King County Real Estate Analysis

Group 8: Mariapaola Ambrosone, Claire Gloss, and Dominic Scerbo

May 9, 2022

Abstract

King County, Washington is the thirteenth most populous county in the United States. It is home to Seattle, Washington and borders Elliot Bay. The housing market in this county contains many different types of properties, from rural farms, to suburban homes, to urban apartment complexes. In this study, we looked at the housing market data from this county in order to determine if there exist relationships between the various characteristics of real estate. In our analysis, we found that there exist many relationships between property attributes. Specifically, we found that there is a linear relationship between selling prices of King County homes and their property characteristics. It was also found that waterfront status can be predicted with the knowledge of other key features of the property. Our findings can be utilized by home buyers to understand the true value of the property they are looking to buy and can be used by home sellers and builders to maximize their return on sales in the real estate market of King County, USA.

Executive Summary

This study was conducted in order to explore the relationship between the price of homes in King County, Washington and their location and various property attributes. For this real estate analysis, we used the dataset "House Sales in King County, USA[1]," found on Kaggle, which contained real estate data for over 21,000 homes sold between May 2014 and May 2015 in King County, Washington. The dataset provided numerous characteristics for each home, including but not limited to the number of bedrooms, bathrooms, and stories, the square footage of the property's interior living space and square footage of the property's lot, waterfront status, a qualitative score for the quality of view from the property, and a qualitative score for the condition of the home. In this report, we will discuss the methods and approaches used to analyze and assess the relationships existing between the selling price of homes and their various property characteristics, address our findings statistically, and discuss our conclusions in context to the real estate market.

Data Description

The dataset used for this report consists of 21,613 observations with 21 variables including home prices, locality identifiers such as geographical coordinates, and quantitative measures of the property such as square footage of the living space and the lot. The majority of these variables are numeric, with the exception of a few identifier variables, including Zip code, latitude and longitude coordinates, date sold, waterfront status, and view score. The scale of these variables are aligned to real estate measures. For example, the number of bathrooms is not limited to integers, instead the bathroom count is in multiples of 0.25. This is due to the fact

[1] https://www.kaggle.com/datasets/harlfoxem/housesalesprediction

that in the property management jargon, a bathroom is considered a full bathroom if it has all four components: toilet, sink, shower, and bathtub; each component represents a quarter of the bathroom. Similarly, the number of stories is not limited to integer values in the case of homes with attic spaces. Further details will be discussed for each variable in the next section of this report.

In order to assess the relationships between these variables, we loaded the dataset into RStudio and performed some data transformations. This included class conversions for categorical variables, such as waterfront status, in order to capture these variables as factors. We then checked for any missing values and ensured that our dataset was ready for visualizations and regression analysis.

Variables Overview

Table 1 is provided in order to describe all the variables contained in the King County Real Estate data file that were assessed in the modeling and analysis efforts. The table also highlights the key variables used in each model as well as additional variables that were generated specifically for the modeling taskings (notated by an asterisk*).

Table 1

Model		**Variable**	**Description**
GLM	**LM**		
-	**-**	**id**	**Unique id value for each home sold**
-	**-**	**date**	**Date of the home sale**
-	**-**	**price**	**Price of each home sold**
P	R	***price.log**	**Log transformed price of each home sold (derived from price)**
P	**P**	**bedrooms**	**Number of bedrooms**
P	**P**	**bathrooms**	**Number of bathrooms, where 0.5 accounts for a room with a toilet but no shower**
P	**P**	**sqft_living**	**Square footage of the apartment's interior living space**
P	**P**	**sqft_lot**	**Square footage of the land space**
P	**P**	**floors**	**Number of floors**
R	**P**	**waterfront**	**A dummy variable (binary) for whether the apartment was overlooking the waterfront or not**
P	**P**	**view**	**An index from 0 to 4 of how good the view of the property was**
P	**P**	**condition**	**An index from 1 to 5 on the condition of the home**
P	**P**	**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**

P	**P**	**sqft_above**	**The square footage of the interior housing space that is above ground level**
P	**P**	**sqft_basement**	**The square footage of the interior housing space that is below ground level**
P	**P**	**yr_built**	**The year the house was initially built**
P	**-**	**yr_renovated**	**The year of the house's last renovation**
-	**P**	***renovated**	**Binary for whether a home was remodeled (derived from yr_renovated)**
-	**-**	**zipcode**	**What zip code area the house is in**
-	**P**	***city**	**What city the house is in (derived from zipcode)**
-	**-**	**lat**	**Latitudinal location of the house**
-	**-**	**long**	**Longitudinal location of the house**
P	**P**	**sqft_living15**	**The square footage of interior housing living space for the nearest 15 neighbors**
P	**P**	**sqft_lot15**	**The square footage of the land lots of the nearest 15 neighbors**

P: Candidate Predictor R: Response Variable

Modeling and Analysis

Feature Visualization and Analysis

Outliers

Figure 1 was created to visually assess outliers and/or error in the data related to the number of bedrooms and bathrooms versus the square footage of the homes. As one can see in the chart showing square feet versus number of bedrooms, there is a relatively smaller home that has 33 bedrooms. This is concluded to be an error in the data that will be removed from the data set. Front the square feet versus number of bathrooms chart one can see that there are also homes with 0 bathrooms which for this analysis will be removed as well.

