NY Accident Analysis - Priyanka Patil, Pranav Rane

Abstract:

The objective of this project is to study the accidents that occured in June-July for the years of 2017 and 2018. We intend to identify the most accident prone areas in the borough of Brooklyn in NYC. Such information will be displayed in terms of cars and pedestrians. The end goal is to make data trends visible for the citizens of New York City to improve their safety. This report will also help New York City Management to make better decisions about city planning.

Overview:

New York City (NYC) is the most populous city in the world, with population of 8.6 million. About 1.4 million households out of 3.1 million households own a car. Everyday the people and cars interact with each other. The scale of this interaction makes NYC very unique. Hence such a city requires a deeper dive into data trends.

The neighbourhoods in NYC are highly varied. As the neighbourhoods change, behaviours of people change. Hence we have decided to study only one borough. We are studying data from June-July for the years of 2017-18 in Brooklyn. Brooklyn has the number of accidents in all boroughs for the specified time period. The demographics of Brooklyn are also adequately diverse. This has allowed us to study behaviour of adults and minors.

This report will contain Year over year changes for the months of June and July. The most prominent accident prone areas will be highlighted. The most common days for accidents will be specified. The overall accident report will be specified using a heatmap.

Data Preparation:

Originally the data had a lot of missing cells. We wanted to ensure that every row has at least borough information and GPS coordinates. To ensure that we used two libraries namely geocoder and uszipcode.

For records that had some form of address geocoder was used to convert address to Zip Code, Latitude and Longitude. For records where GPS information was present, but borough was absent, we used uszipcode to convert Latitude and Longitude to Zip Code. Finally we designed a code that converted all Zip Codes to Boroughs and removed Zip Codes outside the 5 boroughs. For other columns lie CONTRIBUTING FACTOR VEHICLE and VEHICLE TYPE CODE, empty values were replaced by 'Unspecified' or 'Unknown'. After all this every cells had values in them.

We have not quantized data into regions. Data was filtered for time and borough. There were collisions(<2000 rows) where predicted Borough did not match with Existing Borough. Those values were removed.

The data for months(June 2017 vs June 2018 and July 2017 vs July 2018) was not comparable initially due to mismatch in number of weekends. Weekends are significant in terms of accidents. So we have made an adjustment. June 2017 for all calculations contains one extra day from July 2017. This makes allows June 2017 and June 2018 to have 5 weekends. July 2017 and July 2018 have 4 weeks.

Results:

Initially we are comparing heatmaps for June July 2017 and June July 2018. It is difficult to observe insights directly, so we deep dive into data.

