# Student Name

1. Winfred Kinya Bundi
2. Carol Mundia
3. Paul Muniu
4. Dennis Mwenda

FULL TIME HYBRID WESTLANDS

# INTRODUCTION

In the fast-paced world of real estate, agencies must provide clients with precise information. Whether they're looking to become homeowners or investors, clients rely on real estate companies for guidance on important decisions such as pricing, market trends, and property evaluations. Real estate agencies can benefit from a sophisticated regression-based tool to meet this need.

This tool uses various property variables like the number of bedrooms, year built, floor count, living area, condition, location, and amenities to accurately predict property prices. By employing regression analysis, agencies can offer clients more precise pricing estimates, leading to better-informed decisions. Ultimately, this tool aims to improve client satisfaction, streamline decision-making processes, and drive success for real estate agencies.

# BUSINESS UNDERSTANDING

The dataset offers valuable insights for real estate agencies:

**Market Trends**: Agencies can spot trends in property demand and identify up-and-coming neighborhoods with rising values.

**Property Pricing**: They can accurately price properties by understanding how features affect sale prices

**Targeted Marketing**: Insights from the data help tailor marketing to attract buyers looking for specific property types or in certain areas.

# PROBLEM STATEMENT

Real estate experts in King County need help understanding what factors influence property values and market trends. This study aims to analyze property features, locations, buyer preferences, and market changes over time. By gaining insights from this analysis, real estate professionals can make informed decisions about buying, selling, and positioning themselves in the dynamic King County market.
The goal is to provide practical advice to help them succeed in this ever-changing real estate landscape.

# OBJECTIVES

MAIN OBJECTIVE

The primary aim of this project is to develop a predictive regression model to support real estate agencies in advising clients on house prices. This model is intended to anticipate potential changes in property value based on property characteristics, furnishing clients with valuable insights to facilitate informed investment decisions.

## SPECIFIC GOALS

Identify Factors Influencing House Prices:

Analyze property features like bedrooms, bathrooms, and square footage to understand their impact on sale price.
Investigate location-related attributes such as zip codes and geographic coordinates to further understand their effect on property prices.

. Evaluate Model Performance:
- Use metrics like mean squared error, R-squared values, and residual analysis to assess the model's accuracy in predicting house prices.

. Provide Actionable Recommendations:
- Offer practical suggestions to real estate agencies to improve profitability and market presence.
- Utilize insights from the model to optimize marketing strategies and enhance decision-making processes.

# DATA UNDERSTANDING.

King County, Washington, situated in the northwest of the United States, is known for its vibrant housing market centered around Seattle. The county has experienced significant growth due to its strong economy and cultural importance, attracting a large number of residents and creating high demand for housing in both urban and suburban areas. Seattle, with its impressive skyline, is especially sought after by tech professionals and city lovers. King County's real estate market is competitive, offering a range of neighborhoods to suit different preferences, from historic areas to modern suburban developments.

| | | |
|---|---|---|
| 1 | * `date` | Date house was sold |
| 2 | * `price` | Sale price (prediction target) |
| 3 | * `bedrooms` | Number of bedrooms |
| 4 | * `bathrooms` | Number of bathrooms |
| 5 | * `sqft_living` | Square footage of living space in the home |
| 6 | * `sqft_lot` | Square footage of the lot |
| 7 | * `floors` | Number of floors (levels) in house |
| 8 | * `waterfront` | Whether the house is on a waterfront |
| 9 | * `view` | Quality of view from house |
| 10 | * `condition` | How good the overall condition of the house is... |
| 11 | * `grade` | Overall grade of the house. Related to the con... |
| 12 | * `sqft_above` | Square footage of house apart from basement |
| 13 | * `sqft_basement` | Square footage of the basement |
| 14 | * `yr_built` | Year when house was built |
| 15 | * `yr_renovated` | Year when house was renovated |
| 16 | * `zipcode` | ZIP Code used by the United States Postal Service |
| 17 | * `lat` | Latitude coordinate |
| 18 | * `long` | Longitude coordinate |

The dataset contains 21,597 entries and 21 features. Here's a brief overview of the data:

# DATA CLEANING

To streamline our analysis, we're removing some columns from the dataset. This simplifies our data and makes it easier to focus on the most important features. By dropping unnecessary columns, we reduce noise and improve the clarity of our findings.

