Causality & Mitigations - Analysis of NYC accidents

SIADS 591 Milestone I Project

Amit Jha & Sahil Pujari

September 18, 2021

Note: For best experience, please view this report in ArcGIS StoryMap - https://storymaps.arcgis.com/stories/bd7e45b6ae874c45a3642ca2632a8a19

Motivation:

Approximately 1.3 million people die each year on the world's roads, and between 20 and 50 million sustain non-fatal injuries [1]. Road traffic crashes cost most countries 3% of their gross domestic product. Road traffic injuries are the leading cause of death for children and young adults aged 5-29 years [2].

We come from India and have first hand experience of frequent road accidents and related fatalities. In 2019 alone there were 449 thousand crashes causing 151 thousand deaths and 451 thousand injuries [3].

That means, a crash is happening every one minute causing a death every three minute in India. This is huge.

Geo tagged data, needed for analysis is not yet available for India and hence we decided to analyze the traffic crashes for New York.

With the power of data science, we want to be able to study factors that contribute towards accidents in the city of New York and suggest strategies that can be taken to reduce them. The concepts taught in the MADS program so far in our curriculum can be applied well towards the datasets we have found for this task.

Data Sources:

We have used public datasets for this analysis. A brief overview of the datasets and the variables contained is provided below:

1. **NYC Open Data** | Motor Vehicle Collision: The data source contains of 3 datasets - crashes (crash info detail and factors contributing to the collision), vehicles (fields related to vehicles involved in the collision), and persons (fields related to people involved in the collision). The data is obtained by records maintained by New York State for accidents reported between the period of 2015 - 2020.
2. **Traffic Speed Data:** When we started we selected dataset from the common source - NYC Open Data, but provided dataset through this was of 22 GB and contained 52 mn records as of 15-Aug-21, hence was difficult to process. Fortunately we found another source - NYC REAL TIME TRAFFIC SPEED DATA FEED(ARCHIVED) for the same data provided by data.βetaNYC.

   This dataset provides monthly compiled data through different links on the site. For reproducibility, we decided to do webscrap the site to download data for five years programmatically.

   This dataset contains 'real-time' traffic information on five minute intervals from locations where DOT(Department of Transportation) picks up sensor feeds within the five boroughs, mostly on major arterials and highways. There is one meta location dataset provided with traffic speed dataset to be able to combine the traffic speed data with the linkinfo(polyline road segment) location data.
3. **Hospitals Locations:** This is a shapefile dataset, provided by Homeland Infrastructure Foundation-Level Data(HIFLD) which provides the location details for the hospitals available in the USA. This dataset is used to find if a crash incident was close enough to a hospital to get the required treatment in time.
4. **Traffic Volume:** This dataset provides traffic volume counts collected by DOT for different roadways at different times of the day.
5. **Population:** Have used two population datasets - first provides the USA population for different years. And the second dataset provides USA population by ZIP code provided by 2010 US decennial Census accessed via Kaggle.
6. **Driving Cost:** This is the most interesting dataset used in the project and was not available anywhere in the compiled form. So we created it ourselves by compiling the data from a global news and lifecycle publication INSIDER.

7. **Traffic Related Death Rate by Countries:** This dataset provides traffic related death rate for various countries sourced from the WHO - Global Status Report on Road Safety 2018.

Data Manipulation Methods:

Traffic Volume Dataset: While this dataset provides traffic volume at different times of the day, unfortunately it does not provide it for all the days. So we decided to find average traffic volume per day for each road by using pandas group by and computed mean over grouped feature to see any correlation between traffic accidents and different roadways. Also as traffic volume is high compared to crashes data, we computed the log of the volume to better visualize the relation. Further as volume would differ for different road types like expressways/highways, we created a field to differentiate road type - expressway/highway etc

Traffic Collision - Crashes: This is one of the most important dataset for the project. We started by printing dataset info by using pandas info() method to see data type and missing values data. And we saw pandas were not able to assign efficient data types to fields. Also interestingly coz of data size being huge 1.8 mn records it doesn't show missing values info. Hence we wrote our custom utility to print missing values percentage for the dataset.

