# Capstone Project 2: Milestone Report 1

House Real estate Price Prediction

## Problem statement:

Using the data provided I would like to predict the price of a house. My clients are people who are thinking of putting their house up on the market and are confused about the starting price. I am using data from kaggle which is already provided for this problem. I am in the market for a house very soon and I would love to know what factors influence the price of a house more and use that knowledge when I do purchase a new house.

I am thinking of refining by data first, feature engineering, outlier detection etc.. and then use linear regression and sklearn models to predict the price of a house.I am also going to do more EDA on the data to glean insights for a person who is deciding to put his house up on the market. I am thinking of comparing different features and their effects on the price.

I am planning to deliver data cleaning, feature engineering and modelling solutions in a python notebook and a document which describes my steps.

## Data Description:

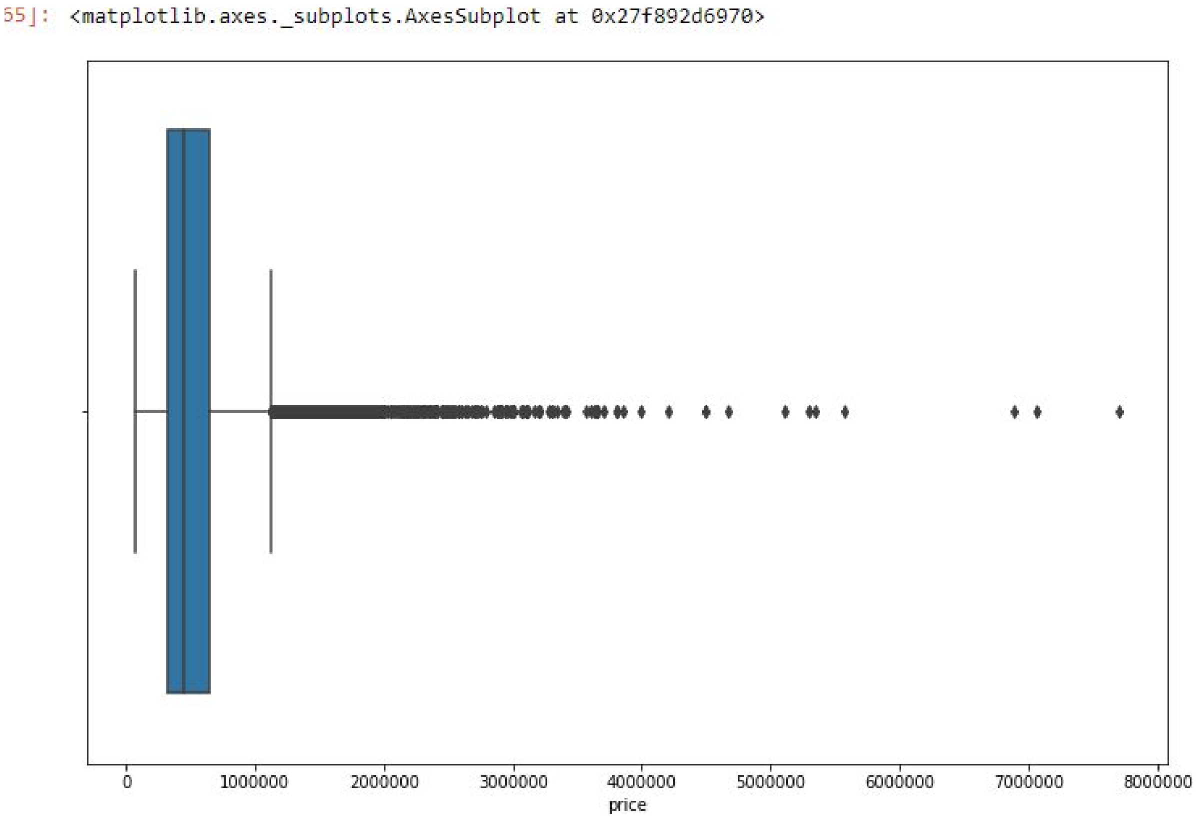
This data is of the prices of the houses in Kings County , Washington from may 2014 to May 2015. Below is the description of each attribute.

| **Variable** | **Description** |
|---|---|
| Id | 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 |
| Waterfront | Whether or no the house has a waterfront of 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 |
| Zipcode | What zipcode area the house is in |
| Lat | Lattitude |
| 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 |

