Investigation of factors associated with traffic incidents in NYC and review of *Vision Zero* effectiveness

CUNY SPS - DATA 698.4 Capstone

Sreejaya Vasudevannair, Vuthy Nguy and Suman K Polavarapu

Abstract

This paper investigates the factors associated with traffic accidents in NYC and reviews the results of the *Vision Zero* plan, an initiative created in 2014 in an attempt to eliminate pedestrian deaths in NYC. Today in New York City, approximately 4,000 New Yorkers are seriously injured and more than 250 are killed each year in motor vehicle accidents. Being struck by a vehicle is the leading cause of injury-related death for children under 14, and the second leading cause for seniors. On average, vehicles seriously injure or kill a New Yorker every two hour. The majority of those killed are people on foot—they could be people headed to work, walking home, out playing, heading to a subway, or waiting for a bus. Pedestrians account for **56%** of all New York City traffic fatalities. Dangerous driver choices—such as inattention, speeding, failure to yield—are the main causes of these incidents. The Vision Zero Action Plan is the City's foundation for ending traffic deaths and injuries on NYC streets.

From our exploratory study, we noticed that even though fatalities appeared to be slightly reduced after the year 2014, there is no such clear indication in case of injuries. Human factors such as Driver Distraction, Improper Driving, violation of laws is without doubt the major reasons for the accidents. And over 50% of the violations are caused by younger drivers. Though Manhattan had the most traffic violations, Brooklyn and Queens had more major accidents that caused injuries and/or deaths. The exploratory study also revealed that Passenger vehicles was most likely to be involved in the accidents that led to injuries or deaths. Exploratory study with pedestrian incidents combined with specific vision zero initiatives showed mixed results - there was an immediate decrease in the years following VZ initiative, however the incidents increased in the year 2017.

The hot spot clusters study indicated mixed results in our vision zero analysis. Though pedestrian hot spots reduced post vision zero initiative, the multi vehicle accident hot spots appears to be increased even after the post VZ initiative (in *Staten Island* and *Queens* boroughs). The time series analysis (done with *driver distraction data in Brooklyn borough*) forecasted gradual decrease in incidents.

Keywords

"traffic tickets, motor vehicle collision, vision zero, spatial data, clustering, DBSCAN, time series, ARIMA, pedestrian incidents, three vehicle incidents"

Introduction

In an effort to reduce traffic fatalities and injuries, NYC released the Vision Zero Action Plan in 2014. The City of New York is no longer regard traffic crashes as mere "accidents," but rather as preventable incidents that can be systematically addressed. The Vision Zero Action Plan is the City's foundation for ending traffic deaths and injuries on NYC streets. The plan seeks to improve street safety by - 1). Increasing the enforcement of moving violations. 2). Improving street designs. 3). Holding public outreach sessions. 4). Increasing penalties for dangerous drivers. 5). Reducing speed limits and 6). Increasing the use of enforcement cameras. Vision Zero program discourage dangerous behavior on roads and streets by combining better enforcement and roadway engineering with improved emergency response and public campaigns on safe driving.

There are 14 different Vision Zero initiatives that are in place today. The NYC administration strongly believes that this multifaceted approach of Vision Zero, incorporating education, engineering, enforcement and legislation, is expected to have a measurable impact on the traffic incidents over time. Our primary objectives include - Does these initiatives actually help decreasing the number of traffic incidents? Can we identify traffic hot spots (areas with recurring traffic incidents)? In this study, we explored the accidents dataset and applied unsupervised data mining algorithms like *DBSCAN* to identify clusters in the spatial data. For each of the accident record, we tried to identify if there is a vision zero initiative in place by iterating thru each of the vision zero shape files and also identified the hot spots for these incident datasets before and after the vision zero initiative (by taking the centroid of the identified clusters as hot-spots). We have also attempted to conduct time series analysis on the incidents dataset.

This study might also help in finding answers to several questions, such as:

- Who are involved in majority of the traffic violations?
- What are the key contributing factors to the traffic accidents?
- Where are the hot spots? When is the majority of the traffic incidents occur?
- Why are some locations more prone to accidents than others?
- How can we prevent traffic accidents from occurring?

Finally, we believe this research would help in providing valuable insights into the effectiveness of the Vision Zero's intended goal of eliminating/reducing injuries and fatalities from motor vehicle accidents.

Literature Review

We have reviewed several articles, news items, journal papers related to the traffic incidents and factors causing the collisions. The research paper (Yuan, Zhou, and Yang, 2017) analyzed motor vehicle crashes from 2006 to 2013 in the state of Iowa, and tried to predict whether an accident will occur, for each road segment in each hour. The journal article (Chang and Chen, 2005) focused on the data mining of tree-based models to analyze freeway accident frequency, leveraging the CART model and a negative binomial regression model. The research (Eisenberg, 2004) study investigated the relationship between monthly precipitation and monthly fatal crashes, using the negative binomial regression approach. We also reviewed the journal article (Jackson and Sharif, 2005) which focused on the impacts of the rainfall on traffic safety in Texas, by analyzing the relationship between crashes and rainfall. The study (Liling Li, 2017) investigated the fatal accident dataset (FARS) to identify the relationship between fatal rate and other attributes including collision manner, weather, surface condition, light condition, and drunk driver. Association rules were discovered by Apriori algorithm, classification model was built by Naive Bayes classifier, and clusters were formed by simple K-means clustering algorithm. This study identified that the southern region seemed to have higher fatal rates compared to north east. This study indicated that south is much riskier compared to rest regions. North east is the safest region and followings are Midwest and west.

Methodology

Data Exploration

The datasets we used for this analysis are from publicly available data from NYPD motor vehicle collision, which is a breakdown of every collision in NYC by location and injury, and *traffic tickets* data from NY data site. We extracted ALL 14 *vision zero* shape files from NYC vision zero data feeds. As part of the data exploration, we performed *Exploratory data analysis* to better understand the relationships in the given data including feature distributions, correlations and basic summary statistics.

Data Preparation

In data preparation, we fixed the data issues noticed in the exploratory analysis, which involve treating the outliers (if any), missing data, invalid data etc. We divided the full incident records into logical groups of data like pedestrian data - before and after the vision zero initiative, multi vehicle incident data before and after the vision zero. We loaded the vision zero initiative shape files and enriched the incident records with vision zero initiative flags. For example, SPEED ZONE is a vision zero initiative, we loaded the shape file for that initiative, and checked to see if the incident record(s) point (latitude, longitude) fall into the shape, if so, we marked the new feature (say 'SPEED_ZONE') for such incident records as 1, else 0. We repeated this process for ALL the vision zero initiatives and added each of the vision zero initiative as a feature into the dataset.

Finding hotspots & performing time series analysis

In addition to data exploration and investigation of the factors associated with the traffic incidents and understand how effective vision zero initiative is, we also developed an application with varying hyper parameters to visualize the ***hot spots*** using the clustering method DBSCAN for spatial data. This would aid in visualizing and comparing the hot spots *before* and *after* the vision zero initiative, and also individually analyses the hot spots for each of the vision zero initiative based on whether the vision zero initiative was in place or not. The application would also provide the ability to filter further by a NYC borough.

A time series analysis on the incident records was also done, which would help in analyzing the incident patterns, and possible forecasting.