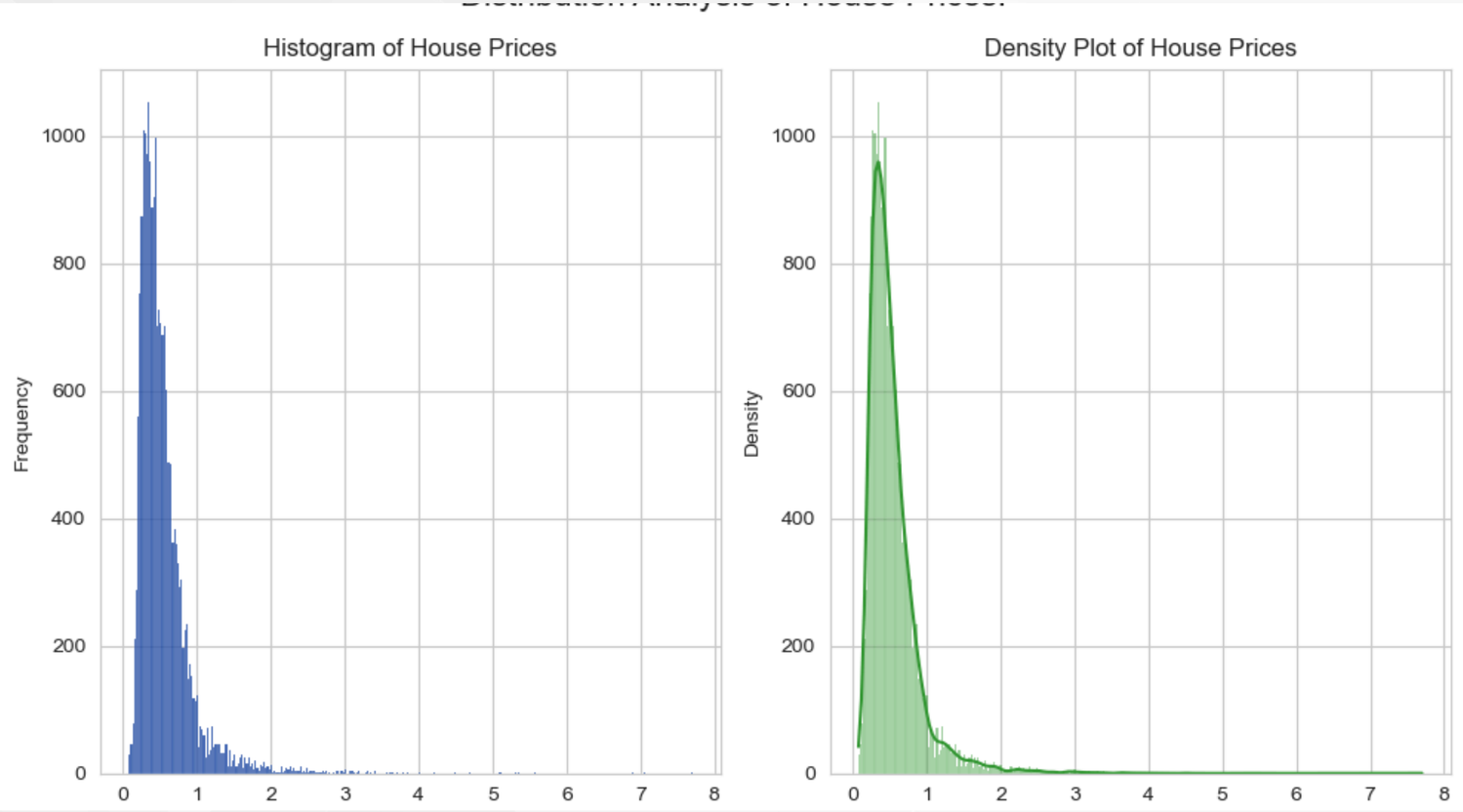
Feature Engineering:
Create additional features that might be informative for our modeling:
**House Age**: Calculate the age of the house from the 'yr_built' column to the current year.
**Renovation Age**: If a house has been renovated ('yr_renovated' > 0), calculate the years since the renovation.
**Total Square Footage**: Sum up 'sqft_living', 'sqft_lot', 'sqft_above', and 'sqft_basement' for a total square footage feature.
These new features could reveal deeper insights into the housing prices and help improve the performance of our statistical models.

# EXPLORATORY DATA ANALYSIS

This process will involve examining and understanding the structure, patterns, and relationships within the dataset. It will aid us uncover insights, detect anomalies, and inform subsequent analysis and modeling decisions.

