Web Scraped Data Analysis and Visualization

Purpose:

The point of this project was to scrape data. The data was scraped from real estate listings from multiple zip codes in NJ getting the data from Realtor.com. They were analyzed to understand how housing features such as number of bedrooms, bathroom, and square footage affected prices. The project also used machine learning to make a predictive model of housing prices based on these features.

Data:

Listing data from five NJ zip codes, each listing pulled price, number of bedrooms and bathrooms, square feet, and city. Data was then stored in a data frame it was then cleaned so there were no missing or zero values. Price per square foot was also an important part of this project, making it easy to compare homes of different sizes.

Tools Used:

The tools used were Selenium, Undectected ChromeDriver, WebDriver Manager, pandas, numpy, seaborn, matplotlib, and modeling tools like scikit-learn.

Preliminary Analysis:

The box plots and bar plots helped understand the different features of houses. Looking through the data it showed that price per sqft did vary and houses with more bedrooms and bathrooms led to higher prices.

Prediction Modeling:

To predict housing prices a multiple linear regression model was made using a pipeline that used one-hot encoding to make the city feature categorical. The data was split into two sets one training and one testing and they were using Linear Regression from scikit. Model performance was evaluated using R-squared, mean absolute error, and root mean squared error. The model had an r-squared value of 0.90. That is a usual score meaning that the model is accurate.

Future Study:

Other housing features can be included like year built, property type, listing age, time listed. Also adding other cities to expand the data can be done. Using advanced models for the predictive model. Seasonal effects should also be studied to understand the pricing trends over time.