Business Understanding

House market is a big yet complex market and is often affected by the overall economy, policies and neighborhood, making house prices difficult to predict. Besides, information asymmetry in the house market makes the prediction of future house prices more challenging.

Thus, the application of advanced data analysis and machine learning plays an important part in house prices prediction. The goal of this project is to apply machine-learning models to predict future house prices with accuracy and classify houses in order to provide a healthier house market by reducing information asymmetry with the application of data science.

Data Overview

The first dataset is house prices in King County, WA, USA from 2014 to 2015.
The second dataset is King County median household income.
Dataset:

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

The given dataset has no missing values and contains 21 variables.
However, there are 4 variables need to be converted into the correct data type.

- *date: the date when the house was sold*
- *zipcode: zip code where the house is located*
- *yr_built: the year when the house was built*
- *yr_renovated: the year when the house was renovated*

The data type of variable "*date* " is object, and we need to convert it to date time format otherwise we won't be able to conduct Time Series model. We also need to convert variables "*zipcode*", "*yr_built*" and "*yr_renovated*" into categorical data type from numeric number since those variables represent the area code and year rather than the values.

Exploratory Data Analysis (EDA)

Distribution of Housing Prices in King County

First, look into the distribution of housing prices using distribution plot and Empirical Cumulative Distribution Function (ECDF).

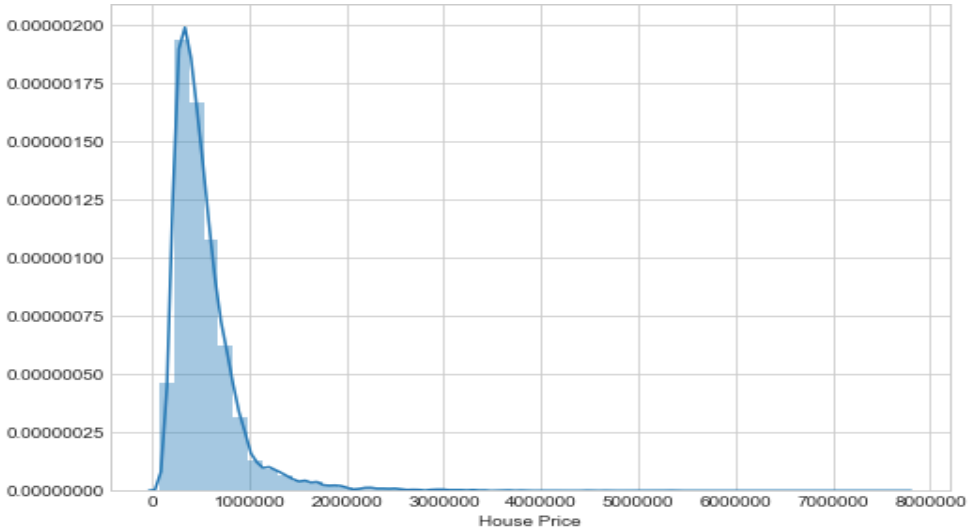


Figure1 Seaborn Distribution plot

The distribution of housing prices is deviated from normal distribution and is skewed to the right; we should take log of house price for further analysis.

