Exploration of Housing prices in King County

April 4, 2024

UBC STAT 306

Annabel Lim
Yilin Long
Gary Wu
Edmond Ye

1. Introduction

1.1 Data

The dataset House Sales in King County, USA, contains house sale prices for homes sold between May 2014 and 2015 for King County, including Seattle. The data is found on Kaggle with the methods of data collection or how the variables were measured not disclosed. However, further research indicates that its shape file with the zip code zones for King County is from King County GID Open Data.

1.2 Variables

The dataset consists of 21,613 observations with the following 21 features. The variable `price` will be used as the response variable.

Variable	**Description**
id	Unique ID for each home sold
date	Date of home sold (year month day)
price	Price of home sold ($ USD)
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms
sqft_living	Size of living area (square feet)
sqft_lot	Size of lot (square feet)
floors	Number of floors
waterfront	1 - if property has waterfront, 0 - otherwise
view	Scale from 0-4 of how good property's view is
condition	Rank from 1-5 of house condition
grade	Rank from 1-13 of quantity level of construction and design (1-3 - falls short, 7 - average, 11-13 - high)
sqft_above	Size of interior housing space above ground level (square feet)

sqft_basement	Size of interior housing space below ground level (square feet)
yr_built	Initial year house was built
yr_renovated	Year of last renovation
zipcode	Zip code area of house
lat	Latitude coordinate
long	Longitude coordinate
squft_liv15	Average size of interior housing for closest 15 houses (square feet)
squft_lot15	Average size of land lots for closest 15 houses (square feet)

1.3 Research Question and Motivation

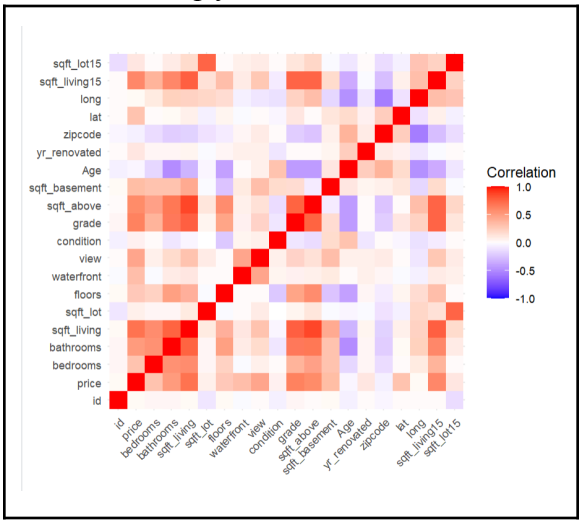
With rising house prices in Vancouver, Canada, there is interest in studying alternative housing markets near Vancouver such as King County and Seattle. By studying the factors driving house sale prices in King County and Seattle, this data analysis aims to provide insight for individuals affected by Vancouver's high housing costs. By gaining valuable insight into market trends, individuals may make informed decisions in real estate transactions and/or consider relocation. This paper investigates what key explanatory variables from the model are associated with predicting the response variable, house sale prices in King County and Seattle.

2. Analysis

2.1 Visualization Summary

The heatmap in Plot 1 on the next page visualizes the correlation between different features of houses within the King County dataset. Warmer colors (red) indicate a positive correlation, while cooler colors (blue) show a negative correlation. Features that are highly correlated with the price, such as `sqft_living` and `grade`, are clearly of interest. Notably, `sqft_living15` and `sqft_above` also show strong positive correlations with several features. On the other hand, `id` shows little to no correlation with other features, suggesting it might not be useful for predictive modeling. The presence of multicollinearity, evident between features like `sqft_living` and `sqft_above`, could impact the performance of some regression models

Plot 1. Heatmap for Correlation Between Features



The histogram in Plot 2 shows a right-skewed distribution of house prices, with a majority of homes in the lower price range and few high-priced outliers. The skewness indicates more affordable homes and some luxury properties.

