



# NYC HOUSING PRICE PREDICTION

BY  
HEATHER ADLER

## PRESENTATION PURPOSE

This capstone project aims to predict house prices in New York City using a combination of historical data and advanced machine learning techniques. By leveraging extensive datasets and performing comprehensive exploratory data analysis (EDA), we developed a robust predictive model that provides accurate price estimations for various property types across different boroughs.

## PRESENTATION GOALS

Provide strategic insights and recommendations for investors interested in the New York City market.

## CHALLENGES

- Complexity of Data
- Model Generalization

## PROBLEM STATEMENT

How can predictive analytics drive success in the New York City real estate market by accurately forecasting house prices and understanding the key factors influencing property values across different boroughs and neighborhoods?



# DATA WRANGLING

The dataset for this project was downloaded from Kaggle and has been filtered and cleaned to include the following key data features:

## 1) Property Characteristics:

- Square Footage (sqft)
- Number of Bedrooms
- Number of Bathrooms
- Year Built
- Property Type (Single Family, Condo, etc.)
- Heating and Cooling Features

## 3) Financial Details:

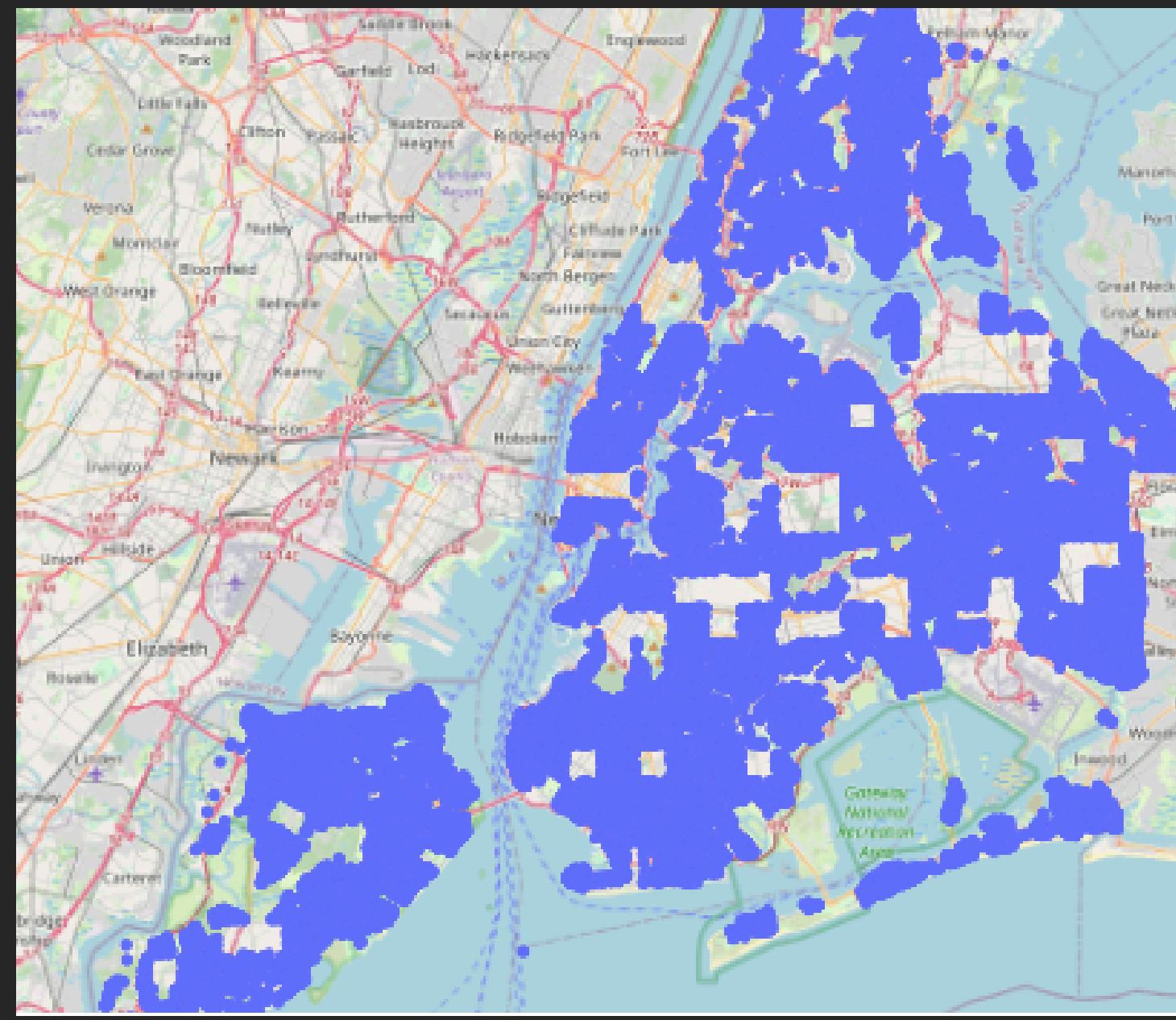
- Sale Price
- Assessed Value
- Tax Amount

## 2) Location Information:

- Borough (Bronx, Brooklyn, Manhattan, Queens, Staten Island)
- Neighborhood
- Latitude and Longitude
- Zip Code

