TOPOS DATA SCIENCE INTERNSHIP ASSIGNMENT

A Study of the correlation between construction permits issued and traffic collisions reported.

By
Shabarish Kumar Rajendra Prasad
Texas A&M University
shabarik@tamu.edu
(346)-332-6237

Introduction

Construction works are always a hinderance to vehicle traffic. Most often construction works lead to accidents in a neighboring area. The aim of this investigation is to find how a building under construction could affect the flow of traffic by finding whether there is any correlation between the number of traffic collisions and the number of constructions works in progress. If there is any correlation, I shall present the evidences that support my claim and also list any shortcomings of the methods that I use.

Datasets

I started this investigation with two datasets. Both these datasets are available on the NYC Open Data website.

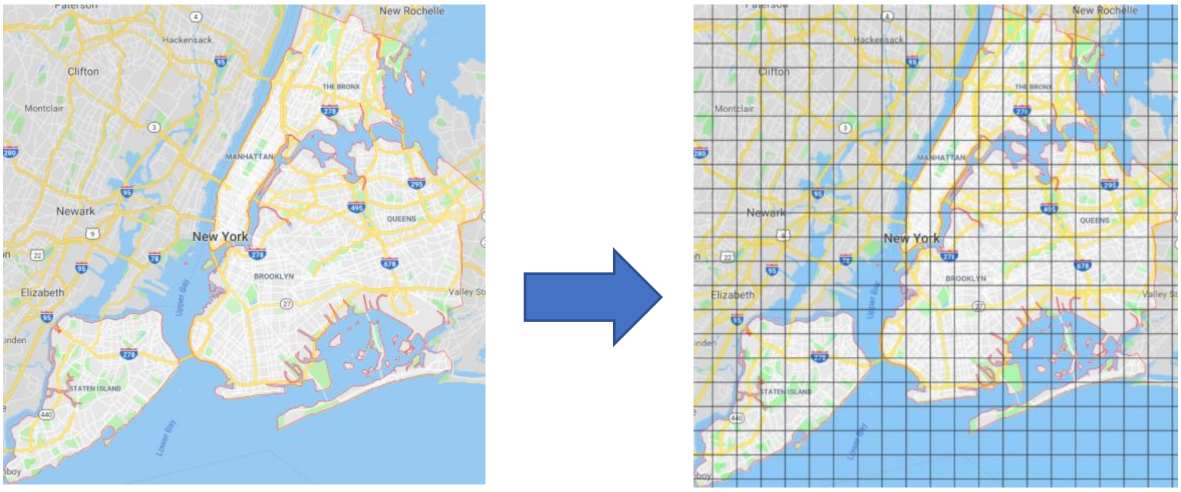
- DOB Permit issuance
- NYPD Motor Vehicle Collisions

Methodology

The DOB Permit issuance dataset contains the record of all jobs filed, requesting permit for construction in New York City from the year 1987 to 2019. Whereas the NYPD Motor Vehicle Collisions dataset contains the record of all the traffic collisions reported between 2012 and 2019, all over the state of New York. In this investigation, we shall focus only where these two datasets overlap. So, the scope of this investigation is narrowed down to the construction works and traffic collisions reported in New York City between 2012 and 2019. I have further filtered the permit issuance dataset to consider the records that are either '*ISSUED*' or '*RE-ISSUED*', in the assumption that in such case, construction work would be active from the date that the permit issued till the expiration date on the permit.

First, I tried mapping the number of constructions active on a particular road with the number of traffic collisions reported on the same road, given a calendar month and year, for every road in the Vehicle Collisions dataset. The vehicle collisions dataset classified roads in three ways as whether the collision occurred on the road, across the road or off the road. We consider that the road is involved in the collision if the road name is present in either of the three fields. But due to some inconsistencies in naming the roads, I could not map about 40% of the DOB Permit Database.

So, in addition to mapping the construction works in an area and the traffic collisions reported in that area based on the roads that incidents are taking place, I also wanted to come up with a more efficient way of mapping the two activities. The next evidence that would determine whether the two activities take place within a relatively close distance is the latitude and longitude of the exact geographical location of the two activities mentioned on the datasets. Hence, I thought of dividing the latitude values between 40.499031 and 40.912869 and the longitude values between -74.254886 and -73.700736 into 20 bins each. The intersection a latitude bin and longitude bin will be a block, and thus dividing the entire map of New York City into blocks based on latitude and longitude and I can examine any correlation within any single block. The blocks I devised are shown on the map below.