As pandas is not able to assign efficient data types, we assigned appropriate data types for the columns through the preprocessing utility we wrote for this dataset.

Traffic Speed dataset: Because of its huge size we decided to not combine monthly traffic speed data into one file, and instead created yearly files for five years from 2015-19 and used them as and when required for analysis. Also we assigned appropriate data types for fields to avoid running into memory issues.

Analysis and Visualizations:

In this project we wanted to answer several questions related to traffic crashes/fatalities. We expand upon all of these in our notebook but for the purpose of size limitations of the report, we focus on the ones highlighted in bold -

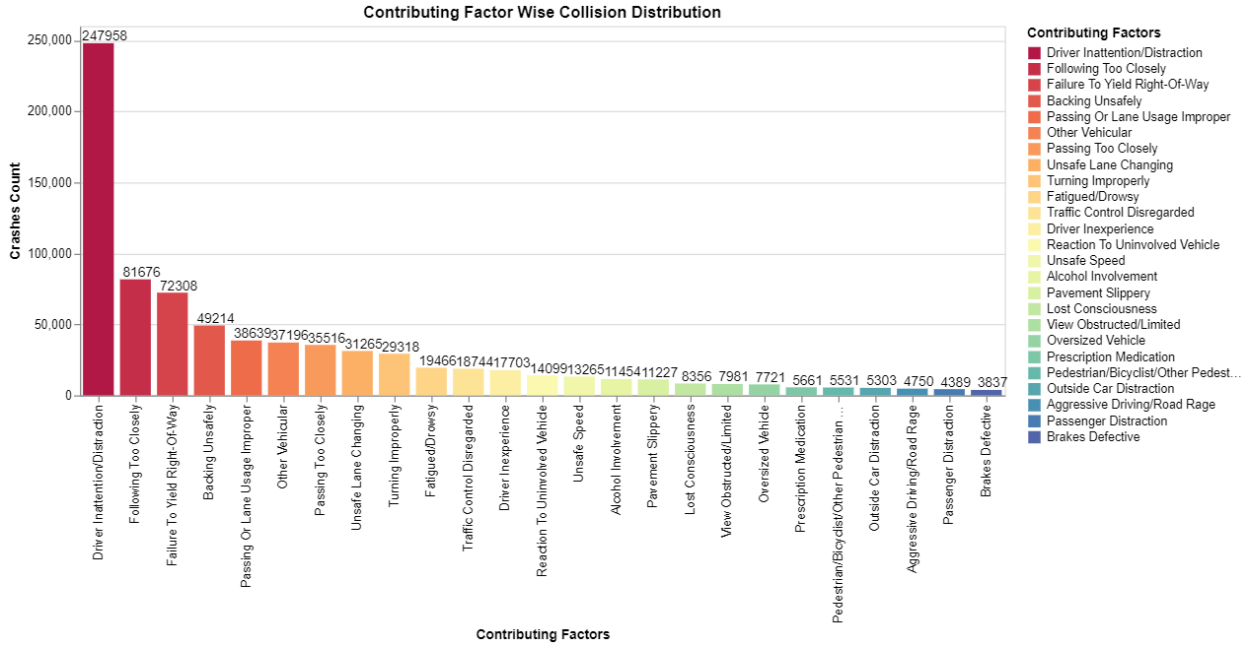
1. **Most common causes of traffic crashes in New York?**
2. **How does driving skill (proxied by license cost) affect crashes and their injuries / deaths?**
3. **Traffic count / time relation with accidents for each zip code**
4. **Traffic speed / time relation with accidents for each zip code**
5. **Could lives be saved by quicker medical attention, i.e. hospitals closer to accident sites? Are there any locations, which do not have hospitals nearby enough?**
6. Are there any locations which are more prone to traffic accidents than others?
7. Are there any sections of society which are more prone to accidents, causal/victims like pedestrians, cyclists or bike users in specific locations?
8. Attributes of vehicles with most accidents

Most Common Causes of Traffic Crashes:

To be able to prevent crashes it is important to first understand the most common causes for crashes.

To visualize this we used Altair library, which provides declarative visualizations in python. It is built on top of the Vega-Lite high-level grammar for interactive graphics.

