Alysa Pugmire, Angleo Orciuoli, Khalil Goddard, Maryam Ali

Group 11

Behind the Price Tag: King County Housing Market Analysis

STAT 6021 Project 2

**Analysis of Pricing for Homes in King County Washington**

This report explores the various factors that contribute to home pricing in King County Washington through data visualization and modeling. Specifically, it builds two models, one that explains how prices of homes in King County are based on other variables, and one that predicts if a home in King County is of good quality.

**Section 1**

**Section 2**

This section provides an overview of the dataset used in the analysis and a description of the variables.

[Description of data set here. Maybe include a broad picture of was the data is about as well as the number of observations and variables. Could also maybe include summary statistics about price].

The following variable were included in the original dataset:

1. id: Unique identifier for each house listing.
2. date: The date the house was sold
3. price: Sale price of the house.
4. bedrooms: Number of bedrooms.
5. bathrooms: Number of bathrooms
6. floors: Number of floors/stories.
7. sqft\_living: Total interior living space (in square feet).
8. sqft\_lot: Total area of the land lot.
9. sqft\_above: Square footage above ground (excludes basement).
10. sqft\_basement: Square footage of the basement.
11. sqft\_living15: Living space (sqft) of the 15 nearest neighbors.
12. sqft\_lot15: Lot size of the 15 nearest neighbors.
13. waterfront: if the property is on a waterfront (0 = no, 1 = yes).
14. view: Integer rating of the quality of the view on a 0–4 scale.
15. condition: Overall condition of the house on a 1–5 scale.
16. grade: Construction and design quality rating on a 1–13 scale.
17. yr\_built: Year the house was originally built.
18. yr\_renovated: Year of the last renovation (0 if never renovated).
19. zipcode: Postal code of the property.
20. lat, long: Latitude and longitude coordinates.

The following variables were created an added to the dataset

**Section 3**

This section discusses the steps taken to identify and fix any data entry errors and transform problematic variables in the dataset.

**Data Validation**

We started by identifying any observations that had clearly incorrect or suspicious values. These included homes with zero or a large number of bedrooms and bathrooms, 0 square footage, and any unusual pricing or year values.

We identified 17 properties with clearly incorrect or extreme values, including homes with 0 bedrooms, 0 bathrooms, or an unrealistic count of 33 bedrooms.

Instead of deleting these rows right away, we looked each one up manually using the King County Parcel Viewer to verify what the correct values should be. For most of them, we were able to confirm the actual number of rooms and made the necessary corrections. For instance, one home listed with 0 bedrooms and 0 bathrooms was corrected to 4 bedrooms and 4 bathrooms after verification, and the home listed with 33 bedrooms was updated to 3, which made more sense given its size. One property turned out to be a studio-style layout with 0 bedrooms and 1.5 bathrooms, so we kept that one as-is. In three cases, we couldn’t find any record of the property, and since the information couldn’t be confirmed and looked suspicious, we decided to remove those rows. Overall, this process helped us clean the dataset in a careful and meaningful way, using outside sources to make informed decisions instead of relying only on assumptions or automatic removals.

After correcting the data entry for the number of bedrooms and bathrooms, we checked for homes where the living square footage or loft square footage was 0, as a house cannot have 0 square feet of living space or land area. We did not find any observations that met these conditions.

Next we checked for home with a sale price of $0 or anything that was built or renovated after 2015, as this data set only includes homes sold in 2014 and 2015. Again, we found no observations meeting these criteria

**Variable Transformations**

After correcting any data entry error or removing uncorrectable errors from the dataset, we turned our attention to the variables. We identified five variables that could not be used in their current states:

1. Date: the date variable was stored as a character string. To fix this, we extracted the month and year from the string and stored then as integers in their own columns
2. Zipcode: there were over 70 unique values. Using this variable as is would resulted in 69 dummy variables, which would have cluttered model and increased the risk of overfitting. Instead we extracted the first three digits of the zip code and store this value in a new column which reduced the 70+ values into 2 regional clusters.
3. Renovated: the renovated variable was stored as years when the home was renovated and 0 if the home was not renovated. We changed this variable to be binary, where 0 = never renovated and 1 = renovated, for easier use in visualizations and modeling.
4. Waterfront: the waterfront variable was stored as integers where 0 = not waterfront view and 1 = waterfront view. We changed the data type to a factor for easier modeling and visualizations.
5. Latitude and longitude: these variables were stored as integers in two separate column. Instead of trying to include both latitue and longitue as separate variables, we created a new variable called distance\_to\_downtown. This variable measures how far each home is from downtown Seattle, using the coordinates of Pike Place Market (47.6062, -122.3321) as a reference point. Homes closer to downtown have smaller distance values, and those farther away have higher ones. This new variable helped capture the impact of location on things like price or home quality, in a way that’s much easier to work with than raw coordinates.