Experimentation and Results

Data Exploration

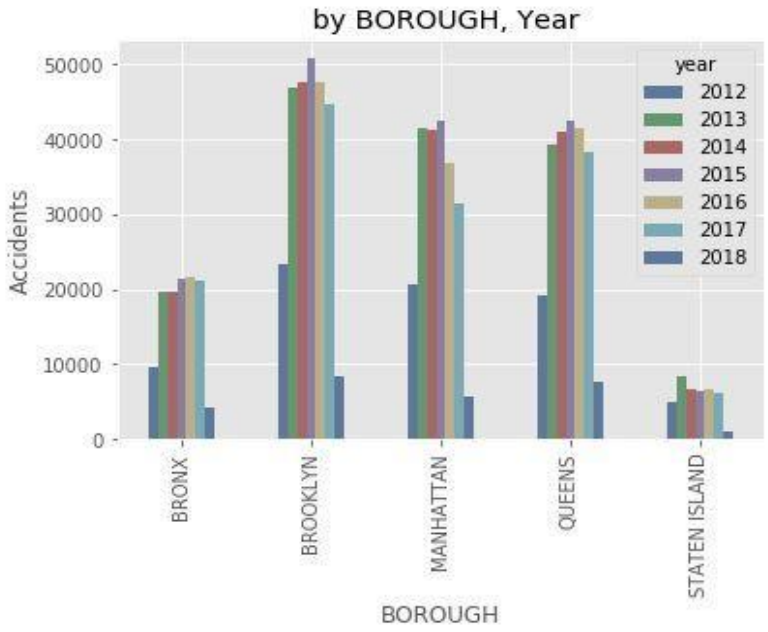
Datasets

The below are the data sets we use to explore the NYC traffic incidents:

NYC Traffic Tickets, NYC Motor Vehicle Collision, Vision Zero

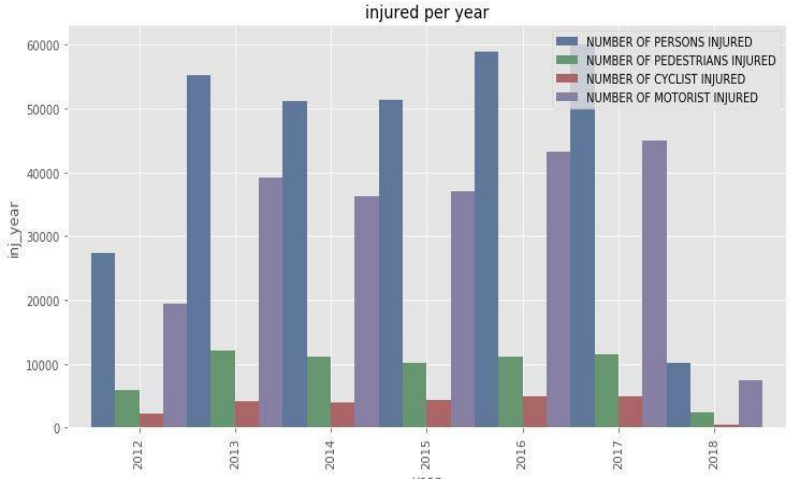
There are 1226374 observations from the collision dataset. The key data elements include - DATE, TIME, BOROUGH, ZIP, LATITUDE, LONGITUDE, ON STREET, CROSS STREET, OFF STREET, INJURY DATA, FATALITY DATA, CONTRIBUTING VEHICLE DATA.

By grouping the accidents by borough and year, we can see the accident counts in each borough of NYC by year, as shown in fig. *Accidents by borough*:



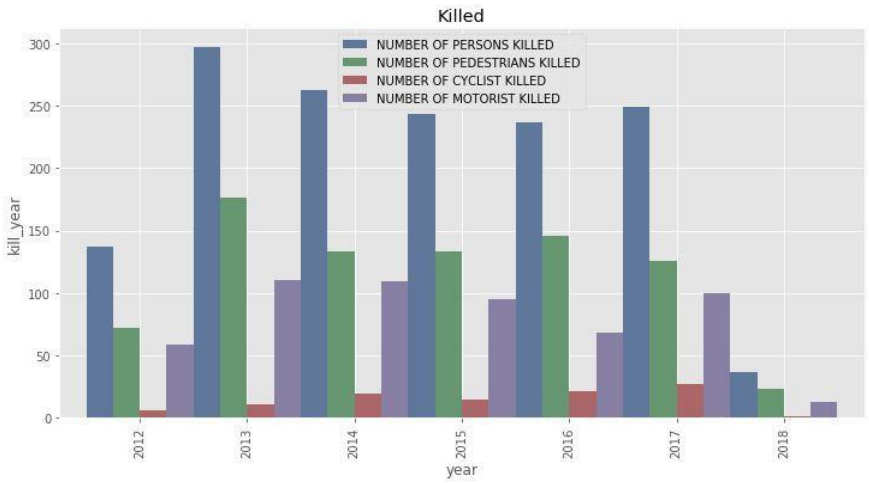
Accidents by borough - NYC Motor Vehicle Collision

Plotted the injuries by year by combining all injuries 'NUMBER OF PERSONS INJURED','NUMBER OF PEDESTRIANS INJURED','NUMBER OF CYCLIST INJURED','NUMBER OF MOTORIST INJURED':



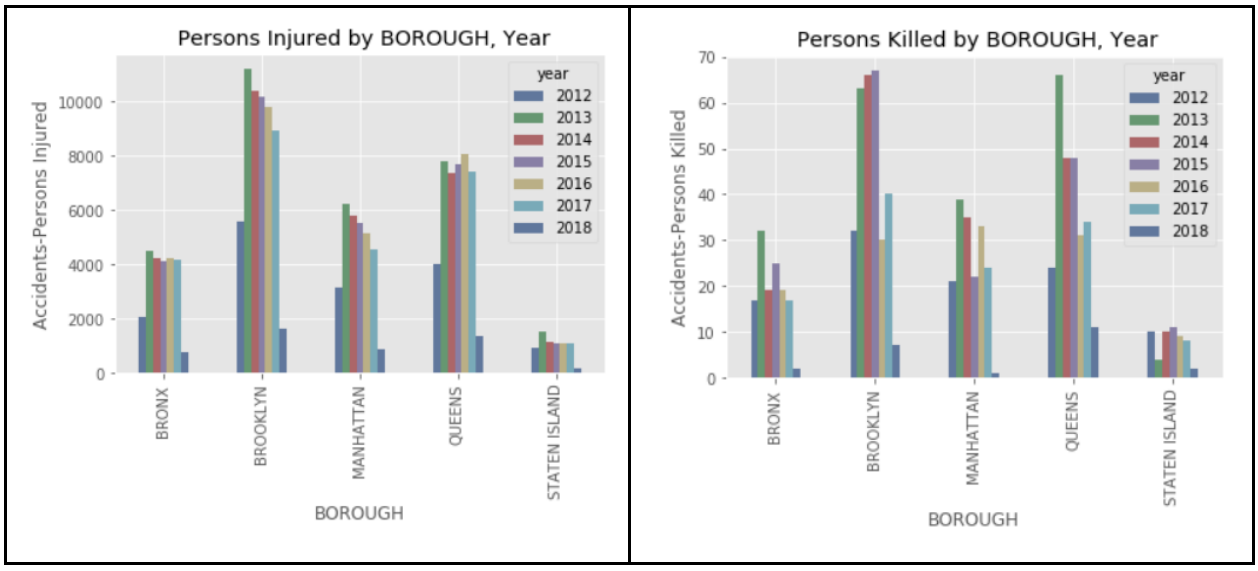
Injuries by Year - NYC Motor Vehicle Collision

The fatalities plot (which include all of 'NUMBER OF PERSONS KILLED', 'NUMBER OF PEDESTRIANS KILLED', 'NUMBER OF CYCLIST KILLED', 'NUMBER OF MOTORIST KILLED') shows the fatalities by year.



Fatalities by Year - NYC Motor Vehicle Collision

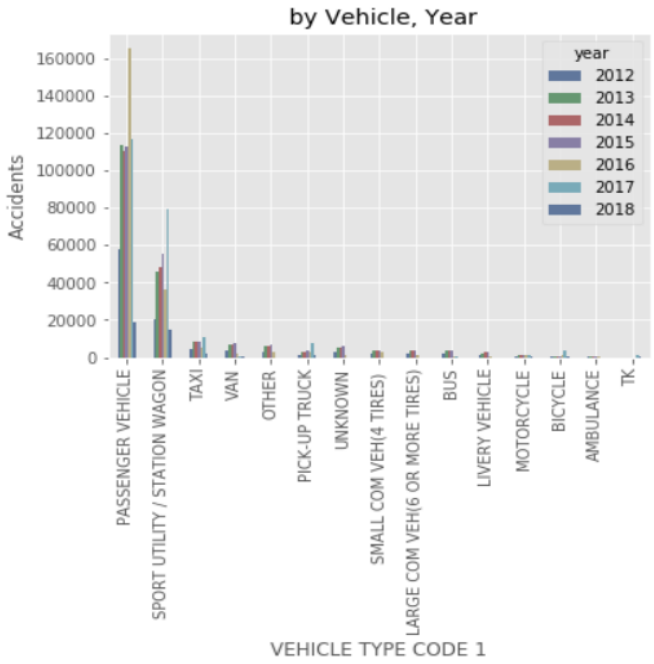
From the above plots, the fatalities appear to be slightly decreased after 2014, however no such indication in case of injuries. So, there is no clear indication if the *vision zero* is the cause for the above reduction in fatalities. We need to further evaluate the same with specific vision zero data feeds.



