Project Proposal - Housing Cost Based on Features

Data Source:

https://www.kaggle.com/datasets/yasserh/housing-prices-dataset/code

In recent years, the cost of home ownership has become increasingly difficult for everyday working class people. With skyrocketing inflation and rising government interest rates, the price of homes have become unpredictable, with large fluctuations in the market in just a few years. Because of this, it has been difficult for the consumer to accurately determine what the true value of a home should be. This leaves many buyers wondering; *What are the top 3 features of a home that contributes to its price, and how can home features be used to accurately determine a fair value cost for the property?*

The sourced dataset provided by Kaggle offers many different insights on home value. This dataset gives us the price and size in area, as well as information pertaining to the features within each house. These features include things like number of bedrooms and bathrooms, number of stories the house has, and amenities such as a finished basement, hot water heating, and central air conditioning.

The dataset contains 545 rows with unique homes, with 13 total features for each house, giving us a good amount of data to work with. To start, I will clean up the data through the data wrangling process, converting yes / no columns to boolean type (True/False), while also handling any missing values or detecting outliers within the data. After the data is cleaned up, I plan on conducting an exploratory data analysis. This will give an idea on the relative trends within the dataset, and give a general idea as to what features are most commonly present.

From here, I plan on engineering features, such as cost per area (by dividing the total cost of the home by the total area), and total number of rooms, by taking the sum of the bedrooms, bathrooms, and guest rooms per house, although this may be misleading since the total number of rooms may not be an accurate representation of this number.

Next, I will create a model using the train:test split function from SciKit-Learn,where I plan on testing different models for the problem statement. These models include a multiple linear regression model, which will help predict home cost based on the features within the dataset. I'm also planning on testing a random forest regression model, which will create multiple decision trees to predict home prices that way.

Once the model is complete, performance will be assessed by calculating the brier score of each model, and the model with the lowest brier score will be able to give us the closest estimate to the price of a home. This model will be able to take listed homes with their features, input those values into the model, and accurately predict what the cost of each given house should be. The model will also be able to tell us what features contribute most to the cost, giving buyers the option to disregard certain features that drive up the price, depending on their budget.