## 4) Additional Features:

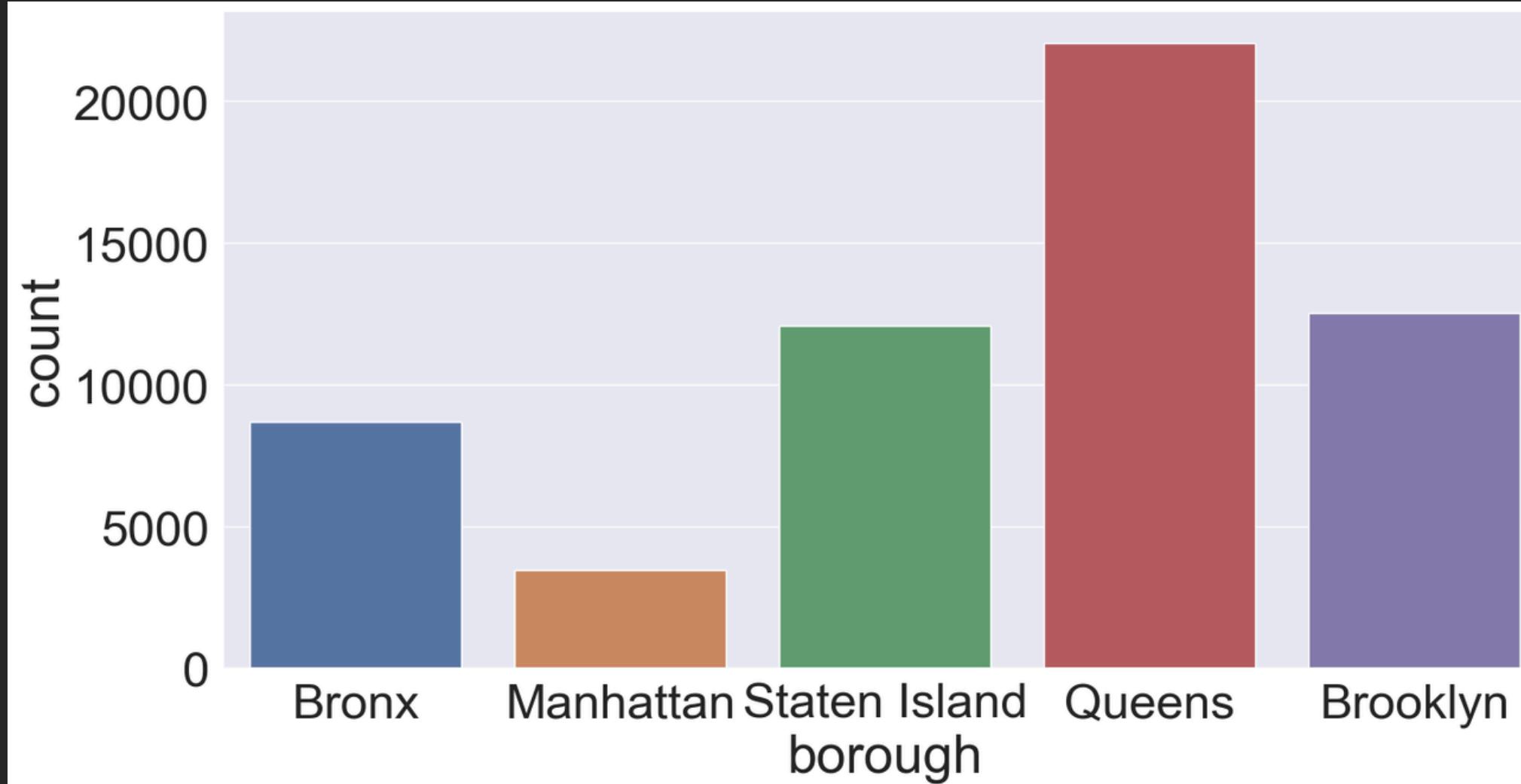
- School Ratings
- Parking Availability
- Basement Presence
- Lot Size



