PHASE 2 GROUP PROJECT

START

GROUP MEMBERS

01 FIDELIS WANALWENGE

02 SYLVIA MURITHI

03 KHADIJA ALI

04 ELIZABETH NYAMBURA

PROJECT OUTLINE

Project overview

In this project, we will make use of everything we have learned about pandas, data cleaning, and exploratory data analysis. We will employ multiple linear regression modeling to analyze house sales data in the King County area. By using statistical techniques, we aim to identify key factors that impact property sales in the region and provide valuable insights to guide our recommendations.

Additionally, we will explore the potential of remodeling specific areas of a property to increase its value and make it more attractive to potential buyers. This approach can help homeowners add functionality and beauty to their property while simultaneously boosting its resale value.

Objectives

1. To create a complex model using several different independent variables that can swifty and effectively achieve pricing estimates closer to realized housing prices

2. To evaluate different models that ultimately lead to selecting our best model for predicting house prices

3. To provide insight on house features that have the biggest impact on sale price

Business Understanding

This project aims to provide valuable insights for a real estate agency operating in King County, Washington, USA. Specifically, the agency seeks to provide accurate advice to homeowners on how home renovations can potentially increase the estimated value of their properties and homes, and by how much. This information will help the agency guide their clients towards making informed decisions on home renovations, which can maximize their return on investment when selling their properties


SOLD

The problem at hand is to provide homeowners with accurate advice on how specific home renovations can impact the estimated value of their properties and homes, and the amount by which it can increase. This information is crucial for the real estate agency to guide their clients towards making informed decisions on home renovations, which, in turn, can help homeowners to maximize their return on investment when selling their properties.

Therefore, the primary objective of this project is to analyze the impact of home renovations on the estimated value of properties and provide recommendations that can help the real estate agency and their clients to make sound investment decisions.

Data Understanding

The King County House Sales dataset, available in "kc_house_data.csv," is the primary data source for this project. However, one of the main challenges we may encounter is the ambiguity or incompleteness of the column names in the dataset. Nonetheless, with thorough research and careful judgment, we can extract the necessary insights to make informed decisions about which variables to use in our analysis.

The dataset contains information on house sales in King County, including the price, design, square footage, location, and more. A comprehensive list of the column names can be found in the "Property Schema".

We will also explore the general areas in which renovations are typically undertaken in properties to identify the potential impact on the estimated value of a property.

Modelling

We first import standard packages, load the data into a pandas data frame, check the datatypes, and their descriptive statistics, and check for duplicate and missing values.

yr_renovated has the most missing values, followed by waterfront and view. waterfront and view are categorical, so we will replace those null values with 0s. I will assume null values in yr_renovated mean that the house has not been renovated and will be replaced with 0s.

The date, waterfront, view, condition, grade, and sqft_basement columns are stored as type objects, we will need to remove or convert them to numerical data type before modeling.

We then clean the dataset by removing irrelevant columns to my analysis and trimming the dataset of null values. We also remove outliers and convert the remaining categorical columns containing strings into numeric datatypes.

For this model, we create a model with all features to serve as our baseline.The dependent variable being predicted is "price," which is the house price.
The model's R-squared value (0.593) indicates how much of the variance in the target variable (house prices) is explained by the features.

