**Big Data Analytics Project Report**

**Team members – Tejashri Joshi, Surbhi Goyal**

### King County Housing Prices

### **Introduction**

In the given problem we have been given the data about houses sold in the king county area in the period of May 2014 to May 2015. The data set includes various aspects related to homes selling such as price, date, number of bedrooms etc. King is the most populated county in Washington, and the 13th most populated in the United States. Our objective is to predict the house prices based on the given data in the King county, WA, U.S.A.

This is a great data set in terms of understanding the market situation for real estate. As a customer any person who wants to buy a house always wants to know the market situation and prices of houses based on different factors. This data will be useful for people who are interested in buying a house where they can focus on features such as room count, area , view etc. At the same time this can be useful for the real estate agents so that they can guide their customers correctly.

Housing prices are based on a number of different factors like economy, demographics, geographical area, interest rates, government policies, job market. Any major change in one of these factors can have a large impact on the housing prices. To predict the prices accurately, we should consider all these factors which is pretty challenging. However as per the dataset, we can consider demographic factors. And have to assume that the overall economy and demographics won’t change. And also the house demand remains constant.

**Data Description**

Data population - Data population in the given case is features related to house sale and price in the king county region. The feature values can vary for different years and for another county.

Data structure - The provided data is rectangular and given in the form of CSV.

Scope and temporality - The Given data has values for a period of 1 year, from May 2014 to may 2015 in the area of king county.

Granularity - The data collection has 21613 total records, consisting of 21 columns where each record represents a house's details such as bedrooms, bathrooms, sqft living, streets, coastline, etc., sold price and date. The price of a house mostly depends on features like Sqft. area of house, bedrooms, location and view. Other features may have less impact on price.

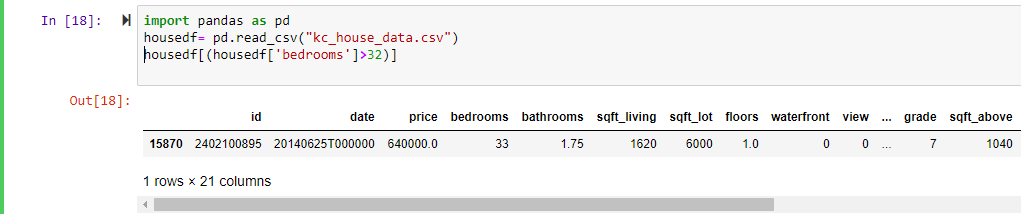
There are several columns present such as price, date , year built, year renovated, Sqft area which are categorized as quantitative data. Some features describe building condition and grade which can be categorized as ordinal data.

Faithfulness - Data originally prepared by the Center for Spatial Data Science but it has some wrong data.Some records have date\_sold values less than the year that the same house was built. Similarly, the date when the house was built is greater than the renovated date. One record has a home with 33 rooms in a 1620sq.ft. living area which seems unrealistic. Some rows do have 0 bathrooms and 0 bedrooms which doesn’t make sense. We chose to exclude these records to make the data consistent.

Examples:

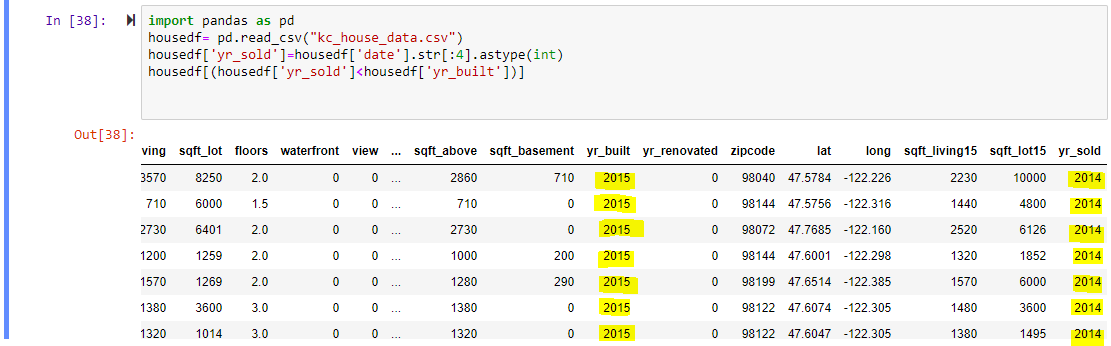
Houses with lot of rooms and having less Sqft area:

There is a record where the number of bedrooms is 33 as shown below.



