Dani DiTomasso
Professor Eleish
Data Analytics
25 November 2024

Data Analytics Assignment 5

1. Creating a derived dataset containing data points from only one of the five boroughs of NYC.

    a. After narrowing the dataset to show only one borough, users can compare and contrast many different data points to draw meaningful conclusions. For example, my first instinct would be to compare whether or not there is a correlation between the “Block” the building is on, and the “Tax Class As Of Final Roll”, which returns what tax classes (options of 1,2, 3, and 4) based on the use of the property. This can then be taken one step further and compare all of the buildings located on the same block, and see if they are all a part of the same tax class.

    b. For an exploratory analysis, a histogram displaying the data for the frequency of buildings that were built between the years of 1875 and 2025. It shows that most buildings were constructed between 1925 and 1975. The next graph that was part of my exploratory analysis was conducting a linear model for the Arverne neighborhood in Queens, NY. The linear model compared sales price to the total units, gross square feet, and tax class at time of sale. When determining the outliers for Sales Price and other variables, I first created scatterplots that compared the sales price variable to year built, gross square feet, and total units without the variables removed. I then repeated these visuals with the outliers removed, and noticed a significant change within the trend of the data, making the results more significant. For example, when comparing the Sales Price and Year Built scatterplots with and without the outliers, the graph including the outliers is almost unreadable, whereas once the outliers are removed the results are much more organized and accurate to the data.

    c. To run a linear model based on price, a random subset of 1000 rows of data was taken from the “queens_data” to make the data more manageable to be run in a sufficient time on the computer. Then, a linear model was run based on PRICE, with the equation of “SALE.PRICE ~ TOTAL.UNITS + TAX.CLASS.AT.TIME.OF.SALE + BLOCK + LOT + ZIP.CODE + YEAR.BUILT.” After running a summary on this linear model, the three variables that were shown to have the most statistical significance, or the variables that had a p-value under 0.05, were “YEAR.BUILT” (p-value 0.00553), “LOT” (p-value: 0.01911), and the most significant variable “TAX.CLASS.AT.TIME.OF.SALE” (p-value 2.83e-05). 2 subsets of data were then created, splitting the variable GROSS.SQUARE.FEET into lowSQFT (all values below 1500) and highSQFT (all values above 1500). Running a linear model on both of these subsets showed that for significantly more buildings with a 4.0 TAX.CLASS.AT.TIME.OF.SALE with GROSS.SQUARE.FEET below1500. For YEAR.BUILT, there is a concentration of buildings constructed from 1920-1970 and past the 2000’s in buildings less than 1500 sq feet, and only from 1920-1960 in buildings greater than 1500. Finally, LOT numbers close to 1000 can be seen in many buildings with less than 1500 sq feet, compared to buildings greater than 1500.

d. Looking at the "queens_data_subset," instead of testing what influenced the sales price of each building, the TAX.CLASS.AT.TIME.OF.ROLL was examined in both a SVM model with a linear kernel, and within an KNN analysis. Essentially, answering the question of whether or not the variables of TOTAL.UNITS, YEAR.BUILT, and SALE.PRICE have an effect on the tax class of the building. Upon conducting the SVM analysis, cleaning was required by first eliminating any columns that had only NA as point values (NTA.CODE, Census.Tract.2020, and EASE.MENT). Then, any remaining rows with an NA value present were then removed, leaving only a dataset with values for all table cells. A similar cleaning process was used for the KNN analysis, and both are converted into a matrix. These matrices are then passed through the "calc_metrics" function, which is a function independently created to calculate the Precision, F1, and Recall and return those values. Upon printing the results as shown in figure 12, the KNN doubles the SVM analysis in both Precision and Recall values, meaning that the KNN model identifies 73.2% of all actual positive cases, making it more accurate and significant than the SVM linear kernel predictions.

2. Analyzing all 5 boroughs within the NYC dataset.

    a. Using the linear models from 1c, the SALE.PRICE of buildings in all 5 boroughs is compared to TOTAL.UNITS, TAX.CLASS.AT.TIME.OF.SALE, BLOCK, LOT, ZIP.CODE, and YEAR.BUILT. After running the summary of this linear model, TAX.CLASS.AT.TIME.OF.SALE and LOT were found to have a p-value less than 0.05, which were the statistically significant variables determined previously for the queens borough data (though an interesting note that YEAR.BUILT was found to be statistically significant for Queens, but is not for all 5 boroughs). Based on the results (as seen in figures 13 and 14), it can be concluded that a higher tax class number correlates to a higher sales price. Taking into account the fact that the data being analyzed is a random subset of 1,000 data points out of the 606,000 found in the original dataset, it can be argued that this is not an accurate representation of all the data, since a subset had to be taken in order for the computer to run the analysis without computational errors or stalling.

    b. The same process from question 1d was repeated for the NYC data subset, and an SVM linear kernel model and a KNN analysis were run for the TAX.CLASS.AS.OF.FINAL.ROLL variable, based on YEAR.BUILT, TOTAL.UNITS, and SALE.PRICE. The results of the matrices for all 5 boroughs were extremely similar to just the Queens borough matrices; where the KNN doubles the SVM analysis in both Precision and Recall values. Given the fact that the closer a F1 score is to 1, the better the model's performance in identifying positives while minimizing false positives and false negatives, the KNN F1 value of 0.7038 indicates that this analysis generalizes to the whole dataset successfully, taking into account the small population of data analyzed from the entire NYC dataset.

c. To ensure a high confidence within the results, it became quickly apparent that the data had to be cleaned. This includes dropping columns that were populated with many NA values, eliminating rows if an NA value was present for specific variables, and removing outliers from the dataset. Given these actions, my personal confidence in the results is high – however, there should be hesitancy around the reduced data subset size, and if 1,000 data points can be considered an accurate representation of over 600,000 data points.

