Predicting House Sales in King County, Washington

Mitch Modlich | February 2023

Problem Statement

King County, Washington contains most of the Seattle metropolitan area, an important technology hub and the largest city in Washington state. This area is home to over 4 million people. Real estate data made publicly available by King County provides the opportunity to analyze and develop modeling techniques that may assist in the process to predict prices for new house sales in the region. Multiple interesting features are included in the dataset for data exploration and regression-based modeling.

What features of a single-family house sold between May 2014 and May 2015 in King County, Washington can be used to predict future sale prices to be used to maximize average sale price for next year for a real estate firm?

Data Wrangling

The dataset was obtained from Kaggle. The original data source is the King County assessor government website.

The data consists of 21,613 sales of single-family houses between May 2014 and May 2015. There are 21 columns which contain the following features:

- id - Unique ID for each sale
- date - Date of sale
- price - Sale price ($)
- bedrooms - Number of bedrooms
- bathrooms - Number of bathrooms
- sqft_living - Area of interior living space (sq ft)
- sqft_lot - Area of land on lot (sq ft)
- floors - Number of floors
- waterfront - Waterfront location
- view - Rating of property view (0-4)
- condition - Rating of property condition (1-5)
- grade - Rating of house construction and design quality (1-13)
- sqft_above - Area of interior living space above ground level (sq ft)

- sqft_basement - Area of interior living space below ground level (sq ft)
- yr_built - Year house was built
- yr_renovated - Year house was last renovated
- zipcode - Zipcode
- lat - latitude coordinate
- long - longitude coordinate
- sqft_living15 - Average area of interior living space of 15 nearest neighbors
- sqft_lot15 - Average are of lot of 15 nearest neighbors

FEATURE DESCRIPTIONS

Number of bathrooms is counted in 0.25 increments. A bathroom can be referred to as a one-quarter, half, three-quarters, or full. Here are examples of each:

- full bathroom: toilet, sink, shower, bathtub
- three-quarters bathroom: sink, shower, toilet. No bathtub
- half bathroom: toilet, sink. No shower or bathtub
- one-quarter bath: toilet only

View, condition, and grade are qualitative ranks based on certain criteria. Condition and grade ranks are described on the King County Assessor website.

DATA CLEANING

Date was converted to a Pandas datetime object from string. Price was converted to integer from float, as all sales are rounded to the nearest integer dollar.

There are 21,436 unique ID values of 21,613 sales. The rows with duplicate ID values have different dates for each unique ID. This suggests that some (177) houses were sold multiple times over the timespan of the dataset.

Sqft_lot has a very large range with minimum and maximum values (520 sq ft, 1,650,000 sq ft), but King County encompasses a very large space of dense urban areas and expansive rural areas.

Summary statistics for the data show that bedrooms has a max value of 33 which seems unusual for a single-family house. Additionally, there are 13 rows which have a zero value for bedrooms and 10 rows which have a zero value for bathrooms. The 33 bedroom value

appears to be an error in data entry when taking into account the other features of the sale. This was changed to 3 bedrooms. The zeroes in bathrooms and bedrooms were replaced with the median values for those features.

Exploratory Data Analysis

House Sale locations (Figure 1) are mostly clustered along the west side of the county in the Seattle metropolitan area. The Tacoma area to the south of Seattle lies in a different county. The geographic scope of this project is within King County.