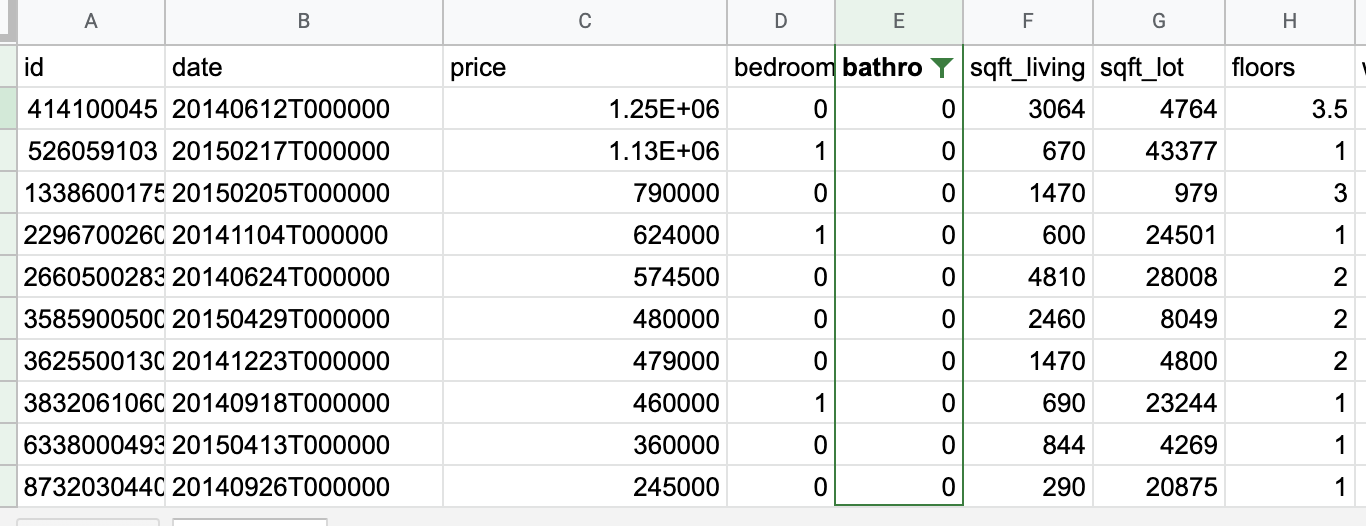
Houses where year sold is less than year built :

As shown below, there are some records where the built year of the house is greater than the year it was sold. The highlighted data given below show such records.



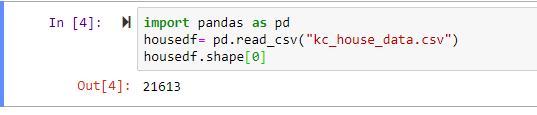
Similarly, some records have 0 number of bathrooms as well as 0 number of bedrooms as shown below.

Houses with 0 bathrooms:

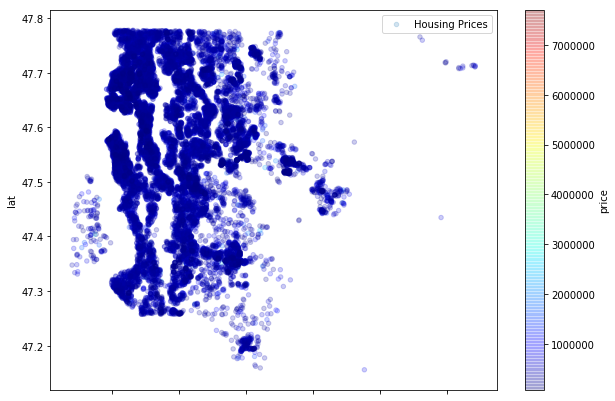


Total Rows in Data :

The total number of raw data in the given data set is 21613 and after removing all the incorrect records, we are left with 21585 records.



Home price on the basis of longitude and latitude :



Longitude

**Related Work**

In the paper [1], the author predicted the price of houses in King County. He has used a number of regression models to compare their performances based on almost all the features. For all predictive analysis, k-fold cross-validation is used to reduce the probability and effect of over-fitting. The data is randomly divided into ten folds and each fold is used once as a test set and nine times as part of the training set. And, RMSE metric is used to measure the difference between predicted and true values. Using 14 features (i.e. only excluding view and yr\_renovated) with knn gives a RMSE of $158,452 and a normalised RMSE of 29.3% which suggests that high accuracy can be achieved by using a larger set of features and more complex algorithm such as artificial neural network.