3. Conclusions drawn from this study about the model type and suitability / deficiencies.

Within this study, multiple analysis types were utilized that allowed for a direct comparison on effectiveness between them. The most prominent example is between the KNN and SVM training models, in which the Precision, Recall, and F1 results were returned for both. Due to the consistency between the results for the Queen's data (1 borough) and the entire nyc subset (5 boroughs) it is clear that the KNN model was more accurate, and gave more positive feedback. For example, the Recall calculation for the SVM linear kernel for all 5 boroughs was only 0.3525, or 35% - whereas the KNN model returned a value of 0.7197, or about 72%, which clearly yields a higher, and therefore more accurate result. The other frequently used model within this study was a linear model, and in my opinion, yielded less successful results than the KNN and SVM models. Though the linear model was extremely effective at determining what variables had a statistical significance, it did not relay whether or not those variables correlated, or had a direct impact, on the target variable being studied. For example, in question 2a, the variable for SALE.PRICE was run through a linear model with multiple other variables, and only LOT and TAX.CLASS.AT.TIME.OF.SALE were shown to have a p-vlaue less than 0.05. However, this does not mean that the lot number or the tax class of the building at the time of the sale had a direct impact on the sales price – rather it proves that there was correlation, but does not mean causation.

Figures

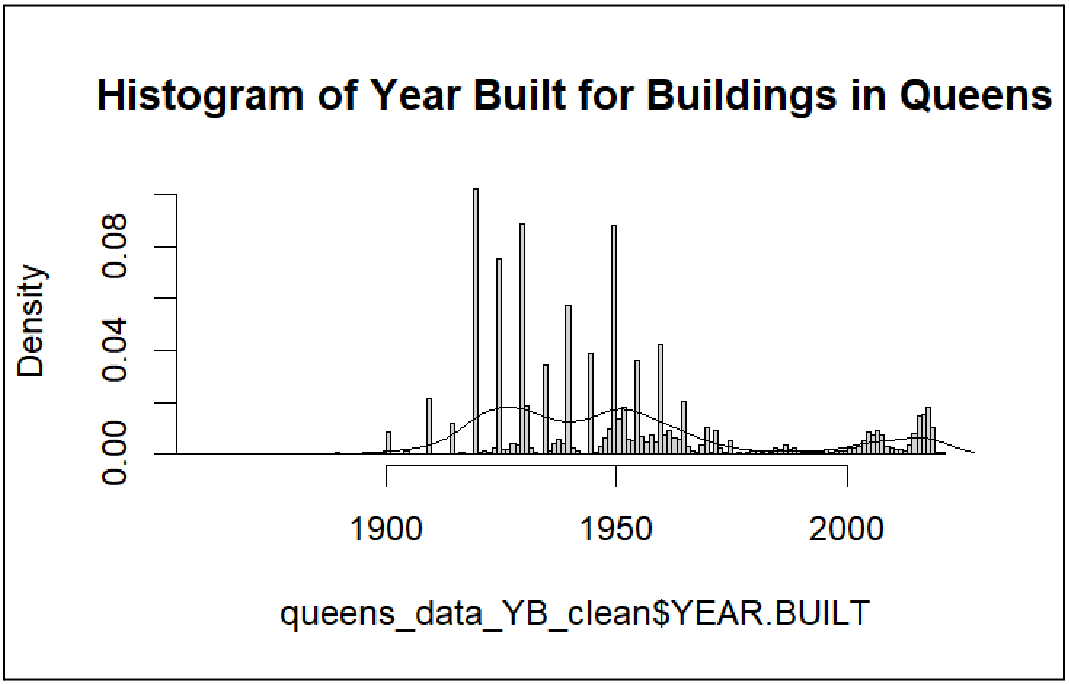


Figure 1: Histogram depicting building construction in Queens based on the year it was built for exploratory analysis. (1a)

