Applied Statistics and Visualization for Analytics: Final Project

Vigna Shree Telukunta

**Data Description:**

Occupants in each network in the area are confronting an exceptional challenge in finding and keeping a home they can bear. Moderate lodging is a basic segment of our locale's foundation, and we should act together, over all degrees of government and all divisions, to address this emergency and guarantee the wellbeing and decency of our networks and the monetary imperativeness of our locale. This dataset contains house deal costs for King County, which incorporates Seattle. It incorporates homes sold between May 2014 and May 2015.

**Data Content:**

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

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.

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

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 Source:** Harlfoxem. “House Sales in King County, USA.” *Kaggle*, 25 Aug. 2016, https://www.kaggle.com/harlfoxem/housesalesprediction

**Research Questions**

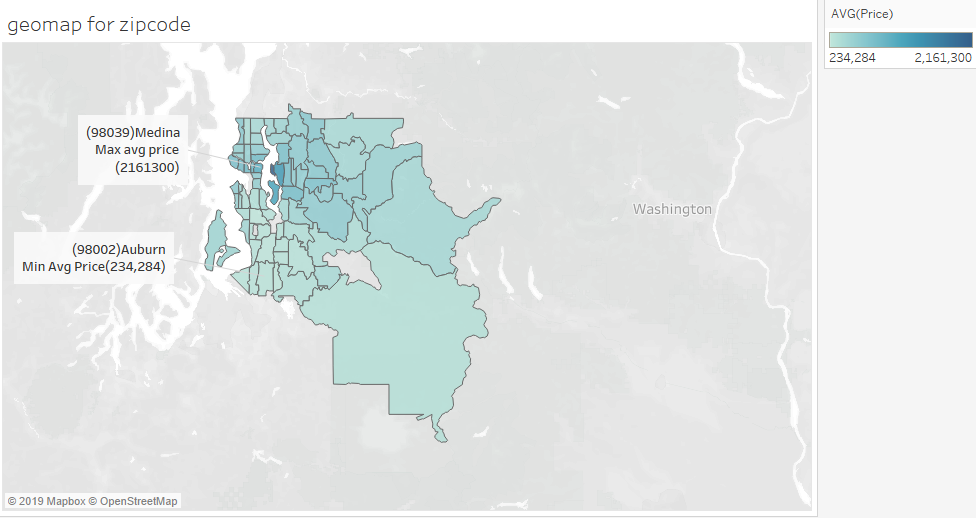
1. When we are buying a new house, which one will be economic (built or renovated)?
2. Which is the costliest place to live in and why is it costly?
3. If we are buying an house mostly which grade houses are available?
4. If we are buying a house at a low price what are the factors that affects the price?
5. Smaller the house lesser the price how do we justify?
6. If we are buying a house, mostly which year it is built?
7. In general, what is the most preferred type of house?
8. Identifying the best statistical model that fits our data?
9. Do we have any outliers? If there are outliers how are they influencing the data?
10. How is the predicted value deviating from the test value?

**Data Preprocessing and Data cleaning:**

Data cleaning or data preparation is an essential part of statistical analysis. To ensure high quality data we performed data preprocessing and data cleaning with our dataset. We removed the inconsistencies in our data. We checked for missing values, this dataset has no missing values. By identifying some incompatible data types for analysis, we transformed the data types and removed few of them namely id, date, lat and long which are not needed for prediction. We found the correlation matrix (heat map) to know the relationship between all variables. We have drawn some of the insights by using visualization techniques.

**Visualization of dataset:**

Research Question: Which is the costliest place to live in and why is it costly?