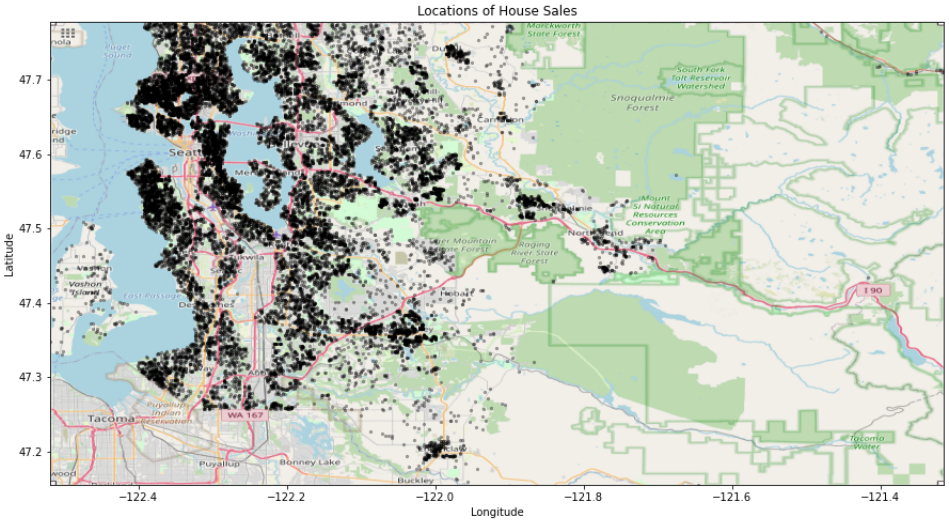

Figure 1: Locations of House Sales in King County

The target or dependent variable of this analysis is house sale price. Figure 2 shows a histogram or count plot of the target variable. The distribution is not normal, and it is right-skewed. Most of the data is in the range of $1,000,000 or less. A long tail of data stretches up to several million dollars.

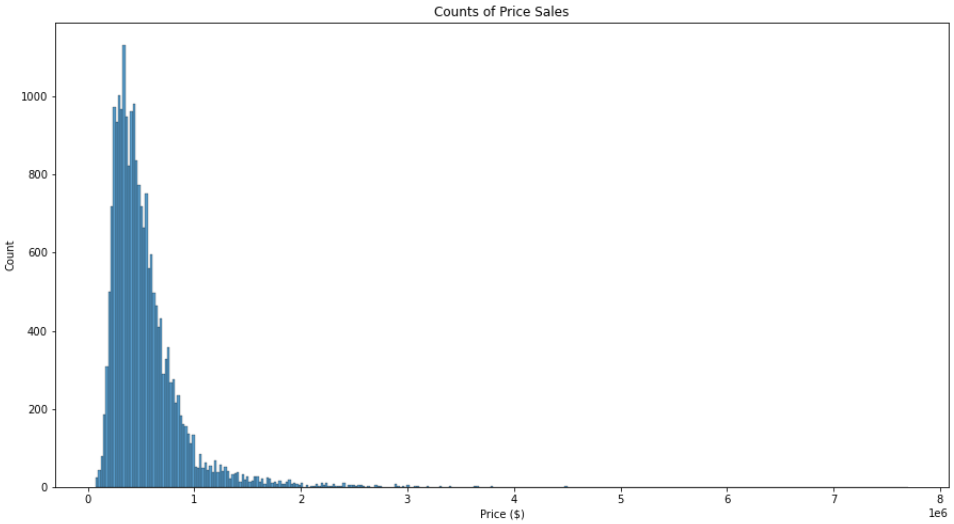

Figure 2: Histogram of House Sale Price

One of the most common and probably important features of house sales is living space (square feet). The distribution of this feature is shown in Figure 3. This is also not a normal distribution, and it is right-skewed. Most house sales had a living area of 4000 square feet or less. A long tail of data stretches to approximately 14,000 square feet.

