Historical Real Estate Pricing Analysis

By: Alex Peterson, Xinyu Wang, Jerry Tsao, Anlun Duan

**Data Description**

King County in Washington, United States is one of the biggest counties in the country with a population of 2,233,163 as of the 2010 census. One reason for such a high population is that King County contains Seattle, one of the largest cities in the country. Seattle Washington is known as one of the most expensive cities to live in the country. In 2018 the median home cost was 623,800 dollars. The data measured comes from the years 2014 through 2015. Variables measured include: price, year built, date purchased, square footage, lot size, neighbor square footage, and neighbor lot size, grade of house, year refurbished, latitude, longitude, number of bedrooms, number of bathrooms, number of floors, whether the lot was a waterfront home, and view (rated 1 through 4).

**Descriptive Statistics**

When looking at real estate sales over a certain amount of time, price tends to be an important area. In this analysis, price is treated as the response variable. The variable price has an interquartile range of 321,950 to 645,000, as well as a mean of 540,088. This trend tells us that the prices of houses in King County between 2014 and 2015 are being leveraged upwards by some significantly high prices. This is most likely coming from houses that have been sold in Seattle. This mean is skewed, however the median of 450,000 is in the expected area considering the interquartile range.

**Question 1**

One area of interest being investigated is which housing characteristics are the most impactful on the price of a home in King County. In order to determine the characteristics, lasso regression and XGBboost Regression were used. These models were chosen in order to combat potential multicollinearity in the explanatory variables. The results of these models were R-squared values of 0.69 and 0.90 respectively. We find from these models that the most important variables are square footage, grade of house, latitude, longitude, and neighbor square footage.

**Question 2**

A saying in real estate says that the value of a home is, “location, location, location.”. To test the accuracy of this long told tale, the effect of a home's zip code was investigated. King County Washington contains a total of seventy different zip codes. We implemented a one-way analysis of variance using zip code with price as the response variable. The results show that zip code is significant at a p-value of less than 0.001. Using Tukey’s HSD we investigated the comparisons between the prices of different zip codes. When looking at the spread of the boxplot, the significance of the analysis of variance is upheld as it seems the prices of different zip codes in King County are quite different from one another (Figure 1).

**Question 3**

Another aspect of purchasing a home is the effect time of the year plays on the price of a home. We implemented an analysis of variance on both the quarter and the month that the homes were purchased. The ANOVA showed that the quarter in which the house was sold is significant beyond p=0.001 and month was not significant beyond 0.05. This information leads us to believe that the quarter, or the season in which a house is being sold will affect its price. However, the individual month is not significant.

**Question 4**

Two of the most interesting variables in this dataset are the neighboring square footage and the neighboring lots square footage. These variables consider the closest fifteen neighbors and the size of their homes and their yards. Rather than building a model with these variables as they were, the variables were adjusted to be relative to each house’s square footage and square footage of lot, meaning that if a house is larger than its neighbors, the result will be above one for neighboring square footage and if the number is lower than one, then it is smaller relative to the neighboring square footage. The hypothesis for this question was that if the house was smaller than its neighbors, then it would sell for less than an equivalent house that was bigger than its neighbors. To test this, a linear model was fit to the data using relative square footage and relative square footage of the lot. The results concluded that relative square footage was significant at a level of 0.001, and relative square footage of lot was found to be significant at a level of 0.005. Then to test the ability to differentiate high price houses and low price houses, a logistical regression model was run using relative square footage and relative square footage of the lot. It was found that relative square footage was significant in the logistic regression at a p-value less than 0.001 however relative square footage of the lot was not found to be significant.

**Discussion**

In the investigation of our research questions, we were able to discover that square footage, relative square footage, relative square footage of lot, grade of house, latitude, longitude, quarter of the year, and zip code are among the most important variables. This finding supports the logical hypothesis that most home buyers are looking for a home that is large, quality grade, in a good location, and good value comparatively to neighbors houses. Our research supports the claim that people are more likely to buy homes that are below the size and value of their nearest fifteen neighbors. Overall, King County Washington is a diverse county ranging on the high side of price when compared to other counties in the US. The insight into its real estate sales is able to show us how and when people buy homes.

**References**

Dataset: <https://www.kaggle.com/harlfoxem/housesalesprediction>

**Figures**

**Figure 1.** Boxplot of zip code vs price for each zip code in King County Washington. Result from question 2.

