King County Housing

2020-12-07

MSIS 510 Purple Team 3

Astha Grover
Dajia Bao
Elizabeth Evans
Freda Li
Ian Kinder-Pyle
Nidhi Pancholi

Objective

The Greater Seattle area has become a real estate hot spot in the past few years. Despite the pandemic, given the historical low interest environment, the median sale price has increased 10.5% year over year and the volume of sales has increased 19.7% year over year.

For an average buyer, it would be very easy for him to enter an endless bidding war. Our goal of this project is to identify the factors based on the historical King County housing data, leading to higher property sales price for our clients. Moreover, our next goal is to establish a model to predict the housing price so that we could use it to compare with the listing price of interesting properties.

Data Exploration

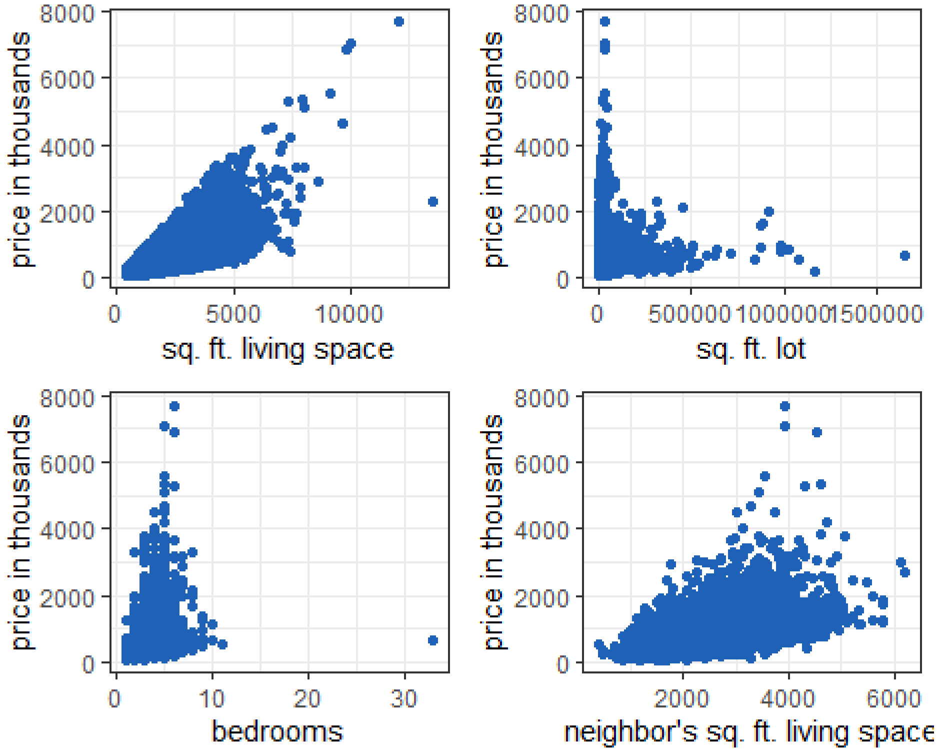
The data source used in this project was found via Kaggle and contains house sales prices for King County, which includes the City of Seattle as well as other municipalities with well-known employers in the Pacific Northwest such as Redmond. The data set includes homes that were sold between 2014 and 2015, with 21 variables and 21612 observations. Below is a summary of this data.

Variable	Type	Description
id	num	Unique ID for each home sold
date	date	Date of the home sale
price	num	Price of each home sold
bedrooms	int	Number of bedrooms
bathrooms	num	Number of bathrooms, where .5 accounts for a room with a toilet but no shower
sqft_living	int	Square footage of the house interior living space
sqft_lot	int	Square footage of the land space
floors	num	Number of floors
waterfront	int	A dummy variable for whether the house was overlooking the waterfront or not