Plot 2. Histogram for the Response Variable Price

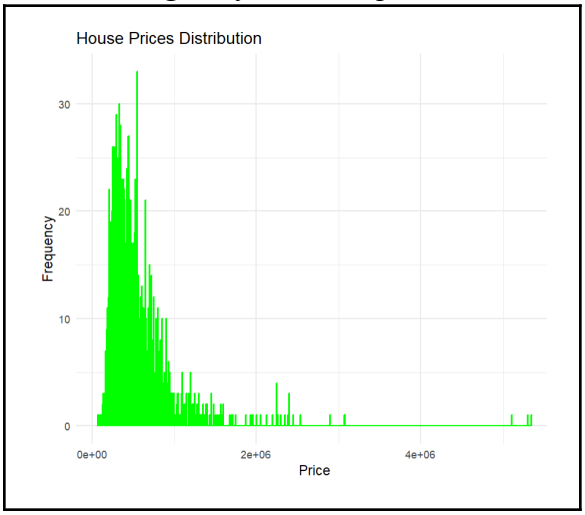
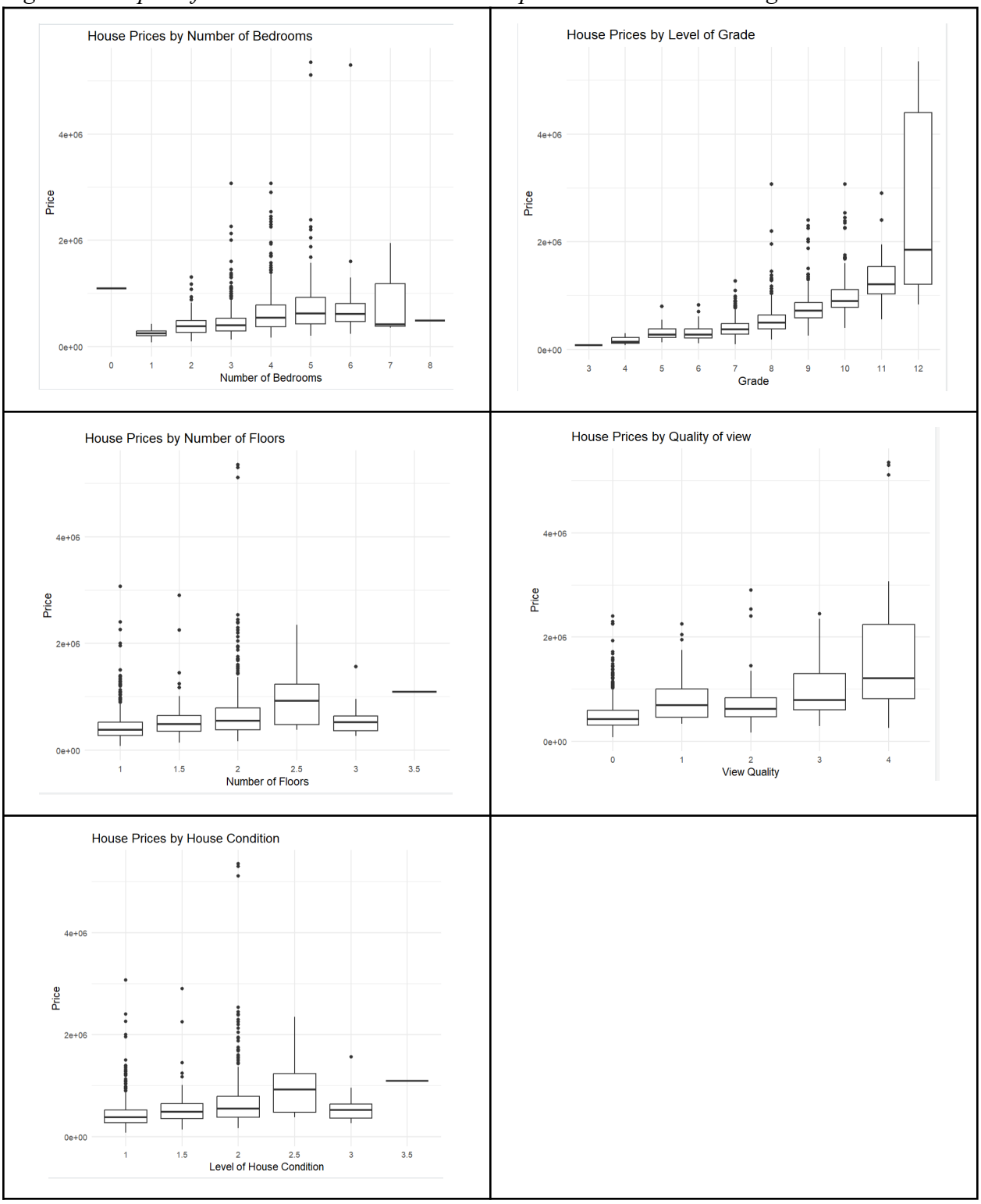


