1. I plotted the data and realized that some of the zipcodes have a very few incidents reported. I came up with a filtered dataset which takes only the zipcodes that have more than 10 incidents. I choose 10 as a threshold because if I choose 100 it only differs by 30 more data points. For me, it was insignificant.
2. I produced python codes that report the 10-fold croos-validated RMSE and R^2 scores. However, RMSE scores do not decrease as the degree of the model increases. R^2 scores look pretty accurate.
3. I plotted everything and I chose the 3rd degree model for the estimation process.
4. I built my final OLS model and predictions are plotted on a graph.

Since the data has some unrealistic incident reportings, I definitely needed to know what incidents mean in these reports. Some places have only 1 incident reported and other have more than a thousand. This created an imbalanced dataset which makes it harder to predict the incidents in other places. If the incidents are categorized, it would be beneficial to run some other models for critical incidents such as accidents, robberies etc.