CS591: Data Mechanics

Utilizing New York Accident Data to Better Place Response Centers

Alan Burstein • Joseph Caluag • Eric Jacobson • Andrew Quan

12th December, 2017

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 (i.e. police, fire, and ambulance dispatch centers). Using data from various municipal databases as well as external constraints and metrics 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 correlation 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. In our project we obtained data regarding automobile accidents from both New York City and San Francisco. We studied New York City accident data both as one cohesive unit, as well as a disjoint collection of five different boroughs. San Francisco was used as our benchmark because it is an average sized city; smaller than New York City. Through our research we hoped to gain insights into which locations in New York City are ideal to place first response stations. Next, by aggregating the data into 24 one-hour buckets, we aimed to determine how heavily staffed the stations should be at certain hours of the day. Finally, we decided to visualize how correlated different cities and their subsets (boroughs, zip codes, etc.) are to each other. By using the calculated consistencies/inconsistencies city planners can 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 to date in 2017. We filtered this dataset to include only the geographical coordinates, which categorical borough it belongs to, the time, and number of casualties of each accident. Using the whole dataset filtered down to geographical coordinates, we conducted our k-means constraint satisfaction algorithm and using the boroughs, time value, and casualties, we generated the correlations found below.

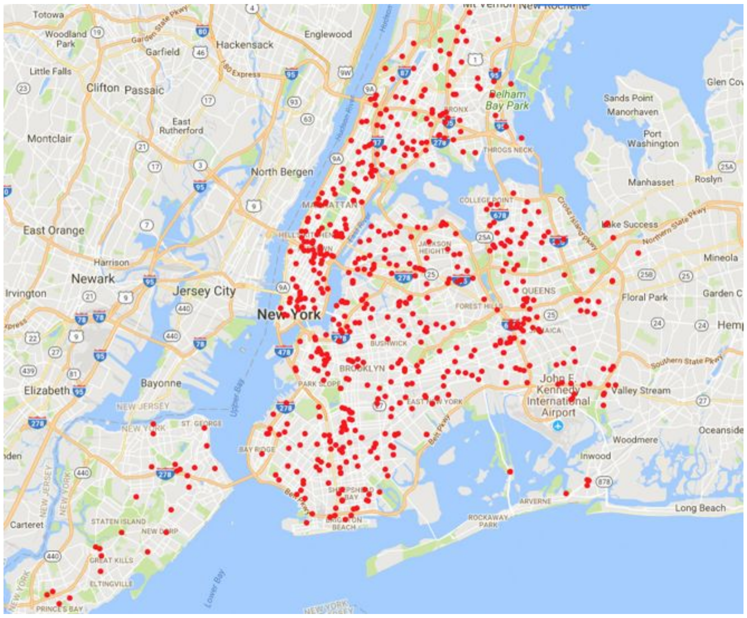
San Francisco Accidents https://data.sfgov.org/resource/vv57-2fgy.json

This dataset is a subset of every vehicle collision in San Francisco between 2006 and 2017. 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 assuming that year of the accident is independent of time of day of the accident. Although this data set spans across more years than the data collected in New York City, the geography and mechanics of the city have not changed substantially enough to result in any non-negligible inconsistencies in our data comparisons.

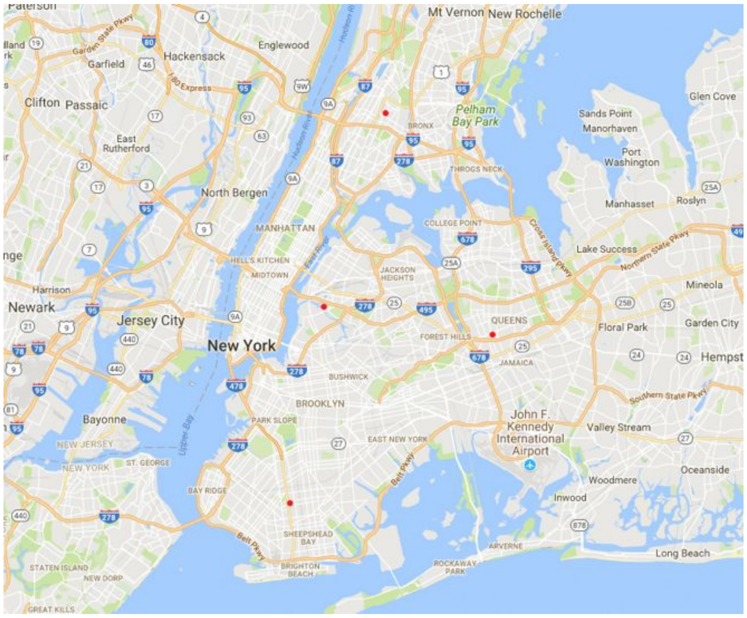
Algorithms:

K-Means with a Constraint Satisfaction

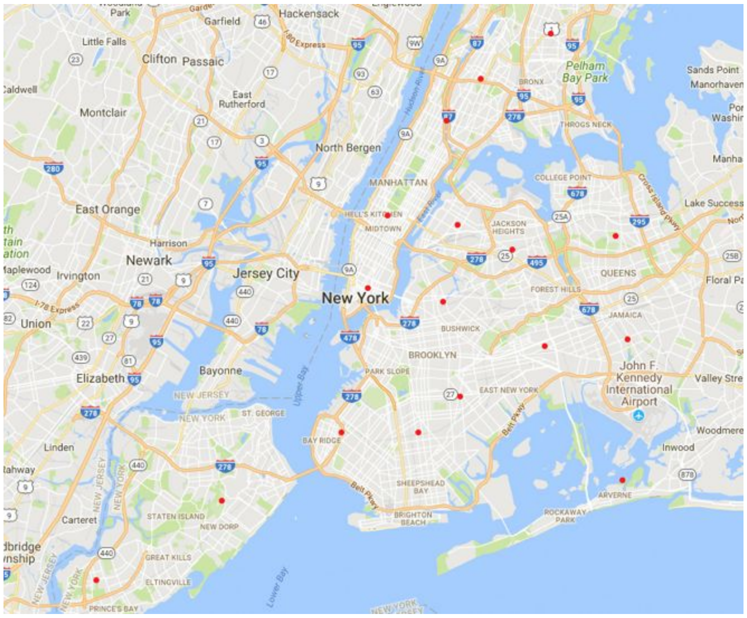
Our first tool is an implementation of the k-means clustering algorithm with an added constraint satisfaction algorithm. We perform this algorithm on the New York City accident data and find the number of means that are necessary to cluster each accident to within a given distance threshold of each mean. 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 is a variable toggle of our design.



Map 1: Plot of Vehicle Accidents in New York City



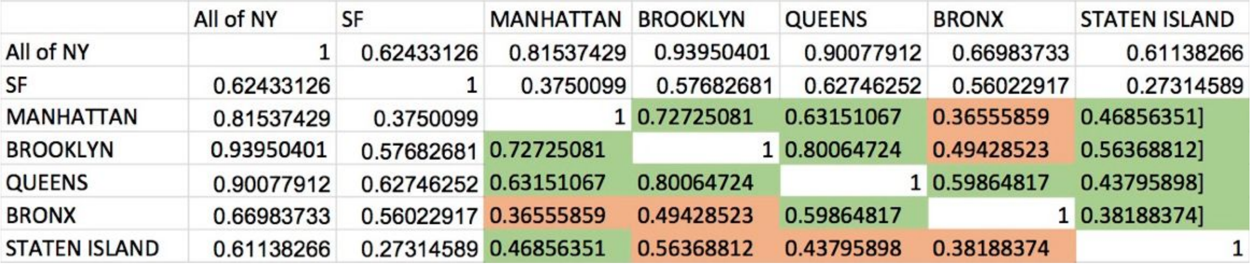
Map 2: Optimal locations for response centers resulting from k-means and a constraint of an average distance of 3 miles



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

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 about the average number of accidents per hour in San Francisco and for each individual borough in New York City. With this data we generate 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.



	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

Conclusions

In our first algorithm 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 group's accidents. At every iteration this algorithm increases the number of means in k-means until the constraint of a n-distance average or worst case is satisfied. We ran the algorithm on our dataset with n=3 miles, 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. In our visualisation we allow the user to play with the variables, freely changing the distance constraint as well as toggling between using that distance as a maximum or an average distance from the centroid. A city planner that may want to predict costs could foreseeably use this tool to calculate how much more money might need to be spent/cut in a development project and the statistical result of their choice.

It should be noted that one very realistic limitation of our findings is that the algorithm takes into account real distance rather than "taxi-cab distance" as a real city network of streets might require. In the same way, the points generated by our algorithms could end up on a location that cannot physically be permuted, e.g a street intersection, a private building, or a historically significant landmark. Despite these limitations, placing response centers in these locations would still be approximately optimal.

In our second algorithm, we found each borough of New York’s correlation with its neighbors. Staten Island was an outlier that was not very similar to Brooklyn, Queens, or the Bronx. This could mean that traffic or accident response policies may need to be more carefully reviewed before a city-wide policy can be applied uniformly to each borough. However, correlation doesn't imply causation, so we cannot truly prove the differences between the time distribution of accidents between boroughs, but only suggest that there exists a correlation or not.

Our project produces a definitive conclusion to optimal placement of response centers *assuming* that any longitude/latitude point is a viable location to build on, but in the future, we could change the algorithm to take into account more realistic variables like street-travel for police vehicles, firetrucks, and ambulances to move through, as well as how to calculate the optimal route from any accident location to the nearest response center. In terms of the correlation matrix, we could also explore how other facets of the accidents are correlated across boroughs, such as the types of cars involved, or develop a metric for how much the accident may have influenced traffic. Another idea that we did not have a chance to pursue, but would be interesting to explore in the future is which metric, average or worst case, might be more appropriate under different circumstances and whether our correlation data can help group the boroughs into either metric.