



PROPERTY PRICE PREDICTION TO AGENCIES

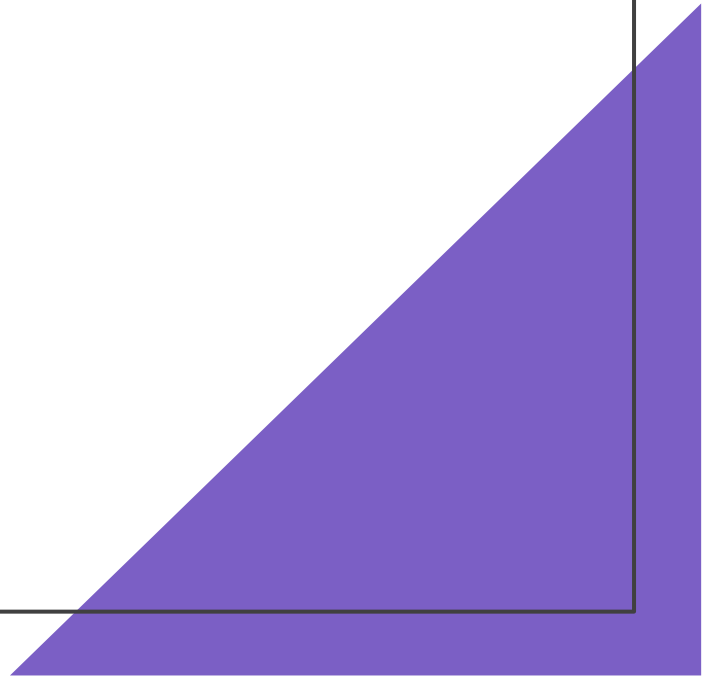
GROUP 6

1. Nicole Bosibori
2. Monica Mwangi
3. Jane Njuguna
4. Shilton Soi
5. Loise Mburuga



OUTLINE

1. Project Overview
2. Objectives
3. Data Understanding
4. Data cleaning and preparation
5. Exploratory data analysis –Bivariate and Univariate analysis
6. Statistical Analysis
7. Modelling
8. Regression Results
9. Conclusion
10. Recommendations



PROJECT OVERVIEW



What is the problem

- A real estate agency in King County seeks to provide advice to homeowners on how renovations could increase estimated value of their homes but lacks data.

How to solve the problem

- Construct a predictive regression model that aids real estate agencies in empowering clients on making informed decisions on property prices



OBJECTIVES

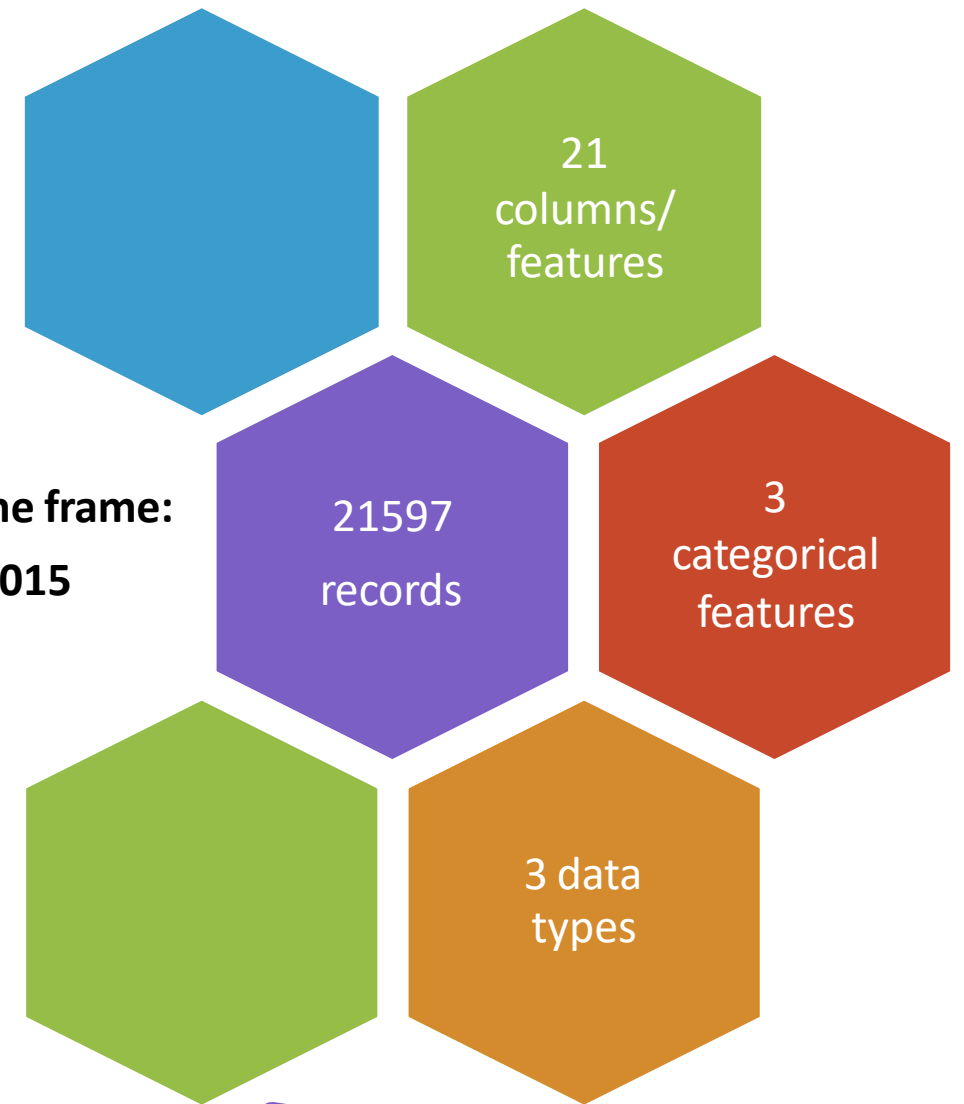
1. Identify Key Factors Influencing House Prices
2. Effectiveness ways to predicting house prices.
3. Provide suggestions to real estate agencies for enhancing profitability and market presence.

