James Kasakyan
Sunday, April 3, 2016
CSC 59927 Big Data Management and Analysis

Final Project Proposal

Motivation

According to the NYS DMV[1], between 2011 and 2013 there were 212,267 motor vehicle accidents in New York City, resulting in 833 deaths. In 2014, New York Mayor Bill de Blasio announced the Vision Zero Action Plan, with the aim of eliminating all traffic fatalities in New York City by 2025.

Year	Total accidents	Fatalities
2011	73,060	268
2012	68,804	271
2013	70,403	294

When the city debated the causes of these accidents and the most effective means to combat them, they turned to a dataset compiled by the DOT, which analyzed all NYC traffic crashes from 2008 to 2012. For this project, I would like to examine this dataset from 2012 to 2013 along with traffic volume and vehicle classification data, and 311 complaints to identify areas around the city that are particularly hazardous for motor vehicles, and display common "symptoms" of these areas.

Project Objectives

The main objective of this project is to identify locations around the five boroughs that are prone to motor vehicle accidents by factoring in data like the number of accidents, fatalities, injuries, and number of vehicles involved (adjusting for traffic volume), and profiling these hazardous areas.

Tasks derived from objectives:

- Develop classification algorithm for classification of areas
- Analyze hazardous areas for "symptoms"
    - (a) Analyze 311 complaints related to street/traffic conditions
    - (b) Analyze type of traffic (commercial, passenger, taxi)
- Plot areas labeled as hazardous, along with profile for area (most common vehicle type, most common 311 traffic/street condition related complaint)

[1] NYS DMV *New York City Crash Summaries* https://dmv.ny.gov/org/about-dmv/statistical-summaries

Hypotheses for profiles:

- Hazardous areas will often have commercial or taxi traffic associated with them
- The most common 311 traffic/street related complaint in hazardous areas will be potholes

Data sources

Name	Years	Provider	Link
NYPD Motor Vehicle Collisions	2012-2013	NYC OpenData	https://datahub.cusp.nyu.edu/dataset/h9gi-nx95
Traffic Volume Counts	2012-2013	NYC OpenData	https://datahub.cusp.nyu.edu/dataset/p424-amsu
Vehicle Classification Counts	2012-2013	NYC OpenData	https://datahub.cusp.nyu.edu/dataset/ae5u-upr6
311 Service Requests	2010-Present	NYC OpenData	https://datahub.cusp.nyu.edu/dataset/erm2-nwe9

Methodologies

This project will make use of the data mining technique of classification. This follows naturally from the objective of the project, which is to identify hazardous areas based on a set of criteria relating to motor vehicle accidents. This will involve building the classifier from a training set. It will be necessary to develop a classification algorithm that will label a given location as "hazardous" or "O.K." based on factors like the number of motor vehicle accidents, the number of fatalities and injuries in those accidents, and the number of vehicles involved, all adjusted for traffic volume in those areas.

The project also incorporates a small amount of descriptive analytics, in the form of profiles for hazardous areas (total number of accidents, fatalities, injuries, most common vehicle type, most common 311 traffic/street condition related complaint).

Big Data Challenges

The data preparation stage of the data analytics lifecycle will provide the biggest challenges for this project. Because of the number of datasets used, data conditioning will play a massive role. There are 146 fields and over 10,000,000 rows over the four proposed datasets. A vast majority of those fields will need to be ignored, and relevant fields will require extensive normalization.

For example, although the datasets for NYPD motor vehicle collisions and 311 service requests have fields for latitude and longitude, the traffic volume and vehicle classification counts sets do not provide these fields, instead establishing location via a roadway name field. Since areas will be split up by ZIP code, it will be necessary to map both roadway name and latitude and longitude to ZIP code. For roadways, a big challenge will be deciding how to deal with roads that stretch across numerous ZIP codes.

Other challenges include factors like variety among values in common fields. The NYPD motor vehicle collisions and vehicle classification count datasets also both provide a field classifying the type of vehicle involved, but use different terms (Auto vs. Passenger, Commercial vs. Small Com Veh(4 tires)

Data cleaning will also be required as many of the datasets have missing values, with the 311 service requests dataset missing over 22% of its values. The vehicle classification dataset provides fields for each hour of the day to classify the amount of each type of vehicle that passed along that road during that hour, but from the sample data it seemed as if the sensors were not active 24/7 like the traffic volume sensors. Most days only had data for a continuous stretch of 10-12 hours.

