Tatiana Leonovich

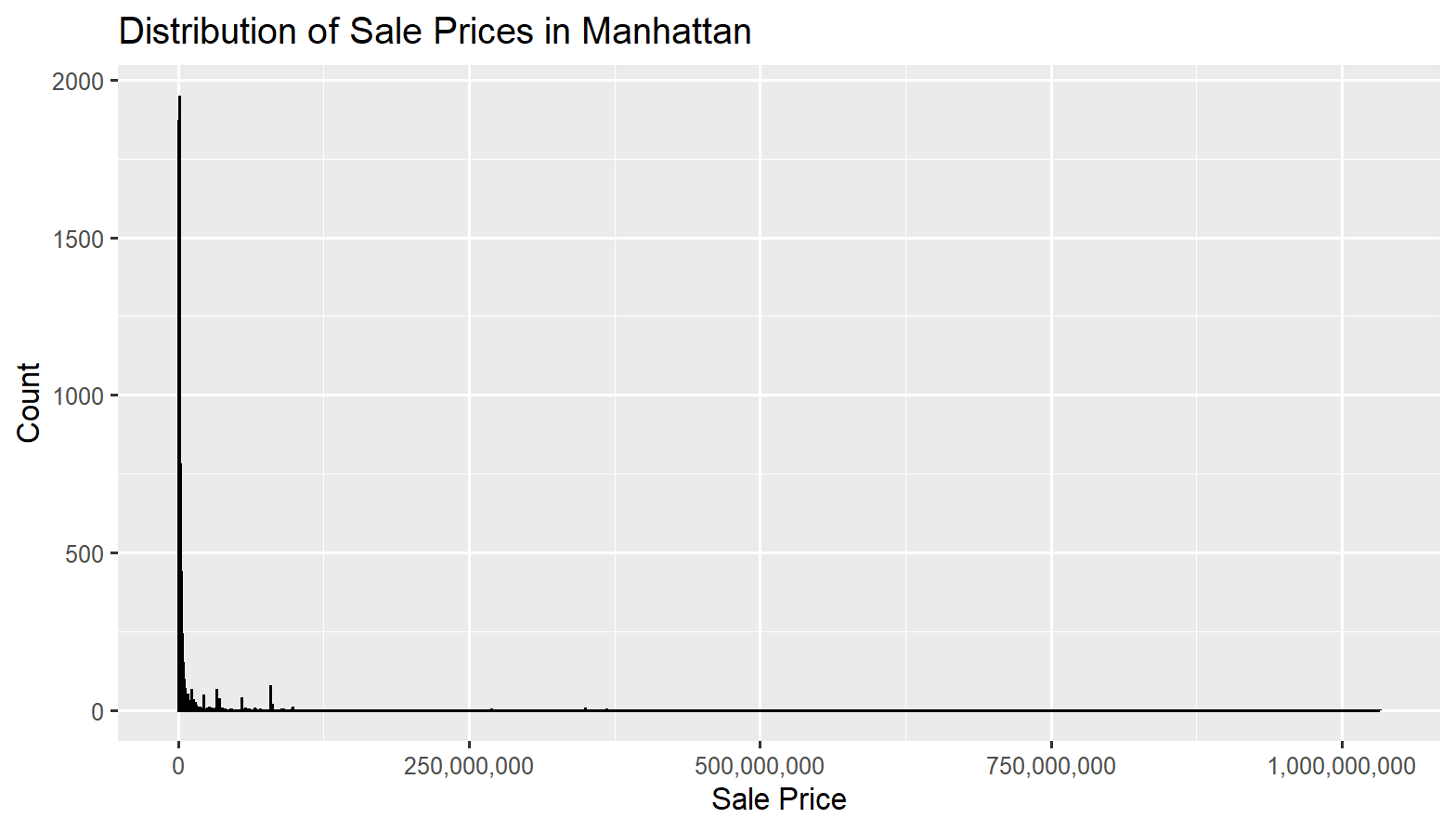
Dr. Ahmed Eleish

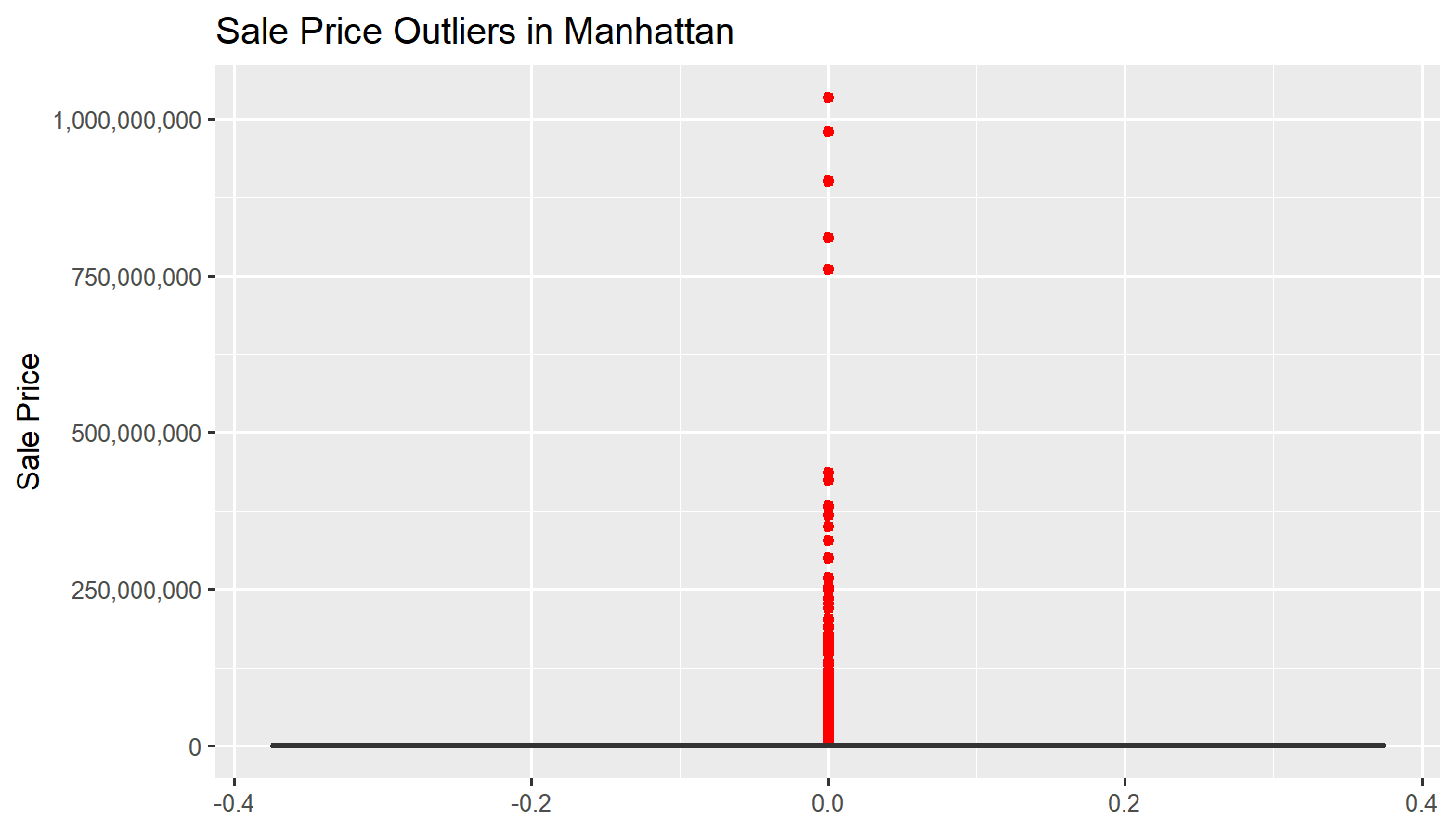
Data Analytics 4000

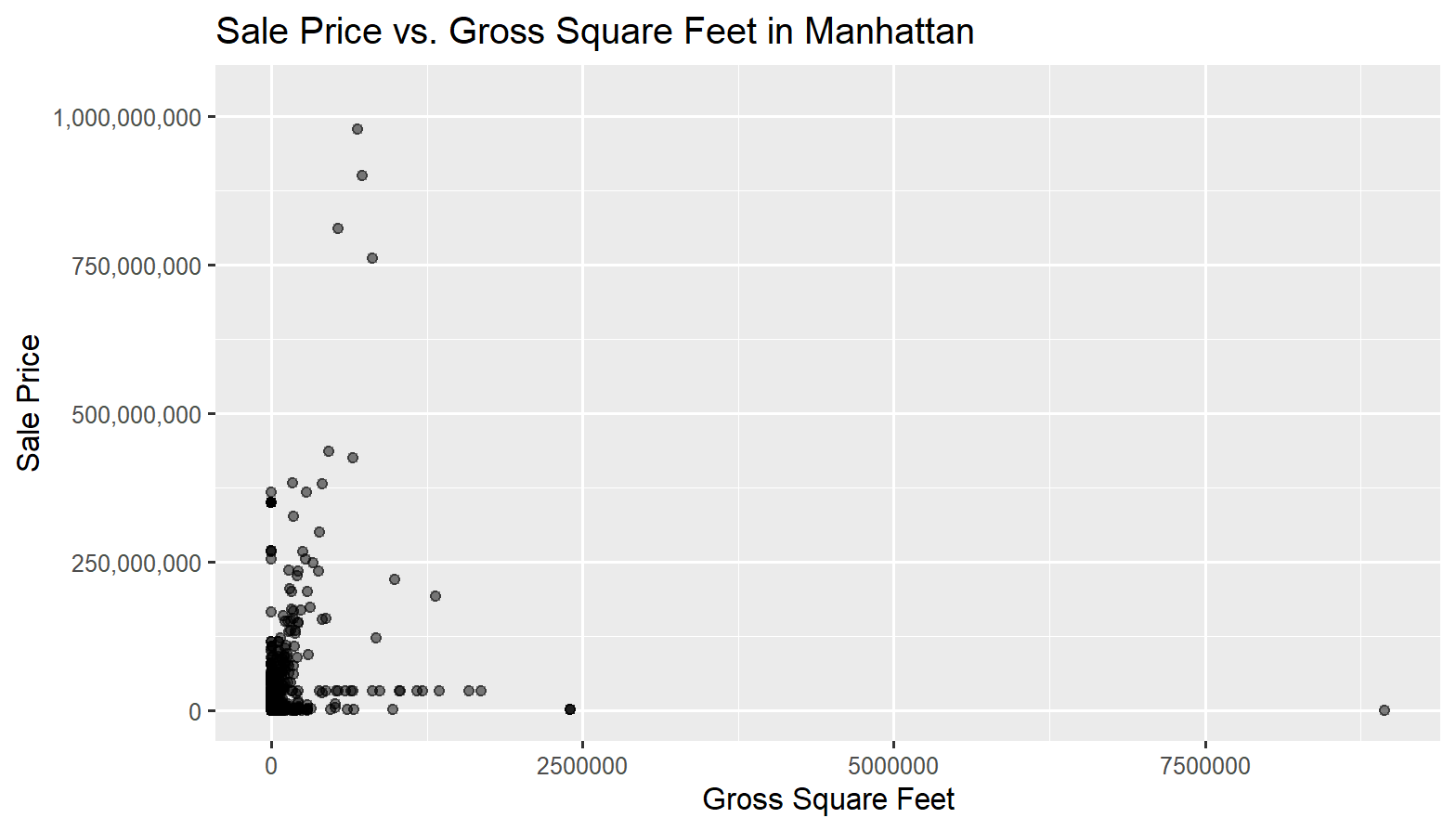
Assignment 5

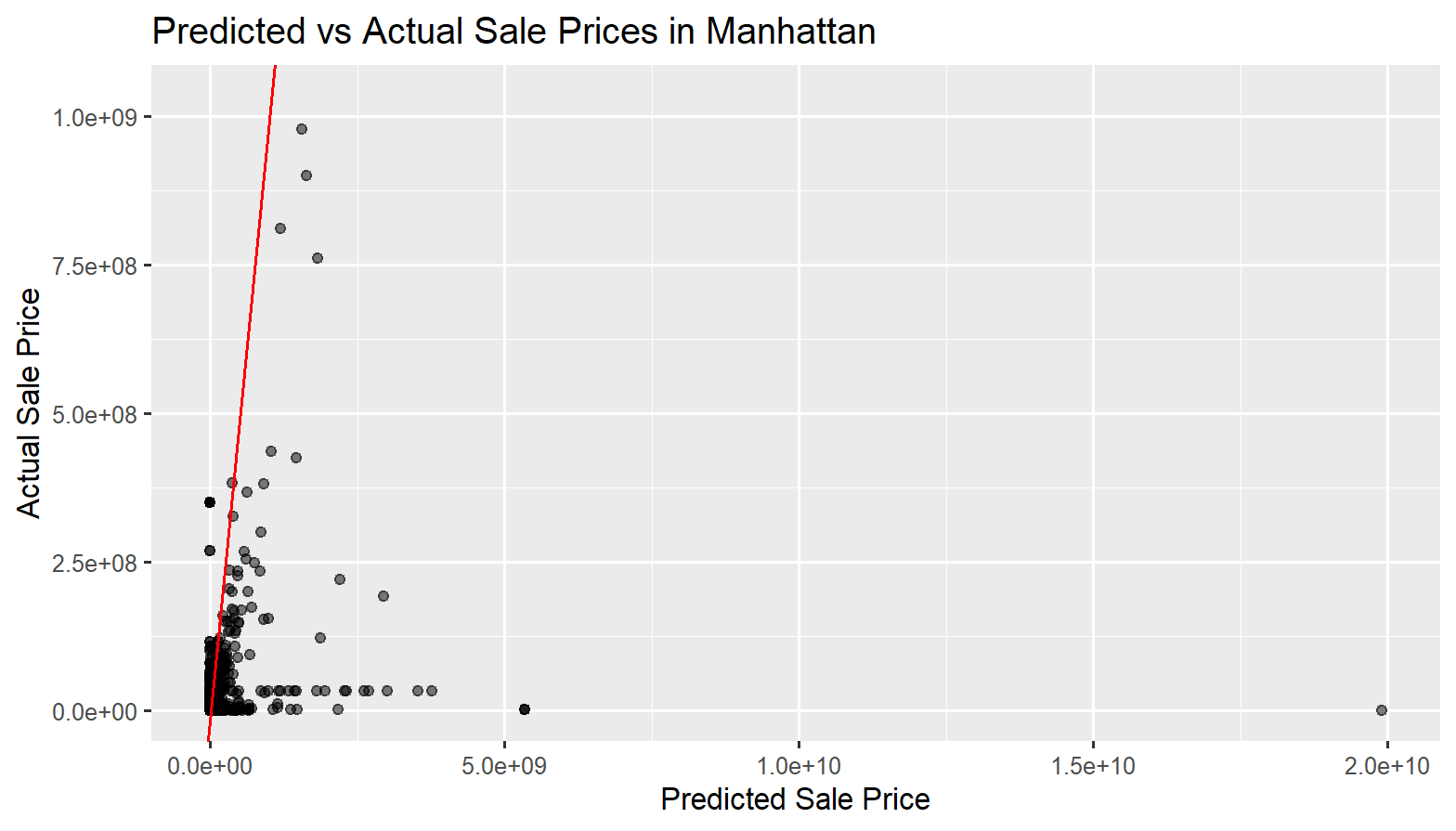
1a. When exploring patterns or trends in the data, I would focus on relationships between the sale price and various predictor variables such as gross square feet, year built, neighborhood, and building class category. For instance, I might look for trends where larger buildings or newer buildings have higher sale prices, or identify whether certain neighborhoods consistently have higher or lower sale prices. To model these trends, I plan to use regression analysis (e.g., linear regression) for continuous relationships, classification models (e.g., k-NN, Naïve