**Capstone Project 1 Final Report**

**Overview:** Even though the Seattle housing market has seen its ups and downs, it’s an exciting time for the King County real estate market since the market is shifting from a seller's market to a buyer's market [2]. In fact, according to a recent King County news publication, King County homes saw dips of 3-6% and condos saw dips of 7% [2]. Therefore, it is predicted that there will be an increase in prospective home buyers [2]. In addition, it is predicted that there will be a huge surge in home purchases, as the largest cohort of Millennials turning 30, will be purchasing their first home [3].

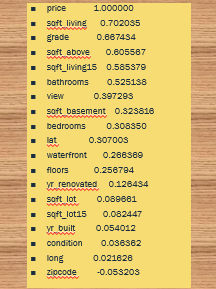
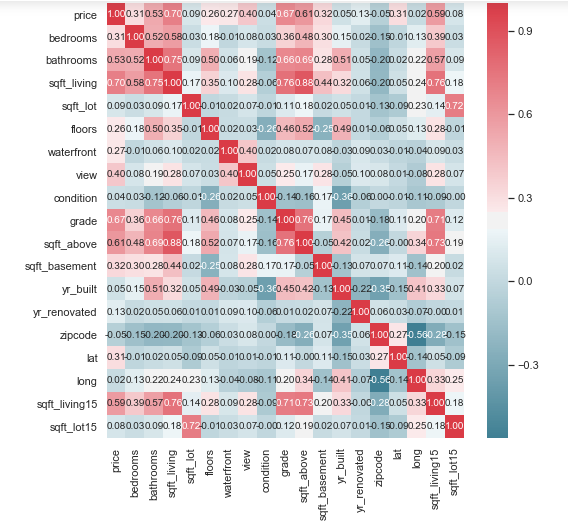
**Problem/Clients:** The problem is that it can be a harrowing experience for new home buyers to know which variables (number of bedrooms, number of bathrooms, number of floors or none of the above) will affect their home price. Moreover, a home purchase can be one of the largest investments of a lifetime with the low end of home prices in King County being upwards of 450,000 dollars. This means that it is important for home buyers to make sure they are choosing a home with good resale value. This also means, understanding the different variables that affect home price will not only be important to home buyers, but also to sellers, real estate agents, appraisers, mortgage brokers, tax assessors, King County officials and lending institutions.

**Dataset:** In this supervised machine learning project, regression classifiers were used to predict house prices in King County, WA which includes not only Seattle, but also the cities of Bellevue, Kirkland and Redmond, WA- just to name a few. The dataset, called House Sales in King County, is from the Kaggle website [1]. The dataset includes homes sold between May 2014 and May 2015 and contains a total of 21 features and 21,613 observations. The original dataset was obtained from the King County public domain website. Not including the target variable of price, the following list includes the names of the 20 house price predictors: date, id, bedrooms, bathrooms, sqftliving, sqftliving15, sqft lot, sqftlot15, floors, waterfront, view, condition, grade, sqft above, sqft basement, yr built, yr renovated, zip code, latitude, and longitude.

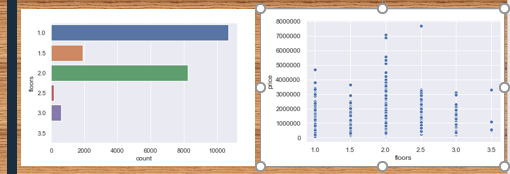
**Overall Methodology:** The project methodology first involved data wrangling techniques such as removing missing values, utilizing the melt function to fix “wide format”, removing unnecessary features and removing any proven outliers. After that, individual features were viewed utilizing different visualizations in order to look for trends and correlations with price. Then, hypothesis testing was performed to look at if there was a difference in the means and if the null hypothesis needed to be rejected. The best features were chosen by implementing different feature selection methods and 10-fold cross-validation testing was also done. A supervised machine learning regression model was utilized and the performance was measured. In order to avoid overfitting, regularization testing was implemented. Finally, the predictor model was plotted visually as well as tested to ensure it was a good predictor of home prices in King County, WA.

**Data Wrangling:** As part of data wrangling, the dataset was read into Jupyter notebook and the data frame was evaluated carefully. Out of the 21 variables, [id] was dropped from the dataset. The data types of the individual columns revealed that all of the columns were either integers or floats except for the [date] feature which was an object. The [date] feature was listed in time series format and was replaced by a newly created column called[year] in order to separate out the 2014 homes from the 2015 homes. The King County.gov website was researched and it was found that there was 108 total zip codes in King County. The dataset included 70 unique zip codes and did not include any outliers or zip codes outside of King County. Thus, the zip codes in this dataset were deemed a good representation of King County. The columns did not need to be melted and there were no missing values in the dataset.