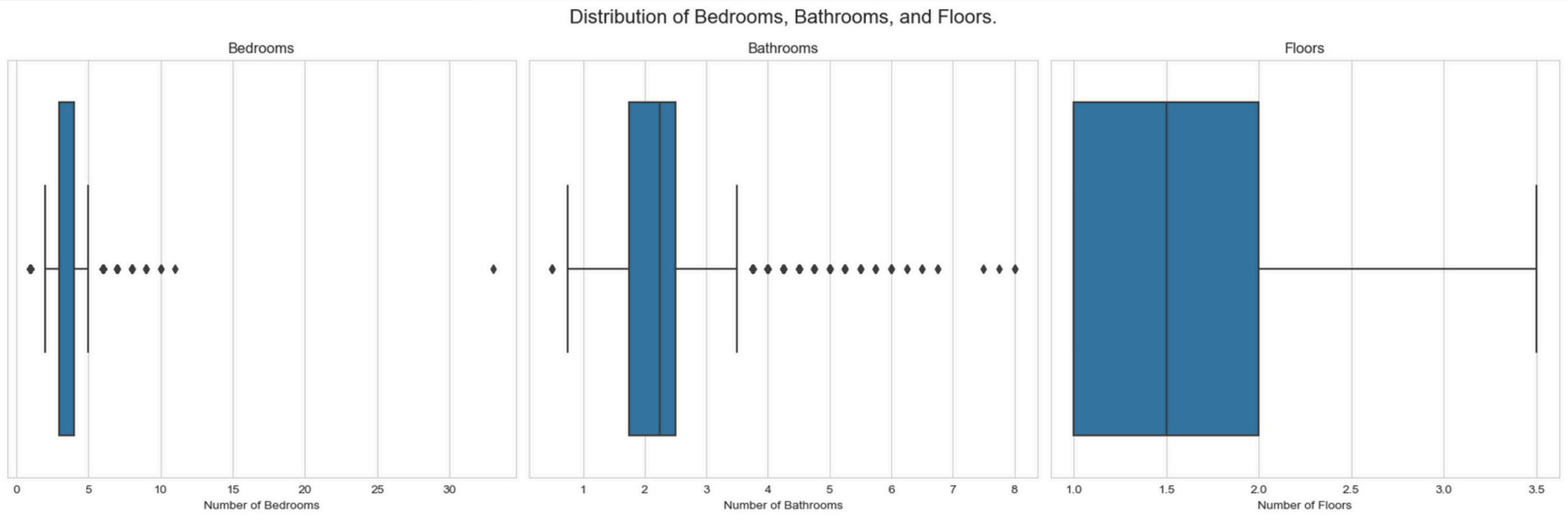
Univariate Analysis

**#Distribution of House Prices.**

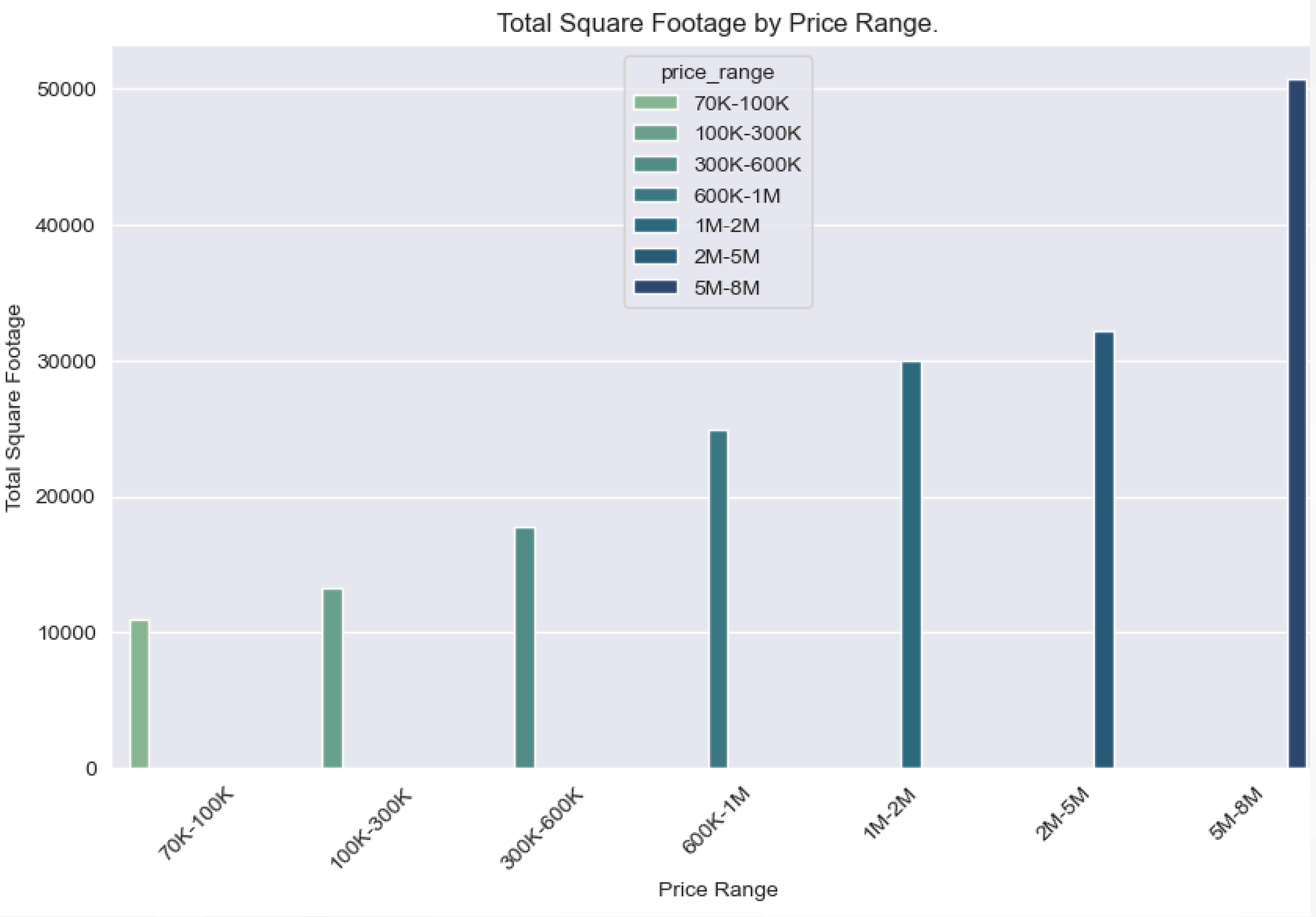



The histogram depicts the distribution of house prices, with most bars are clustered towards the left, suggesting that a significant number of houses are priced lower. The density plot illustrates a curve representing the density of house prices. Similar to the histogram, the curve peaks sharply on the left and gradually tapers off, indicating a right-skewed distribution.

In summary, the majority of house prices are concentrated at the lower end, creating a skewed distribution.

