Jean Chang
CSC 59927
Spring 2018

Final Project Proposal

Motivation

According to the NYS DMV, between 2011 and 2013 there were 232,267 motor vehicle accidents in New York City, resulting in 833 deaths. In 2014, the Vision Zero Action Plan was announced and its goal is to eliminate all traffic deaths and injuries in NYC. When the city debated about the causes of these accidents and how to prevent more traffic deaths and injuries, the DOT released a dataset of all NYC traffic crashes from 2008 to present. For this project, we will examine this dataset from 2012 to 2013 along with traffic volume and 311 complaints data to identify the areas in NYC where it is hazardous for motor vehicles.

Project Objectives

The main objective is to identify hazardous areas around NYC that are prone to motor vehicle accidents by factoring in data like the number of accidents, fatalities, injuries, and number of vehicles involved and profile these hazardous areas. The plan is to develop a classification algorithm for classifying the areas, analyze 311 complaints related to street/traffic conditions and type of vehicle (commercial, passenger, taxi), then plot areas labeled as hazardous with a profile (most common vehicle type, most common 311 traffic/street conditions complaints). The hypotheses are hazardous areas will often have commercial or taxi traffic and the most common 311 complaints will be potholes.

Data Sources

For this project, we will be using three datasets all provided from NYC OpenData website. We choose to analyze the data from 2012 to 2013 is because the data about traffic volume from the DOT is only available from 2011 to 2013.

The datasets are listed in a table below.

Name	Years	Provider	Link
NYPD Motor Vehicle Collisions	2012-2013	NYC OpenData	https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95
Traffic Volume Counts	2012-2013	NYC OpenData	https://data.cityofnewyork.us/Transportation/Traffic-Volume-Counts-2012-2013-/p424-amsu
311 Service Request	2012-2013	NYC OpenData	https://data.cityofnewyork.us/Social-Services/311-Service-Requests/fvrb-kbbt

Methodologies

In this project, we will develop a classification algorithm (supported by Spark) that will label a given location as “Hazardous”, “OK”, or “Safe” based on the factors listed above. We will compute the average and standard deviation in three feature categories for each area (by zip code):

1. Accidents /1000 vehicles
2. Injuries /1000 accidents
3. Deaths /1000 accidents

The idea is that those areas where multiple feature have values greater than one standard deviation from the mean are the most hazardous.

Big Data Challenges

The biggest challenge for this project is the data preparation stage of the data analytics lifecycle. This is because of the number of dataset used as there are 146 fields and over 10,000,000 rows over the three datasets. Only the relevant fields will require extensive normalization. Other challenges included variety among values in common field. For example, the NYPD motor vehicle collisions and vehicle classification count datasets both provide a field for the type of vehicle but using different terms. Data cleaning will also be needed because many of the datasets have missing values.

Project Deliverable

The goal of this project is to identify hazardous areas in NYC by examining motor vehicle accidents and the characteristics of these areas. The type of data visualization we proposed is a 2D choropleth map of the five boroughs partitioned by the zip code.