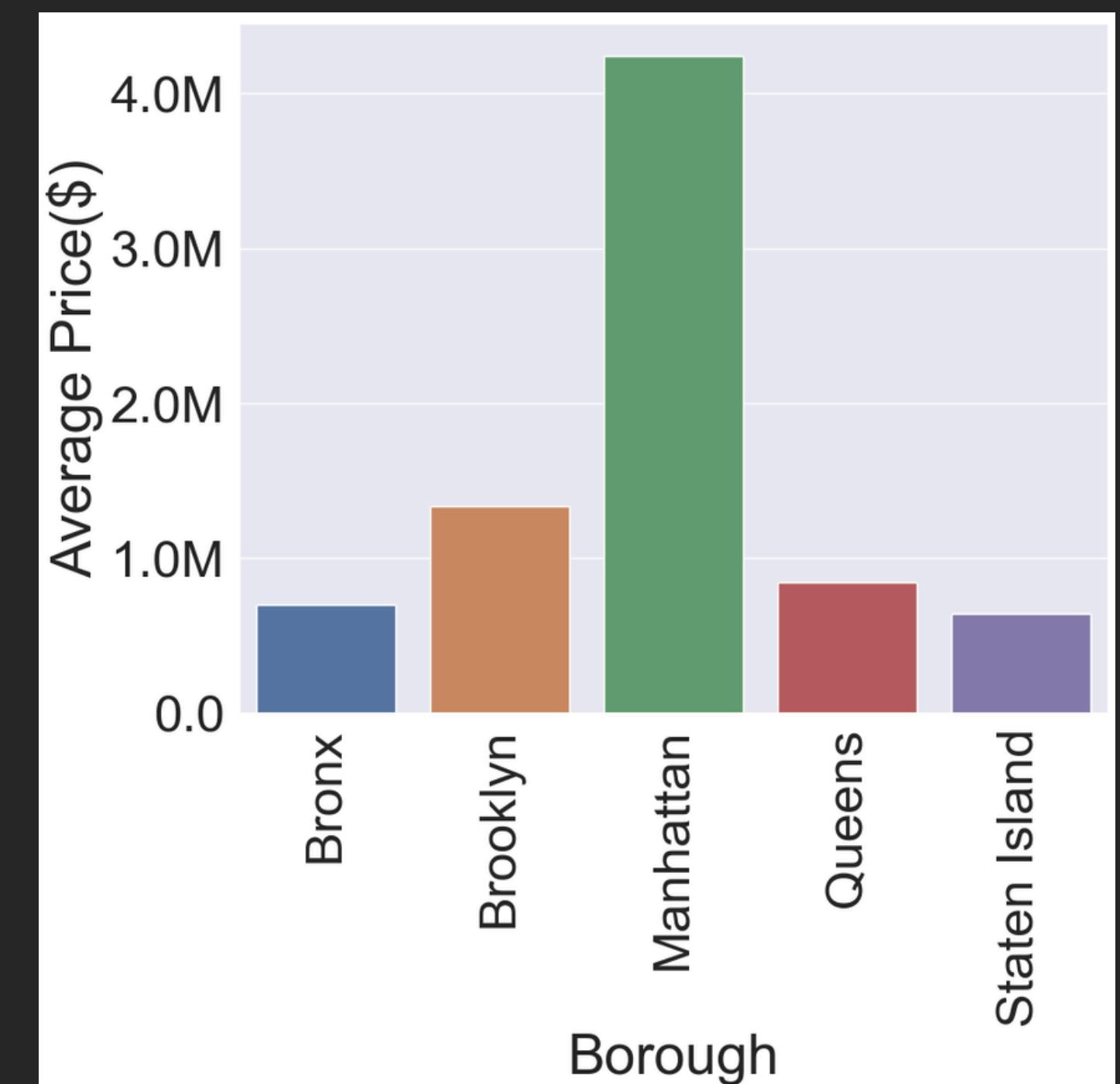
# Data Exploration

