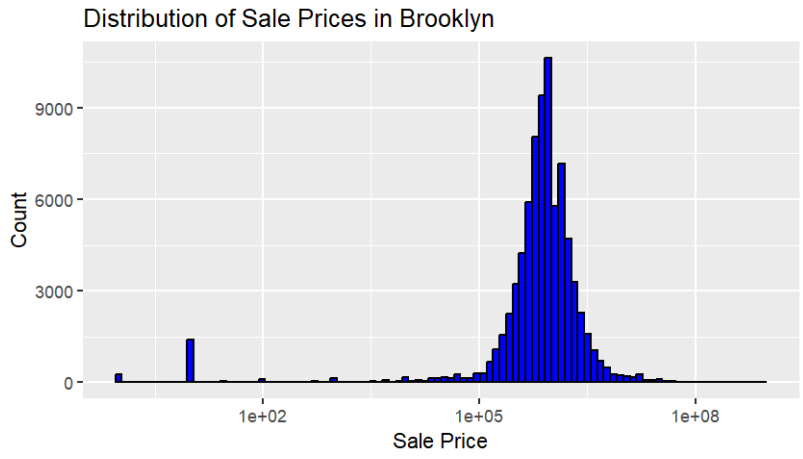
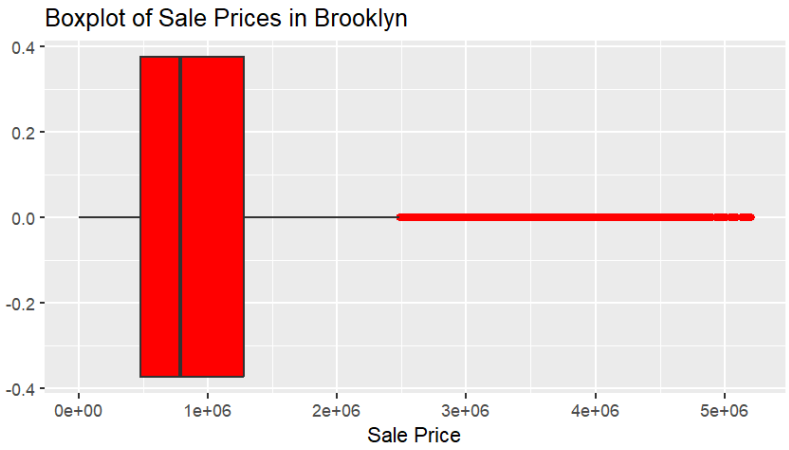
Inah Lee
Data Analytics
3/24/2025

NYC Real Estate Analysis and Neighborhood Classification

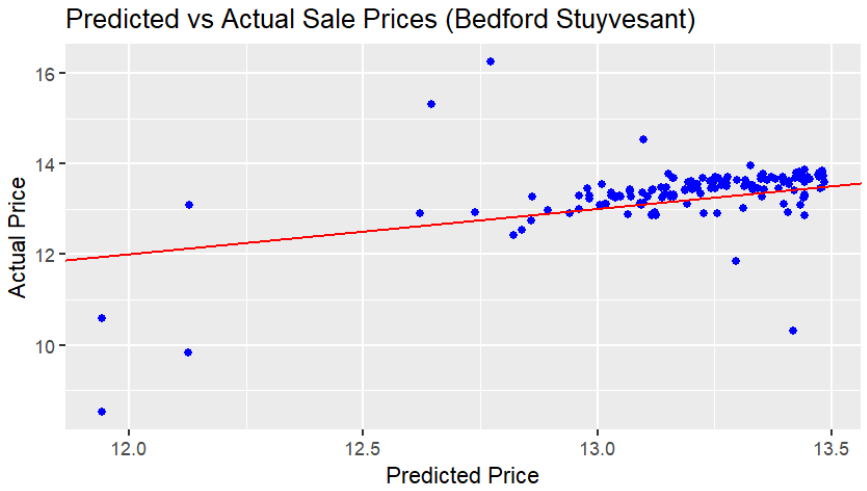
Problem 1

This study will investigate the relationships and patterns influencing sale prices using Brooklyn real estate transaction data. Particular attention will be given to the impact of variables such as gross square footage, building class, and neighborhood on price. The analysis will begin with exploratory data analysis to examine the distributions of key variables and identify outliers. Subsequently, regression models will be employed to estimate sale prices, while classification models will be used to predict neighborhoods based on quantitative features. Finally, the models' generalization performance will be evaluated by applying them to data from another borough, and differences in performance will be analyzed.





I conducted data analysis on the Brooklyn dataset. First, I removed rows with missing or zero sale prices. To examine the distribution of sale prices, I created a histogram with a logarithmic x-axis. The distribution was right skewed, with the majority of properties priced between $100,000 and $1,000,000. I also generated a boxplot using the IQR method to identify outliers. Most outliers were priced above $2.5 million, which could potentially influence the performance of regression models. These outliers will be addressed either removed or treated during the modeling stage.