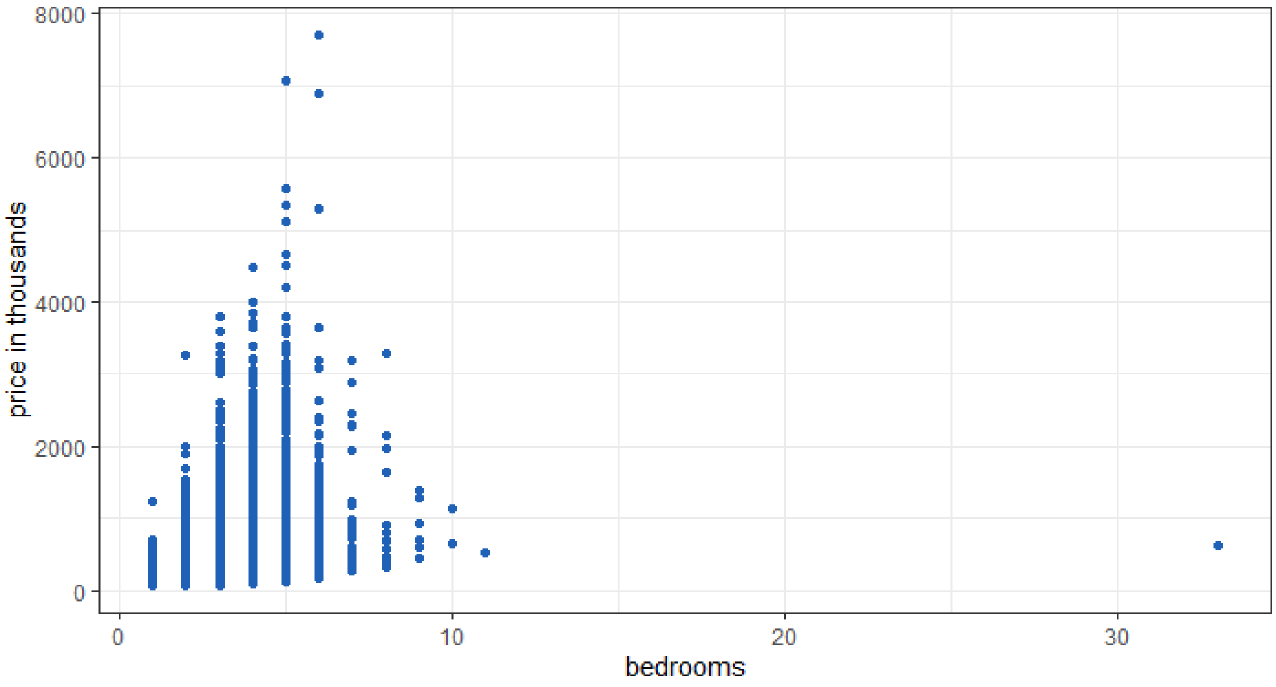
view	int	An index from 0 to 4 of how good the view of the property was (check this predictor)
condition	int	An index from 1 to 5 on the condition of the house, relative to age and grade
grade	int	An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design
sqft_above	int	The square footage of the interior housing space that is above ground level
sqft_basement	int	The square footage of the interior housing space that is below ground level
yr_built	int	The year the house was initially built
yr_renovated	int	The year of the house’s last renovation
zipcode	int	Zipcode area the house is in
lat	num	Latitude
long	num	Longitude
sqft_living15	int	The square footage of interior housing living space for the nearest 15 neighbors
sqft_lot15	int	The square footage of the land lots of the nearest 15 neighbors

Present in this data set are houses sold in 70 zip codes in King County with an average sale price of $540203 and a median sale price of $450000. This would indicate that there are sales present that were sold at a significantly higher price than the median. The average number of bedrooms and bathrooms are approximately 3 and 2, respectively, and the average amount of living space is about 2000 sq. ft. Before our data exploration, we checked the data for blank values and cleared rows that had either zero bedrooms or zero bathrooms.

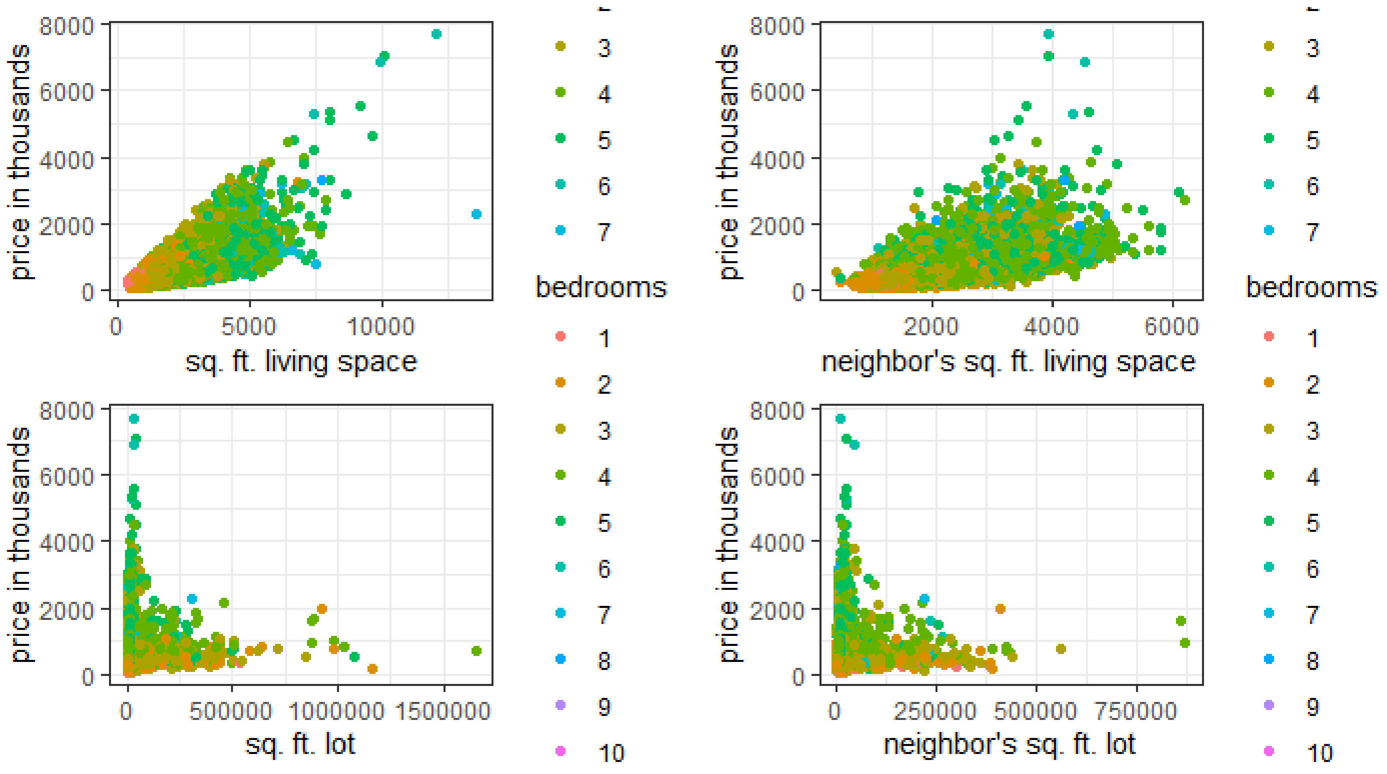
We started with exploring scatter plots for looking for relationships with price and other variables. As you can see in the figure, there is a clear positively correlated relationship with the square footage of living area with higher price, as in the more square footage the higher the selling price tended to be, with a similar but not so tightly correlated relationship with the square footage of the nearest neighbors. This would indicate these houses are located near similarly sized homes.