DATA UNDERSTANDING

This project uses the north western county dataset.

- Integers which include;
id, bedrooms, sqft_living, sqft_lot, sqft_above, yr built , zipcode, sqft living15, sqft lot15
- Float data types include;
price, bathrooms, floors, year renovated, latitudes and longitudes
- Object data type include the columns;
date, waterfront, view, condition, grade and sqft basement.

Data time frame:
1900 - 2015



DATA CLEANING & PREPARATION

- ☐ Dropping irrelevant columns; id, date, longitude, latitude, zip code and waterfront.
- ☐ Check and drop duplicates in the 'id' column.
- ☐ Identify and handle missing values.
- ☐ Check for place holders.
- ☐ Convert data types if necessary.
- ☐ Identify outliers.
- ☐ Feature Engineering by creating new columns i.e 'renovated'.

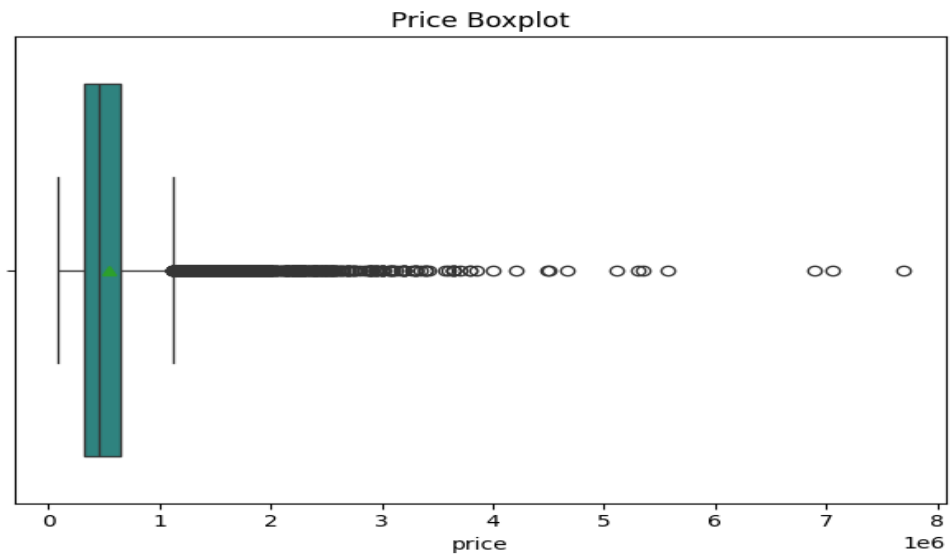
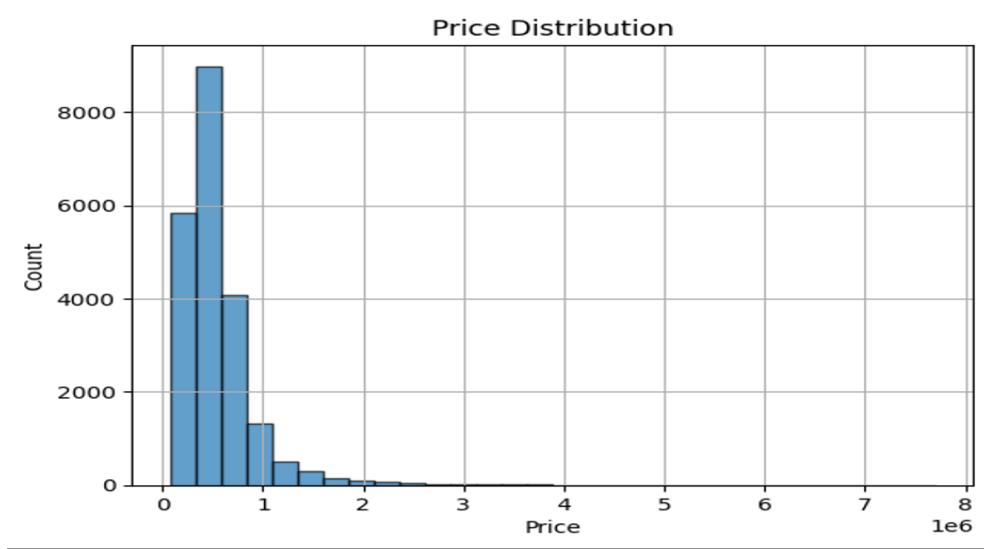