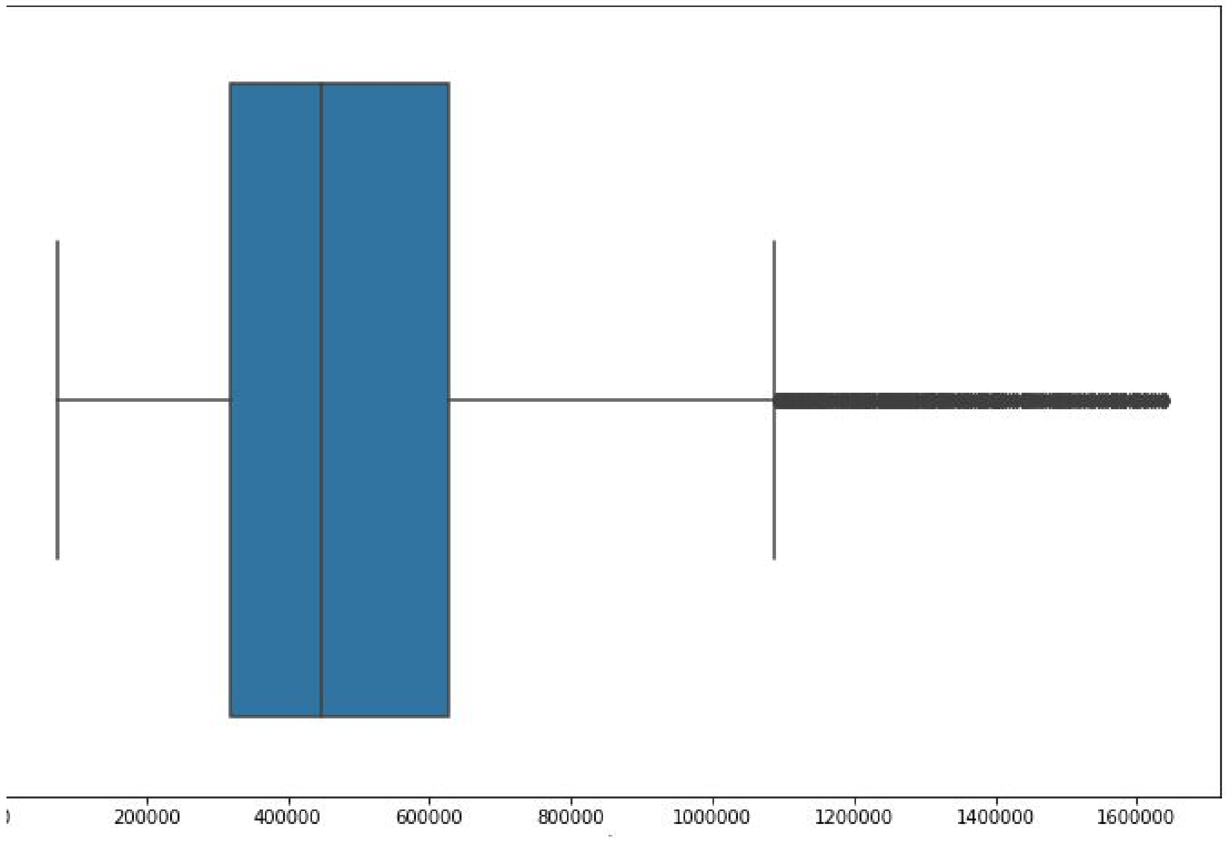
## Data Cleaning:

For data cleaning I dropped the ID and date columns from the dataset as its not a feature we can predict on. I detected outliers in the dataset on the dependent variable price.



```
55]: <matplotlib.axes._subplots.AxesSubplot at 0x27f892d6970>
```