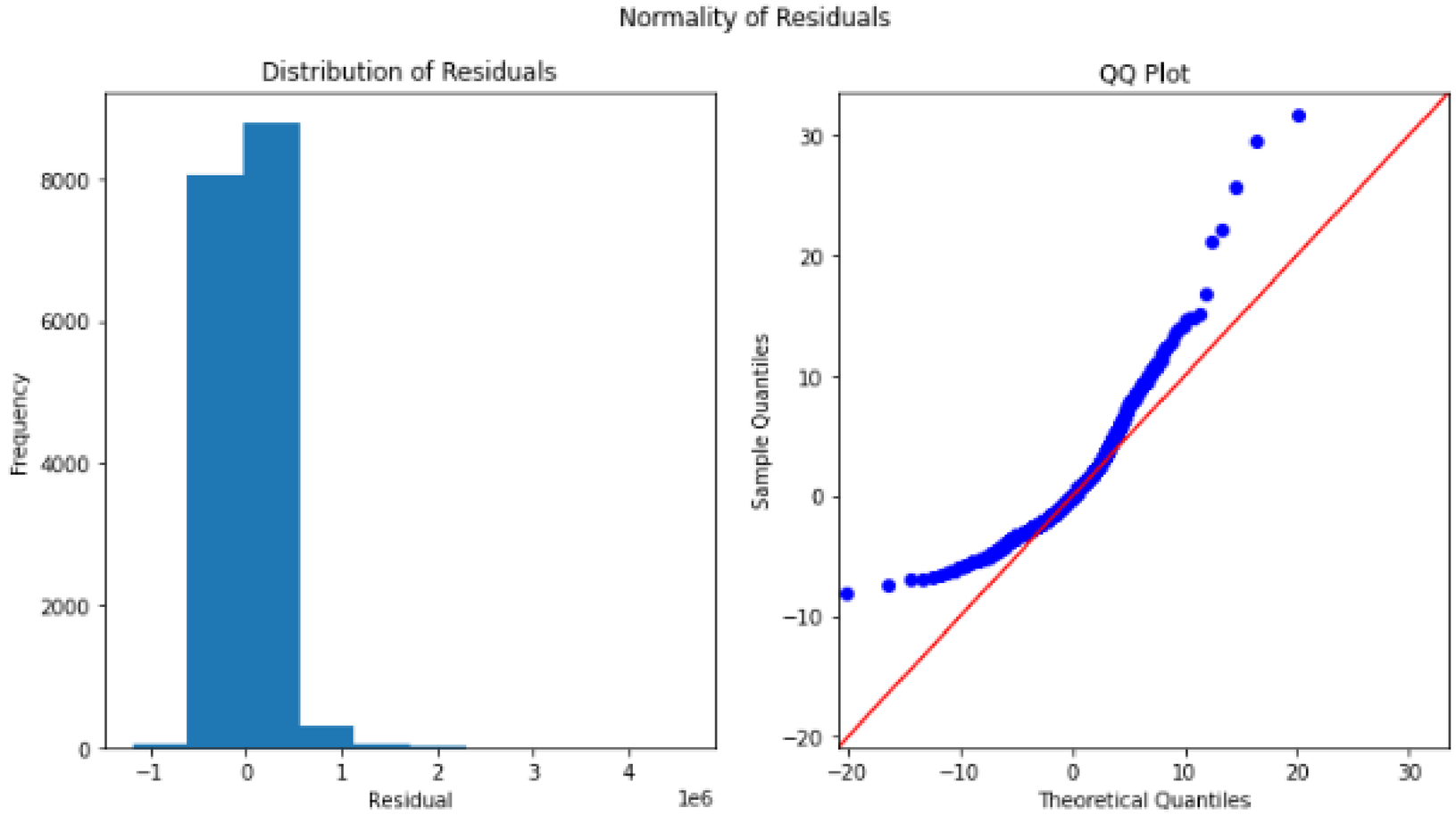
An R-squared of 0.593 is relatively good but suggests that there may still be room for improvement.

The model's adjusted R-squared is the same as the R-squared in this case, meaning there are no penalties for the inclusion of additional features.

The F-statistic (3146) and its associated p-value (0.00) suggest that the model, as a whole, is statistically significant.