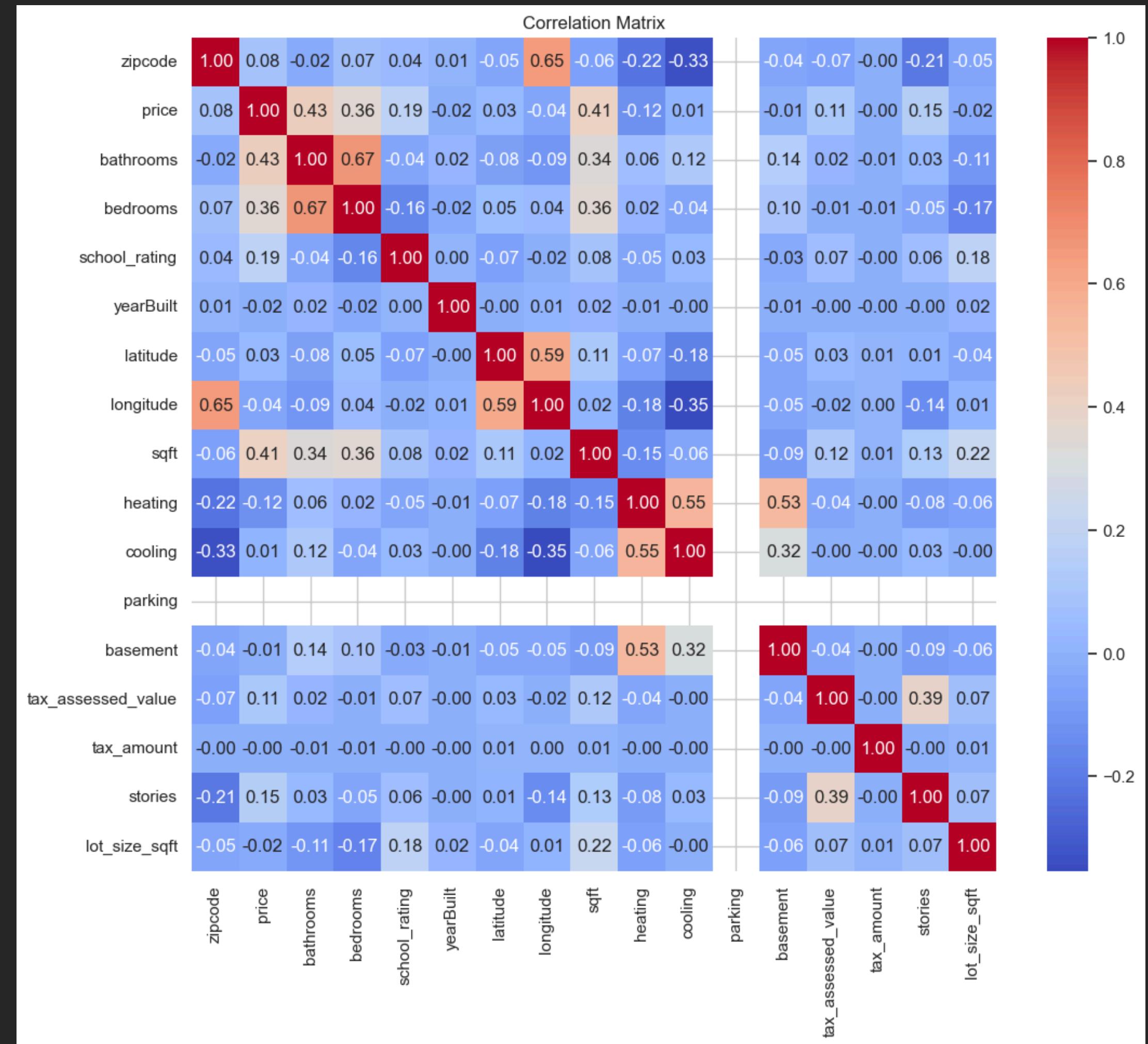


# NUMBER OF HOUSES PER BOROUGH

