John Hsiao

Data Mining Project

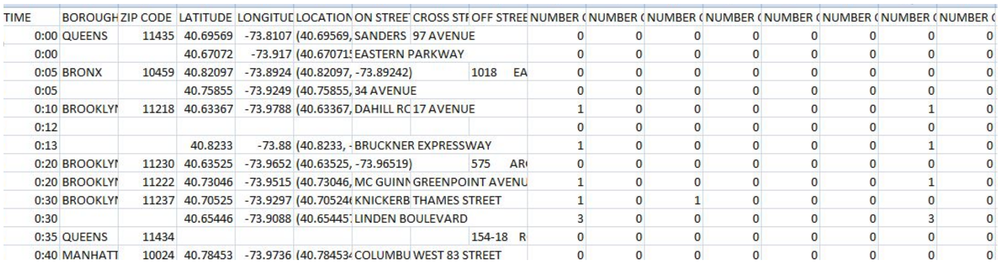
Introduction

In this project, I will be researching the trends that appear in the New York City motor vehicle collisions since 2012. As road safety is an important issue, especially in a big city such as NYC, it is important that we examine this data. Through analysis of past collision history, we can learn more about why collisions happen and try to figure out how to prevent them in the future. Police, the government, and civilians alike should all care about this data in order to provide a safer living environment. The data used in this research project is provided by the NYC government public website, located here. Specifically, it is provided by the NYPD, and it is manually updated every month with information such as the street corner, the vehicle type, the possible traffic infraction, and time of day of the accident.

As outlined above, this project follows CRISP-DM methodology. From a business perspective, companies would care about this information in a variety of ways, such as a safer location, selling traffic safety items, and etc. We must understand, prepare, and model the data in order to analyze it and try to see if a pattern of traffic accidents exist. From that analysis, we can try to evaluate a solution and deploy it to the public.

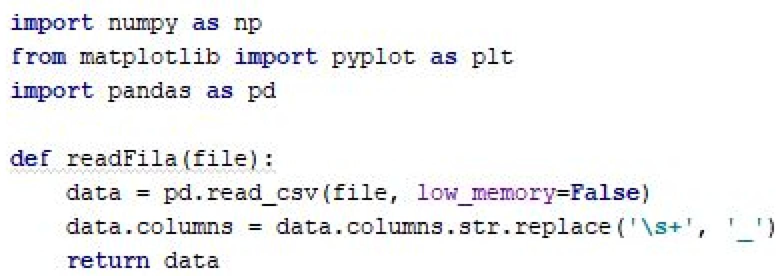
Project Research/Background information

This project provides us with a huge csv file containing traffic accidents that occur in NYC. The data is very comprehensive, including superfluous information such as the latitude and longitude of where the accident occurred. Otherwise, I think there's tons of useful information, such as the types of vehicles that were involved in the accident, the cause of the accident, and stuff that could lead to finding out the specific types of areas in the city that is more accident prone. The New York City Council passed a law in 2011 that this data must be collected, and so the NYC police department provides this data every year, and it is located at: https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95.



