Ryan Tompkins

Dr. Thomas Kinsman

4/30/18

CSCI-420 Final Project Writeup - NYC Dataset

a) Abstract:

This project involved extensive analysis of the 2018 NYPD Motor Vehicle Collisions dataset taken from [1]. The frequency of collisions was visualized in terms of both geographical location and date. The visualizations demonstrate interesting trends about the area that correspond with the changing weather and holidays. The safest months for different mediums of transportation can be determined by computing the graph's minima and maxima.
Analyzing this dataset with K-means and various values of K revealed some interesting locations in NYC when the centroid coordinates were validated in Google maps. Doing K-means analysis also revealed many erroneous data entries that had invalid latitude and longitude coordinates.

b) Overview/Introduction:

The main value in a project of this type is learning how to handle real world datasets that are often incomplete and large. A solid business case could be made in analyzing NYC's dataset to learn safety trends. That is the direction this project took, as the visualizations try to find any possible relationships between safety, location, and time of year, as well as finding the most dangerous locations in each borough.

c) Background and Research

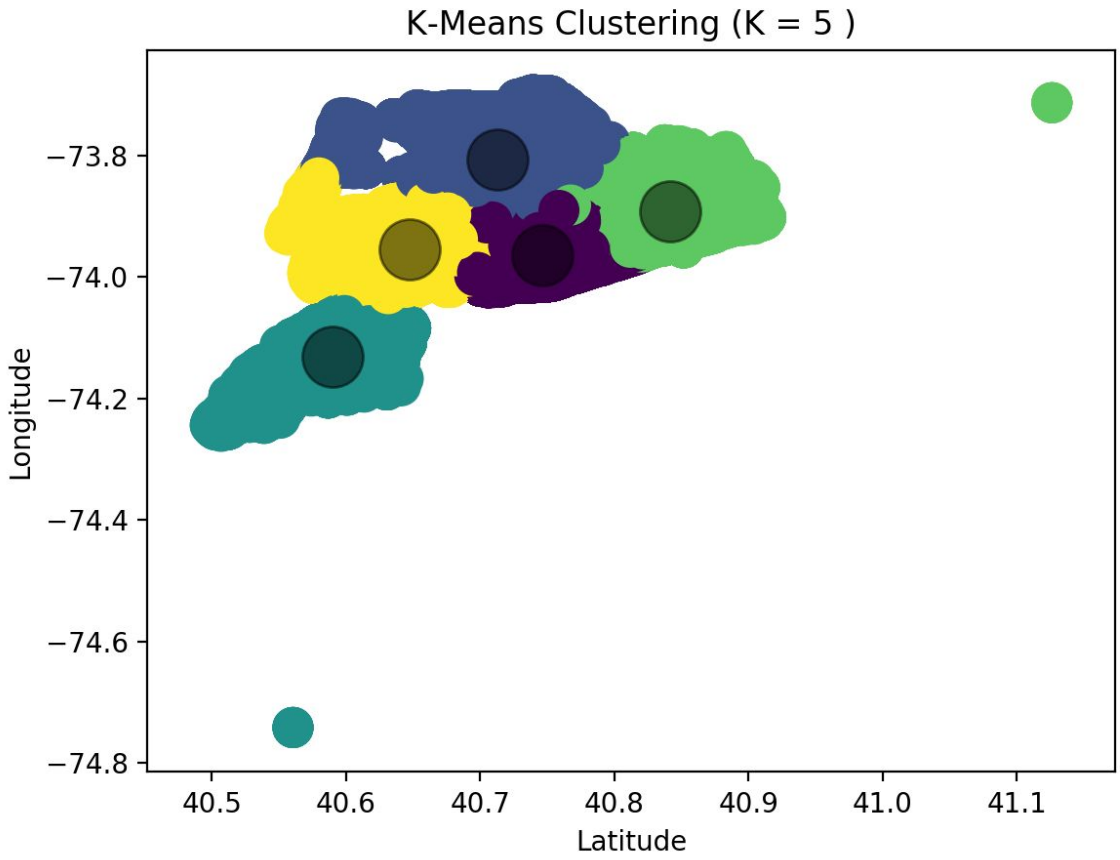
The 2018 NYPD Motor Vehicle Collisions dataset has 1.2 million entries with 29 different attributes. Each row represents a vehicle collision somewhere in New York City. One or more of these 29 attributes could be populated with information relating to an accident such as latitude, longitude, time, date, zip-code, number of injuries, number of deaths, contributing factors to the accident, vehicle type, and many more. It is a good example of a dataset that demonstrates attributes of all types (e.g. nominal → boroughs, interval → latitude/longitude, ratio → number of injuries, categorical → contributing factor).

d) Experiments:

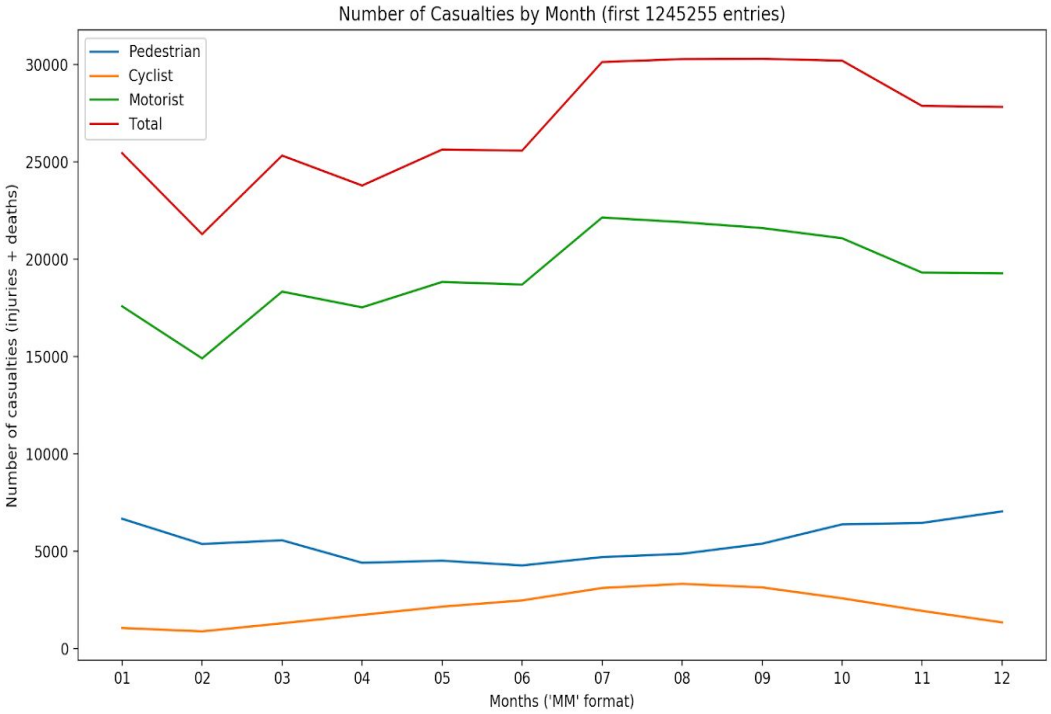
Increasing subsets of the data (e.g. 1000, 10000, 100000…) were visualized using an invented feature called “total casualties” (defined as the total number of injuries and deaths on all columns) on one axis. The total number of casualties per month gave some interesting visualizations and revealed the safest times of the year for each mode of transportation by locating the graphs minima.

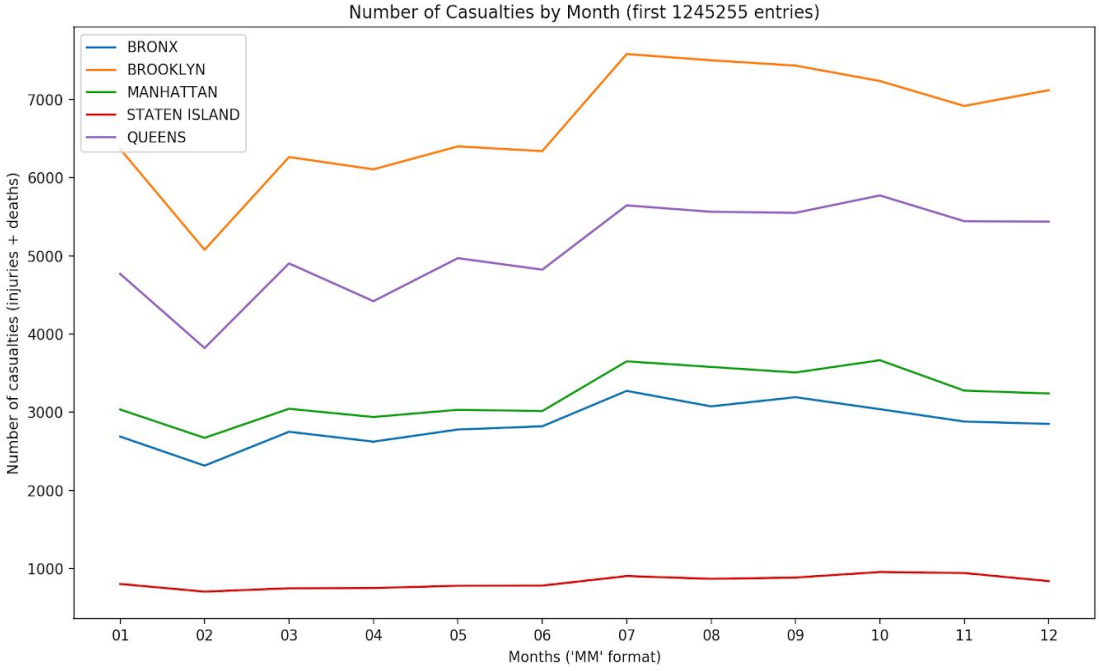
The total casualties for each borough were also graphed throughout the year to see if this would reveal any interesting trends about safety, weather, and temperature for each borough of NYC. These experiments produced graphs where each borough has the same casualty slope, suggesting a higher correlation of season with accidents than borough location. The bias separating the lines corresponds to the expected population differences of the boroughs, with Brooklyn and Manhattan having more people than the other boroughs, and therefore more accidents than a borough like Staten Island. Various visualizations of this type are included in the submission titled *bourough_analysis*.