Baseline Model Visualization

Assumptions Check

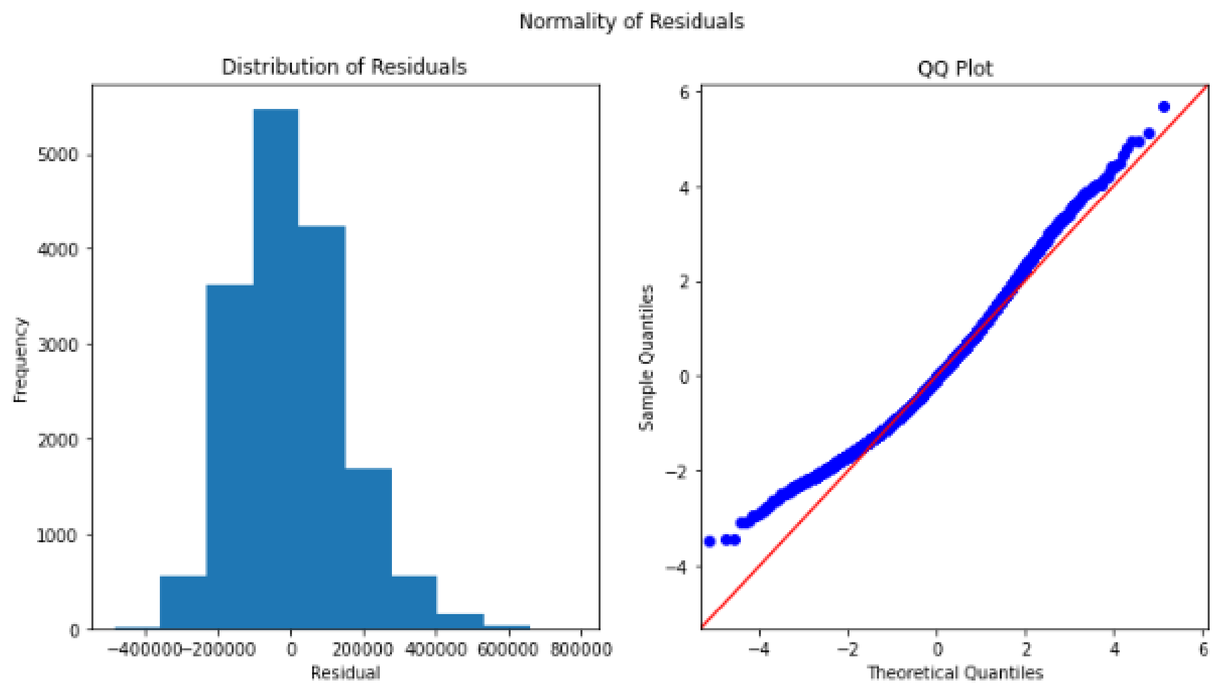


We do a first model where we remove outliers from the price. Overall, while the R-squared value has decreased slightly, the model's predictive accuracy has improved as indicated by lower RMSE values. This means that our model may provide more accurate price predictions, which can be valuable for our business stakeholders when making real estate decisions. However, it's essential to keep refining and iterating on the model to further enhance its performance.

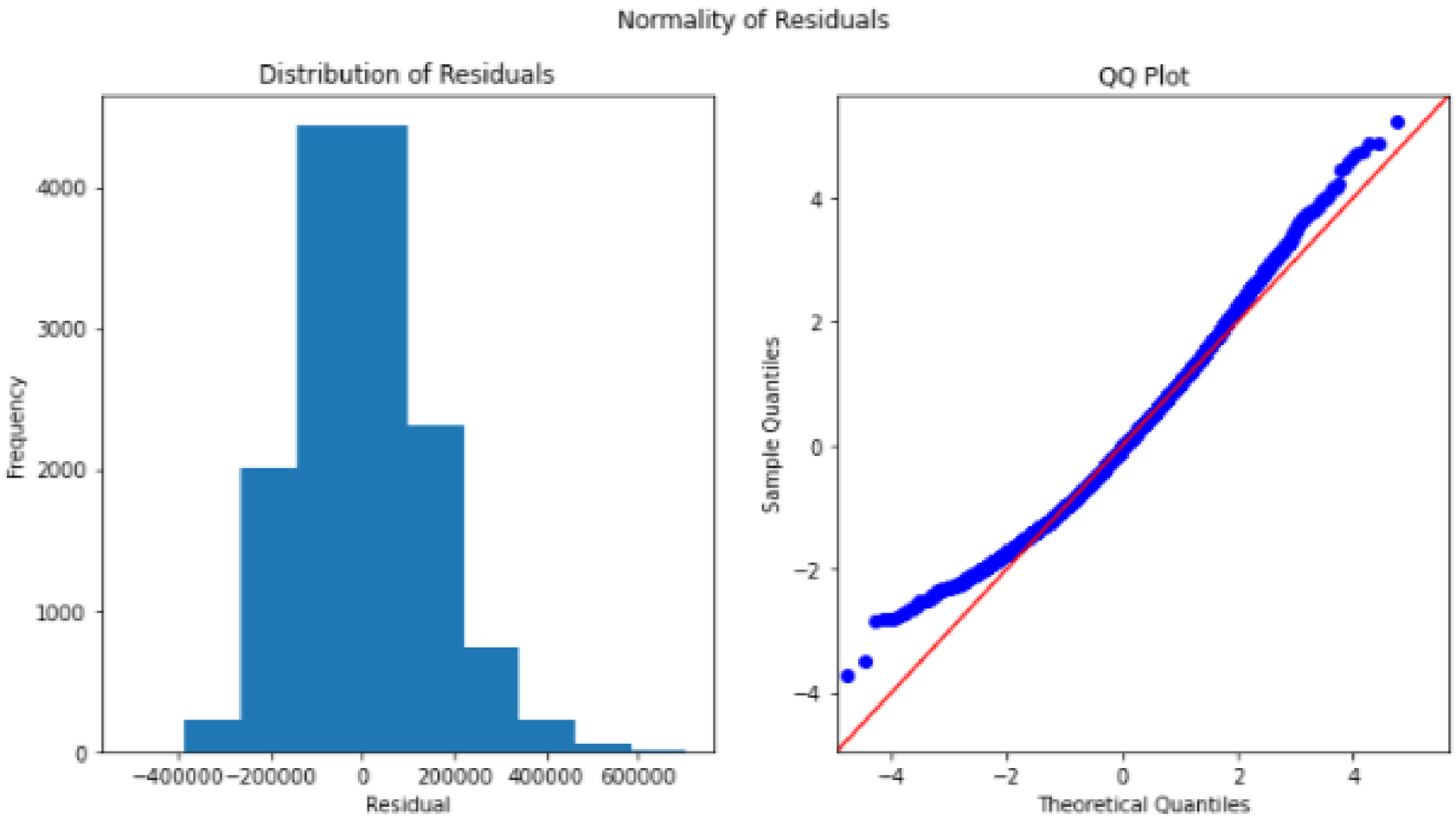
We then do a second model where we remove outliers from predictors. Overall, although the R-squared has decreased slightly, our model's predictive accuracy has improved, as evidenced by lower RMSE values. This suggests that our model may provide more accurate price predictions, which can be highly valuable for our business stakeholders in making informed real estate decisions. We should continue to monitor and refine our model to achieve the best possible results.