Another point of interest from the scatter charts is that there appears to be a data error in the bedrooms plot where a modestly price house has 33 bedrooms. We subsequently removed this row from the data.

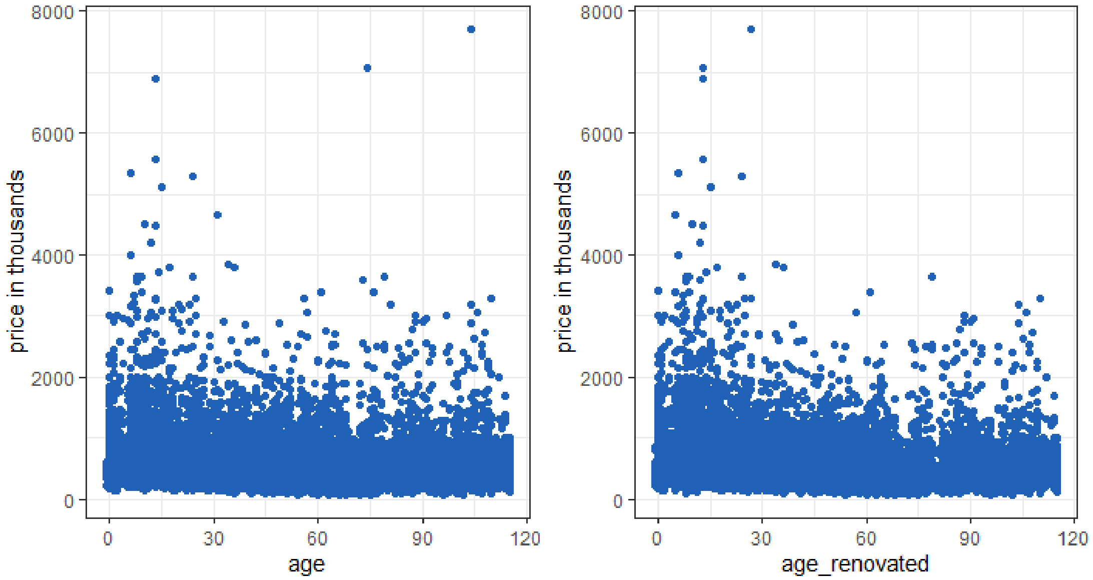


With the data error removed, we then explored the relationship with the square footage of the properties, as well as that of their nearest neighbors, and the price using bedrooms as a discrete variable.



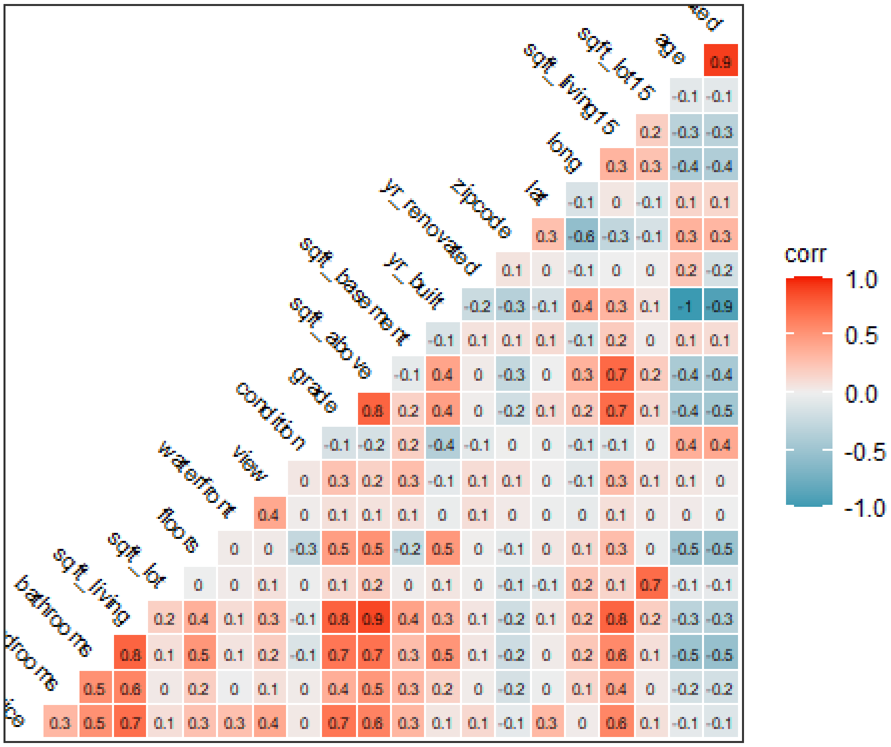
This chart demonstrates a similar relationship between the number of bedrooms and square footage of the property. The variability in the nearest neighbors is also clearer, as the number of bedrooms and the square footage of the nearest neighbors has a much less distinct relationship.