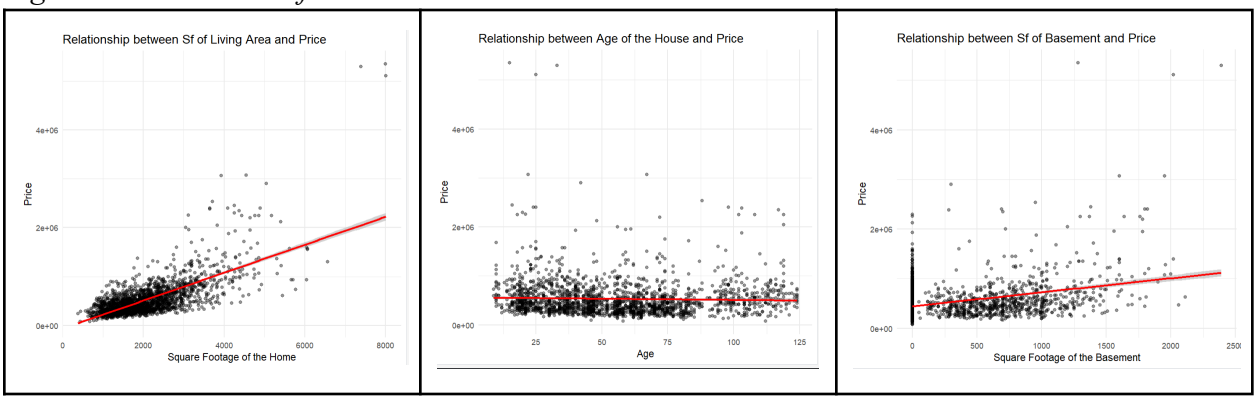
Figure 1 on the next page includes a series of boxplots to analyze the impact of various features on house prices in King County. The number of bedrooms shows a positive trend with price, with larger homes generally commanding higher prices. Grade, indicative of construction quality and design, reveals a strong positive relationship with price. Houses with more floors tend to have higher prices, although this is less pronounced. View quality also affects prices, with homes offering better views fetching higher prices. However, house condition seems to have a less clear impact on price. Together, these visualizations emphasize the significance of size, quality, and aesthetics on property values.

Figure 1. Boxplots for the Relation Between the Response Variable and Categorical Variables



The scatter plots in Figure 2 reveal the relationships between house prices and various factors in King County. The living area square footage shows a strong positive correlation with price, as illustrated by the upward trend line. Conversely, the age of the house appears to have a negligible effect on price, with the trend line being relatively flat. Square footage of the basement also displays a positive correlation with price, although not as pronounced as living area square footage. These visuals highlight that larger living areas and basements tend to increase house prices, while the age of the house is not a decisive factor.

Figure 2. Scatter Plots for Numerical variables and Price



2.2 Feature Transformation and Engineering

Feature transformation and engineering were applied to understand the association between covariates and the response variable, house sale prices and capture non-linear relationships and interactions between variables.

2.2.1 Feature Transformation

- The variable `year_built`, representing the initial year the house was built, was transformed to represent its age to assess its impacts on house sale price. This transformation involved computing the difference between the current year, 2024, and the year the house was built.
- The response variable `price`, denoting the price the house was sold for in US dollars, exhibited a heavily right-skewed distribution, shown in Plot 2. Furthermore, the residual plot of fitted values versus residuals suggested a non-constant variance. To reduce skewness and stabilize the variance, the response variable was log-transformed.
- The variable `view`, indicating how good the property's view is on a scale of 0 to 4, was converted to a factor to be treated as a categorical variable in the analysis.

