Big-Data-Driven Car Speeding Prediction for NYC

Yeqian Yang

New York University

New York, USA

yy1420@nyu.edu

Roni Yosofov

New York University

New York, USA

ry856@nyu.edu

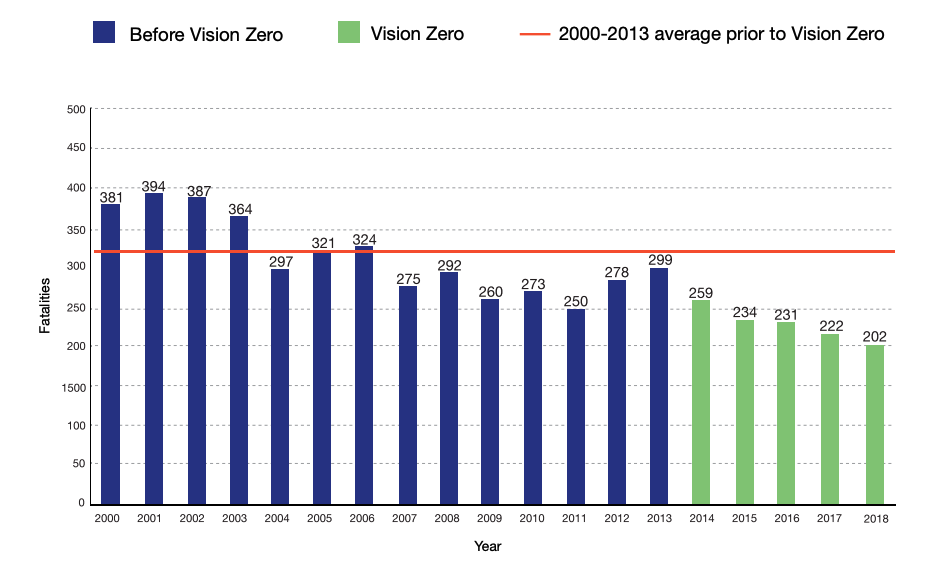
Bilal Munawar

New York University

New York, USA

bm2515@nyu.edu

***Abstract*—**

*The increasing number of satellites, sensors, and cameras on the roads of many urban areas enables authorities and organizations to collect data generated by vehicles. Large traffic data in major metropolis is constantly generated and this paper looks into the correlation between weather, speed limits, congestion and traffic speeds. Our hypothesis is that some or all of these features directly impact car speeding in certain areas and in certain times of the day. An analysis of traffic data to identify the features that strongly correlate with car speeding will allow for optimized police cars placement to stop the violators.*

# Introduction

The increasing number of satellites, sensors, and cameras on the roads of many urban areas enables authorities and organizations to collect data generated by vehicles. With the availability of huge amount of traffic data, there is a surge in demand for systems that can process large-scale data. Chen et al.[1] proposed a method of distributed modeling under a MapReduce framework for data-driven traffic flow forecasting to meet the computation and data storage requirement for this particular application.

However, most existing research focuses on using big data to predict traffic congestion and the application of this kind of research is limited to route suggestions based on traffic conditions. In this project, we decided to utilize some open source datasets and exploit big data analytical tools, such as Hadoop MapReduce, and [Apache](https://hive.apache.org/) Hive to help determine where, when and under which certain conditions speeding violations are more likely to happen.

Based on our findings about the patterns of speeding violations, we then proposed some policy recommendations to optimally place police cars in locations at certain time periods that speeding violations are predicted to happen with higher probability to mitigate violations.

# Motivation

According to NYC Vision Zero Year 5 report[], since the start of the Vision Zero initiative in 2014 which lowered the speed limit, increased enforcement, and designed hundreds of safer streets, New York City has witnessed a continuous decline in traffic fatalities.

Figure 1 NYC Traffic Fatalities

As pointed out by the Vision Zero initiative 2017[], deterring speeding is critical because the stopping distance of a vehicle increases in direct proportion to the speed of the vehicle. This means that the faster a vehicle is moving the longer it takes for the driver to bring the vehicle to rest, and therefore harder for the driver to avoid a crash. In fact, a driver at 40 MPH needs 300 feet to perceive, react and brake to an unexpected event – twice as far as a driver at 25 MPH, who only needs 150 feet. It is important to lower the speed at which a vehicle crashes because a pedestrian who is struck by a vehicle traveling at 30 MPH is twice as likely to be killed as a pedestrian struck by a vehicle traveling at 25 MPH.

