New York University

Center for Data Science

Capstone Project Proposal

Rafael Garcia Cano Da Costa(rgc292), Jirou Xu(jx654)

02/09/2016

Business Problem:

In New York City, serious injury and deaths caused by the daily traffic receive a special attention from the Mayor's office. Starting in 2014, Mayor Bill De Blasio launched a program called Vision Zero. This program has as a core goal the reduction of the number of those serious injuries and deaths to zero. The undertaking foresees the involvement of different institutions and people in the city on an effort to cover all possible contributors when it comes to traffic accidents having people as victims.

The city resorting to an action plan started a myriad of initiatives to propel the decrease of serious injury and deaths. Among them, in partnership with the Limousine and Taxi Commission, registered drivers have to go through training every 3 years in order to keep their work concession. The training comprises orientations on yielding when pedestrians are crossing streets, on respecting speed limits and red lights, and so on. This is not all. Streets are being redesigned as a way to keep pedestrians safer. New crossing points based on pedestrians needs are being created, and intermediate areas called islands where people can wait while crossing streets in case of a light turning red before reaching the next curb are another reengineering example. Beyond this, new laws were enacted reducing the citywide speed limit from 30 to 25 miles per hour, strengthening sanctions to whom hit someone on the traffic and run, and changing automobile design for minimizing drivers' blind spots. Many other approaches are being taken as a means to keep people away from serious injury and deaths on the traffic. Despite all effort in place, some new challenges surface on how to evaluate the effectiveness of the Vision Zero action plan.

This project comes handy in an attempt to quantify how impactful Vision Zero is being on its intent, to visualize metrics before and after the action plan was launched for getting new insights, and to find out hidden correlations that could be addressed potentiating the desired outcome of Vision Zero.

Data Mining Problem:

The data mining goal is compounded by three items. Two aspects relate to changes that were implemented by Vision Zero. This team will create visualizations using the available data. The range of time explored will be attached to the length of time the action plan has been in vogue. For instance, since the initiative took off in 2014, it is being active for almost three years. As a result, this work will plot some metrics derived from the data back three years from 2014 starting in 2011. Once the data is plotted, comparisons among different plots will be performed while looking for insights. In parallel, hypothesis testing will be applied in order to evaluate if changes on trends after 2014 are statistically significant. Finally, a search for hidden correlations among not so obvious databases and their attributes will take place as a means to discover new ways to decrease serious injury and deaths on the New York City' traffic.

Initially, the work will focus on specific sites where the Vision Zero action plan made modifications. For instance, it is expected that a redesigned street will have a decrease on injuries and deaths, and once this data is visualized, it should be easy to identify the new trend. However, just seeing the new trend is not enough for the purpose of this study. Therefore, a hypothesis testing will offer a statistical support for evaluating if the new trend is significant. Once, the new trend is determined, this team will apply the similar method to larger areas to detect the change Vision Zero makes throughout the whole city.

In sequence, this work intends to discover a new approach for making Vision Zero more impactful. This will happen through correlational analysis between data on serious injury and deaths on traffic, and data from not so obvious datasets such as those on wi-fi hotspot locations and on parking violations issued. With the correlational analysis, this team intends to use block or zip code as unit to rank the districts from unsuccessful to successful to show the impact of the enacted plan in different areas.

In the end, this project aims to evaluate the impact of Vision Zero on resolving the issue on serious injury and deaths on traffic, and to suggest new approaches for amplifying the effectiveness of Vision Zero in New York City, especially in those districts that the impact of Vision Zero is not significant.

Data Details:

We are planning to use data from plenty of datasets, which can be mainly divided into 4 categories. First, this team plans to use Taxi data collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP). In addition to this dataset, Yellow Cab data and Citi Bike data will be gathered as a further supplement to explore the traffic condition of streets. Second, NYPD Crime data and NYPD collision data will be used to record the traffic injuries and fatalities along with other accidents related to street safety before and after the enacted plan. Third, Subway Station Locations and Subway Entrance Locations will be used as a resource of street design, to demonstrate the changes before and after 2014. Lastly, the remaining datasets will provide additional features, exploring the correlation between the implement of Vision Zero and relative urban characteristics. Most of the supporting datasets are available on NYC Open Data. The detailed information of each of the datasets are described below:

- Green Cab: ~53436857 records, 21 columns, ~7.92 GB
- FHV Cab: 178469955 records, 18 columns, ~5.49 GB
- Yellow Cab: ~761273804 records, 19 columns, ~119.41 GB
- Citi Bike Data: ~28548005 records, 15 columns, 921.82 MB
- NYPD Crime Data : 1123465 records, 20 columns, 194.1 MB
- NYPD Collision Data : 892927 records, 29 columns, 171.3 MB
- Subway Station Locations: 471 records, 4 columns, 50 KB
- Subway Entrance Locations: 1905 records, 4 columns, 234 KB
- Housing Maintenance Code Violations: 1715366 records, 30 columns, 664.6 MB
- Parking Violation Issued: 39454354 records, 43 columns, ~7.68 GB
- NYC WI-FI Hotspot Locations: 3359 records, 16 columns, 373 MB
- Bus Stop Locations: 13280 records, 11 columns, 1.2 MB
- 311 Complaint Data: different sources of data
- NYC Pluto: 859469 records (split into 5 .csv files, one per borough) 83 columns, 432.9 M

Several of the relevant features are already available in one of the listed datasets. However, a number of features will need to be engineered. For example, Taxi data, Yellow Cab data and Citi Bike data record location by longitude and latitude, while the NYPD Crime records crime location by borough, which indicates that a uniform measure needs to be found. Moreover, the features we anticipate including in analytic dataset are found across multiple datasets. Thus, in addition to feature engineering, we will need to clean and merge multiple disparate datasets to construct our final analytic dataset.

Schedule:

The below tentative schedule will guide this project to be on track:

10/02-10/15:	Code development for data cleaning and small scale testing.
10/16-10/29:	Visualization of data from previous step.
10/30-11/12:	Implementation of previous steps on data from 2011 to 2016.
11/13-11/26:	Evaluating statistics and hypothesis testing on viewed data.
11/27-12/10:	Correlational analysis through cluster and regression.
12/11-12/15:	Report and presentation preparation.