King County Housing:
Predicting Price and Grade through Statistical Modeling

Cameron Joe

DS 100: Special Topics in Computer Science: Information Management
Professor Franks and Professor Kharitonova
June 12, 2020

1. Abstract

Real estate composes a huge portion of any economy and is a cost that cannot be avoided for the average person. The problem that this paper aims to address is to build a model that is capable of accurately pricing houses using relevant variables of interest as predictor variables. Another question of interest is to determine whether it is possible to predict a house's grade -the categorical variable ranging from 1 to 13 indicating the quality of construction and design- when given a house's price. Other relevant sub questions include: "How accurate are the models at their predictions?" and "Which models fit the data best?" The data used was the "House Sales in King County, USA" dataset downloaded off of Kaggle.[1] The data consists of 21,613 entries and 21 variables for each entry. Principle choice analysis was used to explore the variance in the data; The first principle component explained 37.08% of the variance in the data and had the highest correlation coefficient of 0.932 when compared with the "sqft_living" variable. Therefore, when a simple linear regression was conducted to predict price, the predictor variable used was "sqft_living", and the model resulted in a root mean squared error 209,616. The following multiple linear regression model proved to be more accurate with a root mean squared error of 172,483, but neither error could be qualified as satisfactory. For the second question, the "grade" variable was transformed into a binary variable named "grade_bin". Least squares analysis on the new feature provided a root mean squared error of 0.433 which was improved by changing the model's decision rule, resulting in a zero-one-loss of 0.288. Lastly, logistic regression proved to be slightly less accurate with a zero one loss of 0.321. After concluding analysis, it can be said that the multiple linear regression model fit the data better than the simple linear model, but predictions were not accurate enough to be considered satisfactory. In addition, least squares analysis was relatively successful at predicting the binary grade of the houses when given the houses' price.

2. Introduction

The primary goal of the analysis was to create a model that could accurately predict the sales price of a house. The secondary goal was to see if a categorical quality such as the grade of the house could be predicted based on the price of the house. The project of interest is relevant to almost anyone, as a home is often the most expensive purchase an individual makes in their lifetime; understanding what characteristics and qualities affect a house's price can lead to more informed purchasing decisions in the future. The relevance and utility of this area of study can be seen in the amount of research and publications put into the issue. A team of Hartford University students conducted a much more complicated and comprehensive analysis on the same database and constructed multiple predictive models including least squares, stepwise forward, recursive partitioning, and neural network models.[2] In addition, Eric Kim posted a helpful article on

[1] "House Sales in King County, USA," Kaggle (harlfoxem, August 25, 2016), https://www.kaggle.com/harlfoxem/housesalesprediction.

[2] Abdallah Alsaqri et al., "Predicting King County House Prices," LinkedIn SlideShare, September 12, 2017, https://www.slideshare.net/PawanShivhare1/predicting-king-county-house-prices.

Kaggle describing how to predict and visually represent housing prices through several methods including lasso, elastic net, and gradient boosting.[3] Lastly, in an entry in the International Journal of Innovative Technology and Exploring Engineering, Naga Satish and fellow authors go more in depth on prior literature on housing price predictions. The team also addresses the conceptual risks and benefits of several design approaches including: linear regression, multiple regression, and lasso regression.[4] The dataset analyzed was the "House Sales in King County, USA" taken from Kaggle[5]. This dataset was the most appropriate as it contained a rich set of data (with over 20,000 entries each with over 20 variables) and the data was collected relatively recently (2015-2016). In addition, if the models created would be used here in California, it would be more appropriate to choose data in the US, specifically the west coast (as King County is in Washington).

3. Questions of Interest

The primary goal of this project was to build a model that could forecast a house's selling price based on variables such as: square foot living space, number of bedrooms, and condition. The secondary goal of the project was to see if categorical qualities such as grade could be predicted from price alone. Questions of interest include: "Which individual variable would be most accurate at predicting house prices when using a single linear regression?", "How accurate are the models at predicting house prices?", and "Which models are best at predicting house prices?"

4. Data and Methods

As previously stated before, the data examined was the "House Sales in King County, USA" database, taken from Kaggle[6]. The data itself was collected and provided by the government of King County, Washington and the license for the data was classified as CC0:Public Domain. There was no information on the authors of the data to be credited. Variables used for analysis include: date, price, number of bedrooms, number of bathrooms (including halves), square foot living space, square foot lot space, number of floors (including halves), waterfront (either being 0 or 1), view (the property's view rated from 0 to 4), condition (rated from 1 to 5), grade (the quality of design and construction rated from 1 to 13), square footage above ground level, square footage below ground level, year built, year renovated, averaged square foot living area of the 15 nearest neighbors, and average square foot lot area for the 15 nearest neighbors. Variables provided but not used for analysis include zip code, latitude, and longitude. The data was both collected and provided by King County, Washington,

[3] "House Sales in King County, USA," Kaggle.