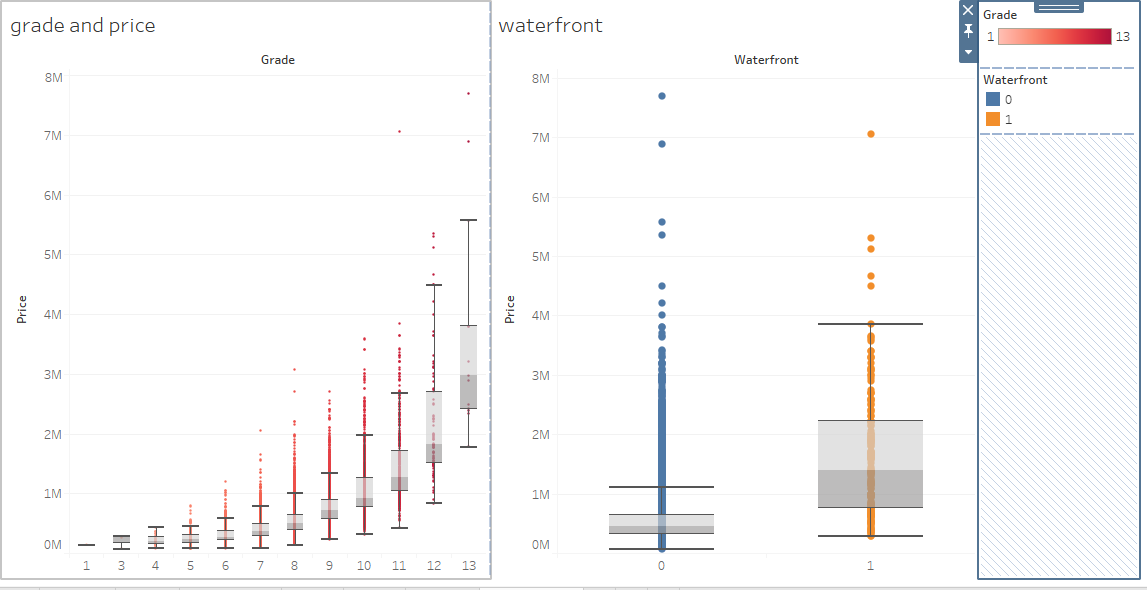


**Fig1: Geomap representing the Average Prices for houses in King County**

From the above Geomap we identified that Medina is the costliest place in King County with an average price of $2,161,300. The cheapest place in King County is Auburn with an Average price of $234,284.