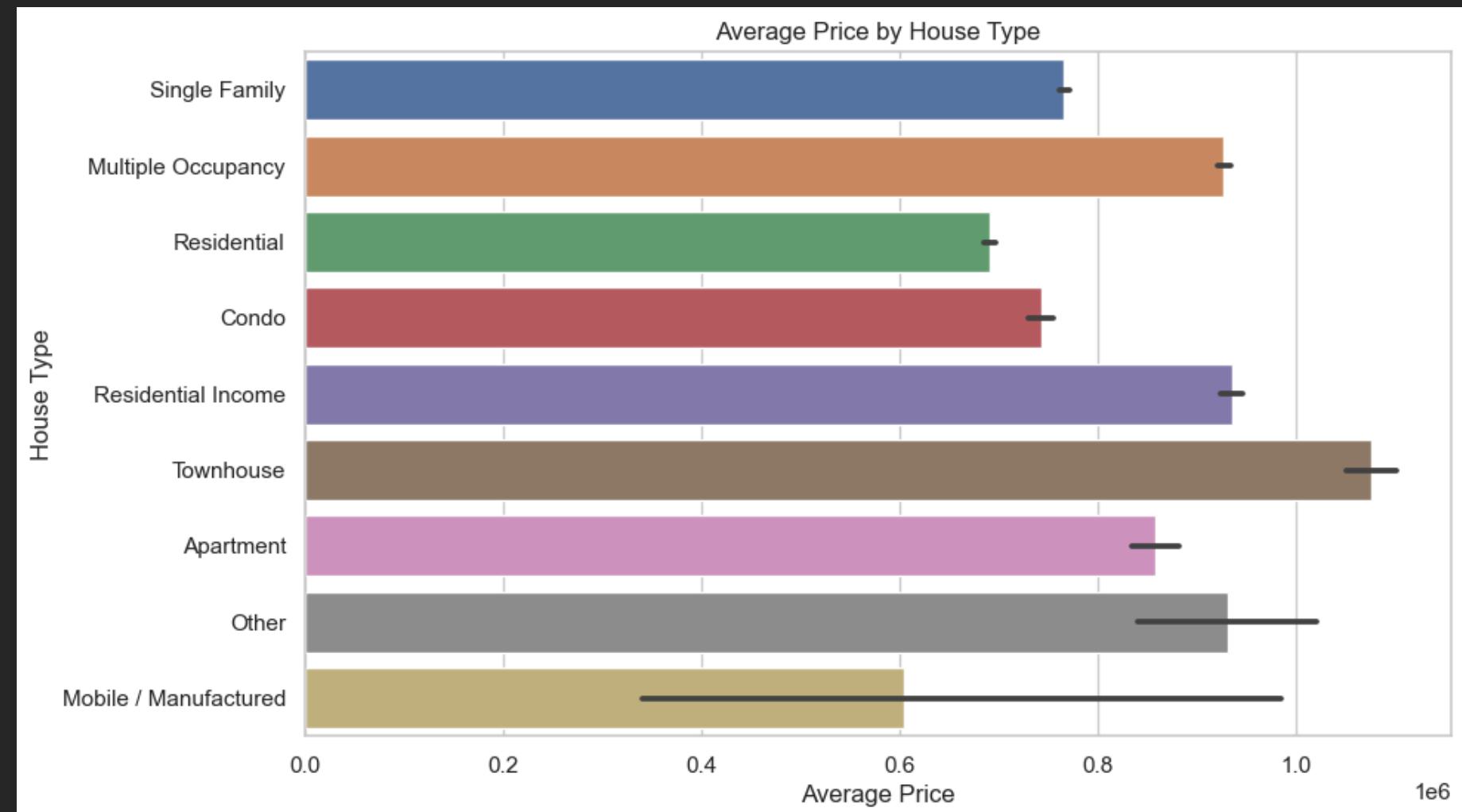
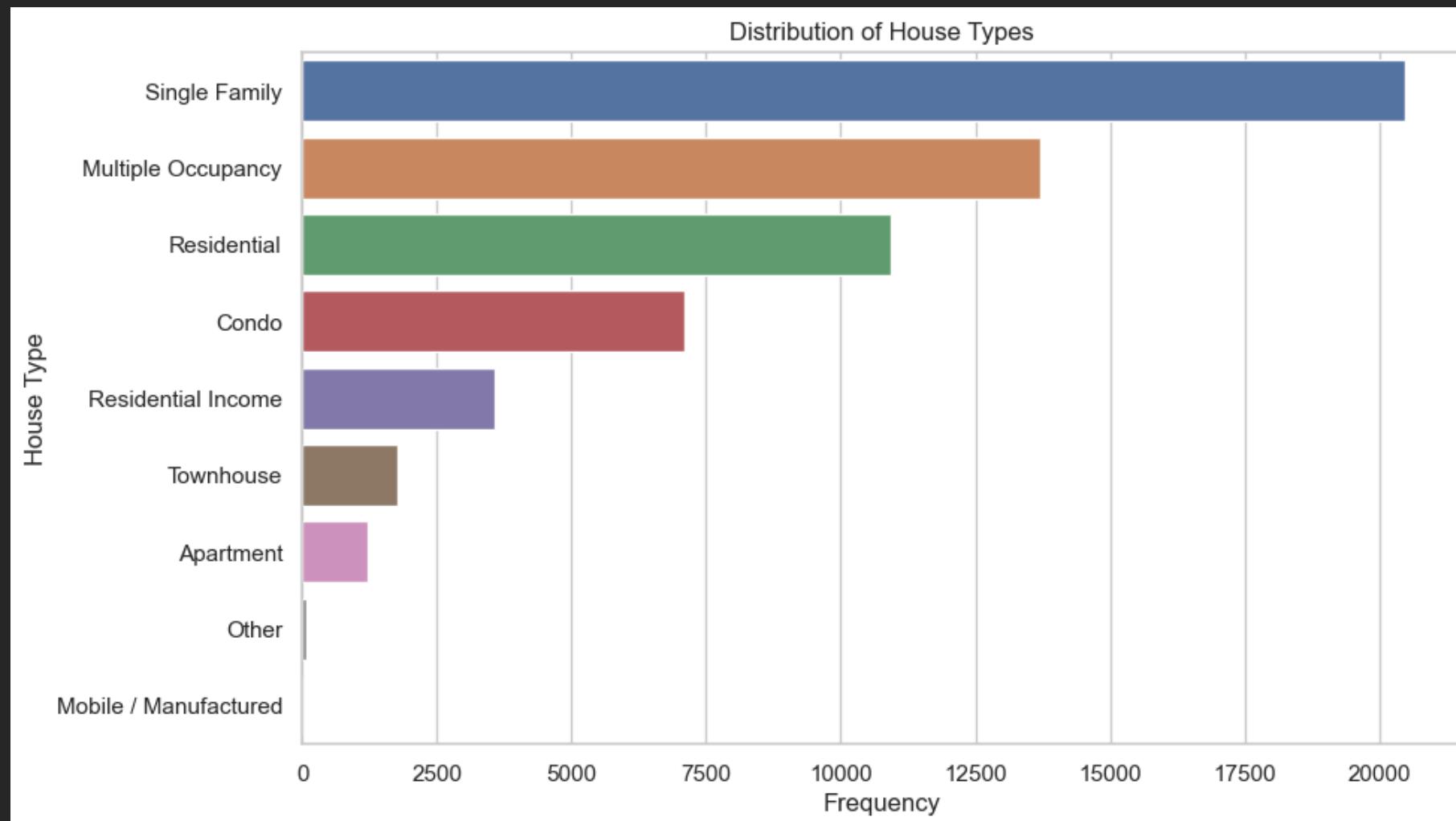