By grouping Vehicle contributing factors into different categories we were able to find out the major contributing factors for accidents. Human factors such as **Driver Distraction, Improper Driving, violation of laws** are without doubt the major reason for accidents. Road conditions and vehicle defects play only secondary roles.

CONTRIBUTING FACTOR VEHICLE 1 year	Aggressive Driving	Animals Action	DUI	Defective Road	Defective signals	Disability	Driver Distraction	Driver Inexperience	Improper Driving	Mechanical defect	Other Vehicular	Pedestrian/Bicyclist/Other Pedestrian Error/Confusion	Traffic Control Device Improper	Unspecified	violaton of law	weather
2012	362	40	791	835	4	1177	21837	1252	4835	535	3231	12	409	59606	5087	524
2013	553	83	1610	1959	17	2253	45503	2384	10317	1029	6486	25	471	119122	10565	1093
2014	610	105	1911	2716	13	2324	51239	2599	12278	1023	8623	29	573	108551	11780	1182
2015	796	103	1965	3112	19	2627	64105	2752	14628	1167	12494	32	610	97234	13858	1356
2016	561	154	1608	1766	16	931	41985	2256	18417	865	6876	562	496	130293	17403	1393
2017	976	203	2643	2654	18	59	57140	4252	44250	1509	11623	1526	684	57673	40189	2268
2018	150	29	455	949	6	11	9361	537	7382	274	2029	191	158	11559	6787	473

Contributing Factors - NYC Motor Vehicle Collision

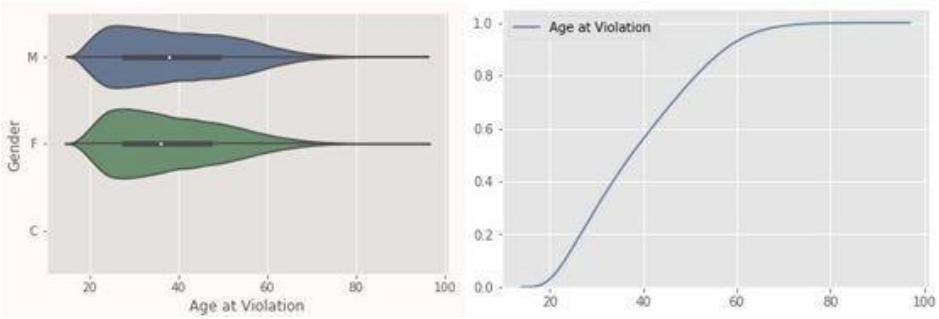


Most of the accidents happened by Passenger Vehicles, Taxies and Pickup trucks also contribute some.

NYC Traffic Violations:

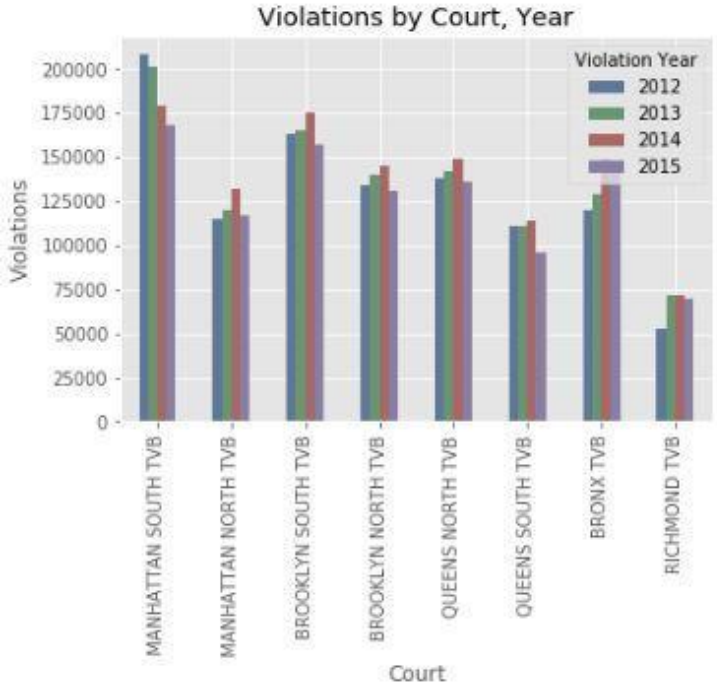
Here, we would like to briefly explore the traffic tickets dataset. NYC Violation data consists of the variables: 'Violation Charged Code', 'Violation Description', 'Violation Year', 'Violation Month', 'Violation Day of Week', 'Age at Violation', 'Gender', 'State of License', 'Police Agency', 'Court', 'Source'. There are approximately half million violation records during the years 2012-16.

By looking at what age group may be causing more violations, it's evident that over 50% of the violations happened in the age group **below 35 years**.



violation age - NYC Traffic violations

Grouping violations by Courts in NYC from traffic tickets data, we see that there is a slight reduction of violations post vision zero initiative across all boroughs. As you can see here, Manhattan has the most traffic violations.



violations by borough court - NYC Traffic violations

Also, one more thing we noticed from the above EDA is, although Manhattan had the most traffic violations, Brooklyn and Queens had more injuries.

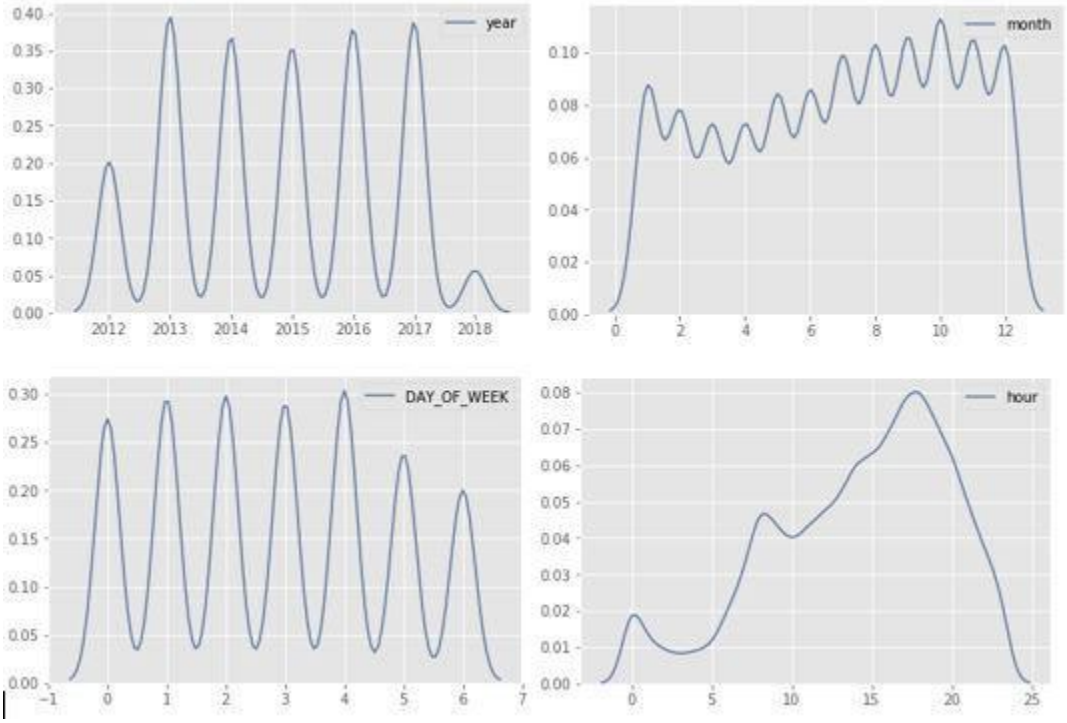
Vision Zero EDA - Pedestrian Incidents

As part of our *vision zero data feeds* analysis, we had a closer look at some specific areas and investigated if there is a change after a few specific vision zero initiatives.

