ML -1 Assignment
Housing price Prediction

Introduction

The housing dataset has been provided and has already been split into train, validation and test datasets for analysis Here our objective is to To fit a model on a Training dataset using supervised learning method and to use the same model for the observations in the Validation dataset and finally to get the unbiased evaluation of the model on a Test dataset.

Objective

- To predict the price of houses based on certain features (predictors) using Mathematical and statistical techniques.
- To perform and exploratory data analysis and draw out some interesting observations
- To draw out interesting conclusions of variation of prices with variables of high correlation and significance as well as how variables are correlated to each other
- To create visuals of the supervised model for better understandability and interpretation

Analyses/Techniques Used

- Multiple Linear Regression
- Correlation
- Standardization
- Encoding

Description

All the 3 datasets provided contain 21 features and as per our objectives and further examination of data we know that the **price is dependent** on various features like bedrooms (which is most dependent feature), bathrooms, sqft_living(second most important feature), sqft_lot, floors etc. The price is also dependent on the location of the

house where it is present. The other features like waterfront, view are less dependent on the price

Tools used

Python, Excel, Tableau

Feature Explanation

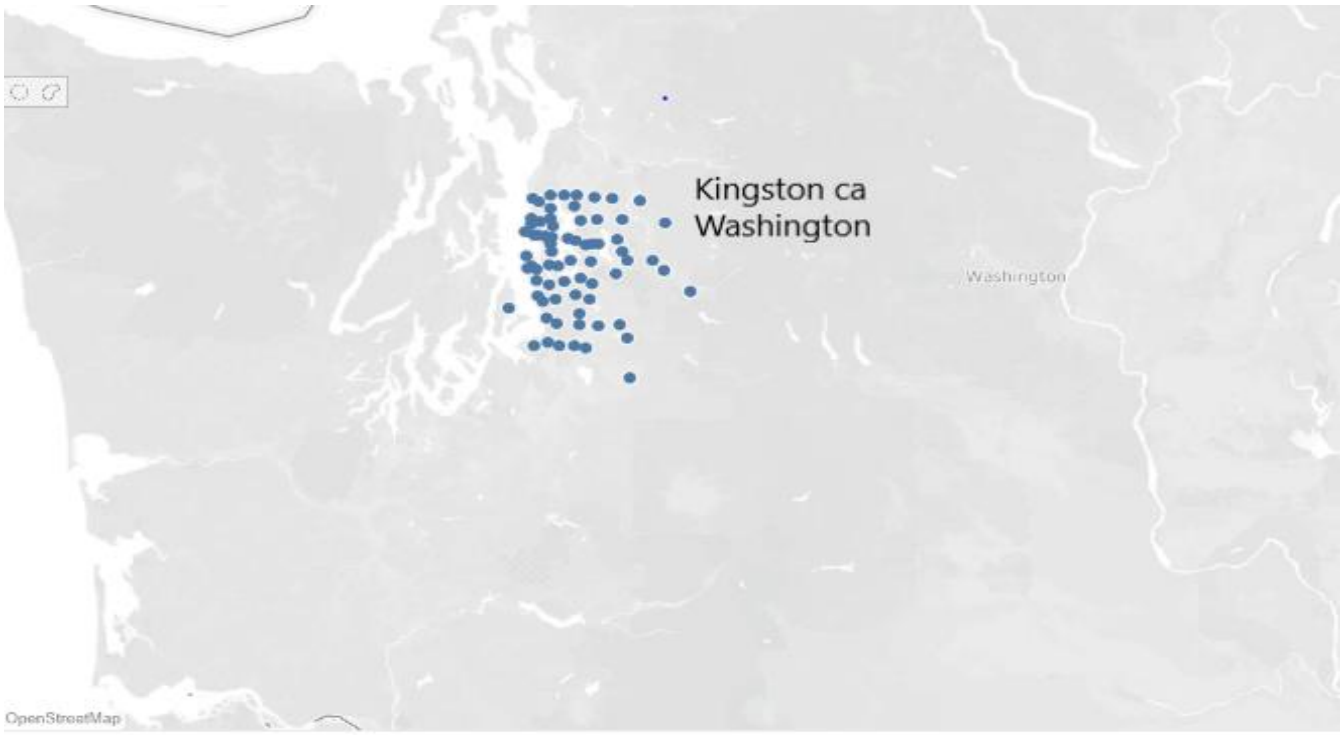
Explanation of each and every variable is given below:

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, where .5 accounts for a room with a toilet but no shower
sqft_living - Square footage of the apartments interior living space
sqft_lot - Square footage of the land space
floors - Number of floors
waterfront - A dummy variable for whether the apartment was overlooking the waterfront or not, 0(no) and 1(yes)
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, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.
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
zipcode - 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

Feature Exploration

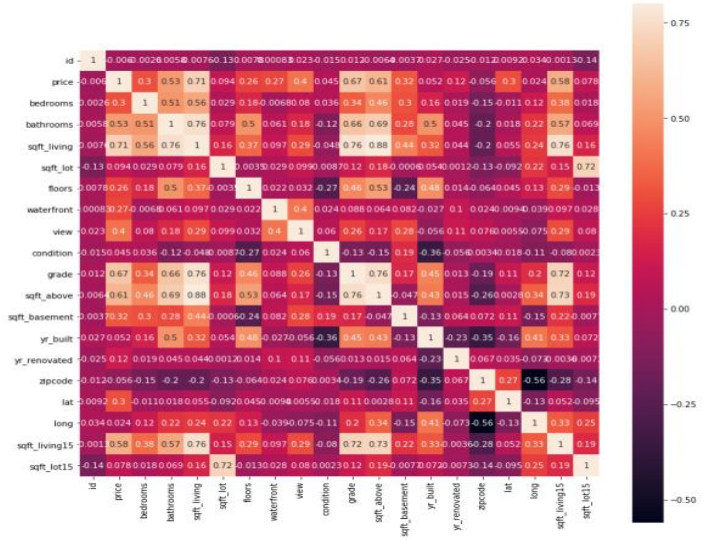
Important variables are analyzed using various scientific libraries in python like scipy, numpy, pandas etc and visuals are created using Tableau and other visual libraries in python (matplotlib, seaborn). The following are the outcomes of investigating the different features .

Plotting world map using tableau

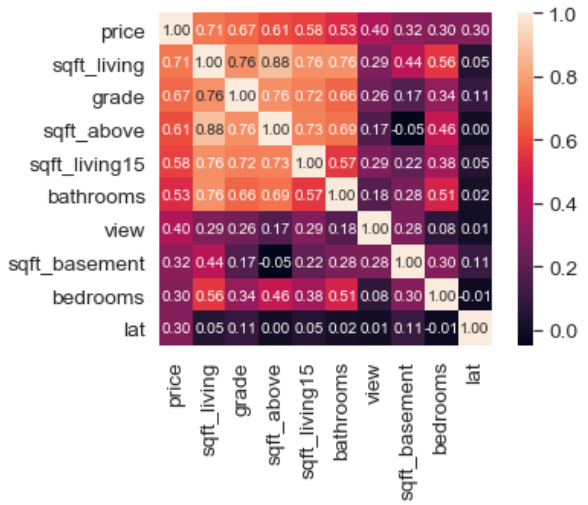


As seen by in the image, the zipcodes plotted using the lat and long coordinate variables, we can see that the area of interest to predict prices is Kingston county CA, state of Washington, USA

Correlation Heat maps



Taking a closer look with top 10 correlated variables



From the above the features: sqft_living, grade, sqft_aboce and sqft_15 features displayed the highest correlation wih the price of the house Also, there is a high correlation of sqft_living with other variables like. number of bathrooms and grade. So

the feature engineering will be a little complex here Also its important to see how variables affect each other

Frequency Analysis

Let's look the frequency of the following:

No of bedrooms, waterfront, condition and grade

This can be done by using the function COUNTA in MS-Excel

no of bedrooms	count
0	6
1	88
2	1255
3	4417
4	3094
5	755
6	116
7	17
8	7
9	3
10	1
11	1
33	1