[4] G Naga Satish et al., "House Price Prediction Using Machine Learning," *International Journal of Innovative Technology and Exploring Engineering Regular Issue* 8, no. 9 (October 2019): pp. 717-722, https://doi.org/10.35940/ijitee.i7849.078919.

[5] G Naga Satish et al., 717-722.

[6] "House Sales in King County, USA," Kaggle.

specifically the King County Housing Authority and Seattle Housing Authority.[7] Looking at the government website for King County, it seems that the government officials are both conscientious and open about collecting and sharing data.[8] On a page regarding public data, they encourage people to "build custom transit maps" using data that they provide.[9]

In recording and using the data for analysis, the most pertinent principle of measurement is relevance. In looking at the recorded variables in the dataframe, it must be asked whether the information provided is relevant to housing sale costs and what other possible relevant variables are absent from the dataframe as housing price can be potentially affected by a plethora of variables. Taking ethical consideration into account, it is possible that building and using a model to evaluate the prices of homes could cause the values of peoples' properties to decrease, negatively affecting them economically. It is also possible that price data used to build the model could be biased by economic or class discrimination, and that a model built from this data could learn from and perpetuate this discrimination. For example, if latitude and longitude data were used for analysis, the model might learn from humans to depreciate the value of a house due to its location in a neighborhood of color. A model is only a predictive tool- it is up to humans to build and use it ethically.

The data itself pertains to housing prices for properties sold in King County between May 2015 and May 2016. Due to the brief description in Kaggle, it is not well known what is over-represented in the data. Possible over-representation could include houses that were sold twice within the allotted time period. (It was assumed that these houses were likely significantly renovated and flipped at a higher price and were therefore kept in the dataframe.) In addition, it could be argued that houses that failed to sell in this time period are not represented. Another key issue of representation involves the intended use of the model once created. Data from this specific source only pertains to King County. If the model is to be used in other parts of Washington, the US, or the world at large, then the data in use is severely lacking in representation and robustness. Ethical issues consist of mainly the previously mentioned harm that could come from a model perpetuating bias and discrimination. The data itself has free access to the public domain and is provided by King County with the thought of protecting the privacy of sellers and buyers of houses through anonymity.

For the first purpose of creating a model to predict housing prices, first a simple linear model will be conducted followed by a more comprehensive multiple linear regression. Linear regression multiplies the recorded variables of interest by weights in order to calculate an estimate for the variable of interest. The linear model itself will calculate weights that will minimize the mean squared error between the actual and predicted values of price. For the

[7] "Health and Housing Data Dashboard," Health and housing data dashboard - King County (King County), accessed June 13, 2020, https://www.kingcounty.gov/depts/health/data/health-housing.aspx.

[8] "Open Data," King County Open Data (King County), accessed June 13, 2020, https://data.kingcounty.gov/.

[9] "Health and Housing Data Dashboard," Health and housing data dashboard - King County.

second question regarding the prediction of binary grade by price, both a least squares analysis and a logistic regression will be conducted.

In inspecting the provided data, some degree of feature engineering was deemed appropriate for later analysis. It was reasoned that if house renovations were significant enough, then it may be more appropriate to use the year of renovation instead of the initial build year. For that reason, a new feature was created denoted “yr” that would substitute the year of renovation in place of the build year if the house was renovated. In addition, instead of using year purchased or the newly created “yr” variable, it may be more accurate to create a new variable named “yr_old” that would subtract the two. The oldest recorded house was built in 1900 meaning that only the past ~100 years are relevant for calculations. For this reason, “yr_old” was thought to be a more appropriate variable to consider for multiple linear regression. In addition, cleaning of the data set was done to remove odd or extreme values. Upon first inspection of the database, one house was found to have 33 bedrooms, but only cost 640,000. These numbers appeared very odd and were removed as it was assumed to be a human error. Also, some properties were listed to have no bathrooms, no bedrooms, or both. These properties were found to have extremely low prices and struggle to be considered houses (as is the intended target of the model) so they were removed. Lastly, when creating a distribution of housing prices (Appendix 1), it can be observed that there is a severe right skew in housing prices; these expensive properties are extreme outliers that could exert a disproportionate amount of influence on models later one. There are two methods to removing outliers from a data frame and both were tried and compared. The first involves simply removing values that exceed a certain value. Using the 68-95-99.7 rule -which assumes a normal distribution which the data does not perfectly fit-, being two standard deviations from the mean was determined to be the cutoff for extreme values. Prices beyond two standard deviations were removed from the dataframe. The second method is winsorizing, which involves removing the “x” most extreme values and replacing those values with the newest most extreme. Winsorizing has an advantage over simply removing extremes, as it preserves some of the impact that the extreme values had. In contrast, if extreme values are due to errors in collecting or recording data, it may be more appropriate to throw out those values altogether. As the extremely expensive houses were not errors but rare outliers, the winsorized data was used in following calculations as it would still maintain some of the impact from these houses.