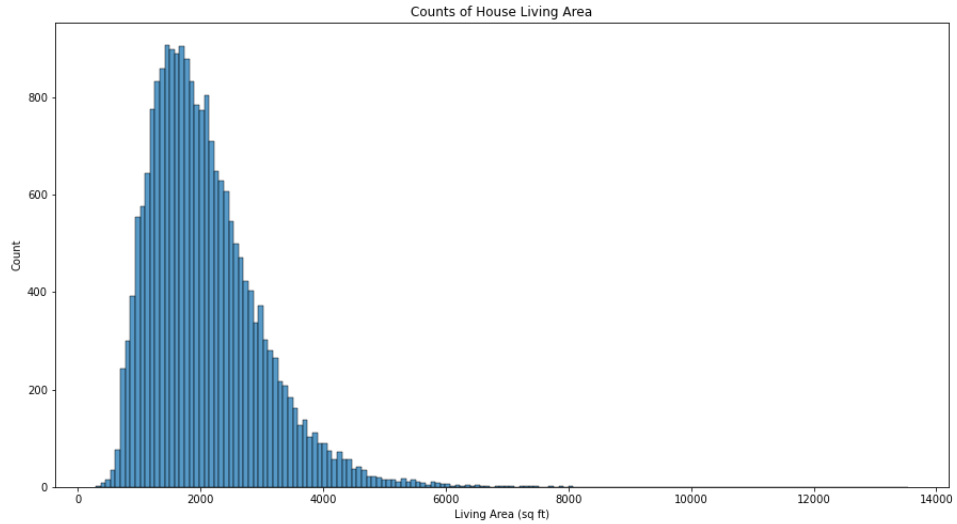

Figure 3: Histogram of House Living Area

Lot area is another interesting feature of house sales. The distribution of this feature is shown in Figure 4. The data is limited to the 99th percentile to better show the distribution as it is extremely right-skewed. Most house sales have lot areas of 20,000 square feet or less. This feature is a good demonstration of the disparity between urban and rural areas in King County. Some rural lots are 1,000,000 sq ft (23 acres) or more.

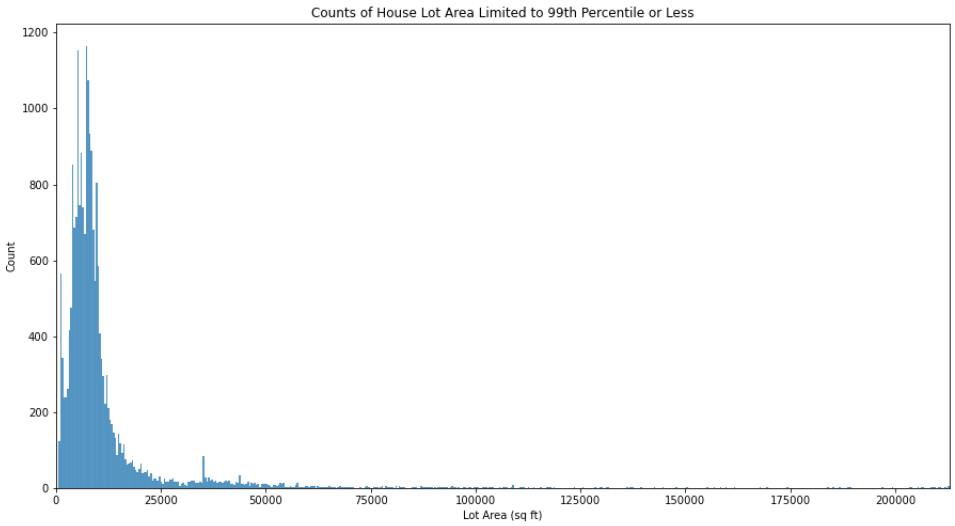

Figure 4: Histogram of House Lot Area Limited to 99th Percentile

Two additional common features of house sales are the number of bedrooms and the number of bathrooms. These distributions are shown in Figure 5 and Figure 6. Most houses sold had 3-4 bedrooms. Bathrooms have a bimodal distribution with centered around 1 and 2.5, which are the most common values.

