KING COUNTY **HOUSE SALES**

Final Project Submission

Student names: Aggrey Timbwa ,Alex Irungu ,Annah Mukethe, Brian Ouko and Petra Kibugu

Group: GROUP 16

Student pace: PART TIME

Scheduled project review date/time: PHASE 2

Instructor name: SAMUEL KARU

PRESENTATION OBJECTIVES

INTRODUCTION	DATA PREPARATION	MODELING	SUMMARY
Project Overview	Data inspection	Feature selection	Visualizations
Business Understanding	Data preprocessing	Model Training	Recommendation
Project Methodology	Feature engineering	Model Valuation	Conclusion

PROJECT INTRODUCTION

The real estate market is a complex and dynamic industry, heavily influenced by numerous factors ranging from location and property characteristics to economic conditions. Accurately predicting property prices is crucial for various stakeholders, including buyers, sellers, investors, and financial institutions. Leveraging historical data and machine learning techniques, we aim to build a predictive model to estimate property prices.

OVERVIEW

The real estate market is complex, influenced by location, property characteristics, and economic conditions. This project analyzes the King County House Sales dataset to understand key factors affecting house prices and to advise real estate agencies and homeowners on how renovations might impact property values.

BUSINESS PROBELM

In the competitive real estate market, stakeholders require accurate price predictions to make informed decisions. Currently, there is a lack of a robust and reliable model that can predict property prices based on historical data and property characteristics. The existing manual or heuristic methods are often inaccurate and time-consuming.

OBJECTIVES

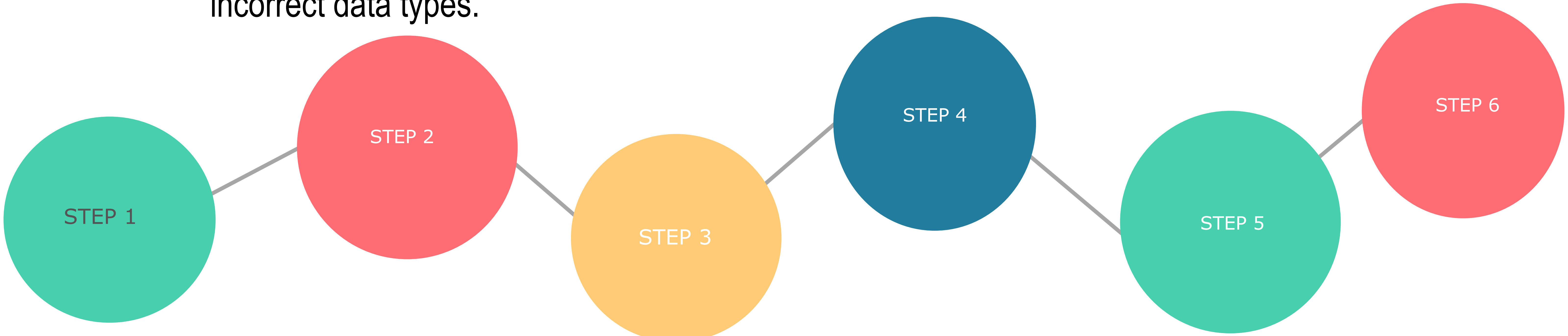
1. **Develop a Predictive Model**: Build a linear regression model to predict property prices using historical data and property characteristics.

2. **Understand Key Features**: Identify and analyze the key features that significantly impact property prices.

3. **Improve Decision Making**: Provide stakeholders with a reliable tool to estimate property prices, enhancing their decision-making process.

4. **Evaluate Model Performance**: Assess the model's accuracy and performance using appropriate evaluation metrics.

PROJECT METHODOLOGY



DATA COLLECTION AND INSPECTION

Gathering the necessary data from the provided dataset.

DATA PREPROCESSING

Cleaning the data to handle missing values, outliers, and incorrect data types.

ANALYSIS

Analyzing the data to find patterns, relationships, and insights.