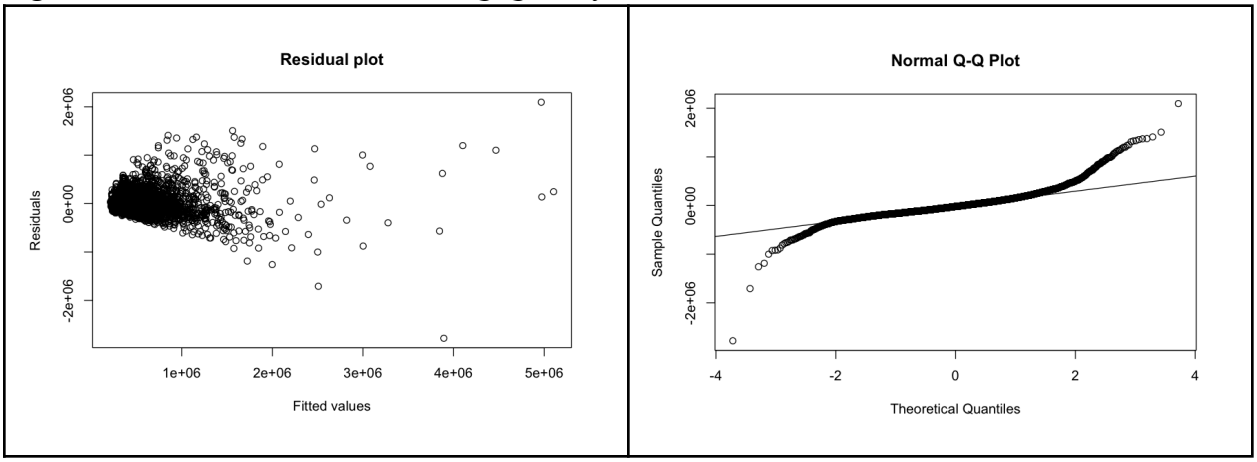
2.2.2 Feature Engineering

Figure 3 provides evidence suggesting non-linear relationships within the data. Specifically, the Q-Q plot points deviate from the expected line, with points lying greater than

expected in the upper tail and falling below expected in the lower tail. The log-transformation of the response variable was not applied in the full model.

- Power transformations of order two were applied to select variables to capture potential non-linear relationships. While higher-order transformations such as power three or square root were possible, limited computational resources restricted the transformation to order two.
- To investigate the combined influence of variables on the response variable, all possible interaction terms were generated and included in the feature selection.

Figure 3. Residual Plot and Normal Q-Q Plot for Full Model



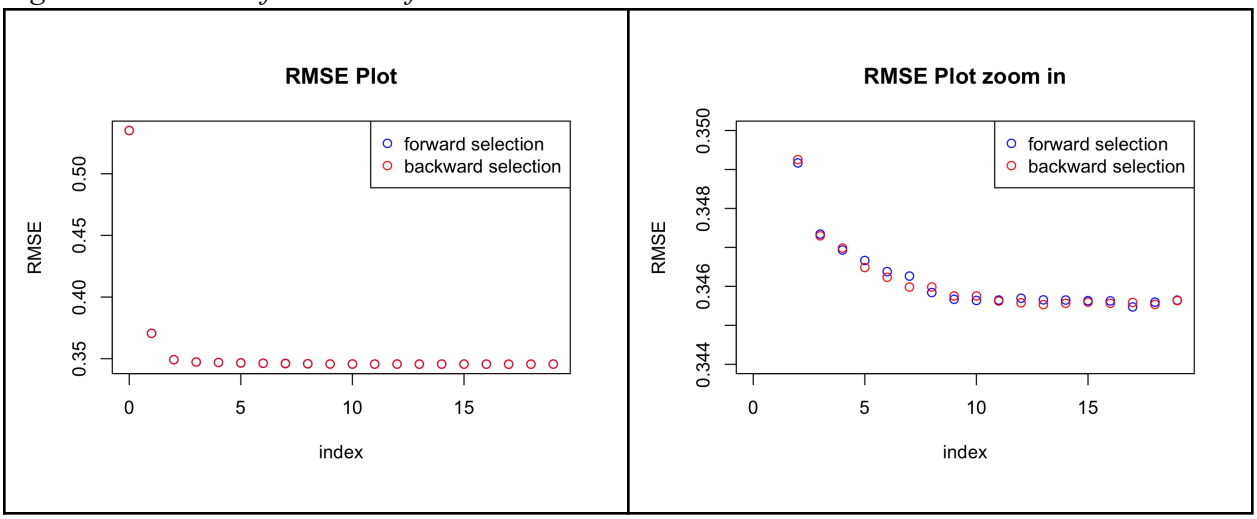
2.3 Linear Regression Modelling

Studying factors influencing house prices in King County and Seattle, this section focuses on model selection and evaluation. Feature selection is a process commonly used to reduce the number of features and decrease the computational complexity of the model. The stepwise procedures, forward and backward selection, were used to identify the most influential predictors for the linear regression model.

The iterative nature of forward and backward selection involved adding or removing features based on their impact on the model's performance. This performance was evaluated using Root Mean Squared Error (RMSE) metrics, run on a k-fold cross-validation.

Figure 4 on the next page, illustrating the RMSE changes as features were added or removed provides insight into predictor importance. Although the two methods, forward and backward selection resulted in similar RMSE performance, backward selection was chosen due to its slightly lower overall RMSE, as the smaller the RMSE, the better the model and its prediction.