However, against all efforts the city has made, according to the statistics from both NYPD, 82 people were killed on the streets of New York between Jan. 1 and June 2 this year which increased 26.2 percent compared to the same period in 2018. One of the reasons for the increase in deaths is speeding. According to the Guardian[], family members of these victims gather at City Hall and held a banner reading: “Vision Zero is in a state of emergency.” Does this mean Vision Zero is starting to fail? We think the initiatives are still good but the problem lies in that the public resources are afterall limited. This paper aims to use easily accessible big data to help determine where, when and under which certain conditions speeding violations are more likely to happen. And therefore, limited public resources such as police cars can be optimally placed and achieve a better enforcement of the speed limit.

# Related Work

With the availability of huge amount of traffic data, distributed systems and tools such as Hadoop MapReduce have been widely used to conduct traffic data analytics. Most existing research focuses on predicting traffic flow congestion.

Hoang et al.[3] proposed a scalable prediction framework based on big data including human mobility data, weather conditions, and road network data. They decompose flows into three components: seasonal, trend and residual flows which catches periodic patterns, changes in periodic patterns and instantaneous changes respectively to model the various complex factors affecting traffic. They build seasonal and trend models as intrinsic Gaussian Markov random fields and exploits spatiotemporal dependence among different flows and regions and also weather condition to build the residual model. Their experiment results showed that their approach is scalable and outperforms baselines significantly. Xia et al.[2] utilized offline distributed training and online parallel prediction by utilizing a MapReduce-based KNN model for traffic flow prediction using correlation analysis (TFPC). KNN was chosen for a few reasons. KNN is a very simple model, no many parameters need to be optimized (risk of overfitting) and it’s capable of processing very complex sets of data. The model tried to analyze the correlation of space and time as an inherent feature of traffic flow in a complex urban transportation networks (e.g. Beijing). The assumption they made was that the traffic flow of the target road segment at the future time interval is closely related to that of the same road segment at the previous and current time intervals. The accuracy of the method used in the model was over 90% in the best case, which significantly improved the efficiency and the scalability of traffic flow prediction.

Joaquim et. Al [4] \*\*need to summarize\*\*\*  
<https://repositorio.inesctec.pt/bitstream/123456789/7047/1/P-00G-T25.pdf> @ Roni still need?

However, there’s limited research on the speeding violation detection using big data. Indeed, several previous papers have identified the role speed limit plays in traffic. Gao et al. [5] investigated the effect of the posted Speed Limit on the Dispersion of Traffic Flow Speed. Three different sections of the same highways with speed limits of 80 km/h, 100 km/h, and 120 km/h along with their traffic-volume and speed are selected for observation. The relationship between the different speed limits and the above-mentioned indicators is studied. The results show that the speed limit has a high correlation with the average speed of traffic flow, with the coefficients of determination as high as 0.84, 0.85, and 0.92, respectively. The results highlight two important findings. Firstly, in restricted traffic flow state, decrease in driving freedom is more likely to affect the driving speed rather than the speed limit value. Secondly, in traffic free flow state, the speed limit has a proportional relationship with the average speed of traffic flow. The results of this study are useful for determining the suitable speed limit under different traffic-flows.

With the advancement of analyzing real-time data to predict traffic volume, there has been research to optimize the enforced traffic speed limit under various traffic and weather conditions. Ye et. al. [6] investigated the relationship between variable speed limit (VSL) and the potential risk of crash. The results of the study show that the speed at which the vehicle is crashed is closely related with the total number of casualties. In fact, driving at 25 MPH or lower significantly reduces the stopping distance, giving drivers and pedestrians more time to react, and hence reducing the impact of collision. Crashes by motor vehicles travelling at a speed of 25 MHP are half as likely to die as those travelling at 30 MPH. The severity of collision can be directly controlled by the travelling speed of the vehicle. Therefore, the author of this paper proposes a proactive VSL control algorithm, which aims to reduce the potential risk and severity of crash. The VSL control algorithm takes into account the responses of drivers to advised virtual speed limit and uses it along with the current traffic conditions to predict future traffic conditions. This framework allows the VSL control algorithm to effectively suggests a dynamic speed limit to achieve a safer and uniform traffic flow. The results suggest that VSL effectively reduces crash probabilities in high-traffic flow state. Therefore, speeding violation detection would be extremely meaningful for NYC to help enforcement of the speed limit and thus protect the safety of people in the street.