MODELING

Building predictive models to estimate house prices based on selected features.

MODEL EVALUATION

Assessing the models' performance using appropriate metrics.

NEXT STEPS

Drawing conclusions from the model results and providing recommendations.

DATA UNDERSTANDING AND PREPROCESSING

GROUP 16-PT7

DATA INSPECTION AND UNDERSTANDING

Dataset Overview:

Source:

Used kc_house_data.csv file which was obtained from the King County House Sales, Washington

Content:

The dataset contains sales prices of houses in King County, Washington, along with various attributes such as the number of bedrooms, bathrooms, square footage, and more. A separate file, column_names.md, provides a description of the column names

Data Size:

The dataset has 21,597 rows and 21 columns

DATA PREPROCESSING

Handling Missing Values:

Identified missing values in waterfront column (2,376 missing), view(63) and yr_renovated (3,842 missing).
We took steps to handle missing values found in the waterfront, view, and yr_renovated columns. This included filling the missing values with the most frequent values.

Feature Engineering:

Generated new variable that take Boolean value showing whether a house was renovated or not renovated based on the column year renovated

Data Transformation:

We handled the date column which is represented as an object by converting it to a date format. Also, we transformed the sqft_basement column by converting from object to numerical value.

DATA ANALYSIS

Descriptive Statistics:

- Calculated summary statistics for the dataset columns.
- Analyzed the distribution of the kc_houses features

Correlation Analysis:

- Examined correlations between key variables such as sqft_living, bedrooms, bathrooms, and sqft_lot against the target variable which is price as well as against other features.

Multivariate Analysis:

- Conducted multivariate analysis to understand how variables like number of sqft_living, yr_built, floors, bedrooms, bathrooms, grade influence the housing price.

VISUALIZATIONS

VISUALIZATIONS

Distribution Plots:

- Used bar plot to show the count of houses renovated and comparing it to those that are not renovated
- Created bar plot to show price comparison of old houses and new houses
- Used bar plot to show the average price by renovation status (either renovated or not)
- Plotted pie chart of distribution of old houses and new houses

Features Analysis:

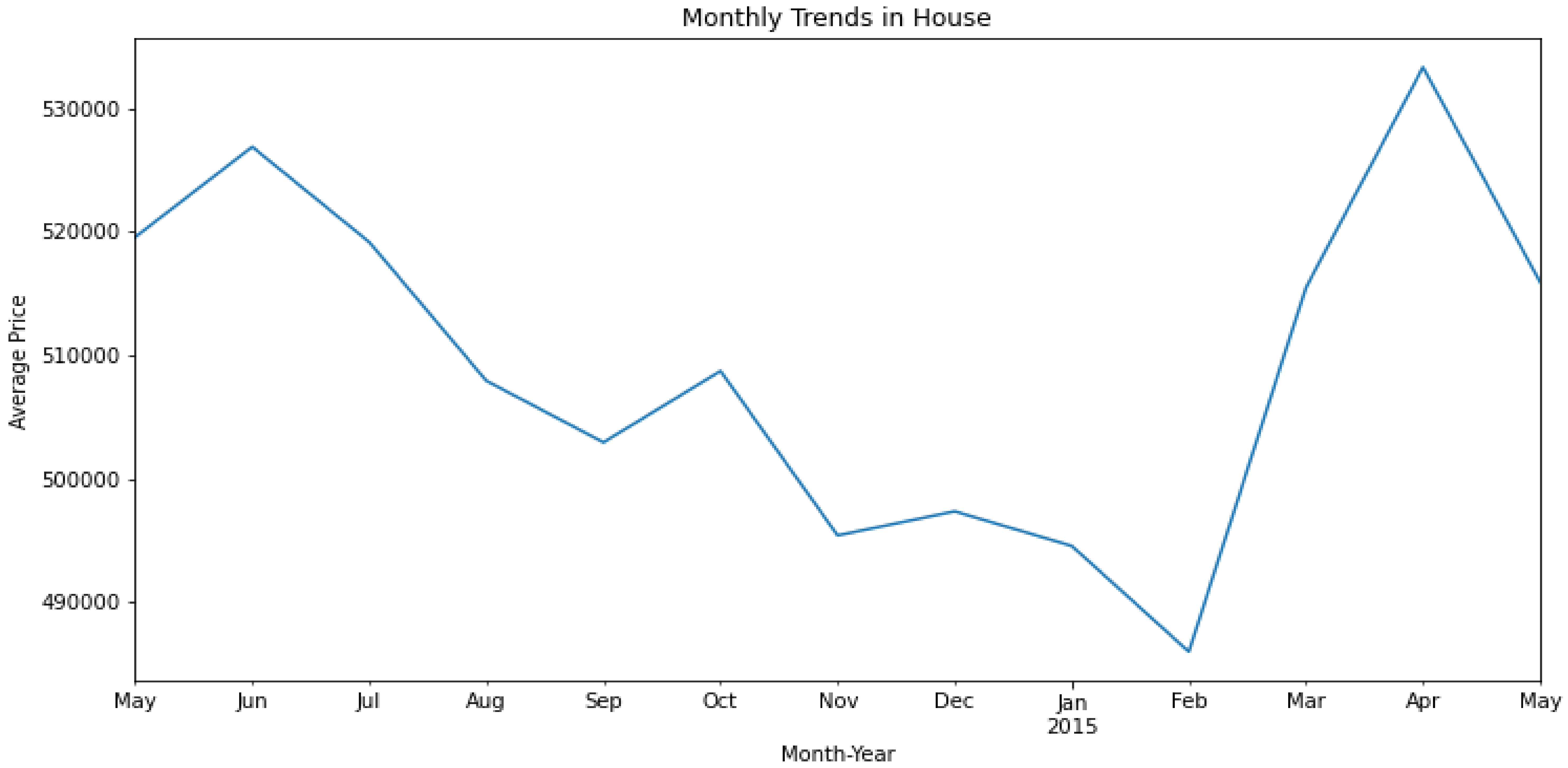
- Used pair plot to understand the correlation between house features and house price
- Also plotted heatmap to get feature that are highly correlated to our target value which was the price

Trend Analysis:

- Plotted the trend of average monthly House sale for the year 2014-2015

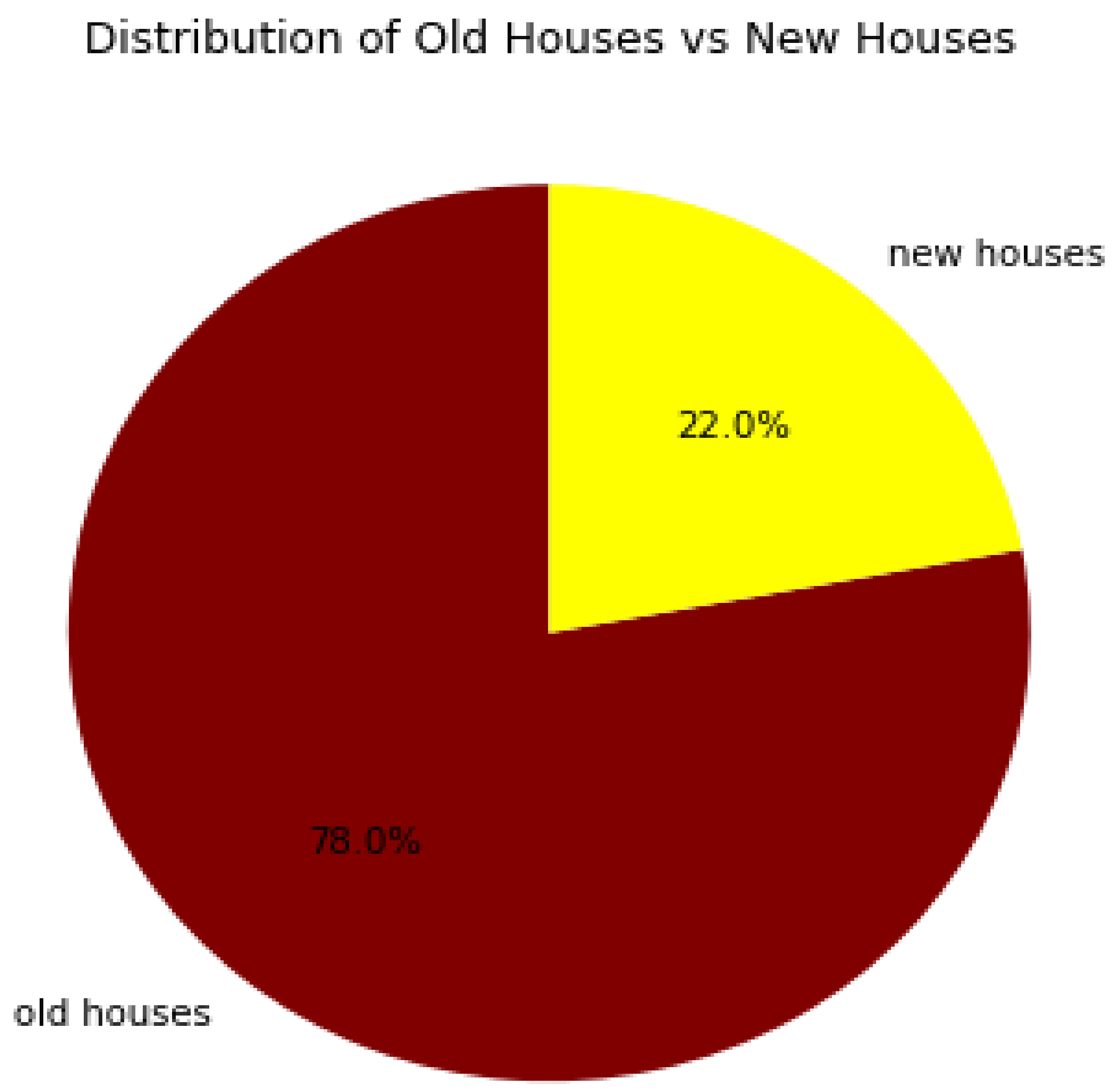
MONTHLY HOUSE SALES

From the trend shows a general decline in house prices from May to January(2014), followed by a significant increase from February to April(2015), and a slight decrease again in May the same year. This could be due to seasonal variations, market conditions, or other economic factors affecting housing prices during this period.