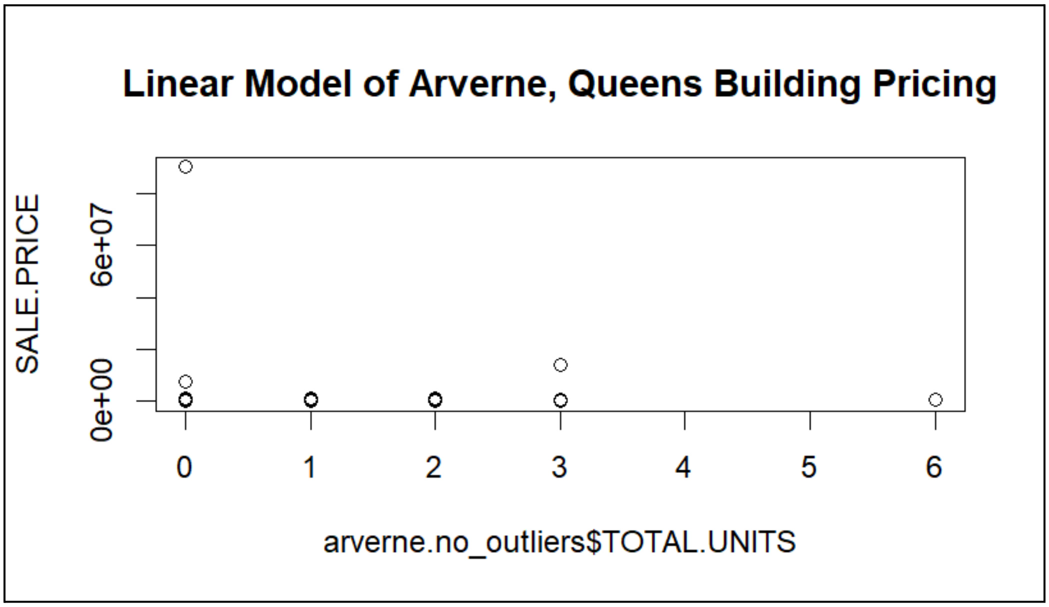


Figure 2: Linear model run on sales price of buildings based on the total units for exploratory analysis. (1a)

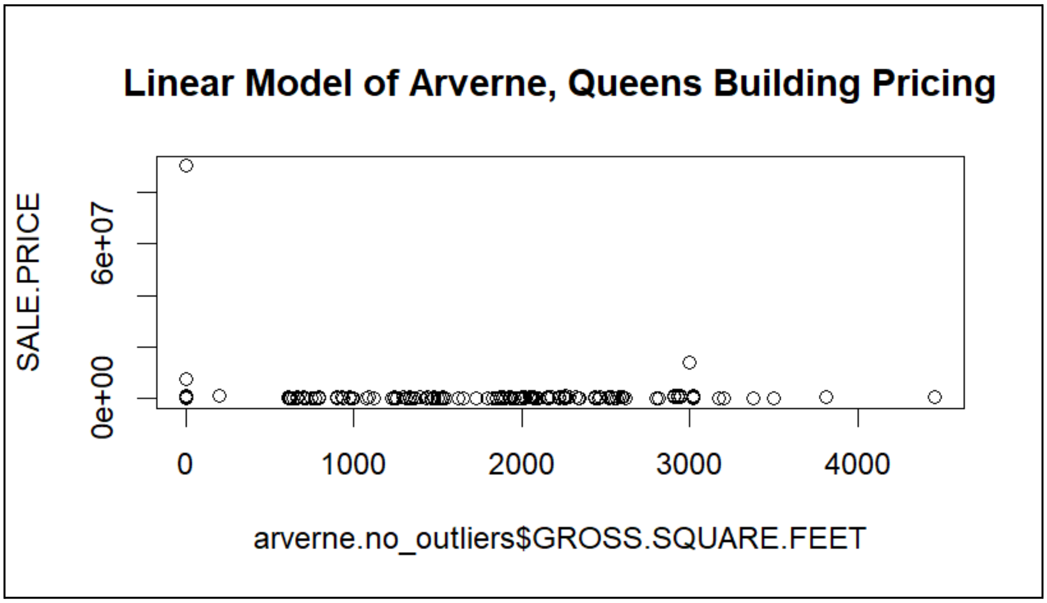


Figure 3: Linear model run on sales price of buildings based on gross square feet per building for exploratory analysis. (1a)

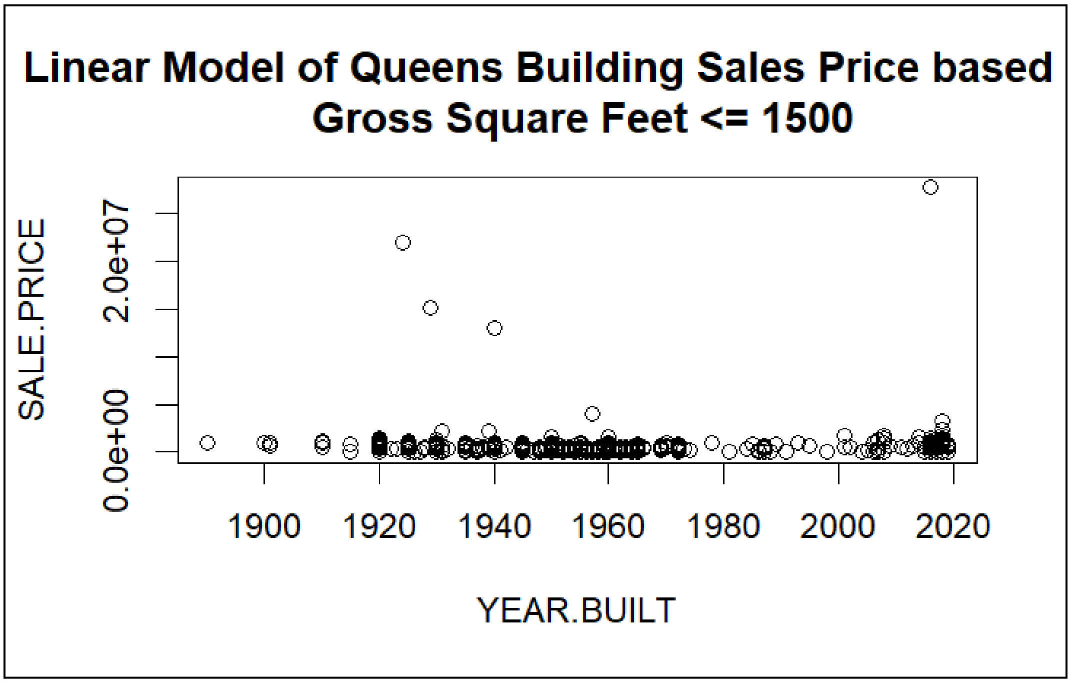


Figure 4: Comparing building sale price to the year it was built for all buildings less than or equal to 1,500 gross square feet. (1c)