Overall, past research has been more focused on predicting the traffic volumes and analyzing the effect of speed limit on traffic accidents, but there’s a lack of research on studying the features that affect traffic speeding. This paper investigates the features that affect traffic speeding to effectively predict traffic speeding, and therefore suggest various measures to detect & stop these speeding violations to reduce potential crashes.

# Design and Implementation

1. Design Details

### 

For the data ETL (extract, transform, load) stage, we downloaded data from multiple sources, moved the big data to HDFS and used MapReduce to extract the columns of interest, deal with the missing values, and transform the data to the desirable format.

We then merged all three datasets into one to be able to utilize all or some of the features that touched all three datasets. We utilized both Hive to merge the Traffic Speed and Weather datasets on the date field (yyyyMMddHHmm had to match). Then, to merge with the Speed Limit dataset, we needed to write a MapReduce job to load the dataset on the fly while having the HMR job go through the initially merged dataset (Traffic Speed + Weather), and merge on matching latitude and longitude coordinates. A threshold for the difference of the coordinates was set up to be 0.0005. This threshold was used to merge Speed Limit data with the (Traffic Speed + Weather) data based on coordinates

The last step was to utilize Hive/Impala to run some analysis on the merged data, find trends and draw conclusions on where and when police forces need to be to increase the chances of catching speeding violators.

(maybe we can add reasons for our choice of the mentioned big data tools)

# Datasets

1. Real-Time Traffic Speed Data (a snapshot)

NYCDOT maintains a map of traffic speed detectors throughout NYC. This data feed contains 'real-time' traffic information from locations where NYCDOT picks up sensor feeds within the five boroughs, mostly on major arterials and highways. The dataset has about 25 million rows of data from 2014 to 2019. The size of the dataset is about 12GB and it requires some big-data tool to even access it outside NYCDOT’s website. We extracted the following information from the dataset:

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Description** | **Type** | **Range of values** |
| Speed (mph) | Average speed a vehicle traveled between end points of the link | Float | Min: 0.0  Max: 186.41 |
| travelTime (secs) | Time the average vehicle took to traverse the link | Integer | Min: 0  Max: 36430 |
| observationTime (GMT) | Last time data was received from link | LocalDateTime (yyyyMMddHHmm) | Min: 201403211523  Max: 201910291313 |
| latitudeStart | Sensor’s initial latitude coordinate | Float | NYC coordinates \* |
| longitudeStart | Sensor’s initial longitude coordinate | Float | NYC coordinates \* |
| latitudeEnd | Sensor’s final latitude coordinate | Float | NYC coordinates \* |
| longitudeEnd | Sensor’s final longitude coordinate | Float | NYC coordinates \* |
| borough | NYC borough | String (Enum) | Brooklyn, Bronx, Manhattan, Queens, Staten Island |
| address | the link location | String | Min length: 16 chars  Max length: 77 chars |

\*NYC Latitude range => [40.495992, 40.915568] NYC Longitude range => [-74.257159, =73.699215]

1. Hourly Central Park Weather

The weather dataset can be obtained from National Climate Data Center[5]. It is hourly surface data observed from the Central Park station during 2005-01-01~2019-10-30. The whole dataset has a size of 22.6 MB and contains 33 features, of which we extracted 6 features to our interest as shown in the following data schema.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Description** | **Type** | **Range of values** |
| ObservationTime (UTC) | Time of observation (approximately every hour) | DateTime (yyyy-MM-dd HH:mm) | Min: 2005-01-01 00:51 GMT  Max: 2019-10-30 05:16 GMT |
| Wind speed | Speed in miles per hour | Integer | Max: 44  Min: 0 |
| Visibility | Visibility in statute miles to nearest tenth | Float | Max: 10.0  Min: 0.0 |
| Weather\_code | The code that denotes a specific type of weather observed | Enum |  |
| Temperature (F) | Temperature in Fahrenheit | Integer | Max: 102  Min: -1 |
| Precipitation | Precipitation for the preceding 1 hour period, in inches and hundredths | Float | Max: 1.70  Min: 0.00 |

