https://www.kaggle.com/datasets/fratzcan/usa-house-prices

Introduction

In the past 20 years, the average home price in the U.S. has grown from about $140,000 to about $420,400 as of the end of 2024. This increase shows the demand for housing and the various economic, social, and structural factors influencing price trends. This project aims to analyze housing data to predict the increase in house prices over time and identify the factors contributing to this trend.

I will be using the USA House Prices dataset from Kaggle. The dataset contains 4140 rows and 18 columns of house-related data, representing various features that could influence house prices. The columns include date, price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, sqft_above, sqft_basement, yr_built, yr_renovated, street, city, statezip, and country. By examining these features, I will explore how different factors contribute to changes in house prices over time and use regression techniques to predict future price trends.

What is regression and how does it work

Regression in machine learning refers to a supervised learning technique where the goal is to predict a continuous numerical value based on one or more independent features. One of the most common types of regression is Linear Regression, which learns from the labelled datasets and maps the data points. This optimized function can then be used to make predictions on new datasets. Linear Regression predicts the continuous output variables based on the independent input variable.

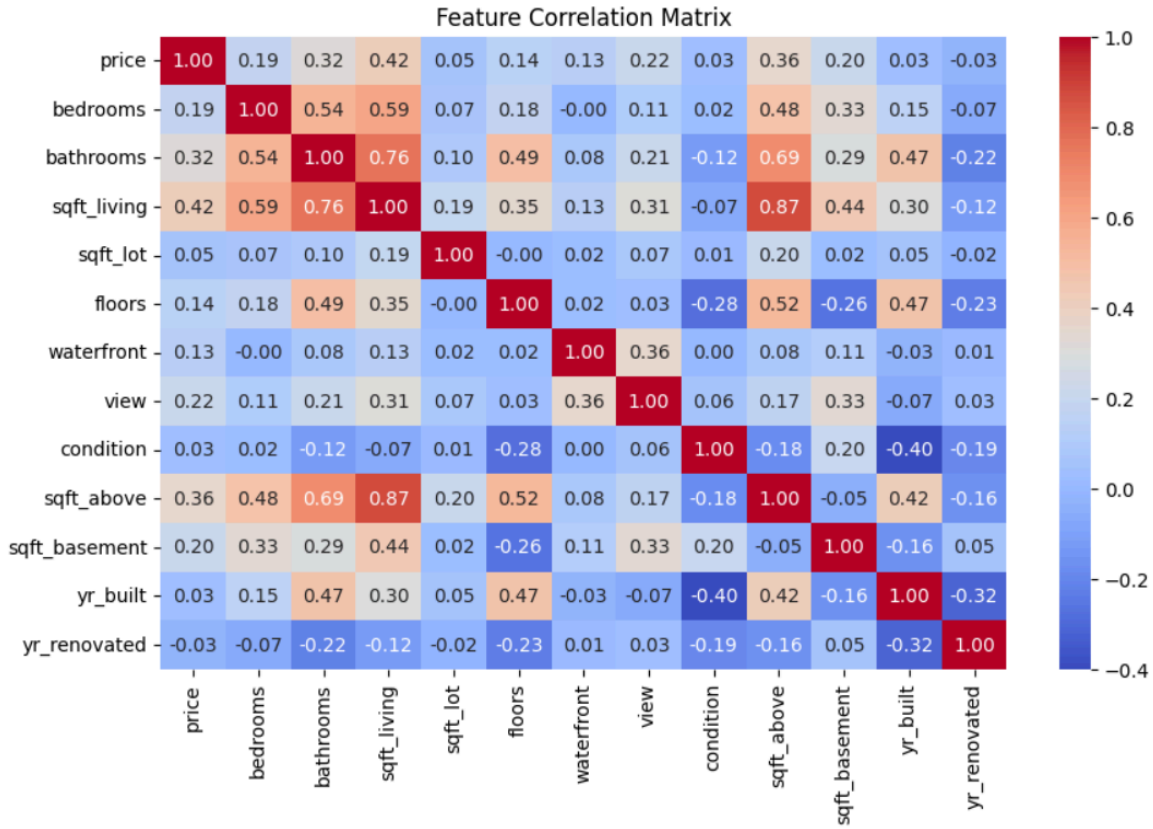
Formula for linear regression:

$$Y_i = f(X_i, \beta) + e_i$$

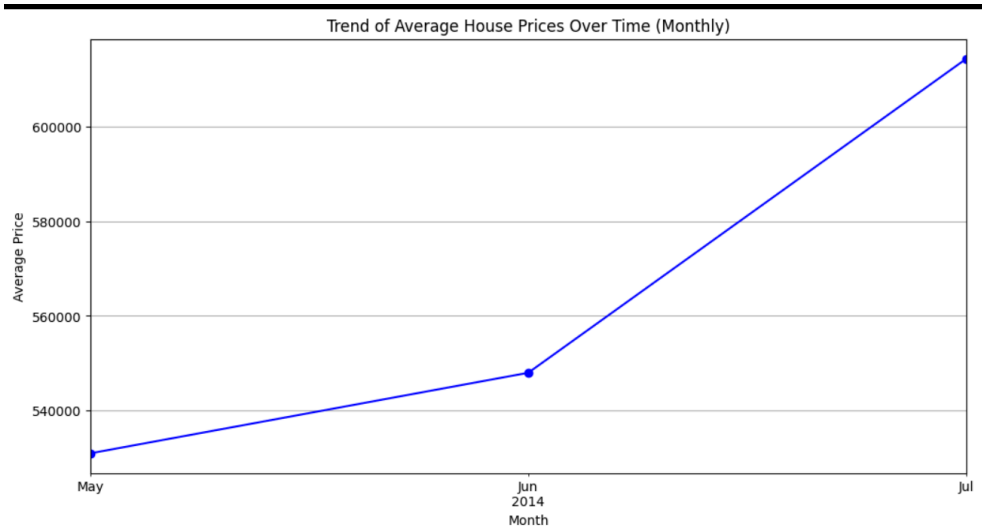
Y_i = dependent variable
f = function
X_i = independent variable
β = unknown parameters
e_i = error terms

Experiment 1: Data understanding

To gain understanding of the dataset I first started with finding null and duplicate values and there were none in this dataset. I then used the df.info() function to find what data types each column is where I found columns price, sqft_living, sqft_lot, sqft_above, sqft_basement, bedrooms, bathrooms, floors, view, condition, waterfront, yr_built, and yr_renovated are numerical values. Columns street, city, statezip, country and date are non numeric values.



I created a heatmap to visualize the correlation between features. From this, I observed that price has the strongest correlation with sqft_living (0.42), sqft_above (0.36), and bathrooms (0.32). There is a moderate correlation with view (0.22), sqft_basement (0.20), and bedrooms (0.19). Price shows little to no correlation with sqft_lot (0.05), floors (0.14), condition (0.03), yr_renovated (-0.03), and yr_built (0.03). Larger homes tend to be priced higher, as there is stronger correlation with sqft_living and sqft_above.



I created a line graph to look at trends in the dataset. The dataset only contained information from the year 2014 thus I looked at the months from that year. The line graph shows house prices increased gradually from May to June, followed by a sharper rise in July.

Experiment 1: Pre-processing

After gaining an understanding of the dataset, the next step is pre-processing so the data is clean and ready for modeling. There are no null and duplicate values in the dataset so nothing needs to be done for this. I dropped the street and country columns because all the houses are in the USA, and these columns do not contribute meaningful information to house price predictions. Additionally, I created a 'year_month' column from the 'date' column to simplify

trend analysis, as all data in the dataset is from 2014. This helps in grouping and visualizing price trends.

Experiment 1: Modeling and Evaluation

```
R² Score: 0.3748
Mean Absolute Error (MAE): $167,422.50
Root Mean Squared Error (RMSE): $256,023.64
```

I created a linear regression model for my first experiment. The R^2 score is 0.3748, meaning that 37.48% of the variation in house prices is explained by the model. The Mean Absolute Error (MAE) is $167,422.50, which indicates that, on average, the model's predictions are off by this amount compared to actual prices. Additionally, the Root Mean Squared Error (RMSE) is $256,023.64, which penalizes larger errors more than MAE. These results suggest that the model is not very strong and could be improved.

Experiment 2: Modeling and Evaluation

```
R² Score: 0.3105
Mean Absolute Error (MAE): $178,337.88
Root Mean Squared Error (RMSE): $268,855.43
```

For this experiment I experimented with different features by removing yr_built, yr_renovated, and condition to see if the model would improve its performance. As the results show, the R^2 score went down and Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) went up. This shows that removing these features didn't improve the model, thus the features contained important information for predicting house prices.

Experiment 3: Modeling and Evaluation

```
Random Forest R² Score: 0.3235
Random Forest Mean Absolute Error (MAE): $168,037.09
Random Forest Root Mean Squared Error (RMSE): $266,317.18
```

For this experiment I did a Random Forest model with the same features as experiment one. The R^2 score is higher than the second experiment but lower than the first experiment. The Mean Absolute Error (MAE) is slightly better than the second experiment but worse than the first experiment. Likewise the Root Mean Squared Error (RMSE) is slightly better than the second experiment but worse than the first experiment.

Impact

A positive impact of this research is that it can help predict house prices for potential buyers, giving them a rough estimate of how much they need to save. This could guide individuals in making better financial decisions before they buy a home. However, a negative impact is that since the models have a high margin of error, they might mislead buyers by predicting prices that are too high or too low. This could affect potential buyers purchasing decisions.

Conclusion

Overall, I learned that feature selection has a major role in model performance, and removing certain features may negatively impact predictive accuracy. While the Random Forest model provided a slight improvement, the models still had high errors, showing that more advanced techniques can be used to further improve predictions.

Citations

https://finance.yahoo.com/news/us-home-values-changed-over-165648270.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAIYPNwooQUDpR9QST7RLgXTD4KLQCVT9gW95pxhIAIFarNkMYHe7WNsQ-gPPvv8OseDTVcvj9FnItOyMMsCSxnFsxGdfNui3VjUVEHPa2J2KyX3pSS5JQZm_vHGAK5v1m2V7oCgMKWQMMU4asxhqhSS4EguVfGnGy5pGZbls8BBP

https://www.geeksforgeeks.org/regression-in-machine-learning/