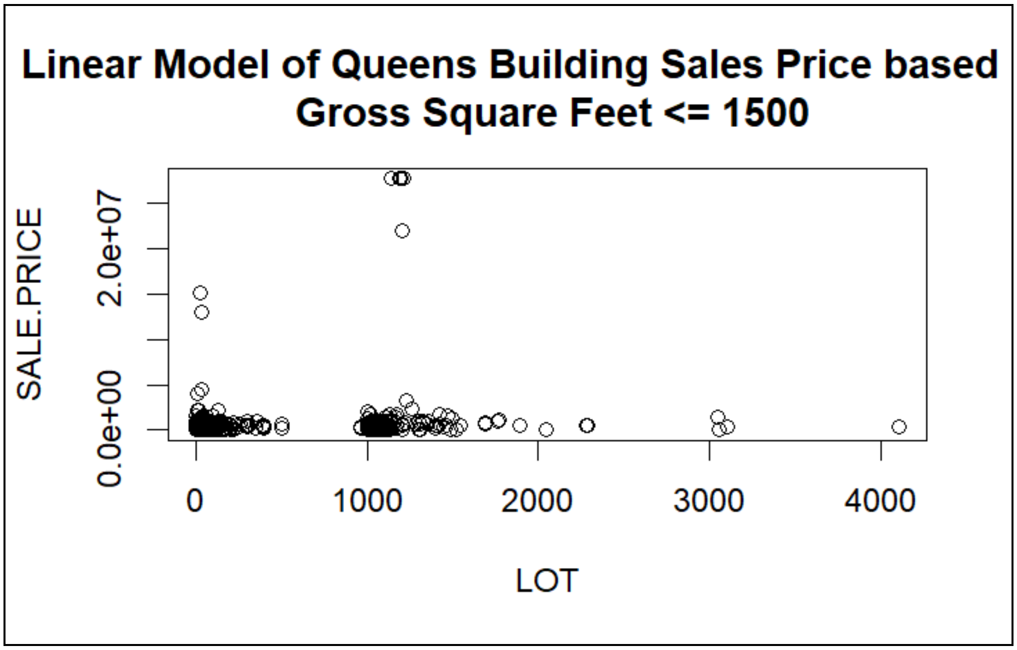


Figure 5: Comparing sale price to the lot number of the buildings for all spaces less than or equal to 1,500 gross square feet. (1c)

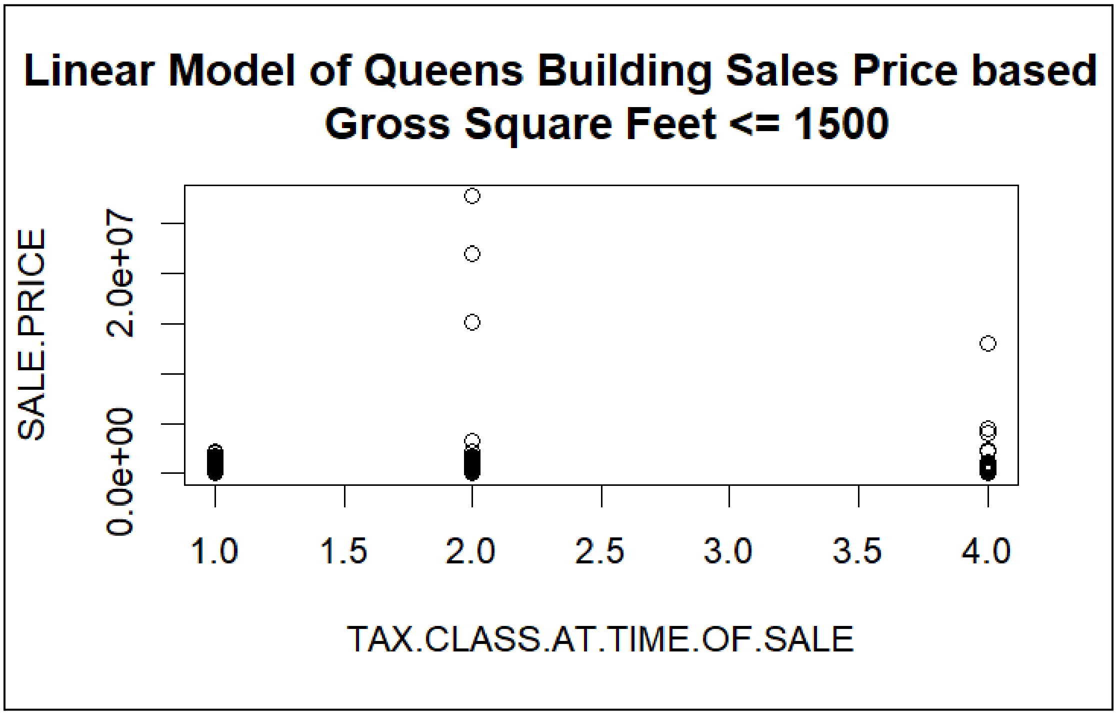


Figure 6: Comparing sale price to the tax class at the time of the building sale for all spaces less than or equal to 1,500 gross square feet. (1c)