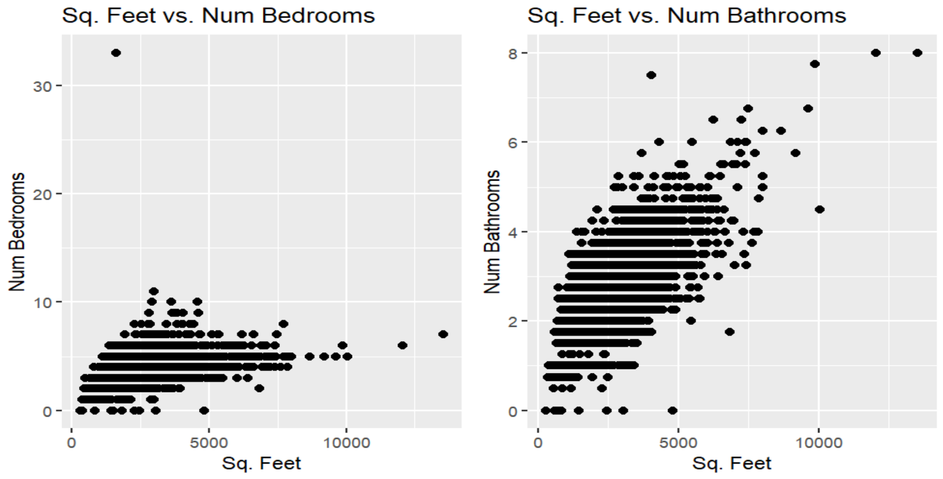


Figure 1

Price vs. Number of Bedrooms

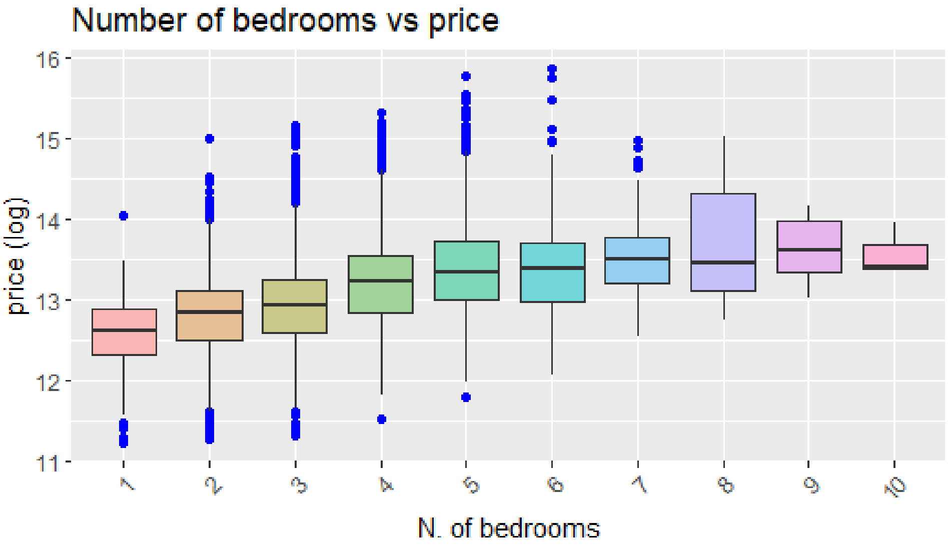


Figure 2

Figure 2 shows the relationship between the number of bedrooms and price of the properties in the dataset. Common belief is that the higher the number of bedrooms, the higher the price of the property. Based on the boxplot, however, we noticed that this may not necessarily be true. Properties with the number of bedrooms between zero and three have similar median prices. The price of the property starts to slightly increase as the number of bedrooms increases over three, where we find similar median prices for the number of bedrooms between four and ten. All classes follow either normal distribution with the exception of the “eight bedrooms” class. For this class we found a higher value for the third quartile compared to the rest of the classes indicating that there may be another variable, for example the square footage of the living space that is contributing to determining the price of a property with the same number of bedrooms. We concluded that the number of bedrooms it’s a factor that, although not drastically, influences the price of the properties.

Price vs. Number of Bathrooms

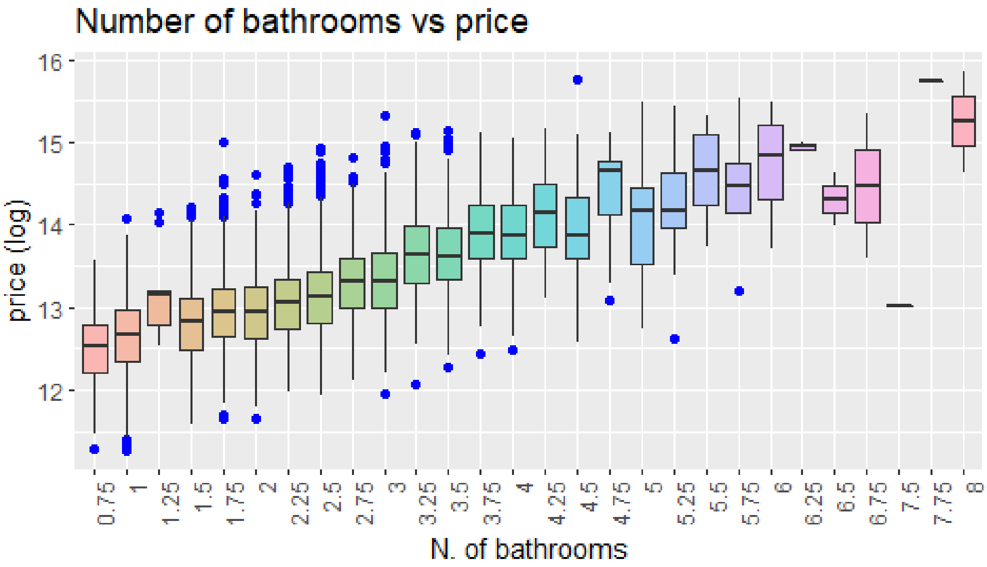


Figure 3

Figure 3 represents the relationship between the number of bathrooms and price of the property. We noticed little variation in the price for properties with the number of bathrooms between one and two and a steady increase in the value of the homes as the number of bathrooms increases.

Price vs. Number of Stories

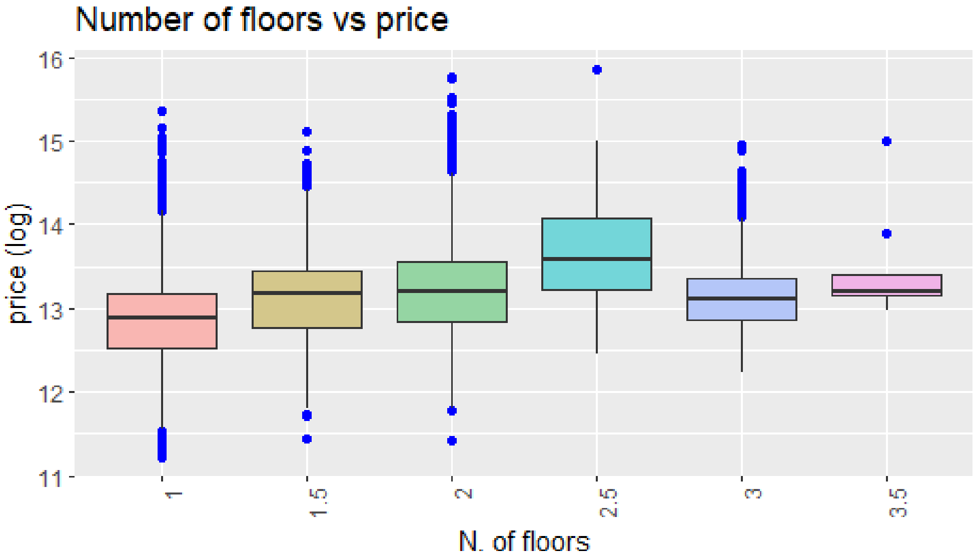


Figure 4

Figure 4 represents the relationship between the number of floors and the price of the property. According to the boxplot, nearly all categories have similar median prices, and there is no immediately apparent trend for those categories. We concluded that this variable is not important in predicting the price. An interesting observation that we made from the boxplot is the slight increase in price for homes with 2.5 floors and a subsequent decrease as the number of floors increases. This is due to the presence of outliers on the “2.5 floors” and “3.5 floors” categories which are pulling the corresponding boxplots towards higher values of price.

Price vs. Grade

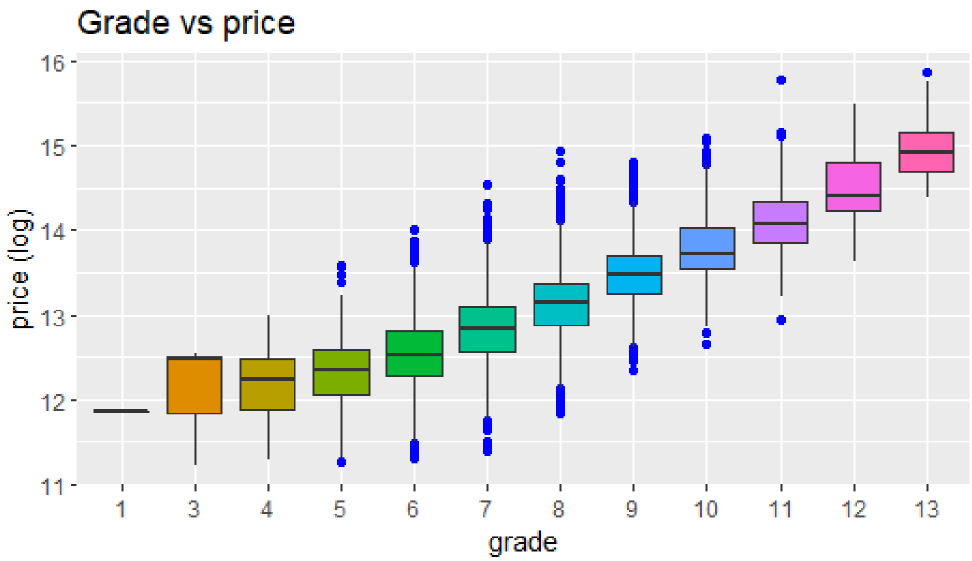


Figure 5

Figure 5 shows the relationship between the grade of the building construction and design, and the price of the properties in the dataset. We noticed a strong relationship between the price and the grade of the house with an increase in price following an increase of the grade value. All classes follow a normal distribution with the exception of the class “3”. For this class we found a higher value for the first quartile compared to the rest of the classes.