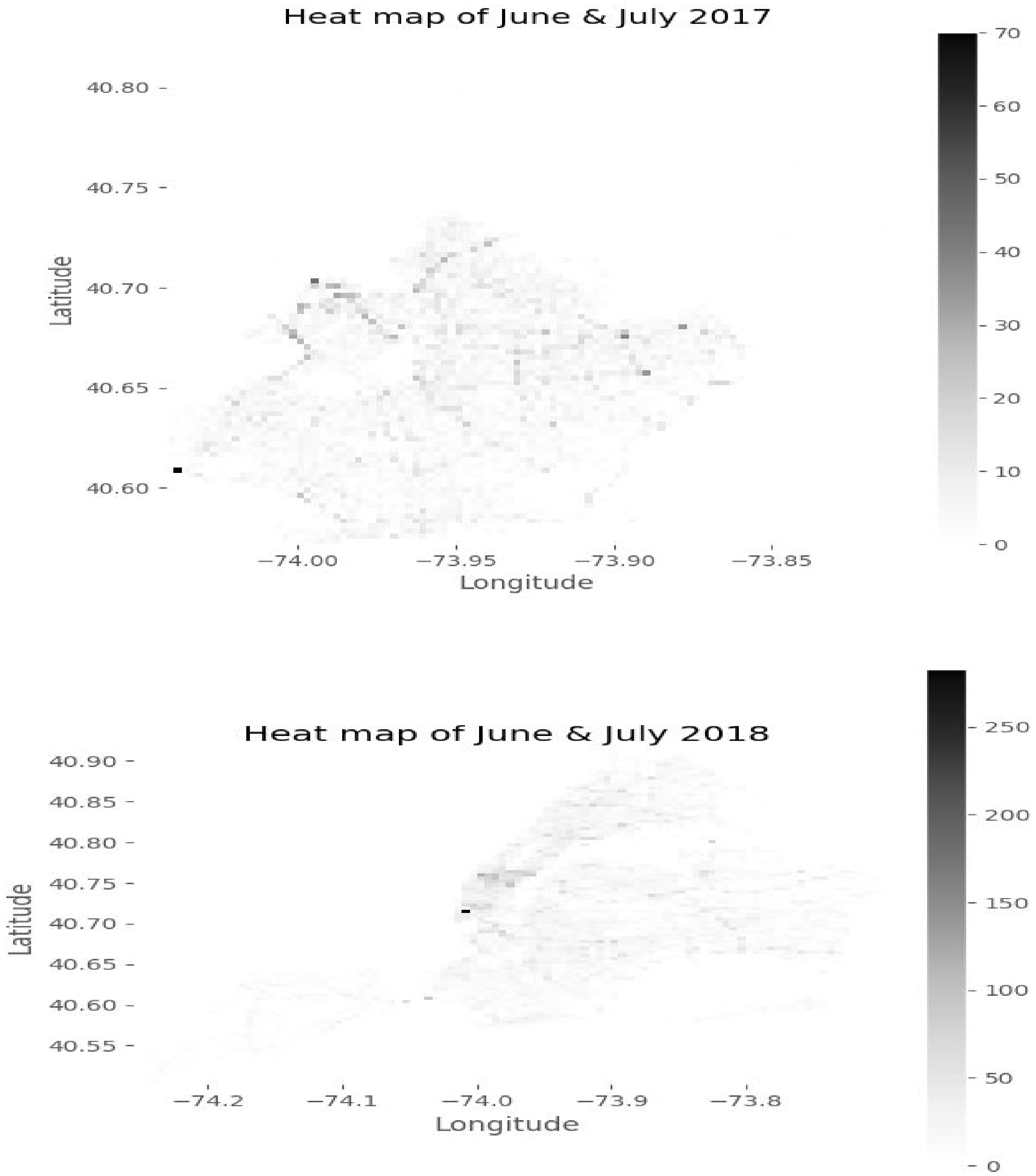

Fig1: HeatMaps for June and July 2017-2018

We have make heatmaps for every month. But it is hard to see values clearly as there are too may values. Here June 2017 has 5 full weekends.