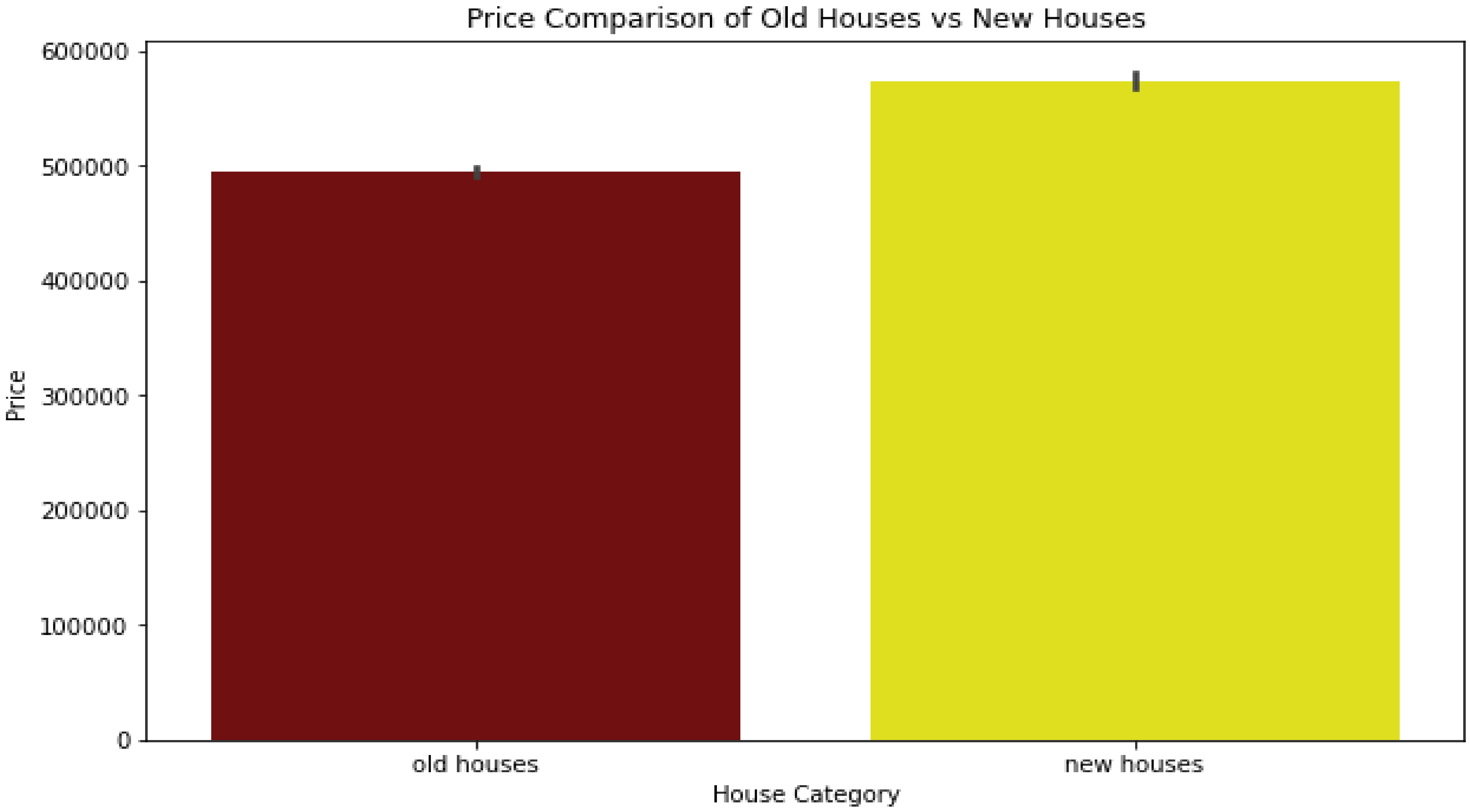
NUMBER OF OLD HOUSES VS NEW HOUSES

From the pie chart above it can be observed that the houses constructed before the year 2000 (old houses) are more as compared to those constructed after the year 2000(new houses)



