Real Estate Price Prediction Project

Introduction to Problem and Data

Problem Statement:

Real estate pricing is a complex problem influenced by a wide range of factors such as property size, location, amenities, and age. These factors interact in various ways, making it challenging to predict property prices. Accurate pricing models are crucial for various stakeholders in the real estate market:

1. **Buyers** can make informed decisions, ensuring they get the best value for their money.
2. **Sellers** can set competitive yet profitable prices, improving their chances of selling properties quickly.
3. **Real Estate Investors** can identify lucrative opportunities, maximizing their returns on investments.

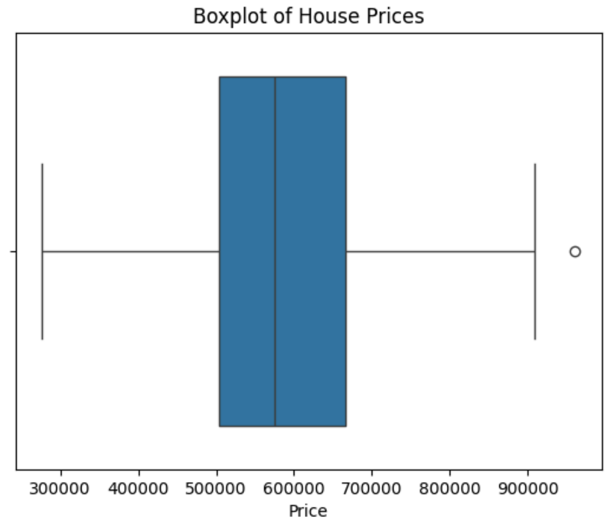
For my final project, I aim to predict real estate property prices using a dataset containing various attributes that influence the value of a property. The ultimate goal is to provide insights into the key drivers of property value and develop a robust, scalable solution that stakeholders can use for accurate property valuation in dynamic real estate markets.

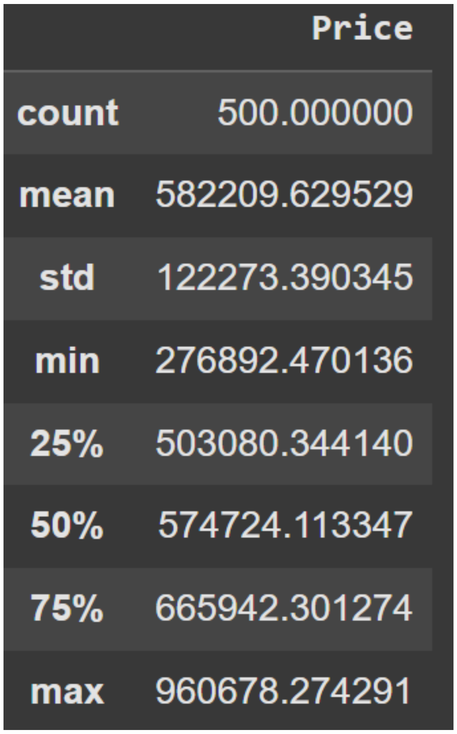
Dataset Description:

Data for this project is sourced from Kaggle in csv format, providing comprehensive information about real estate properties and their varying features. The dataset includes features such as property size (square footage), the number of bedrooms and bathrooms, location score, and proximity to city centers, among other ones.

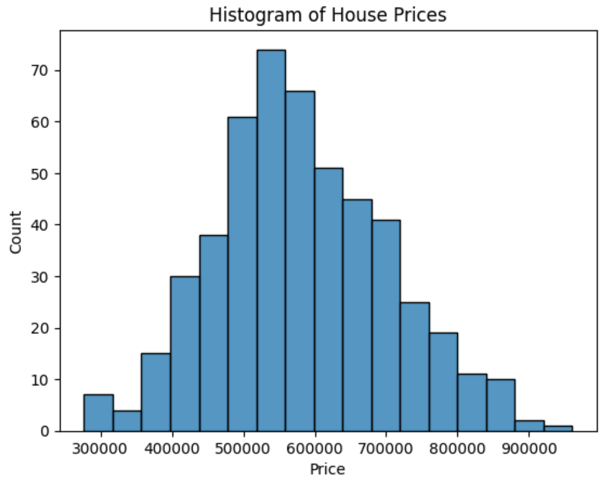
Feature distributions vary widely, with some features being categorical and others numerical. This dataset contains 500 real estate properties and 11 columns representing their different features and attributes (not including ID), which I can use to predict prices of other real estate properties based on these features. The target variable (Price) ranges from roughly $200,000 to $1,000,000.

Exploratory Data Analysis



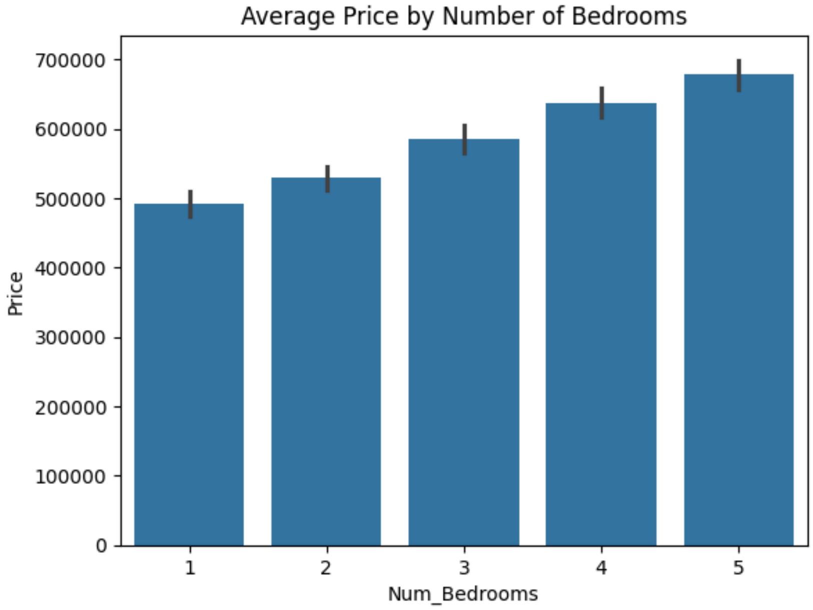


	Price
count	500.000000
mean	582209.629529
std	122273.390345
min	276892.470136
25%	503080.344140
50%	574724.113347
75%	665942.301274
max	960678.274291