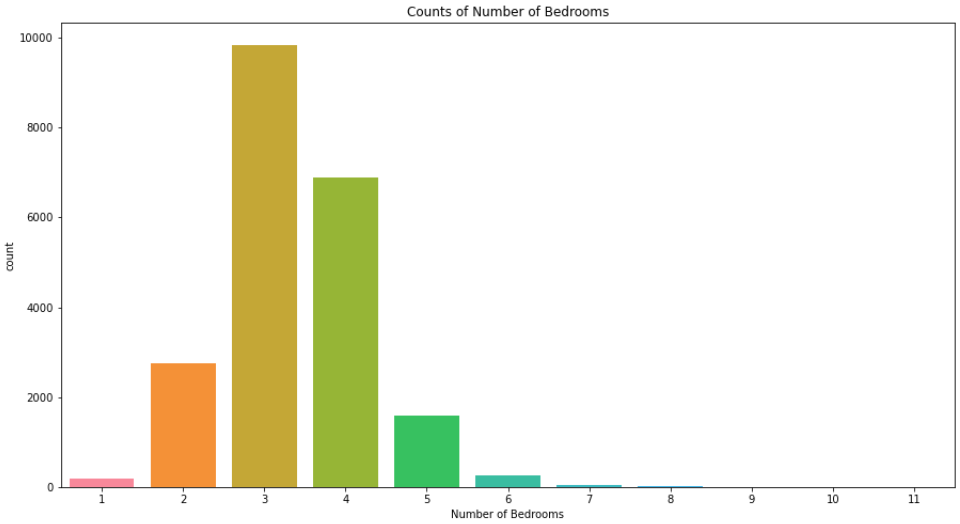

Figure 5: Histogram of Number of Bedrooms

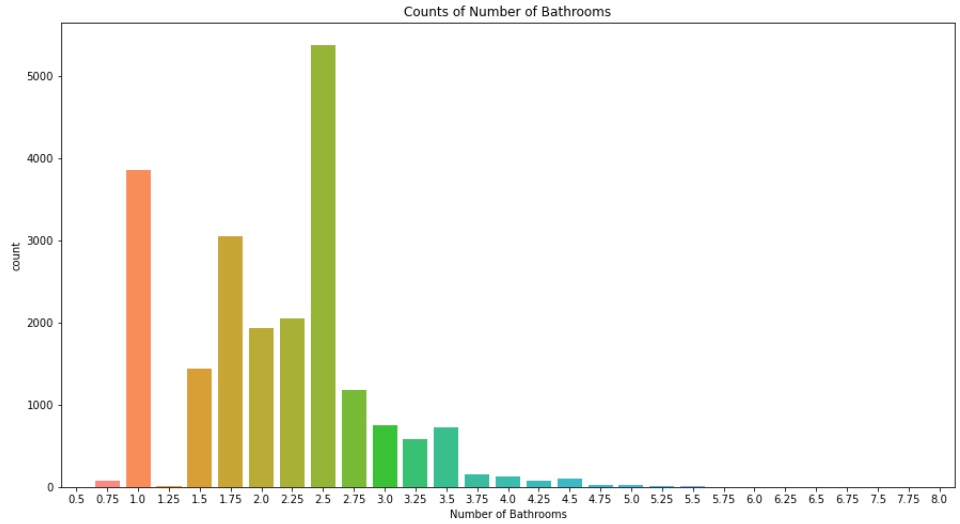

Figure 6: Histogram of Number of Bathrooms

House location or neighborhood is very important to sale price. A good approximation of location is zip code, as these tend to be focused on certain neighborhoods and municipalities. Figure 7 shows the mean sale price by zip code.

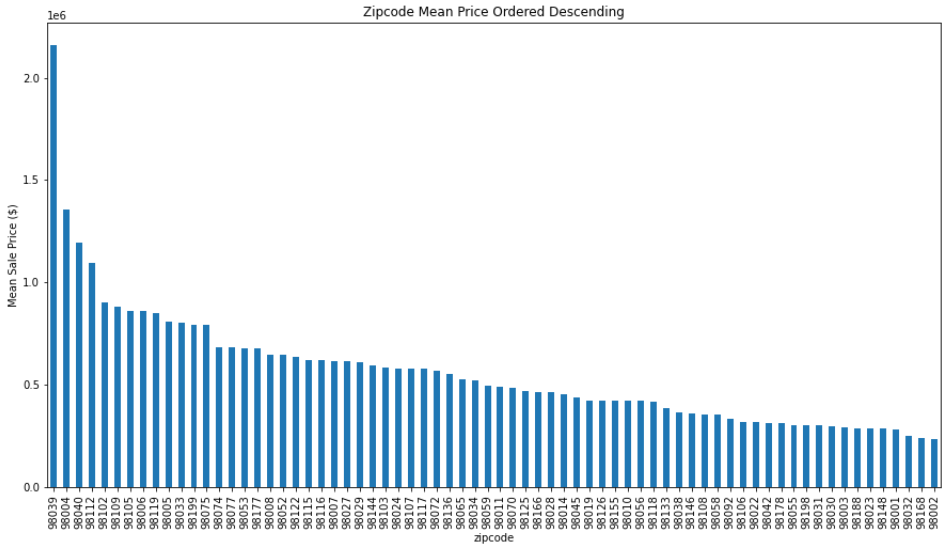

Figure 7: Average (Mean) Sale Price by Zip Code

The top five zip codes for mean sale price are 98039, 98004, 98040, 98112, 98102. This is a mostly contiguous area from downtown Seattle to the east around the waterfront of Lake Washington. The bottom five zip codes are 98002, 98168, 98032, 98001, 98148. This includes the area around Seattle-Tacoma International Airport (SeaTac) and the southwestern portion of the county. Figure 8 shows a map of zip code boundaries in western King County and the Seattle metropolitan area. The top five and bottom five zip codes by mean sale price are shown on the map by yellow and green stars, respectively.