Exploratory Data Analysis

- Exploratory Data Analysis (EDA) is a crucial step in data analysis where the main goal is to understand the characteristics of the data at hand.
- Depending on the number of variables being analyzed simultaneously, EDA can be classified into three main types:
 - 1.Univariate
 - 2.Bivariate
 - 3.Multivariate

UNIVARIATE ANALYSIS – Price Distribution



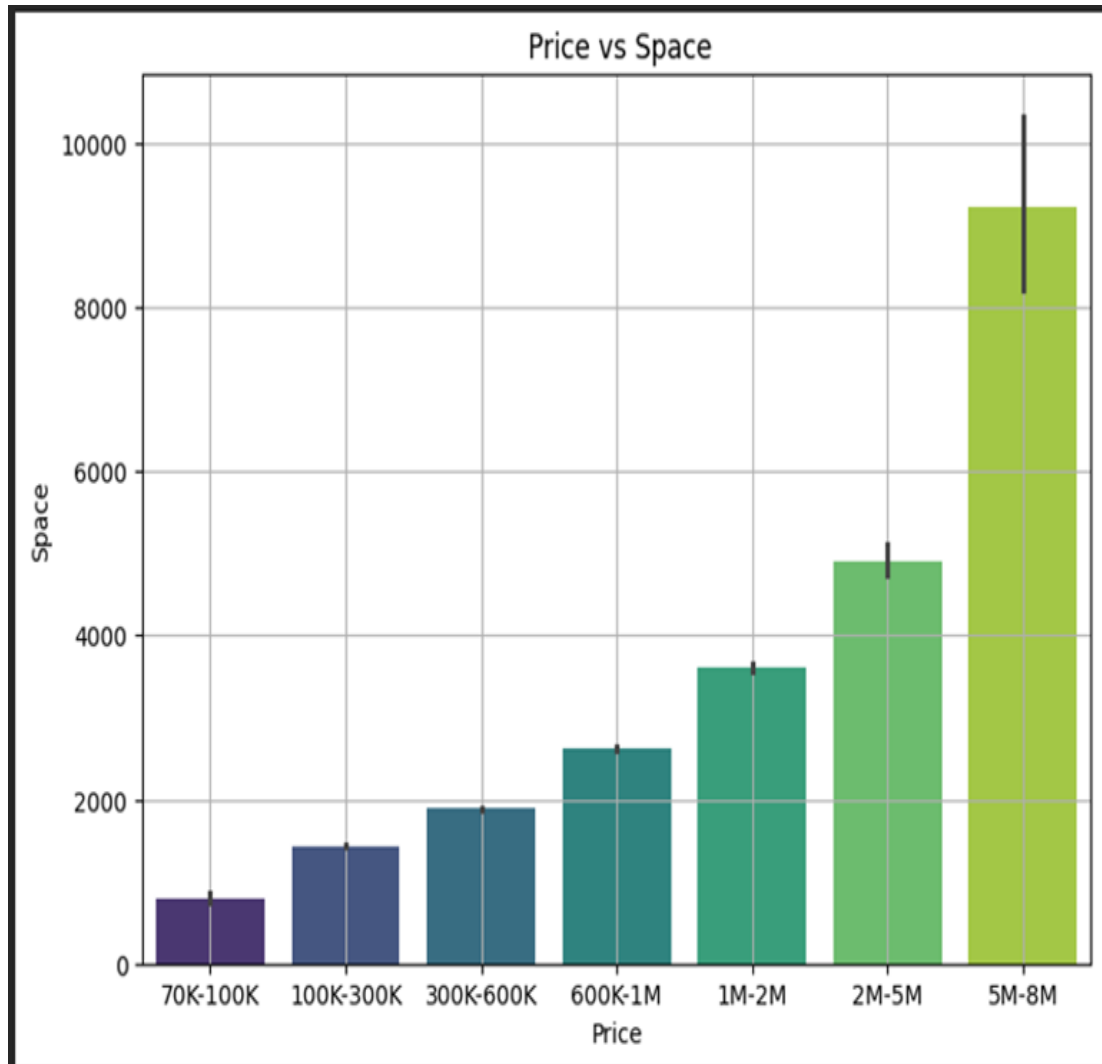
- This graph and boxplot shows the price distribution of houses and shows most houses cost less than 1000000