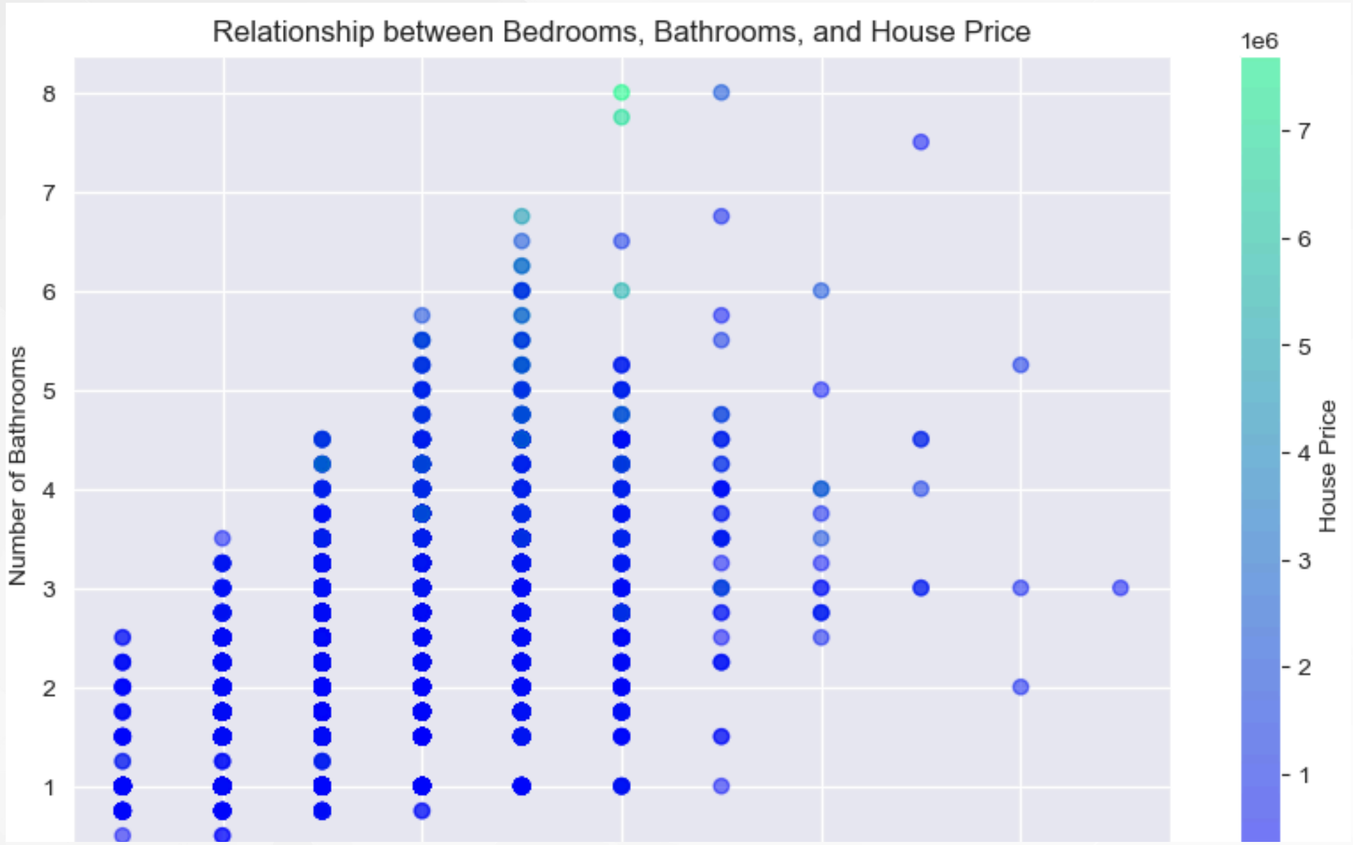



After analyzing the distribution of bedrooms, bathrooms, and floors, an outlier was detected in the bedroom column. To ensure the accuracy of our analysis, the outlier value was identified and subsequently removed from the dataset. The box plot below displays the distribution of bedrooms after excluding the outlier value.



