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.

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?

PRESENTATION GOALS

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

CHALLENGES

- Complexity of Data
- Model Generalization

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

2) Location Information:

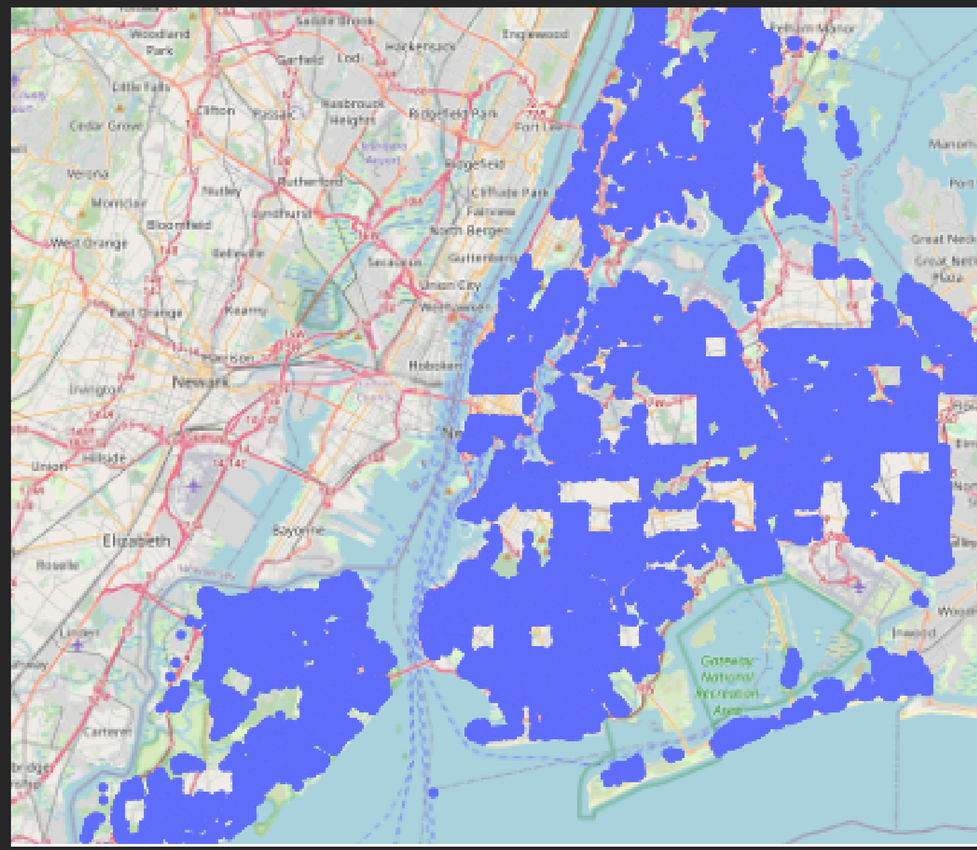
- Borough (Bronx, Brooklyn, Manhattan, Queens, Staten Island)
- Neighborhood
- Latitude and Longitude
- Zip Code

3) Financial Details:

- Sale Price
- Assessed Value
- Tax Amount

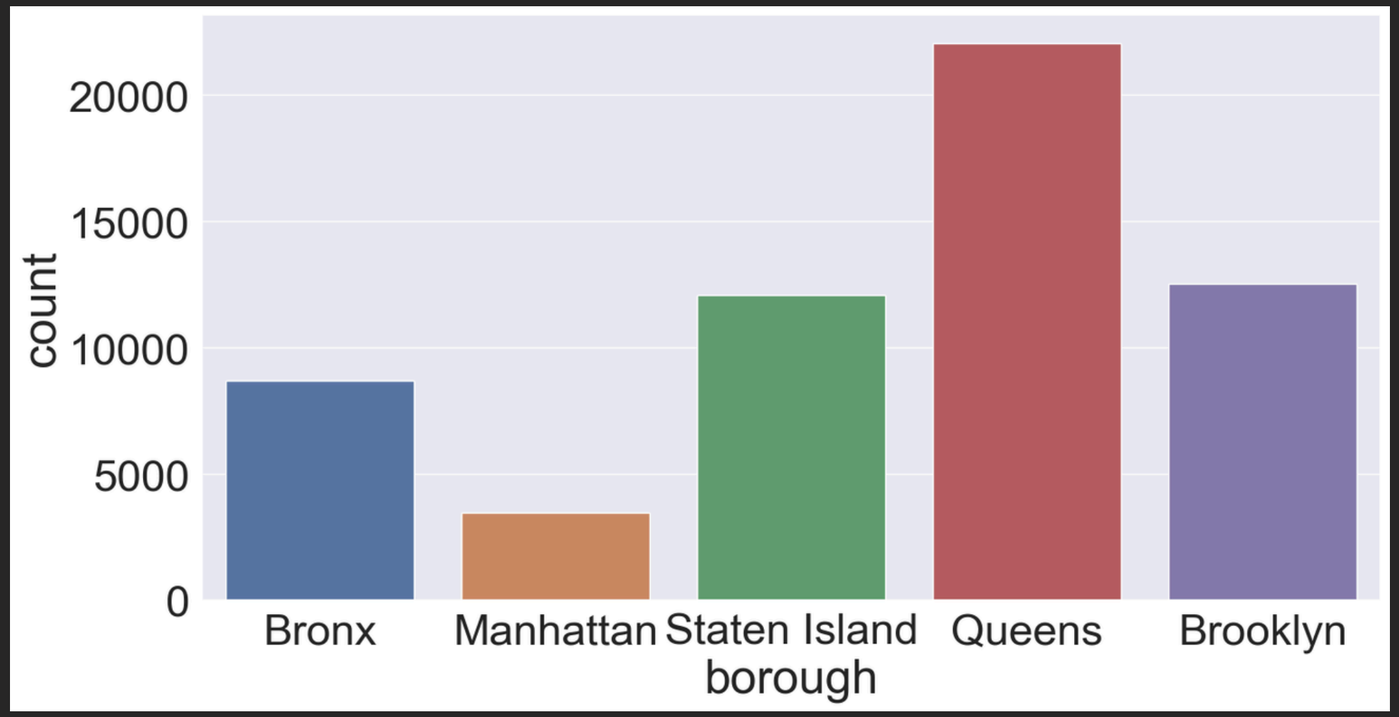
4) Additional Features:

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

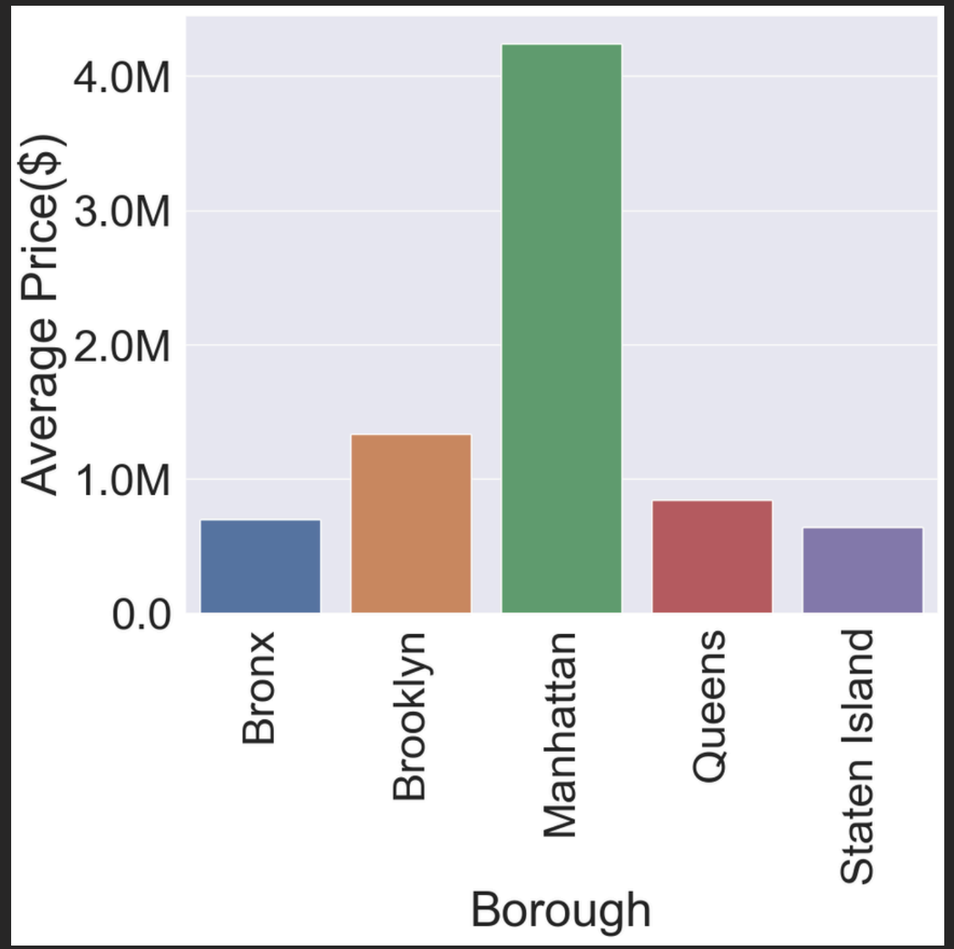


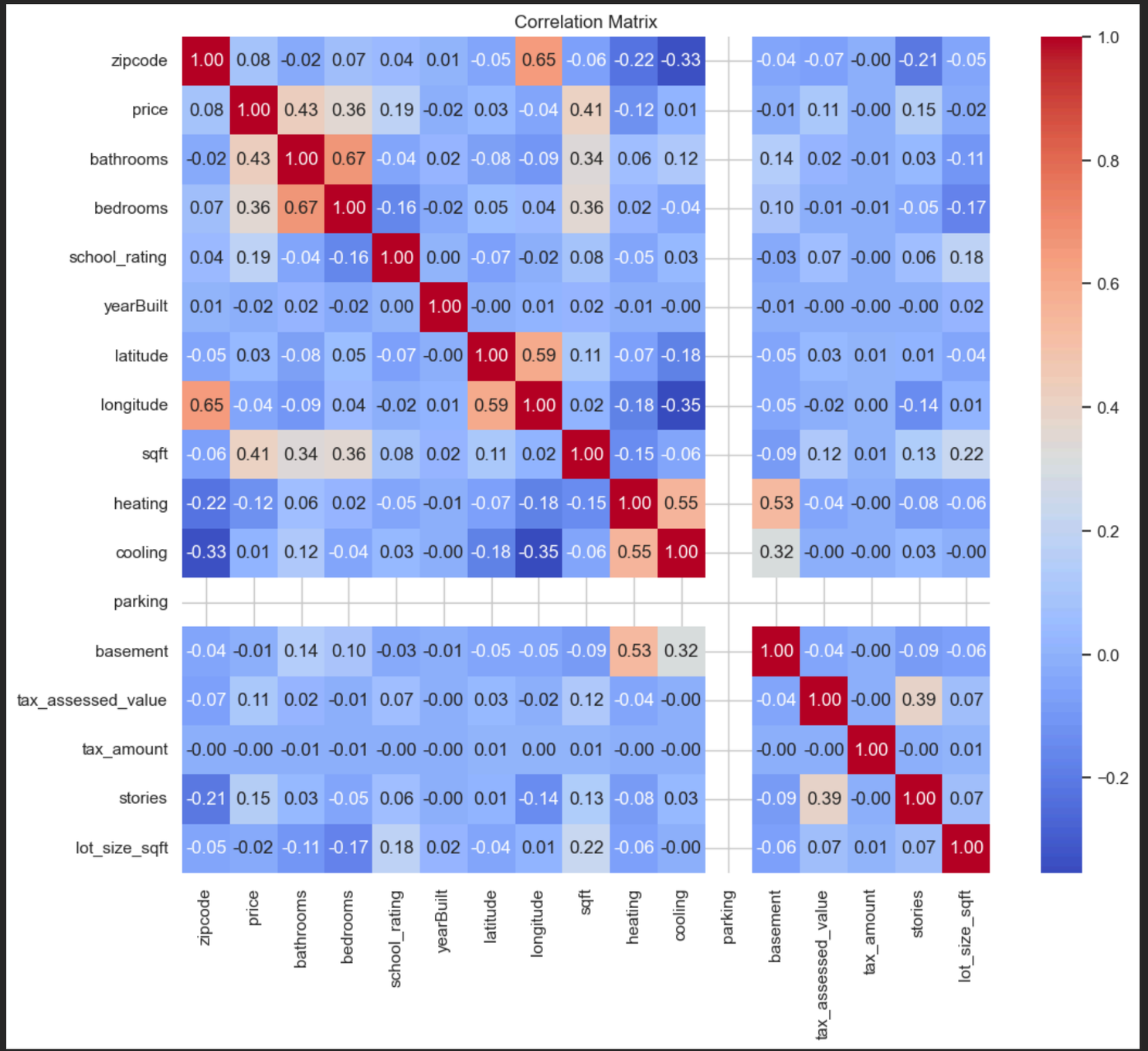
Data Exploration

NUMBER OF HOUSES PER BOROUGH

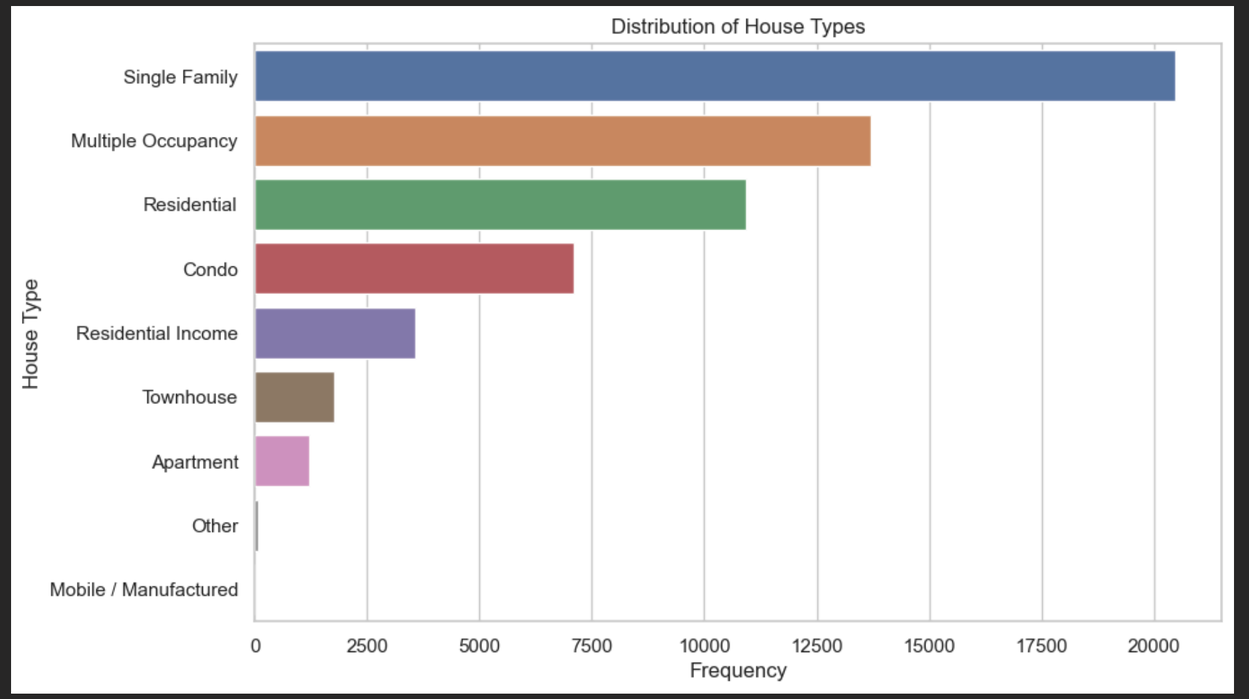


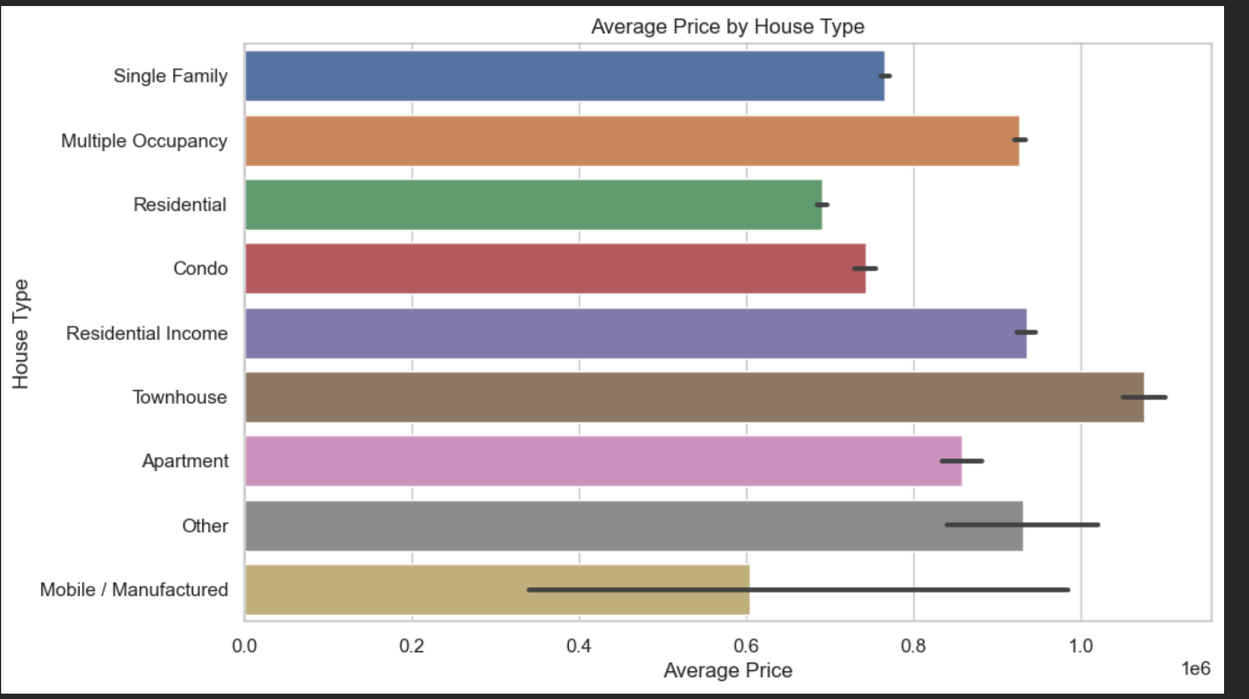
AVERAGE PRICE OF HOUSES BY BOROUGHS

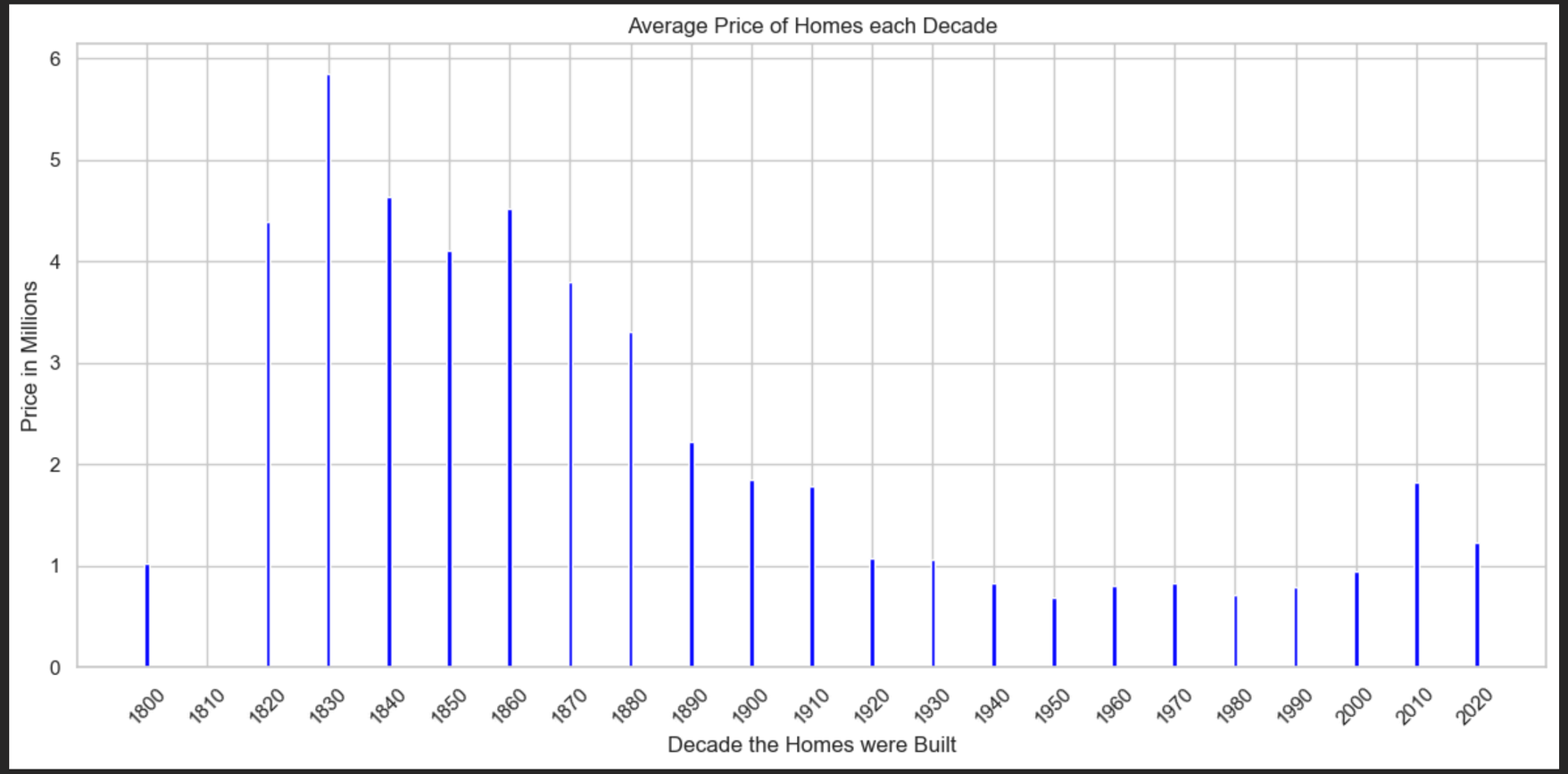


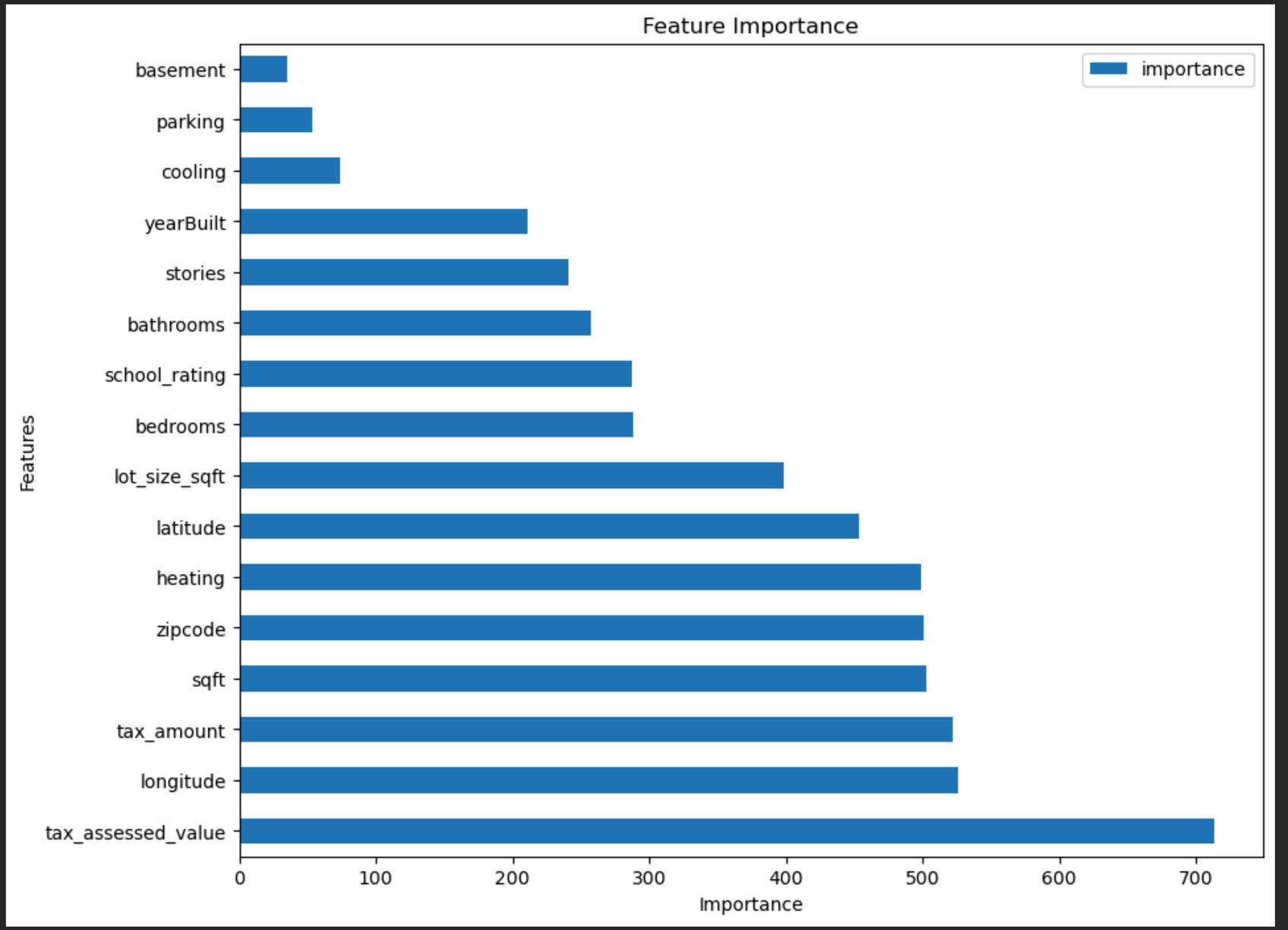

Correlation Matrix
zipcode 1.00 0.08 -0.02 0.07 0.04 0.01 -0.05 0.65 -0.06 -0.22 -0.33 -0.04 -0.07 -0.00 -0.21 -0.05