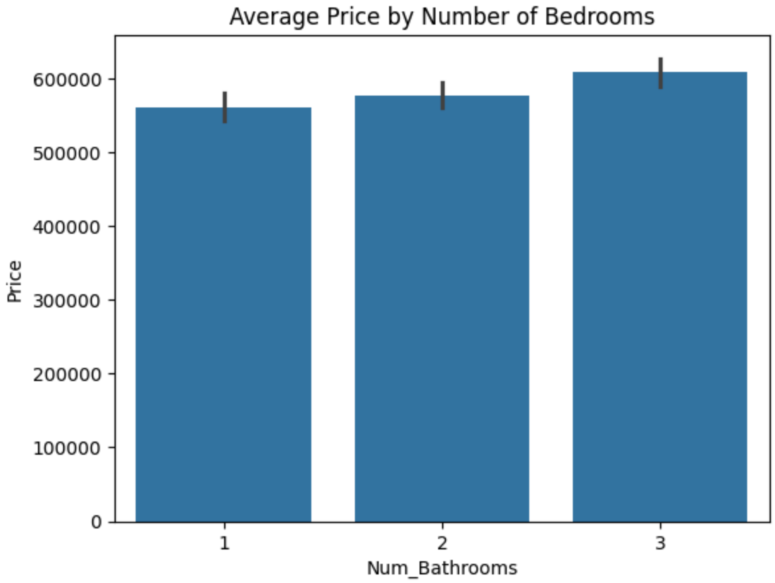


Based on the above models and statistics, the minimum price is roughly $276,000 and the maximum price is roughly $960,000. The distribution of the prices is slightly right skewed and concentrated around the $500,000 to $600,000 price range.

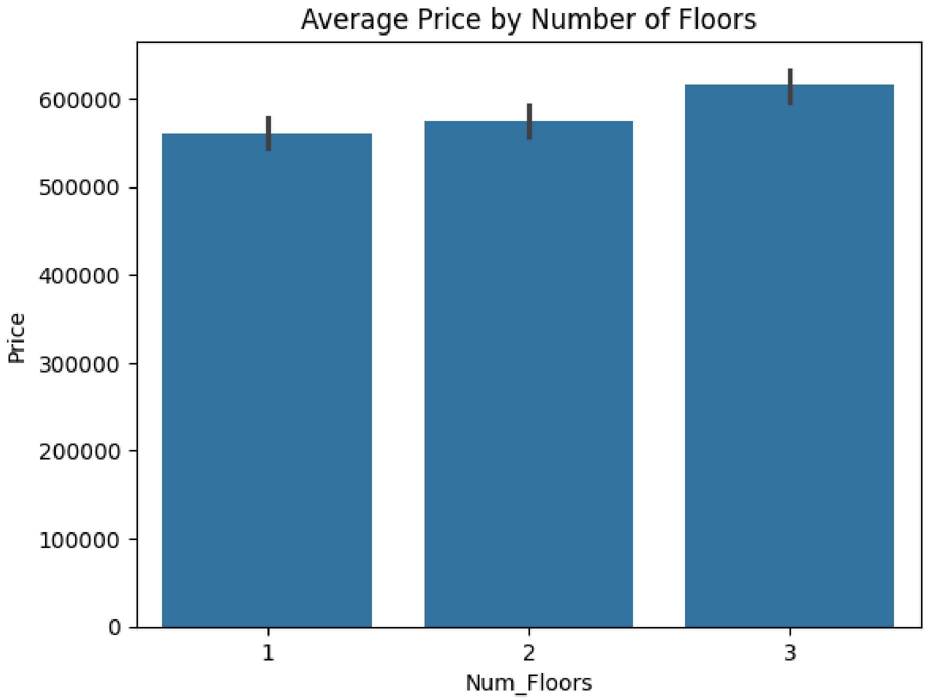
Correlations



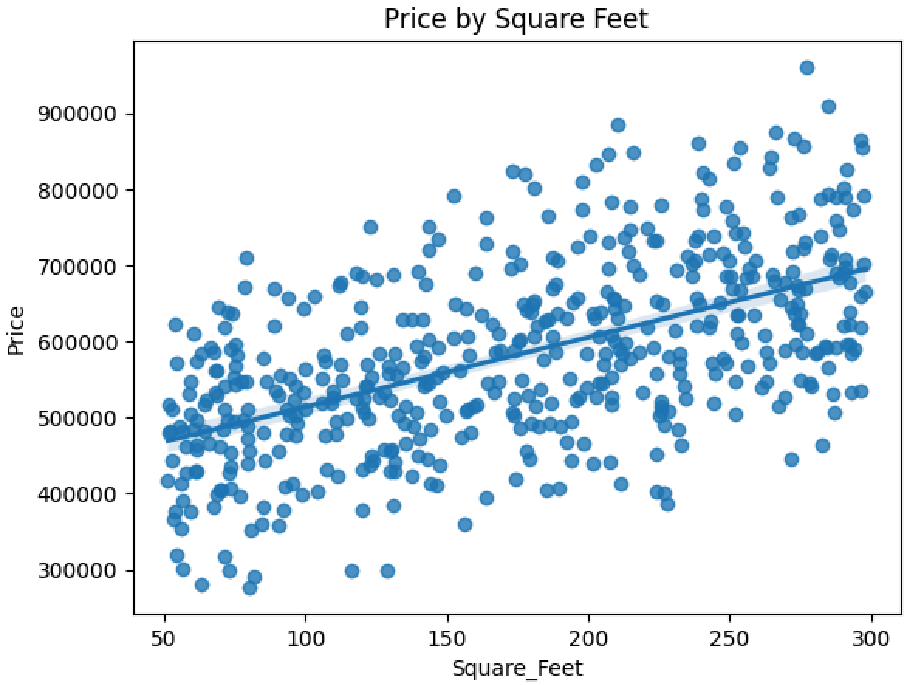
The bar graph shows the relationship between the number of bedrooms and the average price associated with each respective real estate property. It shows that the higher number of bedrooms corresponds to a higher price point, suggesting that housing prices are positively correlated with the number of bedrooms at a roughly even rate.