Bayes) for categorical patterns like building class, and clustering to identify groups of properties that share similar characteristics. By visualizing the data using plots like scatter plots, box plots, and heatmaps, and evaluating the results through cross-validation, I can gain deeper insights into these patterns and test model assumptions.

1b. For the exploratory data analysis, I first examined the distribution of the Sale Price variable to understand its central tendency and spread. I used summary statistics (mean, median, and standard deviation) and visualized the distribution through a histogram to see if it followed a normal distribution or if it had any skewness. Additionally, I created a boxplot to visually identify the outliers in Sale Price by marking the values outside the interquartile range (IQR) using red dots. The boxplot revealed several extreme values that were much higher than the other data points, indicating potential outliers. To confirm the presence of outliers, I calculated the IQR and flagged values that were more than 1.5 times the IQR above the third quartile or below the first quartile. This helped me identify properties with unusually high sale prices that could distort the overall analysis. Further, I plotted residuals from regression models to assess the accuracy of predictions and detect outliers in the predicted sale prices, further confirming the influence of these extreme values on the model.









1c. For my analysis, I conducted a multivariate regression to predict the Sale Price of properties in Manhattan using several predictor variables, including gross\_square\_feet, land\_square\_feet, year\_built, and building\_class\_category. The final model included gross\_square\_feet and building\_class\_category as the key predictors, with building\_class\_category treated as a categorical variable.To test the model’s generalizability, I applied it to two subsets of the Manhattan dataset: one focused on office buildings and the other on residential units.

Office Model:

R-squared: 0.08656 (8.66%) – This indicates that the model explains only about 8.66% of the variance in the sale price.

MAE: 65,087,763 – On average, the model's predictions are off by about 65 million dollars.

RMSE: 124,078,907 – This indicates the average distance between predicted and actual sale prices is about 124 million dollars.

