Phase 2 Project: King County House Sales Prediction: Multiple Regression Analysis

Collaborators:

- Maureen Wambui
- Ann Njoroge
- Phillip Kinga
- Christine Ndirangu
- Caroline Gesaka
- Wilson Wachira
- Eunice Ngunjiri

Project Overview

This project aims to analyze the King County housing data set and to build predictive multiple linear regression models that can be useful in identifying factors that affect the sale prices of houses in King County and provide data-driven insights to homeowners on how home renovations can increase property values.

Business Problem

- What house features have the greatest correlation with price?
- How much do each of these features influence house prices?
- What Machine Learning Model is most suitable when predicting house prices in King County?
- Are there typical house features that are presumed to inherently influence house prices but do not do so in this particular scenario?

Goal and Users

Our main objective is to aid prospective investors, developers, real estate agents, and homeowners in making informed choices by furnishing them with data and insights obtained through an extensive examination of the King County real estate market. By exploring how various features correlate with changes in home values.

Dataset

This project uses the King County House Sales dataset which can be found in the file "kc_house_data.csv", in this repo. The description of the column names can be found in the 'column_names.md' file in this repository.

Work Flow

- Importing of the necessary libraries
- Data Cleaning :
  - Handling Missing Values
  - Correcting Data Types
  - Handling duplicates
  - Handling Outliers
  - Ensure consistency and accuracy across columns.
  - Display final dataset
- Exploratory Data Analysis (EDA)

  To provide potential investors and homebuyers with actionable insights into the King County real estate market, we used a combination of exploratory and predictive analytic to clean, enrich, and visualize the data, thereby establishing a good foundation for predictive modeling.
- Feature Engineering :
  - Conversion of Data types
  - Label Encoding
  - Handling Outliers
  - Feature Selection :
    - Correlation with target variable(Price)
    - Variance Inflation Factor
- Exploring Linear Regression Analysis:
  - Visualizing target and independent variables correlations
  - Overview of the Assumptions of Linear Regression Analysis
  - Recommendations
- Modelling :
  - Formation of Base Models, evaluation and recommendations of the metric :
    - Simple Linear Regression Model
    - Multiple Linear Regression Model
    - Random Forest Regression Model
- Model Evaluation

Analysis

Visualizing target and independent variables correlations

A brief examination of a scatterplot displaying all variables in relation to the price provided initial insights into which features warranted deeper exploration.

Notably, the size of the lot (sqft_lot) and the number of bedrooms (bedrooms) showed minimal correlation with price.

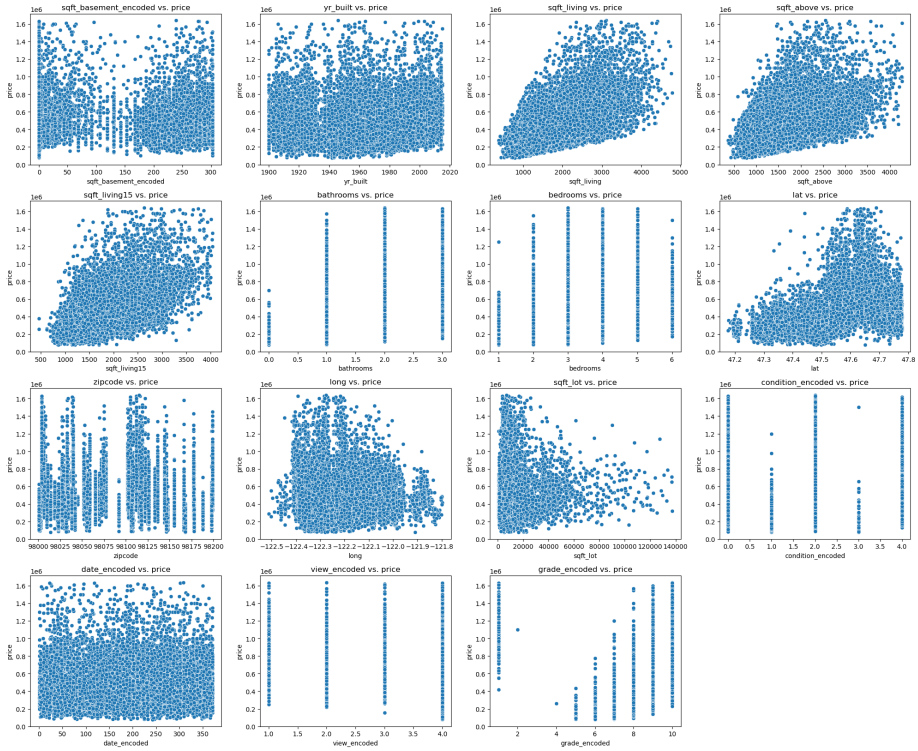

Figure 1: image

What features have the highest correlation to prices?

- The size of the house (sqft_living),
- Number of bathrooms (bathrooms),
- Quality of the construction (grade) and houses in the
- Location which is the northern part of the county had the highest correlation to price, making them the most suitable features for the model.

Modelling What Machine Learning Model is most suitable when predicting house prices in King County?

