Western Governors University / Graduate Capstone

EXECUTIVE SUMMARY

Sunny Lai

Overview

Exploring housing sales within King County, Washington by utilizing multiple linear regression to assess for significant relationships to housing price.

The Problem and Hypothesis

- What housing characteristics significantly influence the price of houses within King County, Washington?
- By determining what housing factors impact price the most, we can predict a price range for prospective homebuyers.
- Analytical data overall has improved the retail process within real estate. Market insight, price analysis, predictive analytics, and targeted marketing are some of the advantages that technical data analysis provides to not only prospective homebuyers, but business stakeholders and investors as well.
- **Hypothesis** - there is a significant relationship between housing characteristics and the price of housing within King County
- **Null Hypothesis** - There are no housing factors that have a significant relationship to the price of housing within King County

Data Analysis Process

- **Data Collection:** The data utilized for this analysis can be found publicly on Kaggle at https://www.kaggle.com/datasets/harlfoxem/housesalesprediction.
- **Data Cleaning:** The data set was cleaned by detecting missing values, duplicates, and outliers.
- **Data Preparation:** Data explorations, Summary statistics, univariate and bivariate analysis, heat map, changing data type

- **Data Analysis:**
    - Normality test: Kolmogorov-Smirnov test
        - KS statistic: 0.11
        - p-value: 0.19
    - Multiple Regression Analysis: OLS
        - Utilized correlation scores <0.3, VIF, and p-values >0.05 to reduce the initial model
        - Residual plot
        - Q-Q plot
        - Cross-validation score
    - Linear Regression Prediction Model
        - Split the dataset into train and test
        - Fit the linear regression model
        - Calculated predictions
        - Mean Squared Error

- The results of our analysis identified the variables of bedrooms, bathrooms, view, grade, and sqft_living15 as significant housing features in the prediction of the price of housing within King County.
- The final model concluded with an improved AIC score of 1.612e+04 from our initial model.
- The cross-validation score for our model resulted in a mean squared error (mse) score and an R-squared score of 0.55.
- The mean squared errors for our train and test set were 1.3 and 1.2 respectfully, these numbers being similar indicate we would get similar results if we imputed a new data set to our model.
- The residual plot exhibited normally distributed residuals and our Q-Q plot distributed along our expected slope

Limitations

- Data set only contains market data for one year. With only one year of market data to analyze, this may cause variables that are highly correlated with each other.
- Our dataset was originally published to the kingcounty.gov official website in 2014, however it cannot be easily accessed or located on their website to date.
- Limited number of independent variables
- Multiple linear regression requires many assumptions to be met for accurate results

Proposed Actions

- For prospective house buyers, evaluate if forfeiting some housing features when it pertains to bedrooms, bathrooms, view, grade, and sqft_living15 can be achieved if the goal is to save money.
- Expand the time range of data collection to exceed one year of data collection
- Add a location variable that is more correlated with the price of houses. Adding districts may provide a more accurate representation of location and how this impacts the price of housing.

Benefits of Study

- Multiple logistic regression will allow for a detailed analysis that can provide insight into what factors impact the price of housing within King County.
- By determining what housing factors impact price the most, we can predict a price range for prospective homebuyers.
- Analytical data overall has improved the retail process within real estate. Market insight, price analysis, predictive analytics, and targeted marketing are some of the advantages that technical data analysis provides to not only prospective homebuyers, but business stakeholders and investors as well
- By determining significant housing factors and predictions of housing prices, we can improve the overall efficiency, decision-making, and provide valuable insight for all stakeholders within the real estate market