Pedestrian and bike Incident Analysis:

Since pedestrians account for over 56% of all NYC traffic fatalities, let's take a closer look at these incidents. We have extracted the pedestrians and cyclist's injuries and fatalities from collision data and treated that as our pedestrian incident dataset, during the years 2012-2018.

Pedestrian Incidents



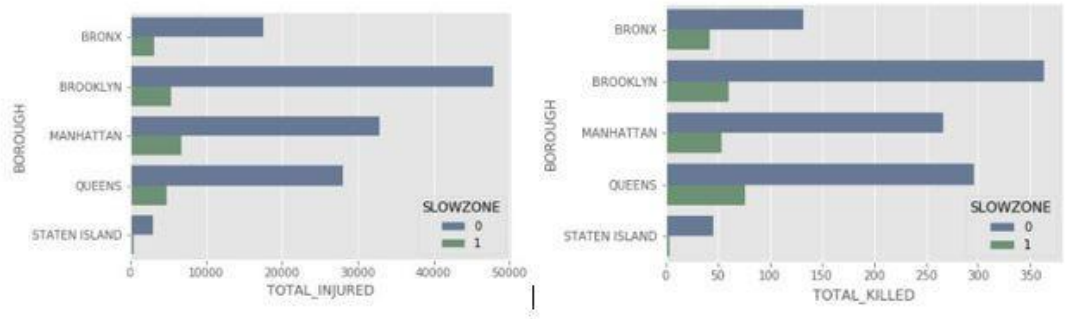
Looking at the pedestrian incidents plot, it's not showing a decrease in the incidents over the years. Also, most incidents are happening towards the later part of the day (**between 3 to 4 PM**), and slightly more chance on **Fridays**. When we looked at overall incidents across all months over these years, **October** had majority of incidents.

We tried gathering the specific vision zero initiative (for example, *arterial slow zones, Speed humps etc.*) and analyzed if there is any pattern before and after the initiative.

Arterial Slow Zone program:

The *Arterial Slow Zone program* uses a combination of a lower speed limit, signal timing changes, distinctive signs and increased enforcement to improve safety on some of New York City's most high-crash corridors. The slow zone geojson data is available here. We loaded the slow zone geojson data and added a feature *SLOW ZONE* to our pedestrian

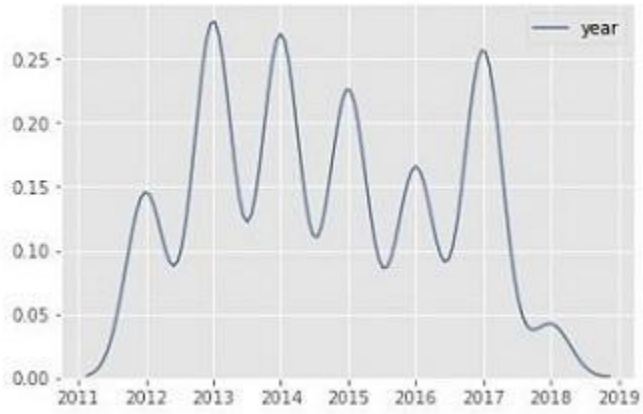
dataset by comparing if the incident point from the pedestrian dataset falls near the arterial slow zone or not. If the incident happened near the *SLOW ZONE*, then we marked it as 1, otherwise 0.



Slow zone Vs Non Slow zone - Pedestrian Incidents

There is a clear reduction in incidents in slow zones compared to non-slow zone areas.

Taking only slow zone areas and plotting incidents year over year produced mixed results, there was a significant reduction immediately following the vision zero initiative, however, the number of incidents shot up again in the year 2017!

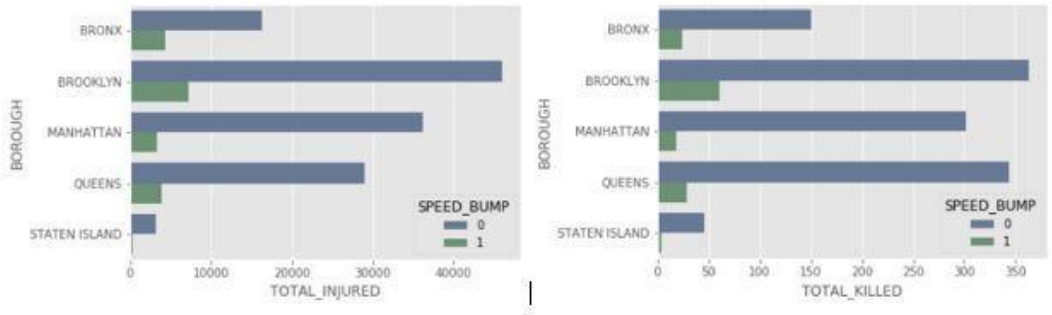


Slow zone - Yearly Pedestrian Incidents

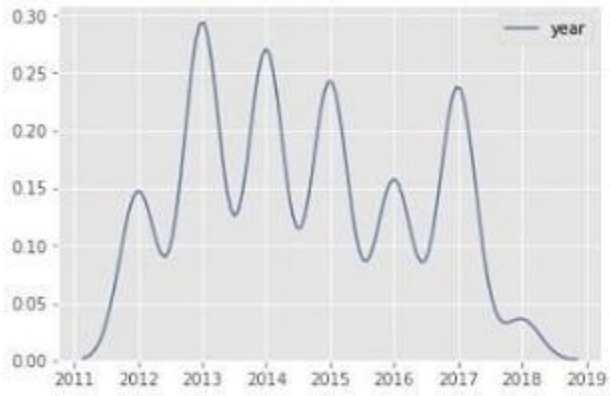
(Kernel Density Estimate Graph)

Speed Humps:

Speed Humps are a raised area of a roadway designed to reduce vehicle speeds. We loaded the speed hump geojson file and tried to mark each of the pedestrian incident with *SPEED_HUMP* as 1 (which represents the presence of speed hump) and 0 (absence of speed hump) by comparing the distance between the incident point from the speed hump location.



Taking only SPEED HUMP locations and plotting pedestrian incidents year over year produced mixed results (very similar to what we have seen with *slow zone* analysis), there was a significant reduction immediately following the vision zero initiative, followed by increased incidents in the year 2017!



SPEED_HUMP - Yearly Pedestrian Incidents

(Kernel Density Estimate Graph)

As part of extending this exploratory research in future, we would plan on joining this incident dataset with weather data to see if the weather has caused the increased incidents in the year 2017.

Augment dataset with ALL vision zero features:

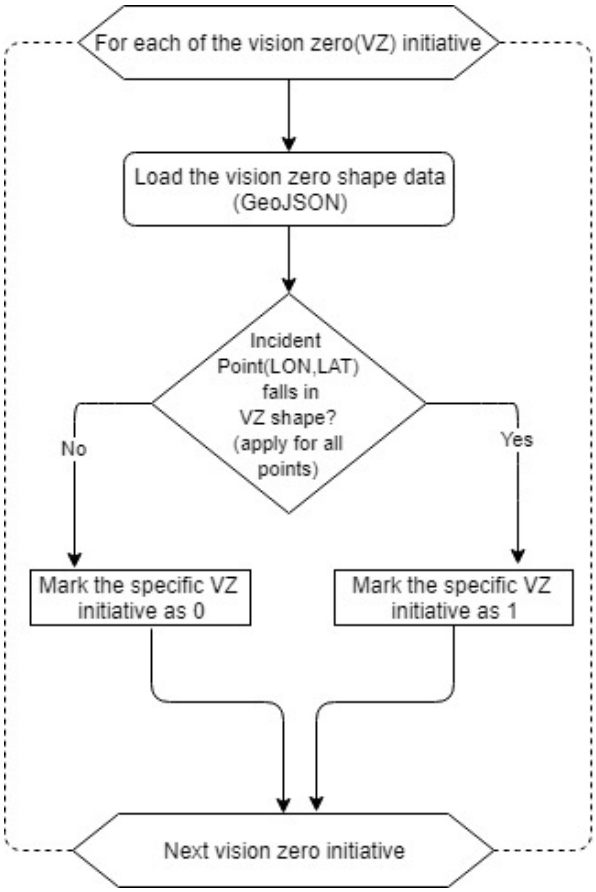
As part of the data preparation, we enriched the incident dataset with the vision zero initiatives. We added all 14 visions zero initiatives as 14 Boolean features each of which indicating if a given vision zero initiative is in place or not.