Price vs. Living Area and Basement

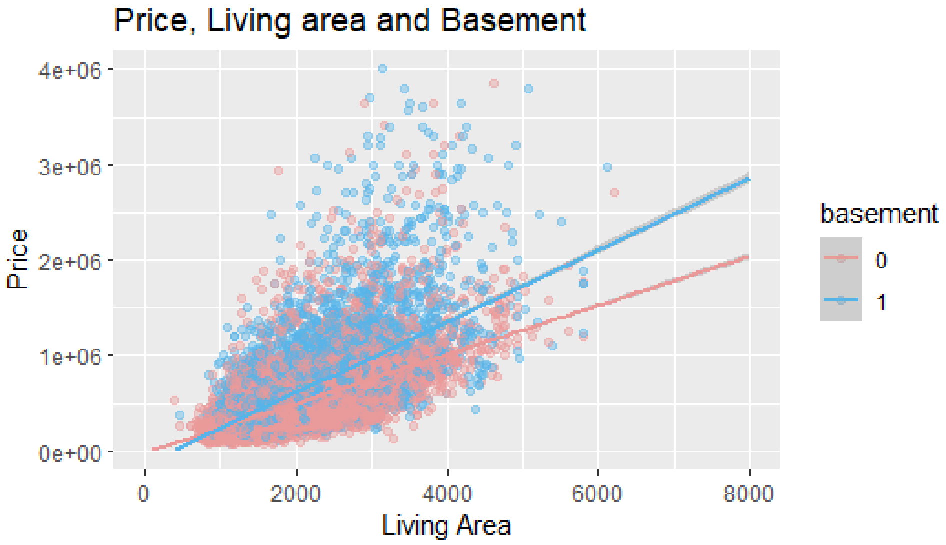


Figure 6

Figure 6 shows the relationship between the square footage of the living area and the presence or absence of a basement in determining the price of the house, where basement value zero indicates no basement and one otherwise. From this visualization we deduced that the value of the house increases when a basement is present and that for homes with similar square footage, the presence of the basement determines an increase in the value of the property.

Price and Waterfront View

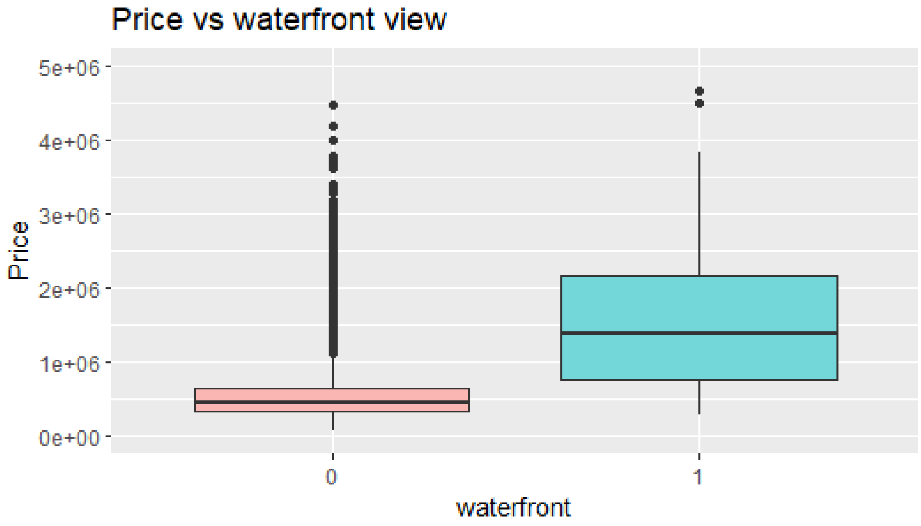


Figure 7

Figure 7 shows the relationship between the access to waterfront and the price of the homes where waterfront takes on the value one for waterfront property and zero otherwise. We noticed an increase in price for those properties that have access to the waterfront compared to the ones that don’t.

Price and View

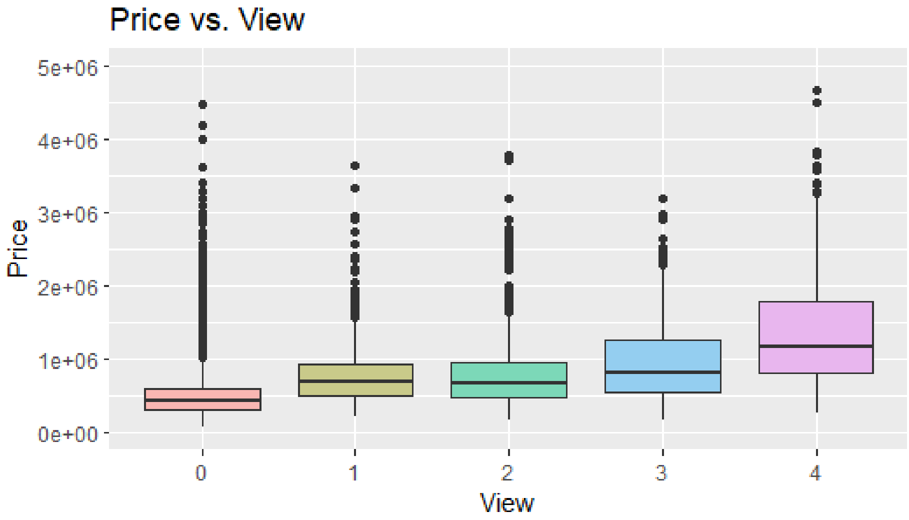


Figure 8

Figure 8 shows the relationship between the view of the house and the price of the house. We observed that categories 1 and 2 have similar median prices and an increase in the value of the property for view categories 3 and 4.

Price and Geographical Location

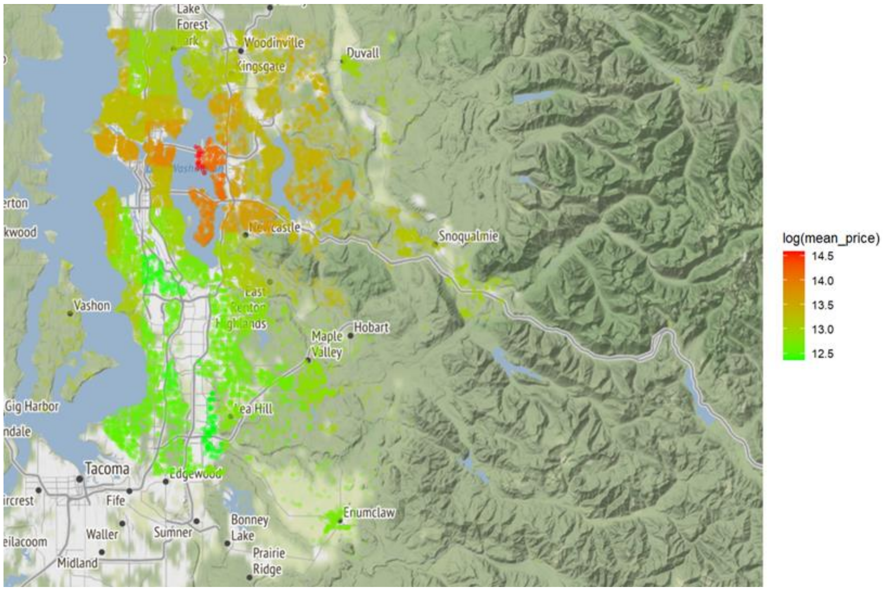


Figure 9

Figure 9 shows the relationship between the geographic location of the house and the relative price. We observed that houses with a lower price are located in the southern part of the map and the majority of these properties are not directly located by the water. On the other hand, we observed that higher prices are located in the middle part of the study area with the highest prices observed right by the water. We concluded that the spatial location of the house is a factor that directly affects its price.