Principle component analysis allowed for more comprehensive exploration of the variation in the data. By creating a scree plot (Appendix 2) it is possible to visualize the percent of variance explained by an individual principle component. It was observed that 37.08% of the variability in the data is explained through the first principle component. This knowledge is not very applicable upon first inspection and further investigation must be done to try and discover what variable(s) the first variable corresponds to. After plotting the first principle component in relation to the predictor variables, it was observed that the “sqft_living” seemed to be the most correlated with the first component. After further investigation, it was found that the “sqft_living” variable and the first principle component shared a correlation coefficient of 0.932

-a very high correlation. Therefore, since “sqft_living” has the highest correlation with the first principle component and that the first principle component explains the most amount of variance in the dataset, it can be guessed that “sqft_living” may be the biggest factor in determining a houses’ price . This was confirmed by looking at the correlation between “sqft_living” and “price” which was found to be 0.760 -moderately high. So in later analyses, “sqft_living” was used as the sole predictor variable in the simple linear regression.

6. Analysis, Results, and Interpretation

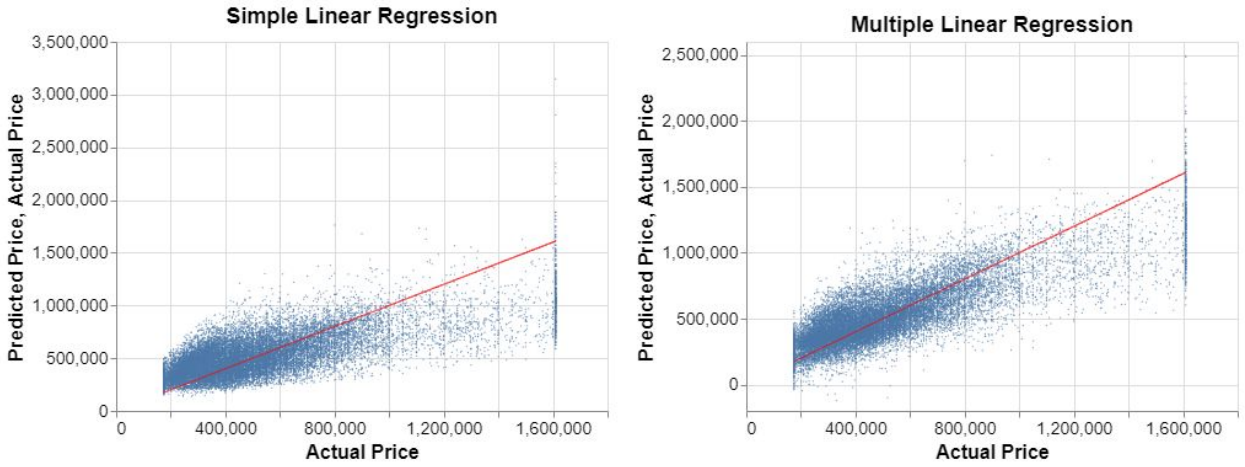


Figure 1.1 and 1.2: Comparing the graphs of the linear and multiple linear regression. All marked points have their predicted price as their y-value and the actual price as their x-value. The two red lines have the actual house price as both their x and y-value.

6.1 Predicting House Price through Linear Regression

Using the linear regression model from Scikit Learn, two linear models were created in an attempt to predict housing prices from the predictor variables. After engineering a new feature, “yr_old”, and removing excess features (“id”, “date”, “yr_built”, “yr_renoved”, ”yr”, “lat”, “long”, and “zipcode”), 15 predictive features remained to be used to be fit to the multiple linear regression model.