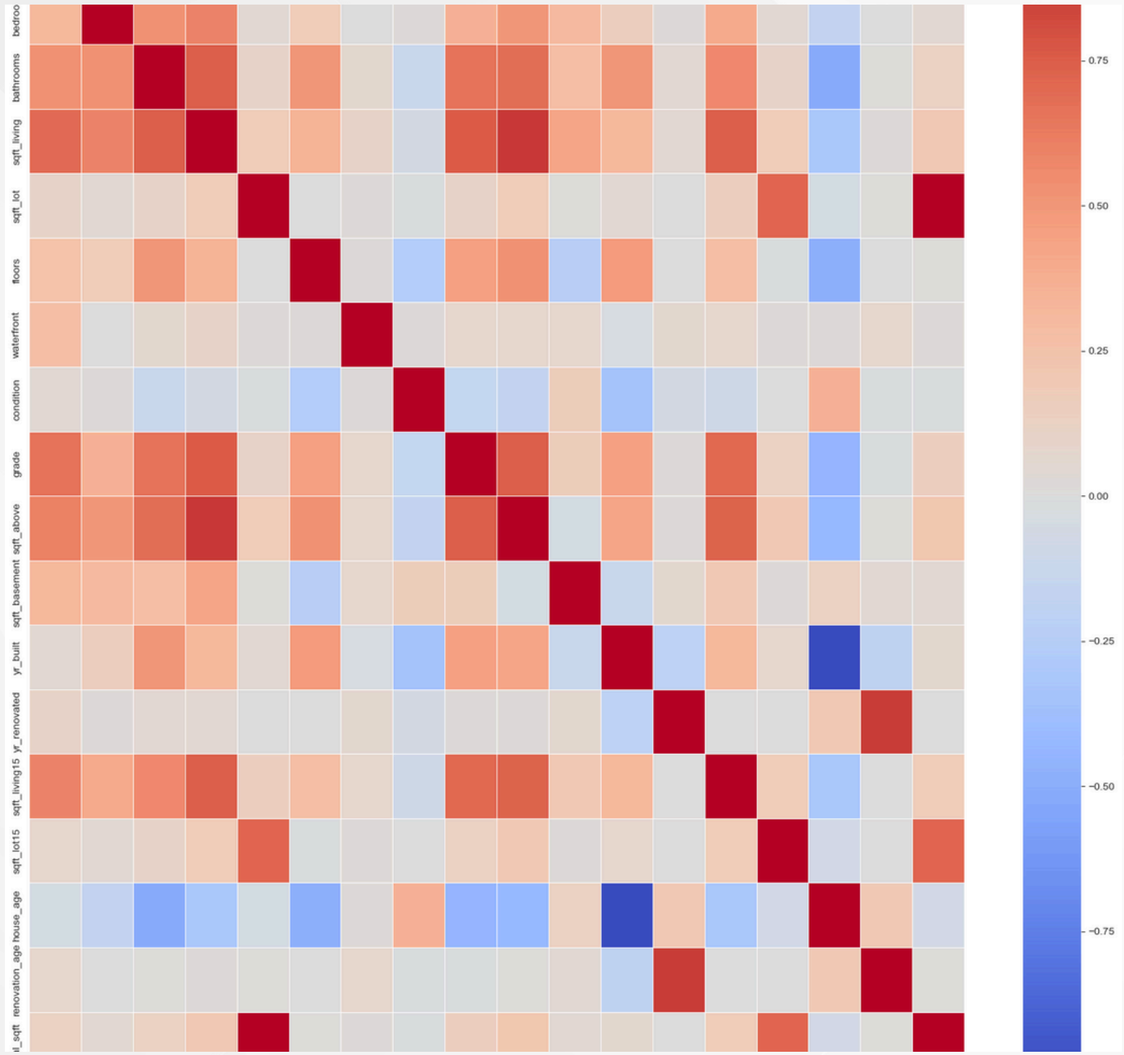

The bar plot illustrates the relationship between price range and total square footage. Each bar represents the total square footage of houses within different price range categories. From the graph, it is evident that there is a positive association between house size and price. Specifically, larger houses, as indicated by higher total square footage, tend to command higher prices.

This suggests that there is a tendency for bigger houses to have a higher price, indicating a positive correlation between the size of the property and its price.




The scatter plot illustrates a positive correlation between the number of bedrooms, bathrooms, and house prices. As the number of bedrooms and bathrooms increases, so does the price of the house. This trend suggests that buyers value properties with more space and amenities. However, there seems to be a point of diminishing returns, where additional bedrooms may not significantly increase the house's value. Understanding this relationship is vital for both sellers and buyers in the real estate market, enabling them to make informed decisions aligned with their preferences and market conditions. A well-balanced combination of bedrooms and bathrooms can enhance a property's appeal and attract a broader pool of potential buyers.

## Multivariate Analysis

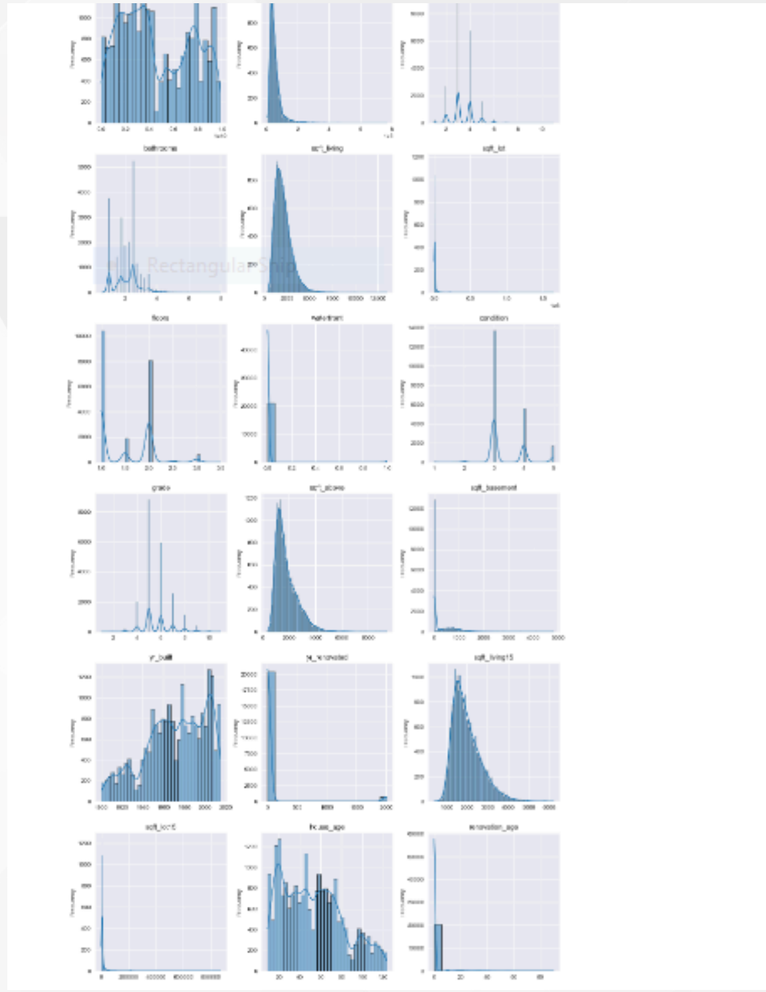



Price has a moderate positive correlation with sqft living (0.70), grade (0.67), sqft above (0.61). This means that as the values of these features increase, the price of the house also tends to increase.