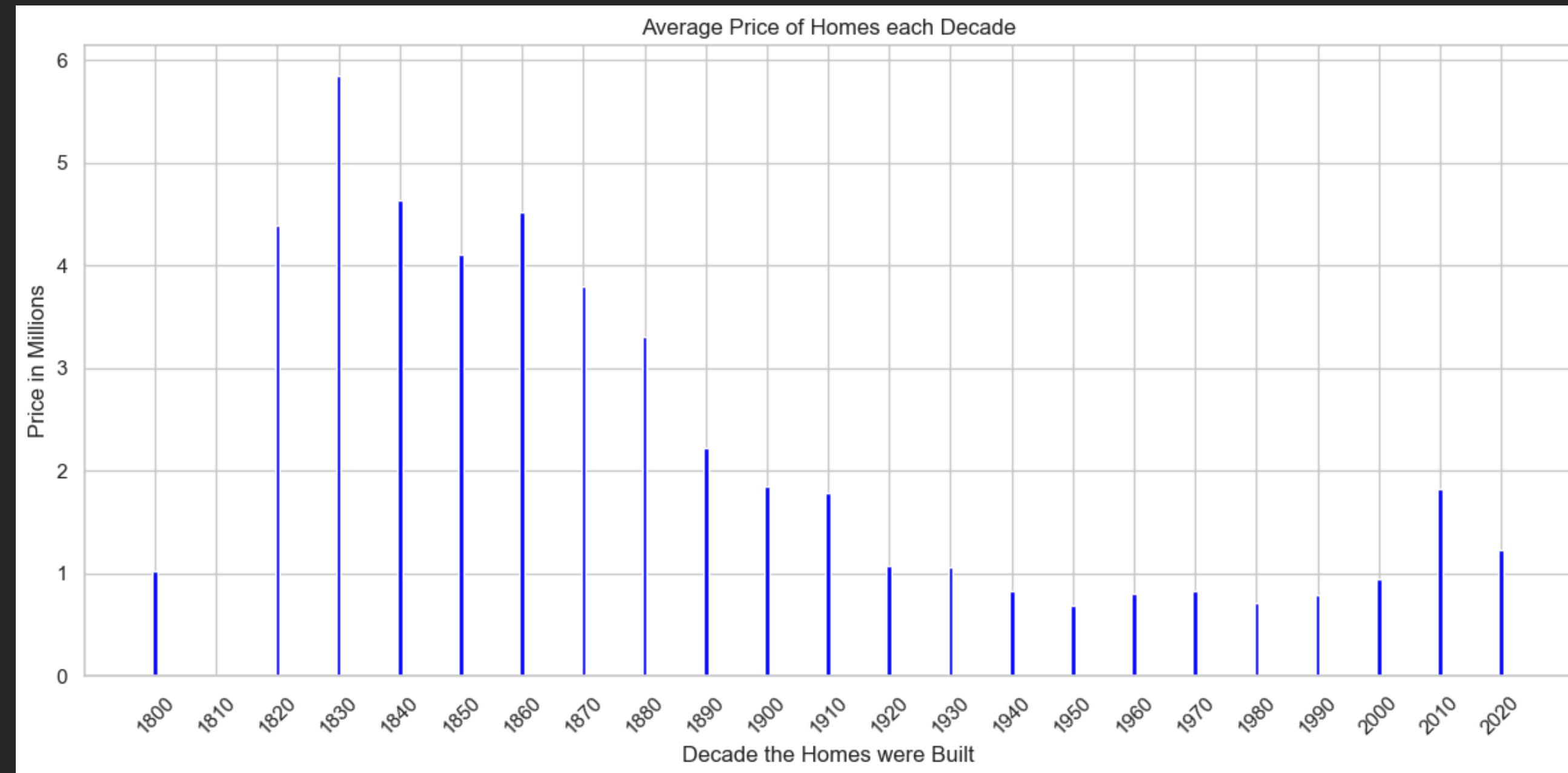
# AVERAGE PRICE OF HOUSES BY BOROUGHS

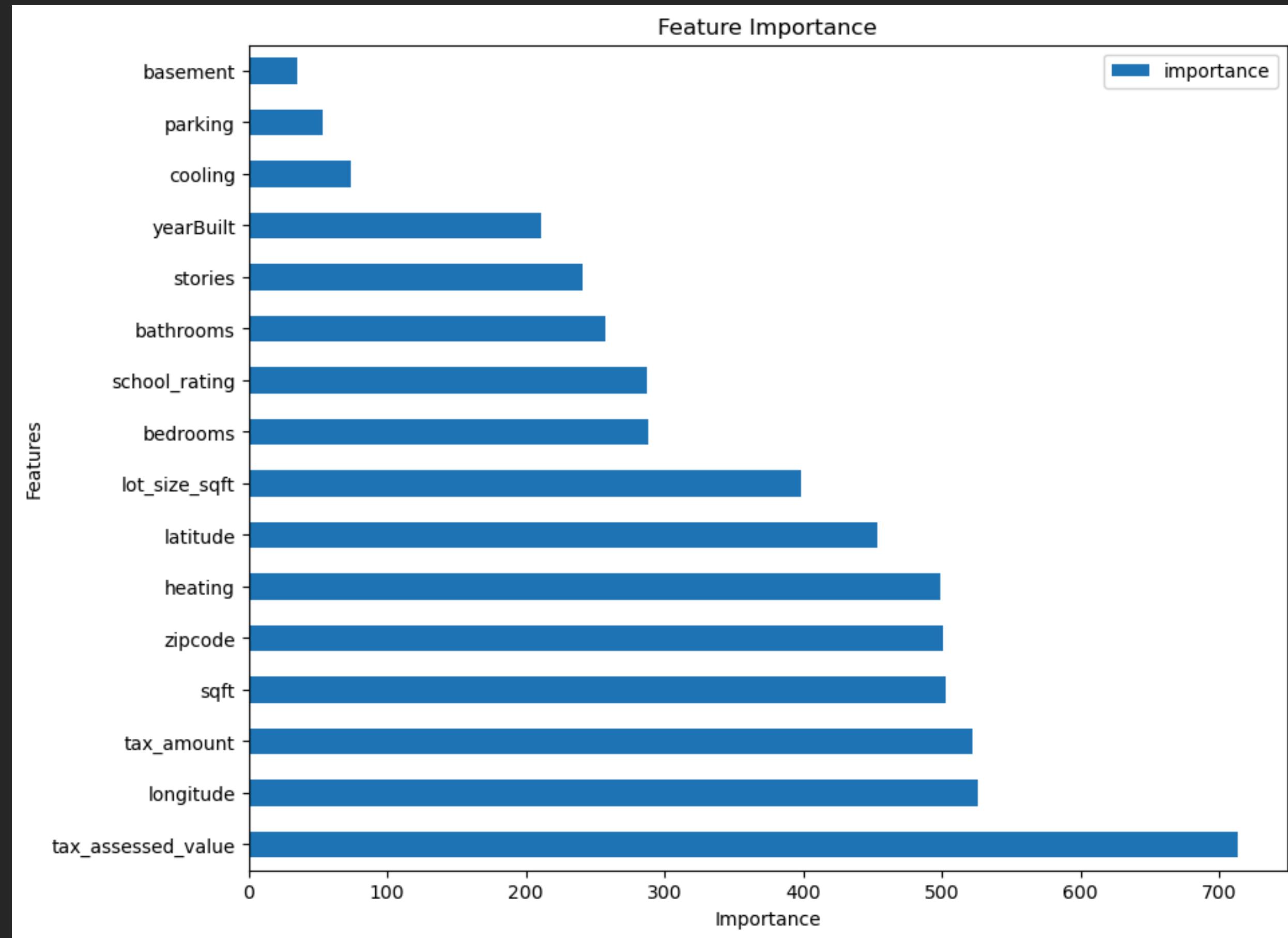




# PROPERTY TYPES







# Data Modeling

An aerial photograph of a major city skyline during sunset. The sky is filled with warm, orange and yellow clouds. In the foreground, there's a dense cluster of skyscrapers of various heights and architectural styles. A prominent bridge spans a river or bay, connecting different parts of the city. The overall atmosphere is one of urban density and modernity.