Conditon	count
1	10
2	78
3	6291
4	2610
5	772

grade	count
1	1
4	17
5	107
6	914
7	4067
8	2707
9	1192
10	532
11	172
12	45
13	7

waterfront	count
0	9679
1	82

From the above tables we can draw the following conclusions

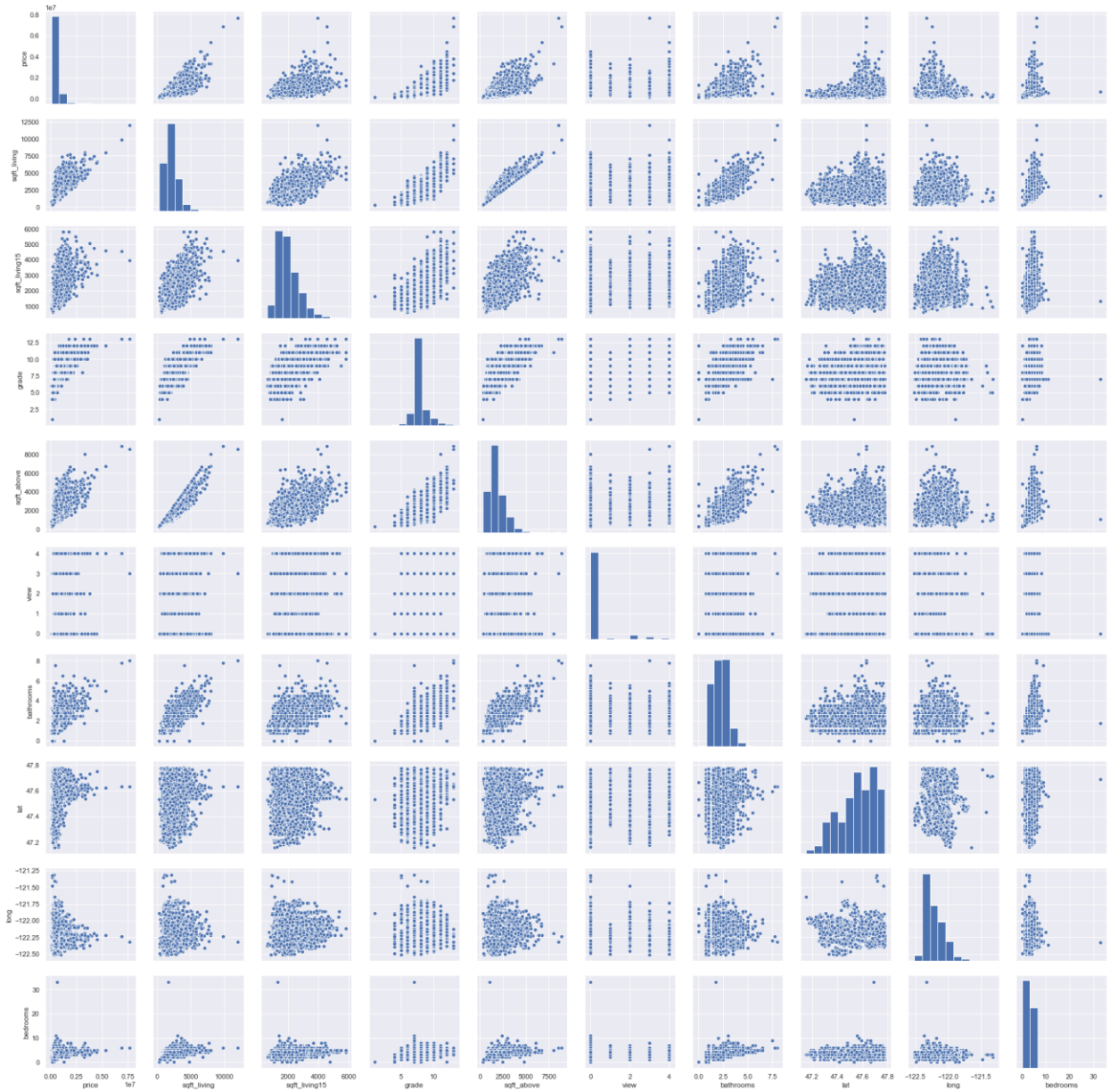
- Only 6 studio apartments have been listed
- There are only 3 houses with more than 9 bedrooms
- Most houses listed in Kingston have a 3/4 bedroom configuration (77%)

- As seen by the grade most houses have average quality of construction (69%)
- Almost all houses don’t overlook a waterfront (99%)

Correlation using Pairplot() and corr()

Using the seaborn package we can visualize the relationship between variables using the pairplot method and we can use the .corr() method to find out the pearson correlation