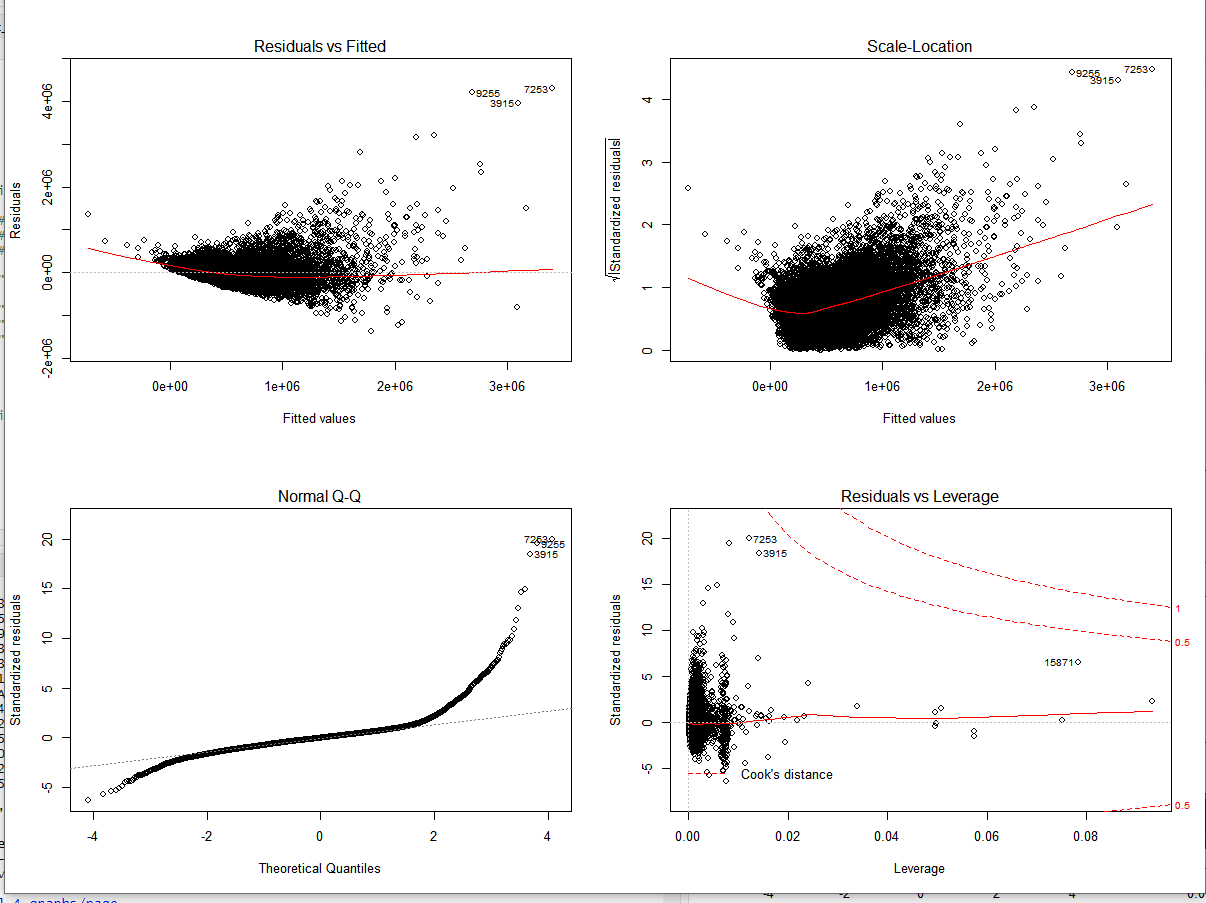
The place Medina is a costliest place in King County because all the houses have a grade which is greater than or equal to 7 and it also has a waterfront which is shown in the below graph. The second costliest place which is Mercer island also has good grade and waterfront.

**Fig2: Boxplot for Grade and waterfront with respect to price.**

From above plot of grade vs price, we observed that price increases as grade increases. From the graph waterfront vs price, we observed that place with waterfront has higher price than place without waterfront.

**Statistical Model:** Research Question: Do we have any outliers? If there are outliers how are they influencing the data?

First we did diagnostic plots and checked outliers using cook’s distance in multiple linear regression model. By doing this we came to a conclusion that there are no outliers in our dataset.

 **Fig3: Diagnostic plots for multiple linear regression model.**

**Residuals vs Fitted plot:**

As the best fit line is in funnel shape but the data distributed is concentrated in the center, from this we found that our data is not linearly distributed.

**Scale Location Plot:**

As the best fit line is in a curved shape so the data does not have equal variance.

**Normal Q-Q Plot:**

We can see that the data is not normally distributed in the beginning and ending of the best fit line.

**Residual vs Leverage Plot:**

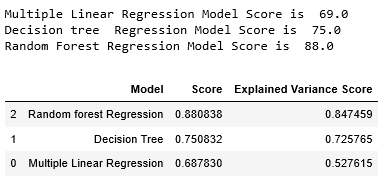
From this plot we find that there are no data points exceeding cook’s distance(>1) so we can conclude that there are no outliers.

By observing he above four plots we conclude that this not the best fit model for our dataset.

**Comparing accuracies for different statistical models:**

Research Question: Identifying the best statistical model that fits our data?

We found accuracy for different statistical models by using python for multiple linear regression, decision tree and random forest regression. We got best accuracy for Random forest regression.

  
 **Fig4: Measuring Accuracies for Different Statistical Models.**

Decision tree is sensitive to data. Even a very little change in the train data set will effect the output drastically. Comparatively, random forest is better as a very little change in train data will not affect the output. This is very important as housing price is a fluctuating data. This is the reason why we have chosen Random forest regression as an optimal model to predict housing prices.