Project Deliverables

Since the goal of the project is to identify hazardous areas for vehicles around the five boroughs by examining motor vehicle accidents, and then examining the characteristics of these areas, it seemed natural that a 2D choropleth map of the five boroughs (partitioned by ZIP code) would be the most appropriate type of data visualization for this project.

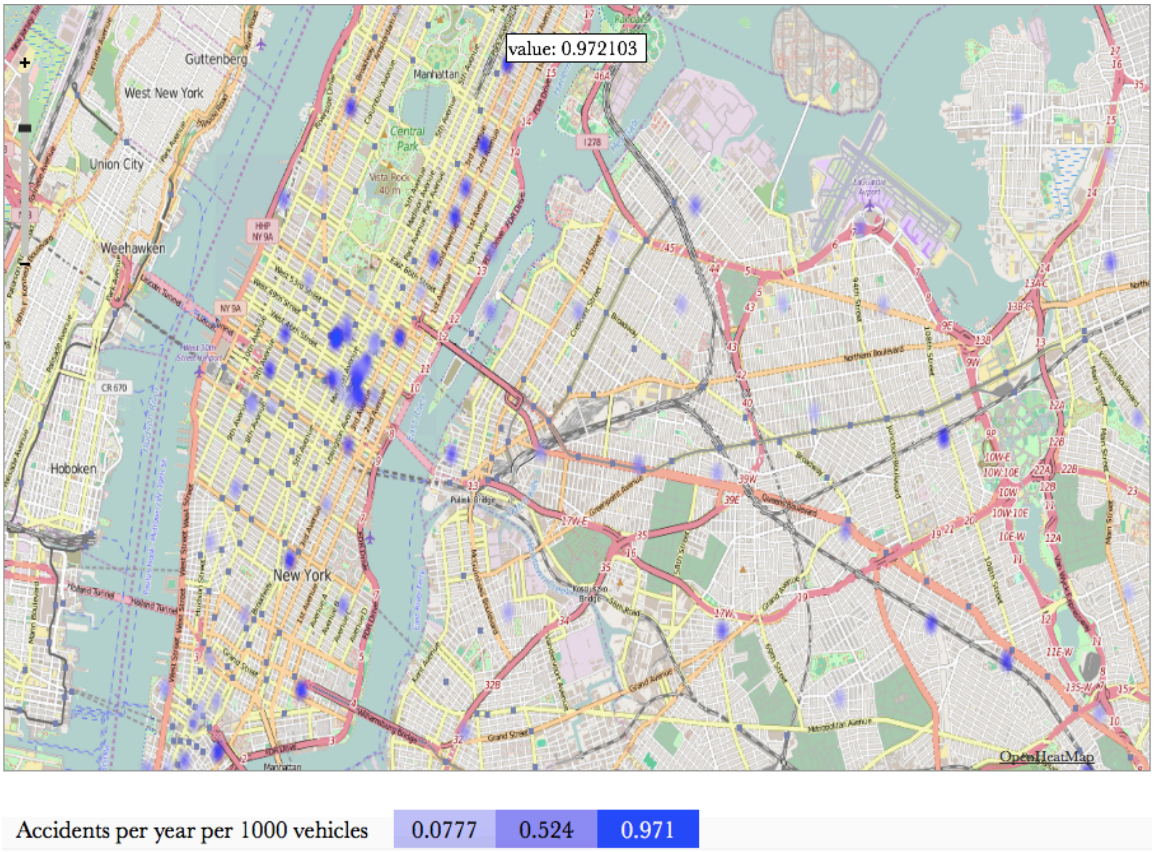

Figure 1- Mock up for visualization

This mock up was generated from an excel sheet using random values between 0 and 1 for every zip code in the five boroughs. While the key implies that the values indicate the number of accidents per 1000 per year, the final visualization for this project will instead display values in a 0 to 1 range indicating the severity of the danger associated with each hazardous area. Mousing over it would display the profile for that area.