We added a few calculated variables to further explore the relationship with the age of the property, when it was last renovated, and the price. For homes that have not been renovated, the time frame defaults to the age of the property.



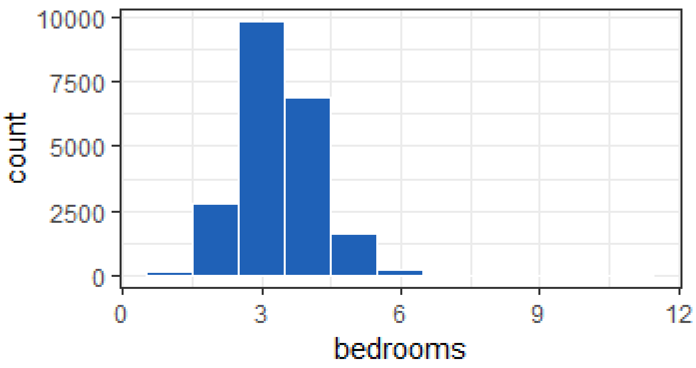
The entire spectrum of ages of homes appears to be well represented in this data set, with a slightly higher density of homes across price points built in the last 50 years. It also appears that higher priced homes tend to have been renovated looking at the shift of these points between the two plots.

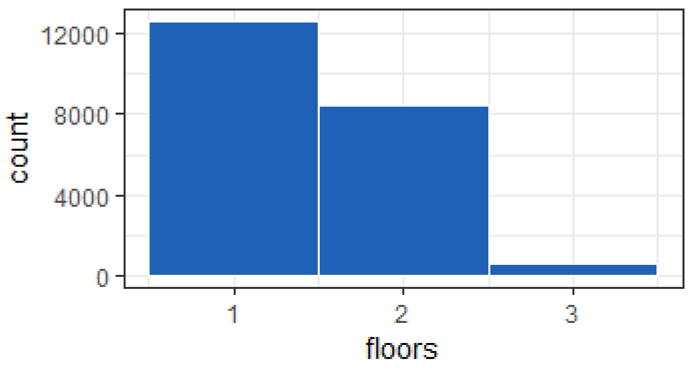
We then plotted a correlation matrix which would be our primary starting point for choosing relevant variables for modeling.

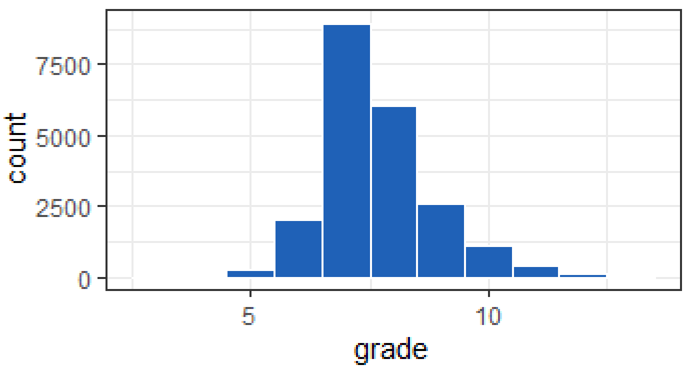


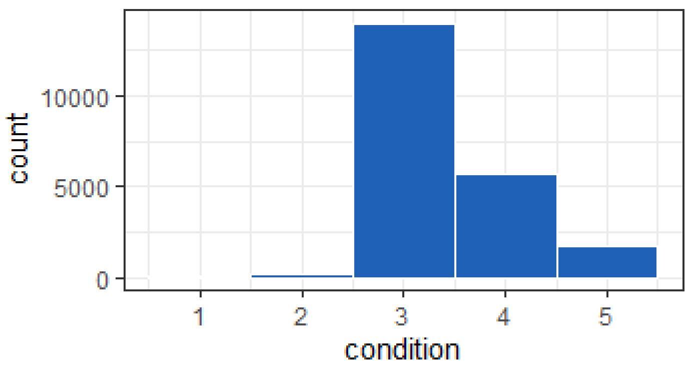
Grade and square footage have the highest correlation with price at 0.7 as compared to all other attributes. The second most important predictors as per the correlation value of 0.6 are the square footage above the basement and average square footage of the 15 nearest neighbors.

We then plotted histograms to observe frequency distribution of discrete variables. As observed in the summary statistics, the most common number of bedrooms is 3. Additionally, most homes have one or two floors and the most common grade in the data set is 7. It is notable that there is a significant number of houses represented with an above average grade.