Vision Zero Feature	**Description**

SLOWZONE	The Arterial Slow Zone program uses a combination of a lower speed limit, signal timing changes, distinctive signs and increased enforcement to improve safety on some of New York City's most high-crash corridors.
SPEED_HUMP	Speed Humps are a raised area of a roadway designed to reduce vehicle speeds
SIGNAL_TIMING	Priority Corridors where the signal progression has been changed to match the 25 MPH speed limit.
BIKE_PRIORITY	Priority Bicycle Districts are neighborhoods with comparatively high numbers of cyclist KSI and few dedicated bicycle facilities. These districts, seven in Brooklyn and three in Queens, represent 14% of the City's bicycle lane network and 23% of cyclist KSI. NYC DOT identified these areas in the 2017 report Safer Cycling: Bicycle Ridership and Safety in New York City. The agency has prioritized these areas for bicycle network expansion.
ENHANCED_CROSSING	Enhanced Crossings are marked high-visibility crosswalks on calm streets with low vehicle volumes and a strong pedestrian desire to cross. Standard DOT toolbox treatments are used (ADA pedestrian ramps, pedestrian warning signs and high-visibility crosswalk markings) to improve the mobility and accessibility of pedestrians.
LEADING_PEDESTRIAN_INTERVAL	Intersections where DOT installs signals that show a walk sign for pedestrians before showing a green light to vehicle traffic. The goal of these signals is to improve street safety by giving pedestrians a chance to establish their presence in the crosswalk before vehicles make turns across that crosswalk.
LEFT_TURN_TRAFFIC_CALMING	Intersections where DOT installs traffic calming measures that guide drivers to turn left at a safer speed and angle, as well as increase visibility for pedestrians in the crosswalk
NEIGHBORHOOD_SLOW_ZONE	The Neighborhood Slow Zone program is an application-based program which takes a

	neighborhood area and reduces the speed limit to 20 mph. Areas are chosen based on crashes, presence of schools and other neighborhood amenities, and community support. The treatments include a mixture of markings, signage, and speed humps.
SAFE_STREETS_FOR_SENIORS	The Safe Streets for Seniors program is an initiative aimed at increasing safety for older New Yorkers. Based on factors such as senior population density, injury crashes, and senior trip generators, DOT has selected and studied Senior Pedestrian Focus Areas. Within these areas, DOT evaluates potential safety improvements and also conducts educational outreach to senior centers.
SIP_INTERSECTIONS SIP_CORRIDORS	Safety-oriented engineering improvements that use multiple treatments (signals, markings, concrete etc.) on both corridors and intersections. Improvements are generally aimed at better organizing traffic, improving travel times, creating shorter, safer pedestrian crossings, and safe routes for bicycle travel.
PRIORITY_CORRIDORS	All corridors in each borough were ranked on a pedestrian KSI (killed and severely injured) per mile basis. Corridors were selected from the top of this list until the cumulative number of KSI reached half of the borough's total. Developed as part of the Borough Pedestrian Safety Action Plans.
PRIORITY_INTERSECTIONS PRIORITY_ZONES	The intersections, zones with the highest number of pedestrian KSI (killed and severely injured) that cumulatively account for 15% of the borough's total pedestrian KSI. Developed as part of the Borough Pedestrian Safety Action Plans.

Below flow chart depicts the high-level steps in identifying whether the vision zero initiative is in place for a given incident's point (longitude, latitude) or not. The steps include: for each of the vision zero initiative, load the GeoJSON (shapefile) from the vision zero website, and for every incident record check to see if the incident location falls in the any of the vision zero shape/coordinates, if so, mark the vision zero initiative for that incident record as *one* else *zero*.



Model Development - Identifying Hotspots in Incident Records:

In the past, not having access to enough spatial data was a major hurdle in answering pressing research questions and/or build more compelling visualizations. However, today the problem is indeed the other way – we have too much data! Too many scattered points on a map can be overwhelming, and processor intensive even to display those many data points on a device. We should be able to leverage the density-based clustering algorithms to compress the spatially redundant data points into a set of representative features.

In this study, we take the collision dataset, and try to apply the clustering algorithms to identify few hot spots.

K Means:

The k-means algorithm divides a set of N samples X into K disjoint clusters C, each described by the mean μ_j of the samples in the cluster. The means are commonly called the cluster "centroids"; note that they are not, in general, points from X, although they live in the same space. The K-means algorithm aims to choose centroids that minimize the inertia, or within-cluster sum of squared criterion:

$$\sum_{i=0}^{n} \quad min(x_j - \mu_j)^2$$

The K Means algorithm minimizes variance, **not the geodetic distance**, so, this may not be an ideal algorithm for latitude-longitude spatial data, which needs geodetic distance to be minimized. Up on further research we found about DBSCAN algorithm (Ester et al. 1996) which works better with arbitrary distances and non-spherical shapes.

DBSCAN (Density-Based Spatial Clustering and Application with Noise):

The DBSCAN algorithm views clusters as areas of high density separated by areas of low density. Due to this rather generic view, clusters found by DBSCAN can be any shape, as opposed to k-means which assumes that clusters are convex shaped. DBSCAN clusters a spatial data set based on two parameters: a physical distance from each point, and a minimum cluster size. So, this works better for non-flat geometry, uneven cluster sizes. The geometry metric used here is the '*Distance between nearest points*'. Hence, this method appears to work better for our *collision data set* with spatial latitude-longitude data:

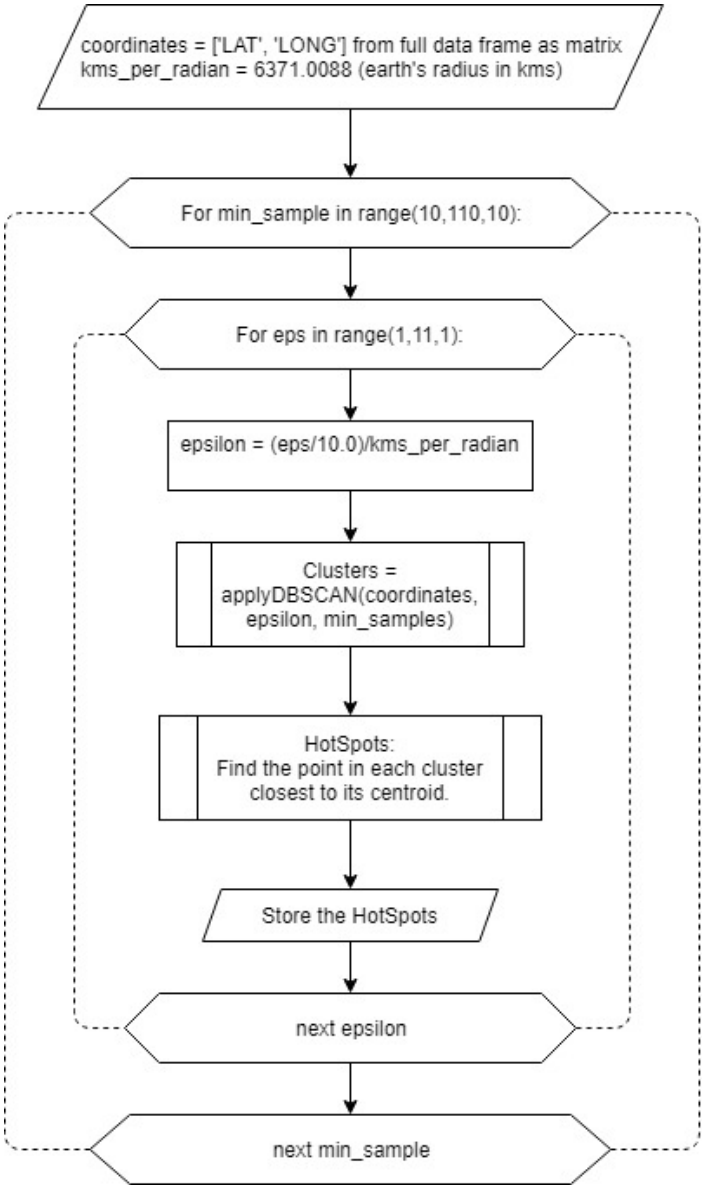
DBSCAN has two parameters, epsilon and minPoints. Epsilon essentially decides the size of the neighborhood. The way DBSCAN works is it randomly picks one point, by looking at the value of the epsilon, it defines the neighborhood size. It assigns all the neighborhood points to the same cluster and repeats the process as long as there are points in the neighborhood defined by epsilon. When it run out of the points in the neighborhood, it jumps to another randomly selected point which has not yet been assigned to any group and repeats the process. If any point which does not have at least "minPoints" in the neighborhood then those are considered to be outliers. They do not become part of any clusters. In this approach we do not tell explicitly how many clusters to make, it will find out the natural clusters in the data. The below is high level steps in the DBSCAN algorithm:

For each $o \in D$ *do*
    If o is not yet classified then
        If o is a core-point then
            Collect all objects density-reachable from o
            And assign them to a new cluster
        Else
            Assign o to NOISE

So, DBSCAN is a human intuitive clustering method - clusters are dense regions in the data space, separated by regions of lower density of points. The DBSCAN algorithm is based on this intuitive notion of 'clusters' and "noise". The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.

To find the hotspots, we used python sklearn package's DBSCAN module. The *epsilon* parameter is the max distance (in km here) that points can be from each other to be considered a cluster. The *min_samples* parameter is the minimum cluster size (everything else gets classified as noise). We have used the *haversine* metric and *ball tree* algorithm to

calculate the great circle distance between points. We had to convert the epsilon and coordinates to *radians*, because scikit-learn's haversine metric needs radian units.



Also, unlike k-means, DBSCAN doesn't require us to specify the number of clusters in advance - it determines them automatically based on the epsilon and min_samples parameters. In the above, we start with the coordinates obtained from the incident records. We took min samples from 10 to 100 (in increments of 10), and the epsilon starting from 0.1 km to 1 km (in increments of 0.1 km) and found the clusters for each iteration. And within each iteration we identified the hotspot by taking the point closest to the centroid in the cluster.

Below is a sample of hotspots located from pedestrian incidents (with epsilon as 0.5 km, and the min_samples size as 50):

Please refer to this app we built as part of this study (click on URL) , for more details & experimentation with varying params: https://data698ssv.shinyapps.io/HotSpotViewer/

The below map shows the hot spots of pedestrian data including ALL data (before and after the VZ initiative, with any of the VZ initiative is in place today):

Cluster analysis is an exploratory method at its core, so review the hotspots with varying *epsilon* and *min_samples* above, using the application..

The app also provides the ability to pick different incident categories, and timeline (before and after vision zero initiative), borough, presence of any/specific vision zero initiative, and the hyper parameters - *epsilon* and the *min points* of DBSCAN.

Model Development - Time Series Analysis of Incident Records:

Time series analysis is a statistical technique that deals with time series data. In Time series data, data is in a series of particular time periods or intervals. There are a few studies that were applied time series technique in modeling and predicting road accidents.

Accidents typically are considered random events, but there are a lot of efforts from researchers to fit different types of mathematical distributions (ranging from stochastic to deterministic) to accident data to predict models.

Time series analysis postulates that future values have a probability distribution that is conditioned by a knowledge of past values; therefore, exact predictions are impossible. Moreover, the reliability of prediction values depends on the characteristics of data.

Make Time Series Stationary

A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time. Most statistical forecasting methods are based on the assumption that the time series can be rendered approximately stationary through the use of mathematical transformations. A stationarized series is relatively easy to predict:

Autoregressive Integrated Moving Average Method (ARIMA)

A commonly used time series model is the Autoregressive Integrated Moving Average (ARIMA) Method, which provides analysts with a powerful tool for describing stationary and nonstationary processes. A random variable that is a time series is stationary if its statistical properties are all constant over time. *A stationary series has no trend, its variations around its mean have a constant amplitude.*

The ARIMA forecasting equation for a stationary time series is a *linear* (i.e., regression type) equation in which the predictors consist of *lags of the dependent* variable and/or *lags of the forecast errors*. That is:

Predicted value of Y = a constant and/or a weighted sum of one or more recent values of Y and/or a weighted sum of one or more recent values of the errors.

If the predictors consist only of lagged values of Y, it is a pure autoregressive model, which is just a special case of a regression model. If some of the predictors are lags of the errors, an ARIMA model is NOT a linear regression model, because there is no way to specify "last period's error" as an independent variable: the errors must be computed on a period-to-period basis when the model is fitted to the data. The problem using lagged errors as predictors is that *the model's predictions are not linear functions of the coefficients*, even though they are linear functions of the past data. So, coefficients in ARIMA models that include lagged errors must be estimated by *nonlinear* optimization methods.

Lags of the stationarized series in the forecasting equation are called "autoregressive" terms, lags of the forecast errors are called "moving average" terms, and a time series which needs to be differenced to be made stationary is said to be an "integrated" version of a stationary series.

ARIMA(p,d,q) model, where:

- **p** is the number of autoregressive terms,
- **d** is the number of non-seasonal differences needed for stationarity, and

- **q** is the number of lagged forecast errors in the prediction equation.

Let y denote the d^{th} difference of Y, which means:

If d=0: $y_t = Y_t$

If d=1: $y_t = Y_t - Y_{t-1}$

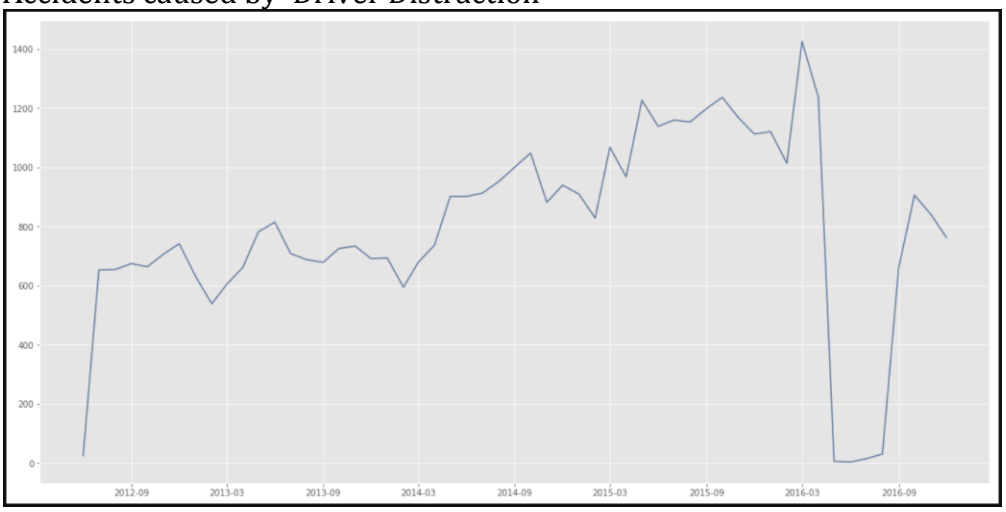
If d=2: $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$ (discrete analog of a second derivative)

ACF (autocorrelation function) and PACF (partial autocorrelation) plots are used to find the AR and MA terms.

ACF is a bar chart of the coefficients of correlation between a time series and lags of itself. PACF plot is a plot of the *partial* correlation coefficients between the series and lags of itself.

In our project we analyzed NYC Traffic accident sequential observations and developed models to fit data, which can be used to produce forecasts of time series that might be expected under various scenarios.