Linear Regression Model

To further explore the relationships between the variables, we decided to explore the potential of creating a linear regression model that would allow us to predict the price value of homes. Data cleaning methods that were used in preparation for the linear regression analysis are as follows. Unnecessary variables were removed from the data set: "id" was removed because we did not need the unique identifier for each home, "lat" and "long" were removed because it essentially provided the same information as "zipcode." "Zipcode" was also removed because we were able to map zipcodes to city names. In using city names, we were able to condense the locality variable into cities, providing a more reasonable identifier variable, "city," to use in the regression model. The zipcode to city mapping was done using data from zip-codes.com.[2] After removing and replacing these variables, we were left with 18 variables to work with in the analysis. Next, we reclassified variables "date," "view," "waterfront," and "city" as factor variables since these variables play no numeric role in a regression analysis. We also decided to change "yr_renovated" into a binary variable, taking the value 1 if the home had been remodeled and 0 if the home had not been remodeled. This variable was also reclassified as a factor variable. We had no missing values to remove from the dataset, but we did remove observations that had "bedrooms" equal to zero. In removing these observations, we are able to analyze the prices of homes and not lots. We also noticed that many homes in this dataset, following suit with what is expected in the home pricing market, were priced on the low end of the spectrum. A log-transformation was applied to "price" in order to control for this skew. We then identified outliers, Figure 10, as being equal to and above log-price values of 14.4198 and equal to and below 11.6262. Observations that were in these ranges were removed from the dataset. Our final data set had 21,266 observations and 17 potential predictor variables to use in our regression model.

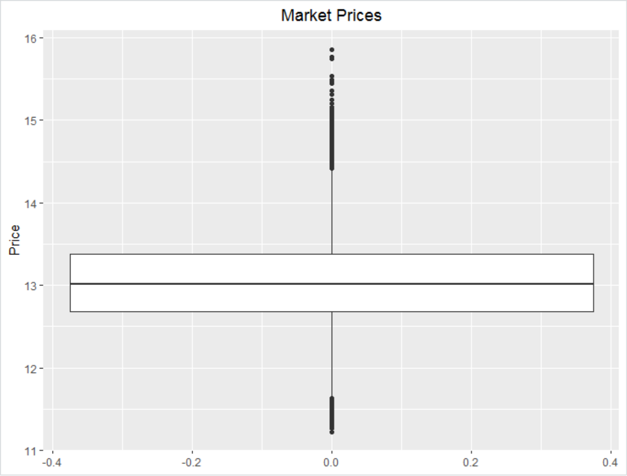
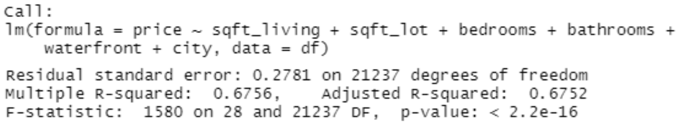


Figure 10

[2] https://www.zip-codes.com/county/wa-king.asp

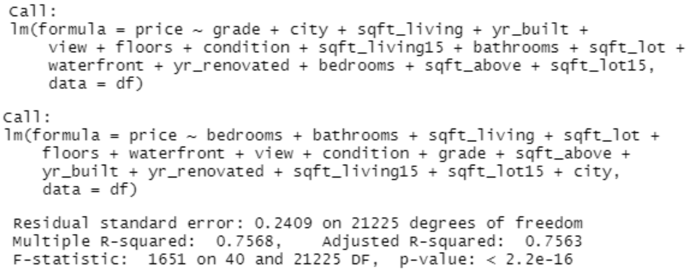


```
Call:
lm(formula = price ~ sqft_living + sqft_lot + bedrooms + bathrooms +
    waterfront + city, data = df)

Residual standard error: 0.2781 on 21237 degrees of freedom
Multiple R-squared:  0.6756,	Adjusted R-squared:  0.6752
F-statistic:  1580 on 28 and 21237 DF,  p-value: < 2.2e-16
```

Figure 11

The first multiple linear regression model that we created to predict the price value of homes used the six most important factors of home buying, “sqft_living,” “bedrooms,” “bathrooms,” “sqft_lot,” “waterfront,” and “city.” R output from this model, Figure 11, shows that the model is significant with F-score = 1580, p<.001, adj. R^2 = 0.6752. This model was used as the base model to compare future models to for the remainder of our analysis.



```
Call:
lm(formula = price ~ grade + city + sqft_living + yr_built +
    view + floors + condition + sqft_living15 + bathrooms + sqft_lot +
    waterfront + yr_renovated + bedrooms + sqft_above + sqft_lot15,
    data = df)

Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + sqft_lot +
    floors + waterfront + view + condition + grade + sqft_above +
    yr_built + yr_renovated + sqft_living15 + sqft_lot15 + city,
    data = df)

Residual standard error: 0.2409 on 21225 degrees of freedom
Multiple R-squared:  0.7568,	Adjusted R-squared:  0.7563
F-statistic:  1651 on 40 and 21225 DF,  p-value: < 2.2e-16
```

Figure 12

We then used a stepwise regression process to create multiple linear regression models. Forward and backward methods were used to provide two regression equations to use. Since these models were both significant, F-score = 1651, p<.001, adj. R^2 = 0.7563 for Model 1 and Model 2 (Figure 12), we decided to use only model 2 to compare for a better fit versus the first model that we created. Since the adjusted R^2 measure is closer to 1 in Model 2 (the model created from the backward stepwise process) than in the original model with only 6 predictors, we concluded that Model 2 was a better fit and provided a more accurate prediction for price value of homes. Thus, we continued our regression analysis with only Model 2, predicting price from the number of bedrooms, bathrooms, sqft living area, sqft lot area, number of stories, waterfront status, view score, condition score, grade score, sqft area above ground level, year built, renovation status, sqft living area average of the community, and the city that the home is located.

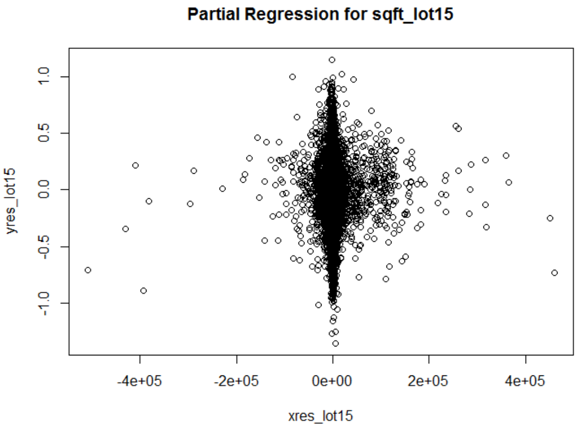
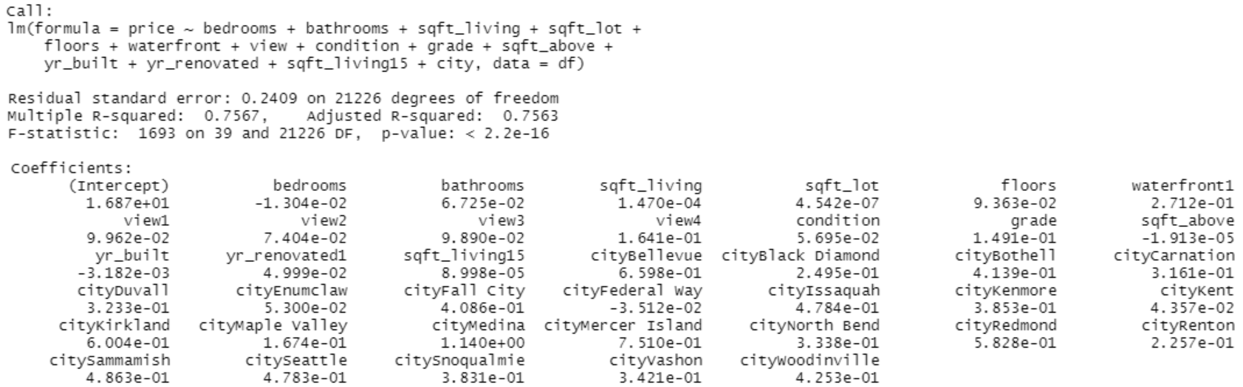