From this we can clearly see that out of the top 10 causes, the top nine are driver's mistakes. This gave us the insight that perhaps driving skill is the most important aspect that needs to be improved to prevent high number of crashes in NYC.



Recommendation: Apart from the #1 factor driver inattention, rest of the factors can be improved by defensive driving education in the public and are directly related with techniques that this driving education entails.

How does driving skill (proxied by license cost) affect crashes:

We tried to find data for training skill but could not find it and so decided to use cost for getting a driving license as a proxy for training skill.

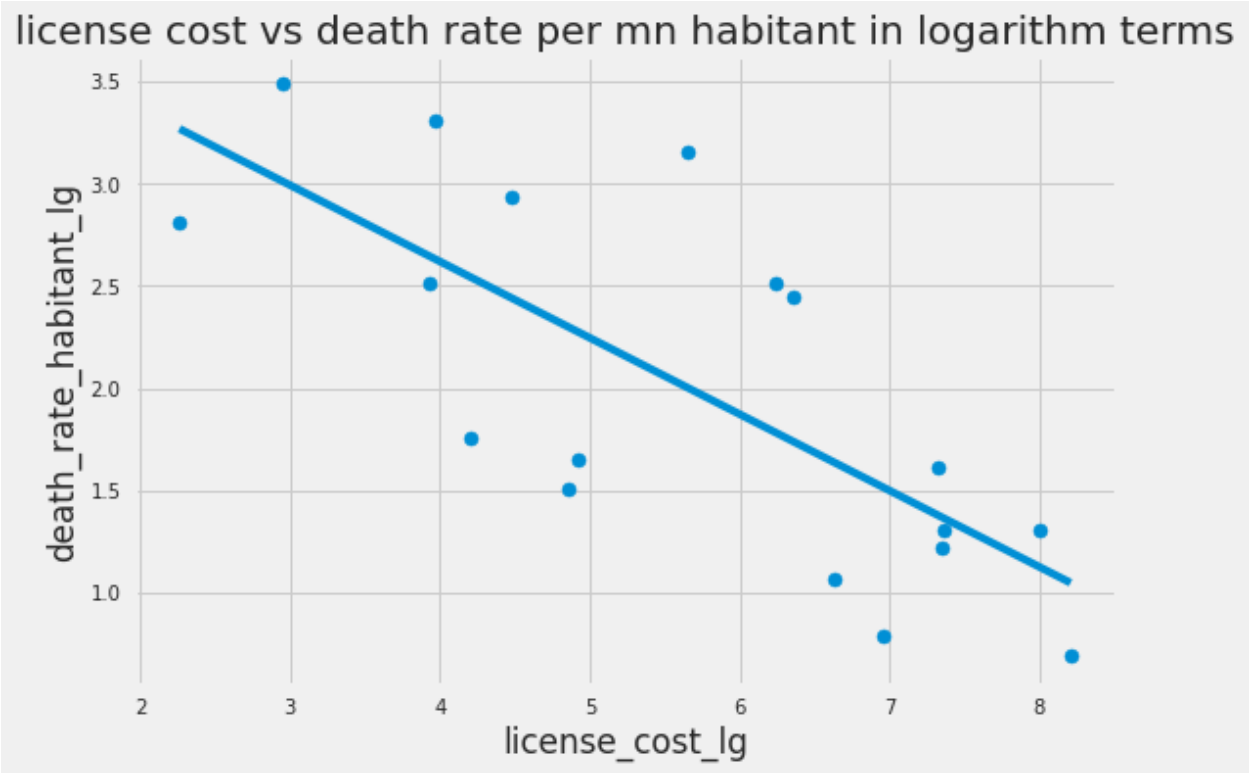
Data for license cost and death rate were not on the same scale and were heavily skewed. Hence we transformed both the license_cost and death rate to logarithm scale and then plotted the scatter plot to see the correlation. We can clearly see the correlation here.

We computed the Pearson correlation(-0.54) and found a negative moderate correlation between license cost and death rate.