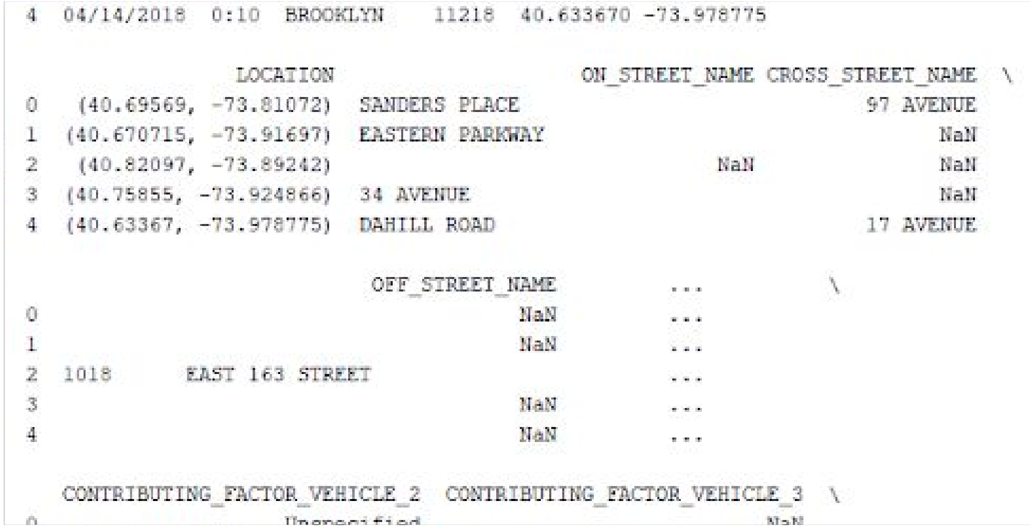
TIME	BOROUGH	ZIP CODE	LATITUDE	LONGITUD	LOCATION	ON STREET	CROSS STR	OFF STREE	NUMBER (	NUMBER (	NUMBER (	NUMBER (	NUMBER (	NUMBER (	NUMBER (	NUMBER (
0:00	QUEENS	11435	40.69569	-73.8107	(40.69569,	SANDERS	97 AVENUE		0	0	0	0	0	0	0	0
0:00			40.67072	-73.917	(40.67071	EASTERN PARKWAY			0	0	0	0	0	0	0	0
0:05	BRONX	10459	40.82097	-73.8924	(40.82097, -73.89242)			1018 EA	0	0	0	0	0	0	0	0
0:05			40.75855	-73.9249	(40.75855,	34 AVENUE			0	0	0	0	0	0	0	0
0:10	BROOKLY	11218	40.63367	-73.9788	(40.63367,	DAHILL RC	17 AVENUE		1	0	0	0	0	0	1	0
0:12									0	0	0	0	0	0	0	0
0:13			40.8233	-73.88	(40.8233, -	BRUCKNER EXPRESSWAY			1	0	0	0	0	0	1	0
0:20	BROOKLY	11230	40.63525	-73.9652	(40.63525, -73.96519)			575 AR	0	0	0	0	0	0	0	0
0:20	BROOKLY	11222	40.73046	-73.9515	(40.73046,	MC GUINN	GREENPOINT AVENU		1	0	0	0	0	0	1	0
0:30	BROOKLY	11237	40.70525	-73.9297	(40.70524	KNICKERB	THAMES STREET		1	0	1	0	0	0	0	0
0:30			40.65446	-73.9088	(40.65445	LINDEN BOULEVARD			3	0	0	0	0	0	3	0
0:35	QUEENS	11434						154-18 R	0	0	0	0	0	0	0	0
0:40	MANHATT	10024	40.78453	-73.9736	(40.78453	COLUMBU	WEST 83 STREET		0	0	0	0	0	0	0	0

However, its comprehensiveness is also a slight problem with the data. As the data set is so big and unorganized, it takes a lot of time to just organize it enough so that the data can be analyzed. Also, since the data is so big, it's hard to verify a hypothesis, so they could be misled due to confirmation bias when in fact the data actually pointed to something else, due to people tending to see what they seek out. This also means that it's difficult to validate the results, as it is impossible to validate results by hand in a timely matter. A quick google search shows that many students have done reports based off of this data. Many of them faced issues such as the data not being very clean, and so they spent the first section of their code cleaning it up so that the data is easier to use in code. The algorithms I used were clustering, k-NN, and k -Means.



```
import numpy as np
from matplotlib import pyplot as plt
import pandas as pd

def readFila(file):
    data = pd.read_csv(file, low_memory=False)
    data.columns = data.columns.str.replace('\s+', '_')
    return data
```