Figure 13

In Model 2, it was noticed that the only variable that was not significant in the presence of the other predictor variables was sqft_lot15, so we performed a partial regression analysis on this variable to test if it truly belonged in the model. Figure 13 shows that there is no significant linear relationship, so we concluded that this variable was not contributing significantly to our model. We then removed “sqft_lot15” from Model 2 and reran the summary results from this model equation. This equation had a slightly improved significance, F-score = 1693, p<.001, adj. R^2 = 0.7563. Since this version of the model was improved, we adopted this model equation as the final regression model equation to use in our analysis. Summary output and regression coefficients are found in Figure 14.



```
Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + sqft_lot +
    floors + waterfront + view + condition + grade + sqft_above +
    yr_built + yr_renovated + sqft_living15 + city, data = df)

Residual standard error: 0.2409 on 21226 degrees of freedom
Multiple R-squared:  0.7567,	Adjusted R-squared:  0.7563
F-statistic:  1693 on 39 and 21226 DF,  p-value: < 2.2e-16

Coefficients:
      (Intercept)          bedrooms         bathrooms       sqft_living          sqft_lot            floors       waterfront1
        1.687e+01        -1.304e-02         6.725e-02         1.470e-04         4.542e-07         9.363e-02         2.712e-01
            view1             view2             view3             view4         condition             grade        sqft_above
        9.962e-02         7.404e-02         9.890e-02         1.641e-01         5.695e-02         1.491e-01        -1.913e-05
         yr_built     yr_renovated1     sqft_living15      cityBellevue cityBlack Diamond      cityBothell      cityCarnation
       -3.182e-03         4.999e-02         8.998e-05         6.598e-01         2.495e-01         4.139e-01         3.161e-01
       cityDuvall      cityEnumclaw     cityFall City   cityFederal Way      cityIssaquah      cityKenmore          cityKent
        3.233e-01         5.300e-02         4.086e-01        -3.512e-02         4.784e-01         3.853e-01         4.357e-02
     cityKirkland cityMaple Valley        cityMedina cityMercer Island    cityNorth Bend      cityRedmond        cityRenton
        6.004e-01         1.674e-01         1.140e+00         7.510e-01         3.338e-01         5.828e-01         2.257e-01
    citySammamish       citySeattle    citySnoqualmie        cityVashon   cityWoodinville
        4.863e-01         4.783e-01         3.831e-01         3.421e-01         4.253e-01
```

Figure 14

We then checked that this model met the linearity assumptions by checking the residuals, acf, and qqnorm plots, seen in Figure 15, Figure 16, and Figure respectively. Each of these showed that the model had no significant violations of the linearity assumptions and so we concluded that our model was appropriate to use in a linear regression analysis of home price values.

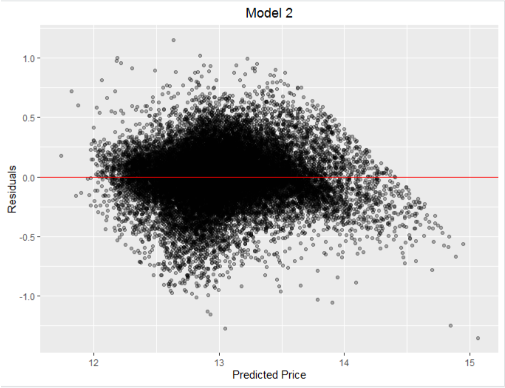


Figure 15

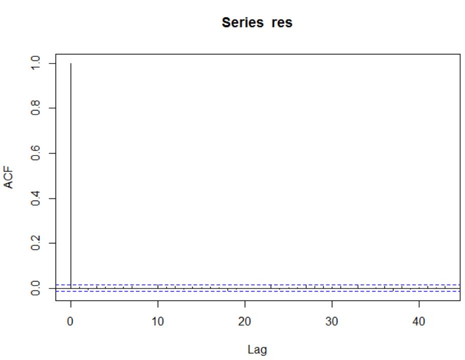


Figure 166

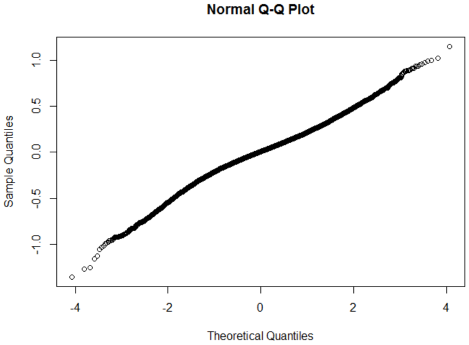


Figure 17

We then checked for outliers using Bonferroni's method and found five outlying observations. The first outlier, observation no. 3957 is a Seattle home with 7 bedrooms, nearly 7 bathrooms, 7K sqft living area, and a 41K sqft lot. This type of property seems to be very rare in the city of Seattle. It is interesting that this type of property is not on a small lot, like many other properties in the city of Seattle with many bedrooms. Perhaps this property is a mansion instead of a typical home listed on the market in King County. The second outlier, observation no. 7198, is a Seattle home with only 2 bed, 1 bath, 1,290 sqft living area, and 5K sqft lot. This house is listed at a very high price, log-price 13.7850, which seems to be oddly high for the type of home. Perhaps there is a locality bubble in Seattle that has extremely high prices compared to the rest of the market in King County. The third outlier, observation no. 12535, is a Seattle home with 3 bed, 1 bath, 1,100 sqft living area, and 5K sqft lot The log-price for this home is 11.7752, which seems consistent with the sqft living area, but inconsistent with the number of bedrooms. Similarly, the fourth outlier, observation no. 18047, is a Seattle home with 3 bed, 1 bath, 1,200 sqft living area, and 7K sqft lot. The log-price of this home is 11.7752, which seems consistent with the sqft living area, but not with the number of bedrooms in our model. The fifth and final outlier, observation no. 20716, is a Bellevue home with 5 bed, 6 bath, 7,120 sqft living area, 40K sqft lot, log-priced at 13.7101. This home also has a view score of 4, and grade score of 12. These scores are very high for a home priced in this range in Bellevue.

In order to test the accuracy of our regression equation, we looked to two homes in King County, Washington that are currently on the market. These listings were found at Zillow.com[3]. In order to plug these homes into our regression equation the following assumptions were made, due to missing information on Zillow.com: “sqft_living” = “sqft_living15” = “sqft_above”; view = 1 for apartment, 2 for small yard, 3 for large yard, 4 for waterfront; condition = 4 (good), grade = 10 (high end of average). The first home that we looked at is a two-story 1915 renovated Vashon home with 3 bed, 1.5 bath, 1,950 sqft living area, 43,124 sqft lot. The listed price of this home is $915K or log-price value of 13.7266. The 95% prediction interval for this home is [13.2098, 14.1854] with fitted value 13.6976. Thus, our model accurately predicts the price of this home. The second property that we looked at was a very interesting listing, and so we were very curious if the model would perform as expected. This home is a single story 2003 Seattle based houseboat with 4 bed, 2 bath, 683 sqft living area, 897.34 sqft lot. The listed price of this houseboat is $490K or log-price value of 13.1021. The 95% prediction interval for this home is [13.0684, 14.0425] with fitted value 13.5555. Thus, our model also accurately predicted the price of the houseboat. Therefore, we concluded that our model was a good fit in predicting the selling prices of homes in King County, Washington.

