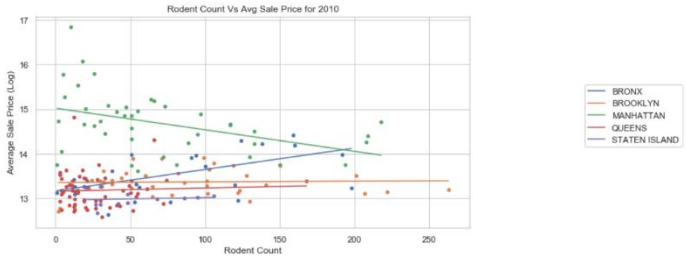
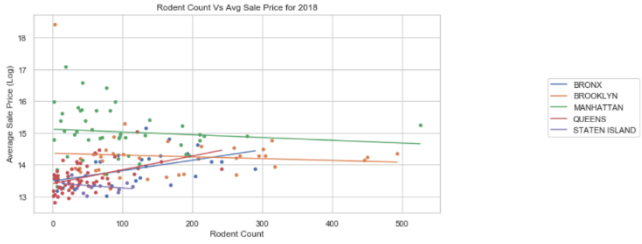
For this project, I wanted to see if there was a better way to predict the price of a house. There are many real estate companies that deploy machine learning models to find a price, however many of them tend to use properties that were recently sold and average the prices of ones that are like them. I wanted to see if construction data would help achieve a better model.

The main features that I focused on for the prediction model are the rodent infestation count, crime report count, and the average prices of the construction costs for properties. For these features, I grouped them by their borough and zip code and did a simple linear regression to see the relationship between those features and the average sale prices of the properties. On the right, you can see a simple regression of rodent data for 2010. It seems that some boroughs have a correlation of some kind with a good p score while other boroughs seem to have no correlation at all. It seems to be the same for the other dates as well, like the example of 2018 right below the 2010 one. These seem to show that it is possible for rodent data to influence the sale prices of some boroughs. It also changes depending on the data. In 2010 rodent reports seem to have little effect on the sales price for Queens, however that changes in 2018.Similar to the rodent data, the crime data also only seems to show correlation for only a few of the boroughs.

```
Correlation for BRONX is: 0.522256 with a p value of 0.007405
Correlation for BROOKLYN is: 0.030956 with a p value of 0.853626
Correlation for MANHATTAN is: -0.418309 with a p value of 0.006495
Correlation for QUEENS is: 0.048196 with a p value of 0.719392
Correlation for STATEN ISLAND is: 0.100614 with a p value of 0.755707
```




```
Correlation for BRONX is: 0.458785 with a p value of 0.021068
Correlation for BROOKLYN is: -0.088937 with a p value of 0.595428
Correlation for MANHATTAN is: -0.109781 with a p value of 0.478085
Correlation for QUEENS is: 0.451622 with a p value of 0.000229
Correlation for STATEN ISLAND is: -0.248796 with a p value of 0.435529
```




```
Correlation for bronx is: 0.634319 with a p value of 0.000501
Correlation for brooklyn is: -0.151978 with a p value of 0.355694
Correlation for manhattan is: 0.263885 with a p value of 0.073083
Correlation for queens is: 0.321417 with a p value of 0.009604
Correlation for statenisland is: -0.039920 with a p value of 0.896978
```

For all three main datasets, they were analyzed using the Pearson correlation (calculated below), however it doesn't really show anything that could be extremely meaningful. It really only shows that there might be a meaningful connection between those features and the sales price of properties. However, it is better that there is some correlation instead of no correlation.

$$r_{pb} = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}}$$

Pearson's correlation