HOUSE PRICING ANALYSIS

June 2, 2023

Project Overview

XYZ Realtors, nestled in the heart of northwestern county, serves as the gateway to turning homeownership dreams into reality. With unwavering commitment to data-driven strategies and analytical insights, the agency seeks to be a pioneering force in optimal house pricing.

Business Problem

XYZ Realtors need help determining the key factors that influence house prices and use this information to guide pricing of houses. The goal of this project is therefore to develop the best model that will help the agency determine the best housing prices.

Project Objectives

To identify key features that significantly influence house prices in the northwestern county.

To develop an optimal pricing strategy using a robust multiple linear regression model.

To identify overpriced or underpriced houses by comparing predicted and actual prices of the houses.

To help improve the agency's annual revenue by leveraging the analytical insights and pricing strategy developed through this project.

Data Understanding

This dataset contains house sale prices for King County,USA. It includes homes sold between May 2014 and May 2015.

- The dataset contains 21 columns and 21,597 rows. This means there are 21 different variables each with 21,597 records.
- There are three main data types in the data; float, integer and object.

Data Cleaning

Dropped missing values.

Created a subset of the main dataset consisting features to use during the analysis.

- Converted the 'grade' and 'waterfront' columns from object type to integer.
- Created new columns ‘year’ and ‘numerical grade’ by modifying the ‘grade’ and ‘data’ columns.
- Removed outliers.

MODELLING

We started off with a simple linear regression as our baseline model and then gradually improved the model through:

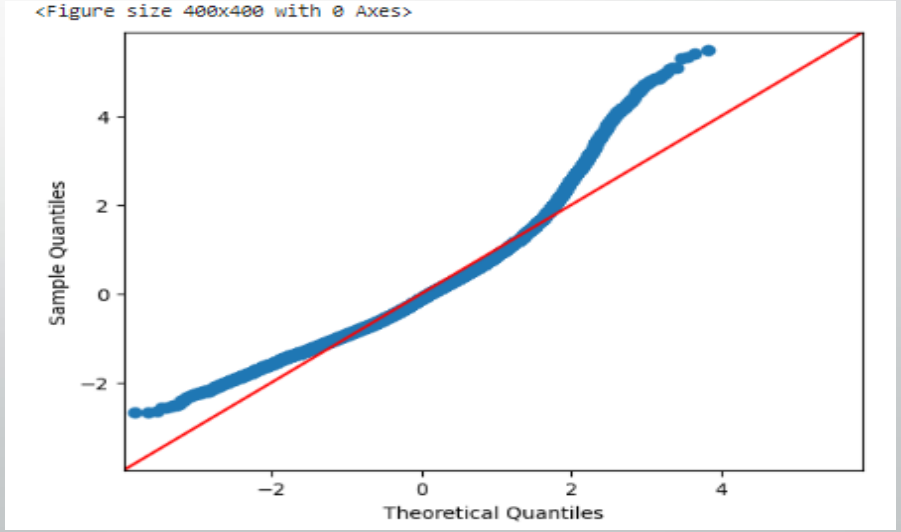
- ✓ One-hot encoding.
- ✓ Log transformation.
- ✓ Dropping statistically insignificant variables.

Baseline Model

Simple linear regression

For our baseline model with square foot living as the predictor variable:

Model Summary:
R-squared is 0.372
P-value is 0.00
Skewness of 1.077



Final Model

Model Improvement:

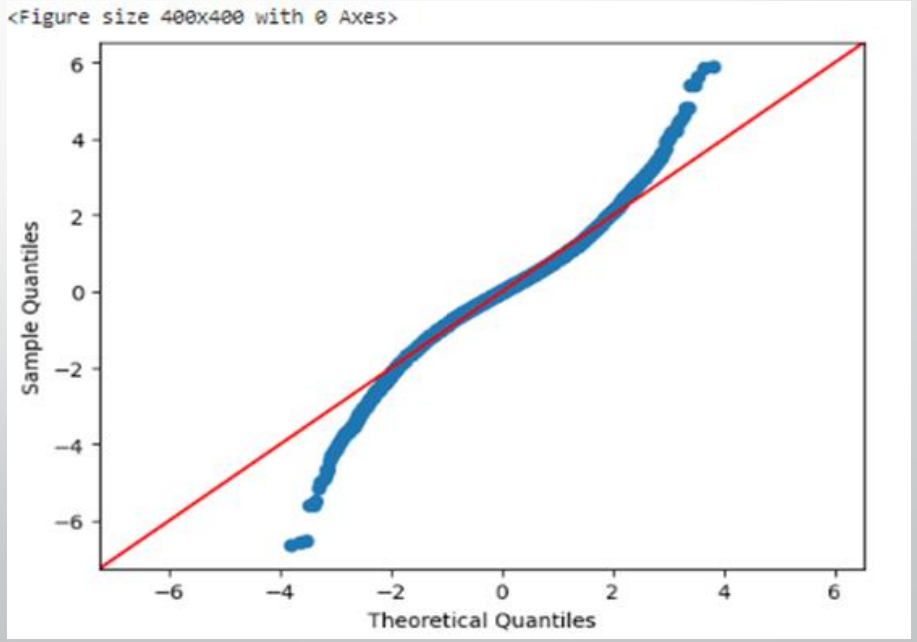
- One hot encoding.
- Log Transformation.
- Dropping statistically insignificant variables.

Model Summary

R-squared is 0.835

P-value is 0.00

Skewness of -0.116



Conclusion

The model identifies key factors influencing house prices and explains 83% of the price variation with a high degree of statistical significance.

Important factors affecting house prices include living area size, waterfront view, number of bedrooms and bathrooms, property grade, sale year, and zip codes.

Recommendations

- Focus on key features analyzed when setting house prices.
- Utilize the regression model to generate an optimal pricing strategy.
- Compare predicted and actual prices to identify overpriced and underpriced houses.
- Leverage analytical insights derived to enhance decision-making.

Next Steps

- Analyze temporal dynamics to capture seasonal trends and long-term fluctuations.
- Collaborate with real estate experts to validate the analysis and gain deeper market understanding.
- Refine the model: Explore more regression techniques to improve our model accuracy.
- Stay updated and iterate: Continuously track new data and industry trends to ensure the model remains efficient.