# STATISTICAL ANALYSIS.

Statistical analysis is crucial for understanding datasets and predicting property values. Key steps include descriptive statistics, correlation analysis, distribution analysis, inferential statistics, and assessing multicollinearity. These steps help uncover patterns and relationships within the data, leading to more accurate regression models.

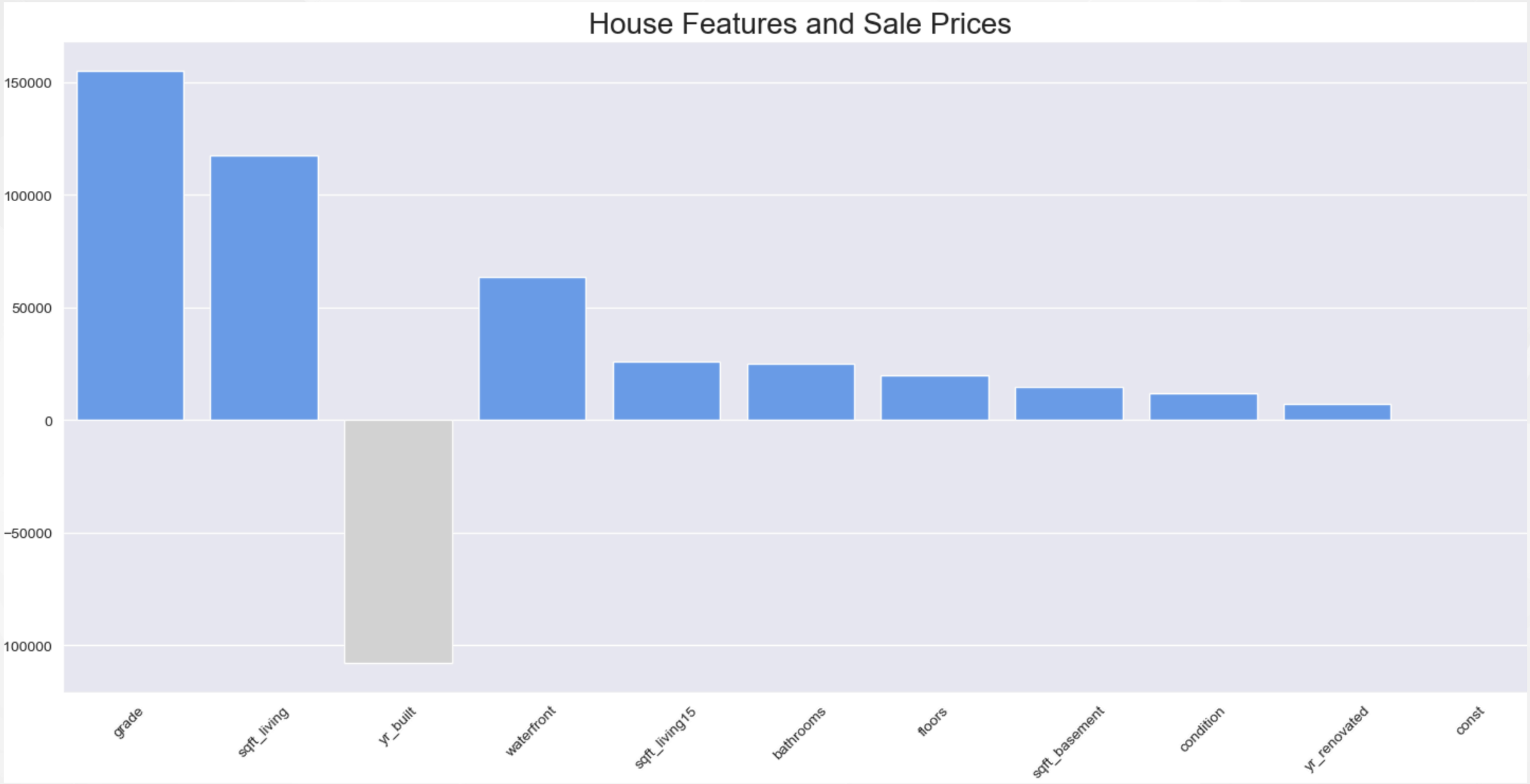
Statistical analysis is vital for predicting property values. It involves steps like looking at data patterns, checking correlations, understanding distributions, and testing hypotheses. These steps help build accurate models for predicting property prices.



-squared (0.48): Indicates that approximately 48% of the variability in house prices is explained by the square footage of living space. It measures how well the model captures patterns in the data. Mean Squared Error (MSE) (68845100756.11): This represents the average squared difference between actual and predicted house prices. Lower values indicate better accuracy, but here, the MSE is quite large, suggesting room for improvement. Root Mean Squared Error (RMSE) (262383.50): This is the square root of the MSE, providing a measure of typical deviation between predicted and actual house prices.