price 0.08 1.00 0.43 0.36 0.19 -0.02 0.03 -0.04 0.41 -0.12 0.01 -0.01 0.11 -0.00 0.15 -0.02
bathrooms -0.02 0.43 1.00 0.67 -0.04 0.02 -0.08 -0.09 0.34 0.06 0.12 0.14 0.02 -0.01 0.03 -0.11
bedrooms 0.07 0.36 0.67 1.00 -0.16 -0.02 0.05 0.04 0.36 0.02 -0.04 0.10 -0.01 -0.01 -0.05 -0.17
school_rating 0.04 0.19 -0.04 -0.16 1.00 0.00 -0.07 -0.02 0.08 -0.05 0.03 -0.03 0.07 -0.00 0.06 0.18
yearBuilt 0.01 -0.02 0.02 -0.02 0.00 1.00 -0.00 0.01 0.02 -0.01 -0.00 -0.01 -0.00 -0.00 -0.00 0.02
latitude -0.05 0.03 -0.08 0.05 -0.07 -0.00 1.00 0.59 0.11 -0.07 -0.18 -0.05 0.03 0.01 0.01 -0.04
longitude 0.65 -0.04 -0.09 0.04 -0.02 0.01 0.59 1.00 0.02 -0.18 -0.35 -0.05 -0.02 0.00 -0.14 0.01
sqft -0.06 0.41 0.34 0.36 0.08 0.02 0.11 0.02 1.00 -0.15 -0.06 -0.09 0.12 0.01 0.13 0.22
heating -0.22 -0.12 0.06 0.02 -0.05 -0.01 -0.07 -0.18 -0.15 1.00 0.55 0.53 -0.04 -0.00 -0.08 -0.06
cooling -0.33 0.01 0.12 -0.04 0.03 -0.00 -0.18 -0.35 -0.06 0.55 1.00 0.32 -0.00 -0.00 0.03 -0.00
parking
basement -0.04 -0.01 0.14 0.10 -0.03 -0.01 -0.05 -0.05 -0.09 0.53 0.32 1.00 -0.04 -0.00 -0.09 -0.06
