get model with low accuracy.
- Ridge Regression model only has 62% test score.

06 Recommendations

Collect more data

- Obtain larger amount of data with more features so that we can fit the predictive model with more data points.

Compare House Prices with Other City

- Compare the Seattle Metro area's data against other metropolitan area that are also have expansions in population and property prices.

Bias and Variance Trade-off

- Balance the trade off between bias and variance when applying machine learning models.
- Overly simplistic assumptions lead to bias
- Too much complexity in training data leads to variance