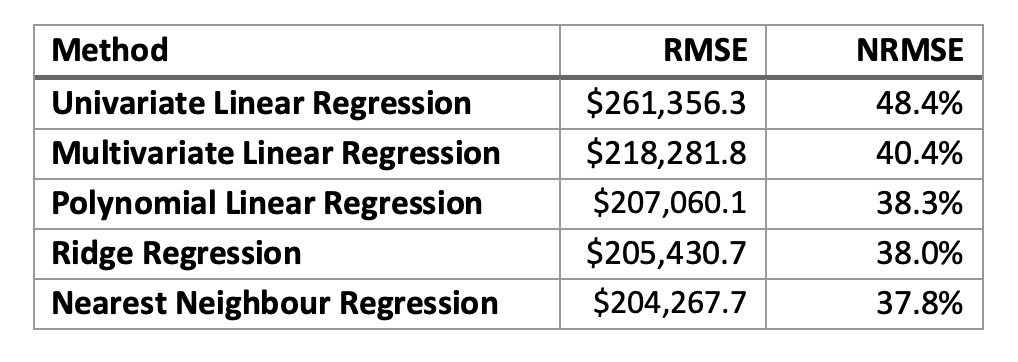
The result set of each method is given below.

Fig 1. Result Set [1]

Advantage & DIsadvantage:

The author has tested multiple Regression models and considered all the features instead of 1 or 2 for prediction. Considering features which are highly correlated will reduce the error further. Author did not try any model other than regression like SVM and ensemble methods like Random forest.

**Data Transformation and EDA**

The table given below shows the complete description of variables in the dataset.

Table#1

|  |  |  |
| --- | --- | --- |
| **Variables** | **Description** | **DataType** |
| ID | Unique ID for each home sold |  |
| date | Date of the home sale | Numerical |
| price | Price of each home sold | Numerical |
| bedrooms | Number of bedrooms | Numerical |
| bathrooms | Number of bathrooms, where .5 accounts for a room with a toilet but no shower | Numerical |
| sqft\_living | Square footage of the apartments interior living space | Numerical |
| sqft\_lot | Square footage of the land space | Numerical |
| floors | Number of floors | Numerical |
| waterfront | A dummy variable for whether the apartment was overlooking the waterfront or not | Nominal |
| view | An index from 0 to 4 of how good the view of the property was | Ordinal |
| condition | An index from 1 to 5 on the condition of the apartment, | Ordinal |
| grade | An index from 1-13, 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. | Ordinal |
| sqft\_above | The square footage of the interior housing space that is above ground level | Numerical |
| sqft\_basement | The square footage of the interior housing space that is below ground level | Numerical |
| yr\_built | The year the house was initially built | Numerical |
| yr\_renovated | The year of the house’s last renovation | Numerical |
| zipcode | What zip code area the house is in | Numerical |
| lat | Latitude | Numerical |
| long | Longitude | Numerical |
| sqft\_living15 | The square footage of interior housing living space for the nearest 15 neighbors | Numerical |
| sqft\_lot15 | The square footage of the land lots of the nearest 15 neighbors | Numerical |

*Missing Data*

Fortunately, the data set that we are using doesn’t have any missing information except the last 2 lines are empty. It shows 21613 records but there are only 21611 records.

*Outlier Detection*

Using boxplot, we have found outliers in the data set and analyzed it. The outliers detected in the price variable correspond to other variables like number of bedrooms, number of bathrooms, grade and sqft\_living. So we have decided to keep these outliers in the dataset.

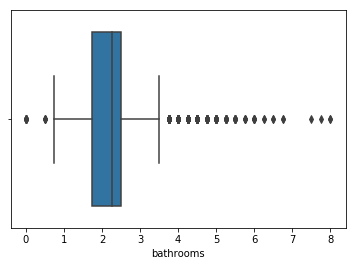
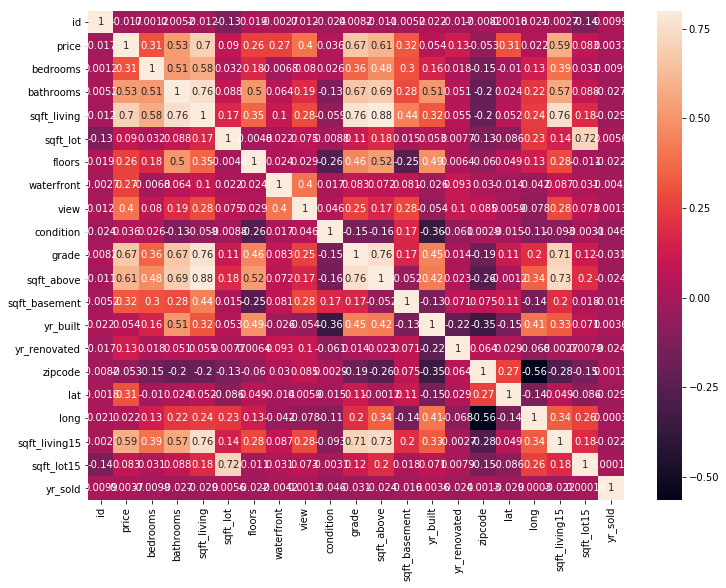