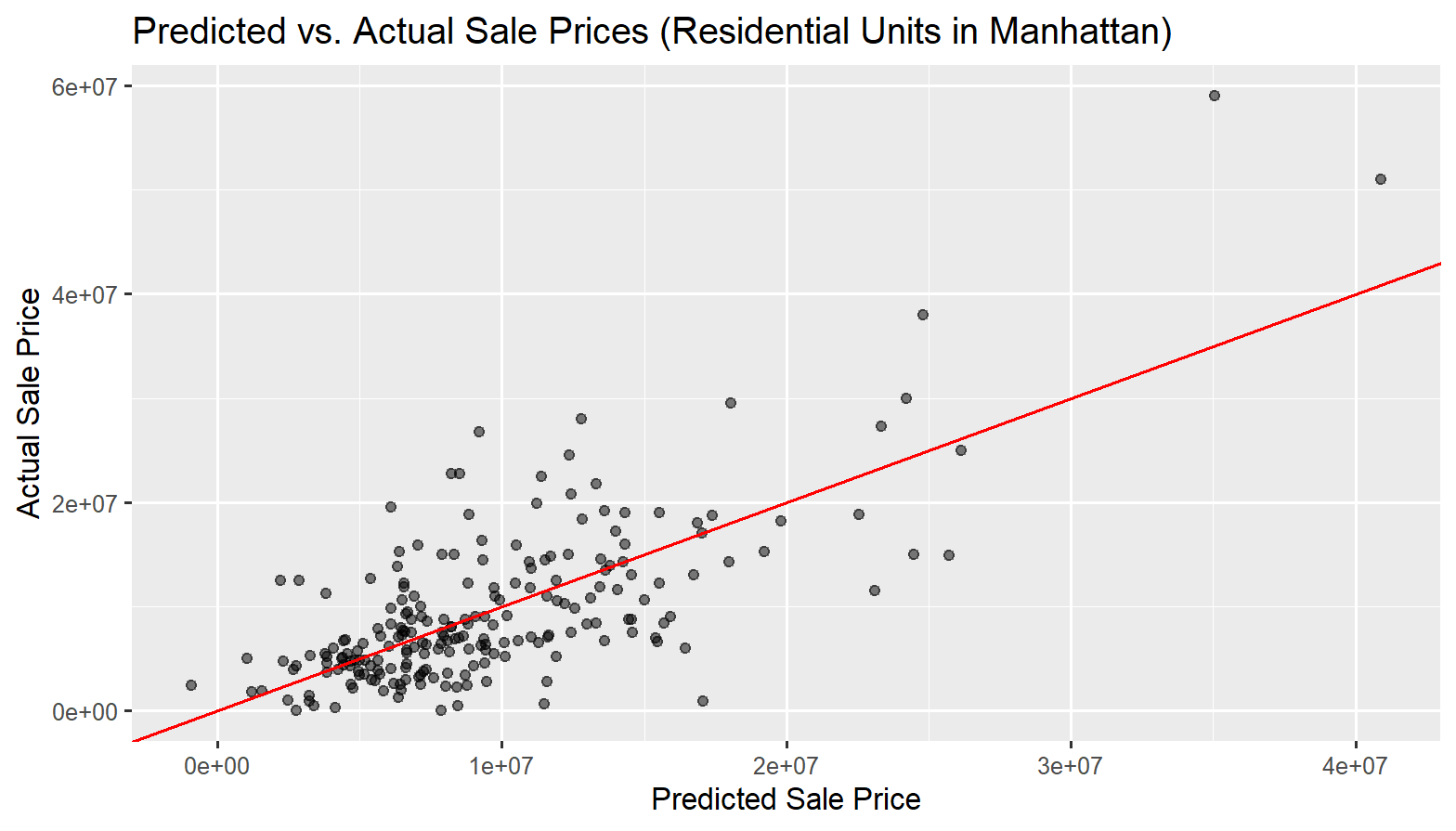
Residential Model:

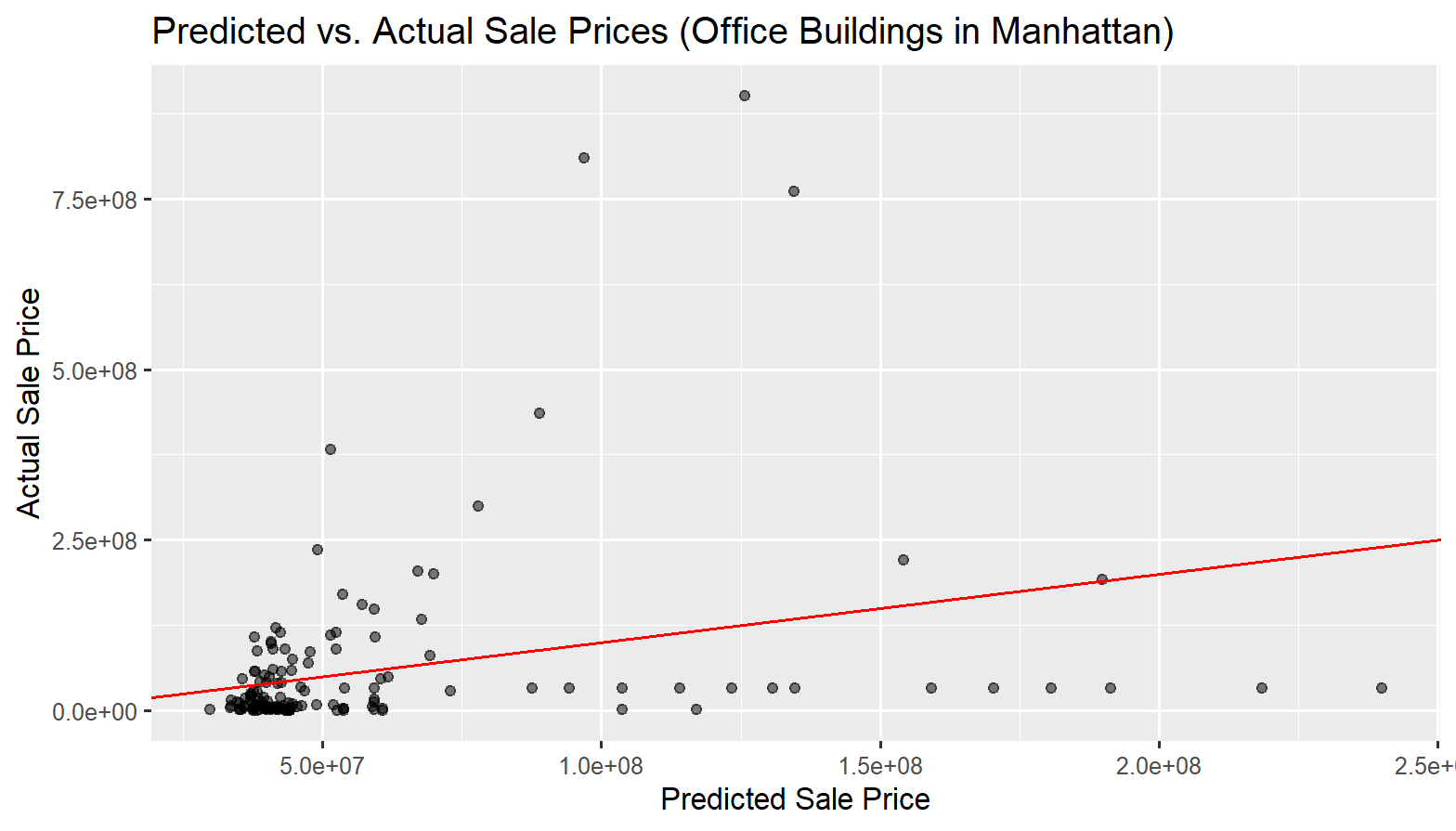
R-squared: 0.5187 (51.87%) – This indicates that the model explains 51.87% of the variance in the sale price, which is much higher than the office model.

MAE: 3,922,246 – On average, the model's predictions are off by about 3.92 million dollars.

RMSE: 5,415,901 – The average distance between predicted and actual sale prices is about 5.42 million dollars.

The Residential Model clearly outperforms the Office Model in all three metrics (R-squared, MAE, and RMSE). It explains more of the variance in sale price, and its predictions are much closer to the actual sale prices on average. The Office Model has a much lower R-squared and higher errors, indicating that it does not generalize well to predict sale prices in this dataset.