BIVARIATE ANALYSIS – Price vs Rooms



- According to this graph, house prices increases with number of rooms, therefore houses with 5 or more rooms sell at around 5,000,000 while those with 2 or less rooms sell at around 70,000.
- We can also conclude that most houses have less than 4 bedrooms and the ones that have 4 or more bedrooms cost more than 1,000,000

BIVARIATE ANALYSIS –Price vs Space



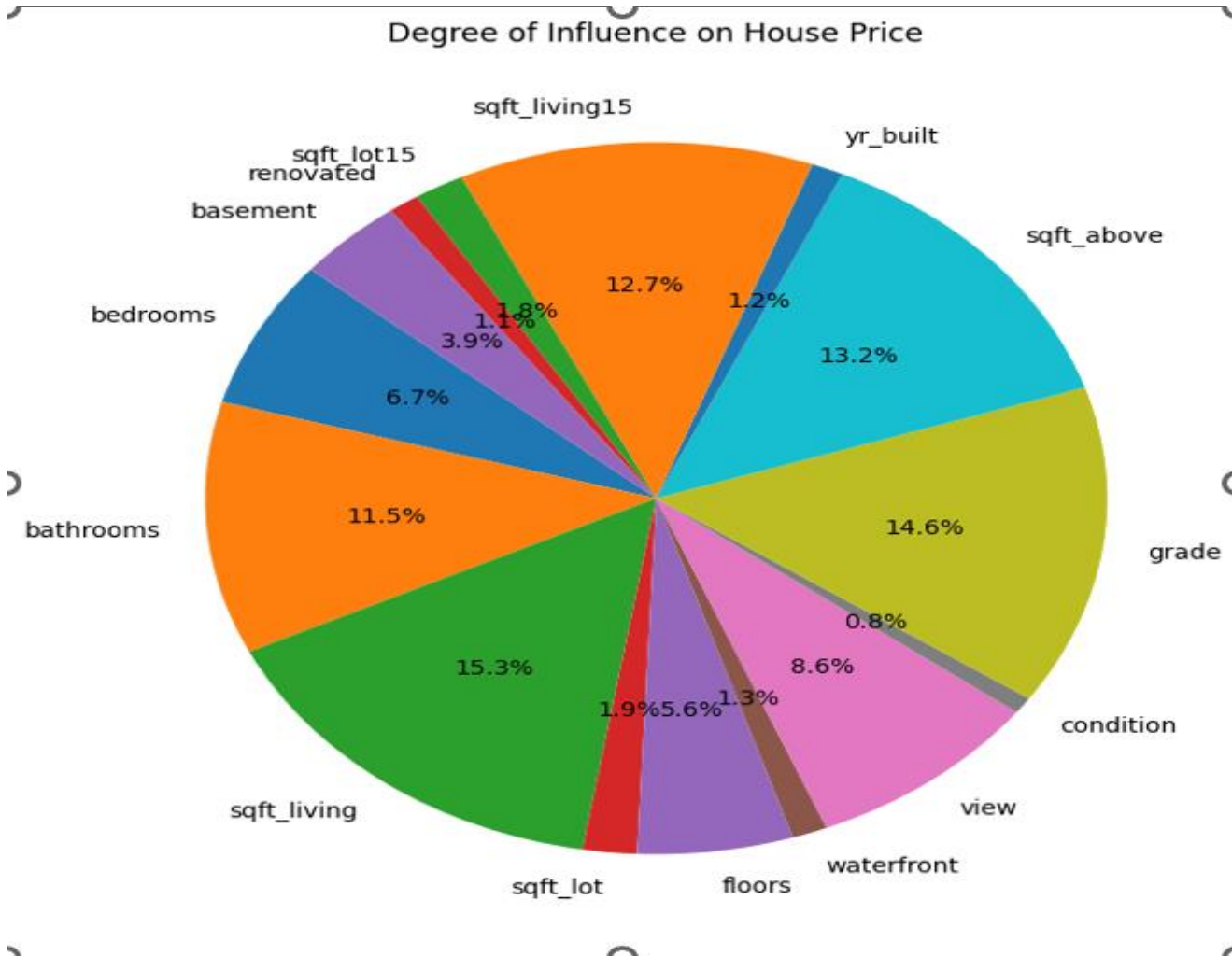
- This graph shows that bigger houses tend to have a higher price, so a house worth about 90,000 is expected to be smaller than a house worth about 6,000,000.
- We know that most houses cost less than 1,000,000 therefore, this indicates that most houses have less than 3000 square feet of living space and Houses with more than 4000 square feet of living space cost from 2,000,000