Fig. Outliers in Bathrooms variable

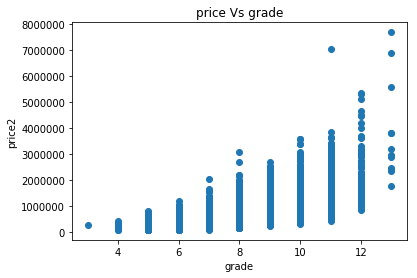
*Correlation Matrix*

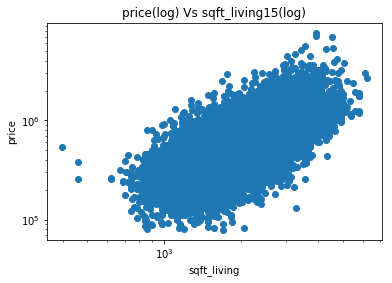
This matrix shows that sqft\_living and grade has the highest correlation between them and price. Also, sqft\_above and sqft\_living15 have significant correlation with price of the sold house.

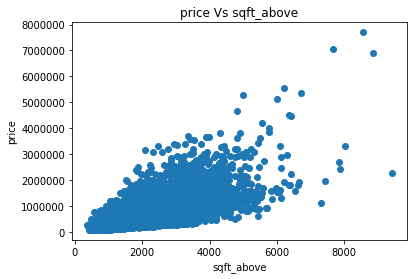


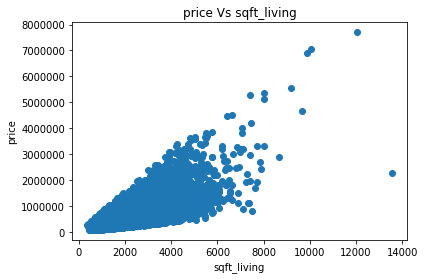
*Features*

As mentioned above, the price of a house depends on a number of factors. So, we have checked each and every variable relationship with price and found that sqft\_living, sqft\_living15, sqft\_above and grade has a linear relationship with price. Below given figures show the clear relation between the price and other features.









**Data Readiness -** As mentioned above, we have removed the rows with inconsistent data such as rows with 0 bedrooms and bathrooms , rows having yr\_sold greater than yr\_built. Also as features we have kept columns sqft\_living , sqft\_living15,sqft\_above , grade. With these preparations our data is ready to use for training machine learning models. In order to increase the consistency of our model, we have added a new column, called the ***Yr\_sold***, which is derived from the date of sale of the home.

After removing all the incorrect records, we have a total 21585 rows.

**Models** :

As the label in our problem is numerical value, this is a regression problem. For regression, there are many machine learning algorithms available. We are proposing the below prediction methods.

1. **Linear Regression** - Linear regression is the simple machine learning algorithm which is most commonly used in solving real estate problems. It attempts to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. In regression, we calculate the coefficients for each feature we used in data. These coefficient values are calculated depending on the line which is best fit through data points. As part of evaluation, we will check the root mean squared error (RMSE) which shows the difference in actual and predicted values. Low value of RMSE means our model is performing well.

As linear regression is the fundamental simple method for supervised learning and is most widely used for price detection, we have decided to use this method for this problem.

1. **Gradient Boosting** - Gradient boosting is one of the boosting ensemble methods. This method can be used for regression. This method is sequential where error at one stage is passed to the next stage as input. In this model we generate multiple decision trees sequentially using error from the last stage. While training the model, we set the values like n\_estimators , max\_depth and learning\_rate. n\_estimators are the number of sequential trees that can be used in the model. Max\_depth is the depth of trees. Learning rate is the rate that defines the impact of each tree on the final step. Lower rates are generally preferred. We have to use these parameters such that the model should not overfit.

As this is an ensemble method it tries to correct the error from the previous stage in the next stage, with proper use of hyper parameters we can get a lower mean square error. So we chose this as one of the methods. to implement the gradient boosting method we have used below hyper parameters.

Table# 2

|  |  |
| --- | --- |
| n\_estimators | 300 |
| max\_depth | 4 |
| learning\_rate | 0.01 |
| min\_samples\_split | 2 |

1. **Random Forest** - Random forests is an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression](https://en.wikipedia.org/wiki/Regression_analysis) that work by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees.