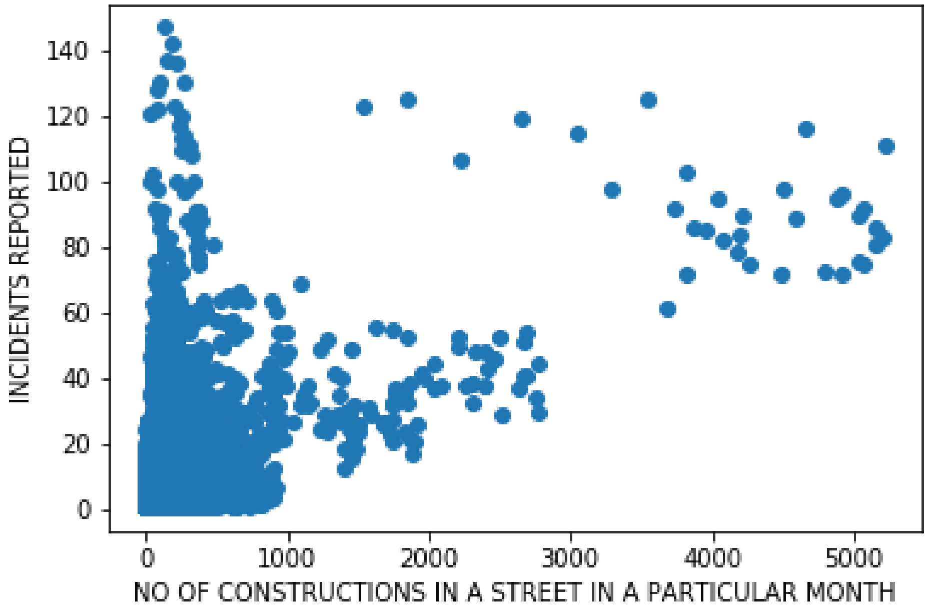
(a) The map of New York City

(b) The same map with grid lines showing different blocks

Dividing the map this way allowed me to map all data from both the datasets.

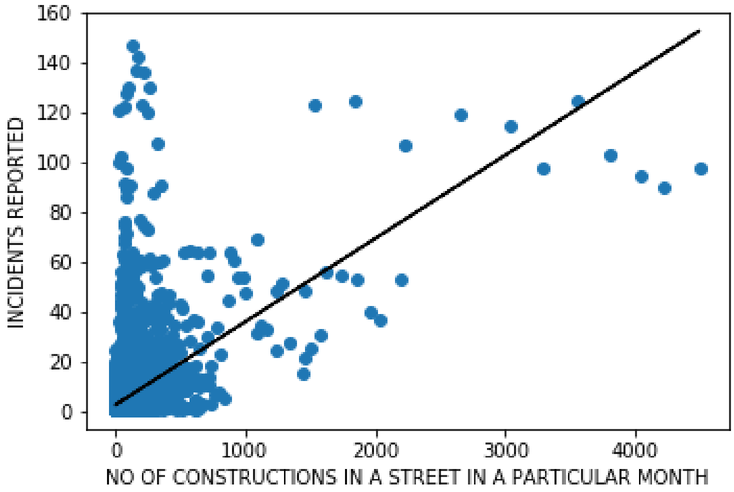
Results

I found a positive correlation between the number of construction works in progress in a road and the number of vehicle collisions reported on the same road, given a calendar month and year. The scatterplot between the two activities.

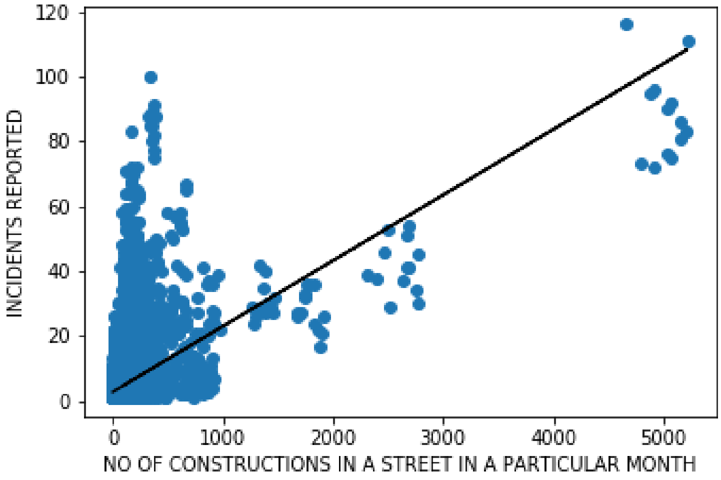


I also tried separating the plot for each year. But, due to the inconsistencies in naming mentioned above, I cannot get any data for years between 2012 and 2015. But we have the plots for the all the years after 2015 till 2018. These graphs are shown below.