BIVARIATE ANALYSIS- Price vs Grade



- Houses with higher grades tend to be of higher quality therefore they also have a higher price, for instance houses worth about 80,000 tend to have a lower grade (5 - 7) while houses worth about 6,000,000 tend to have a higher grade (12 - 13).
- Since most houses cost less than 1,000,000, from this graph we can conclude that most houses have a grade of 7(average) and 8 (good)

MULTIVARIATE ANALYSIS- PRICE VS ALL FEATURES



- The pie chart shows what buyers in this area prioritize when looking to buy a house.
- It indicates how much influence house variables have on the house price.
- sqft_living has significant influence on price at 15.3%.
- Renovation status has less significant influence on price at 1.1%



STATISTICAL ANALYSIS

Statistical analysis is used to understand relationships within the dataset, identifying patterns, and gaining insights. In regression modeling project for predicting property value based on home renovations, here are the key steps in statistical analysis:

- 1.Descriptive Statistics
- 2.Correlation matrix
- 3.Distribution Analysis
- 4.Inferential Statistics using Hypothesis Testing and Analysis of Variance
- 5.Multicollinearity

DESCRIPTIVE ANALYSIS

Understanding the characteristics of the data. The following were the conclusions from Mean, Mode, Variance and standard deviation

Price Distribution:

The prices of houses in the dataset vary widely, with a mean price of approximately 540,296.6 to 367,368.1. The prices range from 78,000 to 7,700,000.

Property Characteristics:

The dataset contains information on various property characteristics such as the number of bedrooms, bathrooms, square footage of living space, and lot size. For example, the average number of bedrooms is approximately 3.37, with a standard deviation of about 0.93.

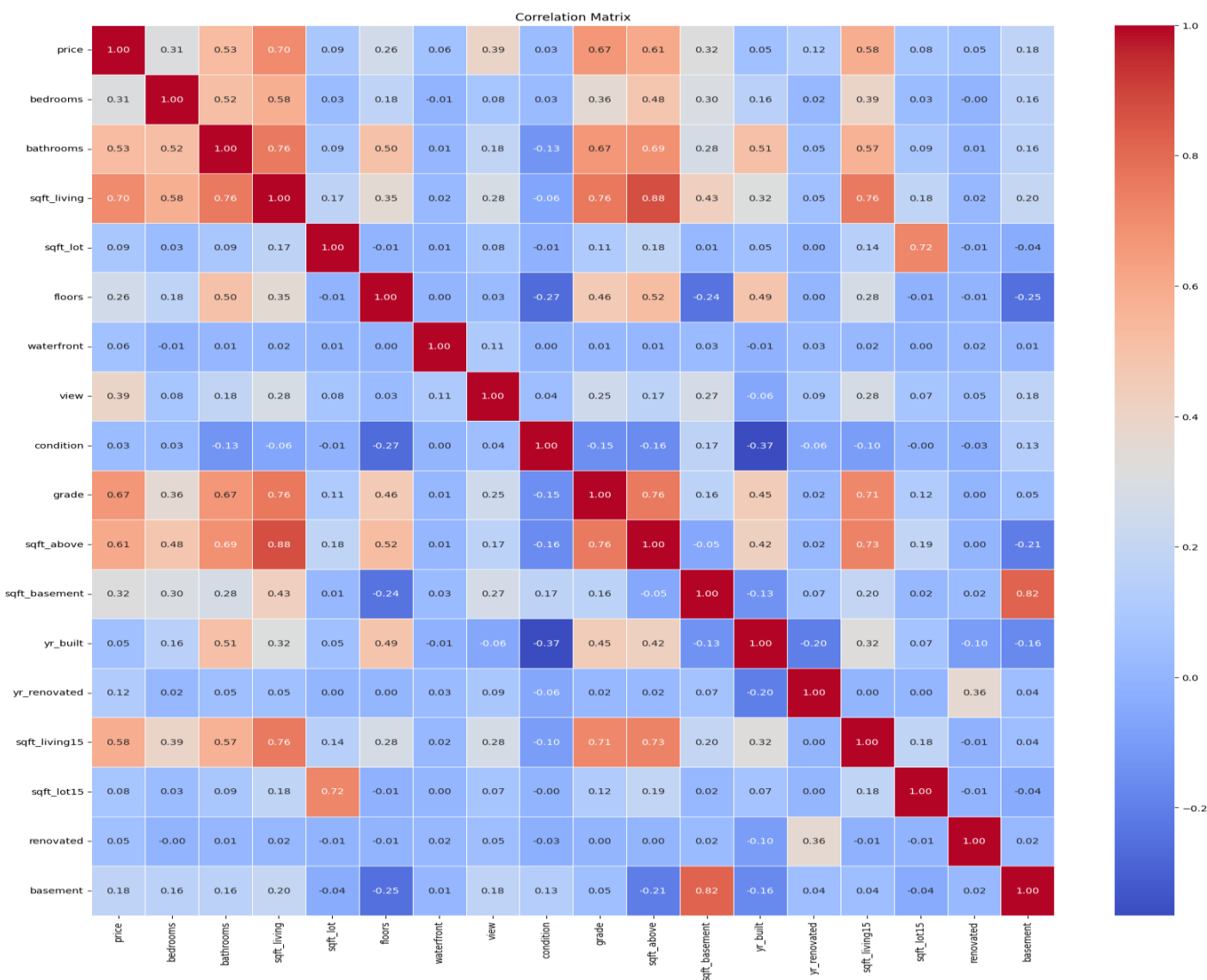
Year Built:

The houses in the dataset were built between 1900 and 2015, with an average year of construction around 1971. The standard deviation indicates that there is some variability in the construction years.

Renovation:

The majority of houses have not been renovated, as indicated by the median value of 0 and the 75th percentile value of 0. The maximum renovation year is 2015.

Correlation Matrix



- Price has a moderate positive correlation with sqft living (0.76), sqft above (0.61), sqft basement (0.32), yr_built (0.45) and sqft living15 (0.71). This means that as the values of these features increase, the price of the house also tends to increase.
- There is a weak positive correlation between price and bedrooms (0.31) and bathrooms (0.53).
- Price has a weak negative correlation with yr_renovated (-0.05). It is important to note that correlation does not imply causation. Just because two features are correlated does not mean that one causes the other.

Distribution Analysis

We performed distribution analysis using Log transformation for a skewed distribution. Below are the conclusions;

- Overall, the log transformation appears to have effectively reduced the skewness in the numerical variables, making their distributions more symmetric and suitable for statistical analysis.
- However, it's important to note that the transformation alters the scale and interpretation of the variables, so further analysis should be conducted accordingly.

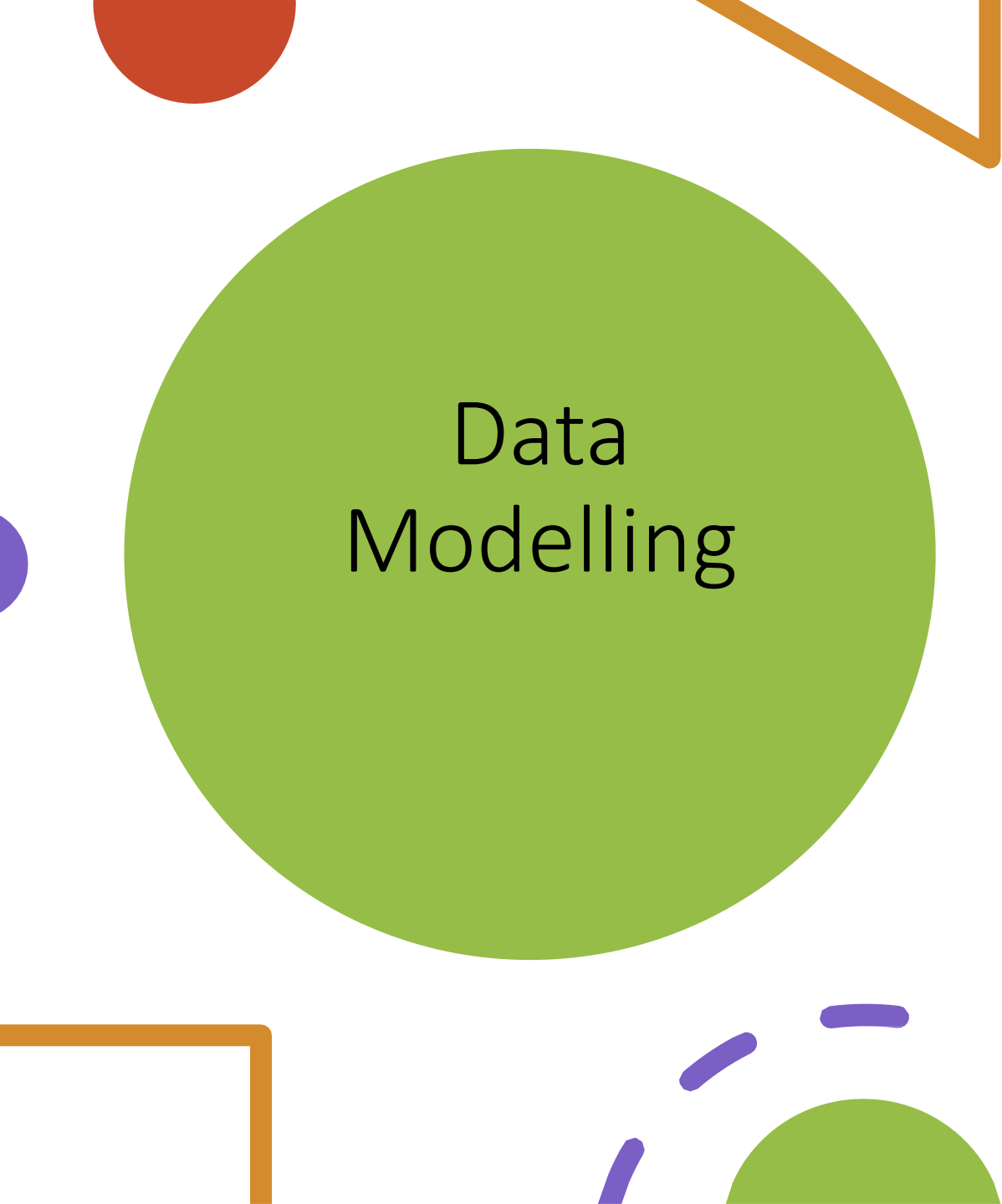
Inferential Statistics using Hypothesis Testing and Analysis of Variance

We performed a hypothesis test and results were as follows;

- The features listed (bedrooms , bathrooms, square foot living, floors, year built, square foot above square foot living 15, basement, grade , view, condition,) under "Reject the Null Hypothesis" have a statistically significant relationship with housing prices.
- On the other hand, features listed(square foot lot, renovated, square foot lot 15, waterfront) under "Fail to Reject the Null Hypothesis" do not show a statistically significant relationship with housing prices based on the ANOVA test.
- These features are important predictors of housing prices in the given dataset.

Multicollinearity

- **Conclusion**
- The analysis reveals significant multicollinearity among several features, notably 'sqft_living', 'sqft_above', 'grade', 'yr_built', and 'sqft_basement', with VIF values exceeding commonly accepted thresholds.
-
- This indicates strong correlations among these variables, potentially leading to unstable coefficient estimates and reduced interpretability in regression models.
- Consideration should be given to dropping or combining these features, implementing dimensionality reduction techniques, or applying regularization methods to mitigate multicollinearity effects and improve model performance.



Data Modelling

- We demonstrated the three models in the notebook as stated below but we will recommend only polynomial because it is the most suitable regression model to analyze the prices for a more accurate output.