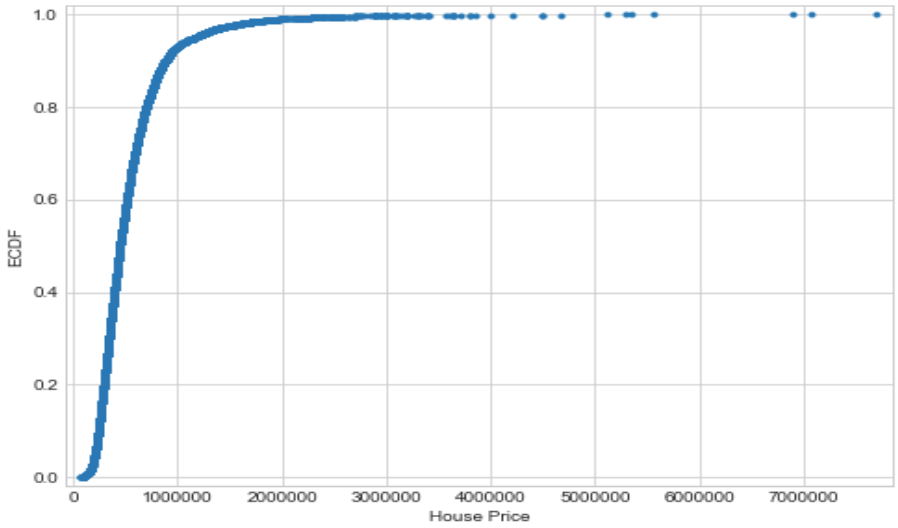


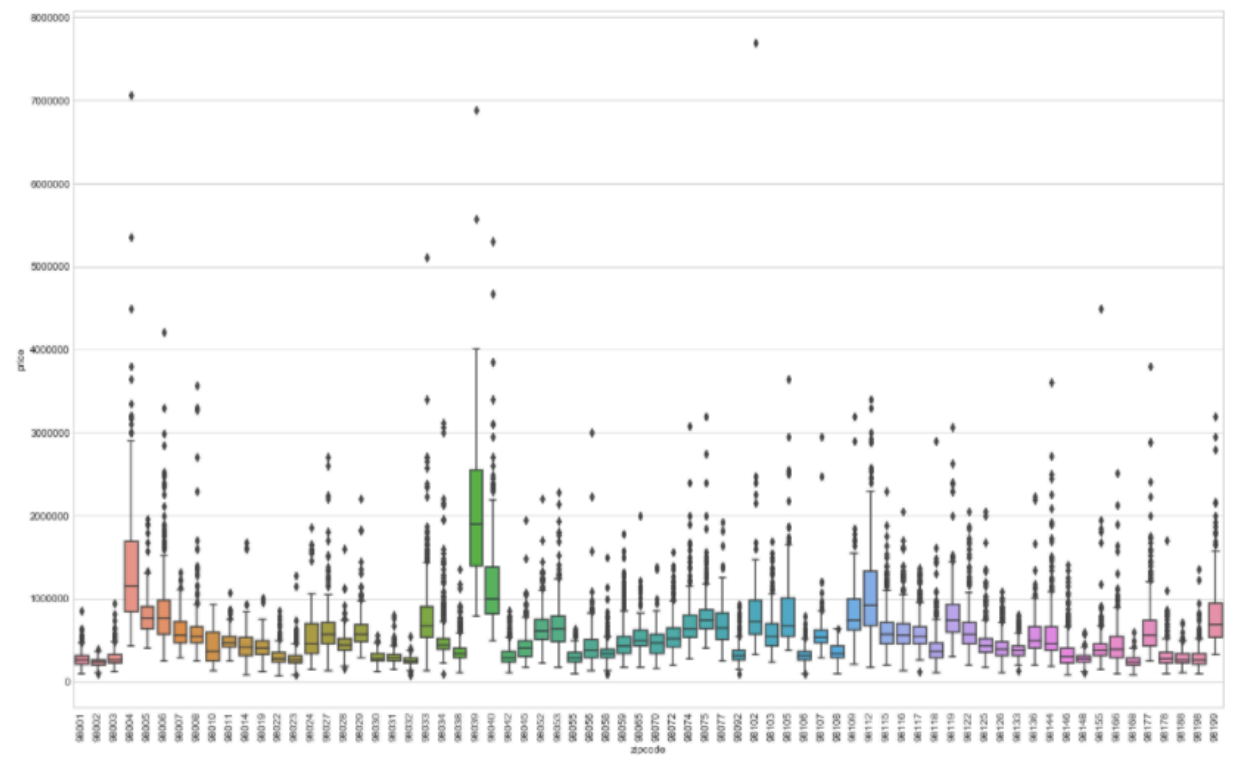
Figure2 ECDF Plot with x-axis "House Price" and y-axis "Probability"

From ECDF plot, we can see that there are 3 houses with prices over $6M, which deviate from normal price distribution.

Zip code and Housing Price

Location usually plays an important role when it comes to housing price. We use boxplot to visualize the distribution of housing price based on zip code and the plot confirms our assumption that zip code does reflect housing price and can be a good indicator when searching for houses in certain locations.

Figure3 Boxplot – House price



Income and Housing Price

We want to know if areas with higher housing price also have higher income.

- Adding King County's median household income in the past 12 months.
- Performed Linear Regression (OLS) model with median household income as independent variable.
- From OLS regression results, the house price and income are statistical significant and the coefficient between house price and income is 0.967.
- From the regression plot, we can see that median house price and median income are positively correlated.

