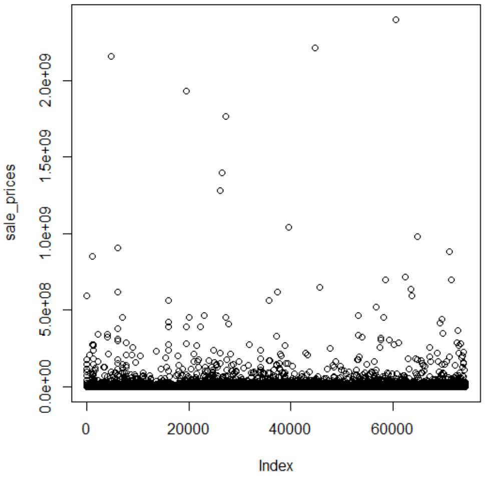
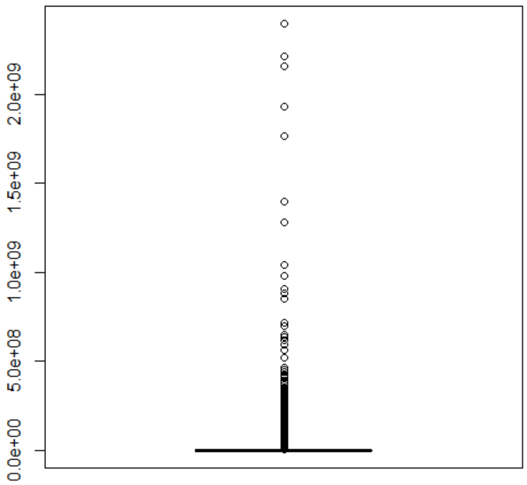
Ryan Kaplan

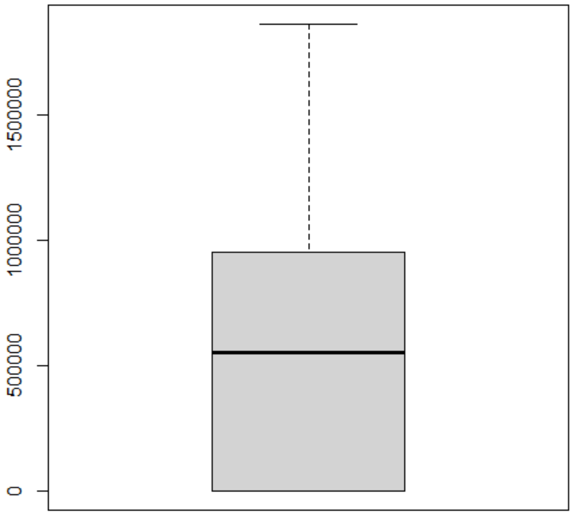
1a) In terms of a pattern, what jumps out to me is looking at how the neighborhood affects the price. I chose the Manhattan dataset, and would certainly expect the average sales price in Harlem to differ greatly from, say, the Upper East Side. Similarly, I would expect building type to matter quite a lot as well, it makes sense that prices for apartments and offices would be different. For exploratory data analysis, I used summary and plots to look at the distribution of the sale price, the variable we care most about. The raw plot is here:



It's tricky to tell as much as we'd like to just from this graph, since the vast majority of our data points are at the bottom of our graph. The average sale price was 3,018,000, with a first and third quartiles of 83,240 and 1,860,000, respectively. The overall range was from 0 to 2,398,000,000. Again, with raw data its a little hard to tell, as shown by the raw box plot:



However if we cut off examples above the third quartile, we can see a much more reasonable distribution:



1b) There are some clear outliers for sale price, the properties that sold far more than the vast majority, and the properties that sold for nothing. There are a decent number of 0 values, and I would argue that since we're talking about the sale of property, a price of 0 should be considered an outlier, even though there are 16,073 of them. These must be examples of people inheriting or being gifted them, as property should be quite expensive, certainly not something to be given away or abandoned for no reason. Since there are certainly cases of people basically giving property away, with only a small charge for legal purposes, we can remove all examples of a sales price of less than a hundred dollars. We have a massive IQR here of 1,776,760 dollars, which gives combined with our Q1 and Q3 of 83,240 and 1,860,000 gives us a lower outlier gate of -2,581,900 and an upper gate of 4,525,140. While I would still argue that <$100 values are still an outlier, as explained previously, we do have values past the upper gate, 7,312 of them in fact. Removing all outliers, the zeros and upper gated values, leaves us with 50,291 data points, or 67% of our original Manhattan data set.

1c) For the vast majority of data points, we're missing either gross square feet or land square feet. While I did sampling on all non-outliers available, this does mean that without selectively picking samples, I was unlikely to get one with both gross and square feet listed. My first sample had a gross value of 0, and a land value of 789, and the model predicted a sale price of 2,207,293. Considering the actual price was 2,150,000 this is a pretty close estimate. The second sample had a gross value of 428 and a land value of 0, and the model predicted a sale price of 1,147,500. This is compared to the actual price of 725,000 which is off, but within the same quartile at least. The third sample had a gross value of 638 and a land value of 0, and the model predicted 1,074,897. The actual value was 1,033,523 so this is another solid prediction.

1d) Cleaning was absolutely necessary. As described previously, there are many outlier properties, particularly in the example of absurdly low sale prices. It makes sense that many factors can come together to fetch an astounding sale price for a property, but much more difficult to believe that a piece of real estate in Manhattan is genuinely thought to be worth $0. Additionally, the null values had to be swapped for scores of 0 in order for numeric methods to work. I chose to do a KNN model, so I was also unable to include non-numeric categories. I included all other variables in the model. In the absence of a compelling alternative, and since my computer was able to handle it, I chose a k-value of 187, i.e. the square root of my dataset.

2a) I used the KNN model to predict sale price, as a nice throughline to the rest of the assignment. I didn't want to make pre-judgements about the usefulness of the features, which is why I included as many as possible. At first I thought some of them were useless, such as apartment number, but as I thought about it that too seemed useful. My initial thought was that it didn't seem relevant, but apartment number often serves as an indication of what floor the apartment is on, and that certainly could affect how expensive it is.

2b) I cut my data into a train and test set, with a 70/30 split, and the results seem good. I of course shuffled the data beforehand, and got an average of 60.8% accuracy in 5 trials. This accuracy refers to predicting the normalized sales price to the requisite degree, so I was open to the possibility of quite low accuracy, but the model seems to have turned out quite well. I should note that this is having removed the many apartments with a sale price of 0, so we don't have that as a source of either easy accuracy points, or a source of confusion.

2c) I feel good about the results. I wish that the gross and land square foot data was more complete, but I suppose that makes the accuracy more impressive so perhaps I shouldn't complain. One thing to note with the KNN model is that it does require a good number of diverse variables to function. You have to make sure that there aren’t ties, and this doesn’t just mean making sure k is even. If k = 5, but 2 neighbors point to one class, 2 point to another, and 1 points to a third class, you still have a tie. This is best avoided by including as many variables as possible, and ensuring k is large, at least as large as you can computationally afford.