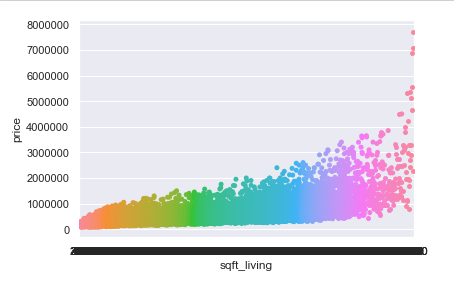
**Visualizations/Data Exploration*: The correlation visualizations came across many important feature trends:***

The trend that was discovered from the following visualizations was that sqft living, grade, sqft above, sqft living15 and bathrooms were the highest correlated features with price. There was also a high correlation on the heatmap between the number of bathrooms and sqft living which makes sense since larger homes could have more bathrooms. There was also a high correlation between grade and sqft living and, according to a publication, this could be due to new homes being built larger [4]. An interesting strong negative correlation was found between the number of floors and building condition. According to a publication, this could be due to two story homes requiring more repairs [5]. So does this that mean King County home buyers do not buy as many two story homes? The count plot found just that; it found that the majority of homes in King County have one floor even though, according to a publication, 69% of pacific state homes have two bedrooms [6]. The discrepancy in King County, could be due to the county's building codes or to the high cost of construction in King County [7]. In conclusion, floors did not have a strong correlation with price since the scatterplot showed that one floor could have the same price as multiple floors.



Another interesting trend found was that although building grade was the second highest correlated feature with price, the average building grade in King County was only 7.5, and a building code of 6 is barely passing. When plotted, building grade showed a nice, normally distributed histogram. In fact, out of the 21,613 observations only two homes had a building grade rating of 13 and a building condition of 4! There were zero homes with building grade 13 and building condition 5. In addition, there were only 13 homes with building grade rating 13 while there were 1701 homes with building condition rating 5. This might explain why building grade is such an important price predictor: there’s not a lot of high building grade homes! Looking at the data frame, it almost seemed that there was an indirect relationship between grade and condition and this was proven by the negative correlation between the two features on the correlation heatmap. To clarify definitions, grade refers to the overall building construction and adherence to building codes and condition refers to the functionality of things such as painting, roofing, plumbing, or heating.

Next, violin plot visualizations of homes with and without basements were analyzed for trends. The violin plots showed that there were almost double the homes in King County without a basement as there were with a basement. What was unclear was if the marked increase in home price with basements was due to the basement or to the home being larger and the high correlation between sqft\_living and price as the plot below shows. 

According to kingcounty.gov, the lack of homes with basements in King County may be due to the constant rain in the winter causing a less stable soil composition [8]. Whatever the reason may be, it appears that basements may not be the strongest predictor of home price in King County.

Since views was listed as the 6th highest correlated feature with price, a scatter plot was created that looked at how many views a home got from prospective homebuyers against price and building condition. The plot essentially helped us to determine if the condition of a home determined how many views a home got.  The first finding was that the number of prospective home buyer views was highest when the price of the home was also the highest regardless of the condition. The reason for this may be that other attributes such as good school systems or proximity to work outweighed the importance of building condition for home buyers. The second finding showed that there were high views of lower priced homes in good condition which makes sense since home buyers want good condition homes that are cost-effective.

Next a scatter plot of yr built against yr renovated and price was plotted to see if there were any trends and if these features were correlated with price. The findings were that homes built in the early 1900's were just as highly priced as those built in the 2000's. According to the Seattle Times, the reason could be that, "The number of homes on the market is at a low point, despite the big increase in population"[9]. The second finding was that the majority of home renovations were done on homes built before the 1980’s. In conclusion, yr renovated did not show much correlation with home price even though yr built did show a slight correlation.

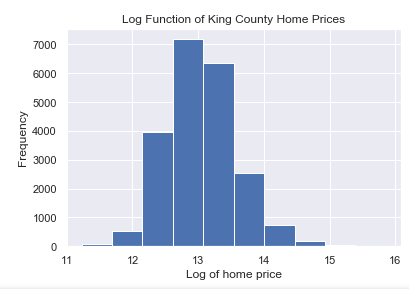
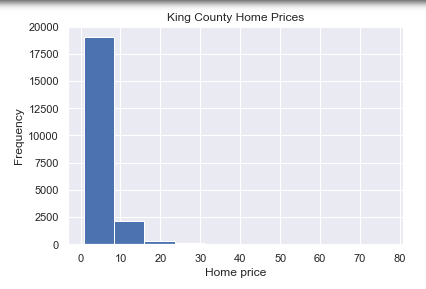
Even though latitude was not rated highly on the correlation list, a scatter plot with latitude was plotted against price to see if there were any important trends. The scatter plot findings were interesting since they showed that the highest priced homes were within the latitude of 47.6. The latitude of 47.6 not only includes the cities of Seattle, Medina and Bellevue, WA but is also home to Amazon and T-mobile! The second largest group of highest priced homes, in the latitude of 47.7 not only contains the cities of Kirkland and Redmond, WA but is also home to Google and Microsoft! King County house price stats indicated those areas are, indeed, highly priced, thus latitude may be an important predictor of home price [10].