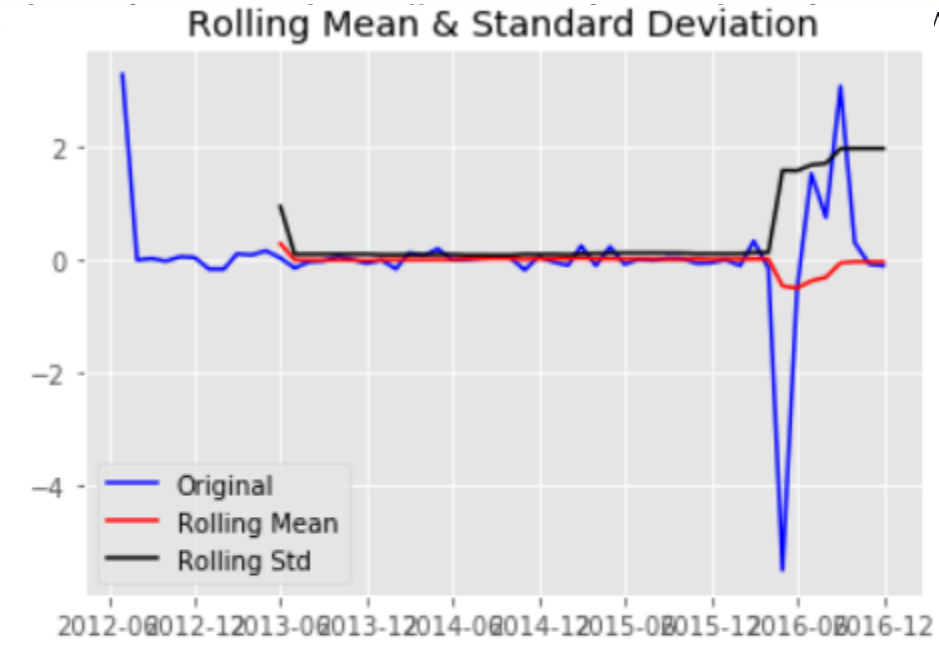
Brooklyn has the highest number of accidents. So, we subset our dataset for Brooklyn and Accidents caused by 'Driver Distraction'



Check whether the Time series is Stationary

Here we use Dickey-Fuller Test to see whether the TS is Stationary or not: The test results comprise of a Test Statistic and some Critical Values for different confidence levels. If the 'Test Statistic' is less than the 'Critical Value', then the series is stationary.

we get

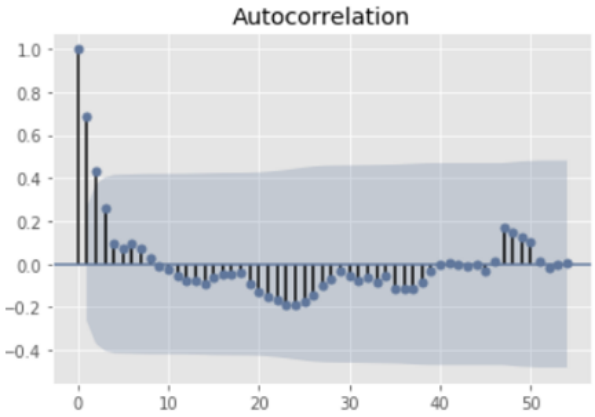



```
Results of Dickey-Fuller Test:
Test Statistic                  -4.242286
p-value                          0.000558
#Lags Used                       5.000000
Number of Observations Used     48.000000
Critical Value (1%)             -3.574589
Critical Value (5%)             -2.923954
Critical Value (10%)            -2.600039
dtype: float64
```

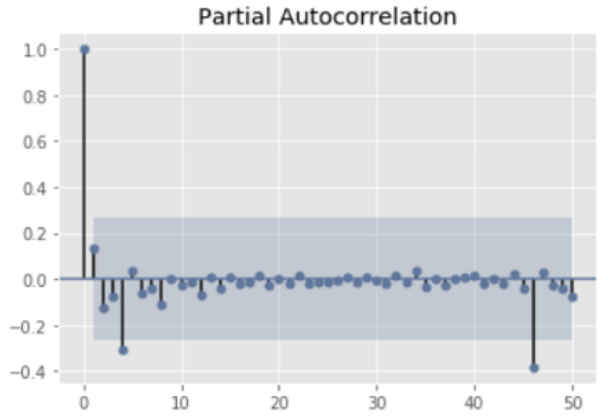
In this case we see that the mean and variance have small variations with time. Also, the Dickey-Fuller Test Statistic is less than the 1% critical value, thus the TS is stationary with 99% confidence.

Parameters p,d,q can be found using **ACF** and **PACF** plots.

ACF

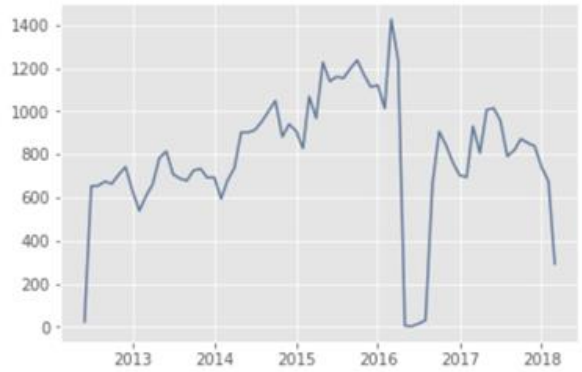


PCAF

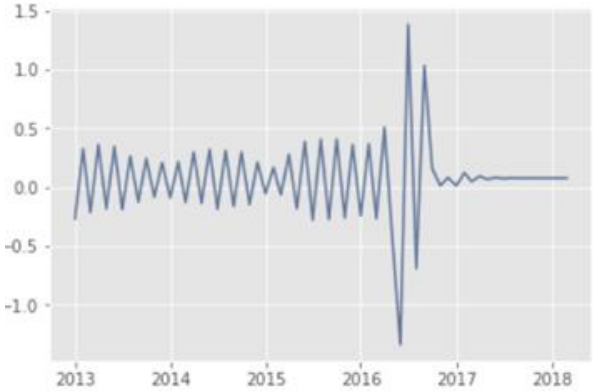


The p,d,q values can be specified using the order argument of ARIMA which take a tuple (2,0,1).

Original



Forecast



Our model shows that there is a gradual decrease in accidents due to DISTRACTED DRIVING from 2017 onwards. But in reality, there is no trend. We will continue our study with other Accident Factors also.

Conclusion

Overall, we noticed mixed results from this study. Although fatalities appeared to be slightly reduced after the year 2014, there is no such clear indication in case of injuries. Human factors such as Driver Distraction (texting or emailing while driving etc.), Improper Driving and violation of laws (speeding) are without doubt the major reasons for the accidents. And over 50% of the violations are caused by younger drivers. Most of the pedestrian incidents appears to be occurring between 3 pm to 4 pm with a slightly more chance of occurring on Fridays. October month shown majority of the pedestrian incidents. Though Manhattan had the most traffic violations, Brooklyn and Queens had more major accidents that caused injuries and/or deaths. The exploratory study also revealed that Passenger vehicles were most likely to be involved in the accidents that led to injuries or deaths. We noticed that there was an immediate decrease in pedestrian incidents in the years following VZ initiative, however the incidents surged in the year 2017.

The hot spot analysis with pedestrian data indicated 7 hotspot regions (2 in the areas where some sort of VZ initiative in place, and 5 in the areas where there is no VZ initiative in place - all 5 are in Queens borough). Prior to the VZ initiative timeline, there were 17 hotspots! In case of multi vehicle (3 vehicle) incidents, system resulted in 11 hotspots (6 in the areas with some kind of VZ initiative in place and 5 in the areas with no VZ initiative in place). Strangely prior to the VZ initiative system only indicated 4 hotspots for multi vehicle accidents compared to 11 after VZ initiative. Please refer to the appendix section for the detailed hotspot maps showing before and after vision zero initiative.

Time series analysis done with the Brooklyn accidents caused by 'Driver Distraction' data forecasted a gradual decrease in the incidents starting 2017. We will continue our study with other Accident Factors also.