Initially, only a simple single linear regression was made based on the “sqft_living” variable (figure 1.1). As explained previously, “sqft_living” is the most appropriate single predictor as it has the highest correlation to “price” and as it is the most correlated to the first principle component, making it the largest source of variance in the dataset. The simple linear model has a slope of 228.50 and an intercept of 52,841.53. Looking at figure 1.1, the ideal situation would have the cluster of points along the red line, but it is clear that the simple model overestimates at lower prices and underestimates at higher prices. The root mean squared error was found to be 209,616. With the mean price of King County houses being 540,000, the model is clearly insufficient for any kind of use. This is to be expected as a property’s value is not simple enough to be predicted solely off of one variable alone.

After completing the single linear regression, a multiple linear regression was conducted taking into account more factors such as the number of bedrooms, waterfront status, square footage of the basement, and more. Upon first inspection, the multiple linear regression (Figure 1.2) looks much more promising than the simple linear regression with the point cluster being more aligned with the slope of the red line indicating the true house prices. But in looking at the error of the model, the multiple linear regression proved to be marginally better than the simple linear regression with a root mean squared error of 174,483.

In hindsight there could have been several explanations for the error in the models. The first being the exclusion of some features provided in the data. In retrospect, latitude and longitude could have an impact on price (although its usefulness in a linear model is questionable). In addition, there are multiple factors that play into a property's price that were not provided in the dataset; features such as having a pool, proximity towards public areas like schools or parks, or kitchen layout can impact a house's price in ways that could not be accounted for. The last source of error in the model could be winsorizing. Winsorizing was initially done to remove extreme values while preserving some of their weight in the data. But there is a real possibility that these expensively priced houses were reflective of house purchasing prices. And by artificially applying a ceiling to prices, the integrity of the data was weakened which could have made the model less accurate.

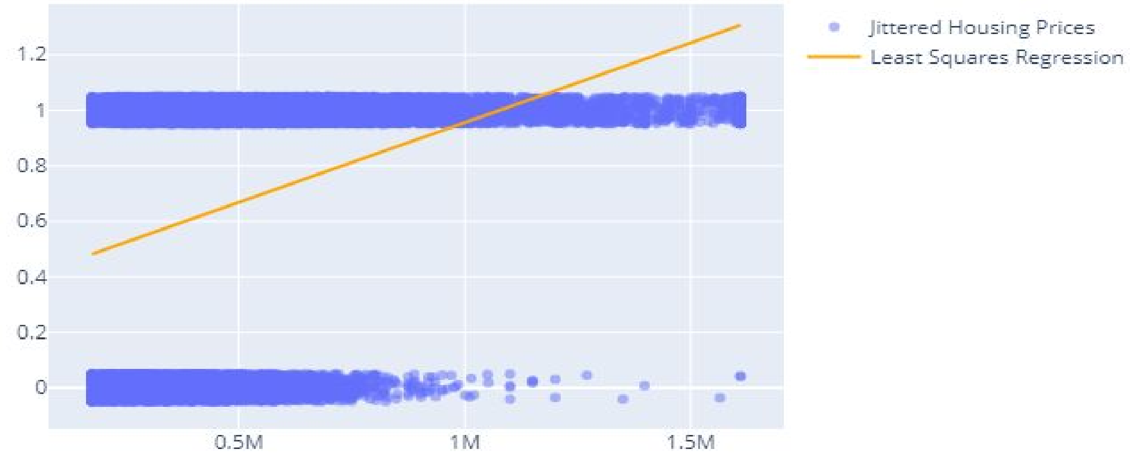


Figure 2: Graph of least squares regression with the x-axis being house price, and the y-axis being their binary grade. The yellow line indicates the probability that a house with the given price was graded as a 1.