Model	R-squared
Simple Linear Regression	0.340
Multiple Linear Regression	0.618
Random Forest Regression	0.847

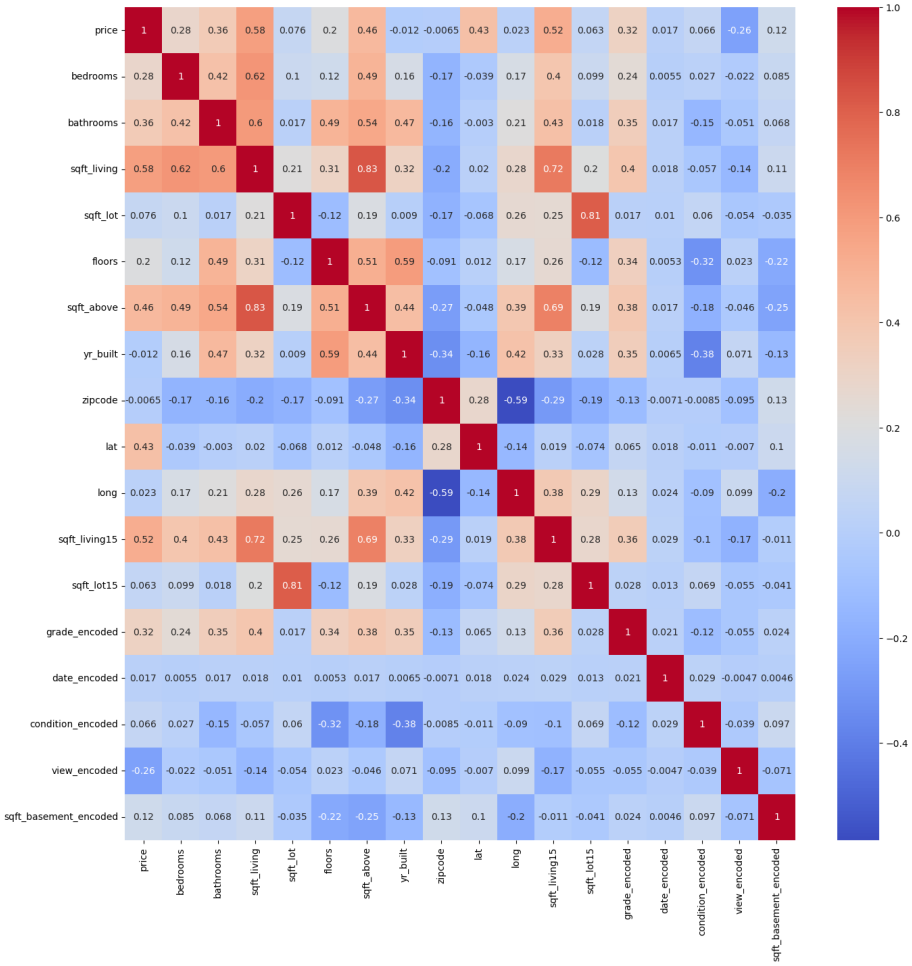

Figure 2: image

Overall, the Random Forest Regression model seems to perform reasonably well based on these evaluation metrics. These performance metrics suggest that the Random Forest Regression model performs relatively well in predicting house prices based on the given features. The high R-squared value indicates that the model captures a significant portion of the variability in house prices, while the relatively low MSE suggests that the model's predictions are generally close to the actual values.

Observation :

- sqft_living: This feature has the highest importance with a value of approximately 0.314. It suggests that the square footage of living space is the most influential feature in predicting house prices.
- lat: The latitude of the location comes next in importance, with a value of approximately 0.398. This indicates that the geographical location, represented by latitude, plays a significant role in determining house prices.
- long: The longitude of the location follows, with a value of approximately 0.062. Longitude is also an important geographical feature in predicting house prices.
- sqft_living15: This feature represents the average square footage of interior housing living space for the nearest 15 neighbors. Its importance is approximately 0.054.
- grade_encoded: The encoded grade of the house has an importance value of approximately 0.031.
- sqft_lot: The square footage of the land lot has an importance value of approximately 0.023.
- yr_built: The year the house was built has an importance value of approximately 0.023.

Conclusion

When considering the purchase of a home in King County, it's essential to recognize that the size of the property emerges as the most significant predictor of its price. Other critical factors include the number of bathrooms, the specific geographic location within the county, and the quality of construction.

However, it's important to acknowledge a significant limitation of this analysis: the dataset used is specific to King County, a single county within a major metropolitan area. Consequently, the insights derived may not be directly applicable to other cities or counties.

Next Steps

It would be beneficial to evaluate the model's applicability to various locations across the United States. Additionally, exploring alternative predictions beyond house prices, such as property valuation or market trends, using the same

dataset could yield valuable insights. Furthermore, investigating interactions between variables and exploring polynomial relationships could uncover additional nuances and improve the accuracy of the model.

For developers operating in King County, focusing on larger builds while ensuring quality construction is essential for maximizing sale prices. Recognizing the significance of location cannot be overstated. As a budget-conscious buyer, exploring areas with potential for growth and considering properties that offer value for money could be prudent strategies to pursue.