We also tested a NULL hypothesis by setting alpha at 0.05 - that is "Death rate is same across different license cost" and got a p-value of 0.0024. That is, the NULL hypothesis is rejected and the death rate is not the same across different license costs. **Now this may be misleading.**

While there is inverse correlation between average license cost and death rate, this does not mean there is causal relationship. But what we can say, is that more training and tests require additional cost and hence cause an

increase in license cost which further causes reduction is traffic crashes because of more skilled drivers on the road.

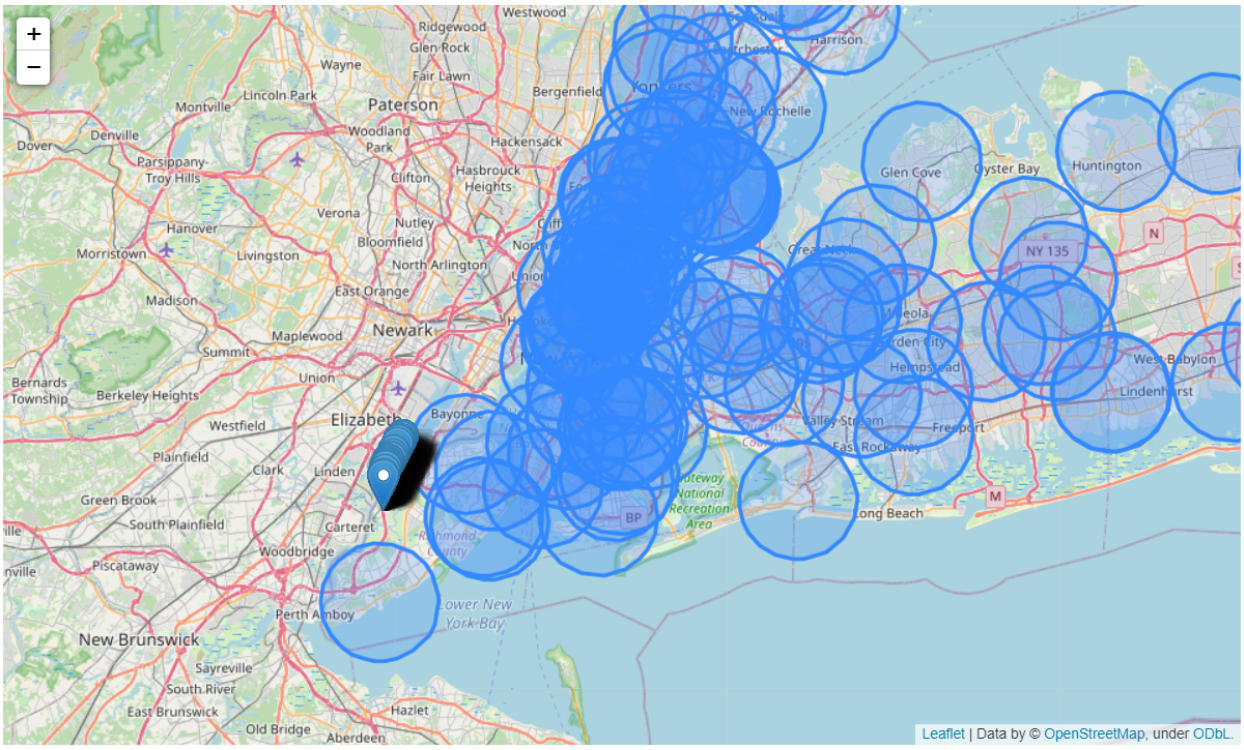


Are Hospitals Close Enough to Crash Sites:

As per WHO report, much of traffic related fatalities can be prevented by timely treatment of accident victims. In that endeavour, we analyzed the hospital coverage of accidents and checked if new hospitals are required and at what locations, and if this can prevent some casualties.

However, still there is scope of improvement and 245 crashes, we see happened outside the buffer range.

This area of *West Shore Expressway* can be further equipped with hospitals/primary care centers to prevent casualties.