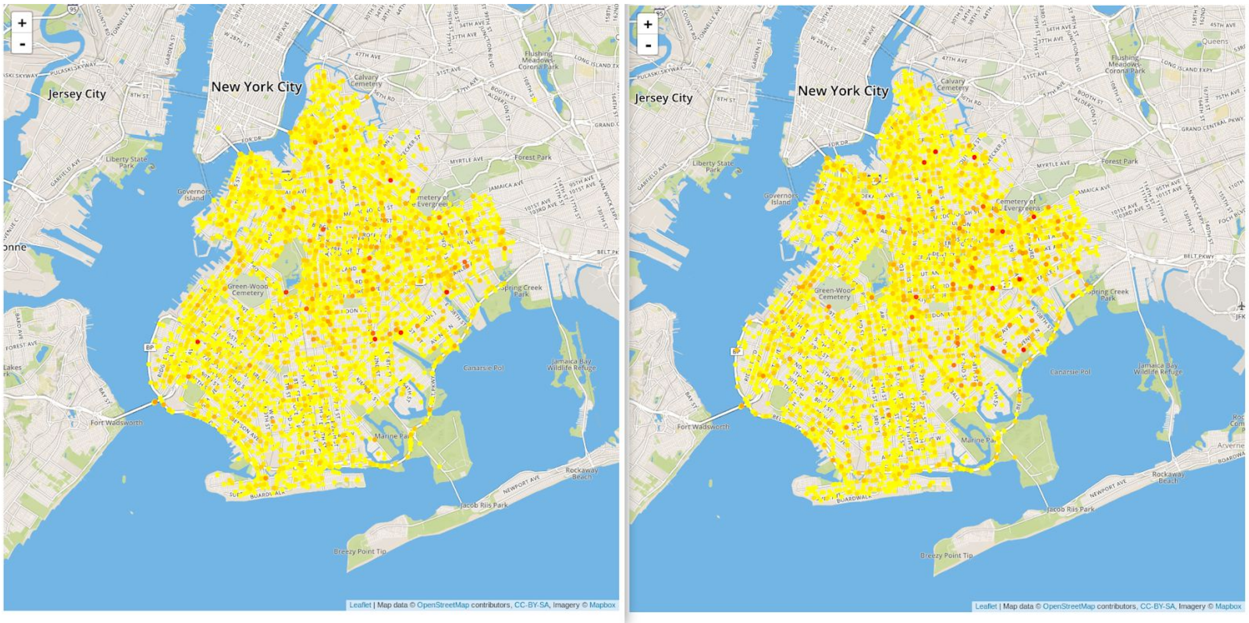

Fig2: June 2017 vs June 2018 heatmap, All Data

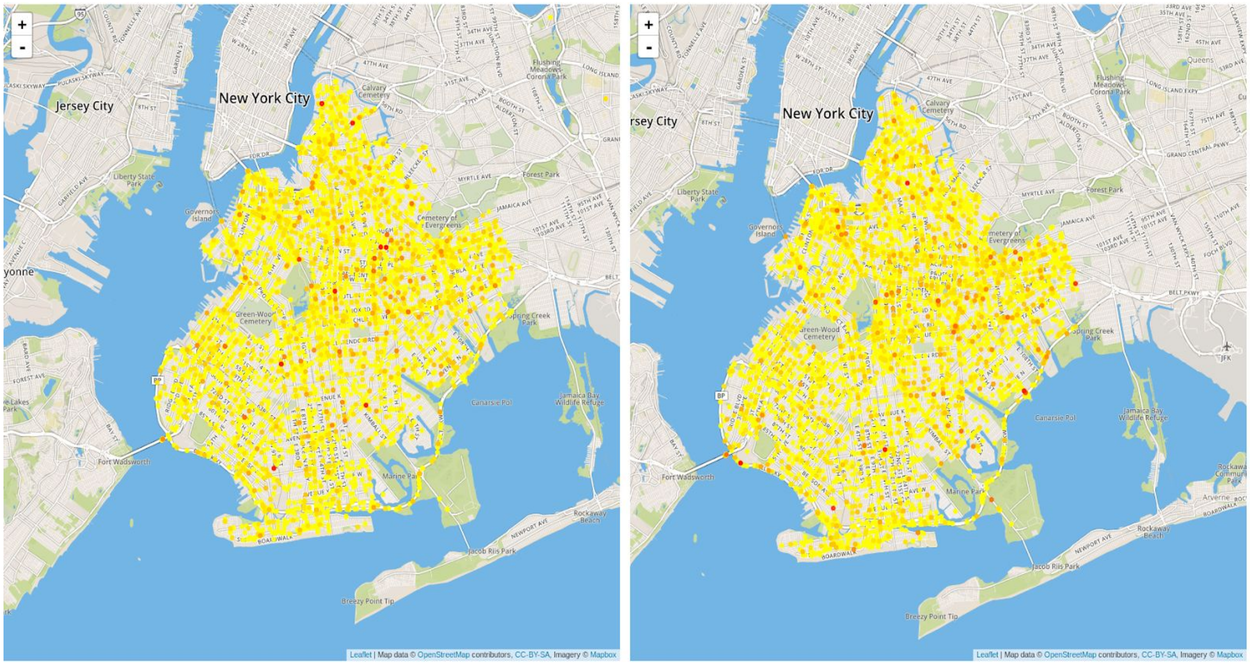

Fig3: June 2017 vs June 2018 heatmap, All Data

June 2017 has more accidents than June 2018. June 2017 had 21,197 incidents while July 2018 has 19,802 incidents.

