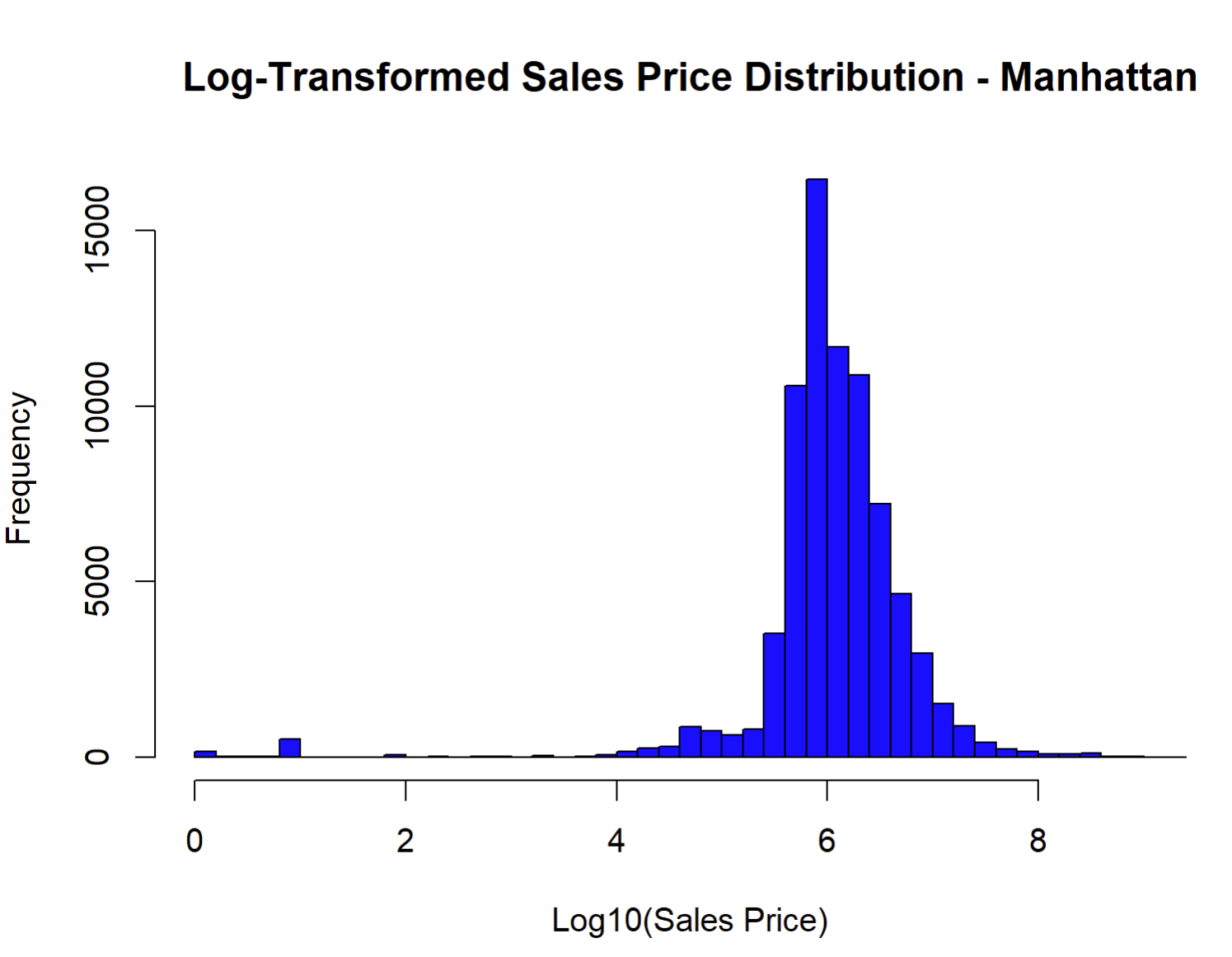
Notice of code used:

I used a lot of code from the class presentations. Much of the code I used came from my labs, which also came from the class presentations associated with that lab topic. But in addition, much of the code from the iris\_classifier example was used in this assignment.

a). Describe the type of patterns or trends you might look for and how you plan to explore and model them. Min. 3 sentences (0%) ;-

I think we are going to see a lot of distinction in housing prices around different neighborhoods. One way were are going to be able to see this is by clustering around different variables like cost of the houses and cost/ft^2. Additionally, even within that, I think we are going to see some huge outliers. Just thinking about Manhattan especially there can be some neighborhoods with average (for NYC) housing costs. But then, there could be one new building in the mix with a spike in the cost of all its units. And on top of that, there could be a penthouse in that building worth an obscene amount of cash that makes everything else in the neighborhood look like small fries. But in reality, that single unit is the outlier.