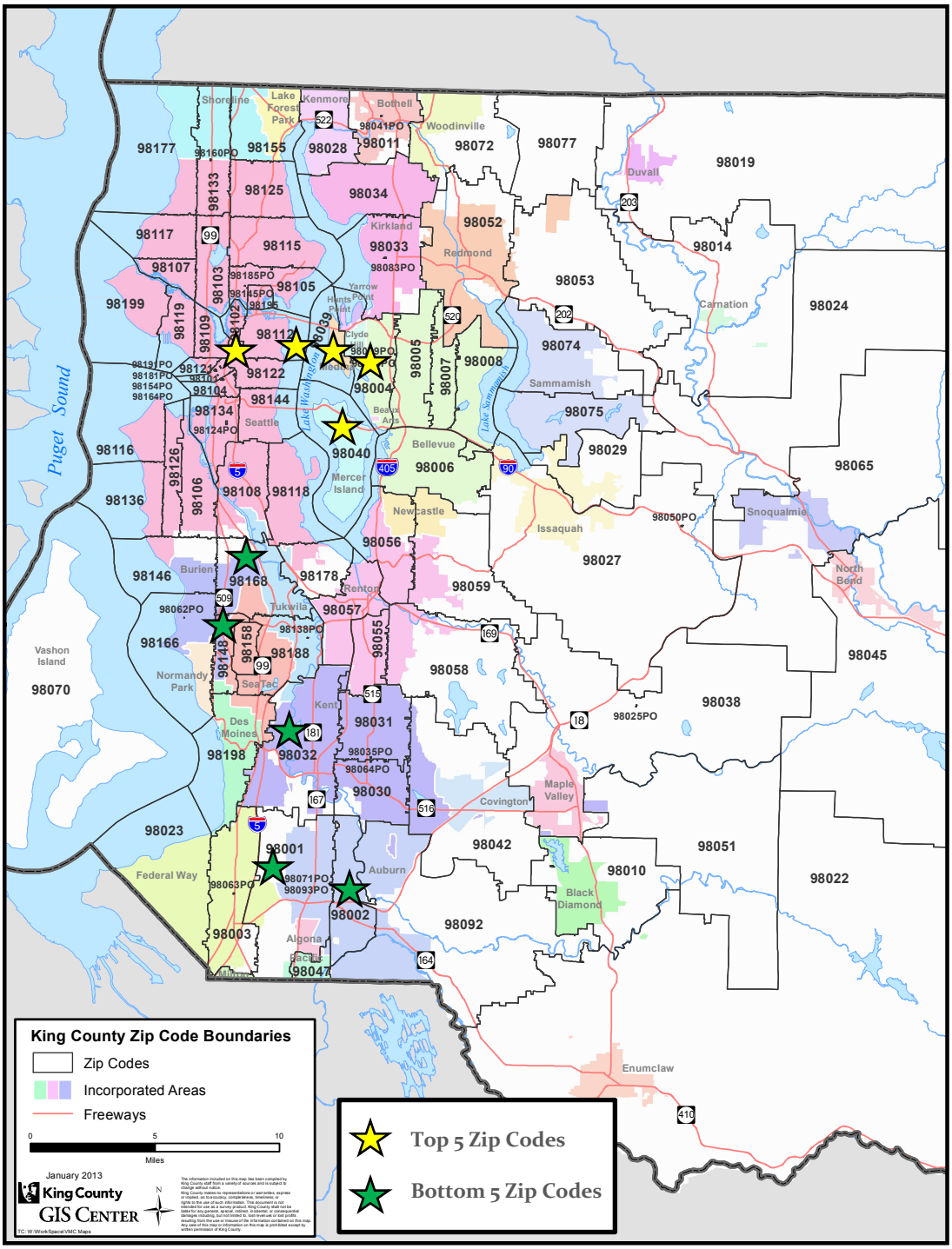

Figure 8: Zip Code Boundaries Map of Western King County

Another way to visualize the house sale price geographically is sale price per square foot of living space. This is shown in Figure 9. Darker circles represent higher price per square foot. In general, house sales with the highest price per square foot are clustered around the previously delineated area around downtown Seattle and Lake Washington. The lowest price per square foot of house sales are to the south of Lake Washington and the bottom southwestern portion of the county.