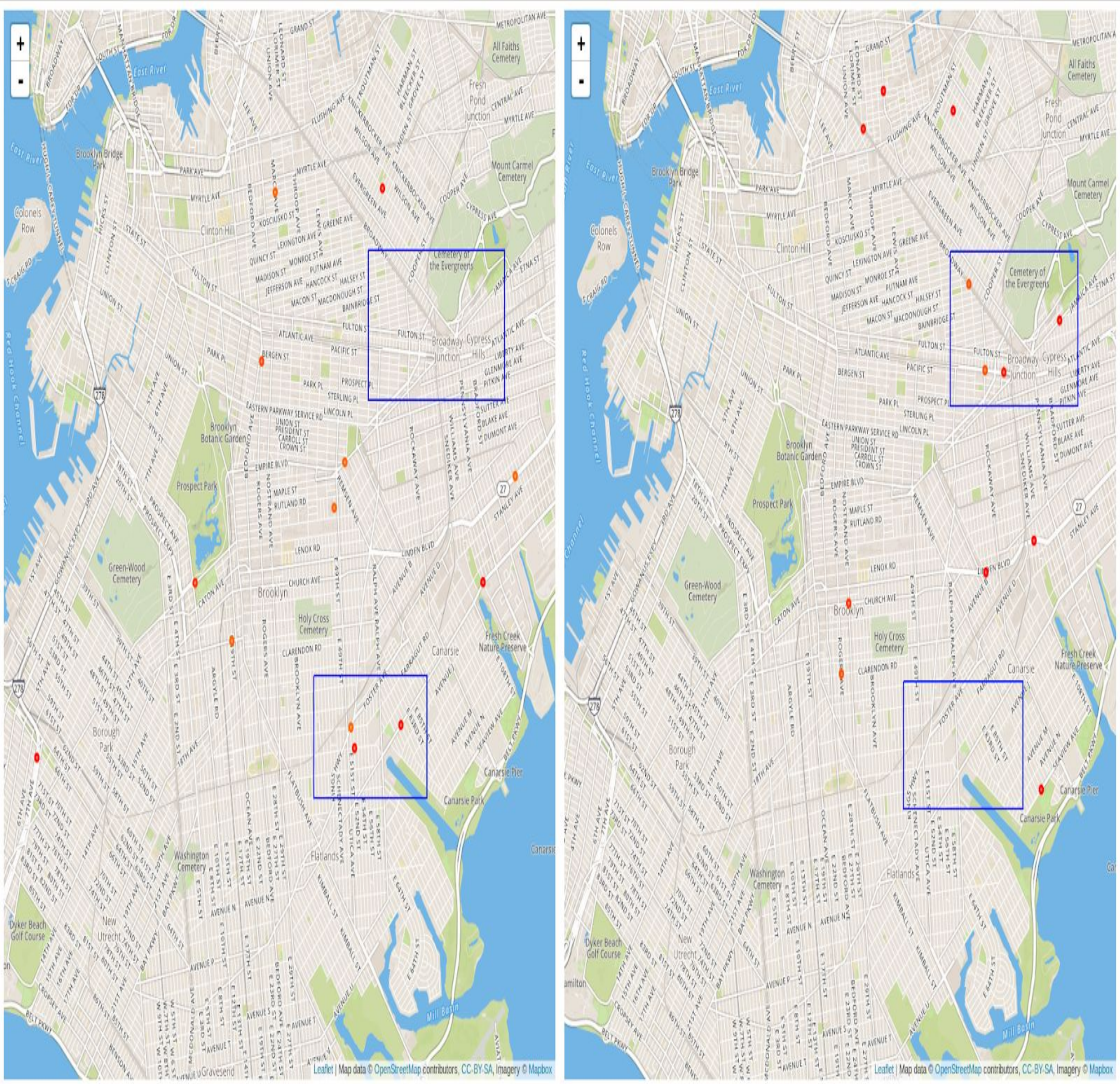
Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox
Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox

Figure 4: June 2017 vs June 2018, Highest Accident Areas Shown, Changes Highlighted

Some new accident hotspots are created in June 2018, while some areas not as accident prone as before. Atlantic Ave and Broadway have more accidents, while Utica Ave. has lesser accidents.

June 2018 has lesser accidents than July 2018. July 2017 had 18,897 incidents while July 2018 has 19,591 incidents.