Table. Weather Data Schema

C. Speed Limit Data

NYCDOT contains a map of all the different Speed Limits imposed in New York City. The Speed Limit is imposed based on the categorization of the road. For example, the Speed Limit imposed in a school-nearby area is expectedly different from speed limit generally allocated on highways. The dataset spans all of New York City and contains a map of all the coordinates (start, end) of streets, highways, and bridges in NYC along with its speed limit. Unless otherwise stated, all streets in New York City operate on a speed limit of 25 mph. The dataset contains 152749 rows and has a size of 350 MB. It contains 9 columns, all of which we have extracted to our interest as shown in the following data schema.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Description** | **Type** | **Range of values** |
| Street | The name of the street | String | Max String Length: 35 |
| SpeedLimit | The maximum speed allowed | Integer | Max: 50 mph  Min: 0 mph  Average: 25.73 mph |
| latitudeStart | The initial latitude coordinate for which the speed limit is imposed | Float | NYC coordinates \* |
| longitudeStart | The initial longitude coordinate for which the speed limit is imposed | Float | NYC coordinates \* |
| latitudeEnd | The final latitude coordinate for which the speed limit is imposed | Float | NYC coordinates \* |
| longitudeEnd | The final longitude coordinate for which the speed limit is imposed | Float | NYC coordinates \* |
| isSigned | Specifies if the region has a sign with the speed limit | Boolean | Signed Yes: 4,3228  Signed No: 109,526  Total: 113,854 |
| shapeLeng | Length in meters of the geometry | Double | Max: 10248.67  Min: 3.102  Average: 269.90 |

\*NYC Latitude range => [40.495992, 40.915568] NYC Longitude range => [-74.257159, =73.699215]

Table. Speed Limit Data Schema

We merged datasets [A], [B] and [C] and this is the schema we decided to go with:

|  |  |
| --- | --- |
| **Name** | **type** |
| id | Int |
| speed(mph) | Double |
| travelTime(secs) | Double |
| observationTime | Bigint |
| month | String |
| dayOfWeek | String |
| hour | Int |
| latitudeStart | Double |
| longitudeStart | Double |
| latitudeEnd | Double |
| longitudeEnd | Double |
| Borough | String |
| address | String |
| windSpeed | Double |
| visibility | Double |
| temperature | Double |
| pcp | Double |
| speedLimit | Double |
| isSignShown | String |
| differenceSpeedAndLimit | Double |

# Results

(In this section, you can describe: Your experimental setup/issues with data/performance/etc. Describe your experiments, describe what you learned. Did you prove or disprove your hypothesis? Were some results unexpected? Why? What actionable insights did you discover?)

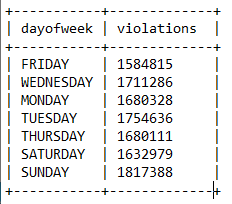
Our hypothesis is that speeding violations happen on specific roads, at specific times and under specific conditions. As a result, we think that a more optimized law enforcement presence (e.g. police cars) can increase the effectiveness of catching speeding violations. For our analysis we started with simple queries and built on top of them with depth and complexity to discover useful insights.

In our analysis we utilized a few very important fields:

* **id**: the unique identifier of each link that recorded traffic speed.
* **differenceSpeedAndLimit**: indicates by how many miles a driver was over the speed limit when passed by a link. Each such occurrence was considered a violation by us.