In our linear regression analysis, we were able to find that log-transformed price is linearly related with several predictor variables: bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, sqft_above, yr_built, yr_renovated, sqft_living15, city. We found that this equation was a better fit than the model based on only six predictors: sqft_living, sqft_lot, bedrooms, bathrooms, waterfront, city. The better model, with more predictors, proved to hold true to linearity assumptions and was accurate in predicting selling prices for homes in King County, Washington that are currently listed for sale on the market. Practical implications of our analysis include utilizing this model to estimate the selling price of one’s home. This can also be a way to identify which factors of the property can be improved in order to maximize selling price. One can also utilize this equation to estimate the buying price of a home and identify potential high-ball prices that could be reduced to a more reasonable market price.

After running this analysis, we believe that additional information would be interesting to include in the dataset in order to better estimate the selling prices of homes. We would like to see future data include the following: time of year that the house was put on the market, duration of home on the market, number of times the home has been sold, school districts, commuting and walking scores for the community, and average income of the residents in the neighborhood. These are all factors that play a role in the valuation and final selling price of a home and could potentially have significant influence on the prediction model that we created in our analysis.

[3] https://www.zillow.com/homedetails/8929-SW-184th-St-Vashon-WA-98070/48852331_zpid/
https://www.zillow.com/homedetails/1700-Westlake-Ave-N-40-Seattle-WA-98109/2064232670_zpid/

Logistic Regression Model

The logistic regression model serves to predict if a property has a waterfront view by on key features. We found that this model could have value in the appraisal market where there is a significant increase in the appraised home value based on the type of property. Therefore, it can be important for a new home builder to determine if their prospect home is more representative of a waterfront or non-waterfront home in the area. Based on this knowledge they can determine if this aligns with their expected comparable and reassess their plans and budget as needed.

To develop the logistic regression model the following approach was applied. We first started by visually assessing the relationships between variables excluding ones that proved to intuitively be poor predictors of waterfront view or appeared to provide no added benefit in a regression setting based on the nature of the variable. We then split the data into a training dataset and a test dataset in order to assess the performance of the model on a dataset that was intentionally held-out and not seen prior. Automated regression methods were applied to derive best subsets of regression models which could be used as a proxy to jump-start the model building process. The preliminary models were evaluated accordingly and used to begin a our testing and evaluated process. During the testing and evaluation we discovered insignificant predictors to drop from the model as well as predictors that directly influenced the response variable. As modifications were applied to the model key metrics were then analyzed in order to validate the performance to ensure appropriate outcomes were produced. In order to properly assess the performance of the model the test, evaluation, and validation procedures were applied in an iterative fashion.

Prior to building the model data cleaning procedures (i.e., the removal of the outliers found in the feature analysis and the log transformation of price) were applied. We were able to intuitively drop several parameters from the list of feasible predictors based on intuition and the scope of the analysis. These features included id, date as well as the geospatial properties ZIP code, latitude, and longitude. Additional investigations of the relationships between the variables showed that several variables indicated a clear distinction between the waterfront and non-waterfront properties. As shown by Figure 22, Figure 20, Figure 19, Figure 21, Figure 18, and Figure 17 illustrate the log transformed price of the home, the square footage of the living space of the home as well the 15 nearest neighboring homes, the quality of the view, the year it was built, and the grade of the property respectfully which all depicted differentiable attributes associated with waterfront homes. Under normal circumstances and based on our prior assessment of the feature this relationship shoed validity. In the United States, the cost of high demand and low supply waterfront homes is much higher than inland homes. Homeowners also will typically rate the quality of the view for homes with more scenic views such as coastal, waterfront, mountains, hills, and agricultural surroundings which, based on demand, come at a higher cost. We also discovered that the waterfront homes are on average slightly larger than non-waterfront homes. It seems intuitive that waterfront homes were typically built during earlier years than non-waterfront homes due to the availability and its decrease over time. The grade of the property was a less evident feature of the dataset that served as a possible predictor of waterfront homes.

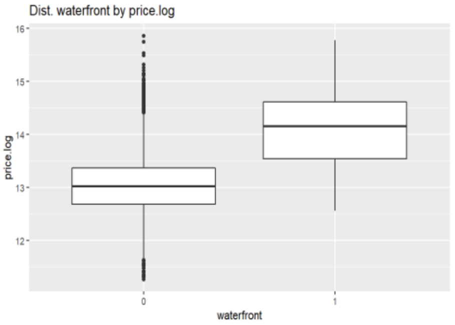


Figure 22

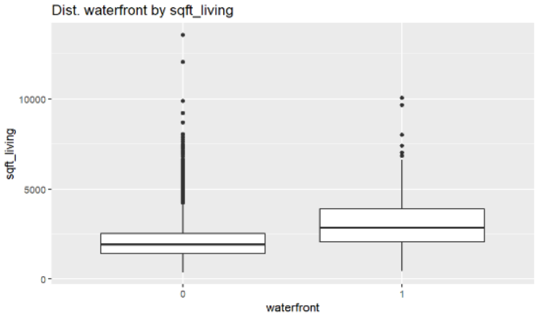


Figure 20

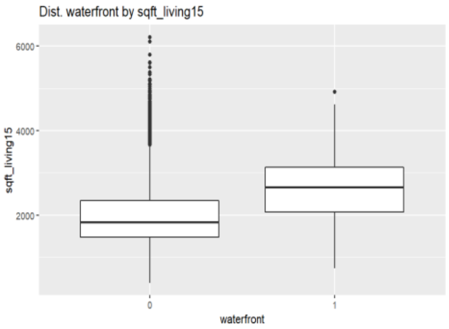


Figure 19

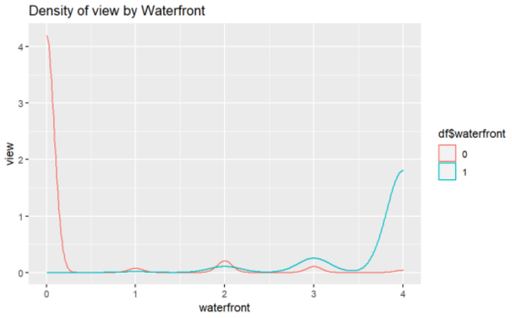


Figure 21

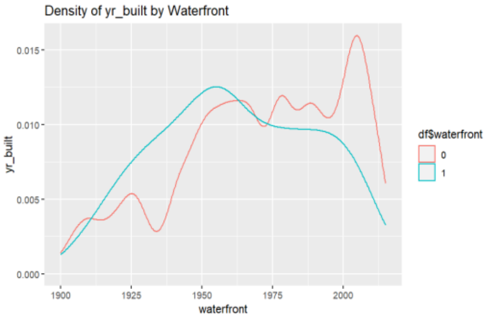


Figure 18

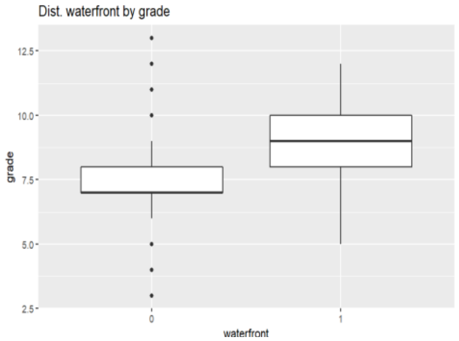


Figure 17

As our first initial test we investigated the relationship between the waterfront homes and the log transformed price. Generating a model between price and waterfront we saw that price had a high statistical significance with a p-value of virtually 0. This led us to generate a sequence of new data ranging from the minimum to the maximum value of home price which we then run through the model. The resulting probabilities were plotted (as seen in Figure 23) which showed the positive relationship between the probability of waterfront being 1 and the increase in price. We found this model to be more descriptive in a general sense since the truth data does not exactly depict a clear break and separation between the two classes, therefore, we will be required to incorporate additional features in order to generate a more prescriptive model.