PRICE COMPARISON OF OLD HOUSES VS NEW HOUSES

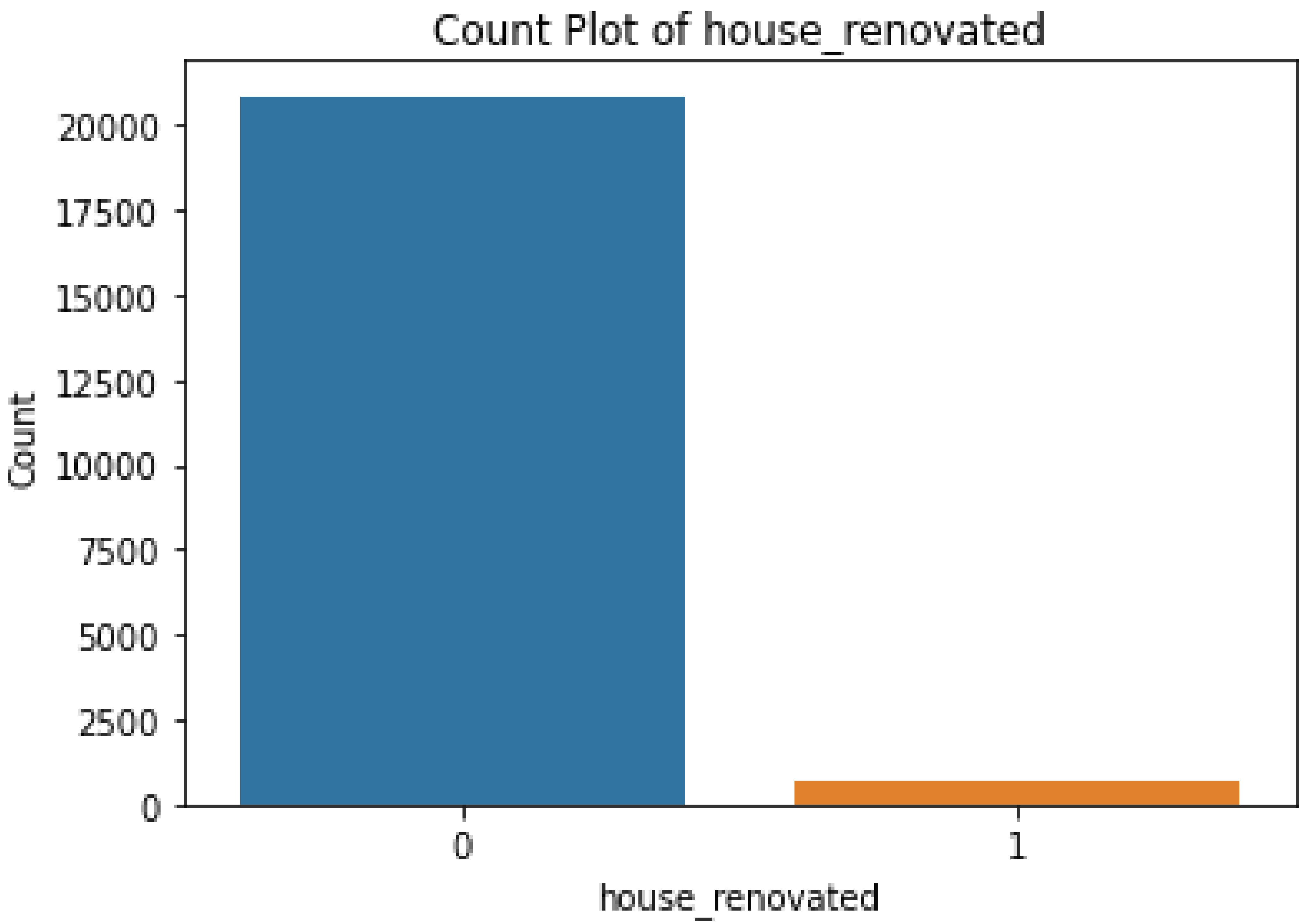
In comparison to the average prices of the old and new houses, it can be seen that the new houses cost more as compared to the old houses despite the new houses being fewer in number.