Figure4 OLS Regression Results

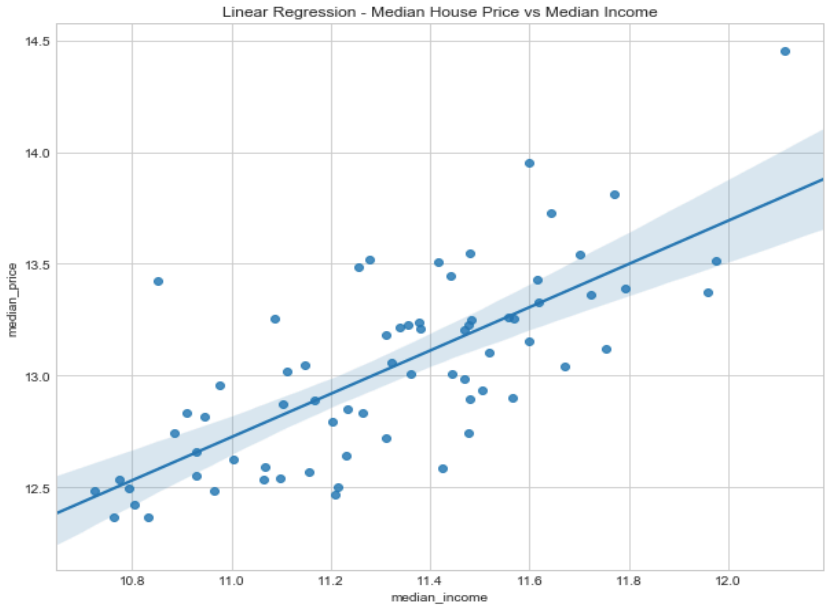
OLS Regression Results

Dep. Variable:	np.log(median_price)	R-squared:	0.542
Model:	OLS	Adj. R-squared:	0.535
Method:	Least Squares	F-statistic:	80.39
Date:	Sun, 11 Feb 2018	Prob (F-statistic):	3.89e-13
Time:	16:43:32	Log-Likelihood:	-10.606
No. Observations:	70	AIC:	25.21
Df Residuals:	68	BIC:	29.71
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.0811	1.222	1.703	0.093	-0.357	4.519
np.log(median_income)	0.9677	0.108	8.966	0.000	0.752	1.183

Omnibus:	5.648	Durbin-Watson:	1.366
Prob(Omnibus):	0.059	Jarque-Bera (JB):	4.957
Skew:	0.634	Prob(JB):	0.0838
Kurtosis:	3.301	Cond. No.	408.

Figure5 Linear Regression – Median Housing Price vs Median Income

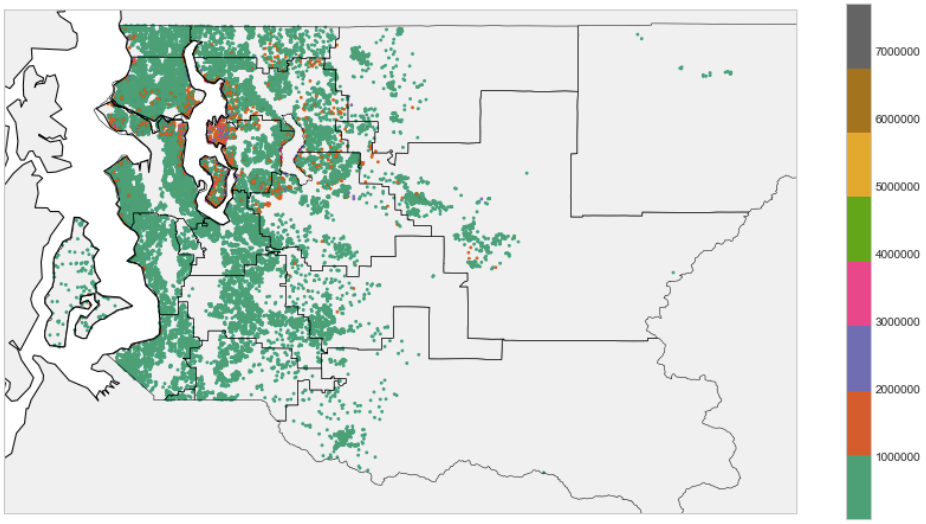


Where are those houses located in?

Visualize the locations of houses on a map with housing price as an indicator so that consumers can easily search houses in certain areas with price range.

- Create a map of King County neighborhoods using Python Basemap.
- Each dot is a house with a price, and the bar at the right is price range.