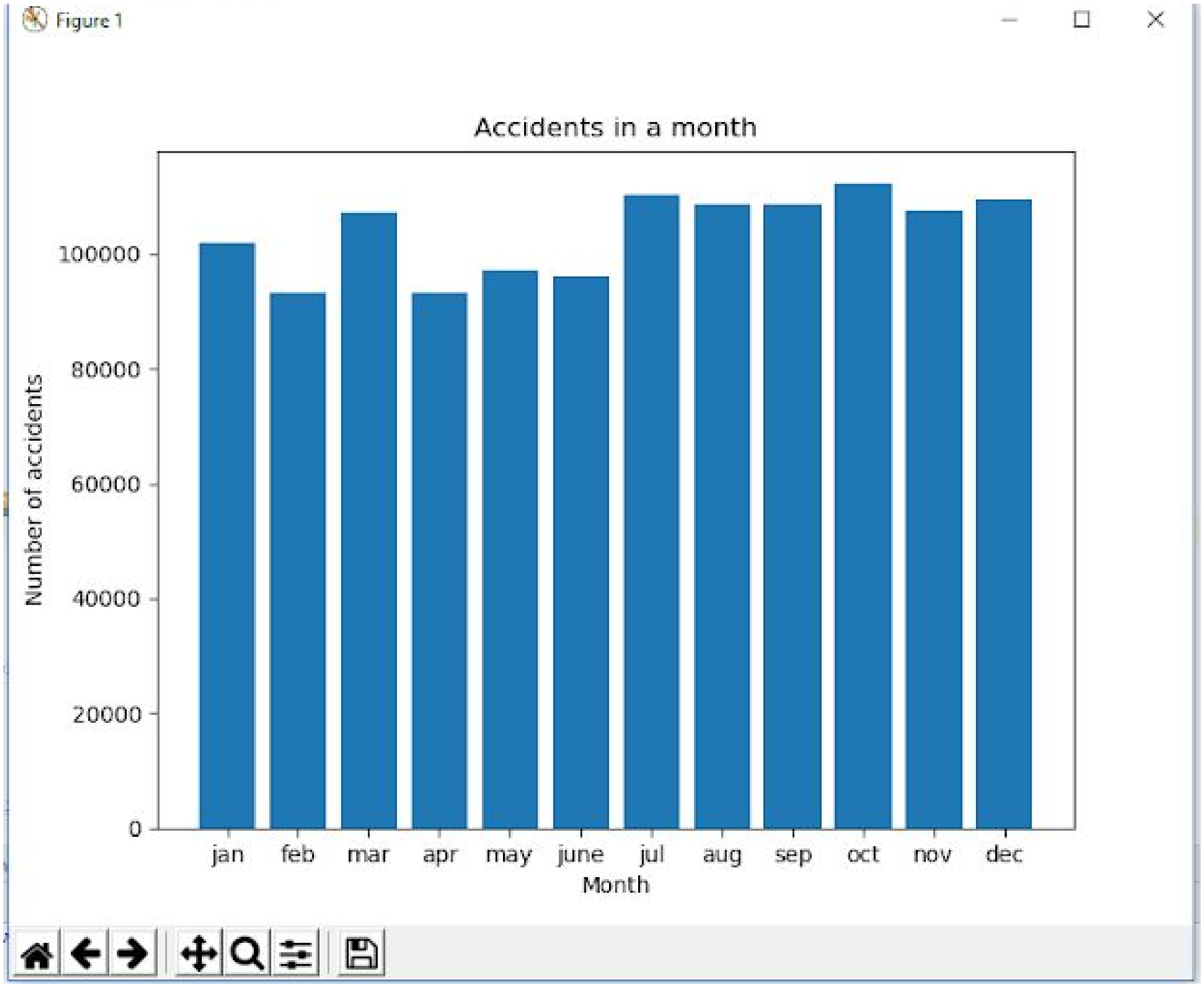
I started out by importing the three modules I thought most important to analyze the data, which was pandas, matplotlib, and numpy. I then read it into a DataFrame, and then cleaned the data so that it would be accessible through the DataFrame.