# DEFINING & TRAINING THE MODEL

**Model Selection:** We explored several regression models to identify the best fit for predicting NYC house prices:

- GradientBoostingRegressor
- Random Forest Regressor
- XGBoost Regressor

**Ultimately, the XGBoost Regressor Model was selected as it was the best performing model with the lowest RMSE of .2870797035074488**

# MODEL EVALUATION & PREDICTION

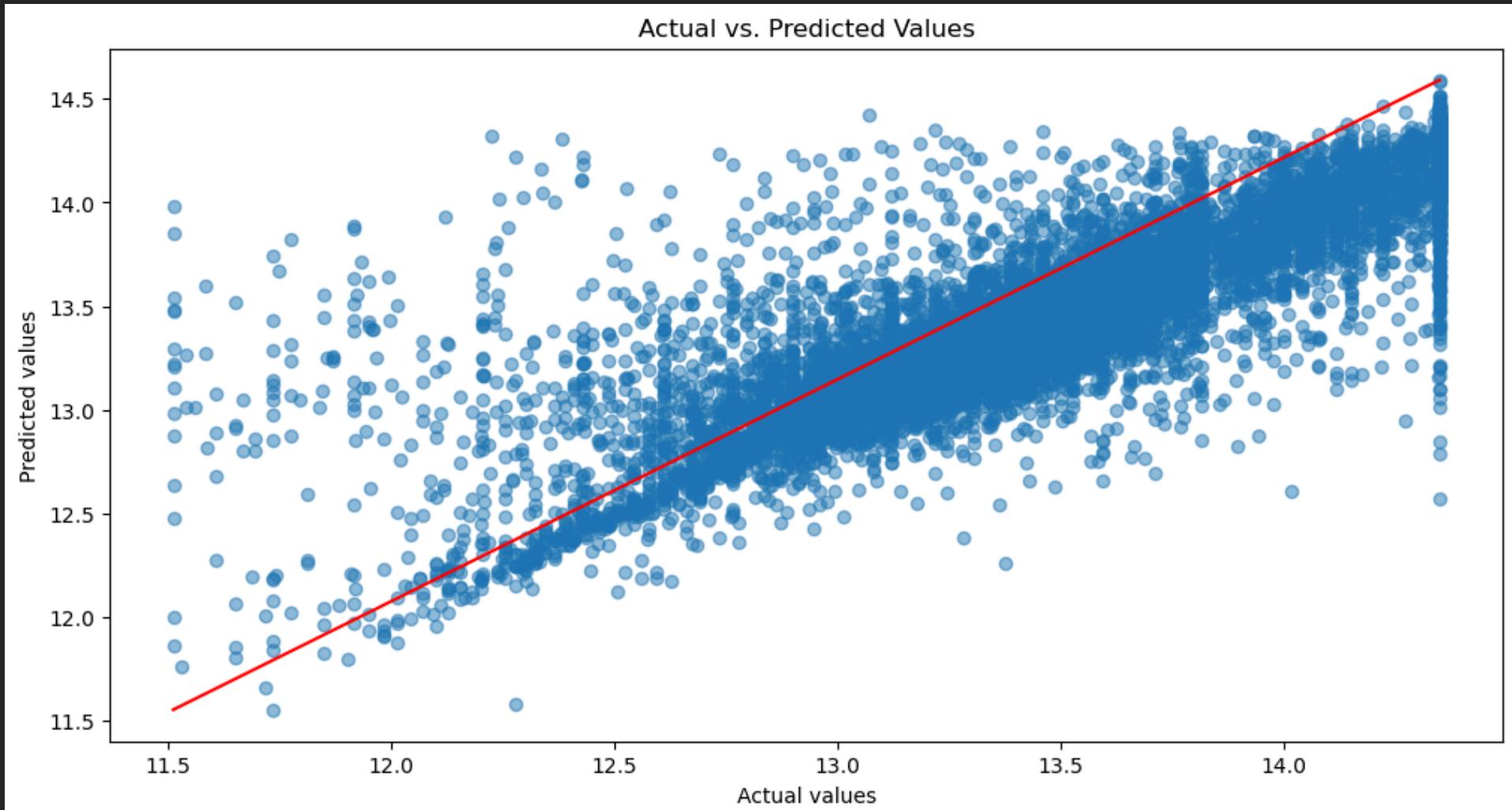
## Reasoning for RMSE Metric for Evaluation

- **Interpretability:** RMSE is in the same unit as the target variable (price), making it easier to interpret and understand the magnitude of errors.
- **Sensitivity to Outliers:** RMSE penalizes larger errors more than smaller errors due to the squaring of the error term. This makes RMSE a good metric when large errors are particularly undesirable.
- **Standard in Regression:** RMSE is a commonly used metric in regression tasks, providing a standardized way to measure the performance of predictive models.
- **Direct Comparison:** RMSE allows for direct comparison between different models and helps in selecting the model with the least error, ensuring the most accurate predictions.