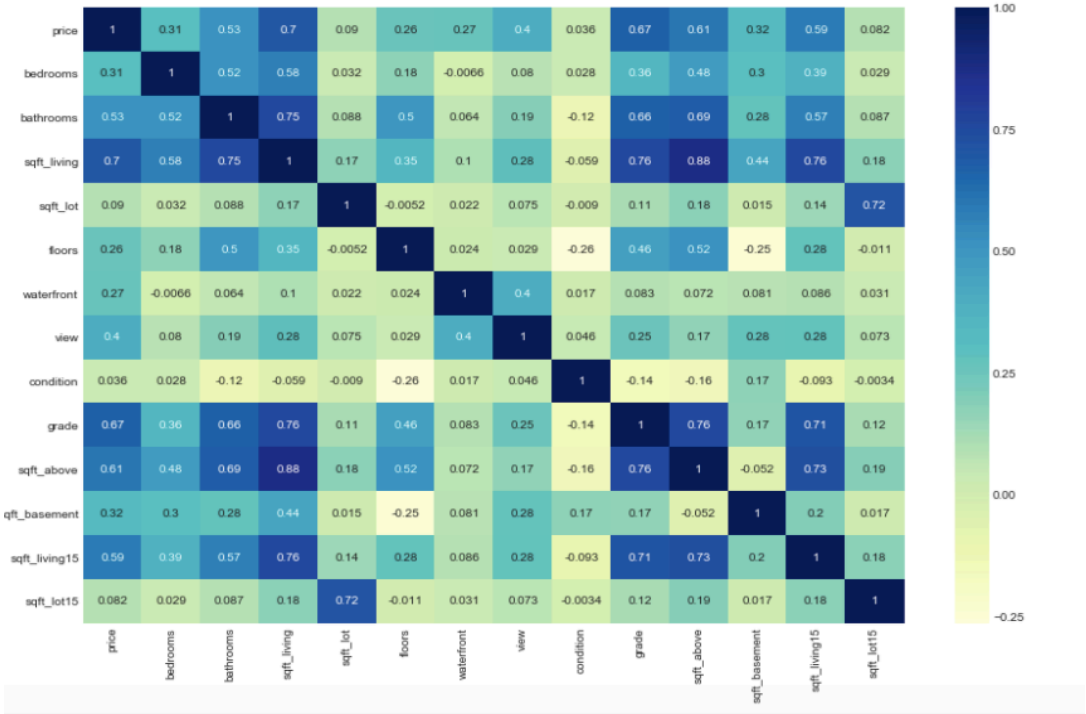
Figure6 Location of each house and its price



Correlation Heatmap

We created a correlation heat-map presenting the correlation between each variable in order to check if there's multi-collinearity. We should avoid multi-collinearity issue since multi-collinearity makes it very difficult to assess the effect of independent variables on dependent variables. Usually, if two variables have over 90% correlation, we considered multi-collinearity issue.