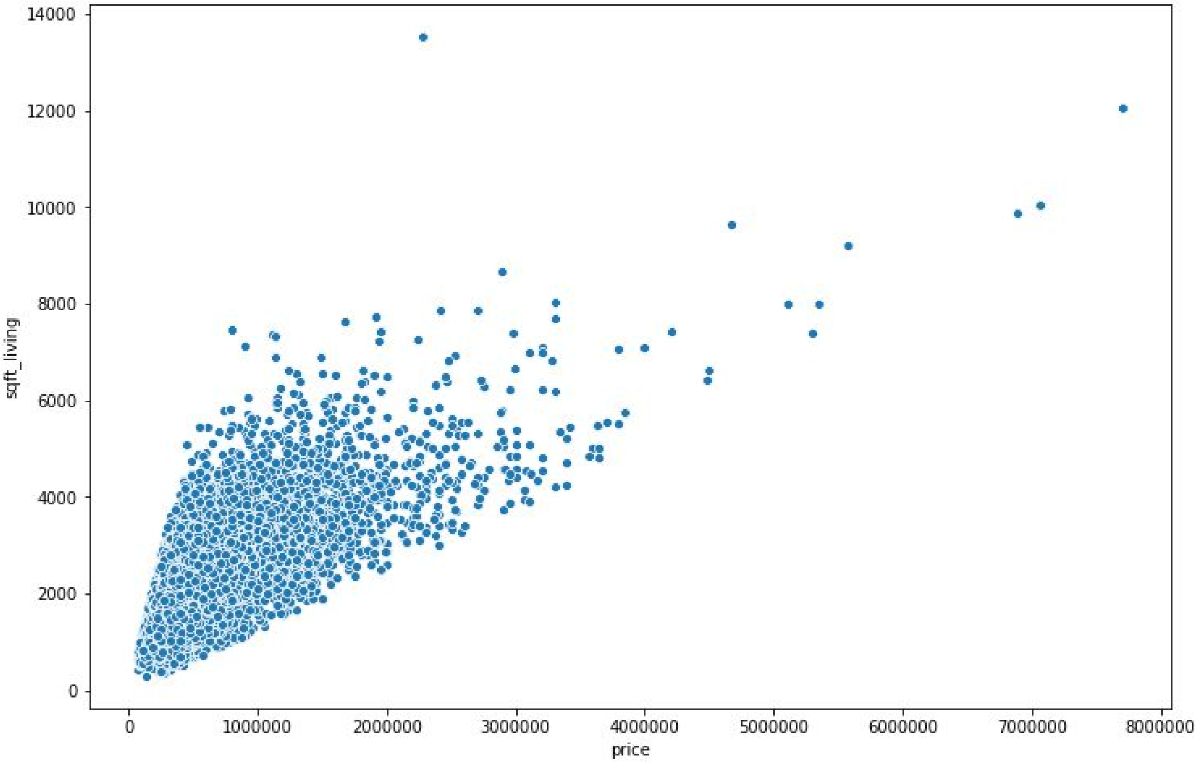
I was able to get the zscores of all the prices and then remove all values greater than 3 standard deviation. After cleaning the box plot looked like this



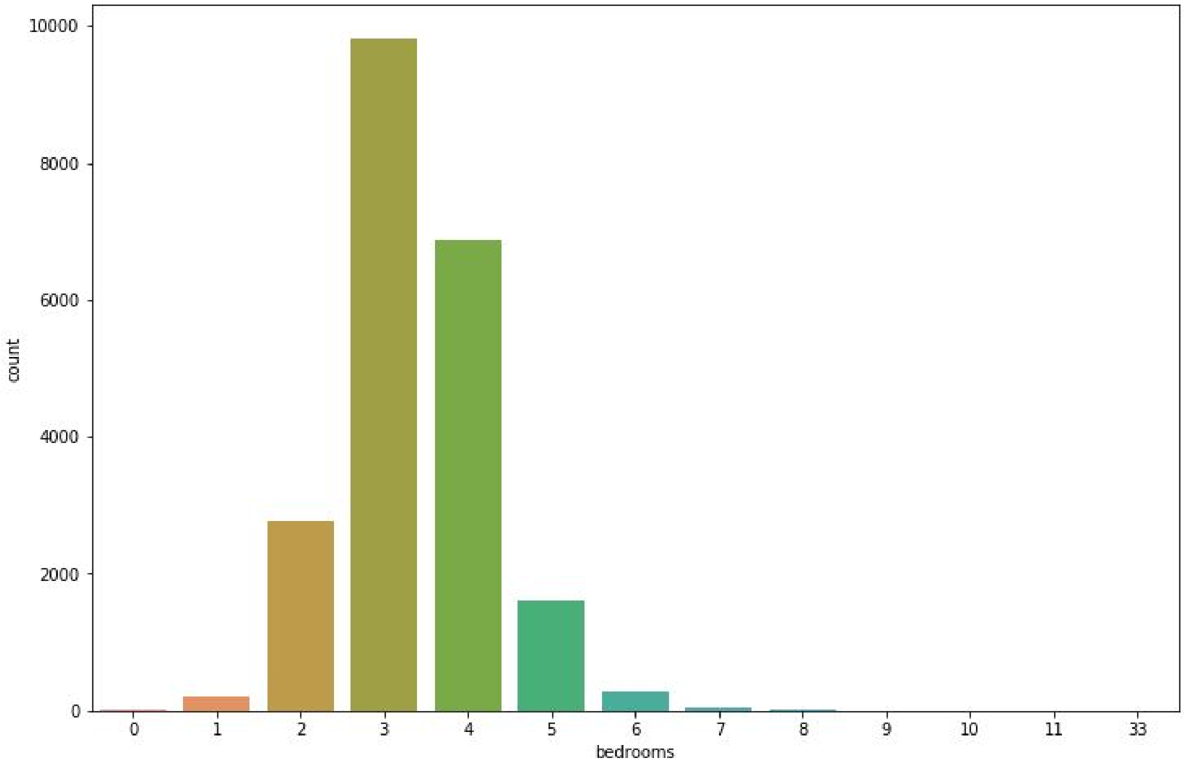
## Exploratory Data Analysis:

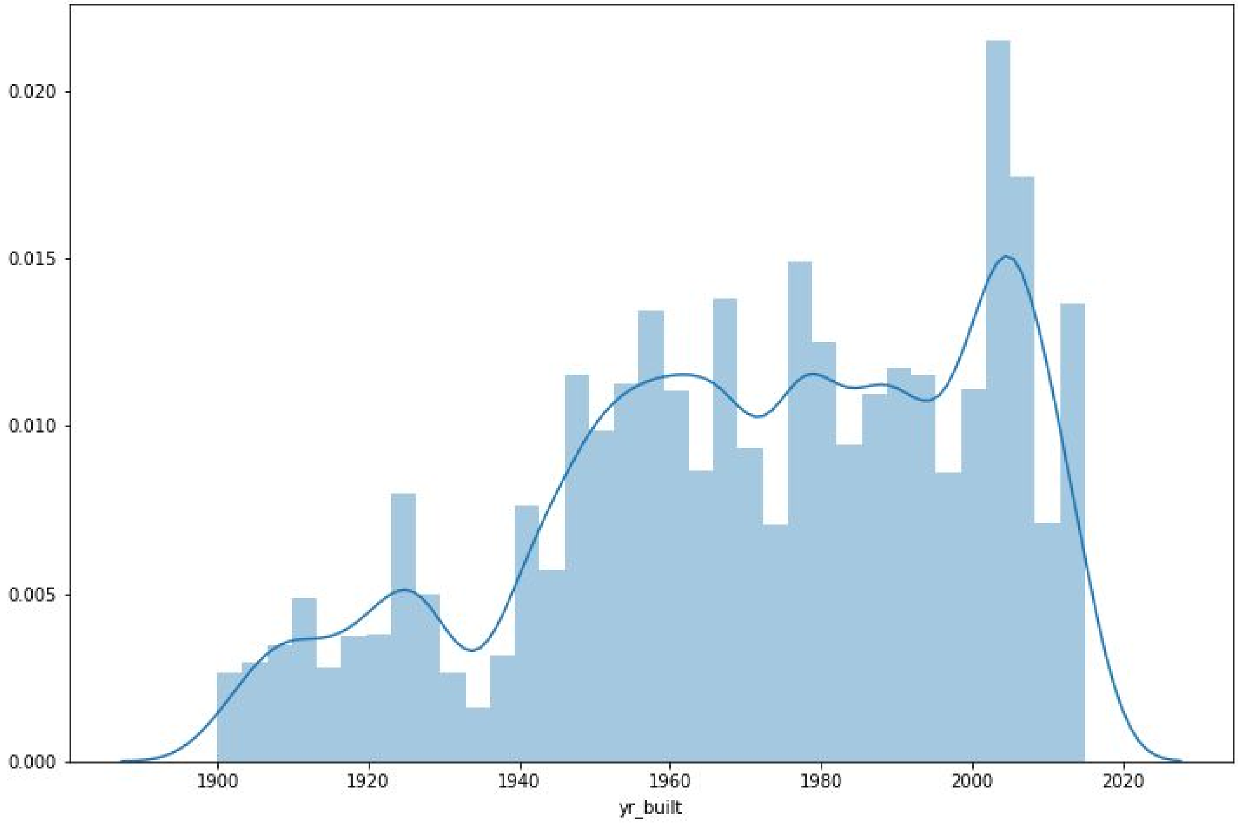
**We find that there is a strong correlation between price and sqft_living.**

```
<matplotlib.axes._subplots.AxesSubplot at 0x27f89f8d880>
```