1. Baseline Model

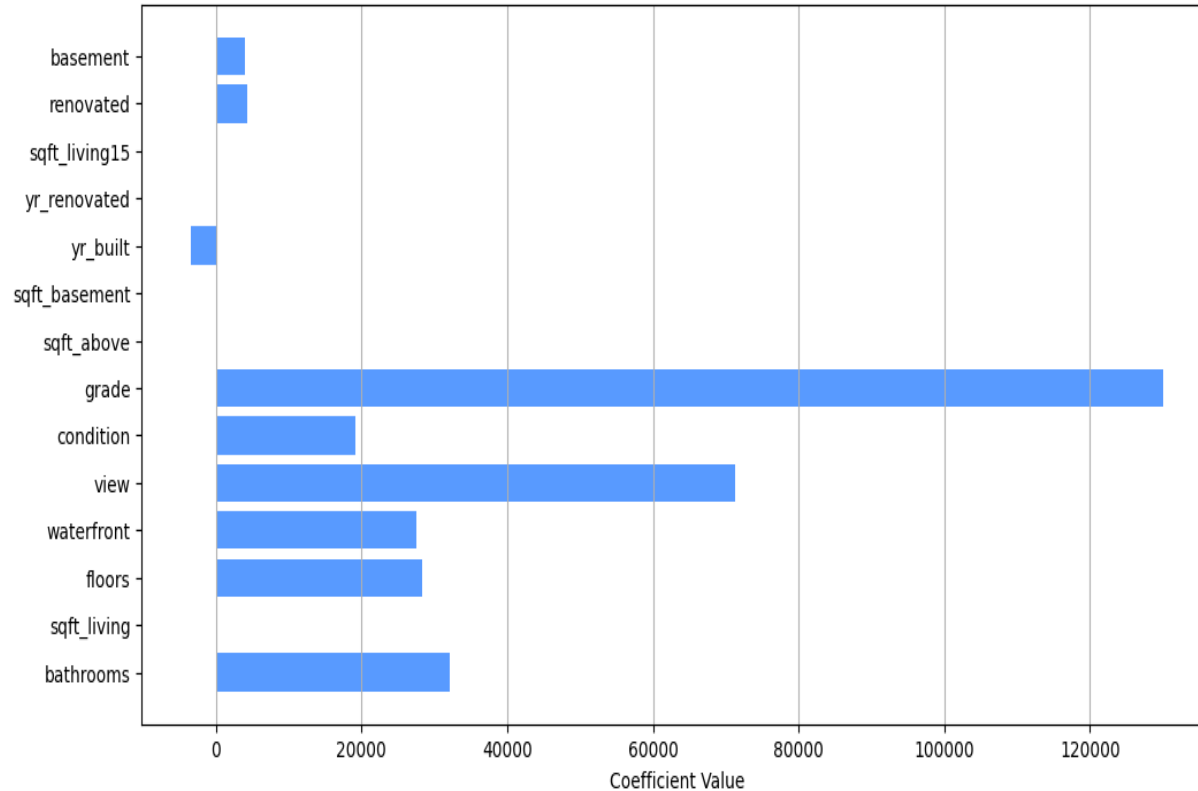
2. Polynomial Regression

3. Log Transformation

Polynomial regression

- Polynomial regression involves converting features into higher-order polynomial terms. Key features demonstrate significant impacts on prices near relationships by utilizing multiple linear regression techniques

Multiple Linear Regression Coefficients



- From the graph, a high R-squared value, attention is warranted regarding potential multicollinearity issues indicated by some coefficients' p-values.
- Moving forward, leveraging these findings can enhance decision-making processes, refine marketing strategies, and empower clients to make well-informed investment choices in the dynamic real estate market.

Regression Results

Conclusion

- Polynomial Regression is the preferred model because from the evaluation it has the highest R-squared value of 0.71
- The features below impact price such that an increase will cause an increase in the price of the property. bedrooms, bathrooms, square foot living, floors, waterfront, view, condition, grade, square above, square foot basement', year renovated', 'sqft_living15', 'renovated', 'basement'.
- On the other hand year built has a negative impact on price since the older the house the lower value it has which is shown by its negative coefficient.

CONCLUSIONS

Limitations

- 1.The dataset could have more property based characteristics.
- 2.Multicollinearity: The presence of correlated predictors (e.g., square footage and number of bedrooms) can lead to multicollinearity issues, making it challenging to interpret the individual effects of each feature accurately.
- 3.Assumption Violations: Polynomial regression assumes linearity between predictors and the target variable, which may not hold true in all cases. Violations of this assumption can lead to biased estimates and unreliable predictions.
- 4.Overfitting: Polynomial regression models, particularly those with high degrees, are susceptible to overfitting, where the model fits the training data too closely and may not generalize well to unseen data.

Overall the model was the best fit model for this prediction



RECOMMENDATIONS

- 1.Feature Enrichment: Enhance the dataset with additional property-based characteristics like property to amenities and architectural style to provide a more comprehensive understanding of factors influencing house prices.
- 2.Multicollinearity Management: Address multicollinearity by employing techniques such as feature selection, principal component analysis (PCA), or regularization methods like ridge regression or Lasso regression to prioritize important predictors and stabilize the model's interpretability.
- 3.Assumption Validation: Before using polynomial regression, verify the linearity assumption between predictors and the target variable. If non-linearity is observed, explore alternative regression techniques like generalized additive models (GAMs) or spline regression for better capturing complex relationships.
- 4.Overfitting Prevention: Prevent overfitting by balancing model complexity and generalizability through techniques such as cross-validation, regularization, or model selection criteria. Consider collecting more data or using bootstrapping to improve robustness and reduce noise in the model.





RECOMMENDATIONS TO THE STAKEHOLDERS

Consider the influence of neighboring properties

Renovations can add value

Pay attention to the year built

Consider the number of bedroom

Focus on the square footage

Maintain the condition of the property

NEXT STEPS



Stay updated on market trends

Educate clients on the impact of features

Stay informed about regulations and policies

Collaborate with appraisers

Conduct thorough market analyses

Provide renovation recommendations

Thank you!