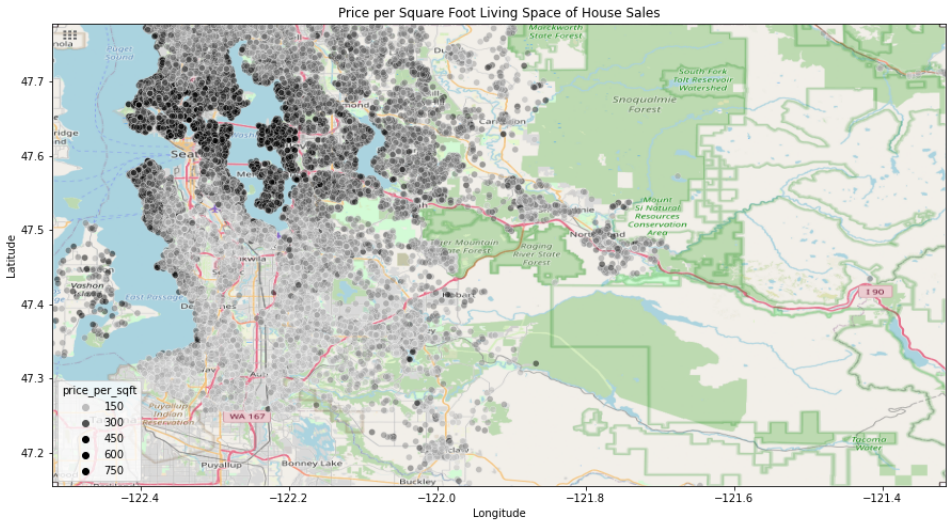

Figure 9: Price Per Square Foot Living Area of House Sales

Figure 10 shows bivariate plots of various features against the target variable price. Sqft_living, sqft_above, and grade appear to have strong correlations with price. Bathrooms also shows a moderate correlation with price.