Figure7 Correlation Heatmap

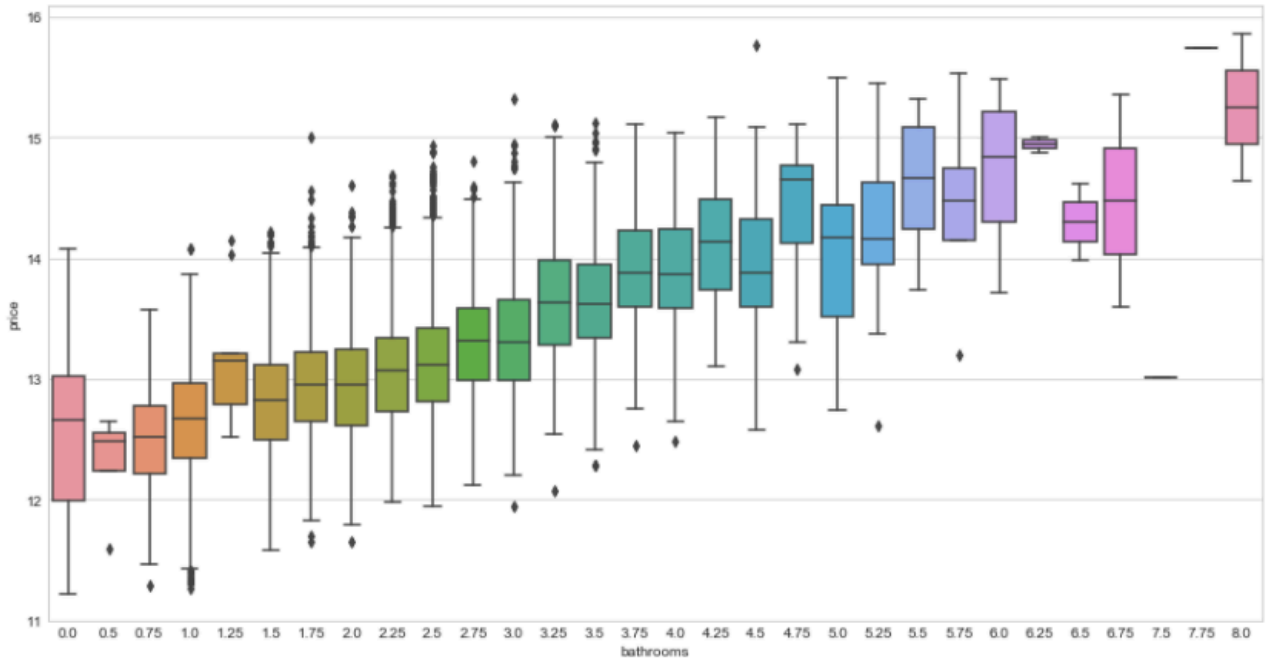


Based on the heatmap above, correlations between variables are < 90%, we can consider there's no multi-collinearity issue.

Housing Price vs. Bathroom

The boxplot gives us information that there are houses with 0 bathroom, which is not reasonable, and should be dropped out. Also, houses with more bathrooms have higher housing price.

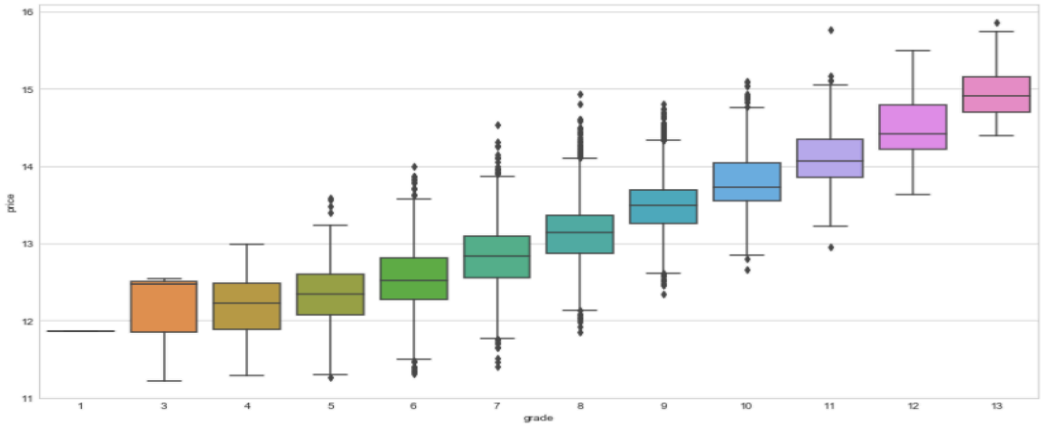
Figure8 House price vs #bathrooms



Housing Price vs. Grade

Grade is a good indicator regarding housing price because houses with Grade labeled as 10 have the highest value and 1 as the lowest.

Figure9 House price vs Grade



Feature Engineering

Dummy Variables

Convert variables into dummy variables before implementing predictive machine learning model. We convert values into 0 and 1, since it's a yes or no question with 0 means "No" and 1 means "Yes".