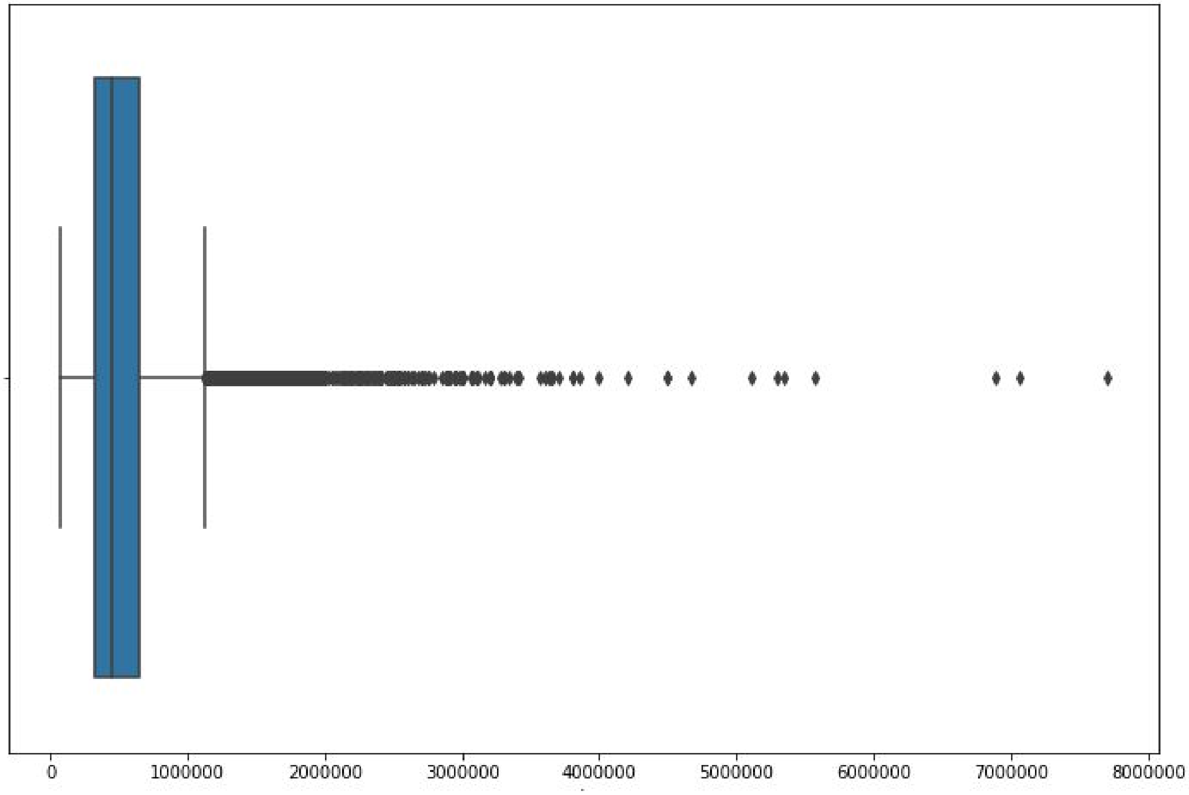


**We found that we have a lot of 3-4 bedroom houses in our dataset and houses built in 2000's**

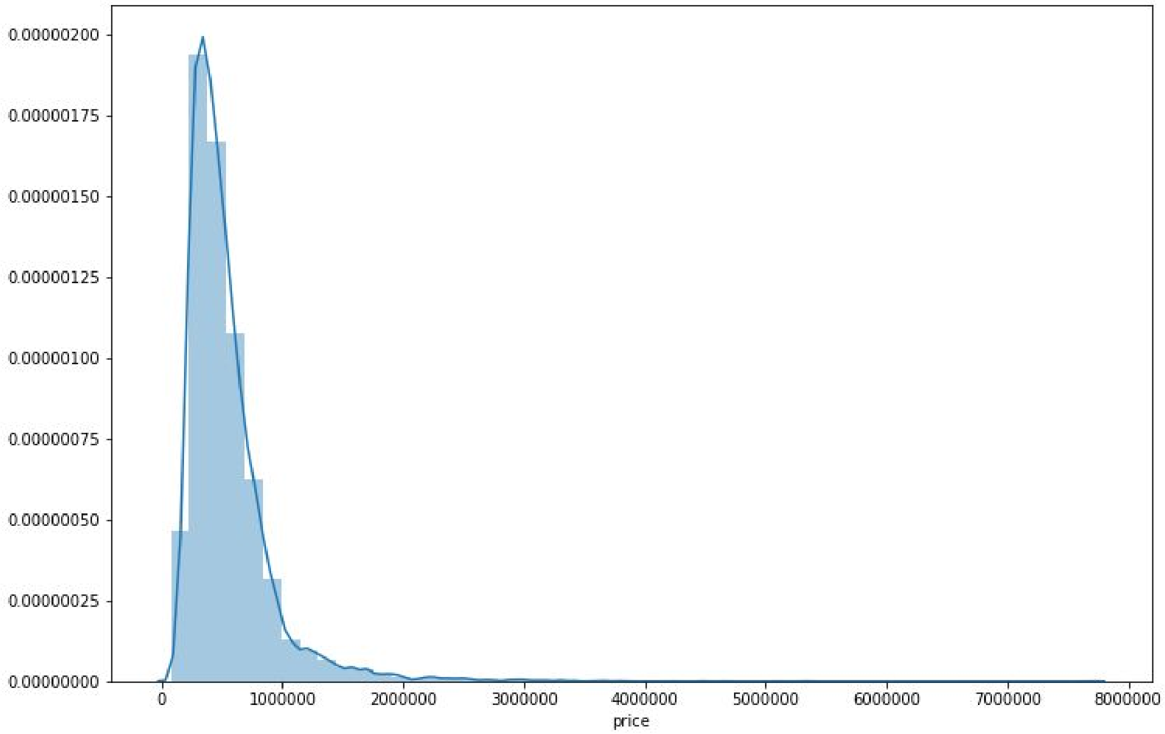

10000
8000
6000
4000
2000
0
count
0
1
2
3
4
5
6
7
8
9
10
11
33
bedrooms


0.020
0.015