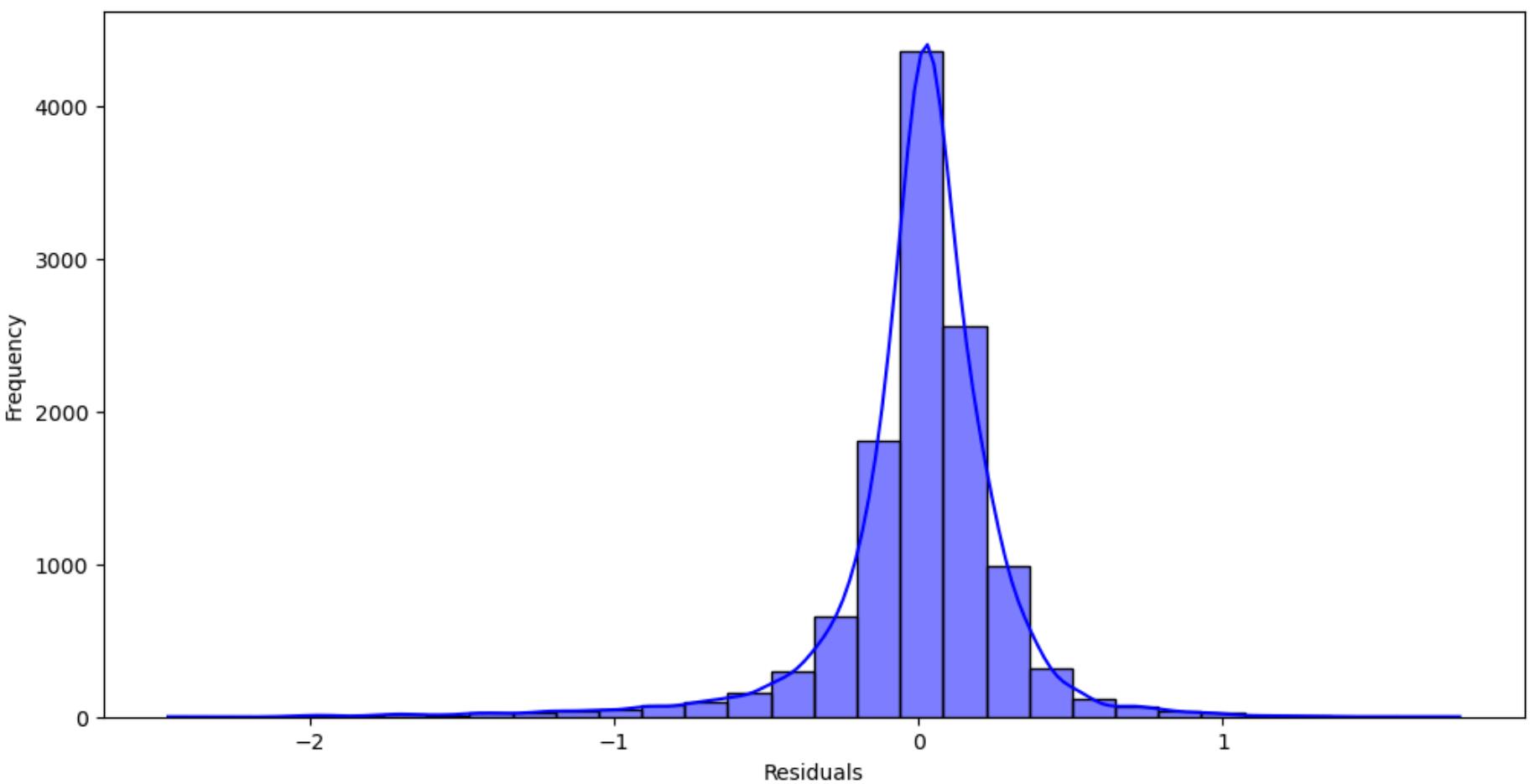
## Comparison with Other Metrics:

- **Mean Absolute Error (MAE):** RMSE squares the errors before averaging, which means it penalizes larger errors more than MAE. In real estate, large prediction errors can be particularly costly or misleading, so a metric that emphasizes these is beneficial. RMSE is also more sensitive to outliers than MAE.
- **MSE (Mean Squared Error):** While both MSE and RMSE penalize larger errors, RMSE is in the same unit as the target variable (price). This makes RMSE more interpretable than MSE, which is in squared units of the target variable.

Actual vs. Predicted Values



Distribution of Residuals



# CONCLUSIONS

- The XGBRegressor model outperformed GradientBoosting and RandomForest models with the lowest RMSE, indicating its superior predictive accuracy for the dataset.
- RMSE for the best model (XGBRegressor) on the test set is approximately 0.287, indicating the model's predictions are close to the actual values.
- The model can be used to predict house prices with reasonable accuracy, which can assist stakeholders in making informed decisions.
- Further refinement and tuning of the model, including exploring additional features or handling data imbalances, can potentially improve accuracy.

# MORE IDEAS TO IMPROVE THE MODEL IN THE FUTURE

- **Feature Engineering:**
  - Incorporate additional features such as proximity to amenities ( parks, public transport), neighborhood crime rates, and economic indicators (employment rates).
  - Create interaction features that might capture complex relationships between existing features, such as combining location with house attributes.
- **Handling Imbalanced Data:**
  - Use techniques like SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset if there are imbalances in price ranges or property types.
  - Explore weighted loss functions in the modeling process to handle imbalance better.
- **Model Interpretability:**
  - Enhance model interpretability using tools like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to understand feature contributions better.