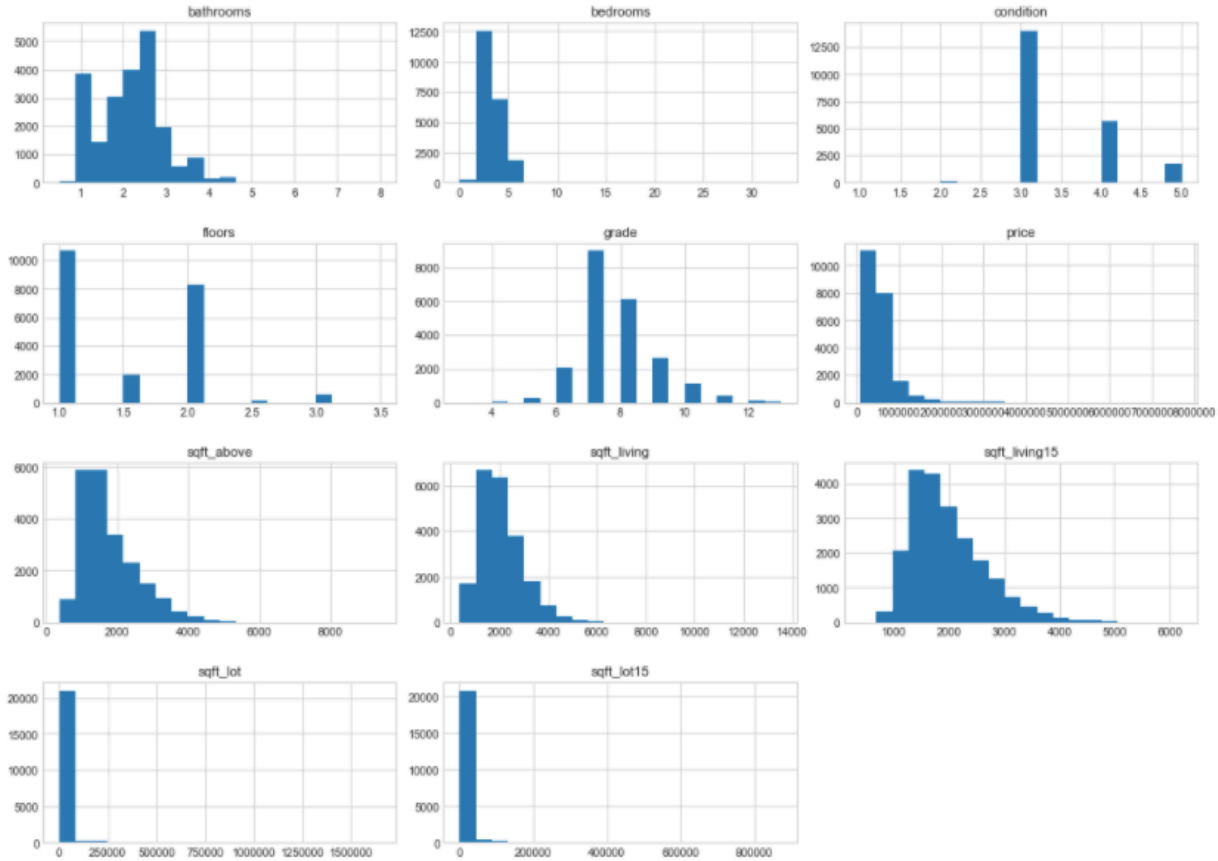
- *sqft_basement: house with basement '1', without basement '0'*
- *yr_renovated: house had renovation '1', no renovation '0'*
- *view: house has view '1', no view '0'*

Normality Test

Check the normality of variables before applying models in order to reduce the chances of getting misleading results and to generate higher accuracy.

- Plot out histograms to look into the distribution of each variable.
- Use skewness to check normality. Skewness close to 0 is normal distribution; skewness > 0 means skewed to the right.

Figure10 Histogram- Most of the variables are skewed to the right



Skewness

The normality test results show variables are skewed to the right (skewness > 0) and thus need transformation to make data normally distributed.

	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

Box-Cox Transformation

First we apply Python Scipy's box-cox transformation to get transformed data then use probability plot to check if the transformed data is normally distributed.