tax_assessed_value -0.07 0.11 0.02 -0.01 0.07 -0.00 0.03 -0.02 0.12 -0.04 -0.00 -0.04 1.00 -0.00 0.39 0.07
tax_amount -0.00 -0.00 -0.01 -0.01 -0.00 -0.00 0.01 0.00 0.01 -0.00 -0.00 -0.00 -0.00 1.00 -0.00 0.01
stories -0.21 0.15 0.03 -0.05 0.06 -0.00 0.01 -0.14 0.13 -0.08 0.03 -0.09 0.39 -0.00 1.00 0.07
lot_size_sqft -0.05 -0.02 -0.11 -0.17 0.18 0.02 -0.04 0.01 0.22 -0.06 -0.00 -0.06 0.07 0.01 0.07 1.00
zipcode price bathrooms bedrooms school_rating yearBuilt latitude longitude sqft heating cooling parking basement tax_assessed_value tax_amount stories lot_size_sqft
1.0 0.8 0.6 0.4 0.2 0.0 −0.2

PROPERTY TYPES






Average Price of Homes each Decade
Price in Millions
0
1
2
3
4
5
6
1800
1810
1820
1830
1840
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
Decade the Homes were Built


Feature Importance
importance
basement
parking
cooling
yearBuilt
stories
bathrooms
school_rating
bedrooms
lot_size_sqft
latitude
heating
zipcode
sqft
tax_amount
longitude
tax_assessed_value
Features
0
100
200
300
400
500
600
700
Importance