The RMSE is approximately 262,383.50 units. Intercept (540631.16): Estimated house price when all independent variables are zero. It's around 540,631.16 units, suggesting a baseline value. Coefficient (259767.82): This represents the change in house prices for a one-unit increase in square footage of living space, with other variables held constant. For every one-unit increase in square footage, house prices are expected to increase by approximately 259,767.82 units.

## MULTIPLE LINEAR REGRESION



| | | | |
|---|---|---|---|
| Dep. Variable: | price | R-squared: | 0.641 |
| Model: | OLS | Adj. R-squared: | 0.641 |
| Method: | Least Squares | F-statistic: | 3768. |
| Date: | Wed, 01 May 2024 | Prob (F-statistic): | 0.00 |
| Time: | 13:34:57 | Log-Likelihood: | -2.9014e+05 |
| No. Observations: | 21142 | AIC: | 5.803e+05 |
| Df Residuals: | 21131 | BIC: | 5.804e+05 |
| Df Model: | 10 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 6.604e+06 | 1.41e+05 | 46.979 | 0.000 | 6.33e+06 | 6.88e+06 |
| bathrooms | 3.47e+04 | 3547.740 | 9.781 | 0.000 | 2.77e+04 | 4.17e+04 |
| sqft_living | 129.1822 | 3.765 | 34.313 | 0.000 | 121.803 | 136.562 |
| floors | 3.576e+04 | 3886.909 | 9.201 | 0.000 | 2.81e+04 | 4.34e+04 |
| waterfront | 7.828e+05 | 1.88e+04 | 41.703 | 0.000 | 7.46e+05 | 8.2e+05 |
| condition | 1.788e+04 | 2568.771 | 6.961 | 0.000 | 1.28e+04 | 2.29e+04 |
| grade | 1.319e+05 | 2290.923 | 57.586 | 0.000 | 1.27e+05 | 1.36e+05 |
| sqft_basement | 27.2065 | 4.581 | 5.939 | 0.000 | 18.228 | 36.185 |
| yr_built | -3728.7602 | 71.951 | -51.823 | 0.000 | -3869.790 | -3587.730 |
| yr_renovated | 18.3132 | 4.407 | 4.156 | 0.000 | 9.676 | 26.950 |
| sqft_living15 | 34.6185 | 3.669 | 9.435 | 0.000 | 27.427 | 41.810 |

| | | | |
|---|---|---|---|
| Omnibus: | 16539.863 | Durbin-Watson: | 1.978 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 1296266.479 |
| Skew: | 3.184 | Prob(JB): | 0.00 |
| Kurtosis: | 40.828 | Cond. No. | 3.35e+05 |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.35e+05. This might indicate that there are strong multicollinearity or other numerical problems.

R-squared ($R^2$): The coefficient of determination is 0.641, indicating that approximately 64.1% of the variance in the housing prices is explained by the independent variables included in the model.

Adjusted R-squared: The adjusted R-squared is also 0.641, which adjusts for the number of predictors in the model. It's useful when comparing models with different numbers of predictors.

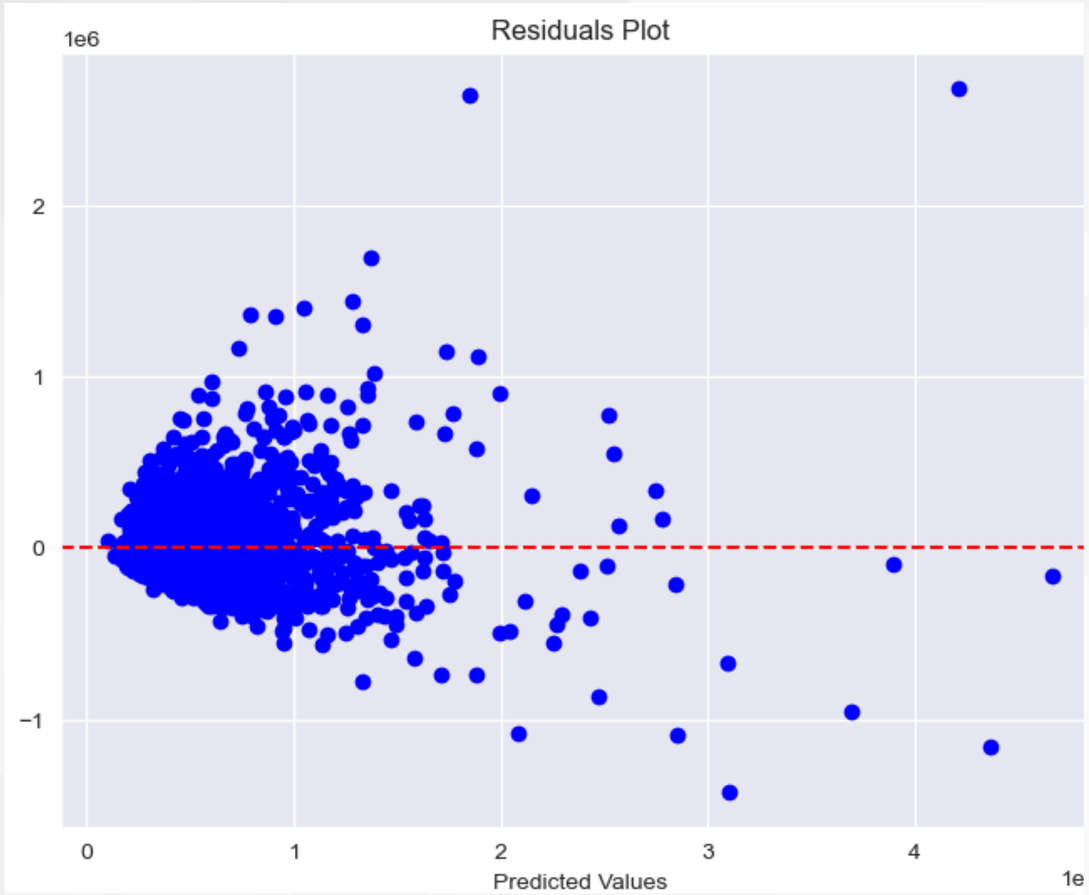
F-statistic: The F-statistic is 3768, with a p-value close to zero, indicating that the overall model is statistically significant.

