Briana Churchill
Student ID: 011009463
Dr. William Sewell
2024 April 26

Executive Summary

Statement of the problem and the hypothesis

Using the USA real estate dataset provided via Kaggle with > 2 million instances of home listings from March 2022 through March 2024, can a reliable multiple linear regression model be created to predict housing prices in Salt Lake County, Utah for investment purposes?

Hypothesis:
H0: A predictive MLR model regarding Salt Lake County's housing market can be made from the research dataset with a model accuracy > 70% based off the R-squared value.

Null Hypothesis:
H1: A reliable predictive model pertaining to Salt Lake County's housing market cannot be constructed using the research dataset with a model accuracy > 70% based off the R-squared value.

Summary of the data-analysis process

Initial data exploration was completed by creating a univariate visualization to provide insights regarding the dependent variable "Price." Then univariate visualizations were created for each independent variable, along with bivariate visuals to demonstrate the relationships between the independent variables and price. Based off the visuals created, the following interpretations were made:

- Most of the listings fall between price range $375,000 to $700,000
- Distribution of bedrooms is almost even
    - 4-6 bedrooms make up 48.4% of listings
    - 1-3 bedrooms make up 51.6% of listings
- Average price for a home with 1-3 bedrooms is about $500,000
- Average price for a home with 4-6 bedrooms is > $600,000
- Listings with 1-2 bathrooms make up most of the listings at 60.4%
- Listings with 3-4 bathrooms make up 39.6% of listings
- Average price for a home with 1-2 bathrooms is $500,000
- Average price for a home with 3-4 bathrooms is > $600,000
- Cities located in the southern half of Salt Lake County make up most of the listings at 58.2%
- Northern located cities make up 41.8% of listings
- Average price of a southern city is $600,000

- Average price of a northern city is $500,000
- Most listings fall between ~1,000 to 4,000 square feet
- In relation to house size, the higher the square footage, the higher the price
- Most listings have a lot size of about 0.3 acres or less
- Greater distribution of price is seen in listings with 0.0 to 0.2 acres as compared to higher acre values
- Higher acreage is associated with higher price

Following the exploratory steps above, the data was normalized, and feature selection was performed. Although measures were taken in the data preparation stages to avoid multicollinearity, there is still a possibility of high correlation among independent variables present in the model. Taking an extra measure of using the VIF value to eliminate features was necessary to ensure a reliable model. Variables with a VIF value of 5 or greater were removed one by one starting with the highest value.

Fortunately for this model, only one variable needed to be removed due to the VIF value being 7.39 for "House Size." Once any possibility of multicollinearity was removed, an additional feature selection method was executed. Backward stepwise elimination was chosen to assess the variables based on the statistical significance of the p-value. Any variables found to have a p-value of <= 0.05 would have been removed due to statistical insignificance. All the p-values of the model were 0.00 so no additional features needed to be removed and the final model was already created. Values related to the residual standard error, R-squared, and Durbin-Watson were assessed to ensure model efficacy. An added measure of plotting the residuals was completed too.

Outline of the findings

Based off the regression results of the model pertaining to RSE, R^2, and Durbin-Watson, the following observations were made:

- The value for residual standard error (RSE) is 0.13
- The value for R-squared is 0.406
- Otherwise, the independent variables explain around 41% of variability in the dependent variable
- The value for Durbin-Watson is 1.346
- A value of 2 indicates no autocorrelation amongst residuals, so this value could be better

The p-values, R-squared value, the residual standard error value, and the Durbin-Watson value indicate that the model does have statistical significance. Unfortunately for the hypothesis of this research question, there is evidence stacking up to accept the null hypothesis. Specifically, the accuracy of the model, which is based off the R-squared value, or otherwise the statistical measure of how well the model is at making predictions on a scale of 0 to 1. Preferably, the final model would have an R-squared value of > 0.70, which would correlate to the independent variables explaining at least 70% of variability in the dependent variable price.

To further analyze the model to provide additional evidence either in support or against the hypothesis, it was necessary to also review the residuals of the regression model. To review the residuals, regression plots were created for each independent variable. For columns Bedrooms, Bathrooms, and City the data points are plotted vertically along the y-axis rather than being centered around the line of best fit, y=0. Only the residuals plot for variable Lot Size shows a differing distribution of data points; with many closer to the line of best fit in comparison to the other variables mentioned previously. However, the visual does not show a normal distribution along the line of best fit, but rather a potential pattern which can skew the analysis. Without multivariate normality, this model cannot satisfy the assumptions for multiple linear regression and is not a reliable model.

Having unsatisfied assumptions for multiple linear regression, irregular residuals plots, and statistical values mentioned above, there was sufficient evidence to accept the null hypothesis. It was not possible to create a multiple linear regression model to predict housing prices in Salt Lake County using the USA real estate dataset provided via Kaggle with an accuracy of > 70%.

Limitations

A major limitation of the model is the lack of additional listing data, such as the duration of listings, type of home (Ie: condominium, townhome, etc.) and other potential factors that could provide more insight to the Salt Lake real estate market. Additionally, although the dataset contained over 2 million instances initially, the cleaning and transformation processes reduced the data frame to only 1,995 rows. Overall, there is not enough data to move forward with utilizing this reduced model to make predictions regarding price.

Summary of proposed actions and expected benefits of the study

There are still benefits to reviewing the results of this analysis for business purposes. Rather than using the model for predictions, an alternative course of action for real estate investors would be to utilize the information provided by the initial exploratory visualizations. Although the regression model may not be reliable enough to make accurate predictions, the bivariate explorations still could prove useful.

There are relationships between the independent variables and dependent variable that could still help investors improve their portfolio(s). Perhaps growing their investments to include more properties in southern Salt Lake County cities, with more bedrooms, more bathrooms, and larger lot sizes. Or in the context of preparing to sell properties that may be undervalued (Ie: located in the northern county area, sparce number of bedrooms, etc.) investors could consider implementing changes to appreciate the value through renovations or small upgrades.