To address potential multicollinearity, we examined the relationships between four square footage variables: ‘sqft\_living’, ‘sqft\_above’, ‘sqft\_basement’, and ‘sqft\_living15’. The correlation analysis showed a strong linear relationship between ‘sqft\_living’ and ‘sqft\_above’ (r = 0.88), and a similarly high correlation with ‘sqft\_living15’ (r = 0.76). The correlation between ‘sqft\_living’ and ‘sqft\_basement’ was moderate (r = 0.43), while ‘sqft\_above’ and ‘sqft\_basement’ showed little to no correlation (r = –0.05).

Given that ‘sqft\_living’ captures the total finished living area, including both above-ground and basement spaces, retaining all three variables would introduce redundancy and likely lead to multicollinearity in the regression model. To reduce this risk and simplify the model structure, we removed both ‘sqft\_above’ and ‘sqft\_basement’, keeping only ‘sqft\_living’ as the most comprehensive and interpretable measure of living space. The variable ‘sqft\_living15’ was kept and will be evaluated for its effect on model performance later.

After transforming the variables above, we were ready to create preliminary visualizations and build our models. We started these processes by splitting the dataset with the set.seed() function into two equal subsets, one for building and training the model and the other for testing it.

**Section 4**

This section provided visualizations that show how different factors are related to the pricing of homes in King Count Washington. The visualizations were created only using the training data set.

To begin the analysis, we looked at features of the home that would potentially impact the price. We created two bar graphs that show the average price by number of bedrooms and the average price by number of bathrooms. As demonstrated is both charts, the number of bedrooms and bathrooms a home has positively impacts the price. Homes with more bedrooms and more bathrooms tend to have higher average prices that homes with less bedrooms and less bathrooms. Interestingly, the homes with the most bedrooms do not have the highest average price, which indicates that the number of bedrooms alone does not impact the price.

Next we looked at the impact square footage has on the price. There were multiple square footage variables to choose from so we opted to use sqft\_living as the square footage of a living space if often used to determine the value of the home. The scatterplot of price against living square footage also shows that a majority of the home are less than 5000 sqft and less than $2,000,000. The scatterplot also shows that there appears to be a positive, linear relationship between the square footage of the living space. As the square footage increases, so does the price. While there does not appear to be any outliers, there potentially a few influential observations.

We used side by side boxplots to compare the prices of homes that overlooked the waterfront or not. The median price of homes that overlook the waterfront is significantly higher than the median price of homes that do not, which indicates that waterfront may be a significant predictor of price. There is more variability in the prices of homes that overlook the waterfront, compared to home that do not. Additionally, there are significantly more outliers with high prices among homes that do not overlook the waterfront.

Finally, we looked at the relationships between the grade of the home and the average price, as well as, the condition of the home and the average price. Both bar graphs show a positive relationship between grade and condition, and price. As the grade increases, so does the average price. And similarly, as the condition increases, so does the average price.

After looking at aspects of the home itself that could impact the price, we turned our attention to external factor, such as when the home was built and location, which could also impact the price.

We created side by side boxplots to compare the prices of homes in different zip code groups. The median price for each group is approximately equal and the variability is also roughly the same. This indicates that zip code group is not a strong predictor of price.

Next we created density plots to see the distribution of homes according to distance to downtown and year built. The density plot of distance to downtown is right skewed, showing that a majority of the home are close to the downtown area. Additionally, the density plot of year built is left skewed, indicating majority of the homes were built after 1950.

The scatterplot of Price Against Distance to Downtown by Condition shows an overall negative relationship between price and distance to downtown. Homes that are closer to downtown ted to be more expensive. However, this relationship does not appear to be linear. Condition does not appear to have a large impact on price when paired with distance to downtown, as evidenced by the mix of conditions on the scatterplot. There seems to be an even mix of conditions among low and high priced homes.

**Section 5**

**Section 6**

**Section 7**