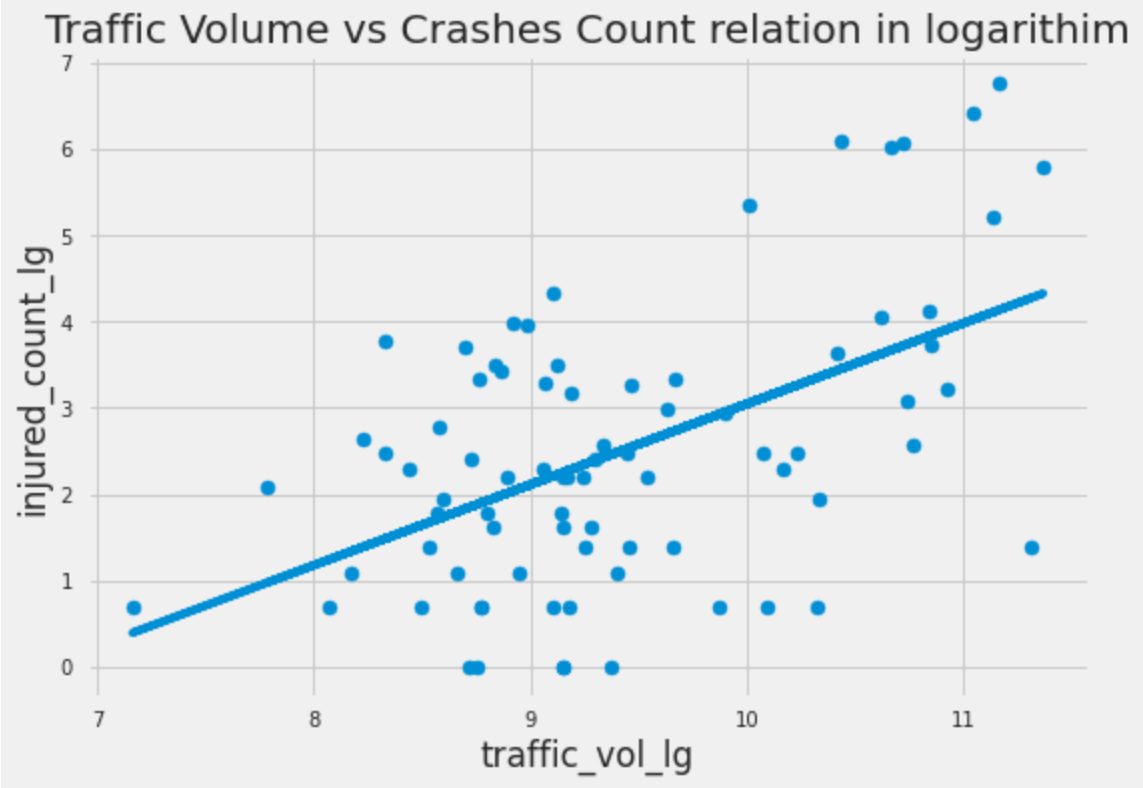


Traffic Volume Relation with Crashes:

We faced multiple hurdles in answering this question. And biggest of this was that traffic volume data provided by NYC Open Data was not complete. While it provides the data at different times of the day, it doesn't provide for all the days/months of the year. Leaving the data incomplete.

We still tried to get around the problem by computing the average daily traffic volume. Further we saw traffic volume and injured counts are highly skewed and so we transformed both the fields by taking their logarithms. With log of traffic volume and injured count plotted we can clearly see moderate positive correlation of traffic volume with crashes count. We computed the Pearson correlation coefficient and got moderate positive correlation of 0.59.

Hence, we would recommend decongesting the roads with most traffic volume to prevent the crashes.