Below are a visuals of top 10 variables from the heatmap as well as other variables



```
]:
price            1.000000
sqft_living      0.702035
grade            0.667434
sqft_above       0.605567
sqft_living15    0.585379
bathrooms        0.525138
view             0.397293
sqft_basement    0.323816
bedrooms         0.308350
lat              0.307003
waterfront       0.266369
floors           0.256794
yr_renovated     0.126434
sqft_lot         0.089661
sqft_lot15       0.082447
yr_built         0.054012
condition        0.036362
long             0.021626
id              -0.016762
zipcode         -0.053203
Name: price, dtype: float64
```

We can see There is myriad of linear correlation between sqft_living, sqft_above, bathrooms and grade. Moreover, what we learned that the above mentioned features have the biggest impact on sale price. One would also expect location to play a role, but as they are in latitude/longitude coordinates, it requires advanced data manipulation to take it into account. Finally, due to many linear relationships we can apply regression models

Missing Value Analysis

As shown by the below table there are **NO** missing values in the data set. This makes our analysis a little easier

	Total	Percent
sqft_lot15	0	0.0
view	0	0.0
date	0	0.0
SalePrice	0	0.0
bedrooms	0	0.0
bathrooms	0	0.0
sqft_living	0	0.0
sqft_lot	0	0.0
floors	0	0.0
waterfront	0	0.0
condition	0	0.0
sqft_living15	0	0.0
grade	0	0.0
sqft_above	0	0.0
sqft_basement	0	0.0
yr_built	0	0.0
yr_renovated	0	0.0
zipcode	0	0.0
lat	0	0.0
long	0	0.0

Feature Engineering

From the correlation plot and heatmap, it is evident that the variables require some type of transformation to make sure that the regression model is fit in a much easier way. Also it's clearly observable that some variables are already encoded (Grade, Water-front) while some require Encoding. In this section we will see the type of feature engineering method used and see how that affects our model. The Feature Engineering was mostly done in Python using various sklearn and other packages.

Standardization

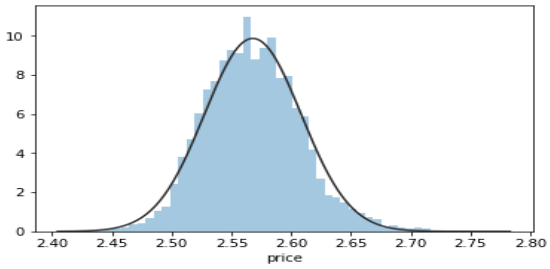
Using the .skew() method and .kurt() method we can check the skewness in python

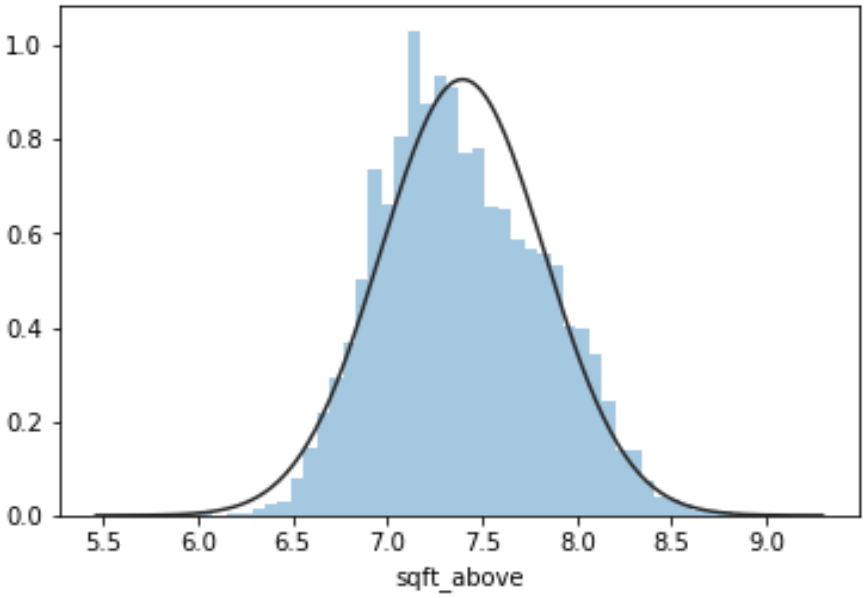
```
Skewness: 4.024069
Kurtosis: 34.585540
```