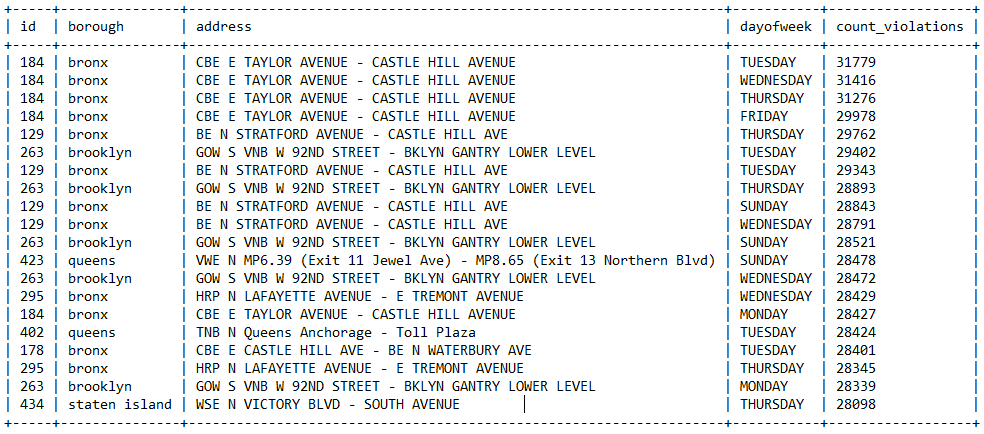
The rest of the fields should be very intuitive to understand as we progress.

Initially, we wanted to see how many times drivers drive by 15 miles or more over the limit on different week days. We grouped the query by day of the week and counted the times the event described above occurred. We clearly noticed that the distribution was not constant and Sunday, Tuesday and Wednesday were the most popular days for violations (please see the table below). Another, query that we ran showed that the violations were uniformly distributed in our dataset, so the findings are pretty reliable.



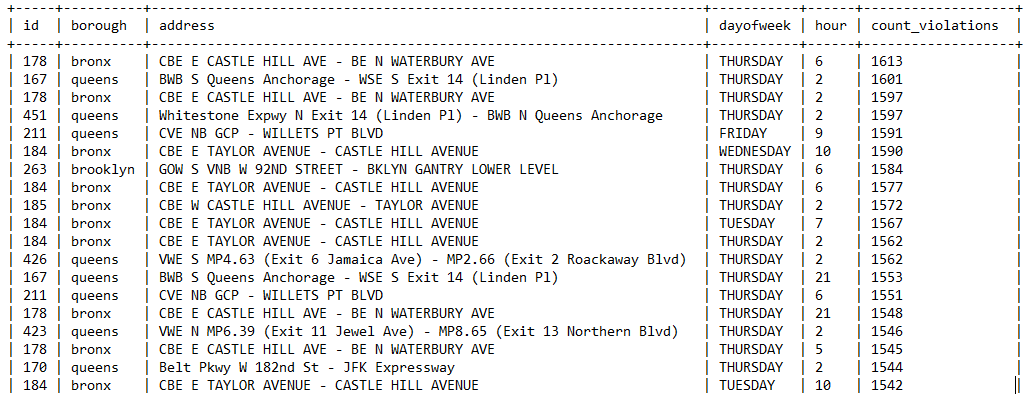
Certainly, the above results don’t tell the whole story, but the first preliminary recommendation is for the police forces is to increase presence on Sundays, Tuesdays and Wednesdays throughout NYC.

Then, we decided to combine violations by location and day of week (the previous query) to see if we could narrow down our search:

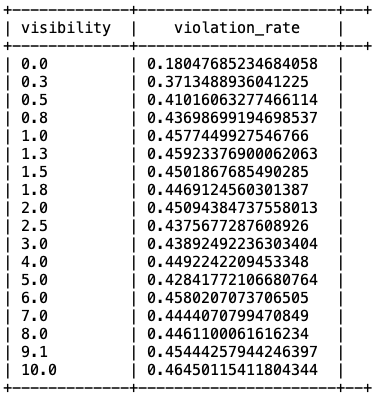
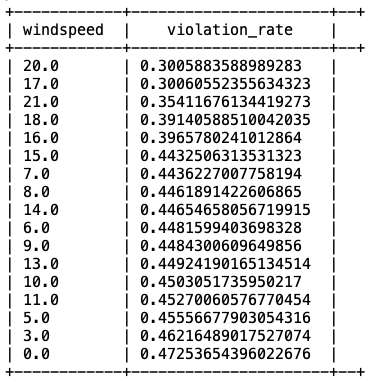


From the top 20 locations/days, we see that the Bronx and Brooklyn are leading and that most violations happen on avenues and usually on Tuesdays, Wednesdays and Sundays. We can see that some roads (e.g. id 184, 129 and 263) are specifically popular. The question that we asked ourselves next was what makes certain roads to be more prone to violations than others?

