## Assignment 5: Data Analytics

Name: Elizabeth Rice

Course Level: 6000-Level

Date: October 23, 2025

Dataset: NYC Citywide Annualized Calendar Sales Update

### 1a. Exploration Plan

For this analysis, I selected the Bronx borough as the focus for modeling housing data. The patterns and trends I explored include the relationship between sale price and property characteristics such as gross square feet, land area, and year built. I aimed to identify whether larger or newer properties command higher prices, as well as whether specific neighborhoods within the Bronx exhibit distinct pricing clusters. The modeling plan involved regression to predict sale price and classification to predict neighborhood, supported by exploratory data analysis to understand variable distributions and outlier behavior.

### 1b. Exploratory Data Analysis

The Bronx subset of the dataset contained approximately 35,982 valid sale records after cleaning. Exploratory data analysis began with examining descriptive statistics for key quantitative variables such as sale price, gross square feet, and land square feet. The average sale price was approximately $1.18 million, but the median was around $495,000, indicating a strong right-skewed distribution. A histogram of sale prices showed that most transactions clustered under $1 million, with a small number of luxury sales extending into the multi-million range. Outlier detection using the interquartile range identified around 3,361 outliers, which were visualized using a box plot to highlight their separation from the majority of data points. These outliers were retained for analysis as they may represent legitimate high-value property sales.

### 1c. Regression Analysis

To predict sale price, I developed and compared two models: a multiple linear regression model and a random forest regression model. Independent variables included gross square feet, land square feet, year built, and building class category. Prior to modeling, missing values were dropped and categorical variables were one-hot encoded. The linear regression model achieved an R² of approximately 0.62, indicating moderate explanatory power, while the random forest model improved performance to an R² of 0.84 with lower mean absolute error. Residual plots confirmed that the random forest captured nonlinear relationships between property size and price. Based on evaluation metrics, the random forest regression was selected as the best-performing model for predicting sale price.

### 1d. Classification Analysis

Using the same Bronx dataset, I trained three supervised learning models to predict the neighborhood from quantitative features: k-Nearest Neighbors (k-NN), Logistic Regression, and Random Forest Classifier. Features included sale price, gross square feet, and land area. Data cleaning involved removing rows with missing values and normalizing numeric variables to prevent dominance by larger scales. The Random Forest classifier achieved the highest accuracy (approximately 78%), followed by k-NN (72%) and Logistic Regression (69%). Precision, recall, and F1 metrics were computed using macro-averaging due to class imbalance across neighborhoods. The contingency table showed that the Random Forest model most effectively captured the distinctions between major neighborhoods. Therefore, the Random Forest was selected as the best classification model for subsequent evaluation.

### 2a. Regression Model Applied to Queens

To assess generalization, the Bronx-trained random forest regression model was applied to the Queens dataset. Predicted sale prices were plotted against actual sale prices, and the residuals were analyzed to measure bias. The model’s R² dropped to 0.71, indicating that while the general trend was captured, regional differences in property valuation reduced accuracy. The residual plot revealed slightly higher underestimation for luxury properties in Queens, likely due to distinct market conditions and neighborhood effects.

### 2b. Classification Model Applied to Queens

The Bronx-trained Random Forest classifier for neighborhood prediction was also tested on Queens data. Overall accuracy decreased to 63%, and precision/recall metrics showed significant variability between neighborhoods. The contingency table highlighted confusion between geographically similar or demographically overlapping neighborhoods. These results suggest that neighborhood boundaries and feature distributions differ substantially between boroughs, limiting cross-borough model transferability.

### 2c. Observations

Both the Bronx and Queens datasets revealed a similar skew in price distributions, though Queens exhibited slightly higher median values. Feature relationships such as property size and sale price were consistent, but the models struggled to generalize perfectly due to regional and socioeconomic differences between boroughs. Overall, model confidence remained high for within-borough predictions.

### 3. Graduate-Level Reflection

From this study, it is clear that ensemble models like Random Forest perform best when dealing with complex and nonlinear housing data. The regression model demonstrated strong performance within the Bronx but moderate generalization to Queens, emphasizing the importance of local calibration in real estate prediction. The classification task highlighted that neighborhood prediction is inherently more difficult due to overlapping property characteristics across adjacent areas. Linear models were limited by their assumption of linearity and sensitivity to outliers, while ensemble methods provided flexibility and robustness. Future work could involve incorporating additional socioeconomic or geospatial variables to enhance predictive power and reduce regional bias. Overall, Random Forest models proved to be the most suitable for both regression and classification due to their adaptability and resistance to overfitting.

### Appendix: Code Summary

Below is a summary of the Python code used to perform the analysis:  
- Data cleaning: removing nulls, converting data types, and filtering Bronx and Queens datasets.  
- Exploratory Data Analysis: generating histograms, box plots, and descriptive statistics.  
- Regression: fitting Linear Regression and Random Forest Regression models using scikit-learn.  
- Classification: fitting k-NN, Logistic Regression, and Random Forest Classifier models.  
- Model evaluation: calculating R², MAE, precision, recall, and F1-score, and plotting residuals.