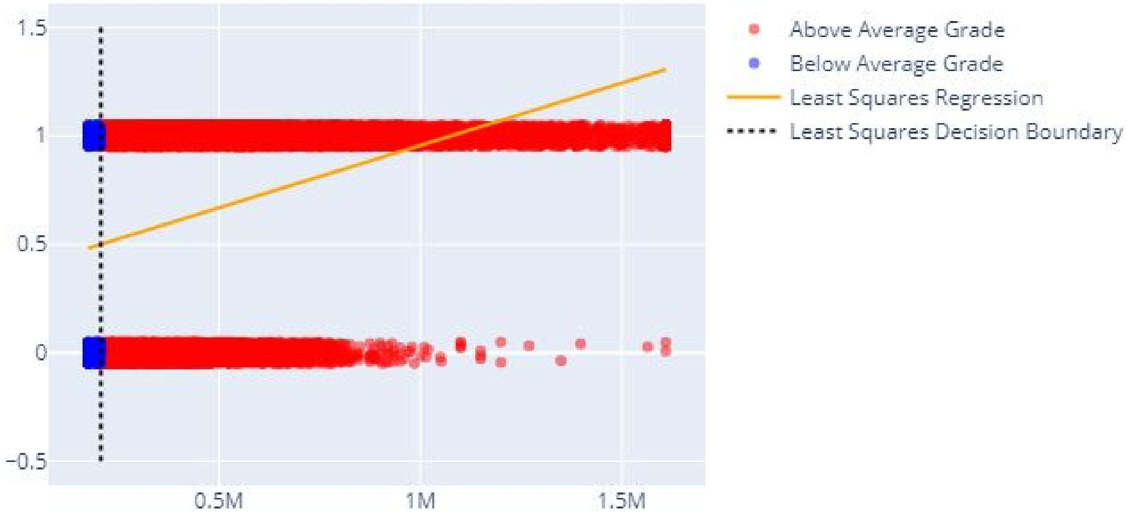


Figure 3: Like figure 2, a graph of least squares regression. Blue points mark house prices where the probability of being a 1 fell below 50%. Red points mark where the probability is above 50%.

6.1 Predicting Binary Grade through Least Squares Regression and Stochastic Regression

When first looking at the dataset it was unknown which categorical variable would be the best predicted variable from price. The correlation between the categorical variables “waterfront”, “view”, “condition”, and “grade” in relation to price was calculated, and price was found to have the highest correlation coefficient with grade. Grade consists of a measurement of a property’s build and design quality rated from 1 to 13. When graphing grade against price (Appendix 3), it was clear that although grade was highly correlated to price, houses with the same grade could vary a good amount in price. Therefore, a new feature called “grade_bin” was created dividing grades into categories: “0” being grades 1-6 that are below average and “1” being grades 8-13 that are above average. Houses with the grade 7 were randomly divided into above and below average.

With a new variable of interest to predict, a least squares model was fit to the price data (figure 2). When predicting binary grade from price, the model had a root mean squared error of 0.433 -too high to be of use. A simple decision rule was applied to the model so that all predicted grade values greater than 0.5 were rounded up to 1, and all grade values less than 0.5 were rounded down to 0 (figure 3). For this new rule, a different type of error would be more appropriate as predicted values could either be entirely correct or entirely incorrect. The zero-one-loss for the model was found to be 0.289. In other words, with the new decision rule, the model would be incorrect 28.9% of the time which is a notable improvement. Lastly, a logistic model was fit to the data and the zero-one-loss was found to be slightly worse at 0.321. Comparing the three versions of models, the simple decision rule least squares regression model was the most accurate at predicting binary grade from house price. It should be noted that randomly splitting all of the houses with a grade of 7 into the two categories introduced randomness into the model that likely caused increased errors. In retrospect, predicting categorical values from predictor variables is most effective when the two are highly correlated.

7. Conclusions and Future Work

In conclusion, it was found that multiple linear regression was slightly more accurate at predicting house prices in comparison to single linear regression. The predictions of the multiple linear regression were fairly inaccurate and should be taken with a good degree of skepticism. Looking back at figure 1.1 and 1.2, it appears that the models often overshoot low prices and undershoot high prices. For future work, it might be interesting to try a logarithmic regression to fit the data. Other possible methods that could be tried include lasso regression or recursive partitioning. For the second question of predicting the categorical variable grade from price, the grade feature was transformed into a binary grade feature. For this feature, the least squares model proved to be most accurate when a simple decision rule was applied. It is difficult to say whether this sort of categorical variable prediction from house price will ever be very accurate as house price is affected by a plethora of factors. The categorical variable in question would require a very strong correlation with price in order to be predicted well.

Bibliography

Satish, G Naga, Ch V Raghavendran, Sugnana Rao, and Ch Srinivasulu. "House Price Prediction Using Machine Learning." *International Journal of Innovative Technology and Exploring Engineering Regular Issue* 8, no. 9 (2019): 717–22. https://doi.org/10.35940/ijitee.i7849.078919.

"House Sales in King County, USA." Kaggle. harlfoxem, August 25, 2016. https://www.kaggle.com/harlfoxem/housesalesprediction.