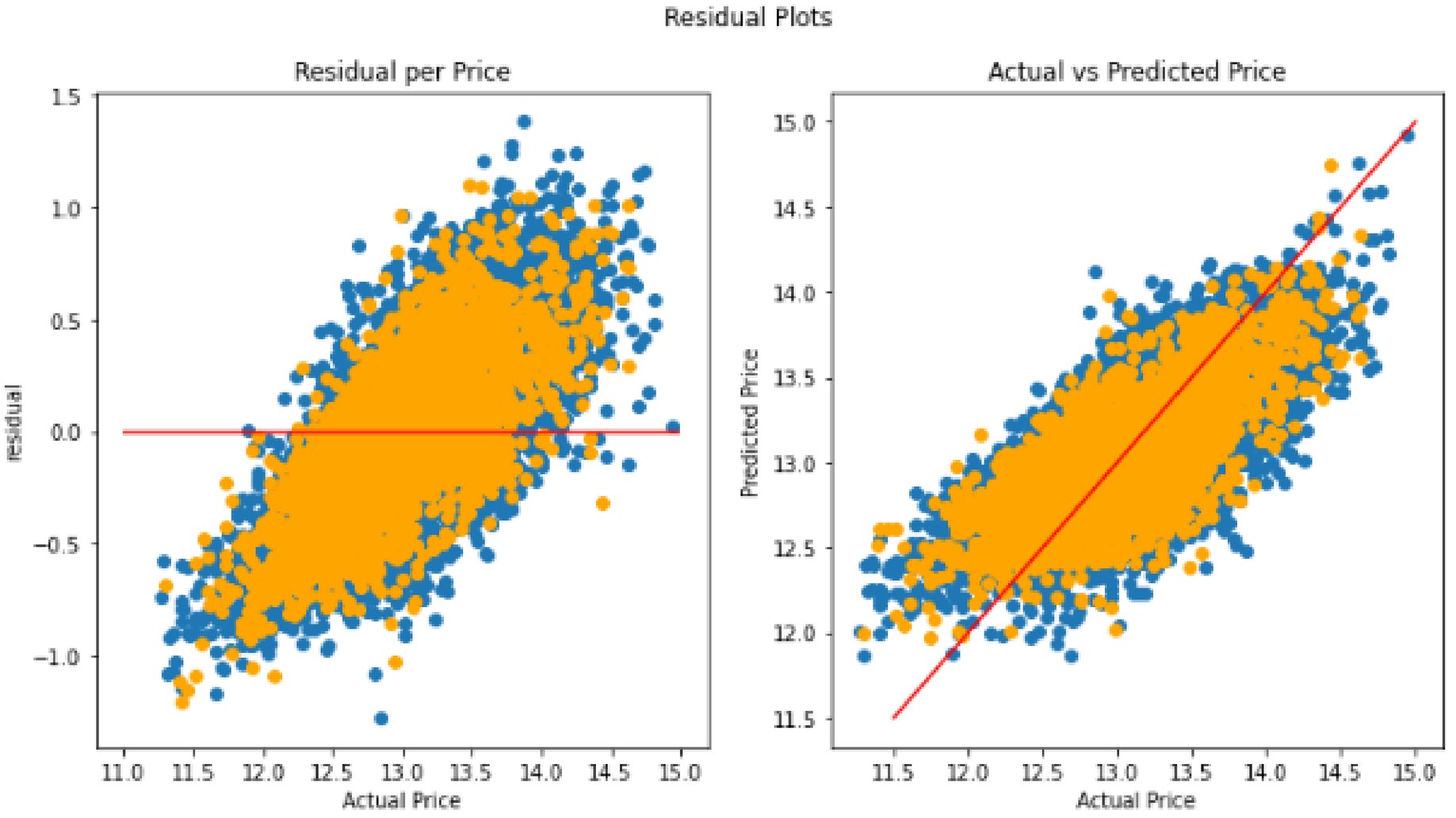
We then do a third model that involves log transformations and finally scale the final model.