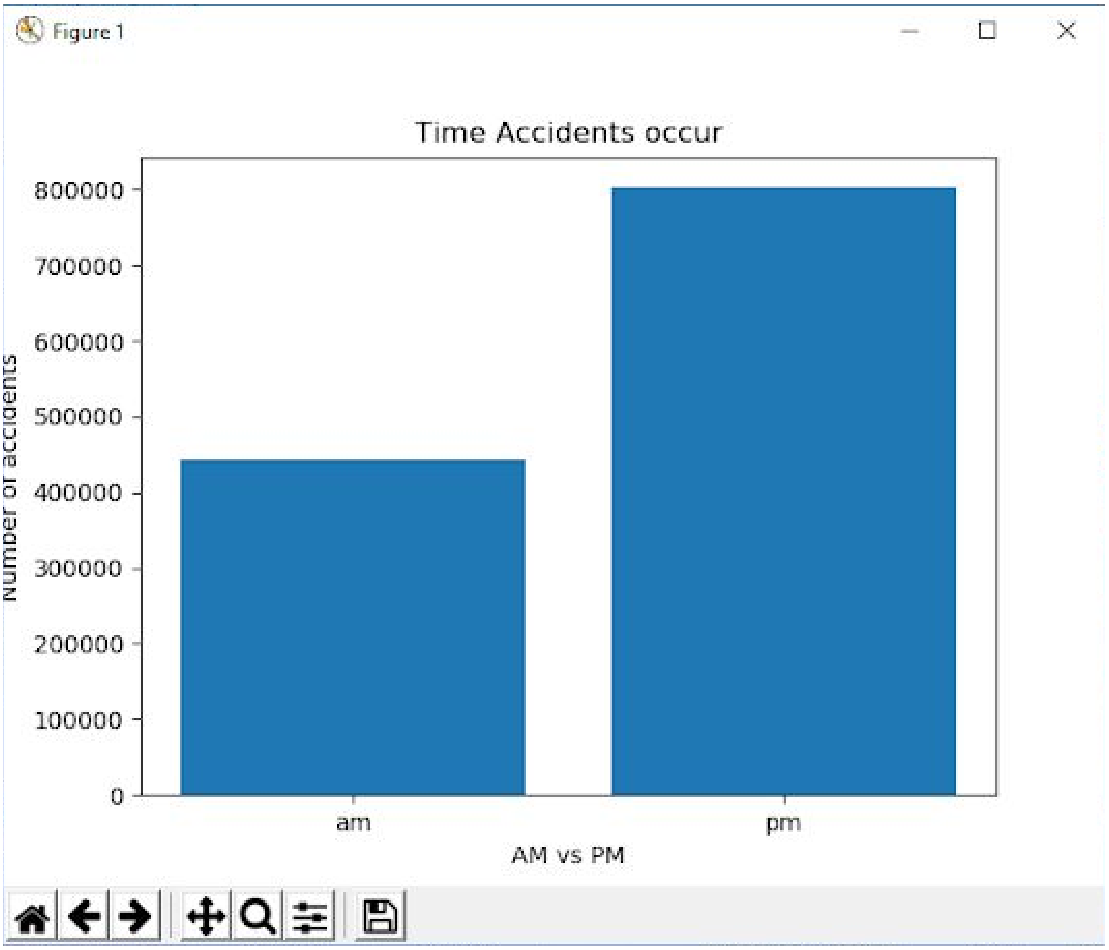
```
4  04/14/2018  0:10  BROOKLYN    11218  40.633670 -73.978775

                 LOCATION                    ON_STREET_NAME CROSS_STREET_NAME  \
0   (40.69569, -73.81072)  SANDERS PLACE                            97 AVENUE
1  (40.670715, -73.91697)  EASTERN PARKWAY                                NaN
2   (40.82097, -73.89242)                               NaN               NaN
3  (40.75855, -73.924866)  34 AVENUE                                      NaN
4  (40.63367, -73.978775)  DAHILL ROAD                              17 AVENUE

                     OFF_STREET_NAME        ...          \
0                                NaN        ...
1                                NaN        ...
2  1018      EAST 163 STREET                ...
3                                NaN        ...
4                                NaN        ...

  CONTRIBUTING_FACTOR_VEHICLE_2  CONTRIBUTING_FACTOR_VEHICLE_3  \
0                   Unspecified                            NaN
```

I added an underscore between the column headers so that the data would be accessible through code, as it’s not possible to access “data.ON STREET NAME”.