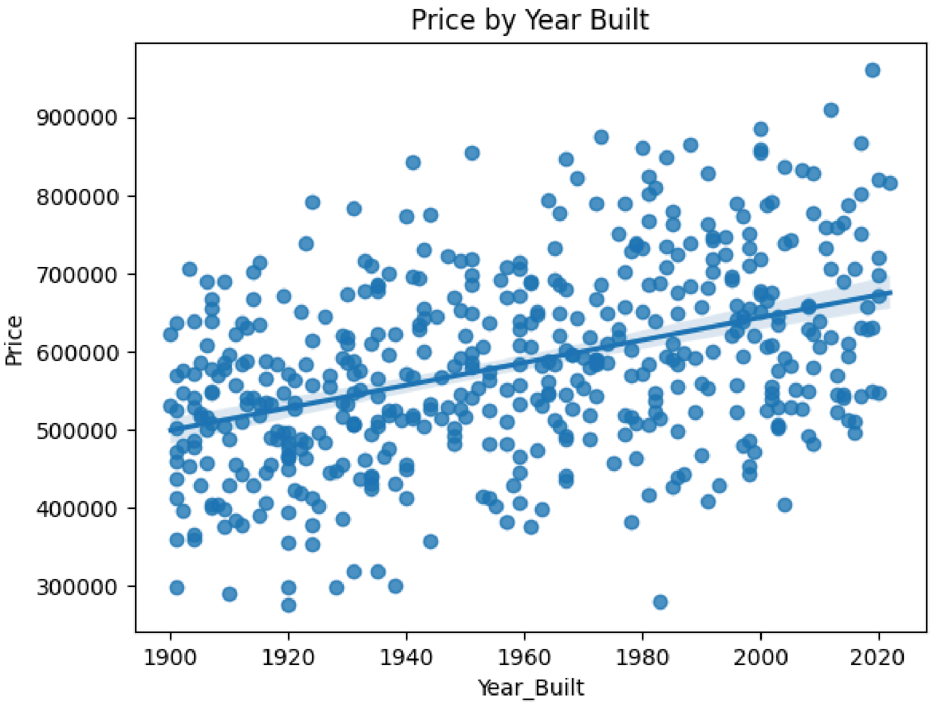
This bar graph demonstrates the relationship between the average price of real estate properties to the number of bathrooms. Like with the number of bedrooms, the more bathrooms results in higher prices, suggesting positive correlation. However, the price points between 1 and 2 bathrooms are much closer to each other than that of 2 and 3 bathrooms, indicating a relatively higher jump in price going from 2 to 3 bathrooms.



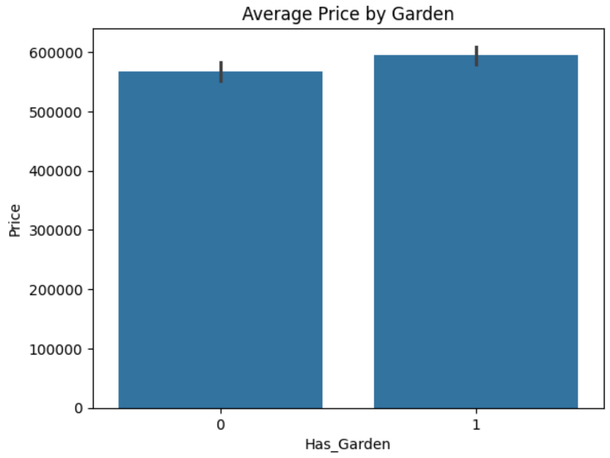
The bar graph compares the average price to the number of floors. Similar with the bathrooms, the more floors results in a higher price, with there being a substantial increase from 2 floors to 3 floors.

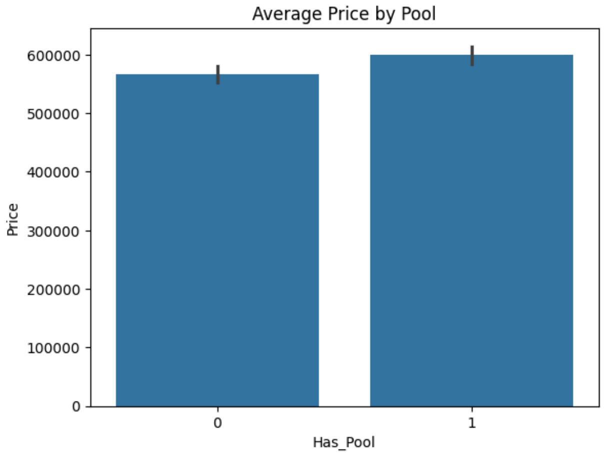


The scatterplot shows the relationship between the square footage of the property and the price of it. There is a general upward trend of more square footage to a higher price.