Next a scatter plot of bedrooms and bathrooms were plotted against home price. The findings of the scatterplot was that although there was a direct correlation between home price and the number of bathrooms, after about 3.5 bathrooms, the home prices, actually, leveled off. The second finding was that price and the number of bedrooms only appeared correlated up to about 5 bedrooms and, overall, did not seem very correlated with price. ECDF functions were also plotted for bedrooms and bathrooms and the first ECDF function showed that the 50% probability mark for the number of bathrooms was around 2 which aligned with the calculated mean of 2 bathrooms. The second ECDF function showed that the 50% probability mark for the number of bedrooms was around 3 which also aligned with the calculated mean of 3 bedrooms.

Since a column was created for [year] in the dataset, an ECDF of 2014 vs 2015 was plotted on the same graph to see if there was any difference in price between the two years. The result was that the ECDF functions for 2014 and 2015 surprisingly overlapped! The 50% probability mark for home prices in 2014 and 2015 were about the same and thus, [‘date’] and [‘year’] were found to not be the best predictors of home price.

Finally, an ECDF plot as well as a box and whisker plot were done for the mean home price to ensure that the dataset was normalized and that there were no outliers. Plotting the home prices on the ECDF plot showed us that the 50% probability mark fell around 500,000 dollars which aligned well with both the mean home price of 540,088 dollars and the median home price of 450,000. The box and whisker plot however, showed that there were some higher priced homes that may be outliers as they were skewing the mean. Next, hypothesis testing was started in order to help decide if these higher priced homes needed to be removed from the dataset.

**Hypothesis Testing:** In bootstrap hypothesis testing, the null hypothesis was that the difference of means between the full data sample and the bootstrap random data sample = 0. The alternate hypothesis was that the difference of means between the full data sample and the bootstrap random data sample != 0. Based on the p-value of >.05, we failed to reject the null hypothesis and concluded that there was no statistical difference between the mean of our bootstrap sample and the mean of our housing price data set. Even though all three different hypothesis tests revealed no statistical difference in the means, the standard deviation was found to be large indicating that there was a high degree of variability of some data points from the mean. However, before the higher priced homes were removed from the dataset, the log function was taken of the home prices to see if that normalized out the data. Based on the histograms shown below as well as the new log mean value of 13.05 and median value of 13.02 aligning, it seems that taking the logarithmic function of home price successfully normalized out the home prices. Hypothesis testing was also redone on the log of home prices and there was no statistical difference found between the difference of the means. The box and whisker plot was also re-done and the median line and mean marker were found to be aligned and there were no outliers on the plot.



**Supervised Machine Learning :** Since price is a continuous and ordered feature, regressor models were utilized for home prediction. The first machine learning classifier used was linear regression and a total of five different feature selection combinations were chosen utilizing different feature selection methods in addition to 10-fold cross validation in order to create the best model for home price prediction.

The first feature combination included the four highest correlated features with price as seen on the correlation heatmap. The four highest correlated features were: grade, sqft\_above, sqft\_living, and sqft\_living15. The linear regression model was evaluated based on the coefficients, the mean squared error, the root mean squared error, the R^2 value and the cross-validation score. The R^2 score was 0.57 and average 10-fold cross validation score was 0.57.

In order to improve the model further, each of the 19 individual features were evaluated utilizing OLS regression results. The features with AIC/BIC values that showed to be the weakest price predictors were: bedrooms, year, floors, sqft\_basement and yr\_renovated. The features with almost zero R^2 value were: long, zipcode, year and yr\_renovated. The highest R^2 value features were: grade, sqft\_living, sqft\_above and sqft\_living15. Thus, the features chosen were: grade, sqft\_above, sqft\_living, sqft\_living15, bathrooms, lat, and view. Utilizing these seven features, greatly reduced the MSE, RMSE and increased the R^2 value to 0.72.

Since adding in additional features improved the model, an additional five features were added, still utilizing OLS regression test stats, for a total of 12 features. Interestingly, even though there was an improvement, it was a very slight improvement with a slight reduction of MSE and RMSE and a very slight increase in R^2 to 0.73 and average 10-fold cross-validation score was 0.73.