Alsaqri, Abdallah, Sree Inturi, Pawan Shivhare, Sakshi Singhania, and Karpagam Vinayagam. "Predicting King County House Prices." LinkedIn SlideShare, September 12, 2017. https://www.slideshare.net/PawanShivhare1/predicting-king-county-house-prices.

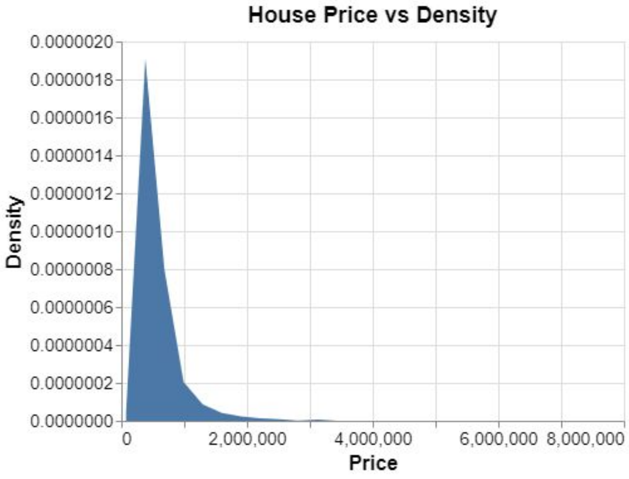
"Health and Housing Data Dashboard." Health and housing data dashboard - King County. King County. Accessed June 10 2020. https://www.kingcounty.gov/depts/health/data/health-housing.aspx.

Franks and Kharitonova. "Lab 10: Binary Classification and Logistic Regression." ds100lsit.ucsb.edu. Accessed June 7, 2020. https://ds100.lsit.ucsb.edu/user/cameronkjoe/notebooks/ds100-s20-content/labs/lab10/lab10.ipynb

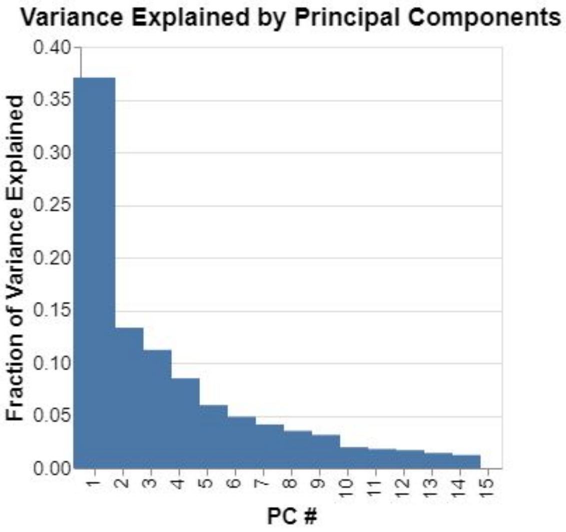
"Open Data." King County Open Data. King County. Accessed June 10, 2020. https://data.kingcounty.gov/.

Appendix

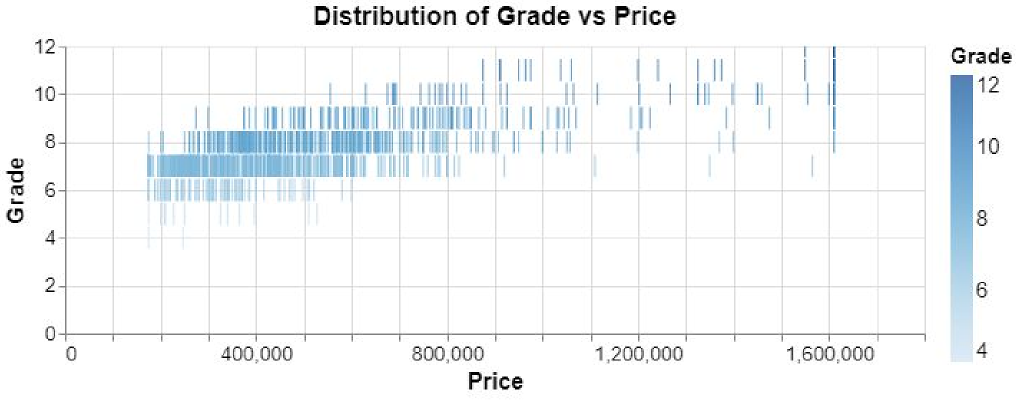
Appendix 1



Appendix 2



Appendix 3 (data was randomly sampled to avoid overplotting)


Distribution of Grade vs Price
Grade
Price
Grade