NUMBER OF RENOVATED VS NON-RENOVATED HOUSES

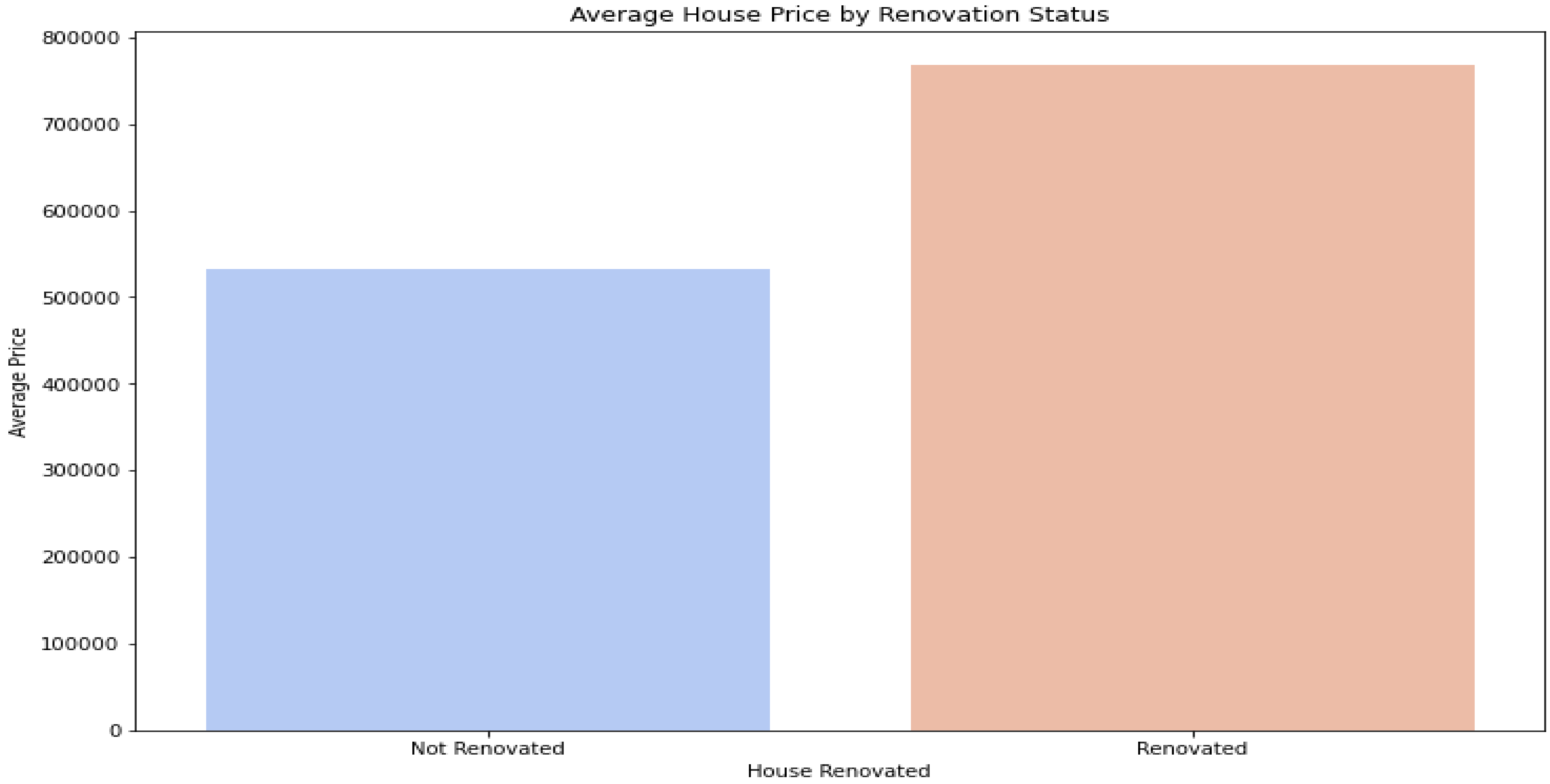
INSIGHT

From the adjacent count plot , it is evident the houses renovated represented by value 1 are fewer than those that are not renovated represented by the 0 value.



AVERAGE HOUSE PRICE BY RENOVATION STATUS

From below plot, we observe that renovated houses have a higher average price as compared to the not renovated ones. This is despite the fact that renovated houses being fewer than those that are not renovated as seen in the previous plot.



MODELING

GROUP 16-PT7

MODELING APPROACH

Data Preparation

- Feature Selection

Numerical: bedrooms, bathrooms, sqft_living, view, floors, sqft_above, yr_built, lat, long

Categorical: grade, condition

Preprocessing

- Standardization: Features normalized for consistent scaling
- Encoding: One-hot encoding applied to categorical features

Data splitting

- The dataset was split into 80% training set and 20 % the test set.

Modeling Techniques

- We started of with simple linear regression using the stats model. We then used the sckit-learn linear regression for further modeling and compared the R2 values.