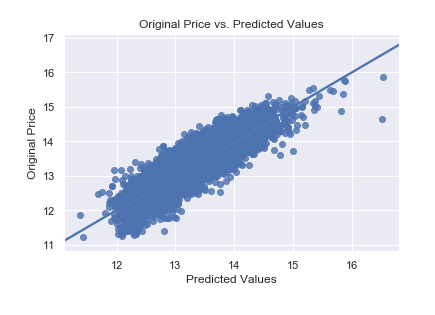
The next feature selection method utilized was sklearn feature selection. Out of the 19 total features, sklearn removed the following 5 features: bedrooms, year, floors, sqft\_basement, and yr\_renovated. Consequently, year, yr\_renovated, floors and bedrooms agreed with data exploratory analysis results as not being the best predictors of home price. But in data exploratory analysis, it was unclear whether sqft\_basement was correlated with price or not. The sklearn feature selection removal of sqft\_basement was an important finding since it indicated that sqft\_basement was not a significant home price predictor in King County. Sklearn feature selection also showed that all location-related features were important to the housing price prediction model. Utilizing these 14 features showed a further reduction in MSE and RMSE and an increase in R^2 to 0.76.

Finally, a combination of features from both sklearn selection, R^2, AIC/BIC, p-value, coefficients, RMSE, MSE,and cross-validation were chosen to see if the model would improve even further. The 15 features chosen were: [grade, bathrooms, bedrooms, sqft\_living, sqft\_lot, floors, waterfront, view, condition, yr\_built, zipcode, sqft\_basement, lat, long, and sqft\_living15]. Sklearn feature selection method removed [year and yr\_renovated]. OLS analysis removed [sqft\_lot15 and sqft\_above]. Combining the feature selections together gave the best model! The RMSE and MSE were further reduced, the R^2 value was 0.77 and the cross-validation score was 0.77.

Next, the random forest regressor model was utilized and gave an accuracy on the training subset of 0.98 and an accuracy on the test subset of 0.88. The same 15 features also created the best home price prediction model.

Next, the gradient boosting regressor model was chosen and gave accuracy on the training subset of 0.94 and accuracy on the test subset of 0.90. This indicated that the gradient boosting regressor model was the best classifier of all the three machine learning models for home price prediction. The same 15 features again, also, gave the best result.

Finally, the Ridge and Lasso regularization tests were utilized and both gave a model score of 0.77. When the original vs predicted price values were plotted on a scatter plot, the result was highly correlated and indicated that the machine learning model is doing a good job of predicting King County home prices.



**Conclusions:** A total of 15 features were utilized from our original dataset of 20 features. Although sqft\_living and building grade were important features to the housing price prediction model, the following features did not have as much impact and were eliminated [year, yr\_renovated, sqft\_above and sqft\_lot15]. Even though sqft\_basement, floors and bedrooms were kept as features, they were not found to be high predictors of home price in King County. The location-based features [zip code, latitude, and longitude] had a surprisingly large impact on the machine learning models as removing them significantly reduced the model's ability to predict home prices! In conclusion, the old adage about real estate is correct:" There are three things that matter in property: Location, Location, Location!"-quote by Harold Samuel.

**Thank you!**

**Capstone Final Report by: Roshnee Raval, June 2019**

**References:**

**Springboard Mentor:** Harsh Vardhan Singh

[**1] https://www.kaggle.com/harlfoxem/housesalesprediction**](https://www.kaggle.com/harlfoxem/housesalesprediction)

[**[2]https://www.seattletimes.com/business/real-estate/seattle-area-home-prices-drop-to-lowest-point-in-two-years-down-116000-since-last-spring/**](https://www.seattletimes.com/business/real-estate/seattle-area-home-prices-drop-to-lowest-point-in-two-years-down-116000-since-last-spring/)

[**[3]https://www.forbes.com/sites/alyyale/2018/12/06/2019-real-estate-forecast-what-home-buyers-sellers-and-investors-can-expect/**](https://www.forbes.com/sites/alyyale/2018/12/06/2019-real-estate-forecast-what-home-buyers-sellers-and-investors-can-expect/)

**[4]https://www.nytimes.com/2016/06/04/upshot/houses-keep-getting-bigger-even-as-families-get-smaller.html)**

**[5]https://www.doityourself.com/stry/the-extra-costs-of-2-story-homes**

[**[6]https://magazine.realtor/daily-news/2016/08/08/one-vs-two-story-homes-which-dominates**](https://magazine.realtor/daily-news/2016/08/08/one-vs-two-story-homes-which-dominates))

[**[7]https://www.boardandvellum.com/blog/cost-of-construction-in-seattle/**](https://www.boardandvellum.com/blog/cost-of-construction-in-seattle/)

**[8]https://kingcounty.gov/~/media/depts/dnrp/solidwaste/ecoconsumer/documents/SeattleTimes\_2013-07-26.ashx?la=en**

[**[9]https://www.seattletimes.com/business/real-estate/why-are-seattle-area-home-prices-so-high/**](https://www.seattletimes.com/business/real-estate/why-are-seattle-area-home-prices-so-high/))

**[10]https://gismaps.kingcounty.gov/iMap/**