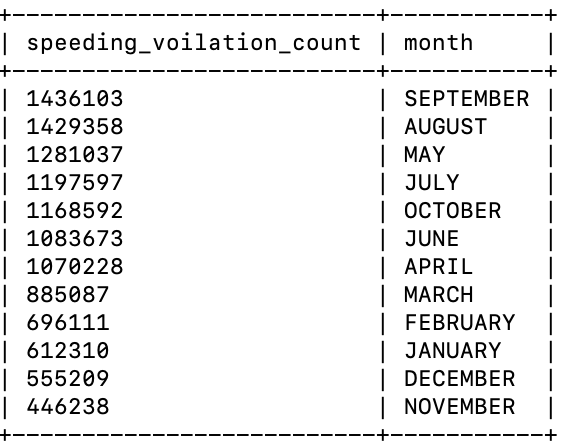
Next, we wanted to go a step deeper to further narrow the presence recommendation of police forces. We decided to look into how the time of the day impacts the query above.



This is interesting because the profile has changed a little when we brought hours into consideration. The roads with most violations stayed consistent, but the most popular day for violations is now Thursday. One reason for the change could be that violations on Tuesday, Sunday and Wednesday are more evenly spread-out throughout the day. Also, we noticed that most violations happen after 8/9pm and all the way to 6/7am.

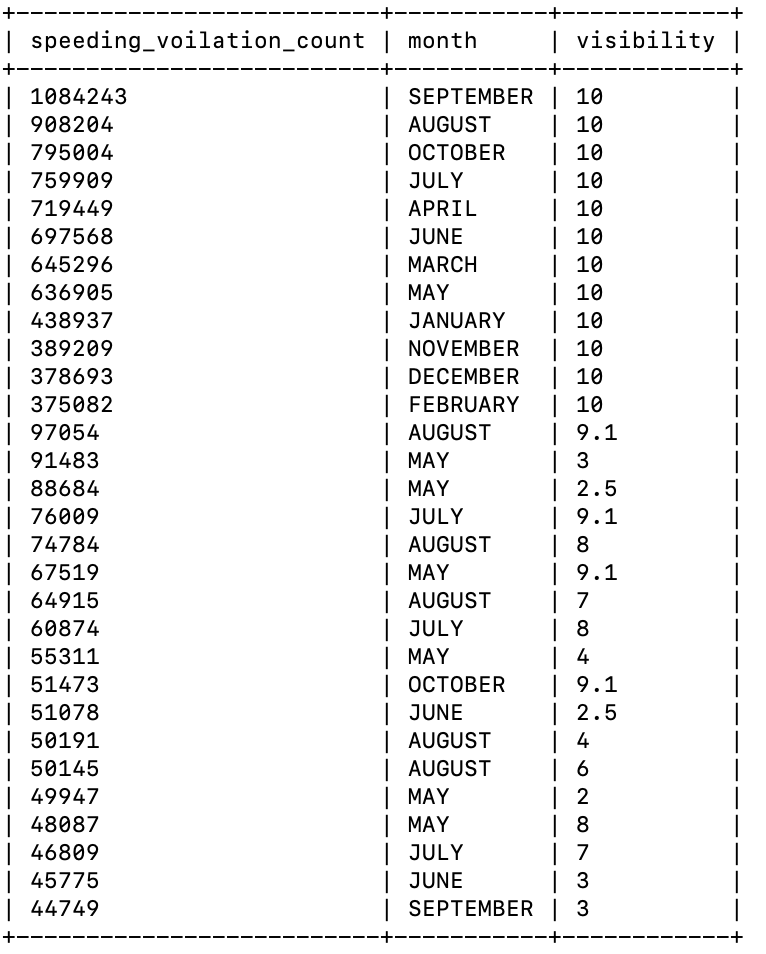
Next, let’s look into the weather and see under what kind of weather conditions speeding violations are more likely to happen. We calculated the violation rates for different wind speeds, visibility levels, precipitation amounts and temperatures. For the wind speed, we find that when it is windy, the speeding violation rate tends to decrease. The negative correlation of -0.02 calculated between the column “windspeed” and “differencespeedandlimit” also supports this finding. For the visibility level, we can observe that when the visibility level is low, the speeding violation rate tends to be low as well and a positive correlation of 0.02 calculated between the column “visibility” and “differencespeedandlimit” also supports our observation. For the precipitation amount, the relationship is not quite clear by sole observation of different precipitation amounts and the violation rates under these conditions. We got a correlation of -0.003 between the column “pcp” and “differencespeedandlimit” which shows a very weak negative correlation. For the temperature, though there’s a negative correlation of -0.05 between the column “temperature” and “differencespeedandlimit”, we find that extreme temperatures have different impact on speeding violations. When the temperature is very low(<=20°F), speeding violation rates are high(>50%) while when temperature is very high(>=85°F), speeding violation rates are relatively low(<40%).