Model 1 visualizations



Model 2 visualizations




Residual Plots
Residual per Price
residual
Actual Price
Actual vs Predicted Price
Predicted Price
Actual Price

Regression Results

1. In our final scaled model, we have achieved an R-squared value of 0.491, which indicates that approximately 49.1% of the variation in house prices can be explained by the selected features. This represents a slight improvement in model performance compared to previous iterations.
2. The Root Mean Squared Error (RMSE) values for both the training and test datasets have decreased as well. The training RMSE is approximately 186,658.39, and the test RMSE is approximately 182,969.03. These lower RMSE values suggest that our model's predictions are more accurate and closer to the actual sale prices of houses.
3. Among the key features, "grade" with a coeff of 0.197 and "sqft_living" with a coeff of 0.187 have the most positive impact on sale price, suggesting that investing in improving the quality of the house and increasing its living space could potentially lead to higher resale values. However, it's crucial to note that the impact of some features, such as "bathrooms" and "bedrooms," appears to be negative, which means that simply adding more of these features may not necessarily increase the resale value.

For every increase in the grade of a home, it increases the price of a home by 19.7% For every square foot of living added to a home, it increases the price of a home by 18.7% For every increase in condition to a home, it increases the price of a home by 6.9% For every floor added to a home, it increases the price of a home by 1.0% For every bathroom added to a home, it decreases the price of the home by 1.4% For every bedroom added to a home, it decreases the price of the home by 3.2%.

Conclusions for King County Real Estate Agents:

1. In order to maximize the price of a home, you should recommend to your clients that they should use great quality products when renovating their home to increase the grade of their home to the highest possible level.
2. If the seller wants to expand the size of their home, creating another floor is a great option to increase the price of their home.
3. Improving the condition of your home to a minimum, average condition will increase your home's value by 6.9%.

Limitations and Recommendations

Limitations of this model include the fact that it still relies on simplified linear relationships between features and house prices. It doesn't capture all the nuances and interactions that could exist in the real estate market. Additionally, while the R-squared value has improved, there's room for further refinement.

Stakeholders looking to remodel houses and maximize resale value should consider that this model provides valuable insights into feature importance but may not account for external factors or market dynamics that can influence pricing. To achieve the best results, they should continue to gather local market information, consult with real estate experts, and consider other factors like location and market demand when making renovation decisions. Ultimately, a holistic approach that combines data-driven insights with market expertise will lead to the most profitable post-renovation strategy.

Next Steps

1. Add more features to our model to see the effects on adjusted R-squared.

2. Create a similar tool for buyers as well that helps them decide what to offer, or what they can likely negotiate down to for a fair price.

THANK YOU