**Random forest regression:**

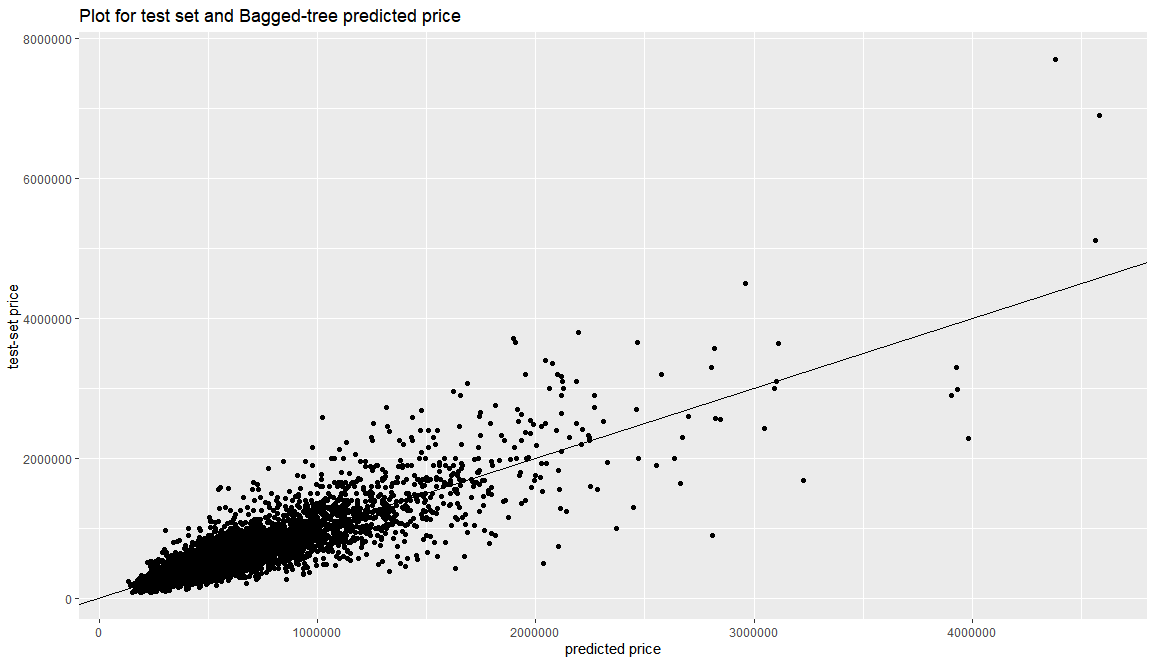
Research Question: How is the predicted value deviating from the test value?

Random Forest is based on ensemble learning which is a type of supervised machine learning algorithm. Random Forest Algorithm combines multiple algorithm of same type which is multiple decision trees resulting a forest of trees. This algorithm can be used for regression as well as classification.

In this algorithm we have multiple trees and every tree is trained on a training set. Even if a data point is introduced it can’t effect all trees as it would only impact one tree. So, it is considered as a stable algorithm. This can be used on both numeric and categorical features of datasets. It performs well even when there is missing data or when data is not scaled properly.

Building the model:

First, we split the data into train and test set, so we can test MSE. We perform bagging by taking mtry=13(attribute split). Then we call randomForest() and then compute the test MSE for 500 bagged trees(default value). Then we plot a graph to observe the deviation between test values and the predicted values.