0.010
0.005
0.000
1900
1920
1940
1960
1980
2000
2020
yr_built

**We found many outliers in the dataset. As highlighted by the boxplot and the distribution plot.**

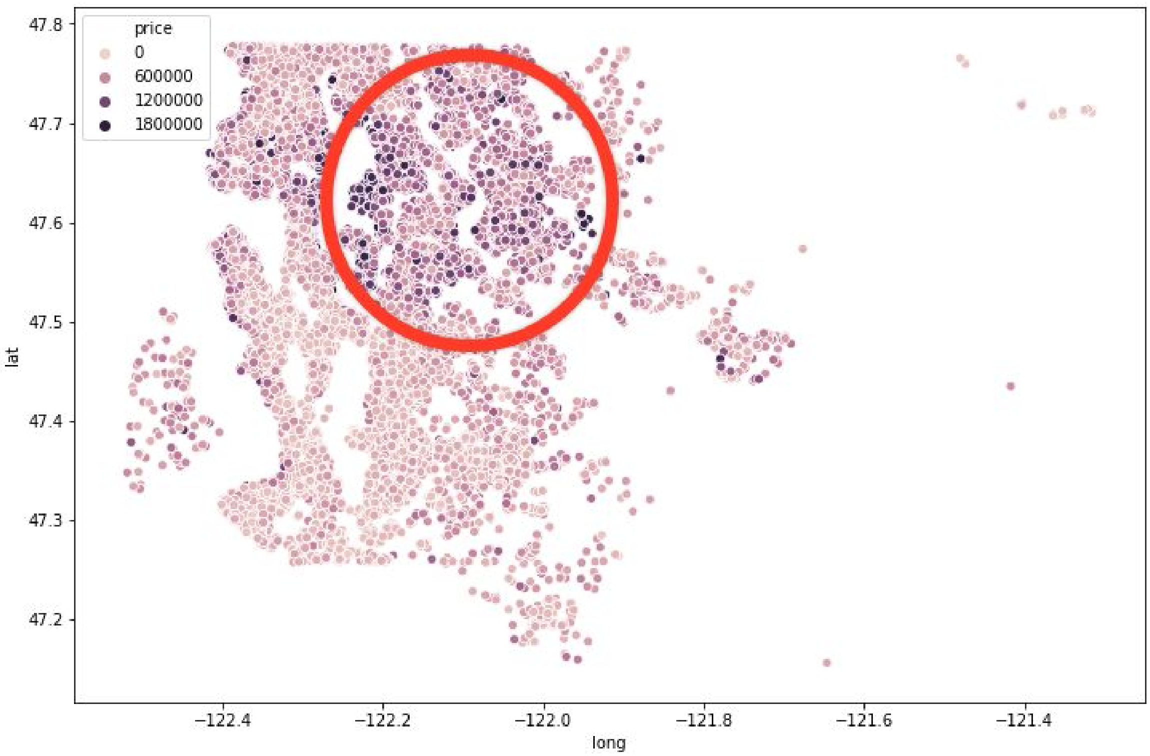


```
<matplotlib.axes._subplots.AxesSubplot at 0x27f89e87ee0>
```

