Part 1a

I started by choosing to work with Queens since it had the most data. I also removed all rows where the price was 0 because since we’re mostly going to be working with the price I only wanted sales where it was an actual number. Looking at the boxplot for it, it’s not very helpful:

Chart, box and whisker chart

Description automatically generated

I would remove the outliers and try again but that’s part 1b, so I’ll wait. I also thought it would be a good idea to see if there’s a pattern around the dates, so I converted the date given to the number of days since the first sale using as.POSIXct and some simple math. The plots for that follow.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Diagram

Description automatically generatedChart, line chart

Description automatically generated

We can see that it’s relatively flat, however there was a dip around 100 days in. I also checked out some plots on the price, but it’s hard to tell without removing the outliers:

Graphical user interface, chart

Description automatically generatedChart

Description automatically generated

We can also look at the frequency of houses sold in the top 5 neighborhoods:

Chart, bar chart

Description automatically generated

Flushing-north greatly takes the cake on this one.

Part 1b

I found the outliers by finding those outside of the upper inner fence (aka. Greater than 1.5 x IQR). The minimum outlier was $1,715,905. I plotted outliers up to 3 million here, but know that they go up to $369,300,000 and there are 1633 of them.

Chart, histogram

Description automatically generated

Part 1c

The predicted values from the regression are here:

Text

Description automatically generated

The fake data is at the top, taken from 7 values that are more or less within the IQR of the respective fields and 3 outliers. The three samples s1, s2, and s3 are all randomly sampled from the sales in queens and of length 2000. I initially did larger sample sizes but it took too long to do the regression, and this way we might see some variability. In this case, the data was pretty even along the first 7. However, and this was to be expected, the outliers vary wildly. Our biggest “house” has a $170 million evaluation from s1, a $41 million evaluation from s2, and a $290 million evaluation from s3. Since the data was only sampled 2000 times and the overall size of the queen sales are more than 34 thousand rows, this makes sense as the amount of outlier data is probably small.

Part 1d

For this part I chose KNN. This is because KNN is a learning model that is for a classification / regression problem, whereas K-means is for clustering, which isn’t relevant to our goals. I debated using decision trees, however they are very non-robust and aren’t as accurate as other models, namely KNN. While doing this, I had to do a fair amount of data cleaning. I removed all columns from the dataset that weren’t numeric. After doing this I realized that the model performs better if I remove all columns that aren’t related to the size of the house, so I only worked with those two land-based columns from now on to predict the price. There was also a lot of data where one of the data points was 0. This doesn’t make sense, as a house wouldn’t be sold for 0 and a house cannot have 0 land. It was confusing the model quite a lot, so I removed them. Finally, I tested removing outliers which greatly improved the model’s performance (actually it’s probably more accurate to say it improved the model’s scoring, not performance). In part 2b I’ll show a few cases of how these changes improved the data.

Part 2a

For this part I’ll only show results from the fully optimized dataset that the model worked on (i.e. everything removed as said in part 1d). Again, I’m using LAND.SQUARE.FEET and GROSS.SQUARE.FEET to predict the SALE.PRICE. All of these graphs have had their x limit and y limit set so the majority of the data is most visible.

First, let’s compare the test\_prices (actual prices) with the predicted prices from the model. Chart, scatter chart

Description automatically generated

Honestly it doesn’t look *amazing*. If the model was perfect, this would be a linear line as shown here:

Chart, line chart

Description automatically generated

So clearly it gets a lot of predictions wrong. Let’s compare this with the regression model that we made earlier.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

On the left is the KNN model, on the right is the regression model. You can see the KNN model is more dense and has distinct lines, which makes sense as KNN works with factors. However from this alone it’s really hard to tell how accurate the model is. So to provide more insight, let’s model each variable vs. price, starting with the land square feet vs. actual price so we have the necessary context.

Timeline, scatter chart

Description automatically generated

This is the raw data, aka. the patter that we expect to see. Our model’s graph looks like this:

Scatter chart

Description automatically generated

So clearly it’s different. However, that being said, it’s not terrible. You can see at a few different points some of the line bevaior at different square footage points that the actual data has. It’s more dense than it should be, but overall relatively accurate. This is especially true if we compare it to the regression model:

Chart, scatter chart

Description automatically generated

Here we can see the downsides of a regression model. It’s clearly taking on a very linear shape (as it has to) and overall doesn’t model the data at all.

A similar pattern happens with the gross square feet:

Actual

Chart, scatter chart

Description automatically generated

Predicted from KNN

Chart, scatter chart

Description automatically generated

Predicted from regression:

Chart, scatter chart

Description automatically generated

Part 2b

To determine how accurate the model actually is I decided to do a mean squared error test.

To start, the MSE of our model was 1.157 \* 1011 and the MSE of the regression model was 7.051 \* 1011. As you can see, the KNN model’s error was almost one seventh of the regression model! This means that it’s generally more accurate.

I want to explain why I cleaned the data as I did. MSE is not perfect, as it is notably strongly influenced by the scale of the data and any outliers that occur. This is the main reason why I removed outliers from the dataset. If we do the same process but with outliers involved, the MSE of our model becomes 8.629 \* 1013 and the regression model’s MSE is 7.761 \* 1012.  All this means is that the regression model might deal with outliers slightly better than the KNN model, but since MSE is so strongly influced by then that’s all we see in the result. If you actually compare the graphs of land vs. price, you can see that the KNN model still shows a structure more similar to the actual data:

KNN on left, actual in middle, regression on right:

Chart, scatter chart

Description automatically generated Diagram

Description automatically generated Chart, scatter chart

Description automatically generated

We can also compare when the KNN model had all numeric columns vs. just the land columns. When it had all numeric columns the MSE was 1.440 \* 1011. As a reminder, MSE with just the land columns is was 1.157 \* 1011. So while it’s not a huge improvement, theres still is one.

Part 2c

Overall I’m very happy with the results of the model. Looking at the graphs, it’s clearly so much stronger of a model than regression. The model is still a long way from being able to predict prices accurately, but I would say it can get a decent ballpark guess for data that aren’t outliers. If you want to have a model that can predict with extreme accuracy, it’s going to need to be 1. A lot more complex and 2. Have a lot more variables. The dataset given is great, and the land is the most important when it comes to predicting price, but in order to make use of say the longitute and latitude varaibles the model would have to be able to “undrestand” that some areas of queens are more valuable than others. This is certainly possible, but not with KNN. Addionally, there’s only so much you can do with data, as you can’t quantify how beatiful a house looks or how annoying the neighbors are (although you can certainly try!).