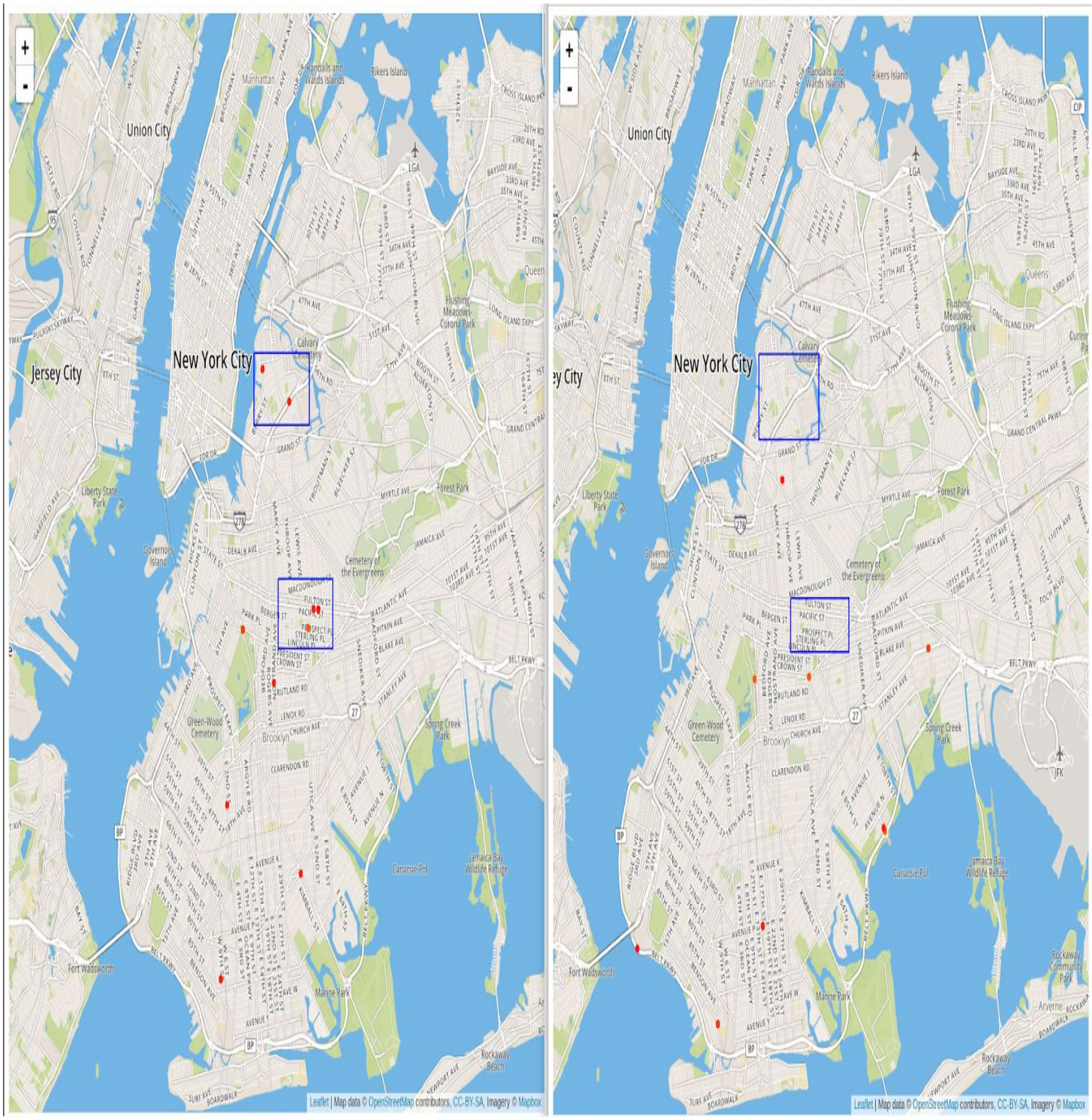
+
-
Union City
Jersey City
New York City
Manhattan
Randalls and Wards Islands
Rikers Island
LGA
Calvary Cemetery
Liberty State Park
Governors Island
Cemetery of the Evergreens
Forest Park
Flushing Meadows Corona Park
Queens
Green-Wood Cemetery
Brooklyn
Spring Creek Park
Canarsie Pol
Jamaica Bay Wildlife Refuge
Fort Wadsworth
Marine Park
Rockaway Beach
JFK
Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox
Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox

Figure 5: July 2017 vs July 2018, Highest Accident Areas Shown, Changes Highlighted

Atlantic Avenue and Meeker Avenue are safer in 2018 compared to 2017 for July.

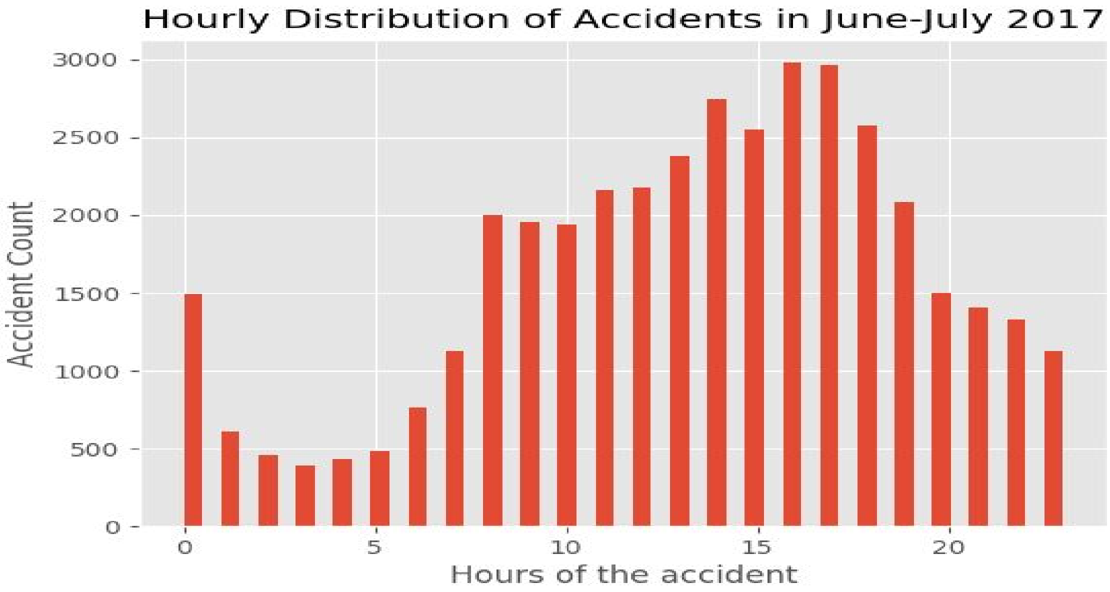

Fig 6: Hourly Distribution of Accidents in June-July 2017

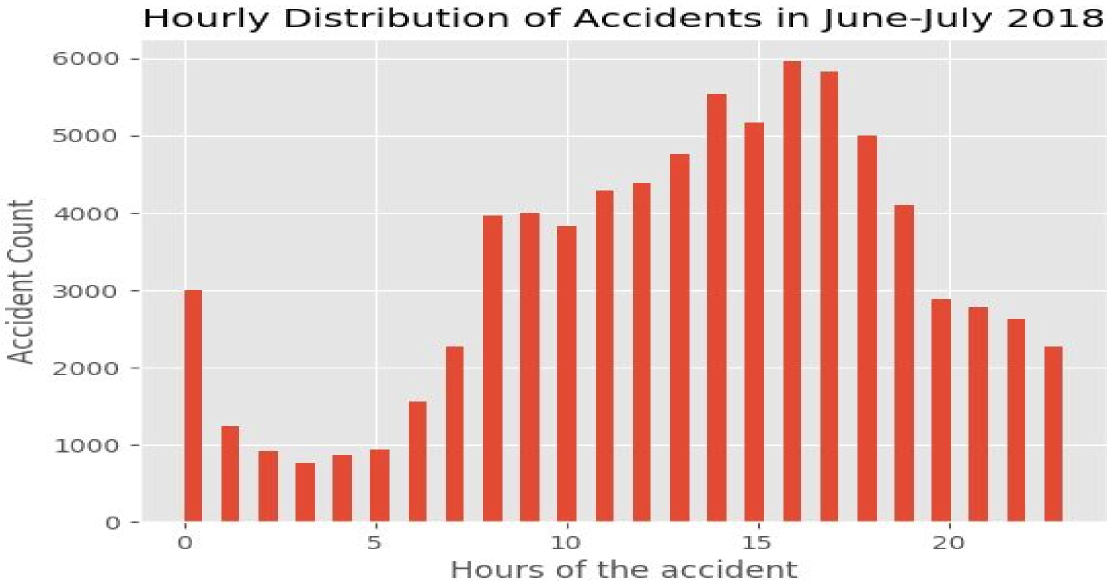

Fig 7: Hourly Distribution of Accidents in June-July 2017

Daily Accident Trends also Stay same for both years. There is a strong spike on Sundays. This is consistent with trends from other cities