Finally I performed K-Means analysis on the entire dataset with increasing values of k. I had an intuition that when k was equal to the number of boroughs (k=5), the centroids of the resulting Locations vs Casualty graph would roughly correlate the centers of each borough with the lowest safety (i.e. highest casualty rate). I validated the longitude and latitudes of each centroid in Google maps and included these screenshots in the dropbox submission. The results of K-means analysis, Casualties over Time (transportation modes), and Casualties over Time (boroughs) are shown in the next three figures respectively.



cluster_centers:
[('purple', [40.74629931 -73.96449019]),
('blue', [40.71284494 -73.80677655]),
('cyan', [40.59008857 -74.13111571]),
('green', [40.84163842 -73.89281049]),
('yellow', [40.64759202 -73.95416445])]


Number of Casualties by Month (first 1245255 entries)
Pedestrian
Cyclist
Motorist
Total
Number of casualties (injuries + deaths)
0
5000
10000
15000
20000
25000
30000
01
02
03
04
05
06
07
08
09
10
11
12
Months ('MM' format)


Number of Casualties by Month (first 1245255 entries)
BRONX
BROOKLYN
MANHATTAN
STATEN ISLAND
QUEENS
Number of casualties (injuries + deaths)
1000
2000
3000
4000
5000
6000
7000
01
02
03
04
05
06
07
08
09
10
11
12
Months ('MM' format)

e) Discussions:

Validating some erroneous centroids in K-means analysis with Google maps revealed a few data cleanliness issues with these records. A minor issue was with the entries where the latitude and longitude had zeroes instead of blanks. Since latitude and longitude (0,0) is just west of central Africa on the equator and clearly not in NYC, this was easy to correct. A more frightening example were a variety of entries that had an accident location of Queensboro Bridge, but with an average latitude and longitude of (40.75837 -201.23706) (the scale of both latitude and longitude is [-180,180]). 40 degrees for a latitude makes sense for NYC'c proximity, but the actual longitude of these entries should have been closer to -73 if they were actually intended to be near Queensboro Bridge. This was most likely an error in data entry, and not anomalous data. A data issue somewhere in the middle of these two extremes were the points that initially threw off my k-means analysis by putting centroids in the middle of the Atlantic Ocean. Further analysis of the data indeed revealed many entries with latitudes and longitudes of (40.55, -47.21) and (40.67,-32.77) with zero injuries or deaths. Perhaps the organization received some erroneous reports of taxies crashing in the middle of the atlantic and the data entry employees are performing their due diligence, or perhaps this is another data entry issue.

Questions)

1) *What's the safest month of the year to walk in NYC?* According this dataset, it's June. This was a little counterintuitive, as I would have thought that pedestrian accidents would increase in the summer. The opposite is true, and the number of accidents spike around the holidays.

2) *What is the most dangerous time of year?* According to these visualizations, there is a sharp spike in July, which makes sense as summer would increase outdoor activity and therefore accidents.

3) *How clean is the data?* Not very, as per the discussion above.

4) *Which borough has the most accidents?* Brooklyn; it also has the highest population of all the boroughs with about 2.6 million people, so higher rates of accidents make sense here.

5) *Which borough has the least accidents?* Staten Island; it has the least number of people, so it has less accidents by the same logic as 4. In fact, the relative positions of the lines in the previous figure correspond the population differences of each borough.

The interesting discussion points for the first two graphs include the sharp spike in accidents beginning in March. This probably correlates to the increasing temperatures and and holidays like St. Patrick's Day when more people are outside. The most surprising part of this analysis were the pedestrian and cyclist slopes that had opposite trends from June to August. This seems to suggest that it is slightly safer to walk during the summer months in NYC than ride your bicycle. The decreasing trend of the cyclist slope in the winter, but increasing pedestrian slope probably correlates to fewer cyclists on the road and more tourists populating holiday attractions like Time Square.

K-means analysis appears to reveal the centroids of high casualties as per my intuition (after a few missteps in picking cluster centers in the middle of the ocean and near Africa). The centroids corresponded to downtown Manhattan, a dense highway region of Queens, the midpoint between CUNY College of Staten Island and some NYC parks, the Bronx Zoo, and the

midpoint between some parks and a hospital in Brooklyn. The Google maps view of these locations are included in the dropbox submission for validation. It seems intuitive that areas of higher population, many tourist attractions, and increased foot traffic would result in more accidents, and these five centroids demonstrate one or more of these qualities.

f) Conclusions:

The often quoted adage that “70-80 percent of time in data mining is spent on data cleaning” rings true in this project. It was only after carefully visualization of the data that I was able to catch the various issues in the entries and get results that made any sense. K-means analysis in conjunction with geographic data from Google was essential in the validation process. The visualizations of casualty rates revealed interesting trends in safety throughout the year, some of which were counterintuitive (i.e. accidents for pedestrians being minimal in the summer). I don’t think a company would pay big money for these results, but the experience of data cleaning, visualization, and a practical application of K-Means was quite valuable to me as a student.

References:

[1] https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95