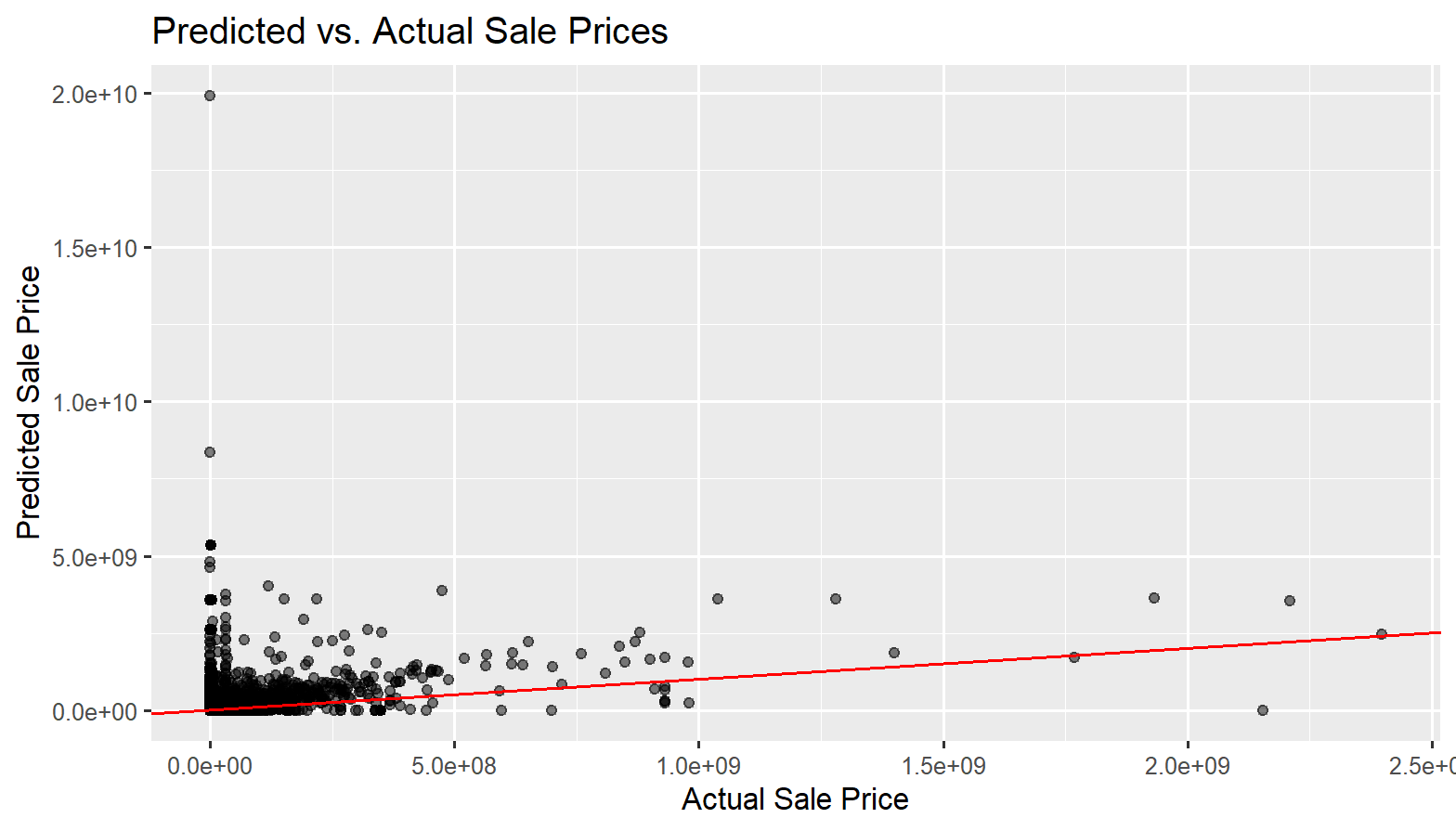


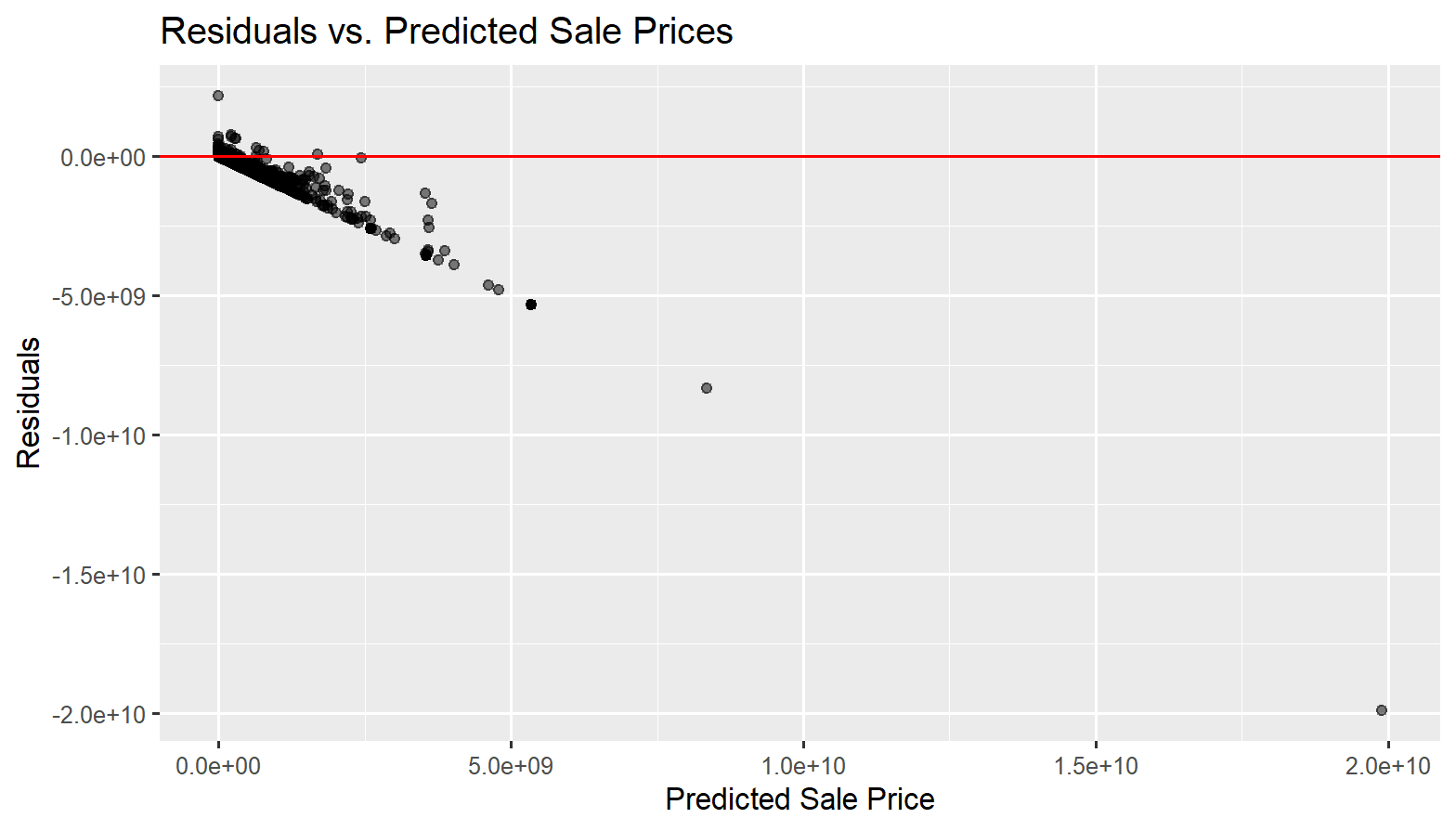


1d. For the classification problem, I focused on predicting the building class using two supervised learning models: Naïve Bayes and k-Nearest Neighbors (k-NN).

Data cleaning was performed to handle missing values and ensure that the features were complete and properly formatted. I removed rows with missing values for the target variable, building class, as well as rows with missing or irrelevant numerical data such as square footage or sale price. After cleaning the data, I applied both Naïve Bayes and k-NN models for classification.

The models were evaluated based on accuracy, confusion matrices, and performance metrics. The k-NN model showed an accuracy of 46.28%, while the Naïve Bayes model achieved a higher accuracy of 65.74%. This suggests that Naïve Bayes performed better in this case, likely due to its ability to handle the distribution of features more effectively compared to k-NN.

2.a 



Mean Absolute Error (MAE): 9,697,439

Root Mean Squared Error (RMSE): 95,070,642

These results indicate that while the model has some predictive power, there is considerable variation in the prediction errors, and the model could likely benefit from further refinement or feature engineering. The R-squared value of 0.5187 indicates that approximately 51.87% of the variance in the sale prices can be explained by the features used in the model (such as gross square feet, year built, etc.). This suggests that the model has some predictive power, but there is still a significant portion of the variance (about 48%) that is unexplained, meaning other factors not included in the model could be influencing sale prices. While R-squared provides a good general measure of fit, it is important to consider that an R-squared value in this range (around 0.5) often indicates that the model captures some of the trends but is not fully accurate. Further improvements, such as adding additional relevant features, could potentially increase the R-squared value and enhance the model's predictive capability.

2.b For the Naïve Bayes classification model, the accuracy was 46.28%. This performance indicates that the model was only moderately successful in predicting the building class category, as its accuracy was significantly lower than the no-information rate (NIR) of 57.1%. This suggests that while the model did better than random chance, it struggled to effectively distinguish between different building classes. In conclusion, the Naïve Bayes model provided a basic benchmark for classification, but its moderate accuracy suggests that more advanced models (e.g., Random Forest, SVM) or refined feature engineering could lead to better performance.