# POLYNOMIAL REGRESSION

Polynomial Model (Degree 2)- MSE: 35637716070.71762
Polynomial Model (Degree 2)- R-squared: 0.7321925161881991

- A polynomial model of degree 2 was evaluated and compared to a previous, presumably simpler, model.
- The polynomial model demonstrated better performance based on two metrics:
- Mean Squared Error (MSE): Lower in the polynomial model, indicating a closer fit between predicted and actual house prices.
- R-squared ($R^2$): Higher in the polynomial model (0.732 vs. the previous model's 0.641), suggesting it explains a larger portion of the variance (spread) in housing prices.

# RESIDUAL PLOT



## REGRESSION RESULTS

From all the models observed,We can see the positive correlation between the house prices and all other features except the year built which indicates a negative correlation.

Polynomial Regression is the preferred model beacuse from the evaluation it has the highest R-squared value of 0.73

The features below impact price such that an increase will cause an increase in the price of the property. 'bedrooms','bathrooms', 'sqft_living','floors', 'waterfront','view''condition', 'grade', 'sqft_above','sqft_basement', 'yr_renovated', 'sqft_living15','renovated', 'basement'.

# LIMITATIONS

- Limited Data
- :The model might not consider all relevant property characteristics. Missing factors could lead to inaccurate predictions.
- 
- Multicollinearity:Correlated features like square footage and number of bedrooms can create multicollinearity. This makes it difficult to understand the independent effect of each feature on the price, reducing model reliability.
- 
- Assumption Violations: The model assumes a linear relationship between features and price. Real-world relationships may not be linear, leading to biased estimates and unreliable predictions. Additionally, heteroscedasticity (unequal variance of errors) could be an issue.
- 
- Overfitting Risk: Polynomial models, especially with high degrees, are susceptible to overfitting. This means they capture random noise instead of true patterns, leading to poor performance on new data.

# RECOMMEDATIONS

1.Invest in Larger Properties: Investors seeking maximum returns should focus on larger houses, as there's a positive correlation between total square footage and price. Such properties have the potential for higher profits upon resale or rental.

2.Upgrade Existing Properties: Homeowners can increase their property's value by investing in upgrades that increase square footage, such as adding extra rooms or expanding living spaces.

3.Optimize Bedroom and Bathroom Ratios: It's essential to find the right balance between bedrooms and bathrooms to maximize property value. Consulting with real estate professionals can help determine the optimal ratio based on market trends and buyer preferences.

# RECOMMEDATIONS

4. Focus on Quality Over Quantity: Prioritize quality improvements that enhance functionality and aesthetics, such as renovating bathrooms with modern fixtures or upgrading kitchen appliances, to add perceived value to the property.

5. Highlight Features in Listings: Emphasize the number of bedrooms and bathrooms in property listings to attract buyers who prioritize space and convenience. Highlight unique features that add versatility to the property.

6. Differentiate Marketing Strategies: Tailor marketing strategies based on property condition and grade ratings. Highlight the benefits of higher-grade properties to attract premium buyers, while emphasizing renovation potential for properties with lower condition ratings.

# CONCLUSION

- Property Size Matters: There is a clear positive correlation between the size of a property, indicated by total square footage, and its price. Investing in larger properties can potentially yield higher returns for investors and increase market value for homeowners.
- Strategic Upgrades Add Value: Upgrading existing properties with strategic renovations and expansions, particularly those that increase square footage, can enhance their market value. Quality improvements that improve functionality and aesthetics are key to maximizing property value.
- Balance is Key: While adding extra bedrooms and bathrooms can increase a property's price, there's a point of diminishing returns. It's important to strike a balance between quantity and quality, optimizing the bedroom-to-bathroom ratio to align with market trends and buyer preferences.
- Marketing Differentiation is Essential: Tailoring marketing strategies based on property condition and grade ratings is crucial. Highlighting the unique features and benefits of higher-grade properties can attract premium buyers while emphasizing renovation potential for properties with lower ratings can appeal to savvy investors.


"Every project is an opportunity to learn, grow, and innovate. Embrace the challenges, celebrate the victories, and never underestimate the power of collaboration.

# THANK YOU

https://github.com/winnycodegurl/dsc-phase-2-projectgroup4/blob/main/student.ipynb