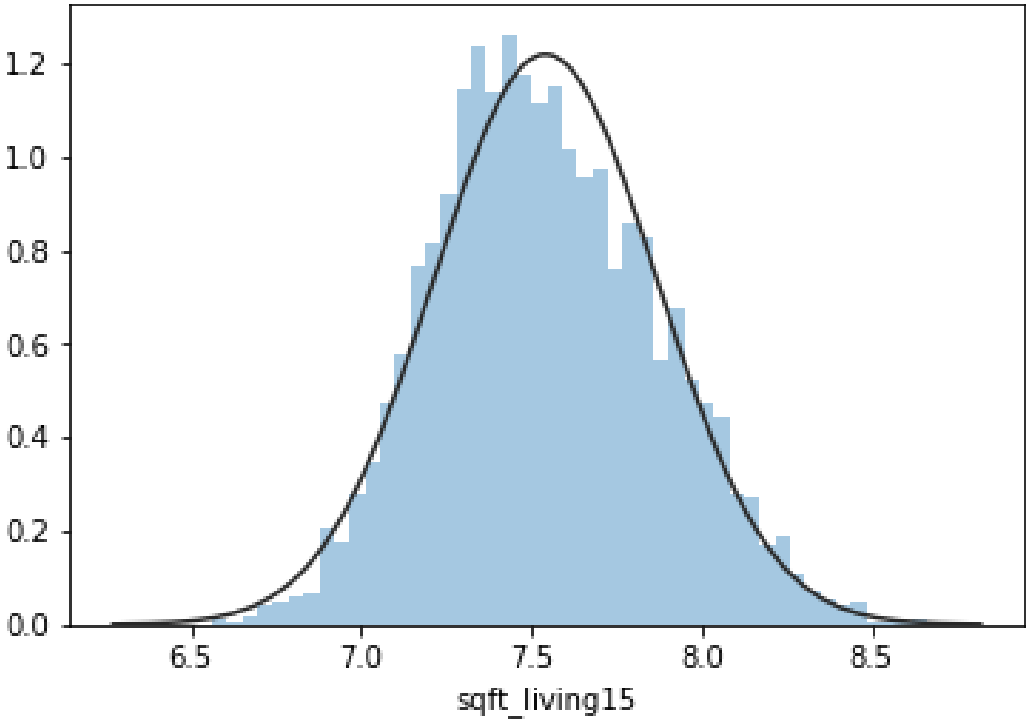
This is quite significant. The distribution is heavily skewed with the tail to the right. The data standardization method from **standard scalar** module of sklearn package can be used to mitigate skewness and kurtosis