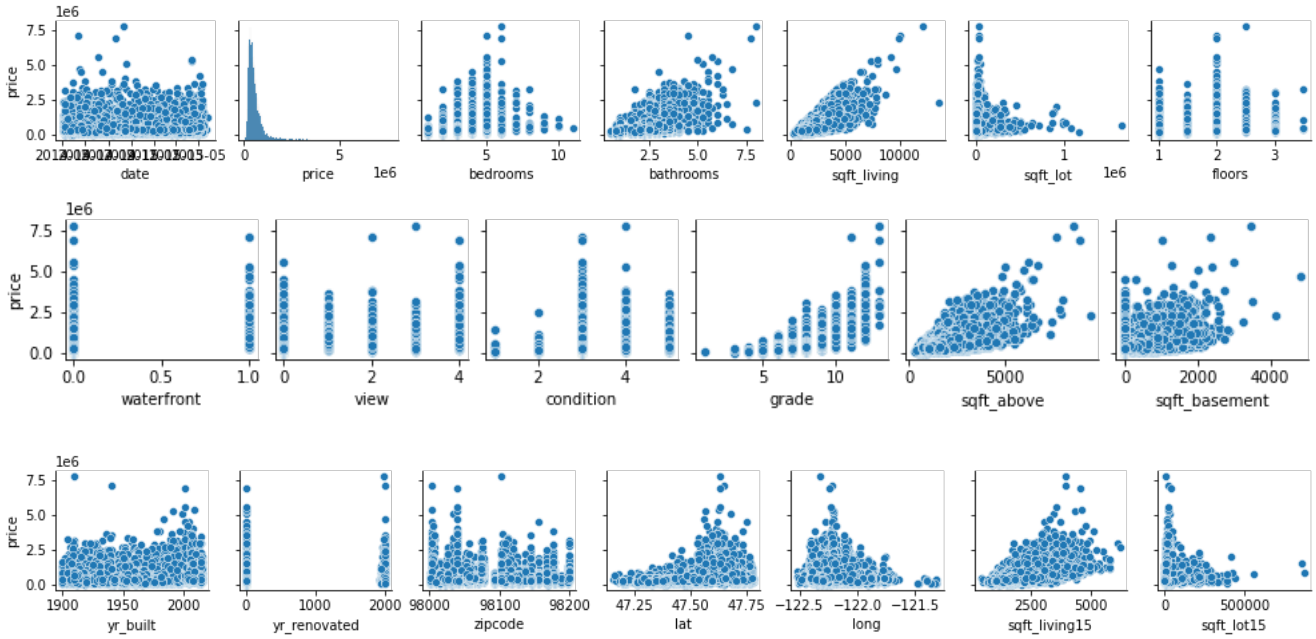

Figure 10: Bivariate Plots of Features against Sale Price

Figure 11 is a correlation heatmap depicting the Pearson correlation coefficients between various variables in the dataset. These range between -1 and 1 for negative and positive correlations. In this diagram, positive correlations are colored red, while blue denotes negative correlations. The strongest correlations with price are sqft_living, grade, sqft_above, sqft_living15, and bathrooms. Some features such as zip code (categorical) and id are not interpretable in this diagram. Some independent variables or features show correlations with one another. These include sqft_living, sqft_above and grade.

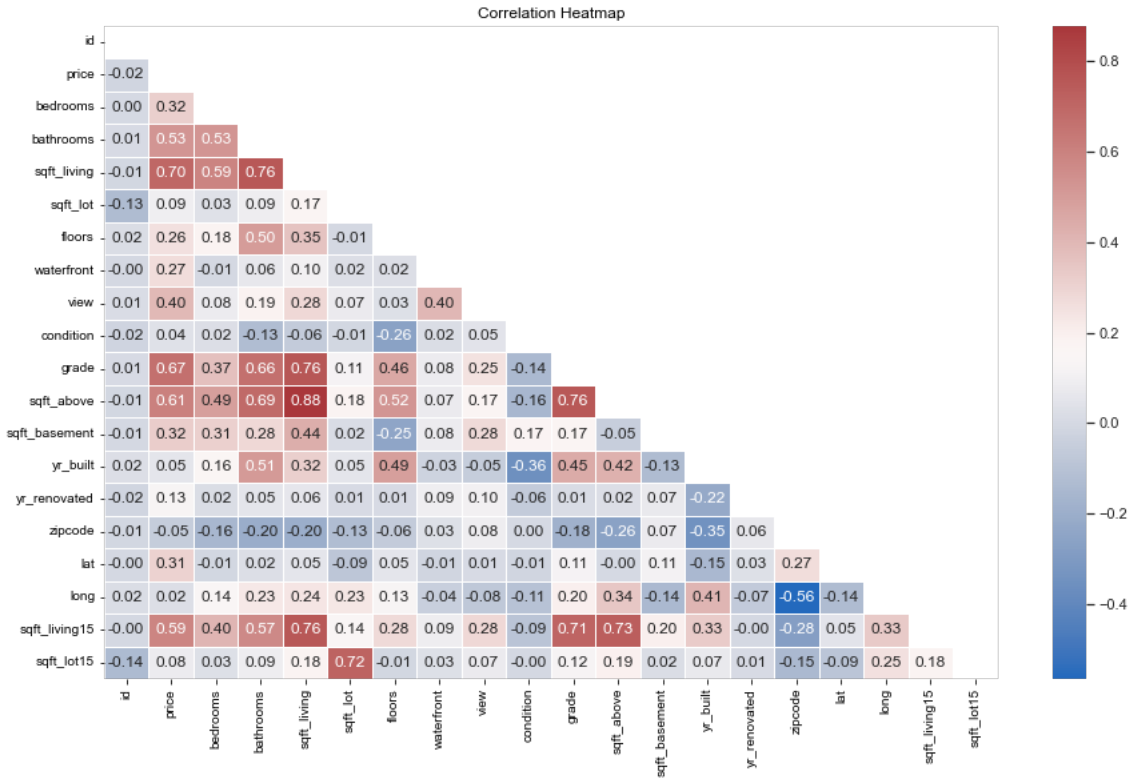

Figure 11: Correlation Heat Map

Data Modeling

A regression analysis approach was taken for prediction of the model of house sale price, and to determine which features are the most important to house sale price. Some features were not included in the model, while others were modified. Model algorithms explored are linear regression (simple (univariate) and multiple (multivariate), ridge regression, and random forest regression.

PRE-PROCESSING AND FEATURE ENGINEERING

1. Binary features were created for renovated, basement, and view. The age of the house (yr_built) was binned into decades (e.g., 1900s, 1950s, 1990s, etc.). This is intended to simplify the interpretation of these features. A feature for month of sale was created to represent the date feature.