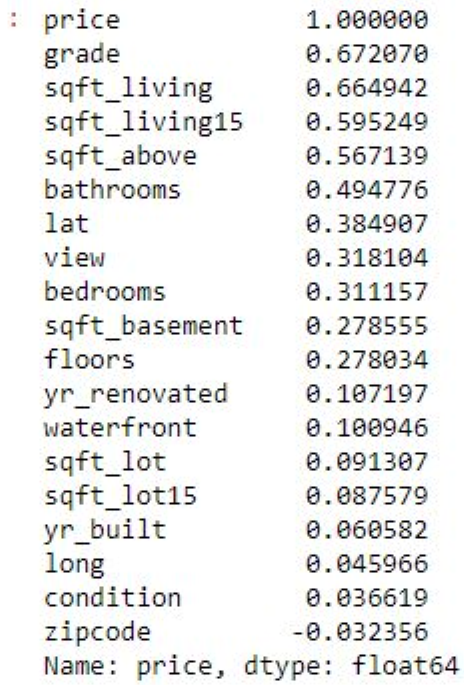


**We found the highest priced houses were in one area of our map. Highlighted below.**

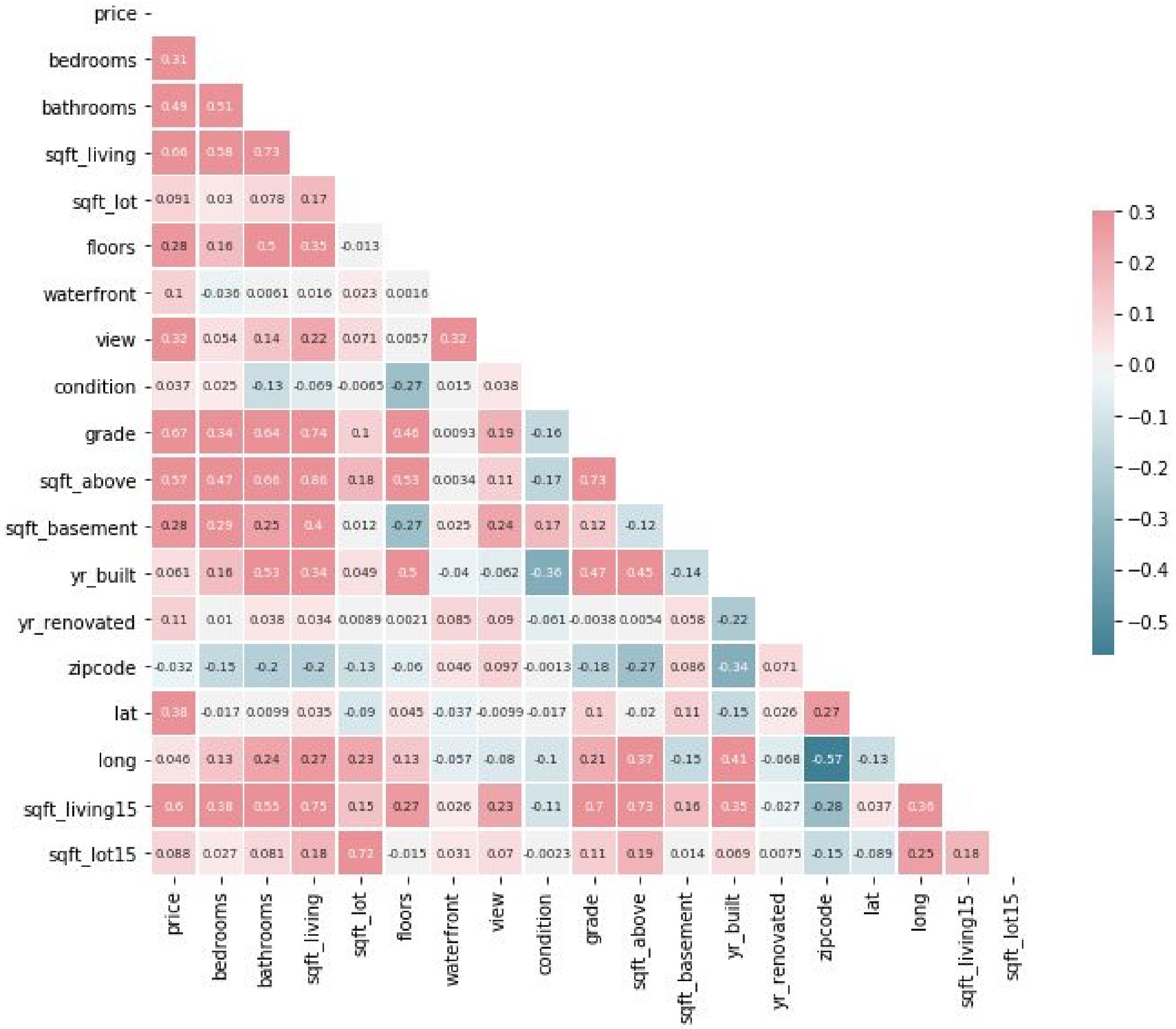
```
<matplotlib.axes._subplots.AxesSubplot at 0x27f8917c6a0>
```