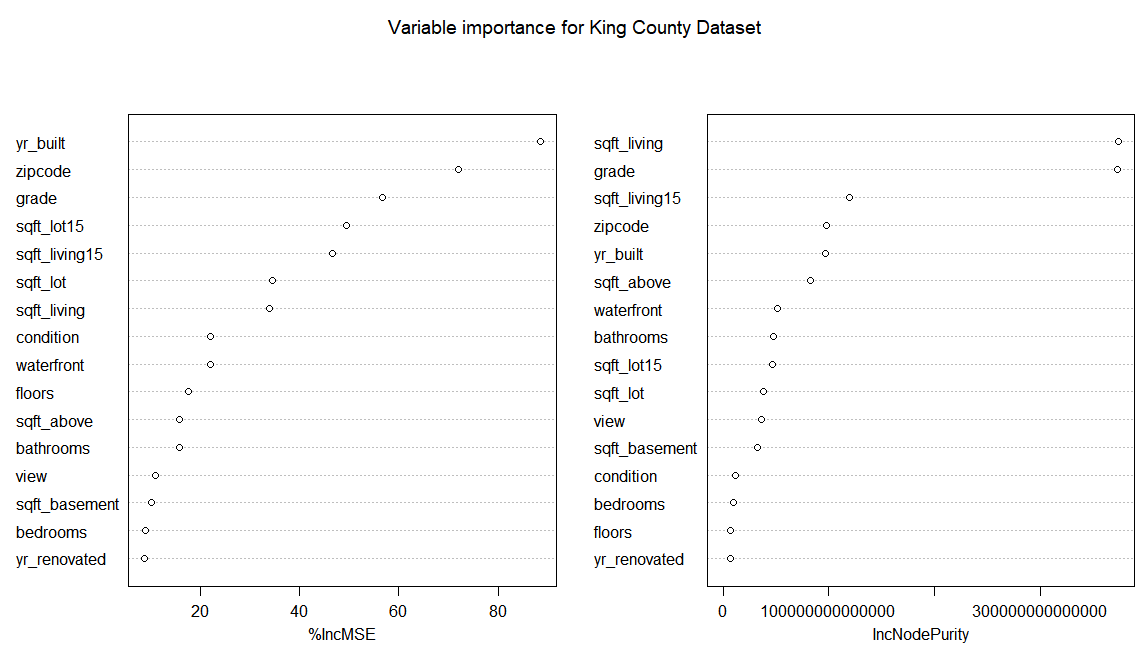


**Fig5: Scatterplot for test set vs bagged tree predicted (price)**

Initially there is not much deviation between the predicted and test set values in the scatterplot, but after a certain point there was a lot of deviation. This deviation cause due to some exceptional cases in dataset.

**Variable Importance plot:**

We plotted the Variable importance for the dataset and then we did random forest by taking mtry=8 and then using four most important variables in our dataset.



**Fig6: Variable importance for King County Dataset**

From the above graph we can say that sqft\_living is the first important attribute in the dataset. Grade is also almost equally important as sqft\_living. Yr\_renovated is the least important variable used to predict the price in this dataset.

Yr\_built and zipcode attributes have highest percentage increase in MSE.

**Conclusion:**

Insights drawn from visualization:

1. We found that places which are closer to the coast are costlier than non-coastal areas.
2. We found that grade and price are directly proportional.
3. We found that as a square feet of living is increasing price is also increasing; price has highest correlation with sqft\_living.
4. We also found that average house prices of built houses are cheaper than renovated houses.
5. The most preferred type of house consists of 3 bedrooms, 2.5 bathrooms and 1.5 floors.

Insights drawn from statistical models:

1. We found that multiple linear regression is not best fit model for our data and we don’t have any outliers.
2. We calculated accuracies for different models and we got random forest is best model as it has accuracy of 88.03%
3. We plotted predicted vs test value and we found that the model fits our data.

**References:**

[1] Harlfoxem. “House Sales in King County, USA.” *Kaggle*, 25 Aug. 2016, https://www.kaggle.com/harlfoxem/housesalesprediction.

[2] Girgin, Samet. “Random Forest Regression in 5 Steps with Python.” *Medium*, Medium, 7 Apr. 2019, https://medium.com/@sametgirgin/random-forest-regression-in-5-steps-with-python-ee4259eca0de.

[3] “Linear Regression R.” *DataCamp Community*, https://www.datacamp.com/community/tutorials/linear-regression-R.