THYNK UNLIMITED

PREDICTING THE HOUSE PRICES IN THE NORTHWESTE RN COUNTY.

Project Overview

This is a project that focuses on analyzing the King County House Sales dataset. The goal of using this data is to find correlating factors with the price of a house so that I can make recommendations to the real-estate company based on the findings. Looking for factors that contribute a large amount to the variation of the price of a home so that better business recommendations can be made based on those factors.

Business Problem

Our stakeholder is a real-estate company who are looking to expand to NorthWestern County. They need a reliable metric for house prices and would like to know which features of houses are more impactful on the price. My task is to provide them with a multiple linear regression model that will provide them with an equation that will include features that are most important in determining housing prices. By applying multi linear regression and analyzing the King County housing dataset, I aim to uncover the significant factors that drive house prices. This analysis will enable real estate agents to navigate the local market and provide accurate pricing recommendations.

Data Understanding

Data preparation and cleaning

For this project, I will be using the King County House Sales dataset, the dataset includes various features such as the price, number of bedrooms, number of bathrooms, living area size, house grade, and condition of the house, just to name a few.

The dataset has 21 columns and over 21597 rows, covering sales of houses between the years 2014 and 2015.

The analysis performed on the dataset included the following steps:

1. Data Cleaning: Rows with missing data were dropped, and duplicate entries were removed.
2. Data Transformation: Categorical data in the `view`, `date`, `grade`, `condition`, and `waterfront` columns were converted into numerical data using label encoding.
3. Exploratory Data Analysis: Various checks were conducted to assess the linearity assumptions between the target and predictor variables and identify any potential multicollinearity issues.

MODELING

THE FINAL MODEL TOOK FOUR FACTORS IN THE END, SQUARE FOOTAGE, SQUARE FOOTAGE OF LIVING15, GRADE AND NUMBER OF BEDROOMS. I CHOSE THESE FACTORS BASED ON THEIR CORRELATION WITH THE TARGET FACTOR, THE HOUSE PRICE. I CHECKED THE COLINEARITY ON ALL THE FACTORS AND EVEN THROUGH SOME FACTORS OUT OF THE MODEL BASED ON THEIR HIGH COLLINEARITY WITH SQUARE FOOTAGE. MY R-SQUARED VALUE WAS 54.1%, MEANING THAT WITH THE MODEL CAN BE ABLE TO EXPLAIN 54.1% OF THE VARIATION IN A HOME'S PRICE. THE MEAN SQUARED ERROR FOR THE MODEL WAS ALSO SIGNIFICANTLY SMALLER THAN THE BASELINE MODEL AND SIGNIFICANTLY SMALLER THAN ALL THE OTHER MODELS. EACH VARIABLE IN THE MODEL ALSO HAS A SIGNIFICANT P VALUE.

Results of the Modelling

The final model has a r-squared score of 0.537, meaning that 54% of the variance in the dataset is described by the model.

The final model had a Mean Absolute Error of 163,441 and a Root Mean Squared Error of 245,084. This is low compared to the other three models.

The model features had a p-value < 0.05 (our alpha/significance level), which tells us that all features have a statistically significant linear relationship with price.

The final model included 5 of the most significant predictor variables of house price. They are Intercept, grade, sqft_living, sqft_living15, and bathroom.

In the final model3 when compared to the baseline model, the R-squared increased from 49.6% to 54%

Recommendations

* In NorthWestern County, square footage, square foot living, grade, and bathrooms have been identified as the most important factors in determining the price of house. When increasing the square footage and improving the grade of the house the home sellers should also consider adding more bathrooms since from the analysis there exists positive relationship between these four factors.

* The real estate market is an industry that is dynamic and constantly changes. To ensure the model validity and continuous accuracy, the models needs to be regularly retrained using the latest data. This will help capture any shifts or trends in the market and maintain the model's effectiveness.

Conclusions

* The more the number of bathrooms, the more expensive the house.

* The better the grade of the house, the more expensive the house.

* Square Footage of Living Space: The square footage of living space has a positive impact on house prices. As the size of the living space increases, the estimated price of the house also increases. This indicates that larger houses are generally priced higher.

* The model provides valuable insights into the factors affecting house prices in NorthWestern County and offers recommendations for homeowners and researchers interested in understanding the housing market dynamics.

* The selected features in our model were statistically significant linear relationships with the price, since their p-values was less than the alpha, thus the assumptions of independence, linearity, and normality were met.

THANK YOU