Milestone 1: Project Proposal

Felipe Rodriguez

Bellevue University

DSC 680 Applied Data Science

Professor Amirfarrokh Iranitalab

March 17, 2024

Topic

The project Real Estate Pricing Analysis will analyze real estate data to understand where homes are being sold and the relationships between location and size of the home.

Business Problem

A real estate agency wants to understand their real estate listings countrywide. Additionally, the company wants to build a model to accurately predict prices based on zip code and size, specifically for the state of New York. The data provided contains historical listing information that will aid in the analysis and modeling. The model and predictions will be used to provide more accurate pricing recommendations to clients.

Datasets

The data is being obtained from Kaggle and is sourced from Realtor.com. The data contains the following information: status, bed, bath, acre lot, city, state, zip code, house size, previous sold date, and price. The data will allow for the analysis of price based on the features provided. The main features that will be used for modeling and predictions are house size and zip code. Additionally, the other fields in the data set can be used to analyze price such as city and state.

Methods

The methods used in this project will include various visualizations of prices, correlation matrices of the variables, and a linear regression model to predict prices. The visualizations will include various scenarios of the prices, which will include geographical analysis and charts

illustrating the prices. The correlation matrices that will be created will help understand the correlation between the features and the target - price.

Ethical Considerations

An ethical consideration to consider is that the model does not have any bias and unfair to the variables used. The two fields being used to create predictions are zip code and city. When conducing analysis, the predictive model must not discriminate based on these two variables. The data needs to be analyzed to ensure that there are no biases that could lead to unfair outcomes.

Challenges/Issues

When conducting the analysis, it is important to understand what the data contains and if it will be sufficient for analysis. For example, when examining the state of New York for pricing predictions, there needs to be a variety of zip codes and prices to properly apply a model. Another challenge that might occur, is a scaling issue with the zip codes. When conducting this analysis, the zip codes need to be pre-processed for modeling, such as handling missing values and encoding the variables. This could be a challenge if there many zip codes.

References

The data used is sourced from realtor.com. Realtor.com provides a monthly report of median prices by state, city, and zip code. Using this set, a comparison can be done of the prices to see the accuracy of the data gathered. One thing to note is that the data contains historic and

current data, while the data from Realtor.com that will be used to validate will contain a month's worth of data.