PHASE 2 PROJECT - GROUP 3

ANALYSIS OF KEY INDICATORS OF HOUSE PRICES

AUTHORS:

- TRIXIE CHEROP
- EVALYNE MACHARIA
- JOSEPHINE GATHENYA
- LAURA MUTHEU
- PRISCILLA VEKE
- MERCY CHEROTICH

Outline

- Project Overview
- Business problem
- Objectives
- Data Understanding
- Modeling
- Regression Results
- Conclusions
- Recommendations
- Thank you

Overview

This project aims to analyze house sales data using multiple regression modeling techniques. We aim to identify and quantify the relationship between various predictor factors and house sale prices.

The analysis seeks to help real estate investors make informed decisions on what type of houses they should invest in. This is in terms of the most impactful features, both positively and negatively, on House prices.

The key components of the analysis include Data preparation, Feature selection and Engineering, Model Development, Evaluation, and Validation.

Business Problem

- Real estate is a highly dynamic market influenced by numerous factors. This makes it challenging for real estate investors to accurately predict house prices.
- Inaccurate pricing models can lead to reduced profitability, missed opportunities, and dissatisfied customers.
- The current pricing strategy in the real estate industry is suboptimal, leading to potential loss of revenue and increased customer dissatisfaction.
- Hence a need for a robust predictive pricing model to enable companies to stay competitive and adapt to market fluctuations.

