BU Department of Computer Science

Utilizing New York Accident Data to Better Place Response Centers

Alan Burstein • Joseph Caluag • Andrew Quan • Eric Jacobson

Introduction

Vehicle accident records can be a powerful tool in determining safety protocols for cities which want to allocate resources in the construction or placement of response centers. Using data from various municipal databases and external constraints that we impose on the problem, a solution can be calculated that satisfies a desired physical objective.

Additionally, by comparing data from different cities, we can determine whether or not their accident data is correlated. This can be used to determine whether or not similar protocols should be used on those cities. We can further apply the rules of independence on sub-regions or boroughs of cities to see if distinct laws should be applied to each area.

Goals

Through our research we hope to:

- Gain insights into which locations in New York City are ideal to place first responder stations.
- Determine how heavily staffed the stations should be at certain hours of the day.
- Visualize how correlated different cities and their subsets (boroughs, zip codes, etc.) are to each other. We can use this consistency/inconsistency to help decide how viable similar protocols might be for any given city.

Data

New York Accidents

https://data.cityofnewyork.us/resource/qiz3-axqb.json

This dataset is a breakdown of every vehicle collision in New York City. We filtered this dataset to include only the geographical coordinates, time, and number of casualties of each accident.

San Francisco Accidents

https://data.sfgov.org/resource/vv57-2fgy.json

This dataset is a breakdown of every vehicle collision in San Francisco. We filtered this dataset to include only the times of each collision. In doing so we were able to compare the number of accidents per hour between New York City and San Francisco.

Techniques

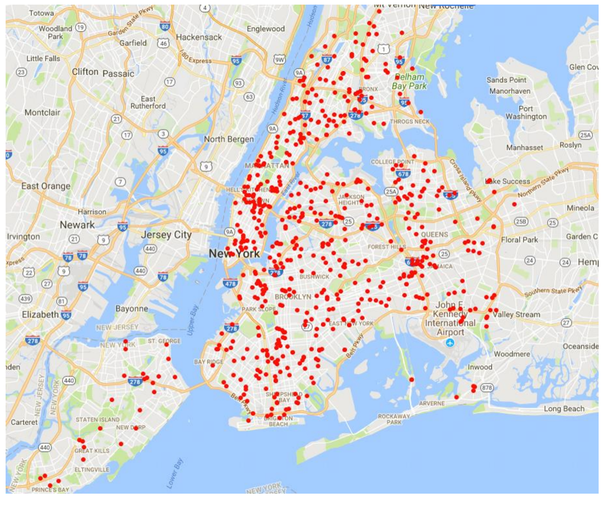
1) **K-Means Constraint Satisfaction**: Our first tool is an implementation of the k-means clustering algorithm with an added constraint satisfaction algorithm. We perform this algorithm on the New York City accident data and find the number of means that are necessary to cluster each accident to within a given distance threshold of each mean. We originally made our metric for constraint satisfaction the maximum distance between the centroid and the points within its cluster, but later decided that an average function would make for a better constraint satisfaction metric because it aligns with the default averaging metric that k-means uses. Using either metric can be a variable function of our design.

2) **Subset Correlations**: In our second area of study we used correlations between two sets of data, as well as subsets of each to determine how correlated the overall and subset data is to the other cities. By putting the results side-by-side in a covariance matrix we are able to visualize how similar their accident rates are at any given hour of the day. We found data on the average number of accidents per hour in San Francisco and for each individual borough in New York. With this data we generated a covariance matrix that shows us how accidents between two areas correlate. We used San Francisco as a baseline because San Francisco is a separate city and we expect that it would be less correlated to a given New York borough than two New York boroughs are to each other. In the table below green signifies that a given New York borough is more correlated to the corresponding borough than it is to San Francisco, and orange signifies that they are less correlated. With this data we can find the higher correlated cities and use them to make predictions or mimic protocol in regards to vehicle accidents.

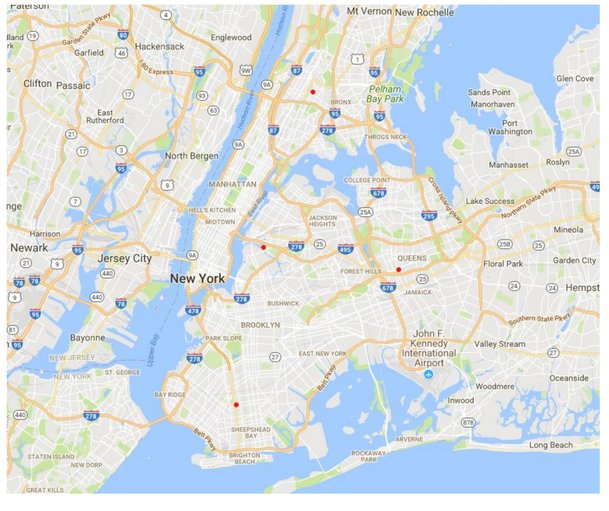
Results

	All of NY	SF	MANHATTAN	BROOKLYN	QUEENS	BRONX	STATEN ISLAND
All of NY	1	0.62433126	0.81537429	0.93950401	0.90077912	0.66983733	0.61138266
SF	0.62433126	1	0.3750099	0.57682681	0.62746252	0.56022917	0.27314589
MANHATTAN	0.81537429	0.3750099	1	0.72725081	0.63151067	0.36555859	0.46856351]
BROOKLYN	0.93950401	0.57682681	0.72725081	1	0.80064724	0.49428523	0.56368812]
QUEENS	0.90077912	0.62746252	0.63151067	0.80064724	1	0.59864817	0.43795898]
BRONX	0.66983733	0.56022917	0.36555859	0.49428523	0.59864817	1	0.38188374]
STATEN ISLAND	0.61138266	0.27314589	0.46856351	0.56368812	0.43795898	0.38188374	1

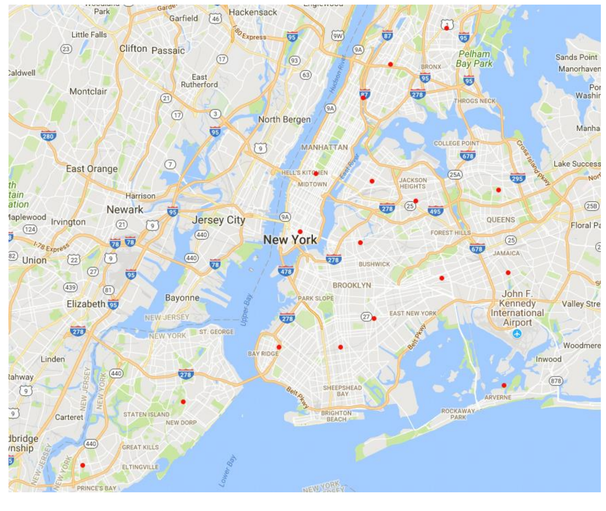
We constructed centroid locations using a constraint satisfaction algorithm whose metric is the average distance between a supposed response center and the location of the grouping's accidents. At every iteration, this algorithm increases the number of means in k-means until the constraint of an average 3 mile distance is satisfied. We ran the algorithm on our dataset, and the result is that it would take a minimum of four response centers (as shown in Map 2) to satisfy the 3 mile average distance constraint. With the 3 mile max distance constraint, the minimum number of response centers required would be 17 (as shown in Map 3).



Map 1: Plot of Vehicle Accidents in New York City



Map 2: Optimal locations for response centers resulting from k-means and a constraint of an average distance of 3 miles



Map 3: Optimal locations for response centers resulting from k-means and a constraint of a worst case distance of 3 miles