Figure11 - 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 × 21 columns

Figure12 – 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

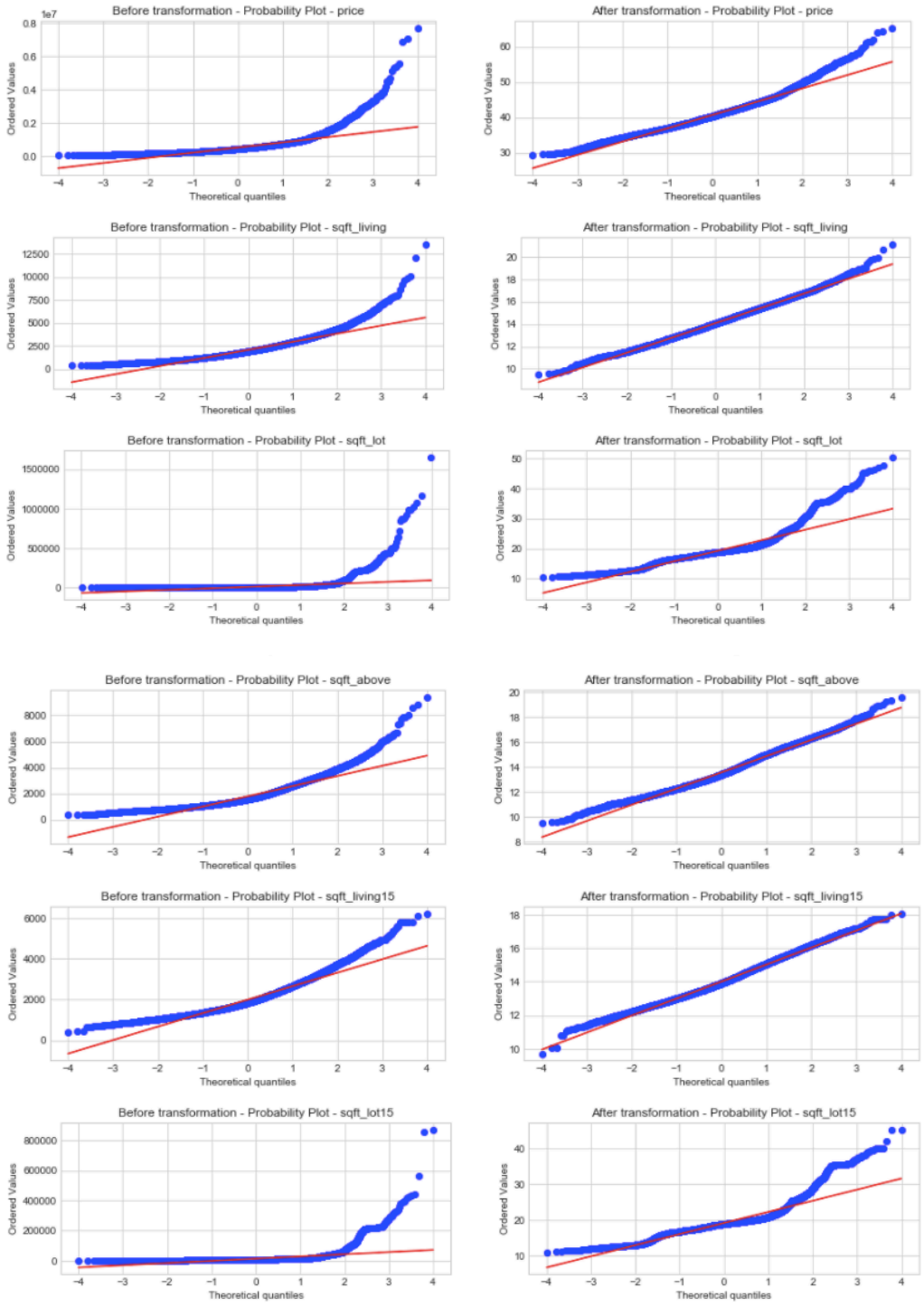
5 rows × 21 columns

Probability Plot

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

- Variables after transformation are closer to normal distribution even though some variables are still not normally distributed.

Figure14 – Probability plot



Modeling and Evaluation

Since the goal here is to predict housing price in King County, machine learning predictive models are top candidates. Using the following models for prediction.

Ridge Regression

Linear least squares with L2 regularization to reduce over-fitting issue.

- Ridge regression didn't perform very well, with 62.9% accuracy rate.

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

Random Forest Regression

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

- Random Forest Regression performs better than Ridge Regression, the accuracy increased from 62.9% to 66.4%.

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

Gradient Boosting Regression

In Gradient Boosting, the decision trees are generated in sequence. Each tree is generated using information from previously grown trees and the addition of a new tree improves upon the performance of the previous trees.

➢ **Grid Search with Cross Validation**

- Tuned parameter using Grid Search with cross validation
- Parameters we tuned:

  #criterion: friedman_mse, mse

  #max_depth: mean squared error with improvement score by Friedman

  #max_features: number of features to consider when looking for the best split

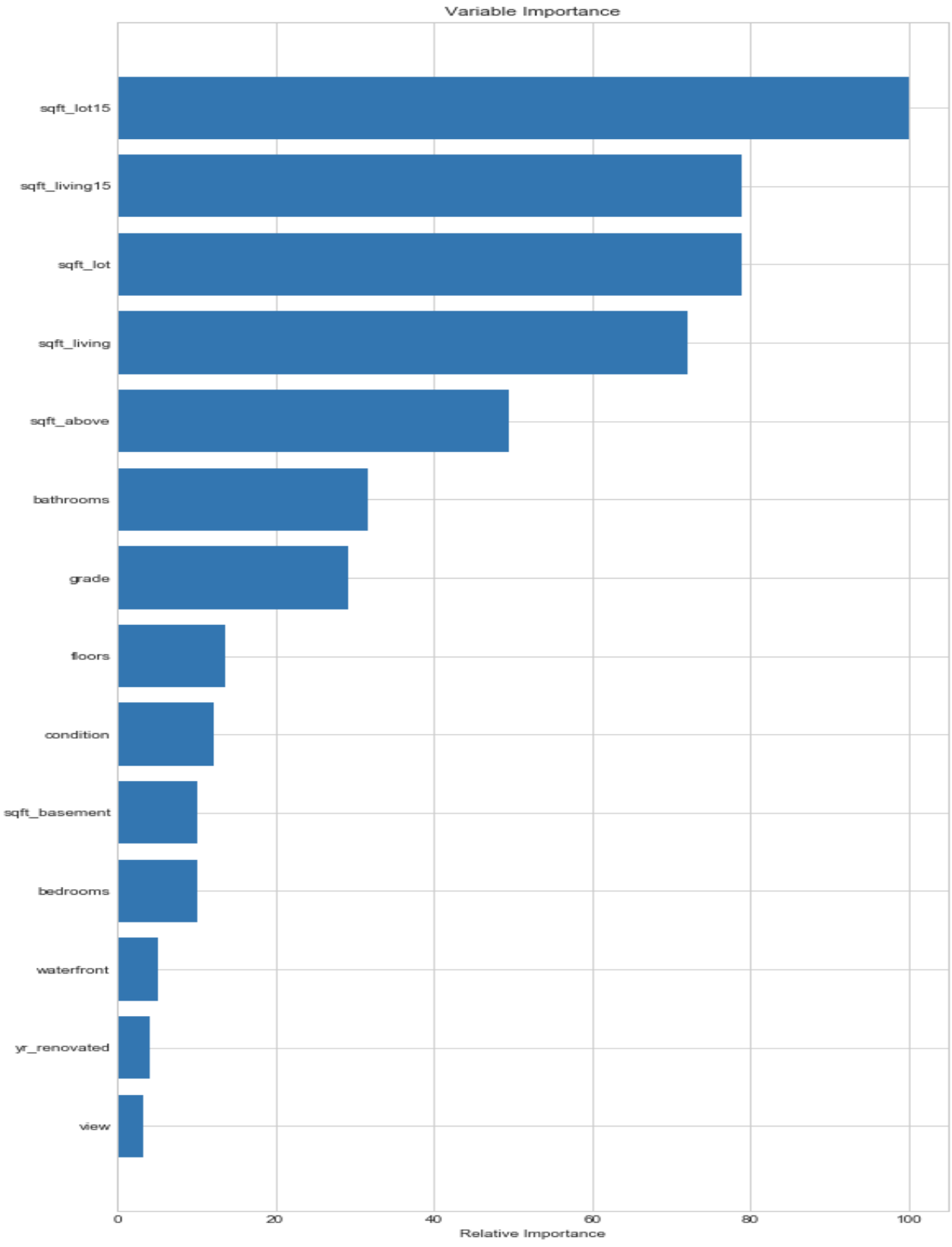
  #min_samples_split: min number of samples required to split an internal node
- Gradient Boosting Regression with grid search so far has better results than Ridge and Random Forest Regression. With 72% accuracy.

```
Best parameter     : {'criterion': 'friedman_mse', 'max_depth': 4, 'max_features': 4, 'min_sample
s_split': 2, 'random_state': 42}
Test data score    :  0.721614432648
Training data score:  0.758956656955
```

Feature Importance

Plot out the important features based on gradient boosting.

- Top important features are features about square foot lot, square foot living and square foot above the ground.



Evaluation & Suggestion

Based on the predictive models above, Gradient Boosting Regression with Grid Search generates the best prediction result, because in boosting, the new decision tree is based on the residuals and is then added to the current decision tree, and the residuals are updated.

By fitting small trees to the residuals, we gradually improve the overall model in areas where it does not perform well. Therefore, Gradient Boosting outperforms than Ridge and Random Forest Regression, but Gradient Boosting costs more time to run in order to find the best parameters.

The prediction result will be better if we have larger amount of data with more features because we can fit the predictive model with more data points. However, we will also need to avoid over-fitting problem when dealing with large amount of data by applying regularization with optimal parameters. With relatively small amount of the data we have high chances of getting biased result due to overly simplistic assumptions and thus lead to under-fitting the data and low accuracy.

There is trade off between bias and variance when using machine-learning models. Therefore, we need to try different predictive models with different regularization in order to balance the trade off between bias and variance in machine learning models.