Future Work

- Include more incident data sets for the clustering & time series analysis.
- Combine incident records with weather data to study the impact of weather in fatal accidents. This study might also help in determining the reason for sudden surge of incidents in the year 2017, and help in figuring out what played vital role in these accidents - Human factors or environmental factors?

References

Chang, L. and W. Chen. “Data mining of tree-based models to analyze freeway accident frequency”. In: Science Direct (2005). URL: https://www.sciencedirect.com/science/article/pii/S0022437505000708.

Eisenberg, D. The mixed effects of precipitation on traffic crashes. 2004. URL: https://www.ncbi.nlm.nih.gov/pubmed/15094418.

Jackson, T. and H. Sharif. “Rainfall impacts on traffic safety: rain-related fatal crashes in Texas”. In: TandF (2005). URL: https://www.tandfonline.com/doi/full/10.1080/19475705.2014.984246?src=recsys.

Yuan, Z., X. Zhou and T. Yang. Predicting Traffic Accidents Through Heterogeneous Urban Data. 2017. URL: http://urbcomp.ist.psu.edu/2017/papers/Predicting.pdf.

Geoff, Boeing. “Clustering to Reduce Spatial Data Set Size”. In: SocArXiv Papers (2014). URL: https://osf.io/preprints/socarxiv/nzhdc/

Martin Ester,Hans-Peter K, Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. URL: https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

Davoud Moulavi, Pablo A. Jaskowiak, Ricardo J. G. B. Campello, Arthur Zimek, Jorg Sander. Density-Based Clustering Validation. URL: http://www.dbs.ifi.lmu.de/~zimek/publications/SDM2014/DBCV.pdf

Liling Li, Sharad Shrestha, Gongzhu Hu, Analysis of road traffic fatal accidents using data mining techniques, URL: https://ieeexplore.ieee.org/document/7965753/

Documentation: Shiny App Web Application Framework, URL: https://www.rdocumentation.org/packages/shiny/versions/1.0.5

Article - ARIMA models for time series forecasting https://people.duke.edu/~rnau/411arim.htm#pdq

Appendix

Team Github Location:

https://github.com/psumank/DATA698

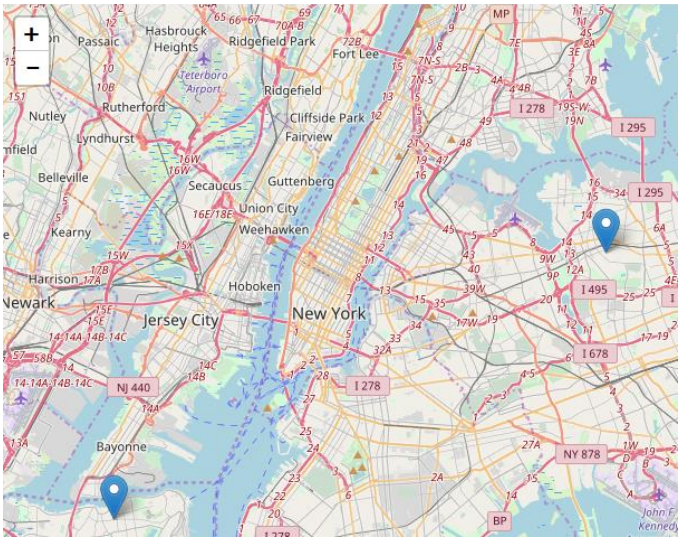
Apps Deployed:

HotSpot Viewer: https://data698ssv.shinyapps.io/HotSpotViewer/

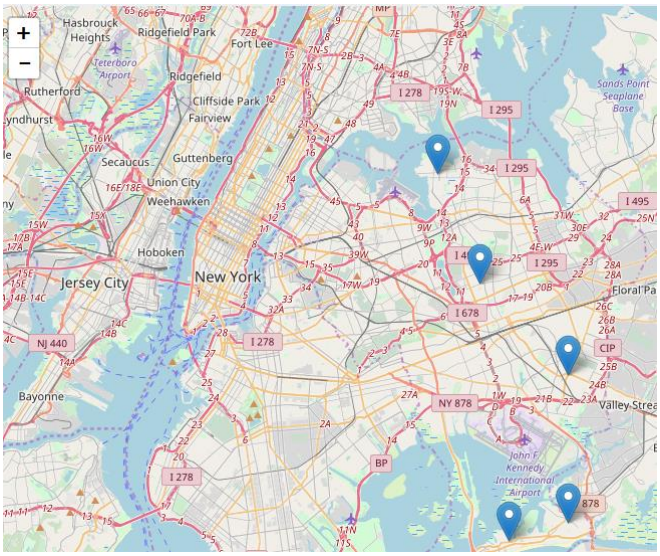
Pedestrian Incidents

Before Vision Zero - Pedestrian Hot Spots: [epsilon: 0.5 , minPoints:50]

After Vision Zero - Pedestrian Hot Spots: [epsilon: 0.5 , minPoints:50]

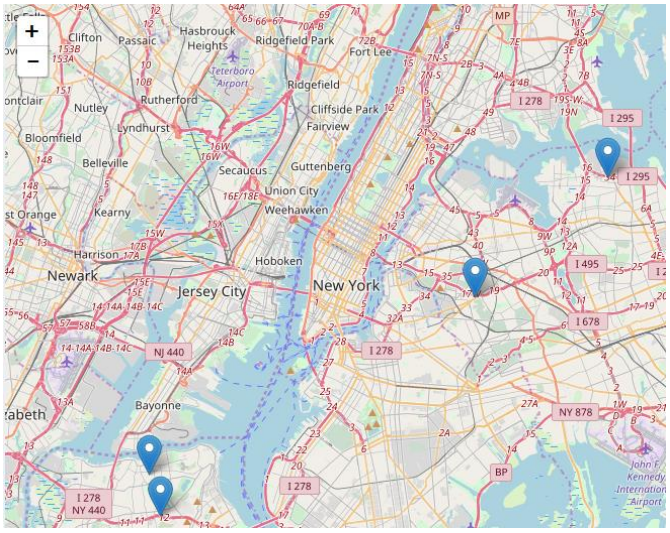


After Vision Zero - Pedestrian Hot Spots: [epsilon: 0.5 , minPoints:50]
(With NO existing VZ initiative in place)

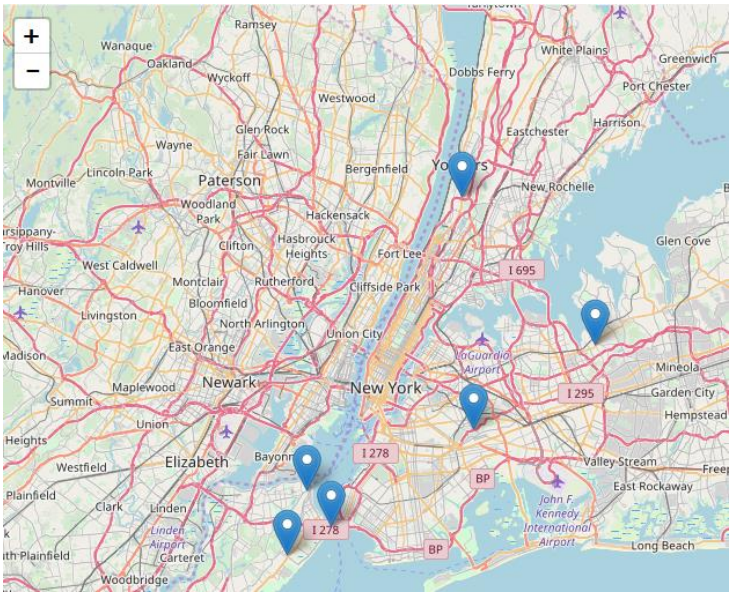


Three Vehicle Incidents

Before Vision Zero - Multi Vehicle Accident Hot Spots: [epsilon: 0.5, minPoints:50]



After Vision Zero - Multi Vehicle Accident Hot Spots: [epsilon: 0.5 , minPoints:50] (With an existing VZ initiative in place)



After Vision Zero - Multi Vehicle Accident Hot Spots: [epsilon: 0.5, minPoints:50] (With NO existing VZ initiative in place)