It uses bagging and features randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. Random forests allow us to tune parameters, either to increase the predictive power of the model or to make it easier to train the model.

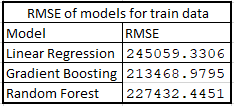
Table#3

|  |  |
| --- | --- |
| n\_estimators | 300 |
| max\_depth | 4 |
| max\_features | 4 |
| min\_samples\_split | 2 |

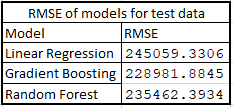
**Experiments :**

Evaluation metric - As the given problem is of regression we have used the evaluation metric as Root Mean square Error. This is the difference between values predicted by a model and the values observed. It tells how concentrated the data is around the line of best fit. Lower the value of RMSE means the model is well fitted.

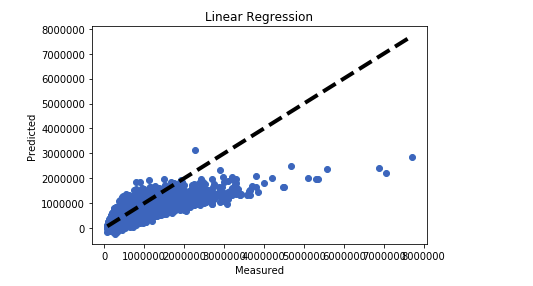
Below table shows the RMSE values for each model against train data.

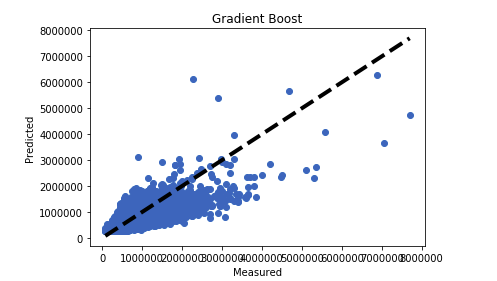


Below table shows the RMSE values for each model against test data.



Below given plot shows actual vs predicted values for the both models - Linear Regression and Gradient Boost. In the graph the points should be near to the best fit line. For lower values of the RMSE the points will be close to the best fit line.





By comparing the RMSE values of the models we can conclude that, performance of gradient boost is better than Linear Regression as the RMSE value is minimum as compared to other.

**Cross validation** - To do comparison between the methods used , we also usedcross validation to split the train and test data.

We have used k-fold cross validation with 5 folds to split the train and test data and the scoring parameter as ‘neg\_mean\_squared\_error’ to calculate the mean square error and standard deviation.

Below given table#4 shows the calculated metrics corresponding to each model.

Table#4

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean Square Error** | **Standard Deviation** |
| Linear Regression | -56890810660.636452 | 5005901737.367044 |
| Gradient Boosting | -48732639597.703529 | 3372717818.256587 |
| Random Forests | -53419707700.304329 | 4022324376.683426 |

**Statistics & P-Value:** To compare the methods, t-test is used to calculate the p value which is shown in the below given table.

Table#5

|  |  |  |
| --- | --- | --- |
|  | **Gradient Boosting** | |
| **Linear Regression** | Statistic | P-Value |
| -2.70017770820947 | 0.027066585661775347 |

As can be seen in table#5, p value is 0.027 which is less than 0.05 so we can reject the null hypothesis and thus we can say that both of these models will behave differently.

We have used the Random Forest model also to compare it with the above mentioned Linear and Gradient Models. Table#6 shows the t-test result set of random forest with linear regression and gradient boost models. As can be seen, both the p values are greater than 0.05 so we can’t reject the null hypothesis.

Table#6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Linear Regression** | | **Gradient Boosting** | |
| **Random Forests** | Statistic | P-Value | Statistic | P-Value |
| -1.1310559172866357 | 0.2907961770710954 | 1.7562978694621363 | 0.11710502857497124 |

**Conclusion :**

We used Root Mean Square Error as a metric to compare the methods. Along with it we used A/B testing as well for the comparison. With the calculation of p value and RMSE values for Linear Regression and Gradient Boosting, we conclude that gradient boosting is the better method. For the third method that is random forest, there is not much difference in RMSE value in comparison with the first two methods. So overall after comparing all the 3 models, it was found that the Gradient Boost model performs slightly better than Linear Regression and Random Forest.

**Learnings:**

While doing the project, we learned the step by step procedure to create the prediction model which includes data cleaning, feature selection, data plotting, model selection , model comparison and evaluation.

**References**:

1. [[www.teacheron.com](http://www.teacheron.com)] Predicting House Prices in King County.
2. <https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab>
3. <https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6>