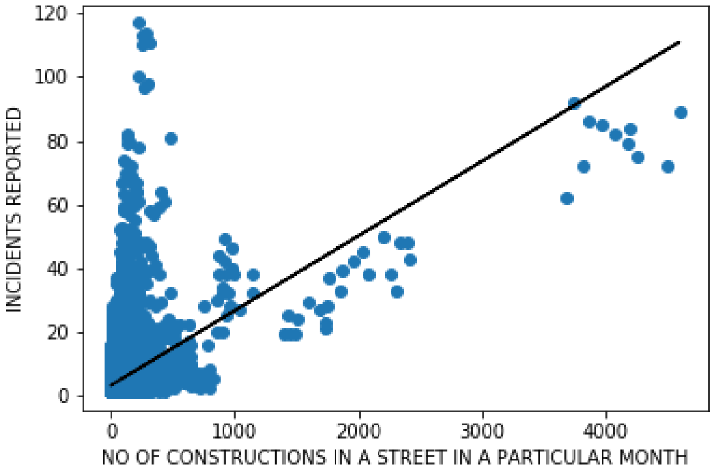
YEAR 2016:



YEAR 2017:



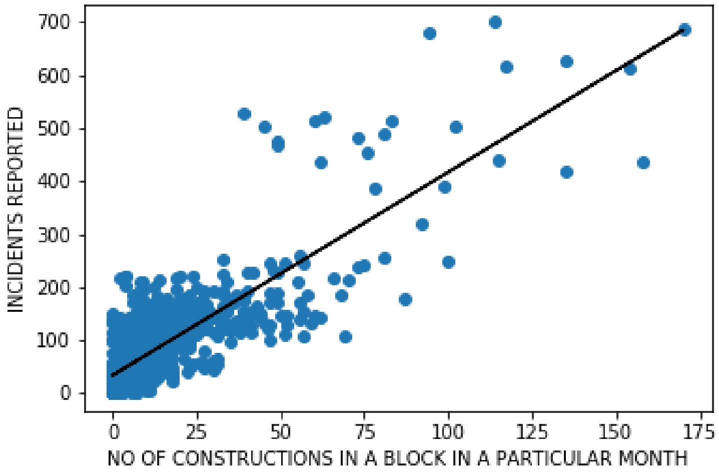
YEAR 2018:



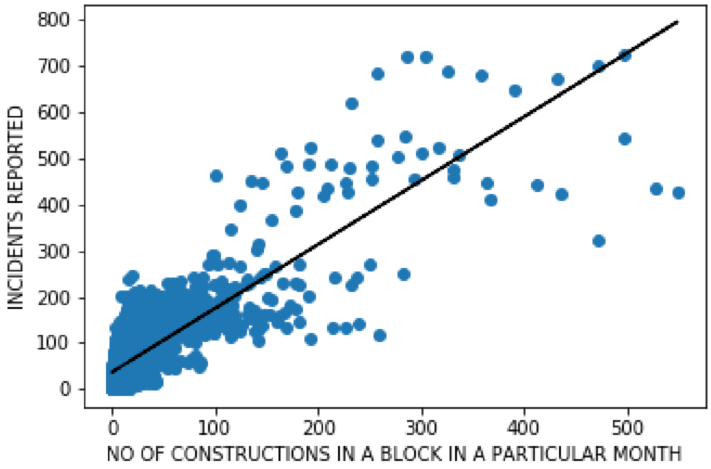
The black lines through the graphs show a linear model trying to predict the number of traffic collisions given the number of construction works in progress. We can see that the predictor shows an upward trend every year. Thus we can say, the higher the number of constructions in work in a road, the higher the expected number of traffic collisions on that road.

This same fact is further cemented when plotting the construction works in progress vs traffic collisions plot for the latitude-longitude bins mentioned earlier. Those plots, year-wise are shown below.