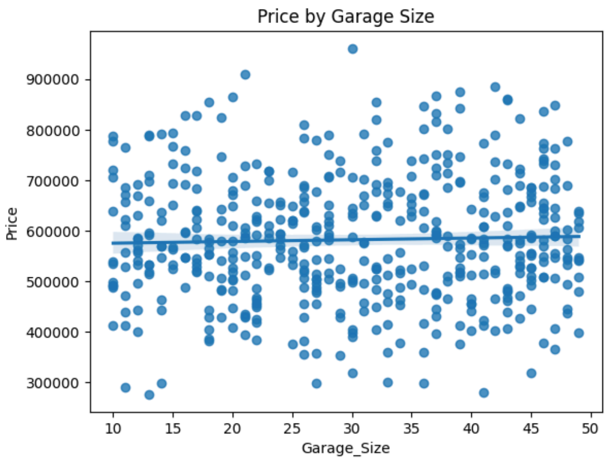


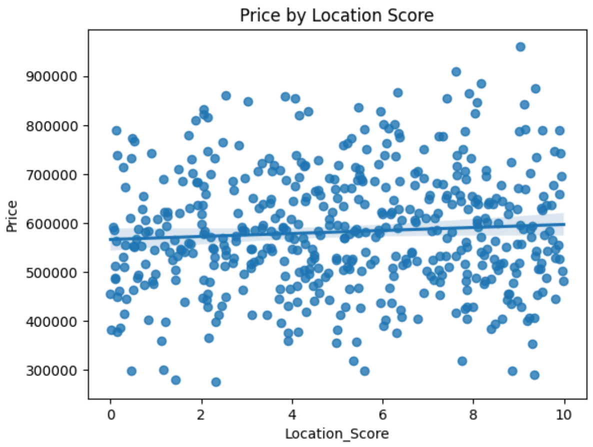
The scatterplot compares the year the property was built to the price of it. More modern properties tend to have a higher price, suggesting a positive correlation between how modern the property is and the price of it.