Model Evaluation

- We evaluated the models based on R2 and MSE metrics.
- Model performance comparisons and results interpretation.

MODEL DEVELOPMENT

1. Simple Linear Regression (OLS)

Formula: price ~ sqft_living_normalized

Results:

R-squared: 0.434

Interpretation: 43.4% of price variance explained by sqft_living

2. Multiple Linear Regression (OLS)

Formula: price ~ sqft_living + bathrooms + grade + sqft_above

Results:

R-squared: 0.544

Interpretation: 54.4% of price variance explained by selected features

3. Scikit-learn Linear Regression

Features: All selected numerical and categorical features

Data Split: 80% train, 20% test

MODELING EVALUATION AND RESULTS INTERPRETATION

Below are the metrics Used:

Mean Squared Error (MSE): Measures the average of the squares of the errors between predicted and actual values, indicating the accuracy of the predictions. Lower values indicate better model performance.

R-squared (R^2): Represents the proportion of the variance in the dependent variable that is predictable from the independent variables. Values range from 0 to 1, with higher values indicating a better fit.

Results Comparison

Model	R-squared	MSE
Simple Linear(OLS)	0.434	N/A
Multiple Linear(OLS)	0.544	N/A
Scikit-Learn	0.744	0.0595

Interpretation

Scikit-learn model outperforms OLS models

- ○ Higher R-squared: Better fit and explanatory power
- ○ Low MSE: Closer predictions to actual values

- Final model explains 74.4% of variance in house prices
- Improved predictive accuracy compared to simpler models

RECOMMEDATIONS AND NEXT STEPS

CONCLUSION

The linear regression model for predicting property prices performs well, with an MSE of 0.0595 and an R^2 score of 0.744. It effectively identifies key features impacting property prices, providing a reliable tool for informed decision-making. Below are the conclusion drawn in relation to our objectives:

1. Develop a Predictive Model:
The linear regression model was successfully developed to predict property prices. The model's performance metrics indicate that it is a reliable tool for estimating property prices. The Mean Squared Error (MSE) of 0.0595 suggests that the model's predictions are reasonably close to the actual values on average, considering the data is normalized.

2. Understand Key Features:
Through feature selection and one-hot encoding, the model includes various numerical and categorical features such as `bedrooms`, `bathrooms`, `sqft_living`, `floors`, `sqft_above`, `yr_built`, `lat`, `grade` etc. The high R^2 score (0.744) indicates that these features collectively explain approximately 74.39% of the variance in house prices, underscoring their significant impact.

3. Improve Decision Making:
The model's high R^2 score and low MSE indicate that it is a reliable and accurate tool for estimating property prices. Stakeholders can use this model to make informed decisions about property investments, pricing strategies, and market analysis. The model's ability to explain a substantial portion of the variance in house prices provides stakeholders with confidence in its predictive power.

4. Evaluate Model Performance:
The model was evaluated using Mean Squared Error (MSE) and R^2 score. The MSE of 0.0595 indicates that the predictions are reasonably close to the actual values, while the R^2 score of 0.744 suggests that the model explains about 74.39% of the variance in house prices. These metrics demonstrate that the model performs well and meets the objective of accurately predicting property prices.

RECOMMEDATIONS

Below are some of the recommendations to the real estate agents, investors and house owners:

1. Real estates agents can focus on the key features such as condition, bedrooms, house size in terms of square feet etc. that significantly impact property prices when making decisions about property investments, pricing strategies, and market analysis.

2. Advise stakeholders to prioritize investment in recently renovated properties or properties with renovation potential.

3. Work with developers and renovators to ensure a diverse portfolio of properties. Monitor market demand and adjust the mix of old and new houses accordingly.

4. Having observed that renovated houses tend to command higher prices, we recommend that homeowners looking to sell their older properties to consider investing in renovations. This can enhance the property's value and help secure a more competitive selling price.

THANK YOU