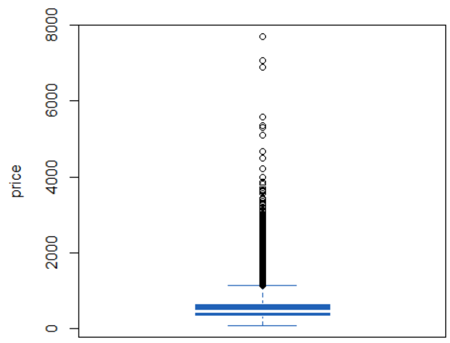


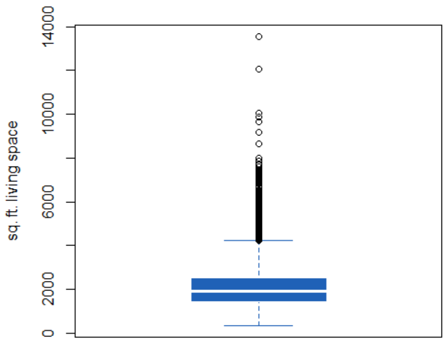


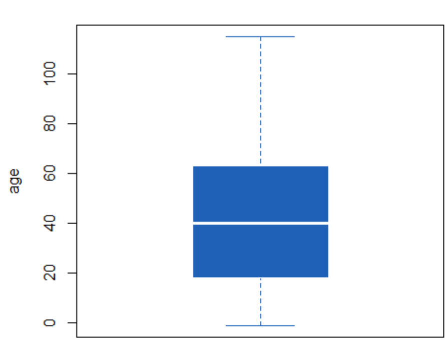




We then started exploring more statistical characteristics of the data set with box plots of the price, square foot living space, and age represented in the data set.