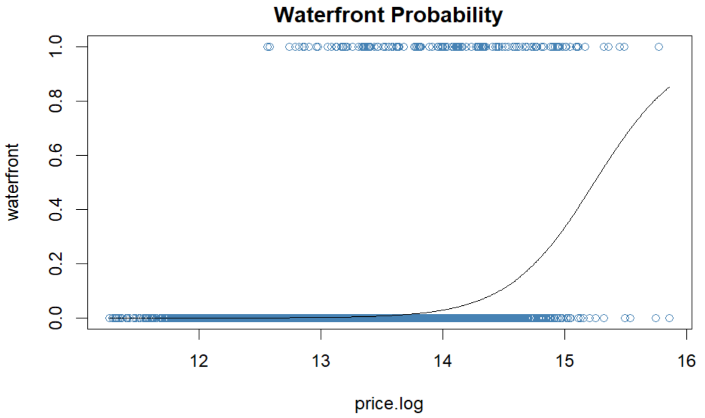
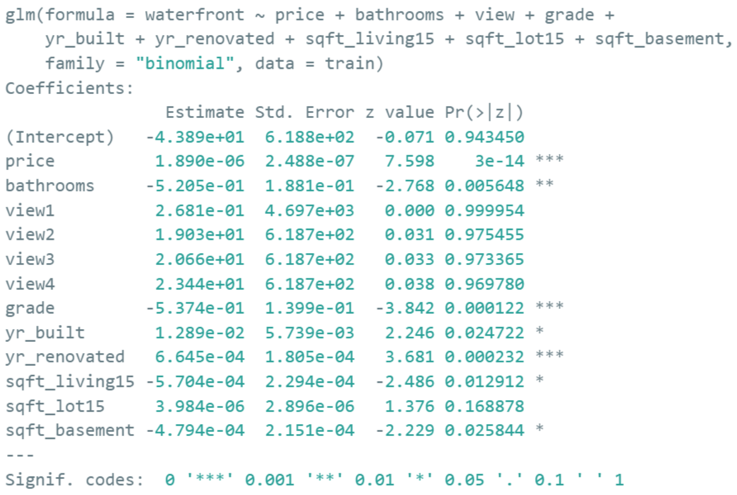


Figure 23

After obtaining some preliminary knowledge about the variables relationships we split the data into train and test sets by taking a sample size of 80% of the data without replacement for the train set and the remaining 20% as the test set. Note that we set the seed to 1 for this effort. We began the model building process using the regsubset method from the leaps package. Using this approach we were able to identify best-fitting models as predictors are added in the multiple logistic regression model. To evaluate the results of the regsubset methods we looked at the adjusted R-Squared, Mallows's Cp, and Bayesian information criterion statistics which all converged on the price, number of bedrooms, the view, grade, the year the property was built and renovated, square footage of the lot, square footage of the 15 closest neighboring homes, and the square footage of the basement being applicable parameters in the model (GLM1 - Figure 24).

GLM1 contains several key parameters that are in direct alignment with features that we discovered in our prior analyses. However, there were the absence of some variables and lower statistical significance that were less evidence initially. Firstly, the view predictor did not hold a significance level statistical significance. Secondly, the regsubsets method determined the year the property was renovated was more significant than the year the property was built.



```
glm(formula = waterfront ~ price + bathrooms + view + grade +
    yr_built + yr_renovated + sqft_living15 + sqft_lot15 + sqft_basement,
    family = "binomial", data = train)
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.389e+01  6.188e+02  -0.071 0.943450
price          1.890e-06  2.488e-07   7.598    3e-14 ***
bathrooms     -5.205e-01  1.881e-01  -2.768 0.005648 **
view1          2.681e-01  4.697e+03   0.000 0.999954
view2          1.903e+01  6.187e+02   0.031 0.975455
view3          2.066e+01  6.187e+02   0.033 0.973365
view4          2.344e+01  6.187e+02   0.038 0.969780
grade         -5.374e-01  1.399e-01  -3.842 0.000122 ***
yr_built       1.289e-02  5.739e-03   2.246 0.024722 *
yr_renovated   6.645e-04  1.805e-04   3.681 0.000232 ***
sqft_living15 -5.704e-04  2.294e-04  -2.486 0.012912 *
sqft_lot15     3.984e-06  2.896e-06   1.376 0.168878
sqft_basement -4.794e-04  2.151e-04  -2.229 0.025844 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 24

Prior to immediately dropping any insignificant predictors from the model. The ROC curve of the model was generated. Figure 25 shows that this model presented a near perfect fit. This level of performance was surprising based on the insignificance of several features in the model; however, we determined that this could be explained contextually by evaluating the distribution of data for view. As we noted previously there is a clear separation between the waterfront and non-waterfront classes in terms of view.

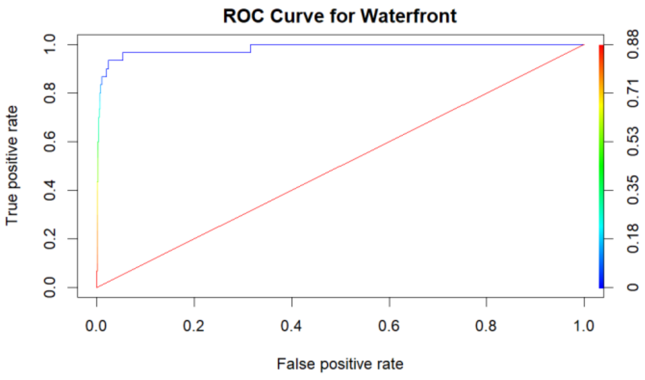


Figure 25

Figure 27 shows a box plot that in conjunction with Table 2 shows that higher rated views are disproportionately separated between the two property types which may be serving as a proxy value or creating inherent bias in the model to categorize the home more easily in the presence of the other variables. In a practical sense it might not be appropriate to include this parameter since it is highly representative of our response variable.

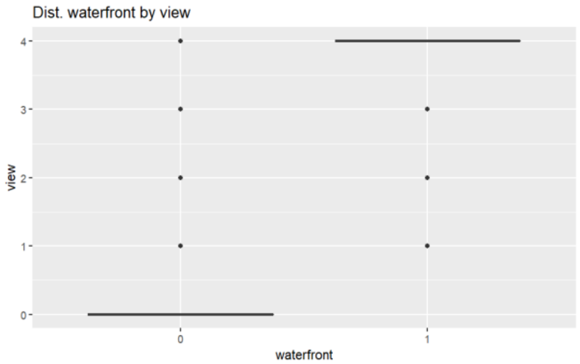
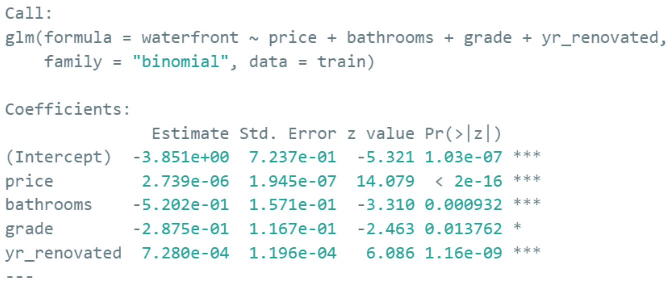


Figure 27

Table 2	0	1	2	3	4
non-waterfront	19470	331	953	491	184
waterfront	0	1	8	19	135

The other parameters that we hypothesized could be dropped from the model are the square footage of the 15 closest neighboring lots and living space and the square footage of the basement. We then carried out an appropriate hypothesis test to see if the coefficients for view, the square footage of the 15 closest neighboring lots and living space and the square footage of the basement can be dropped. Using the output from the reduced model (GLM2 - Figure 26) we formed the following hypothesis test.