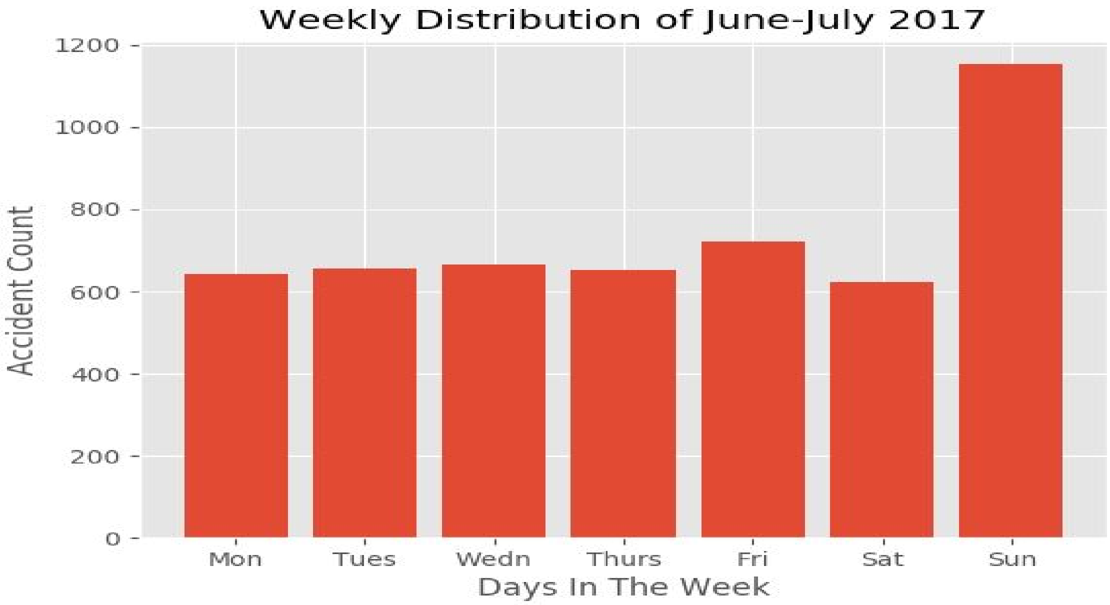

Fig 8: Weekly Distribution of Accidents in June-July 2017

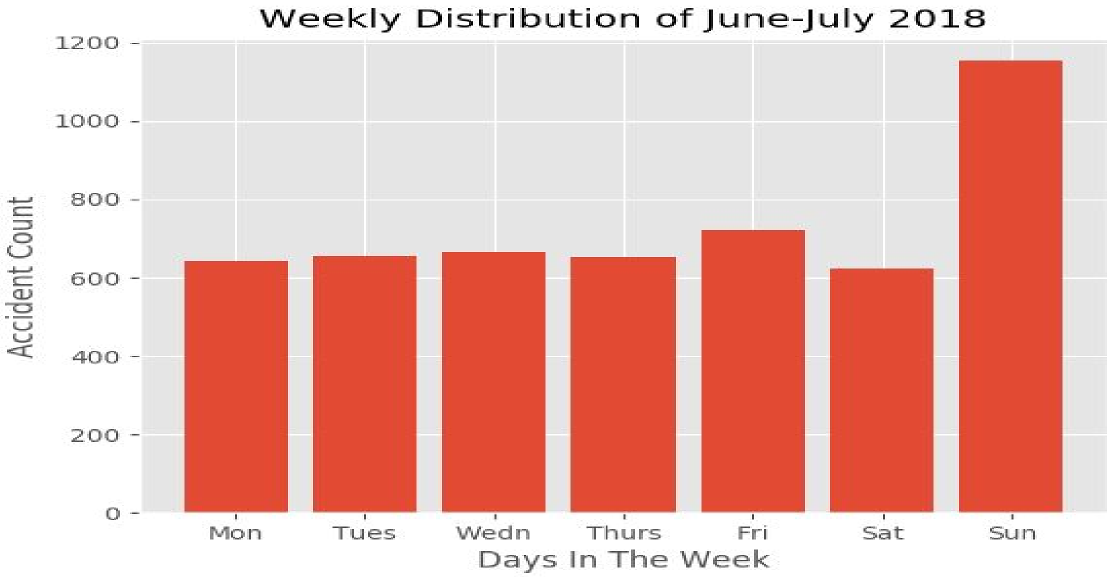

Fig 9: Weekly Distribution of Accidents in June-July 2018

Driven Inattention and Driving too close to each other are the prominent causes for accidents

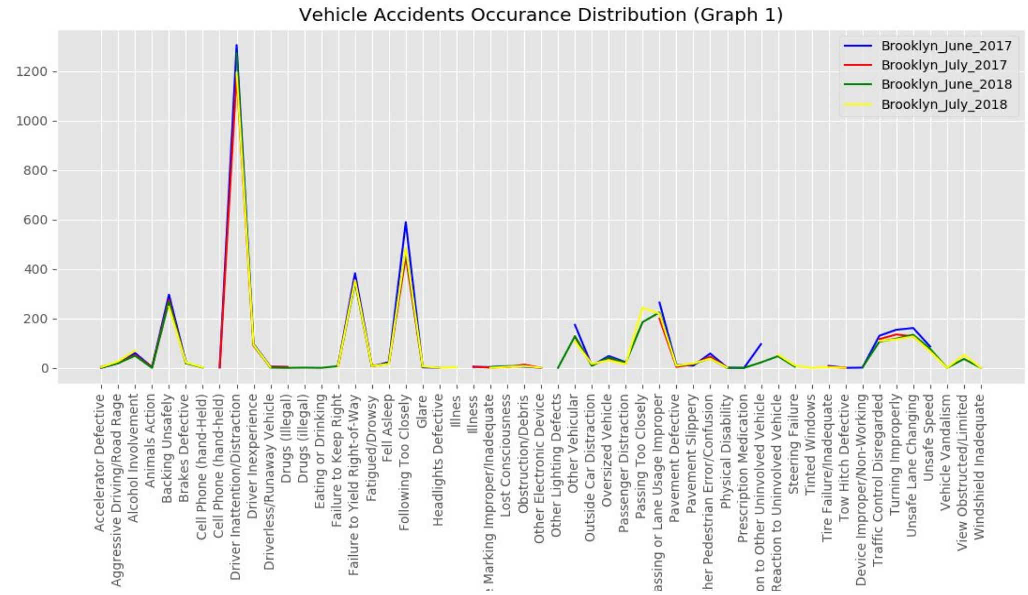

Fig 10: Causes of Accidents

We also generated the 5 most dangerous places for Pedestrians based on accidents. The locations change over the year. But Prospect Park remains a dangerous place for both years.