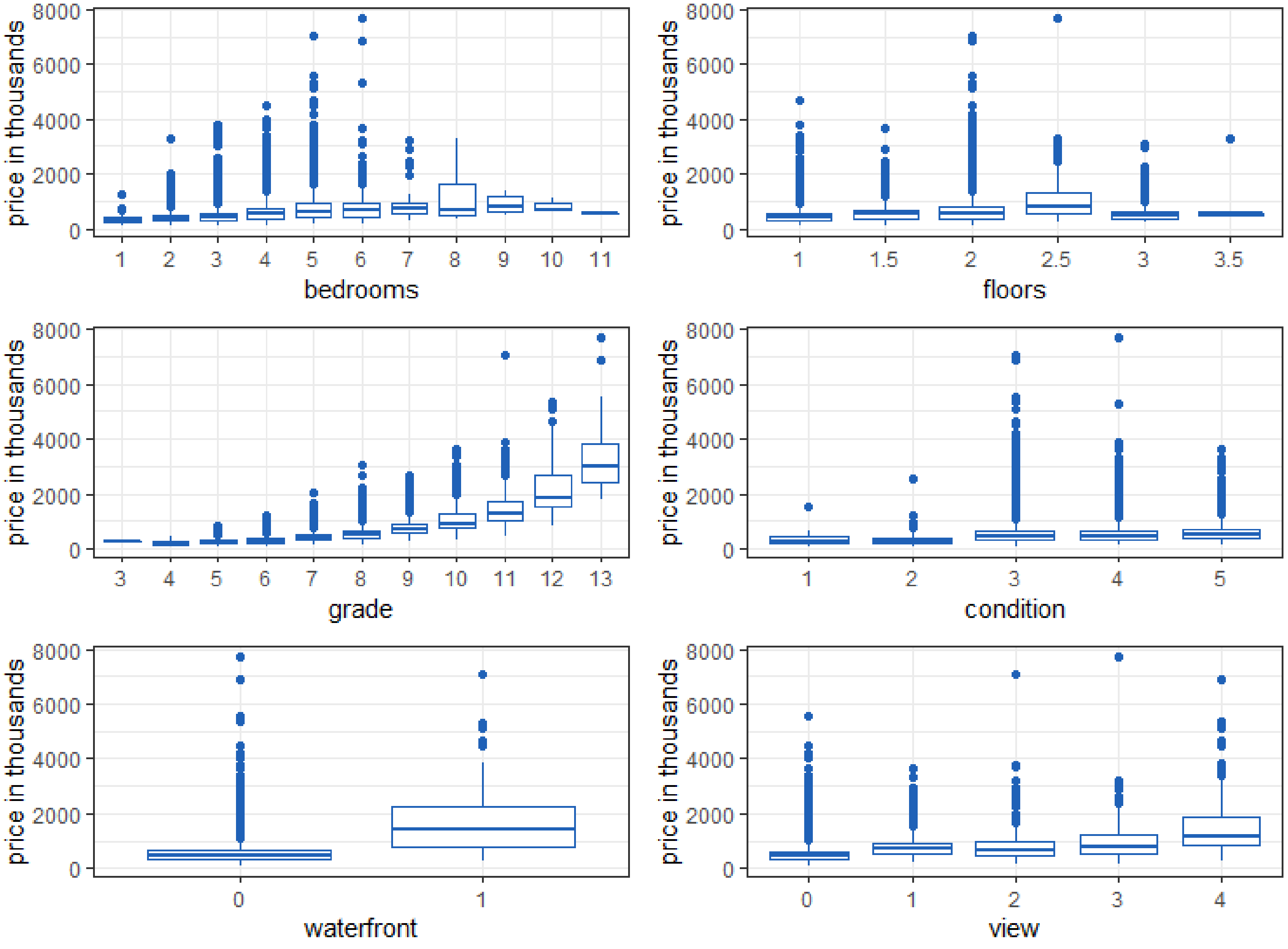




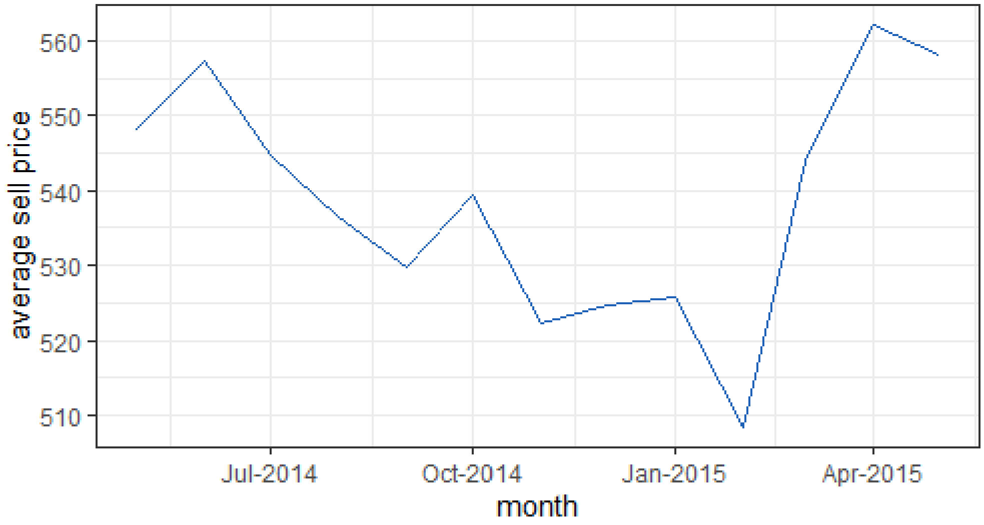


There are many statistical outliers in both price and square foot living space. This would explain the variance between the median price and the mean price and further demonstrates the relationship between price and square footage. As observed in the scatterplots, the age of the properties is well distributed with a slight skew towards younger homes. Even more information about the qualities of this data set and their relationship with price can be seen by subdividing the box plots by factors.

Further exploration of the different features using box plots, we can see that among houses with a lower number of bedrooms have more outliers in pricing whereas once we reach the 8 bedrooms subgroup and above there are no statistical outliers for price. We also observe that if the grade of the house i.e. the construction and design of the house is of superior quality, the selling price for the house is more. As the condition of the house is relative both grade and age of the property, that relationship between grade and price flattens under the condition factor. The box plots for waterfront and view demonstrate that houses with better views and those overlooking a waterfront are costlier as expected. King County has the potential for a lot of good views and the prices in the data set reflect that relationship.



Lastly, we explored the seasonality of sales by plotting the average price per month. Sale price averages ranged between a little over $560000 and $510000, with February seeing the lowest average price and April seeing the highest average price.

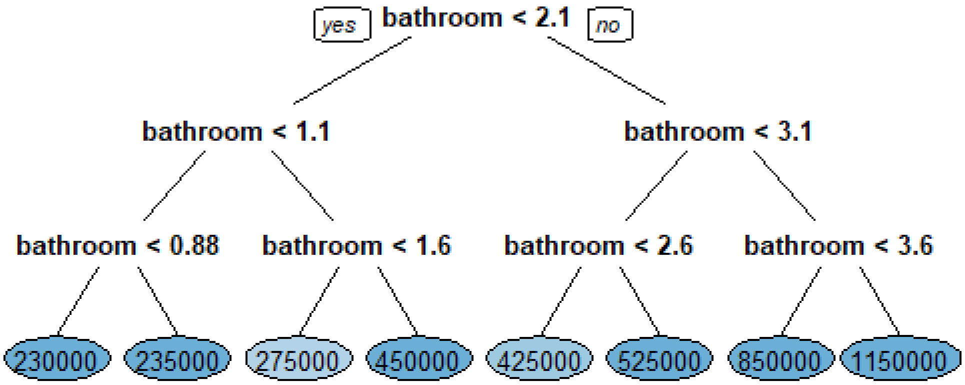


Classification

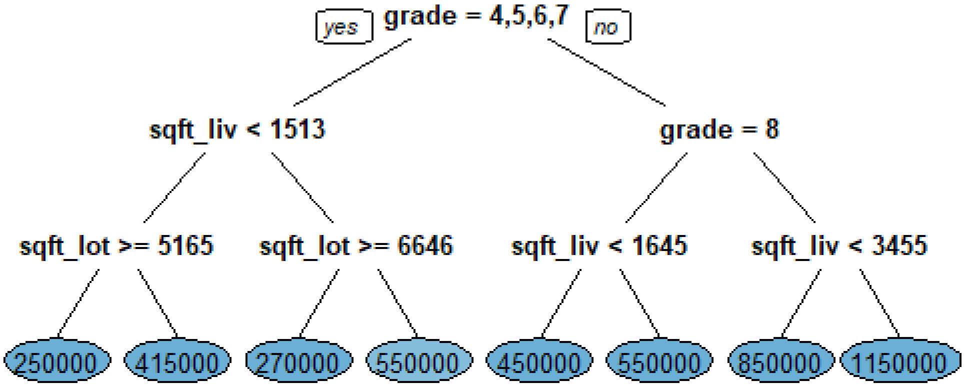
Decision Trees

We used the classification tree model to further validate predictors that are of most value in predicting selling price. As price is a continuous attribute, we divided the prices into bins and added controls such as maxdepth = 3 and minsplit = 1 for computational simplicity. Our first thought was to see how the tree looks like with all attributes, but it turned out that it did not make sense as many of the attributes were of little use.