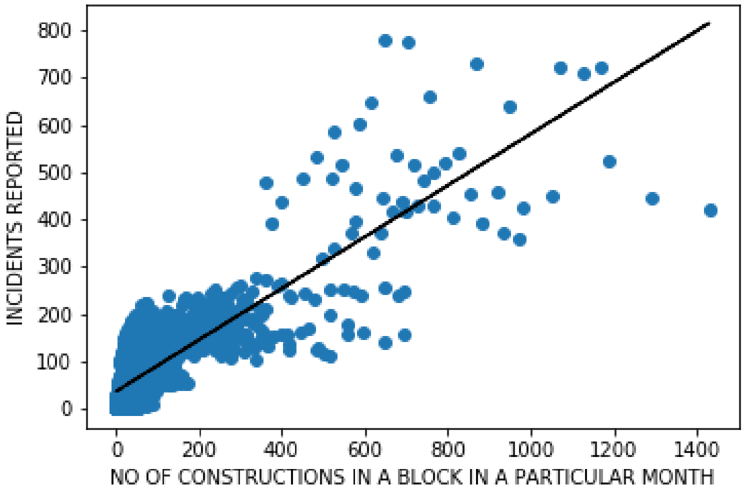
YEAR 2012:



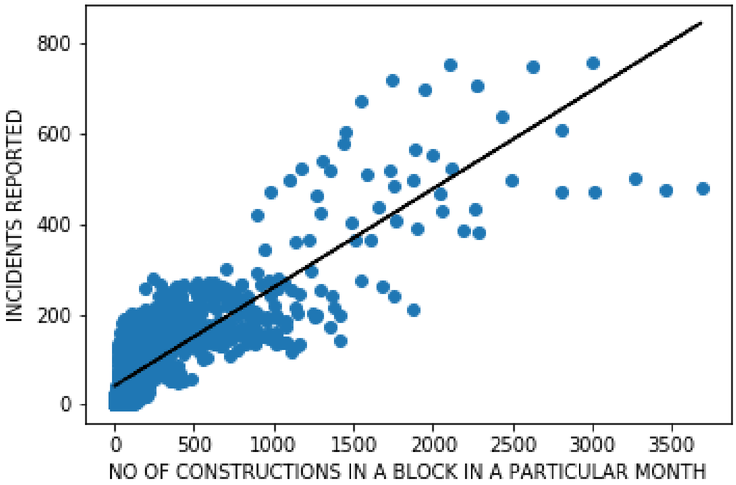
YEAR 2013:



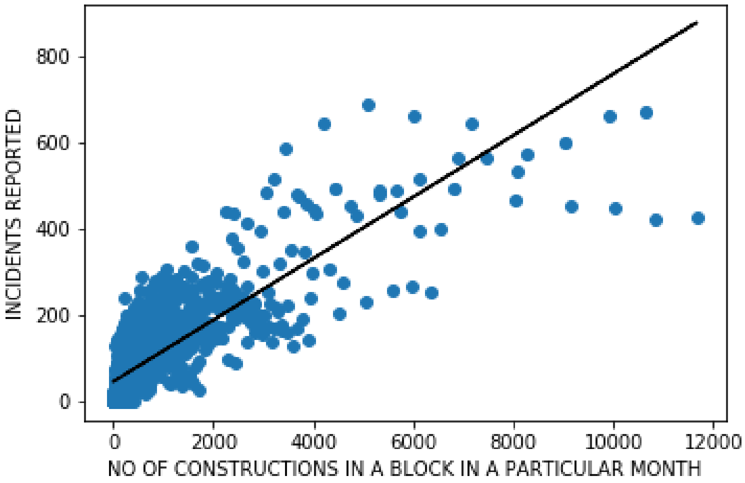
YEAR 2014:



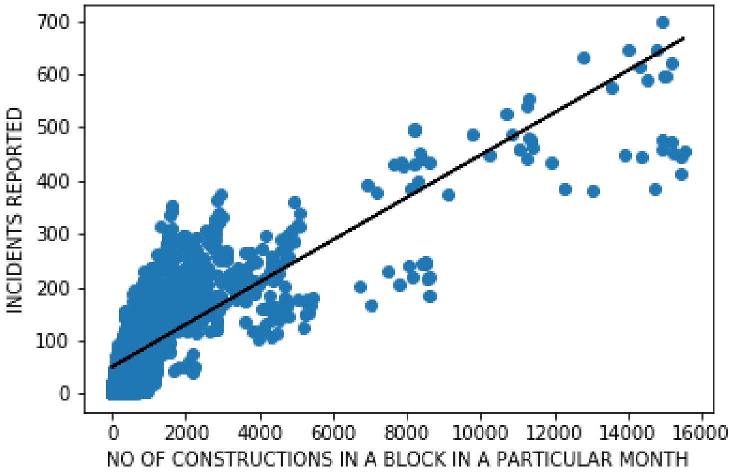
YEAR 2015:



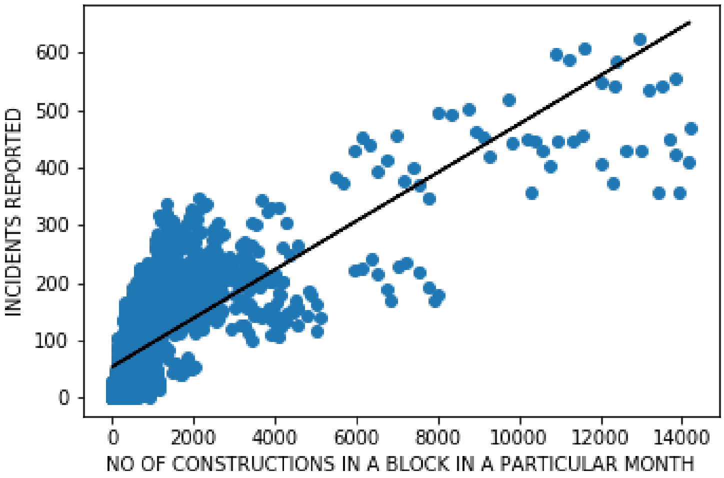
YEAR 2016:



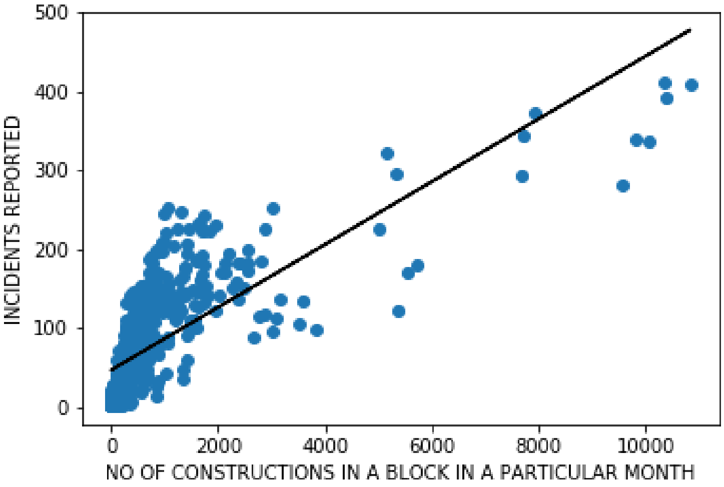
YEAR 2017:



YEAR 2018:



YEAR 2019:



We can see that these graphs also show similar trends to the plots made based on roads. Further, these plots also hint at the fact that the trend has also been the same in the years between 2012 and 2015 and even in 2019.

Shortcomings

Although we found that if the number of construction works in a road is higher in number then the number of traffic collisions reported on the same road are also high, there is no evidence to say that they take place close to each other (say within 500ft), since many roads in NYC run miles long.

The same thing cannot be said about the relationship between the activities in the latitude and longitude bins mentioned earlier, since we know it could not have happened really far away on the same road. But even then, the size of the grids that I used are fairly large in size and in order to make a solid claim, we should examine the numbers in even smaller areas, reducing the size of the grid to city block level or slightly bigger. And even if they happen really close to each other, there could be no evidence to say whether the construction works are the reason for the collision in the datasets.

Conclusion

Apart from the shortcomings mentioned in the section above, I believe that the evidence from this investigation would still hint us that there is a correlations between the number of construction permits issued and the number of traffic collisions reported in a given area. Further, the study of the numbers in the grid space also clearly show that, the lower the number of construction permits issued in a block, the lower the number of traffic collisions reported in that block. Hence, the city of New York should consider issuing a lower number of permits in an area, if they wish to help the flow of vehicle traffic in that area and reduce the number of traffic collisions in that area.