b). Perform exploratory data analysis (variable distributions, etc.) and describe what you did including plots and other descriptions. Identify the outlier values in the data for Sale Price and generate suitable plots to demonstrate the outliers relative to the other data points. Min. 5 sentences (2%)



I think this graph is clearly demonstrating to us where the majority of the houses lie, around the $1,000,000 mark. Additionally, considering that it is Manhattan, it would be a safe bet to remove anything below 3, as it is likely that property being sold for $1,000 or less is likely to be a clerical error. On the high end, I will not be removing anything as I do believe that some of the penthouses, or buildings themselves, can be sold for >100,000,000. So, cleaning steps are going to be to build a dataset without the low sales. And just so it doesn’t skew the data as much, I am also going to remove any sale that was worth more than $500,000,000 as that is getting into the realm of incredulity.

c). Conduct regression analysis on the 1 borough dataset to predict the Sale Price using other variables. After you find a well-performing model test it on a subset of the 1 borough dataset based on any criterion of your choice (e.g. neighborhood). You may have to try multiple models and drop variables with very low significance. Explain the results. Describe any cleaning you had to do and why. Min. 5 sentences (2%)

A computer screen shot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

I know the R^2 isn’t too great, but with lots of testing, this is about as good as I could do. I originally just started with the gross square feet and the land square feet. However, I found that only gave me about 2.8% R^2. I added these other variables one by one. I tried a few other variables, like the year the building was built, but did not find success in doing so. I tried to incorporate neighborhoods, but as they were a nominal statistic rather than an ordinal or interval, I couldn’t find an easy way to do so.

d). Using the same 1 borough dataset, train and evaluate 3 supervised learning models e.g., Naïve Bayes, k-NN, Random Forest to predict the neighborhood based on the quantitative variables (price, area). Evaluate the results using contingency tables & precision/recall metrics. Describe any cleaning you had to do and why. Min. 5 sentences

I had to do some additional cleaning, especially for the knn as that would not work with any nas or nans in any of the prevalent rows. Additionally, I found that there were quite a few duplicates that I had to remove as they were all tying during the knn process. I also found that the variable I was using was not properly sub sectioned to only Manhattan neighborhoods, so was getting to a lot of neighborhoods, outside of Manhattan, for which 0 predictions were being made. This really messed up the accuracy ratings of the functions. In terms of pure accuracy, the accuracy increased in the order I did them: Naïve Bayes, k-NN, Random Forest. Additionally, these took a long while to each run, I suspect because of the size of the dataset, but Random Forest did run the quickest on my machine.

a). Apply the best performing regression model(s) from 1.c to the new dataset to predict Sale Price based on the variables you chose. Plot the predictions and residuals. Explain how well (or not) the models generalize to the new dataset and speculate as to the reason.

The model does not work that well, but it didn’t really work that well on the original. The decline in performance must be relevant to the fact that what makes a building more valuable in Manhattan is not the same (though it is probably quite similar) to what makes a building valuable in Brooklyn. Thus, the model that I developed in question one is overfit to the parameters of the training set, Manhattan, and will need to be better generalized for the city if it will ever be effective in predicting the prices of the other boroughs.

b). Apply the classification model(s) from 1.d to predict the neighborhood in the new dataset. Evaluate the results (contingency tables & metrics). Explain how well (or not) the models generalize to the new dataset and speculate as to the reason. Min. 3-4 sentences

The models generalize the data set pretty well. Though there is a noticeable drop in performance. Additionally, the arrangement of the accuracy has changed among the models with the Brooklyn data group. The random forest and knn are performing nearly identical to one another while the naïve bayes are still struggling. I would say the decrease in performance, again, is because of the key variables in Brooklyn and Manhattan not aligning perfectly with one another. However, they performed not too much worse on this dataset for having been optimized for the Manhattan dataset.

c). Discuss any observations you had about the datasets/ variables, other data in the dataset and/or your confidence in the result. Min 1-2 sentences (0%) ;-)

This dataset was not very well maintained. I am used to more academic datasets, which are purposefully cultivated to be very complete and leave very few rows and columns blank, but this NYC housing dataset is truly a monster in those regards. Not only the number of rows left blank, but the number of rows that were straight copies of another row, from what I would have to assume resulted from a double entry, is kind of insane. My confidence in the results is mixed. I spent a long time messing around with variable combinations for the linear regression, but despite that, I never was able to achieve any sort of substantial R^2 value. I don’t know if it is just that there is not a pattern underlying the sell prices, or if there are housing bubbles/crashes that I need to account for using the selling year better, but I really struggled to get higher R^2 values. Though, it was reassuring to see the miniscule p value I was getting. Overall, it was a cool, but sometimes very frustrating, data set to work with.