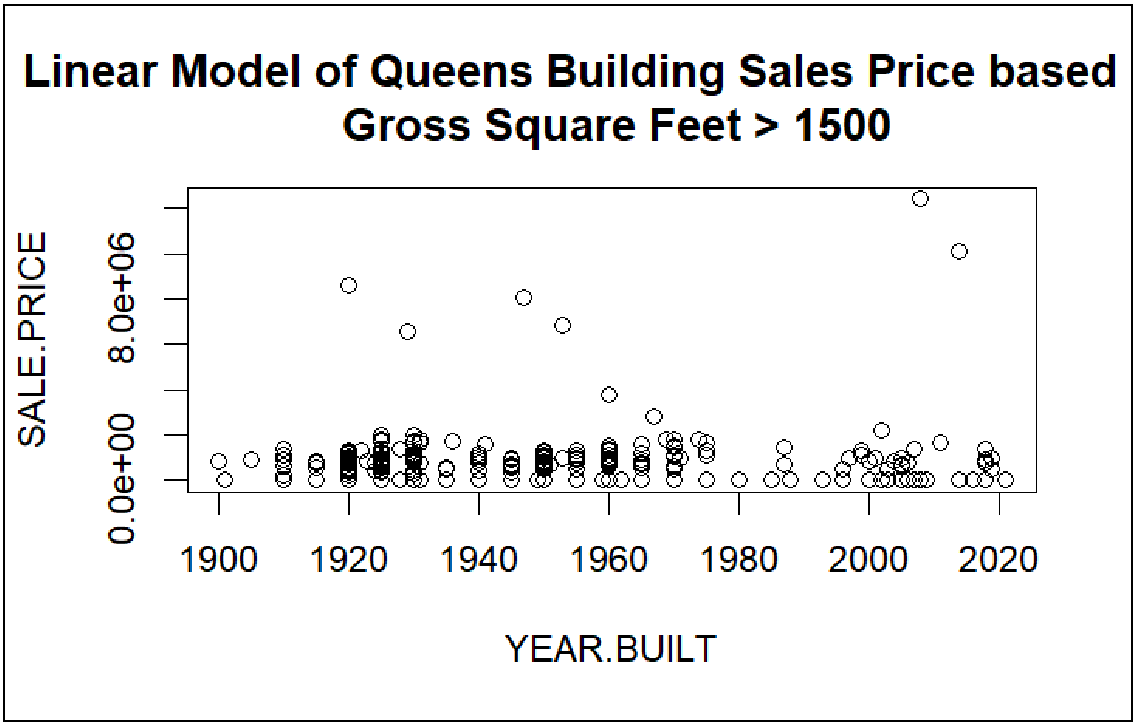


Figure 7: Comparing building sale price to the year it was built for all buildings greater than 1,500 gross square feet. (1c)

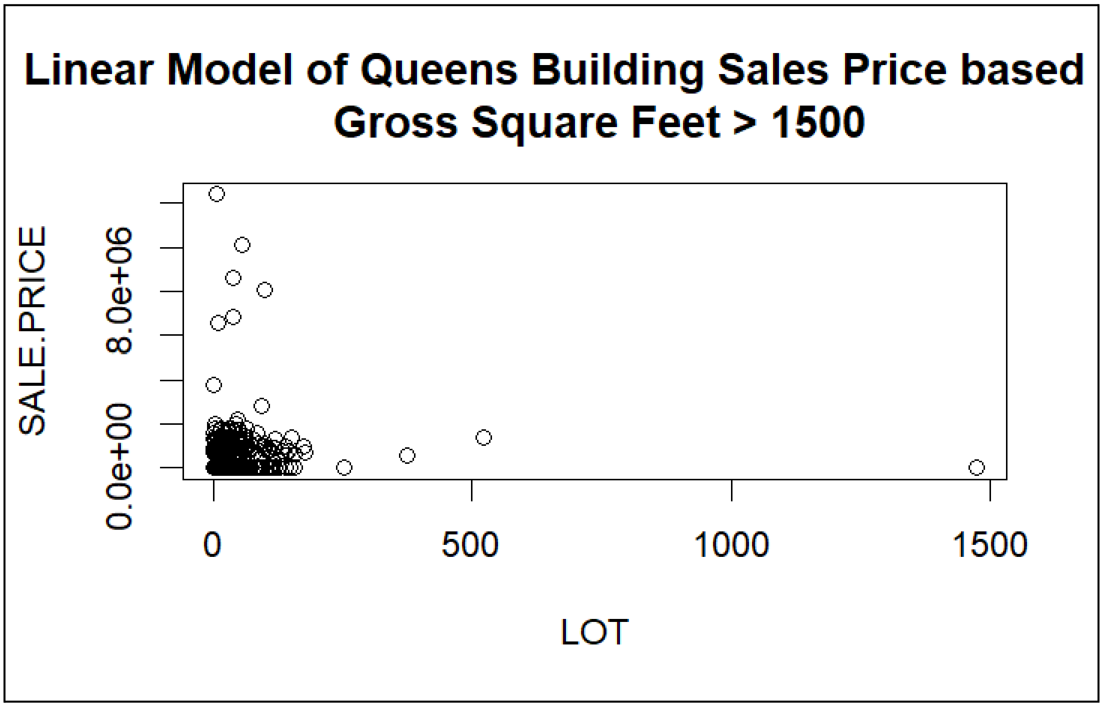


Figure 8: Comparing sale price to the lot number of the buildings for all spaces greater than 1,500 gross square feet. (1c)