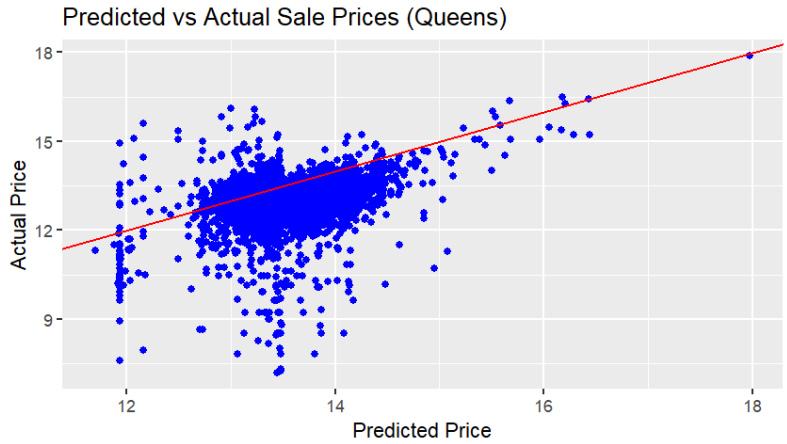


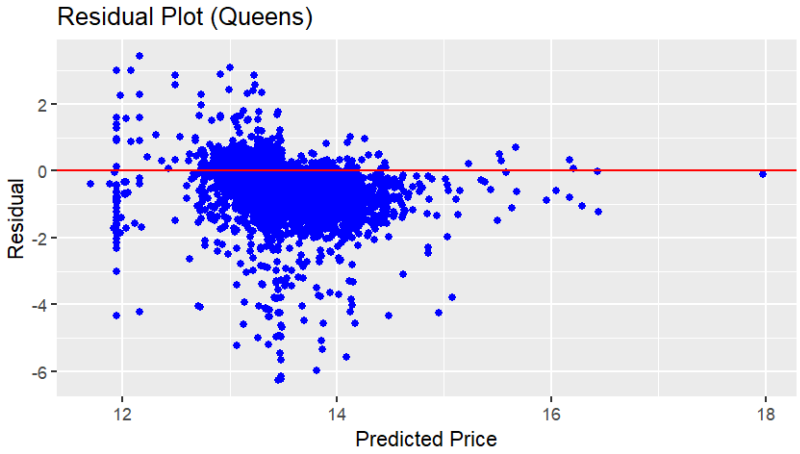
I performed regression analysis on property sale prices using multiple variables. During preprocessing, I removed entries with missing or zero values for sale prices and gross square

footage. Relevant columns were converted to numeric types, and both variables underwent log transformation to reduce skewness. Several models were tested, and a multiple linear regression model using log-transformed square footage and year built demonstrated the best performance. To assess generalization, I selected the Bedford-Stuyvesant neighborhood as a case study. The scatter plot of predicted versus actual values showed that most data points were close to the diagonal, although some outliers deviated, indicating that unmodeled factors may influence sale prices. Overall, the model provided a reasonable baseline but would benefit from incorporating additional predictors for improved accuracy.

For classification, I trained Naive Bayes, k-Nearest Neighbors (k-NN), and Random Forest models to predict the neighborhood using sale price and gross square footage as features. Prior to training, I removed entries with missing or zero values for these variables and applied log transformations to reduce skewness. The dataset was split into 70% for training and 30% for testing. All models showed relatively low accuracy, with Random Forest achieving the highest accuracy (0.2898253) and Naive Bayes the lowest (0.1437372). k-NN performed moderately well but was sensitive to class imbalance. Contingency tables and precision/recall metrics revealed that model performance varied by neighborhood, with certain classes being more difficult to distinguish. The low overall accuracy was attributed to the limited number of input features and the large number of neighborhood categories. I concluded that incorporating additional features could improve classification performance. Due to the length of the confusion matrices and detailed outputs, these results are provided in the appendix at the end of this document.

Problem 2





I applied a regression model trained on Brooklyn data to the Queens dataset, using log-transformed gross square footage and year built to predict sale prices. In the predicted versus actual plot, the points were more widely dispersed compared to the Brooklyn results. Most points were concentrated between 12.5 and 14.5 on the log scale, indicating that the model struggled to capture the price patterns in Queens. The residual plot showed a wider range of residuals at lower predicted prices, suggesting potential bias in the model. These results indicate that the model did not generalize well to the Queens data.

For classification, I attempted to evaluate a model trained on Brooklyn data using the Queens dataset. However, because there were no neighborhoods in common between the two datasets, it was not possible to compare predicted and actual labels, and all outputs were NaN. This

highlighted an important consideration for generalization testing: both the training and testing datasets must share the same label space. This evaluation underscored the importance of verifying dataset consistency before assessing model generalization.

Through this project, I examined multiple models to predict housing sale prices and classify neighborhoods using a New York City dataset. The regression model demonstrated good performance on Brooklyn data but performed poorly when applied to Queens data, likely due to differences in housing market characteristics between the two regions. For classification, overall accuracies were low, although Random Forest and Naive Bayes outperformed k-NN.

One major challenge in Section 2b was evaluating the Queens dataset using models trained on Brooklyn data. Because the neighborhood labels did not overlap between the two datasets, it was not possible to generate meaningful predictions, resulting in NaN accuracy values. This highlighted the importance of ensuring label consistency when designing and evaluating classification models.

Reference

Data_Analytics2025Fall_Group3_Evaluating_Classification_Clustering_Models.pdf
Data_Analytics2025Fall_Group2_Linear_Models_Random_Forest_0.pdf
Data_Analytics2025Spring_Group2_Generalization_Validation_Optimization.pdf
Data Analytics S25 - Classification & Clustering Evaluation | Powered by Box