```
Call:
glm(formula = waterfront ~ price + bathrooms + grade + yr_renovated,
    family = "binomial", data = train)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.851e+00  7.237e-01  -5.321 1.03e-07 ***
price         2.739e-06  1.945e-07  14.079  < 2e-16 ***
bathrooms    -5.202e-01  1.571e-01  -3.310 0.000932 ***
grade        -2.875e-01  1.167e-01  -2.463 0.013762 *
yr_renovated  7.280e-04  1.196e-04   6.086 1.16e-09 ***
---
```

Figure 26

$$H_0 : \beta 3 = \beta 4 = \beta 5 = \beta 6 = \beta 8 = \beta 10 = \beta 11 = \beta 12 = 0.$$
$$H_a : at\ least\ one\ \beta\ is\ not\ zero.$$

Using the chi-square test the test statistic is

$$\Delta G^2 \quad = Residual\ deviance\ of\ reduced\ model - Residual\ deviance\ of\ full\ model$$
$$= 1116.36 - 453.1897 = 663.1707$$

The corresponding p-value is $P\left(663.1707 > {X_2}^2\right) = 0$ using 1-pchisq(663.1707, 5). Since the p-value is 0 we reject the null hypothesis. The data does support the claim that the reduced model is useful, compared to the full model.

To further validate the performance of the model we plotted the ROC curve once again. Interestingly we found that the ratio of the true positive rate and false positive rate had decreased. As we called earlier there was suspicion for a bias in the model based on the view predictor therefore, we generated a reduced model, but only dropping view to analyze the decrease in performance in isolation. As seen in Figure 28 and Figure 29 dropping view had just as big of an impact on decreasing the performance of the model as dropping the other predictors which led to the confirmation of our theory that view created a bias in the model.

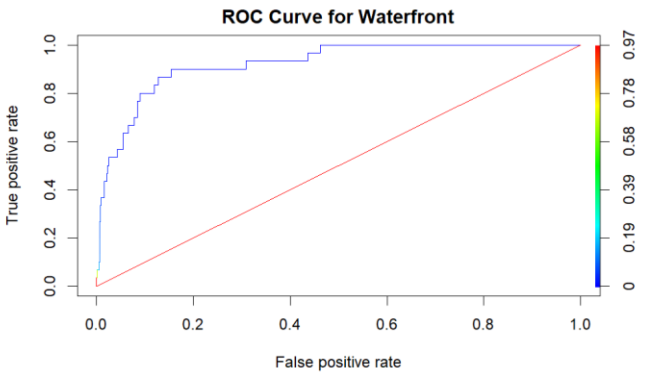


Figure 28

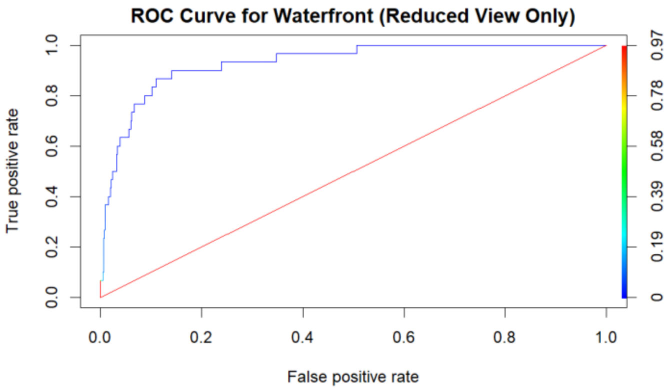
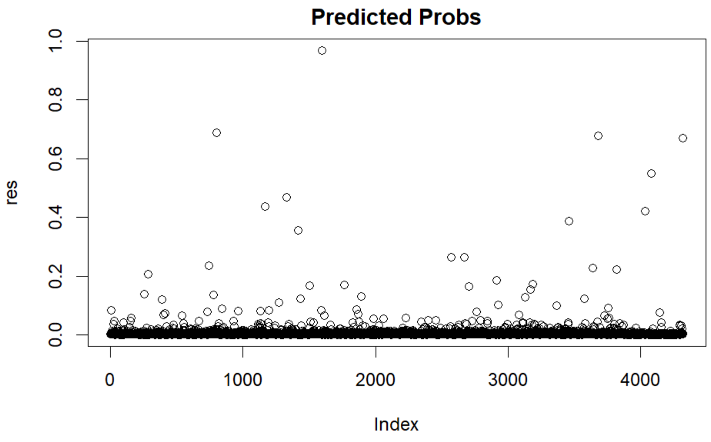
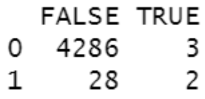


Figure 29

Despite dropping the predictor view we still achieved an AUC value of 0.9226. When constructing the final confusion matrix at a threshold level of 0.80 we achieved an overall error rate of 0.0069; however, visualizing the probabilities of the resulting predictions as seen in Figure 30 most of the probabilities are below 0.1. By reducing the threshold to 0.5 we were able to calculate a final accuracy of 0.9931, a True Positive Rate of 0.9333, and a False Positive Rate of 7e-04.





```
   FALSE TRUE
 0  4286    3
 1    28    2
```

Figure 30

In conclusion we found that using a logistic regression model we were able to predict if a property was waterfront based on a limited number of attributes about the home and performed better than random guessing; however, in context it seems that this model could benefit from further exploiting other variations of the data to provide a more robust utility. In the context of the logistic regression model geospatial properties were not assessed; however, some possible future extensions to the model could include the transformation of geospatial parameters to provide a proxy for key homeowner factors. Those factors could include the population density of the resident or neighboring areas, the distance between the residence and key cities, attractions, or amenities. Despite the challenges with the model, we were able to exploit practice test and evaluation methods. With this model we were able to assess real world challenges in the data that led to high performance but an impractical predictive capability. By contextually evaluating the relationship between variables we were able to resolve said challenges while minimizing the overall loss over the predictive ability of the model.

Conclusion

From our analysis, we were able to see that relationships exist between many property characteristics and how they contribute to determining the price of the properties. These include the influence of the number of bathrooms and bedrooms on the price of the house, the number of stories and the presence or absence of a basement. We considered the role of waterfront view over the price as well as the overall view of the house in influencing the value of the property. As we might have expected, variables related to the size of the homes seems to be responsible for price fluctuations amongst properties. In particular, the square footage of the living space and the presence of the basement, the grade, number of bathrooms and bedrooms, and the location of the homes appear to be good predictors of a house's price in King County. This preliminary exploratory data analysis, allowed us to identify variables of interest and the relationship amongst and between variables was further explored in linear and logistical regression models.

We also found that a linear relationship exists between the log-transformed price value of homes the number of bedrooms, bathrooms, sqft living area, sqft lot area, number of stories, waterfront status, view score, condition score, grade score, sqft area above ground level, year built, renovation status, sqft living area average of the community, and the city that the home is located. This model was found to be significant, F-score = 1693, p<.001, adj. R^2 = 0.7563, and did not have any significant violations to the linearity assumptions. Several interesting outliers were identified, showing that there exist some homes with unusually high or low pricing for the given characteristics. We were able to apply two currently listed homes in King County to our model equation, and found that the currently listed prices fall within the 95% prediction intervals. Thus, we concluded that our model accurately predicts the price of homes in King County. This can help buyers and sellers alike in identifying the best price for their properties of interest.

After assessing the practical implications of adding various predictors to the logistic regression model we were able to classify waterfront and non-waterfront homes in a manner better than random guessing with an overall accuracy of 0.9931 and an AUC 0.9226 while using just the price, number of bathrooms, grade of the property, and the year it was rennovated.

For further analysis, we would like to see additional housing market information included in the dataset. Characteristics that we would like to explore include the time of year that the house was put on the market, duration of home on the market, number of times the home has been sold, school districts, commuting and walking scores for the community, and average income of the residents in the neighborhood. We believe these factors play a role in the valuation of a home and potentially influence the prediction of home prices in King County.