2. Dropped columns include id, date, sqft_above, lat, long, sqft_living15, sqft_lot15. Sqft_above, sqft_living15, sqft_lot15 were dropped for multicollinearity to sqft_living. Lat and long were dropped in favor of zip code.
3. Condition, grade, decade_built, zipcode, and month are one-hot encoded as categorical features.
4. The resulting feature matrix consists of 115 columns (there are 70 zip codes and several other categorical features)
5. The features (X) and target (y) are split into training and test sets randomly.
6. Normalization of numeric features to values of a range of (0,1) is performed before determining feature importance.

REGRESSION MODELS

Three types of regression models were explored:

1. Linear regression
   a. Simple linear regression (sqft_living on price)
   b. Multiple linear regression (all features in pre-processing workflow above)
2. Ridge regression (all features)
3. Random forest regression (all features)

A baseline model of simple linear regression was performed to determine the usefulness of adding more features in a multivariate analysis. Ridge regression is multiple linear regression with a parameter to induce regularization of the coefficients. This can be useful in mitigating multicollinearity in models with many features. Finally, random forest regression was performed to see if the decision tree ensemble approach can improve upon linear regression. Hyperparameter tuning of ridge regression and random forest regression was also explored. The loss function chosen to compare these models is mean square error (converted to root mean square error (RMSE) to terms of price units). After comparing the trained model cross-validation scores and test scores, multiple linear regression is selected as the best model.

Model Scoring Metrics

Model	Validation Set RMSE	Test Set RMSE	Test Set R2 Score
Multiple Linear Regression	149124	155582	0.83233096
Ridge Regression (tuned)	149805	155592	0.83230862
Ridge Regression	149253	155767	0.83193186
Random Forest Regression	148080	168913	0.80236671
Random Forest Regression (tuned)	148453	169926	0.79998967
Simple Linear Regression	256171	272965	0.48388319

The RMSE of each model is somewhat large at approximately $150,000. Multiple linear regression is a substantial improvement over simple linear regression. The test set and validation set RMSE have similar scores, which demonstrates that the model generalizes well to unseen data. Random forest regression has a wider discrepancy in these scores, which suggests that this model is overfitting by comparison and does not generalize as well to unseen data. This is another reason to choose multiple linear regression as the best model. Figure 12 shows the selected model's test values versus prediction values.

Figure 12: Multiple Linear Regression Model Predictions

The red line indicates perfect correlation. The fit of the model is relatively good, especially at test price values of $1.5 million or lower. As the test price increases, the variance of the model increases and the prediction is not as accurate. Notably, the model consistently underpredicts at higher test values.

Feature Importance

After normalizing the numeric features so that all values are in the range of (0,1), the coefficients of regression are sorted by magnitude to determine the most important features in the model. The top ten features by magnitude are grade_13, sqft_living, zipcode_98039, grade_12, zipcode_98004, waterfront, zipcode_98112, zipcode_98040, zipcode_980102, zipcode_98119. These zip codes correspond to the highest mean sale price zip codes previously mentioned in the report. Of moderate importance are the features sqft_lot and bathrooms. Features of low importance include age of house (decade_built bins), month of sale, basement, and renovated.

Conclusions

Multiple linear regression is a good approach for general price prediction for single family house sales over a large and diverse area such as King County. However, the root mean square error for the model of approximately $150,000 means that the model cannot be used for precise price prediction, especially at the lower end of the range of house sale prices. Still, the model may be useful to start the conversation on house pricing between realtors and their clients.

RECOMMENDATIONS

- To maximize house sales in the future for a real estate firm in King County, the most important features to emphasize are:
    - High building grade (11-13 in particular)
    - Living area square footage
    - Location within the county (zip code), particularly the area around downtown Seattle and to the east along Lake Washington.
    - Waterfront location
- Common features of lesser importance include Lot size and number of bathrooms, and number of bedrooms. The month of house sale, the age of the house, and whether it has been renovated, and presence of basement are of low importance to sale price.

FURTHER WORK

This large-scale model generated interesting insights, but it could be improved by creating smaller-scale models that focus on certain neighborhoods and more narrow price ranges (e.g., below $1.5 million sale price). This could reduce the variance and error of the model price predictions, which increase as the house sale price increases.