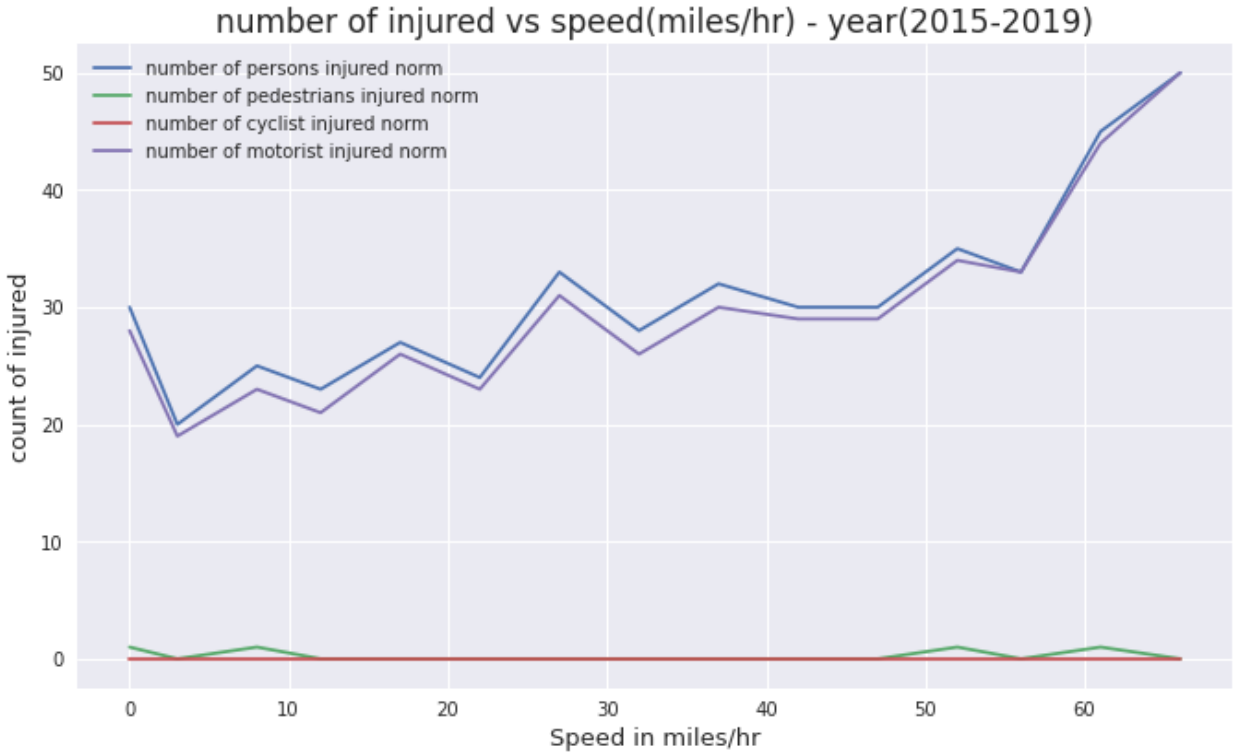
Traffic Speed Relation with Crashes:

Traffic speed dataset was the most challenging to handle. While we would expect the crash dataset to contain the speed of the crashed vehicle this was not the case. And traffic_speed dataset was provided separately as monthly dataset on a five minute interval for different roadways.

For this we used Folium library and instead of computing the minimum distance to each hospital, which would be time consuming, we utilized Folium buffer() method, to create a buffer of one mile around the hospital and then combined the connected buffers to create a unary buffer. Post that we checked if a crash location was outside the buffer range. While we appreciate that New York has good coverage of Hospitals, and that's the reason, even on choosing a small buffer of one mile we don't see many crashes outside the buffer range.

Now combining these two datasets crashes and traffic_speed was a challenge. While the crash dataset contained geolocation points(lat/long), the traffic speed dataset contained multiple geolocation points as linkPoints. Further, we found the linkPoints field to be often corrupted and having junk numeric value.

Google encoded polyline fields available on traffic_speed dataset came to our rescue. We could decode this field using a polyline package and get corresponding geolocation points of the road segment correctly. While this too could not decode all the polylines, we decided to leave the problematic values from the analysis.