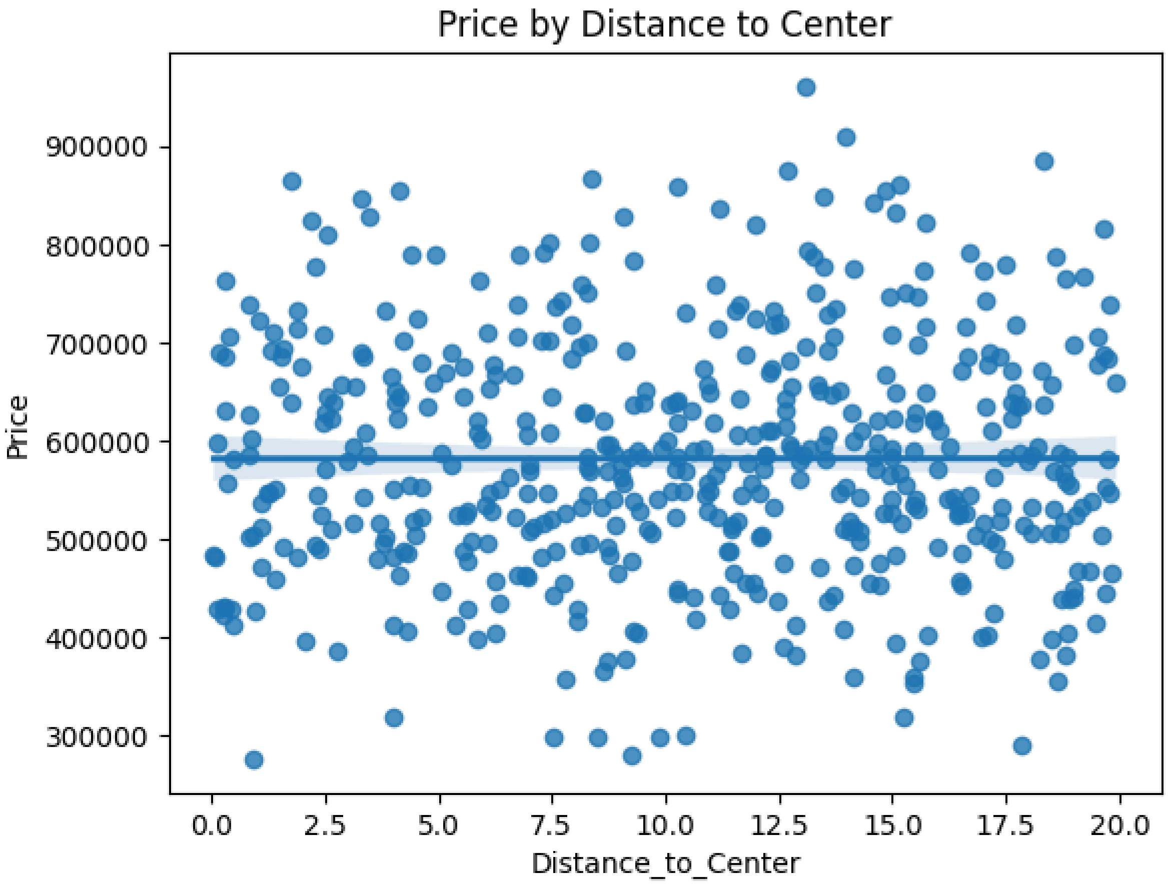




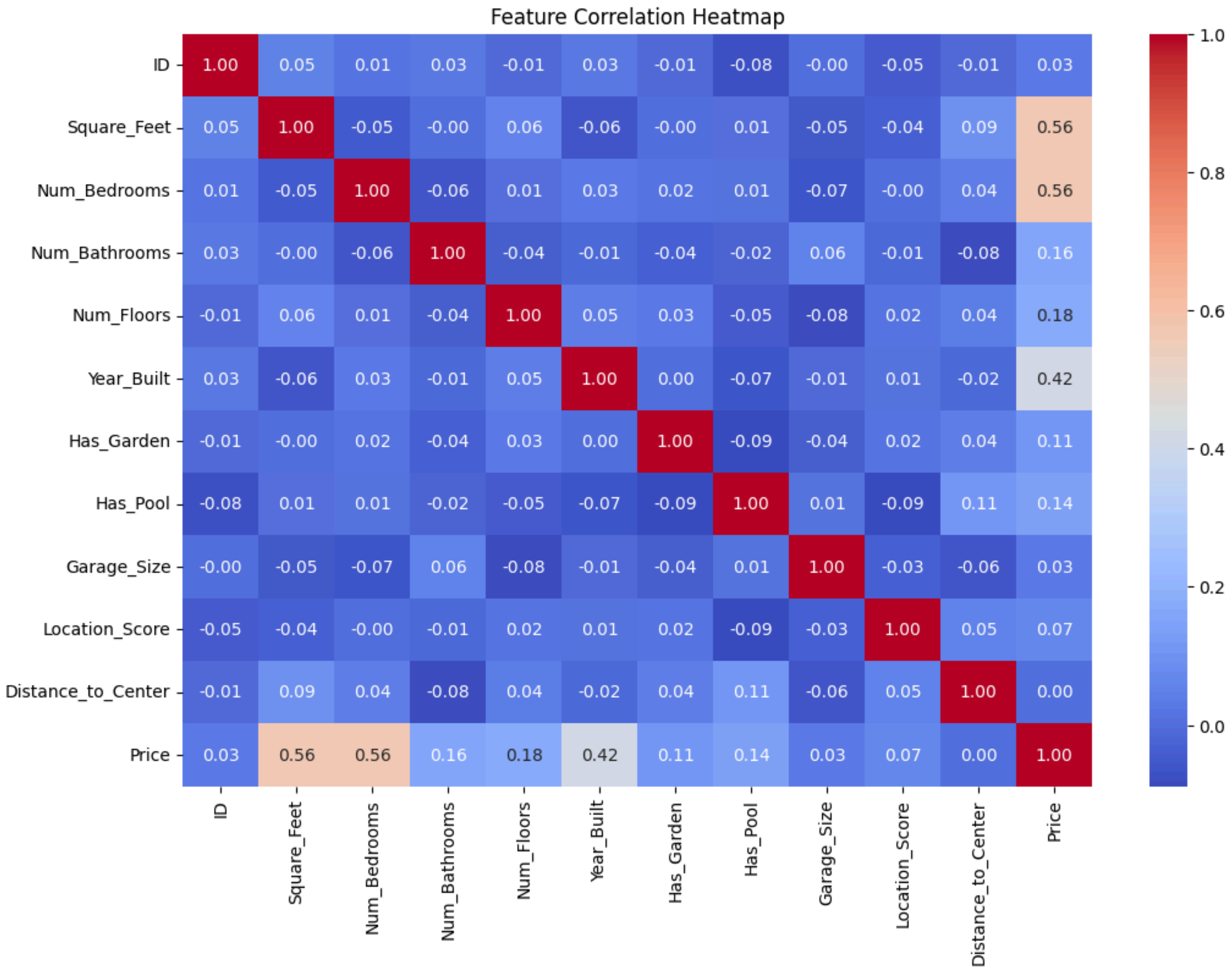
The above bar graphs compare properties that have gardens and/or pools. Both graphs demonstrate a similar relationship; properties with gardens and/or pools tend to cost more.







For the three scatter plots above–comparing price to the properties' distances to the city center, garage sizes, and location scores (a score from 0 to 10 indicating the quality of the neighborhood)– there seems to be little to no linear correlation between these features and the price as signified by the straight linear regression line in each regression plot.



	ID	Square_Feet	Num_Bedrooms	Num_Bathrooms	Num_Floors	Year_Built	Has_Garden	Has_Pool	Garage_Size	Location_Score	Distance_to_Center	Price
ID	1.00	0.05	0.01	0.03	-0.01	0.03	-0.01	-0.08	-0.00	-0.05	-0.01	0.03
Square_Feet	0.05	1.00	-0.05	-0.00	0.06	-0.06	-0.00	0.01	-0.05	-0.04	0.09	0.56
Num_Bedrooms	0.01	-0.05	1.00	-0.06	0.01	0.03	0.02	0.01	-0.07	-0.00	0.04	0.56
Num_Bathrooms	0.03	-0.00	-0.06	1.00	-0.04	-0.01	-0.04	-0.02	0.06	-0.01	-0.08	0.16
Num_Floors	-0.01	0.06	0.01	-0.04	1.00	0.05	0.03	-0.05	-0.08	0.02	0.04	0.18
Year_Built	0.03	-0.06	0.03	-0.01	0.05	1.00	0.00	-0.07	-0.01	0.01	-0.02	0.42
Has_Garden	-0.01	-0.00	0.02	-0.04	0.03	0.00	1.00	-0.09	-0.04	0.02	0.04	0.11
Has_Pool	-0.08	0.01	0.01	-0.02	-0.05	-0.07	-0.09	1.00	0.01	-0.09	0.11	0.14
Garage_Size	-0.00	-0.05	-0.07	0.06	-0.08	-0.01	-0.04	0.01	1.00	-0.03	-0.06	0.03
Location_Score	-0.05	-0.04	-0.00	-0.01	0.02	0.01	0.02	-0.09	-0.03	1.00	0.05	0.07
Distance_to_Center	-0.01	0.09	0.04	-0.08	0.04	-0.02	0.04	0.11	-0.06	0.05	1.00	0.00
Price	0.03	0.56	0.56	0.16	0.18	0.42	0.11	0.14	0.03	0.07	0.00	1.00

According to the above correlation heatmap, the square footage, number of bedrooms, and the year it was built held the most significance in determining the price of the properties. In relation to the previous three scatter plots, the garage size, location score, and distance to center all have correlations below 0.10.

Modeling and Interpretations

For each of the models, I chose to employ an 80-20 train-test split, where 80% of the data would be used to train my models while the other 20% would be used to test the scores of each model.