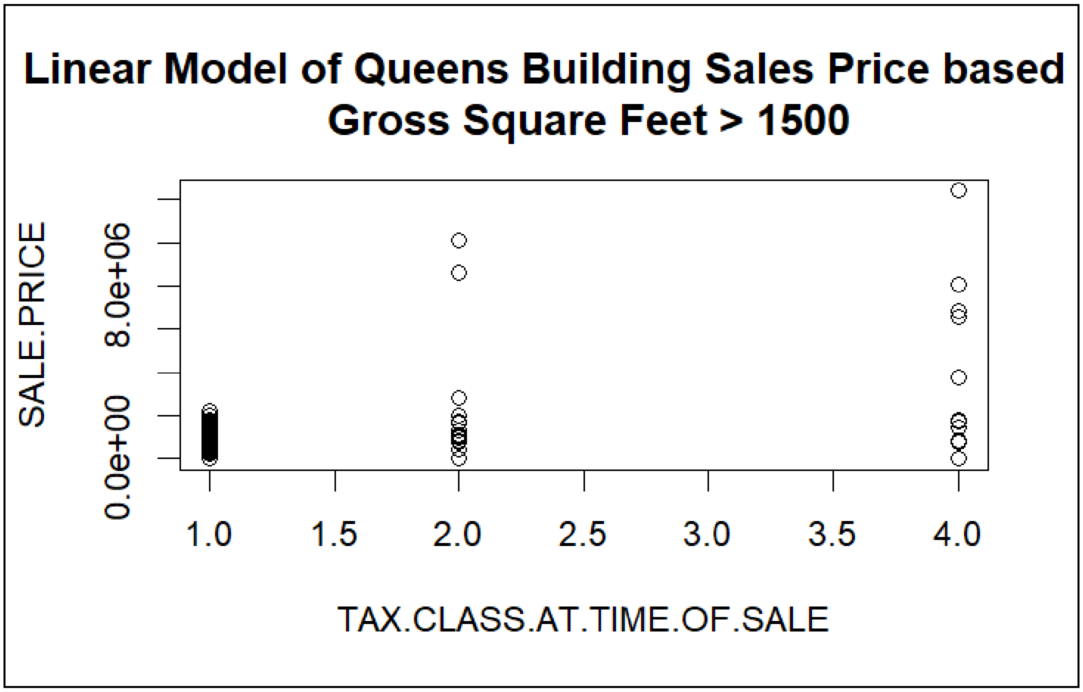


Figure 9: Comparing sale price to the tax class at the time of the building sale for all spaces greater than 1,500 gross square feet. (1c)

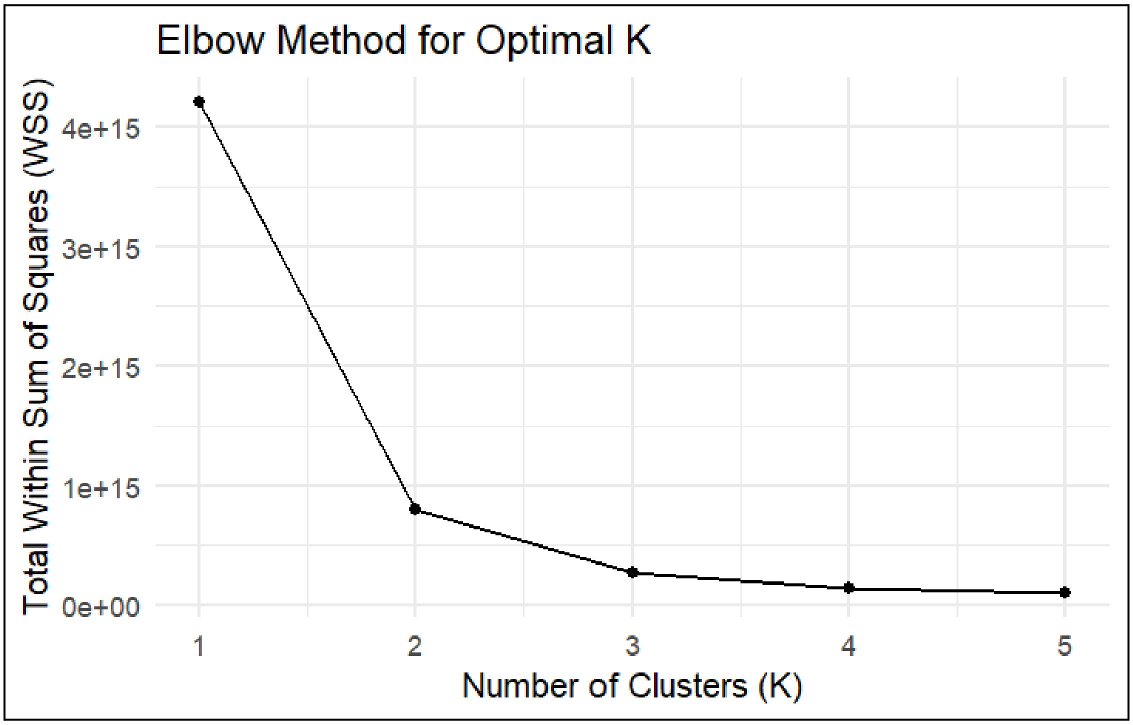
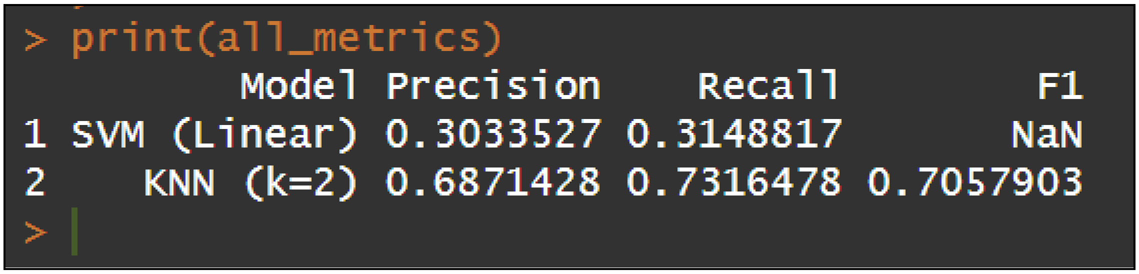