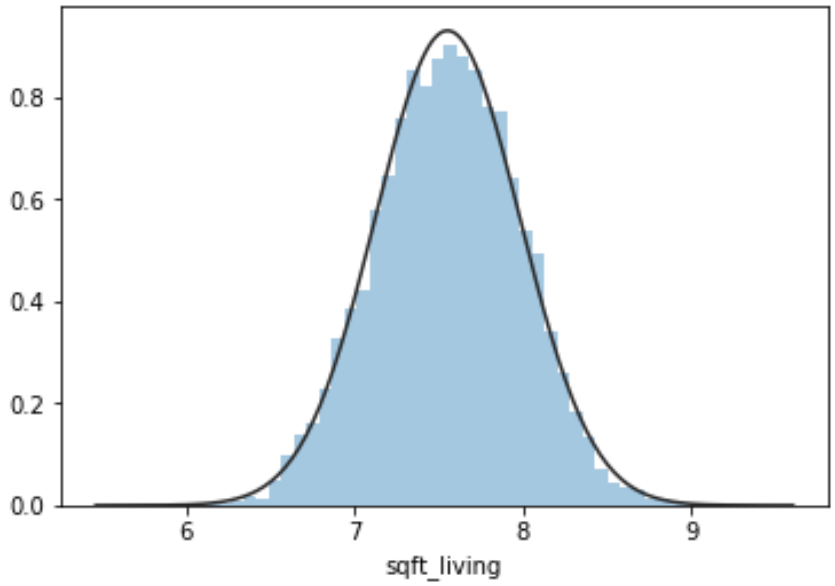
Normalization

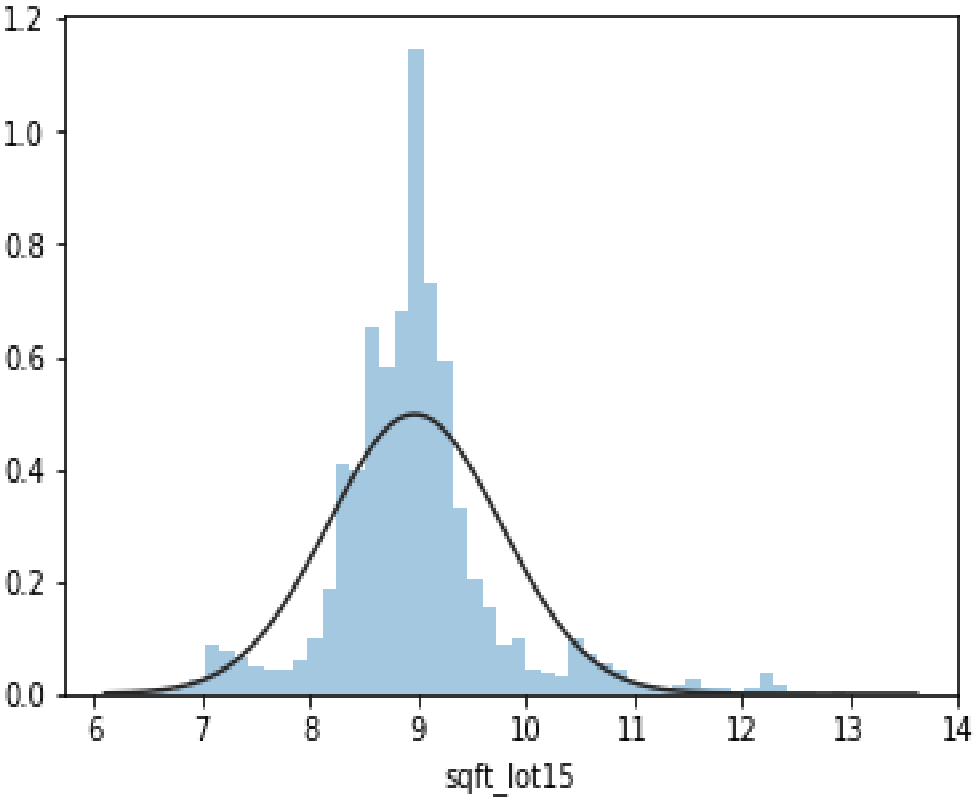
From the correlation plot it is evident that ‘sqft_living, sqft_lot, sqft_above, , sqft_living15, sqft_lot15’ are continuous variables that have a linear relationship with a significant amount of outliers. To treat these outliers we will use log transformation so that data follows a log-normal distribution or an approximate log normal distribution, we will do this using the np.log() method in python. After normalization the following are the graphs. We will also normalize the target variable which is ‘price’ since we have normalized these features and price also has outliers




1.0
0.8
0.6
0.4
0.2
0.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
sqft_above


1.2
1.0
0.8
0.6
0.4
0.2
0.0
6.5
7.0
7.5
8.0
8.5
sqft_living15


0.8
0.6
0.4
0.2
0.0
6
7
8
9
sqft_living


1.2
1.0
0.8
0.6
0.4
0.2
0.0
6
7
8
9
10
11
12
13
14
sqft_lot15

Zipcode encoding

Shoud we use zipcode in our analysis?

As shown the Pearson correlation value Let us see thuse zipcode in our analysis, we can do this by encoding zipcode using get_dummy() method in pandas we will also check if encoded columns have unique values or not. The understanding is that it if one column cell has same value in the corresponding column cell than that column cannot be considered in our analysis to fit the model merely because of multicollinearity (two houses can't exist in a single zipcode). Fortunately they aren't any such

Feature selection

Finally we can go ahead and select the following features based on the correlation plot and pearson correlation values we will fit a linear regression model and see the results and see how we can improve the model

Target: log of price

Predictors: bedrooms, bathrooms, Log of sqft_living, Log of sqft_lot, floors, waterfront, view, condition, grade, Log of sqft_above ,sqft_basement, Log of sqft_living15, Log of 'sqft_lot15

Fitting the model

Using Linear Regression we have fit the model on the training data set and the following are the results of the model

Final Regression Model Results	
Accuracy of training	89%
Accuracy of validation	87%
Squared Mean Error	0.19
R sqaured Training	0.886
R sqaured Testing	0.877

Conclusion

As seen by the above table the Linear Regression model has produces good accuracy and acceptable R squared value for the testing dataset. Hence this model can be used to predict prices of houses with good accuracy