Baseline

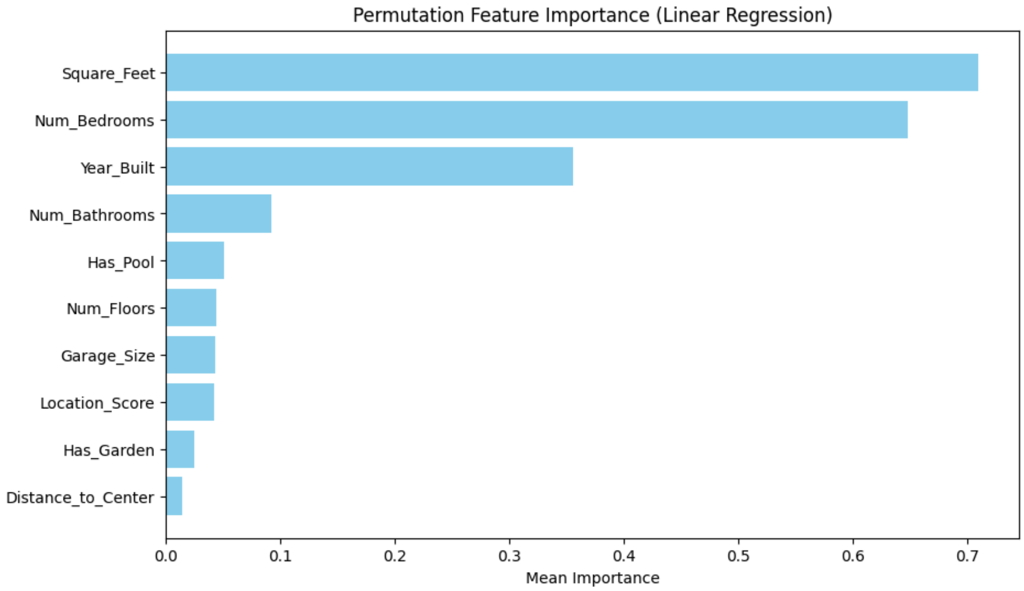
The baseline value that I used for my mean squared error will just be the mean of the dataset.
Baseline MSE (using mean price): 14,920,880,422.51

Multiple Regression Model

I chose to use a multiple regression model to capture potential non-linear relationships between the independent variables (features) and the dependent variable (Price). By extending the regression to include polynomial terms, I aimed to model more complex patterns that a simple linear model might miss, including possible combined effects such as between pools and gardens or multiple bedrooms, bathrooms, and floors.

However, after testing up to the 4th degree, the 1st degree model proved the best with the lowest test mean squared error.
Test MSE: 437,730,359.71

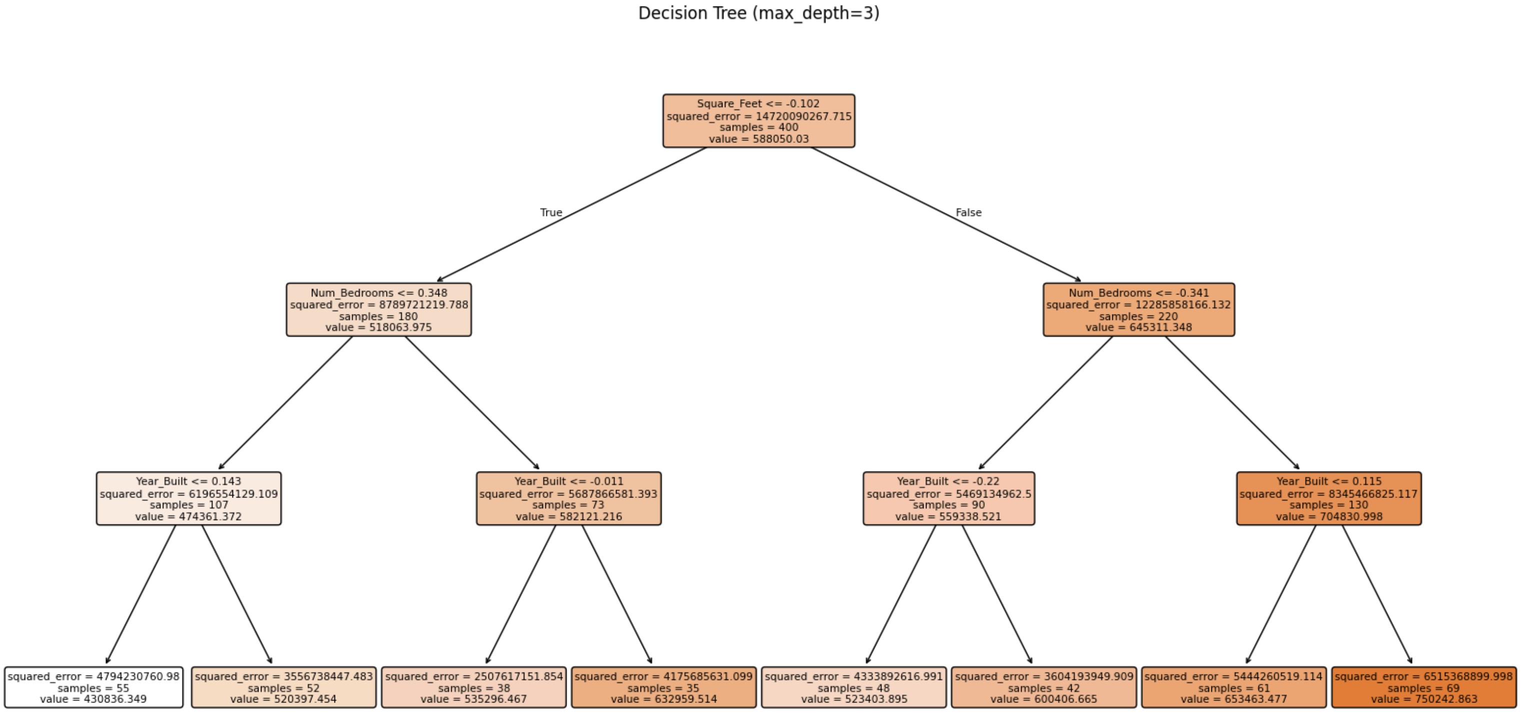
My multiple regression model performed far better than my baseline, effectively utilizing the independent variables to predict the price and create far more accurate predictions than simply relying on the mean price.