Figure 10: Conducting an Elbow Method plot to determine the optional K value for the KNN confusion matrix, k =2. (1d)



```
> print(all_metrics)
          Model Precision    Recall        F1
1 SVM (Linear) 0.3033527 0.3148817       NaN
2    KNN (k=2) 0.6871428 0.7316478 0.7057903
> 
```

Figure 12: Console image of metrics results, comparing Precision, Recall, and F1 values for a KNN model and a SVM model (linear kernel) for Queens data. (1d)

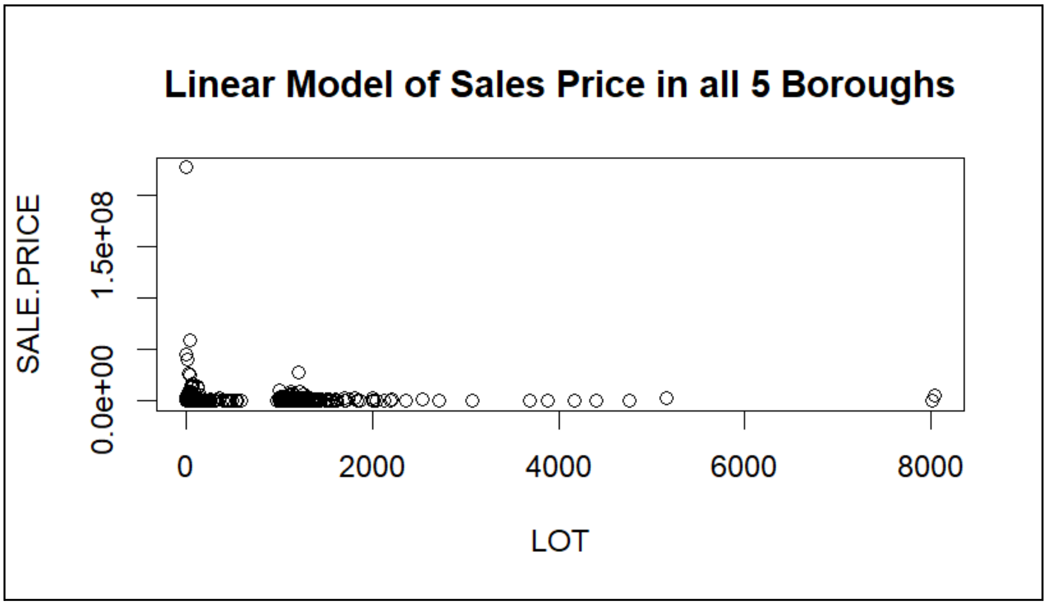


Figure 13: Linear model run on Sales Price per building in the 5 Boroughs based on Lot number, showcasing a large concentration of Lot Numbers around 0 and 1,000. (2a)

