Project Report: Predicting House Sales
By: Sarita Patel

Executive Summary:

Housing market pricing outcomes reflect local economic conditions of any state. On average nearly 6 million houses are sold in the United States Annually with the average cost of each house being $200,000. Additionally, most homebuyers do not have the cash to make an outright home purchase, so they take out 20 to 30-year mortgages to help finance their home. For most home buyers, the purchase of real estate is one of the largest and most impactful financial transactions they will make in their lifetime. This article aims to review house features such as number of bedrooms, number of bathrooms size of living room in square feet, condition of house and many more relevant features to estimate the price of the house. Through our research we aim to answer the question of which features play an important role in predicting the house sale price for king county homes sold between May 2014 and May 2015. This research will seek to estimate house sale price and help identify important component that can help individuals to make good decisions about home buying.

The application of this research from the perspective of buyers who are planning to purchase a new house is that it will help them compare prices and check the condition of homes to avoid renovations problems. Also, buyers who are planning to purchase previously owned homes should get a better overall value for their money. Our study also aims to have a clear understanding about the house prices within a specific range in the same area with similar features. Ideally houses with same number of bedrooms, bathrooms, square footage of the apartments interior living space, square footage of land price should cost the same.

This research will take into consideration several important features about houses such as number of floors, whether there is a water body or not in front of the house, how good the view of home is, what is the condition of home, what is the construction design, what is the square footage of the interior housing space that is above ground level and what is the square footage of the interior housing space that is below ground level and in which year the house is built whether is new or old and when is it renovated and in what latitude and longitude the house is located followed by its zip code and what is the square footage of interior housing living space in 2015 and the square footage of the land lots in 2015. Analyzing these features will give insight into the influence they hold over house prices to help the buyers/sellers decision.

Utilizing all variables mentioned above, we are applying various methods to predict the house price. Our methods include Model building techniques like linear regression and dimensionality reduction techniques like principal component analysis, exploratory factor analysis and cluster analysis techniques like k-medoids and LDA (Linear Discriminant Analysis) etc.…

The data was processed and cleaned prior to applying our methods. We identified that there were no missing values in our data but decided to remove or reclassify some of the variables based on the specifics of the data type and the technique being used. Performed multicollinearity test to avoid collinearity issues. To predict house prices our initial method was to fit a simple linear model to the raw data which explained a large portion of variability in house prices. We then applied principal component analysis to group the various features of houses into separate components based on which were most closely related. Next, Factor Analysis was applied to help us to simplify our data by grouping each data point with other data points that are closely related. This allowed us to focus on groupings of data instead of each individual data point.

Cluster analysis was applied to our numeric data to understand what underlying groups may exist in the dataset. These clusters can offer empirical insights into what types or attributes of houses are similar to one

another. Additionally, understanding these clusters could help confirm with seller that they are accurately pricing their house. If their house is like a house in a higher price range, then the consumer may be more inclined to increase the price of their house. Additionally, consumers could use this analysis to understand if they are making a sensible investment.

A simple linear regression is applied on the cleaned and transformed data to devise statistical relationship between the sale price and other variables. A significant model which can explain 77% of the variance in the sale price of the house is obtained from the model we can come to the conclusion that Square footage of living area of a house and the grade of the house had positive correlation with the sale price of the house. Houses with a view of waterfront tend to sell at higher prices than other houses.

The purpose of running LDA on the kc_house_price dataset was to identify the accuracy of the model from configuration of the IV's that can help us predict the value of the dependent variable. Lastly, we applied Linear discriminant analysis (LDA) that discovers a grouping pattern among different categories of houses and test the accuracy of model that was created. LDA gives the appropriate price estimate and category split which can help assist buyer/seller to look and understand the condition of the house and make an appropriate investment.

Limitations of the research and future work:

Our research was limited by the type of data that was utilized and the scope of the data collected. The data utilized had only a few descriptive types of data (categorical) which limited us from utilizing various other analysis techniques. In future work we recommend expanding the scope of the data collected to include information on other variables that may impact the price of a house. Potential future data should include the level of technology integration of the house, signifying what features are able to be controlled electronically (smart fridge, smart doorbell, smart tv, etc…). Another area of focus should be variables that are not direct features of the house, rather amenities and services that the homeowners would have access to. Quality of local public schools, police force, community centers and public parks are all relevant data types. Lastly, the amount of any potential estate tax should also be accounted for. Additionally, all our data was collected from a single county in a single state. We recommend collecting data from various metropolitan cities and rural villages around the United States. Adding the aforementioned data will help potential homebuyers and sellers gain more insight into this important purchase as well as make the data collected more relevant to a larger percentage of the population.

Final Conclusions:

From our research we were able to isolate multiple features which greatly impact the price of a house. Features such as the number of bathrooms, the number of bedrooms, and the square feet of the living room all were proven to be strongly related to the price of the house. This information can be leveraged by potential home buyers to gain insight into if they are getting a great deal on a house or if they are being overcharged. Conversely, our findings can also be utilized by homeowners looking to sell their house by allowing them to set a competitive asking price based on the features of their home. Equipped with our findings people may be able to save thousands if not tens of thousands of dollars in real estate costs. However, we must keep in mind that our research is based on data from a single county containing a metropolitan city (Seattle) and may not be applicable to all other counties of the United States.