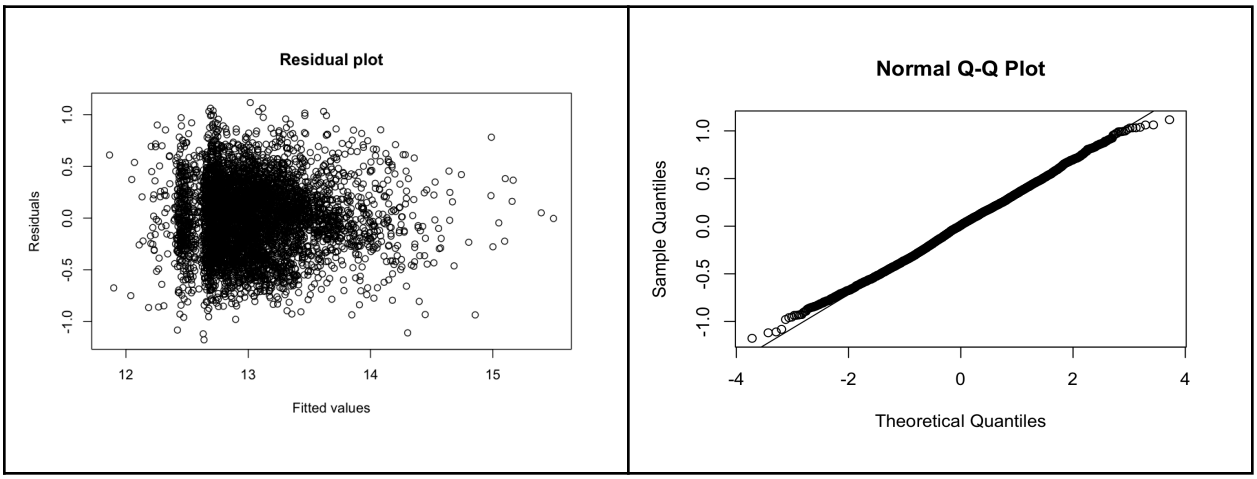
Figure 4. RMSE Performance of Forward and Backward Selection



The final model, selected through backward feature selection, comprises 25 features. With an R-squared value of 0.5824, this value suggests that approximately 58.24% of the variance in house prices can be explained by the selected explanatory variables. The residual standard error, measuring the average deviation of the observed from the predicted values was 0.347, highlighting the model's precision in capturing variability in the data. Predictions on the test set yielded an RMSE of 0.344, indicating the model's out-of-sample prediction performance.

Diagnostic plots shown in Figure 5, including the residual plot and Q-Q plot, were analyzed to assess the final model's adherence to the assumptions of linear regression modelling. The final model's residual plot displays more evenly scattered points and the Normal Q-Q plot points lie closer to the line. These observations suggest that the final model has almost constant variance and the errors follow a Normal distribution, indicating enhanced model fit and reliability.

Figure 5. Residual Plot and Normal Q-Q Plot for Final Model



Moreover, the 95% prediction interval computed suggested with 95% confidence that for a new observation with 3 bedrooms, 1180 square foot living area, the predicted house price will fall between USD 165,833.03 and USD 646,060.12.

3. Conclusion

3.1 Key Findings

The following variables collectively form the foundation of the final model

Direct Effects:

`sqft_above`
`sqft_basement`
`grade`
`view` (accounting for the four categories from `view0` to `view4`)

Interaction Terms:

```
bedrooms:sqft_basement
sqft_basement:sqft_above
bedrooms:view
view:grade
```

Quadratic Terms:

```
I((bedrooms - mean(bedrooms))^2)
I((sqft_above - mean(sqft_above))^2)
I((sqft_basement - mean(sqft_basement))^2)
```

The key finding of this project, aimed at identifying factors that predict house prices in King County, is that a model incorporating both quantitative features and their interactions—particularly square footage of living areas, grade, and waterfront presence—most effectively predicts house prices. Notably, the inclusion of interaction terms between bedrooms and basement size, and between grade and view quality, significantly enhanced the model's predictive power.

This model serves as a valuable tool for potential buyers, sellers, and real estate analysts in King County, offering a data-driven basis for making good decisions in the housing market.

3.1 Limitations and Further Improvements

A potential limitation identified throughout this report is using only RMSE to validate our model. Such a method can be limiting because it weighs larger errors more heavily. This means that outliers, which are not rare in the housing market, might skew our assessment too much. Adding additional metrics with RMSE can give us a better overview of our model's performance. For instance, adjusted R-squared is useful because it adjusts for the number of features, helping compare different models more fairly. By combining RMSE with adjusted R-squared, we can better judge our model's accuracy and its ability to handle the variety in King County's house prices, leading to a stronger model.

References

House sales in King County, USA. (2016, August 25). Kaggle. https://www.kaggle.com/datasets/harlfoxem/housesalesprediction/data

Zipcodes for King County and Surrounding Area (Shorelines) / zipcode shore area. (n.d.). https://gis-kingcounty.opendata.arcgis.com/datasets/zipcodes-for-king-county-and-surrounding-area-shorelines-zipcode-shore-area/explore