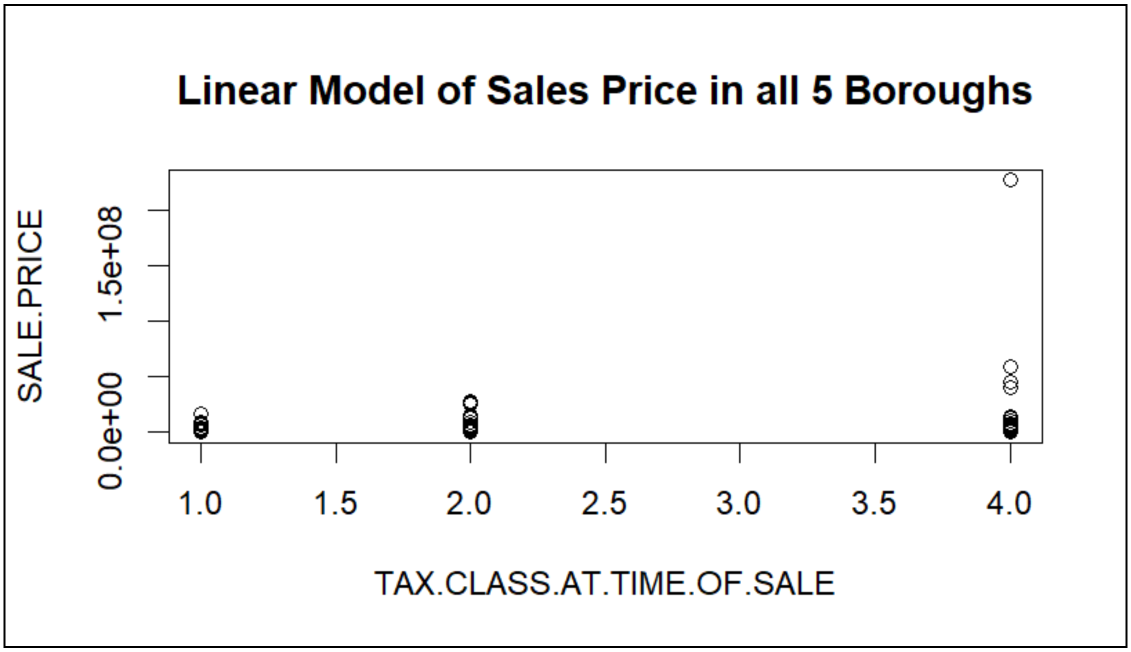


Figure 14: Linear Model run on Sales Price per building in the 5 Boroughs based on tax class at time of the building sale. (2a)

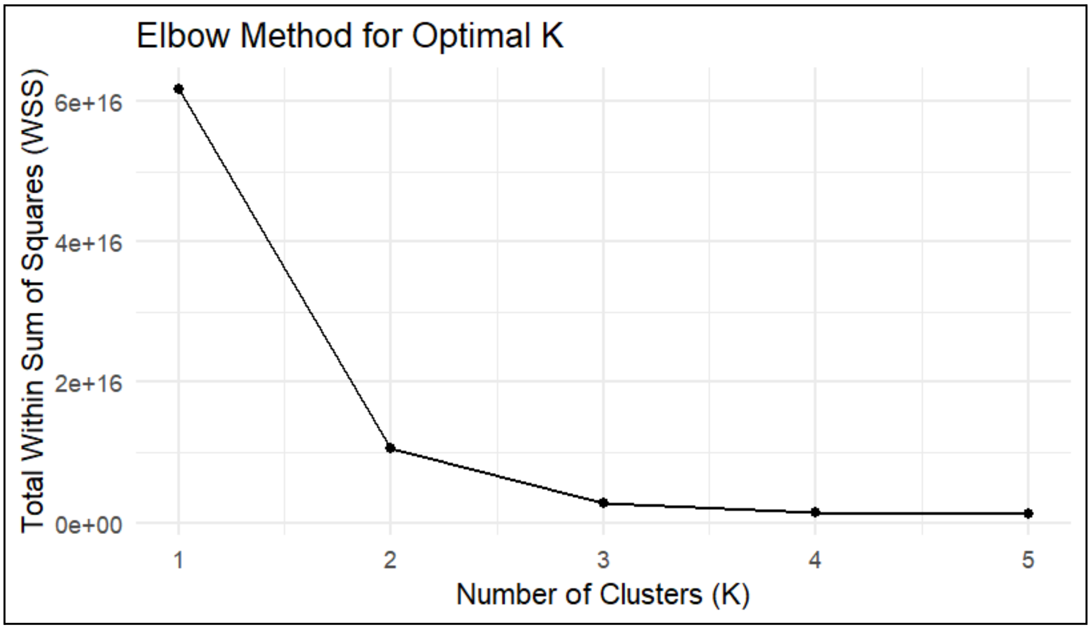
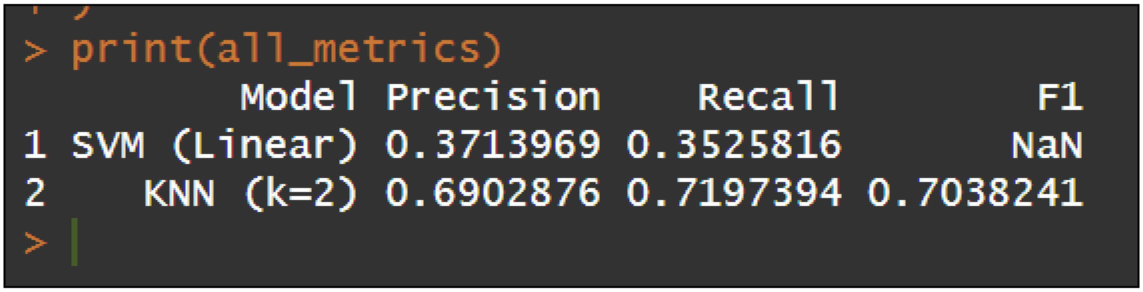


Figure 15: Conducting an Elbow Method plot to determine the optional K value for the KNN confusion matrix for NYC data, k =2. (2b)



```
> print(all_metrics)
         Model Precision    Recall        F1
1 SVM (Linear) 0.3713969 0.3525816       NaN
2    KNN (k=2) 0.6902876 0.7197394 0.7038241
> 
```

Figure 16: Console image of metrics results, comparing Precision, Recall, and F1 values for a KNN model and a SVM model (linear kernel) for NYC data. (2b)