Finally, we would like to investigate if there is any effect of the month of the year on the speeding violations. Here are our results of the total number of speeding violations by each month of the year:



From the table above, we note that speeding violations are most likely to occur in the month of September and August. We also note that the colder months (November - February) are among the bottom four of the speeding violation count table.

Next, we would like to explain why certain months are more likely to have greater speeding violations than the others. We run the same query grouped with the visibility level. Here’s our results:



We come across an interesting result. From the table above, we understand that certain months pointed out earlier are more likely to have greater speeding violation because visibility in these months is higher. This also explains why colder months would be at the bottom end of the speeding violations table. This is because visibility in colder months is greatly diminished because of the weather. In conclusion, we understand that a higher visibility is likely to encourage the driver to drive faster, leading to traffic speed limit violation.

# Future Work

During the data exploration and analysis, we utilized our common sense and our knowledge to discover trends and features. However, machine learning can automate many of these steps and discover even deeper insights. So, we would love to utilize SparkML to try and find trends to answer questions.

# Conclusion

(A few paragraphs about the results and the value/accuracy/goodness of your analytic.)

To conclude our analysis, we should encourage the police department to follow a few steps. Firstly, increase presence on road 184, 263, 129, 295, 402 (top 5 roads with most violations in total). Secondly, spend more time in the Bronx, Queens and maybe also Brooklyn. Thirdly, place more police forces on the roads in second shift (8pm to 7am). Also, have higher presence of cops on Thursdays and Tuesdays. And lastly, police shall pay extra attention for speeding violations in certain weather conditions such as when the visibility is high, when it’s breezeless and when the temperature is extremely low.

##### References

(Add references for all of the papers/texts that you refer to in your paper. You will probably want to include the papers you read that were related to your project. You may have websites to reference, the Hadoop book, the MapReduce paper, the Pig Latin paper, etc. Some references are added below as an example. Links to original datasets and APIs used should be listed here and referenced in this paper where appropriate.)

1. Chen C, Liu Z, Lin WH, Li S, Wang K. Distributed modeling in a MapReduce framework for data-driven traffic flow forecasting. IEEE Transactions on Intelligent Transportation Systems. 2012 Jul 10;14(1):22-33.
2. Xia D, Li H, Wang B, Li Y, Zhang Z. A map reduce-based nearest neighbor approach for big-data-driven traffic flow prediction. IEEE access. 2016;4:2920-34.
3. Hoang MX, Zheng Y, Singh AK. FCCF: forecasting citywide crowd flows based on big data. InProceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems 2016 Oct 31 (p. 6). ACM.
4. Joaquim Barros, Miguel Araujo, Rosaldo J. F. Rossetti. Short-term real-time traffic prediction methods: a survey. 2015 June.
5. <https://www7.ncdc.noaa.gov/CDO/cdoselect.cmd>
6. Gao, Chao, et al. "The Effect of Posted Speed Limit on the Dispersion of Traffic Flow Speed." *Sustainability* 11.13 (2019): 3594.
7. Ye, Huixuan, Lili Tu, and Jie Fang. "Predicting Traffic Dynamics with Driver Response Model for Proactive Variable Speed Limit Control Algorithm." Mathematical Problems in Engineering 2018 (2018).

<https://www1.nyc.gov/assets/visionzero/downloads/pdf/vision-zero-year-5-report.pdf>

<https://www1.nyc.gov/site/visionzero/initiatives/initiatives.page>

<https://www.theguardian.com/us-news/2019/may/07/traffic-deaths-new-york-city-surge-2019>