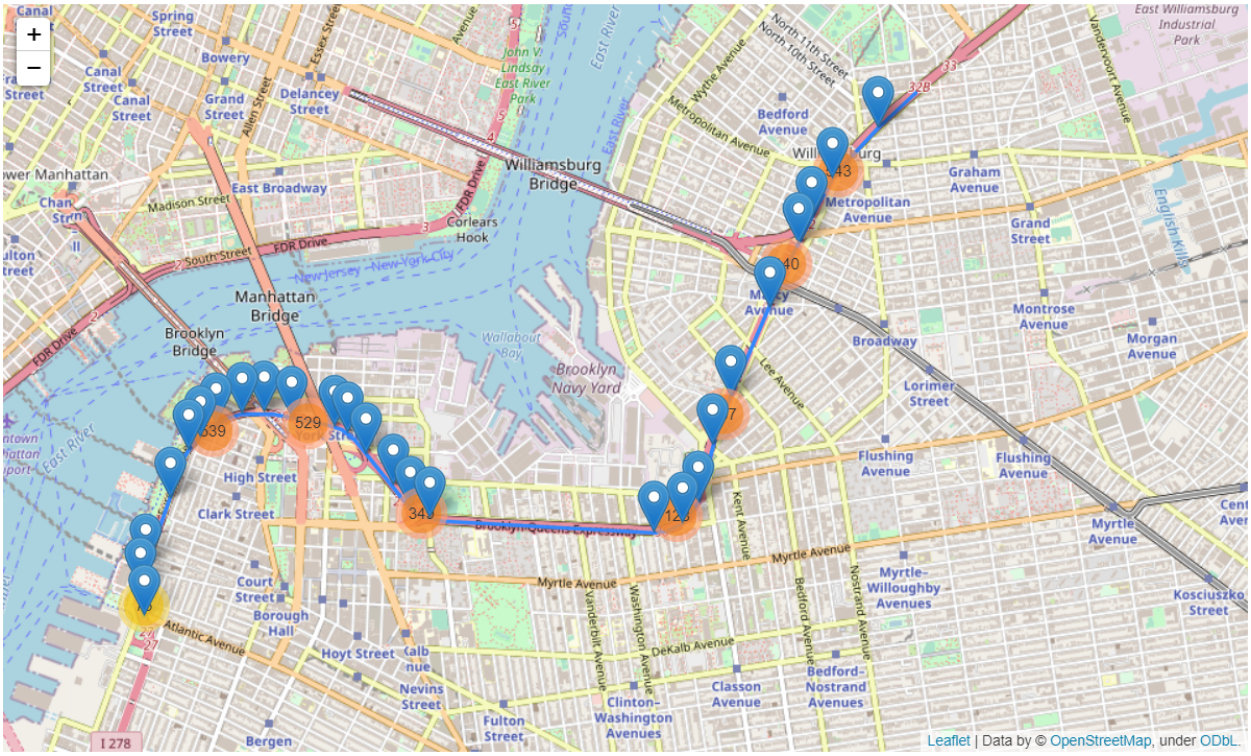
Further challenge was to combine the two datasets. We used the logic:

- Find the max_long/lat, min_long/lat for all the polylines
- Check if the crash point lies between max/min lat/long points.

Further to combine the two datasets, we used the pandas - merge_asof() method, which can be used to find the nearest/closest data point. Further we can provide a field to group by before merging on timestamp.

However, on plotting crash points we found that we were not getting the correct polyline mapping for crash points. So we also included the condition of minimum distance to polyline being 40 meter(a random choice trying to get all possible polyline mappings yet be correct).

On plotting these crash points we could confirm that our crash point to polyline mapping is correct.



Statement of Work

Amit	Sahil
• Knowledge and uses of Folium in spatial maps • Data Source Research and Selection • Data Cleaning • EDA and selection of final visuals • Report writing	• Extensive use and knowledge of leaflets • Data Source Research and Selection • Data Cleaning • EDA and selection of final visuals • Report writing

References

[1] World Health Organization - global-status-report-road-safety-time-action
https://www.afro.who.int/publications/global-status-report-road-safety-time-action

[2] World Health Organization - Road Traffic Injuries Jun-21
https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[3] Ministry of Road Transport and Highways - Road Accidents in India - 2019
https://morth.nic.in/sites/default/files/RA_Uploading.pdf

[4] International Cooperative and Mutual Insurance Federation
https://www.icmif.org/wp-content/uploads/2020/10/lb-app-road-traffic-accident.jpg

[5] The Parrish Law Firm
https://www.theparrishlawfirm.com/wp-content/uploads/2020/01/faqs-what-are-the-most-common-causes-of-car-accidents-in-virginia.jpg

[6] Proximity Analysis by Alexis Cook
https://www.kaggle.com/alexisbcook/proximity-analysis