The most important features were the number of bedrooms, square footage, and the year the property was built, similar to the findings in the correlation heatmap before. On the other hand, the least important features were the distance to city center and whether there was a garden.

Decision Tree Regressor

I chose the Decision Tree Regressor as a model to evaluate feature importance and understand how individual features contribute to the prediction of property prices. Decision trees are interpretable and provide a clear visual representation of decision-making, providing further insight into the features influencing the price along with their non-linear relationships.



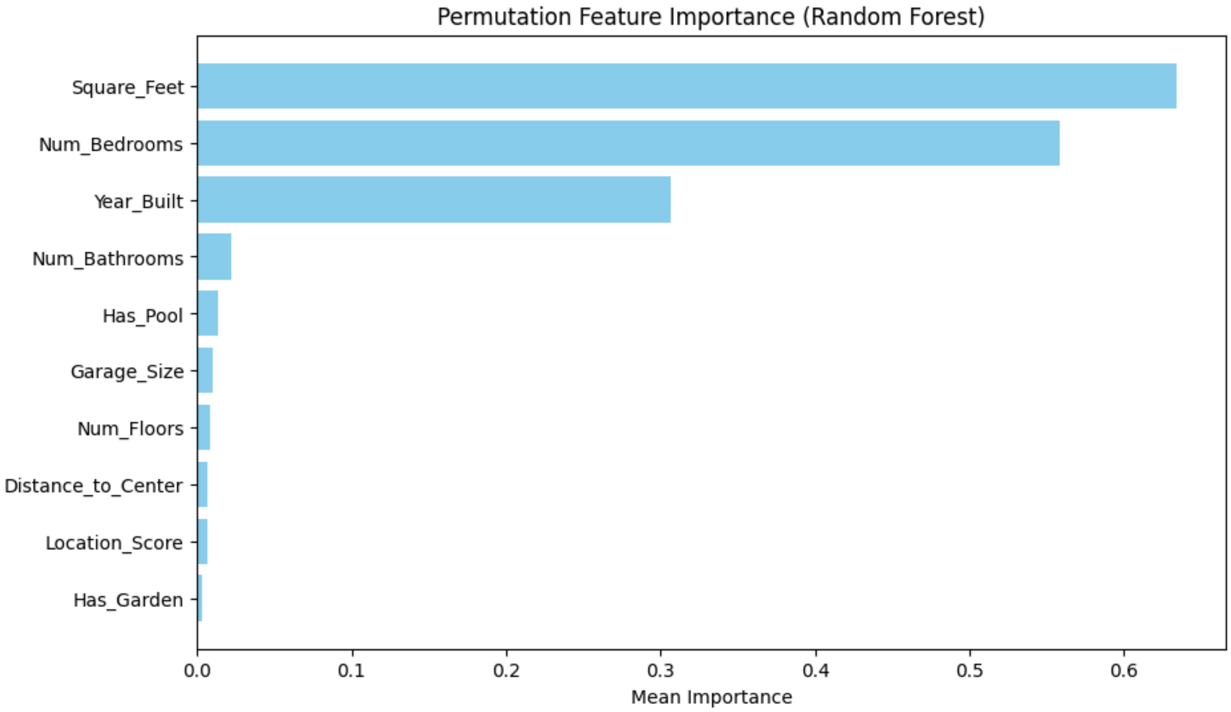
The square feet was the most important factor in the first split followed by the number of bedrooms and the year built of the properties, in accordance with the linear regression model and the correlation heatmap.

While the decision tree regressor had a better test mean squared error than the baseline, it did not outperform the multiple regression model. This may have been due to the dataset having a relatively small sample size (500 samples) and the decision tree regressor had insufficient data for reliable splits. Moreover, there may be high variance since the decision tree makes local splits which are sensitive to small changes in the data.
Optimal max_depth: 6
Minimum Test MSE: 4,851,212,956.41

Random Forest Regressor

Next, I selected the Random Forest Regressor because it is an ensemble method that combines multiple decision trees to improve accuracy and reduce overfitting. Moreover, Random Forest is well-suited for handling non-linear relationships and feature interactions, improving predictive performance over the Decision Tree Regressor.

While my random forest regressor did better than the baseline and the decision tree regressor (with a lower mean squared error), the multiple regression model still beat out the rest, still with the lowest mean squared error so far. It may have suffered from the same predicaments from the decision tree regressor, though it did optimize the best parameters which allowed it to perform better than its predecessor.
Test MSE: 2,683,887,888.89