We generated classification tree 1 using bedrooms and bathrooms and classification tree 2 using attributes that have good correlation coefficient from the correlation matrix and found that grade, square footage of living space, and square footage of the lot are more important predictors as compared to bedrooms and bathrooms.



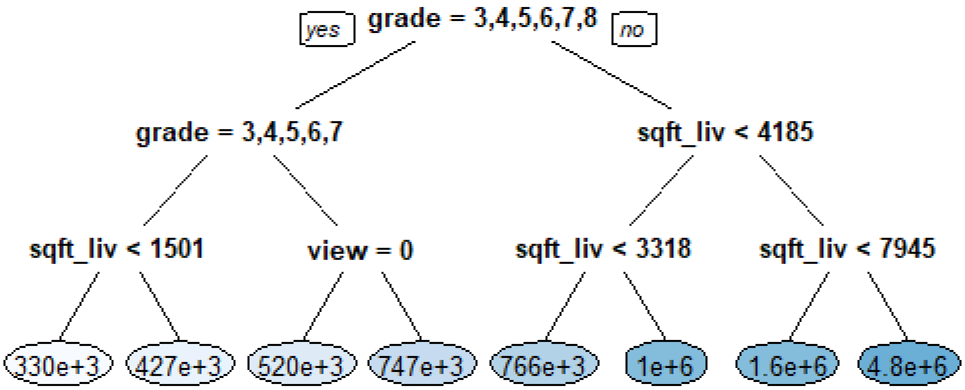
Classification Tree 1



Classification Tree 2

The prediction accuracy when calculated using a Confusion Matrix for classification tree 1 and 2 come out to be 3.21% and 4.15% respectively which is a lot lower from what we expected. We generated a couple of more classification trees with different predictors but the accuracy for all remained between 3% to 4%. This may be due to the limitation of depth for the trees.

We then generated a regression tree, for which we did not convert the prices into bins and found that the most dominant predictors are grade and square footage of living space. This aligned with our insights from the classification trees as well. We calculated RMSE whose value came out to be 257242.2 which is very high, so its accuracy for prediction is low as well.



K-Nearest Neighbor

Our next attempt at using a classification model was with the k-nearest neighbor method. Once again, we split the prices into bins for classification. This time, we also tested various bin sizes and as expected, the larger the bin size, the more accurate the prediction. It is notable that the classification performance improved with the same bin sizes over the decision trees from around 4% to over 40%.

Price bins	K	KNN Result
100,000	Tried with K between 1 to 20	Ranges from 40 to 45%
200,000	Tried with K between 1 to 20	Ranges from 60 to 65%
300,000	Tried with K between 1 to 20	Ranges from 70 to 75%
400,000	Tried with K between 1 to 20	Ranges from 75 to 80%
500,000	Tried with K between 1 to 20	Around 82 to 83 %
600,000	Tried with K between 1 to 20	Around 84%
700,000 and above	Tried with K between 1 to 20	Around 85%

Multilinear Regression

For our first model we chose all predictors except id, date. We then split the data with 70% in the training set and the remaining 30% in the validation set.

The Adjusted R-squared for our first model was 0.7033, and the RMSE of this model was 204826.8. However, when calculating the model's mean-square error with the validation data, we received the warning message: "prediction from a rank-deficient fit may be misleading". Investigating this error led us to find that this model had multicollinearity. Further investigation showed that the square footage of basement is not defined because of singularities, so we took this variable out to refine the model. After running this model again without the basement size, we received the same result but without the warning message.

We next created a model with predictors that had a correlation coefficient with price is greater than 0.3, which came to 11 predictors. The Adjusted R-squared for this model is 0.6615 and the RMSE of is 218882.5. This model also had a warning for multicollinearity.

Lastly, we created a model with predictors that had a correlation coefficient with price greater than 0.5. With this condition, our fourth model had 5 predictors. The Adjusted R-squared was 0.6124, and the RMSE of this model was 228963.8.

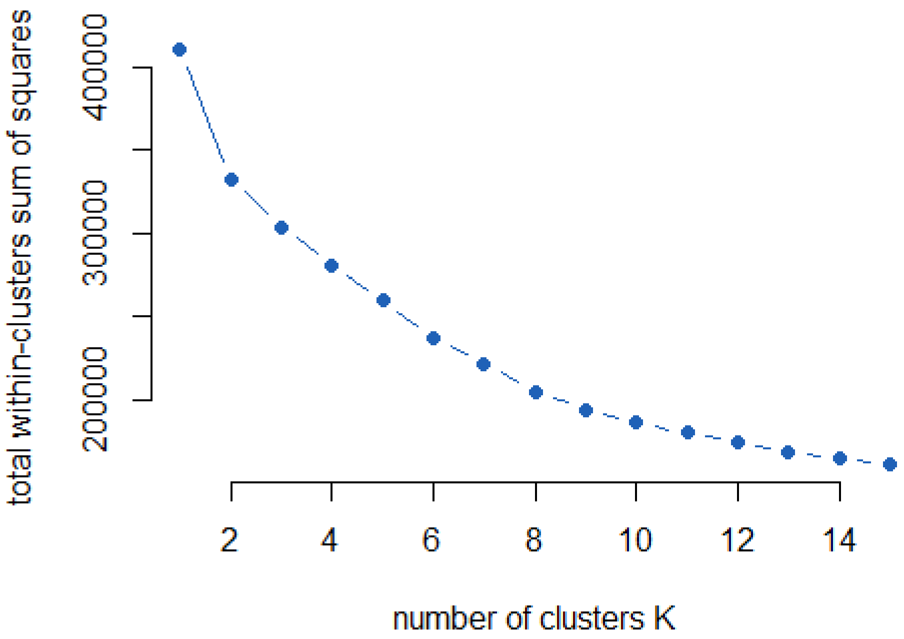
Model Number	Adjusted R-squared	RMSE	Notes
Model 1	0.7033	204826.8	multicollinearity + singularities
Model 2	0.7033	204826.8	BEST
Model 3	0.6615	218882.5	multicollinearity
Model 4	0.6124	228963.8	