Data Modeling

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

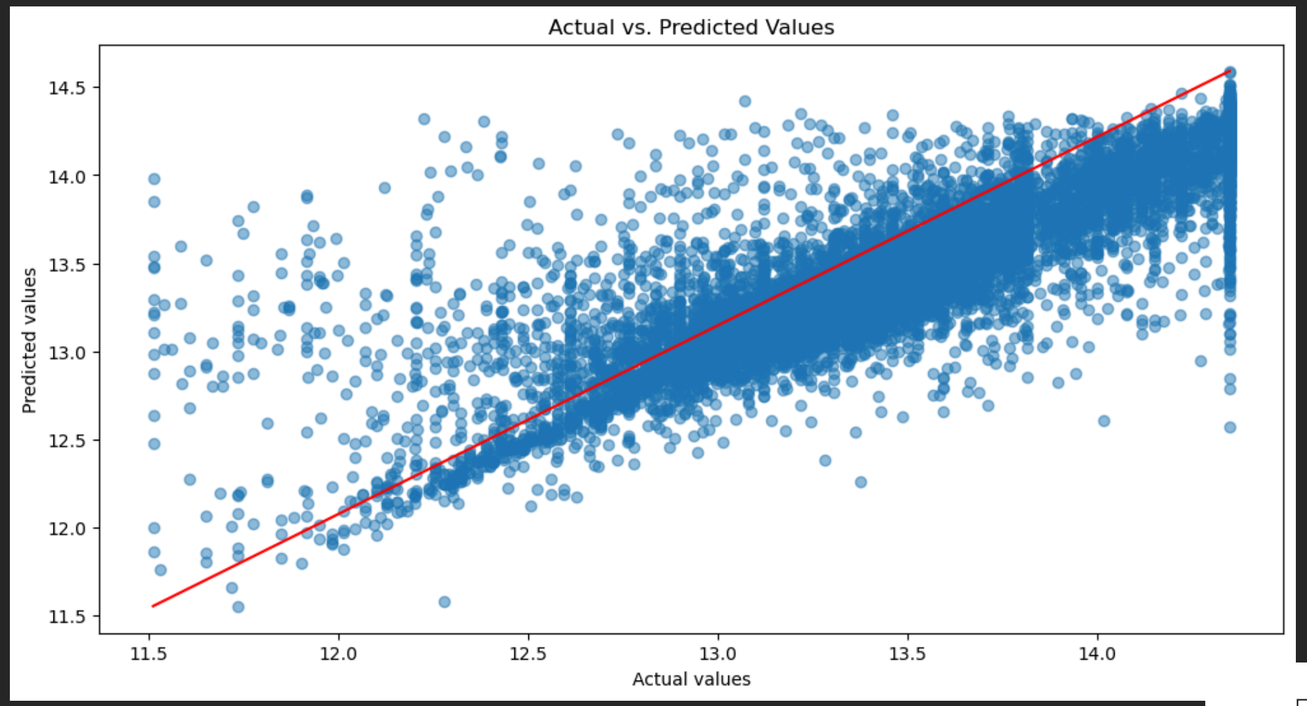
MODEL EVALUATION & PREDICTION

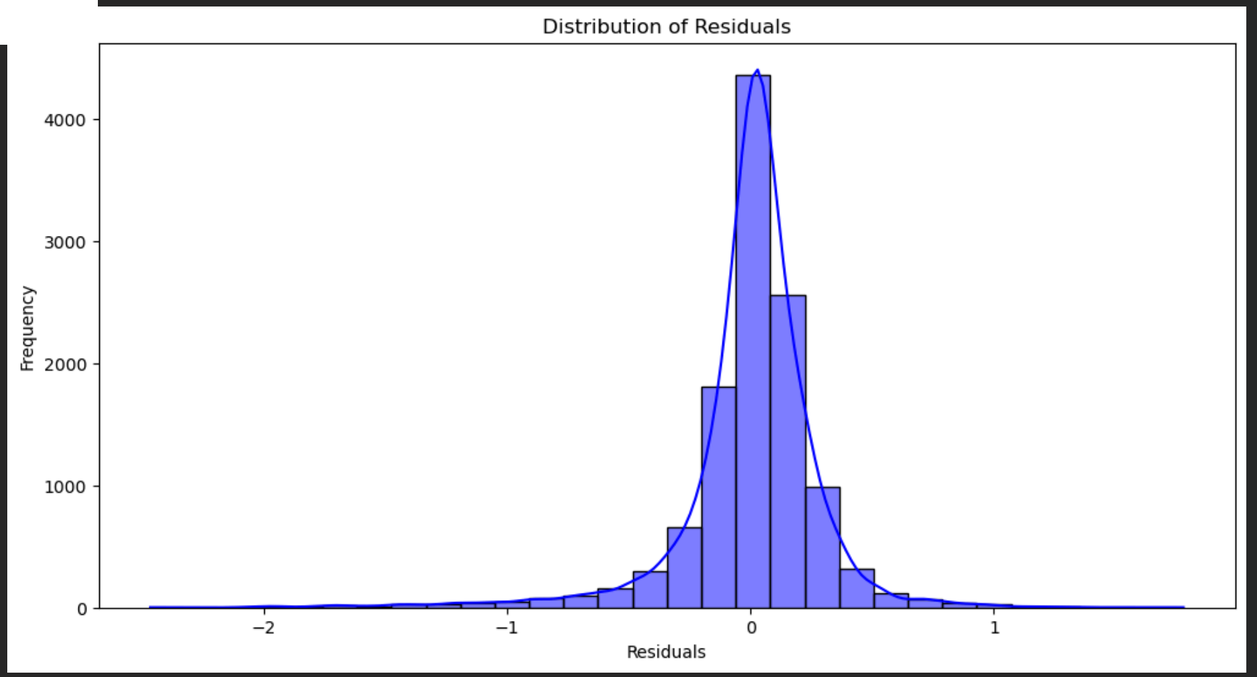
Reasoning for RMSE Metric for Evaluation

- **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.

Compariosn 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): W**hile 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
Predicted values
Actual values
14.5
14.0
13.5
13.0
12.5
12.0
11.5
11.5
12.0
12.5
13.0
13.5
14.0


Distribution of Residuals
Frequency
Residuals
4000
3000
2000
1000
0
−2
−1
0
1

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.