**We found grade , sqft_living, sqft_living15 , sqft_above, bathrooms had the highest correlation with price.**



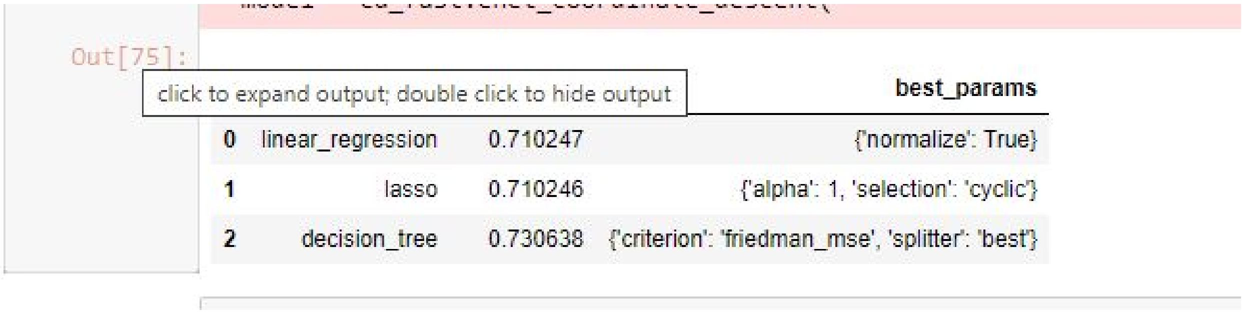
```
: price            1.000000
  grade            0.672070
  sqft_living      0.664942
  sqft_living15    0.595249
  sqft_above       0.567139
  bathrooms        0.494776
  lat              0.384907
  view             0.318104
  bedrooms         0.311157
  sqft_basement    0.278555
  floors           0.278034
  yr_renovated     0.107197
  waterfront       0.100946
  sqft_lot         0.091307
  sqft_lot15       0.087579
  yr_built         0.060582
  long             0.045966
  condition        0.036619
  zipcode         -0.032356
  Name: price, dtype: float64
```



# Modeling

We used K Fold cross validation to measure accuracy of our model. Initially I just got the linear regression score of 70% after which I made the decision to try out other models. Our training test split was 80-20. We split our datasets into 5 for cross validation.

We used GridSearchCV to pass different parameters for different models and compare them.



| | | | best_params |
|---|---|---|---|
| 0 | linear_regression | 0.710247 | {'normalize': True} |
| 1 | lasso | 0.710246 | {'alpha': 1, 'selection': 'cyclic'} |
| 2 | decision_tree | 0.730638 | {'criterion': 'friedman_mse', 'splitter': 'best'} |

Here the first column gives us the Model we used. The second one gives us our score and the third is the best parameters the GridSearchCV has choose to get the score.

Hence based on the above results we can say that Decision Tree gives us  the best score. Hence we will use that for prediction of price for real estate in Kings County, Washington.