Objectives

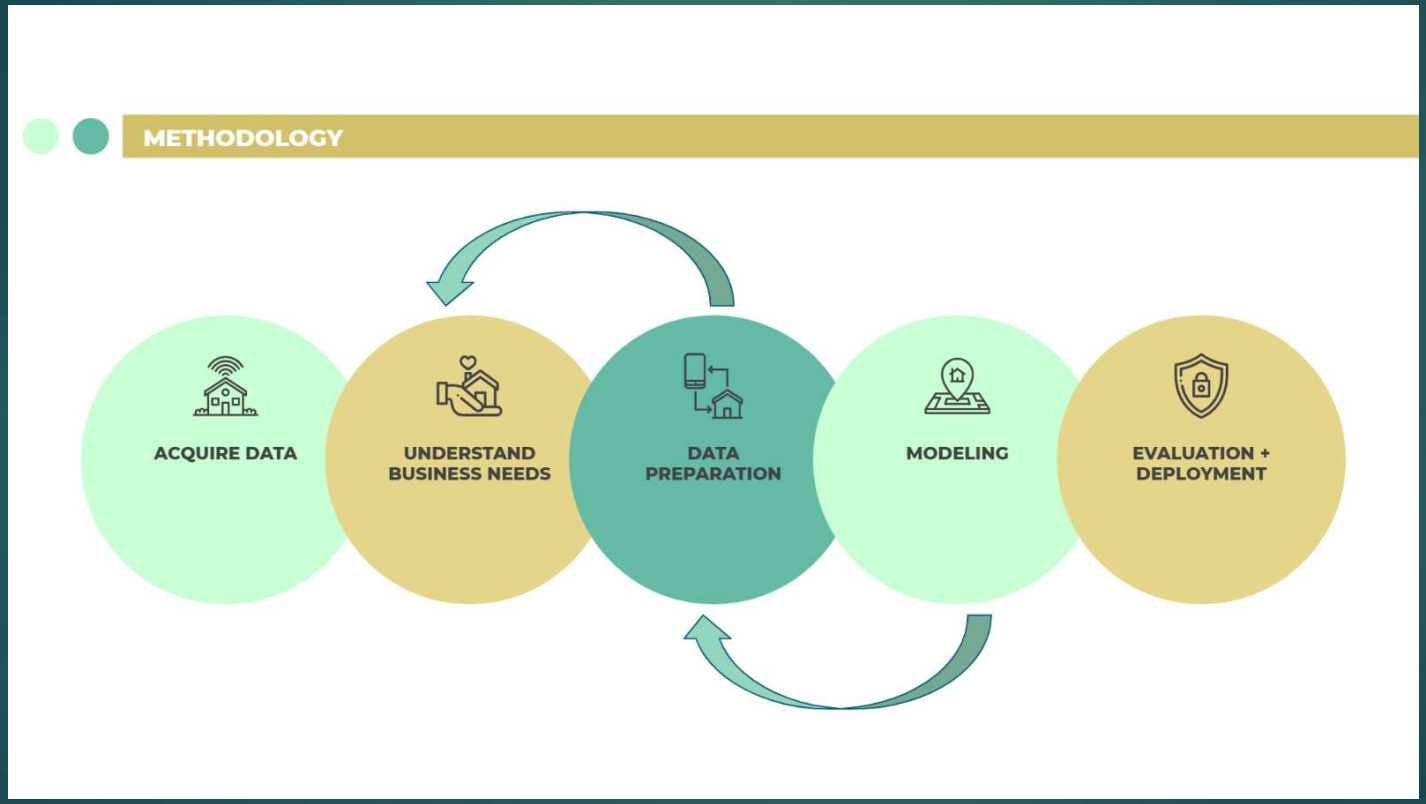
- Main Objective

To build a Linear Regression Model that predicts House Prices

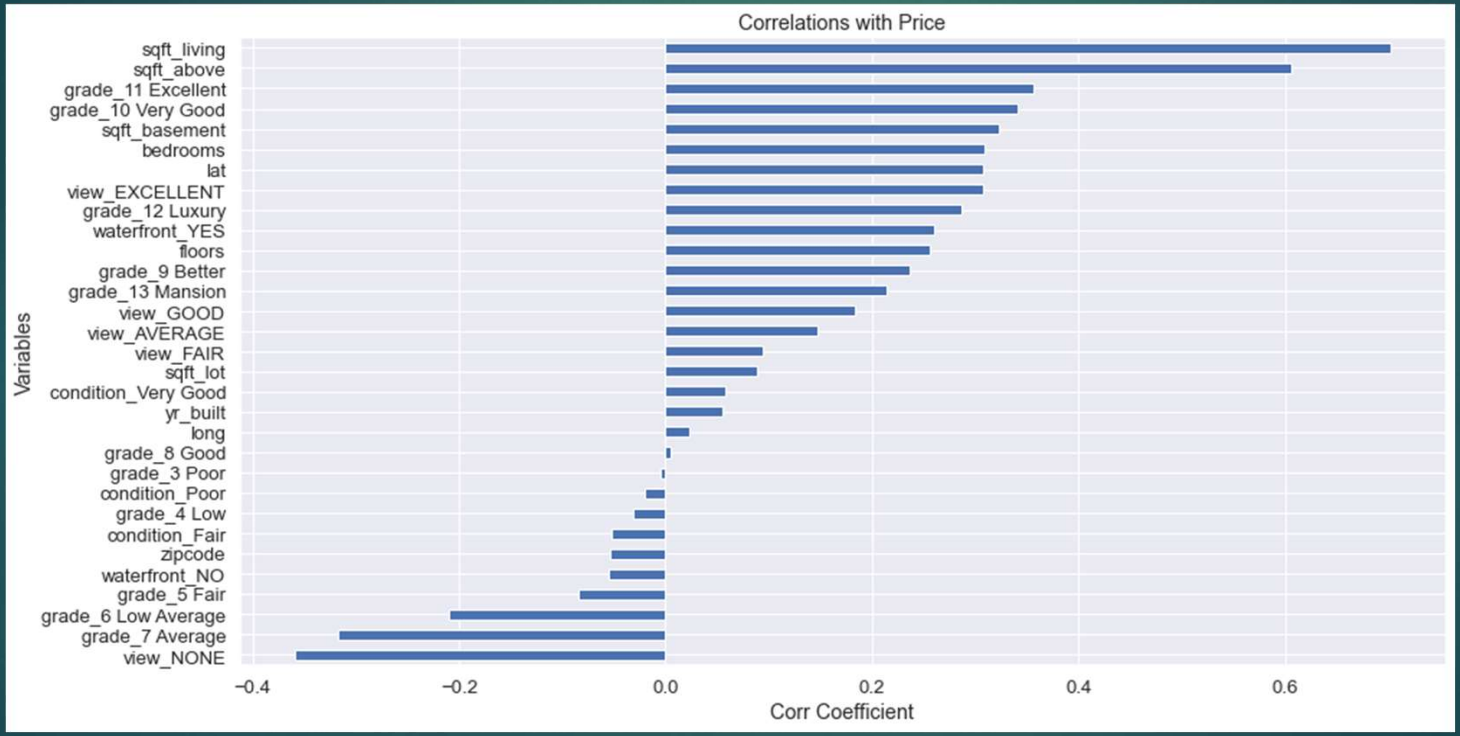
- Specific Objectives

1. To Identify key features that influence house prices
2. To assess the feature with the highest impact on House prices
3. To evaluate and validate the performance of the model

Data Understanding



Correlations between price and different independent variables after data preprocessing. This was done before doing regression analysis.



Modelling

- **Built a baseline Model (Model 1) using the 'sqft_living' feature, since it was the most highly correlated with price.**
- **Introduced more features into our baseline model that formed the basis for the multiple regression model.**
- **Created four models and noticed a significant improvement in the final model performance.**

Regression results:

- From the baseline model it was noted that the R squared was 0.493 meaning that 49.3% of price variations could be explained by square foot living.
- After transforming and adding features the R squared increased from 49.3% in the first model that only had square foot living compared to price, to 56.1% when multiple variables are used.
- The F statistic was 0.00 indicating that the overall model is significant for our analysis.
- NB: The above results were after data preprocessing and analysis.

Conclusions

- The 'Grade' feature was found to have the greatest impact on house prices.
- The features associated with higher-grade classifications (grade_13 Mansion with a coefficient of **1.2596**, grade_12 Luxury with a coefficient of **0.8849**, grade_11 Excellent with a coefficient of **0.6763**, and a larger living area (log_sqft_living) with a coefficient of **0.7078**, had the most positive impact on house prices.
- Houses with a waterfront view(with a coefficient of **0.4086**) were highly priced compared to those without a waterfront.
- Features like lower-grade classifications (grade_7 Average) with a coefficient of **-0.0812** and smaller square footage above ground (log_sqft_above) with a coefficient of **-0.1240**, had a negative impact.

Recommendations

- Real estate investors seeking premium returns should consider the grade of the house.
- Real estate investors should recognize the positive impact of larger living areas, as indicated by the log_sqft_living variable to fetch higher returns.
- Real estate investors should also consider waterfront views as they also positively impact house prices.
- Investors should be mindful of features with a negative impact on house prices, such as lower-grade classifications ("Average") and smaller square footage above ground (log_sqft_above).

THANK YOU