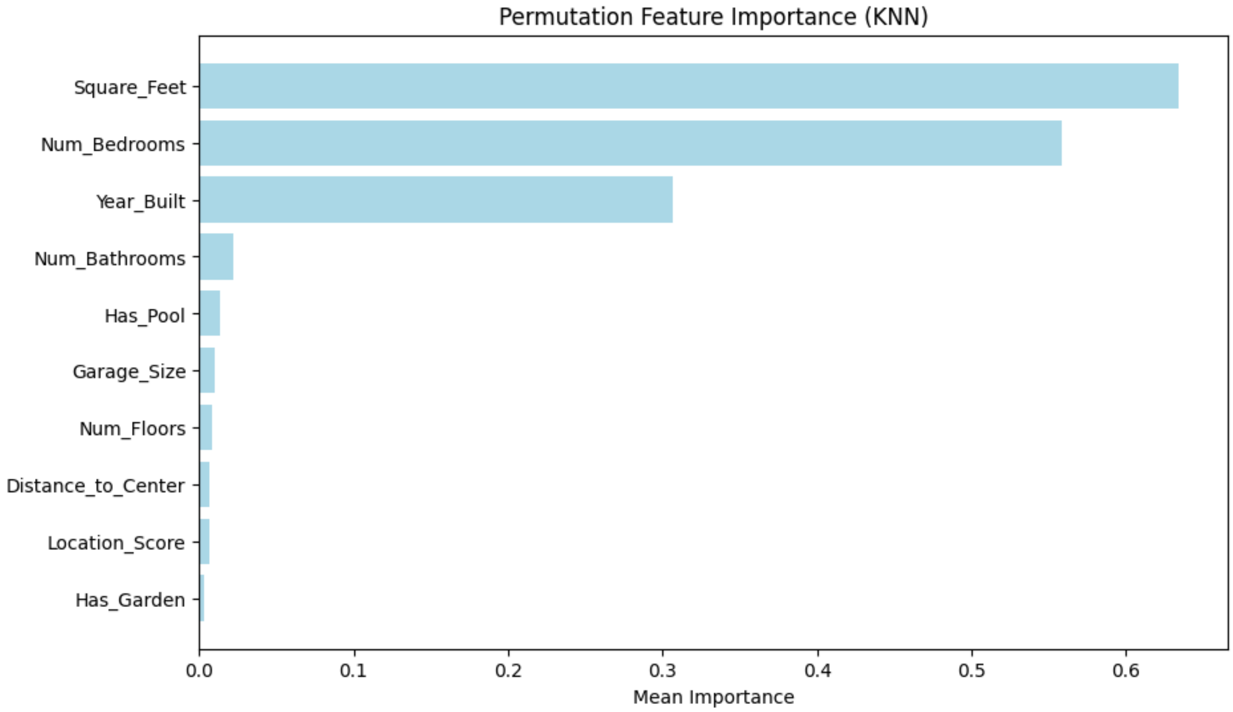
Once again, the square feet, number of bedrooms, and the year the properties were built remain the most important features.

K-Nearest Neighbors Regression Model

Lastly, I chose to try the k-nearest neighbors regression model because KNN makes predictions based on the similarity between data points. KNN can effectively capture local patterns in the data, making it suitable for clusters of similar properties especially by important features such as square feet or year built.

Test MSE: 4,615,926,745.49

My KNN model performed better than my decision tree and baseline predictions but was outperformed by my multiple regression model and random forest regressor. This may have been because the Random Forest Regressor and Multiple Regression Model could more accurately capture complex interactions and non-linear relationships. Random Forest Regressor also allowed it to ignore less important features.



Twice again, the square feet, number of bedrooms, and the year the properties were built remain the most important features.

Next Steps and Discussion

Summary of Findings

In my analysis of real estate prices, all the models I constructed performed better than the baseline predictor. The models ranked in performance from best to worst are: Multiple Linear Regression, Random Forest Regressor, K-Nearest Neighbors Regression, and Decision Tree Regression. A table is provided below with the test MSE and respective models.

Model	**Test MSE**
Multiple Linear Regression	437,730,359.71
Random Forest Regressor	2,683,887,888.89
K-Nearest Neighbors Regression	4,615,926,745.49
Decision Tree Regression	4,851,212,956.41
Baseline (mean)	14,920,880,422.51

Key Findings:

1. Multiple Linear Regression performed the best, producing the best predictive capabilities out of all the tested models.
2. Across most all models, the most impactful features were the square feet, number of bedrooms, and the year built of the properties.
3. While the number of bathrooms and whether there was a pool were consistently the 4th and 5th most impactful features respectively, the impact of whether the property had a garden, the score of the neighborhoods, and distance to the center of the city varied between models.

In conclusion, Multiple Linear Regression performed the best, achieving the lowest test MSE of 437,730,359.71. This result demonstrates the model's ability to effectively capture the linear relationships between the input features and the target variable (price). By leveraging multiple predictors simultaneously, the model was able to account for the collective influence of features such as square footage, number of bedrooms, and year built, which consistently emerged as the most significant variables across all analyses. Despite its simplicity compared to more complex models like the Random Forest Regressor, its generalization ability and performance prove that, in this dataset, linear relationships were dominant in determining property prices. This model provides stakeholders with a reliable, interpretable, and efficient solution for real estate price prediction, forming a solid foundation for further enhancements.

Next Steps

To further enhance the predictive capabilities of the models and gain deeper insights into real estate property prices, I propose to:

1. Incorporate External Data Sources:
    a. Neighborhood and Crime Statistics: Adding crime rates, school quality ratings, and other neighborhood-level data would provide more context for the location score, potentially improving predictions.
    b. Market Trends: Consider real estate market trends, such as historical price changes and economic indicators like interest rates, to capture broader influences on property prices.
    c. Economic Indicators: Integrating regional economic data, such as employment rates, median income, and housing demand, could offer additional explanatory power.
    d. Address sparsity in features like pools, gardens, and multi-floor properties to ensure balanced representation across all property types.
2. Implement Advanced Modeling Techniques:
    a. Gradient Boosting Models: Implement ensemble methods such as XGBoost or LightGBM, which can optimize performance and better handle non-linear interactions.
    b. Regularization Techniques: Use Lasso or Ridge Regression to reduce overfitting while enhancing the interpretability of coefficients.
3. Use Geospatial Analysis:
    a. Conduct geospatial mapping of property prices based on latitude and longitude (if available) to identify trends within specific geographic clusters. Heatmaps could highlight areas with higher property values.
4. Address Data Limitations:
    a. Expand the dataset to include more real estate properties from diverse geographic regions. A larger and more varied dataset will improve model generalization.
    b. Address the imbalance in features like the number of floors or pools to ensure better representation across property types.
5. Expand the Dataset and Further Test the Model:
    a. Gather additional data from various regions and apply the models to improve the diversity and generalizability of the models, further improving and testing the multiple linear regression model.

By incorporating additional data, refining features, and exploring advanced machine learning techniques, the models can be improved further to provide more robust and actionable insights into real estate property pricing. This approach would enhance the accuracy of predictions and the overall utility of the analysis for stakeholders in the real estate market.