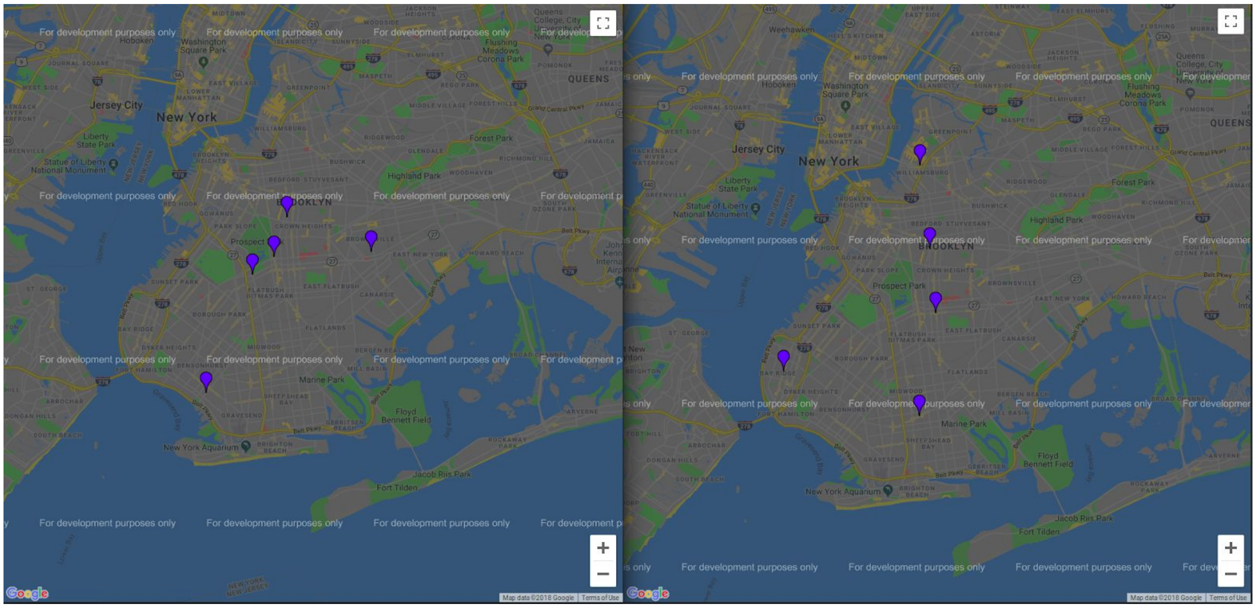

Fig 11: Most Accident Prone Areas for Pedestrians, June July 2017 vs June July 2018

We have not performed clustering for this project. The only commensurate metrics were latitude and longitude. But unfortunately we don’t think clusters generated from GPS are actually insights. We also considered making a feature by assigning lethality score to every accident. But we could never got any substantial results.

We never found any need to normalize data. Almost all data was directly usable and relevant.

Experience and Challenges:

Data Cleaning was a big initial challenge. We did not consider all cases for data errors so we spent a lot of time in trial and error.

Data Clustering attempts also took substantial time and discussion. We spent time discussing if metrics are truly commensurate. But we could not present usable insights about data using clustering.

Finally we experimented a lot with heat maps. We were not getting the exact package, so we had to try out a lot of them. We specifically wanted a package that will allow us to edit the display of most accident prone places. That was the only way to meaningfully display our results.

Overall the dataset is very rich. If more time is alloted, then better insights for the entire year can be generated. There are so many other factors like accidents during winter, impact of sunrise-set on accidents, tourism impact on accidents. We can spend more time studying impact of policy changes. We didn’t really get a lot of chance to explore that.

As far as data mining is concerned, we learnt about exploring or playing with data before writing code for data. This would have saved us a lot of time during the cleaning phase.

References:

1. Data.cityofnewyork.us. (2018). [online] Available at: https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95 [Accessed 1 Dec. 2018, 2 times].
2. Health.ny.gov. (2018). *NYC Neighborhood ZIP Code Definitions*. [online] Available at: https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm [Accessed 5 Dec. 2018, 3 times].
3. Shetty, R. (2018). New York City Vehicle Collision Analysis · rajeshetty. [online] Rajeshetty.com. Available at: http://rajeshetty.com/post/nycvehiclecollision/ [Accessed 7 Dec. 2018, 5 times].
4. Usatoday.com. (2018). [online] Available at: https://www.usatoday.com/story/money/nation-now/2018/05/26/driving-car-crash-deaths-speeding/640781002/ [Accessed 8 Dec. 2018, 2 times].
5. Nytimes.com. (2018). *New York City's Population Hits a Record 8.6 Million*. [online] Available at: https://www.nytimes.com/2018/03/22/nyregion/new-york-city-population.html [Accessed 10 Dec. 2018, 1 time].
6. NYCEDC. (2018). *New Yorkers and Their Cars*. [online] Available at: https://www.nycedc.com/blog-entry/new-yorkers-and-their-cars [Accessed 10 Dec. 2018, 1 time].