Out of these four models, we chose Model 2, because this model did not have multicollinearity or singularities. In addition, the adjusted R-squared is larger and the RMSE is lower than Model 4.

Clustering

K-Means Clustering

Once we had our model, we considered ways we could improve our model with the given data set. Due to the number of outliers in pricing and square footage, we decided to subdivide the data into clusters using the k-clustering method. Based on the elbow chart below, we decided on dividing our data into nine clusters.



Of the 9 clusters, cluster 7 was the most expensive with a normalized price of 3.055 and positive values throughout the selected variables. This cluster most likely corresponds to expensive homes with multiple bedrooms and ample living space. Cluster 4 had a price value of 1.911 and positive values throughout.

Three clusters (3,5, & 6) had negative normalized values for price and a preponderance of negative values throughout. These clusters probably correspond to lower income housing and apartments.

Four clusters have positive normalized values for price but are significantly less than cluster 7. Clusters 1, 2, 8, and 9 have price values ranging from 0.008 to 0.332. These clusters probably represent more middle-income housing environments.

cluster	price	bedrooms	bathrooms	sqft_living	floors	waterfront	view	condition
1	0.008124	0.381366	0.590117	0.485776	0.871982	-0.087205	-0.220478	-0.444187
2	0.057169	0.718130	0.339877	0.417734	-0.826924	-0.087205	0.027204	0.380132
3	-0.179989	-0.491984	0.370323	-0.451546	1.449030	-0.087205	-0.178073	-0.590941
4	1.910642	0.865884	1.549760	2.115501	0.773834	-0.087205	0.789706	-0.233260
5	-0.573031	-0.381740	-0.591880	-0.646430	-0.667614	-0.087205	-0.240725	0.303616
6	-0.446691	-0.801128	-1.122303	-0.893030	-0.698343	-0.087205	-0.206339	-0.097634
7	3.055164	-0.078767	0.730913	1.190848	0.272353	11.466678	4.609266	0.190617
8	0.331820	0.056735	0.135321	0.108291	-0.007876	-0.087205	0.241705	-0.303772
9	0.230472	0.425740	0.018287	0.149076	-0.351963	-0.087205	0.374952	0.853993

Clusters v. Regression

Once we had our clusters, we tested our chosen model individually against each cluster to see how it would perform. After running our model against each cluster, we found that it performs best on the cluster with the highest sell prices, with an Adjusted R-Squared value of .8374. In testing the other models, the degradation of accuracy was similar to testing the models on the complete data set.

Conclusion

Results & Insights

1. On an aggregate level, we observed several factors that have a greater impact on the housing price. The strong predictors are grade, number of bedrooms, number of bathrooms, how large the living space is, and if the property is located on a waterfront.
2. Intuition suggests the location of a house might be important. For example, based on the data, 98039 zip code has a median sale price of 2.16 million whereas 98002 has a median sale price of 234.28 K. Further clustering analysis reveals the property location is indeed very influential. We observe very strong correlation within two clusters, corresponding to the most expensive cluster and the least expensive cluster (R squares above 0.80).

Who would benefit from this model?

1. Home buyers can evaluate the individual listing prices based on the property factors. Historical data might be able to provide crucial insights regarding the asking price, future appreciation of the property, as well as neighborhood listing information.
2. Real estate agents can take advantage of the visualization and models to better understand the historical real estate pricing trends in the King County.
3. Real estate companies can better allocate resources based on the current pricing trends.

Potential Improvements

1. Real estate trends vary greatly among different regions. It would be beneficial to analyze other adjacent areas to better understand the greater Seattle area housing market.

2. The recent tech boom has greatly reshaped King County property trends. Historical data may not be able to capture very recent changes. To ensure the accuracy, re-training our model on a periodic basis will be necessary.
3. Further K means analysis & logistical regression will be able to refine our cluster distribution, improving the accuracy of our model as well as giving us the potential for forming different models based on clustering characteristics.
4. Additional crucial factors might be helpful to include in our modeling exercise, such as neighborhood school ratings, walkability scores, flooding area designation, etc.
5. Another interesting topic is to test the seasonality by isolating the month as a variable in a data set that spans multiple years so that we could better understand the pricing fluctuations over time as well as in relation to the economic performance of local employers.