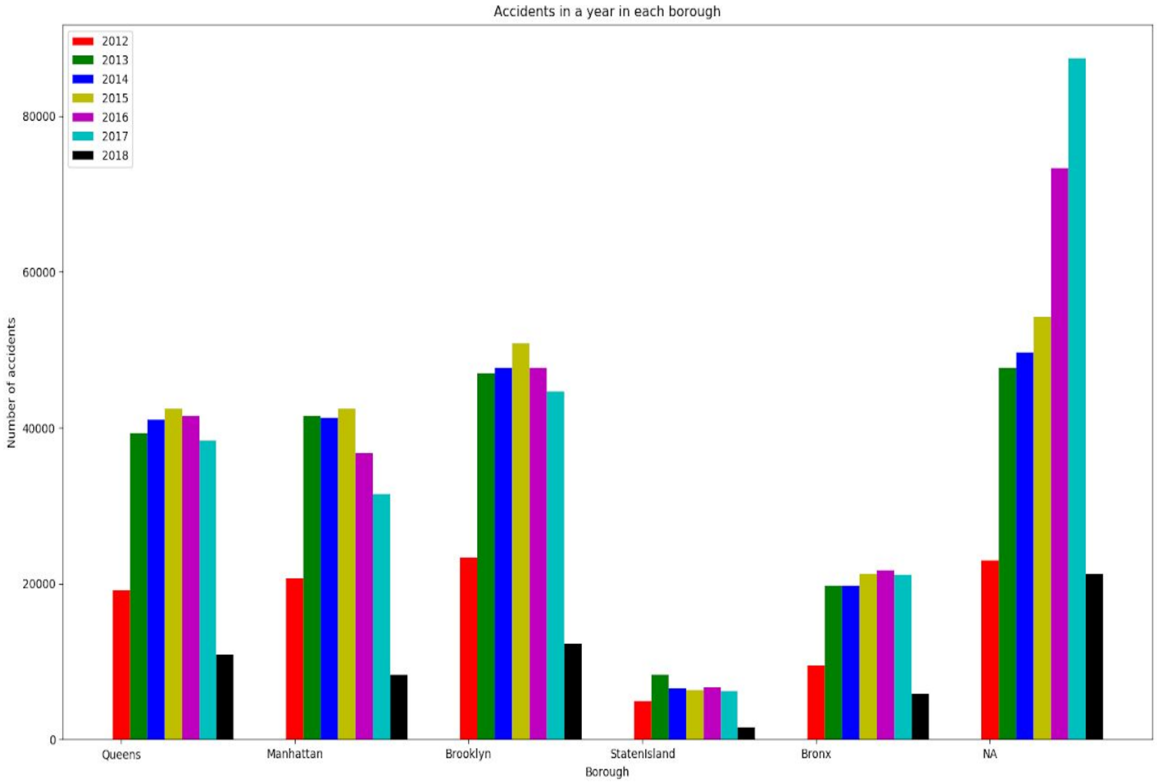
Time Based Metrics



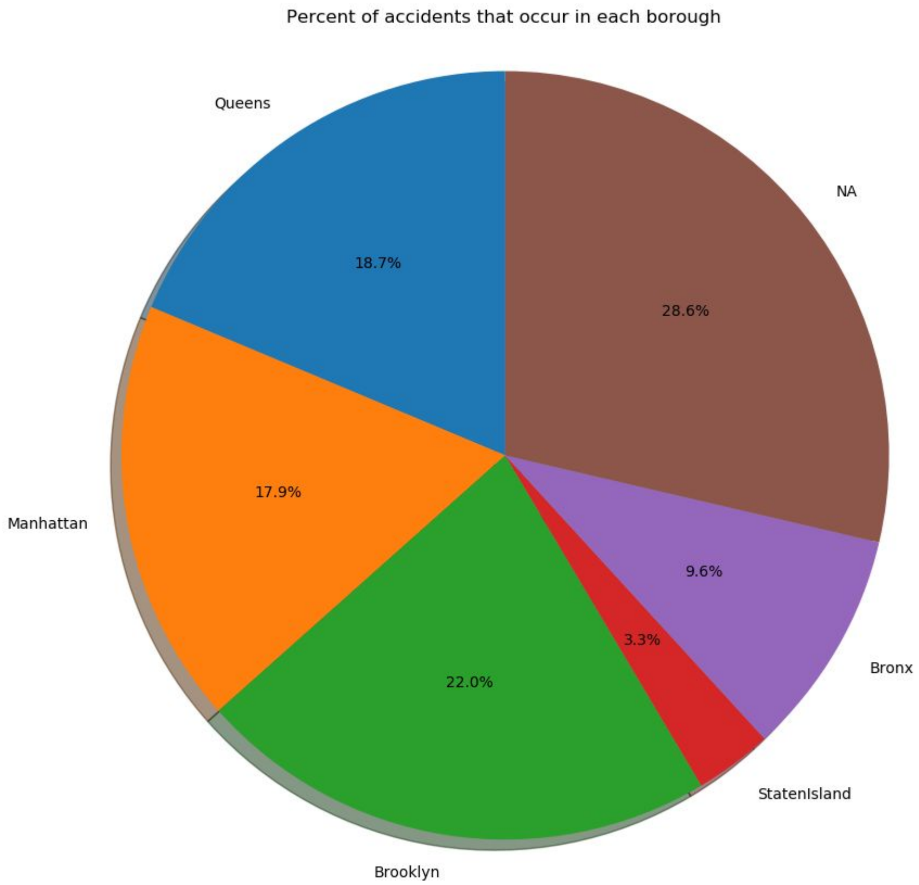
The data shows that the month does not necessarily correlate with how many accidents there are going to be. However, other than March, you can see a small spike in accidents that occurs starting from July all the way until December. This might be because that's when the weather really starts getting nice in New York City, so that's when pedestrians and bikers start becoming more common, causing confusion to drivers. And the reason that it lasts until winter might be because the roads start becoming more dangerous but people aren’t used to the shift yet, so accidents remain higher until drivers adjust to the different weather and road hazards.



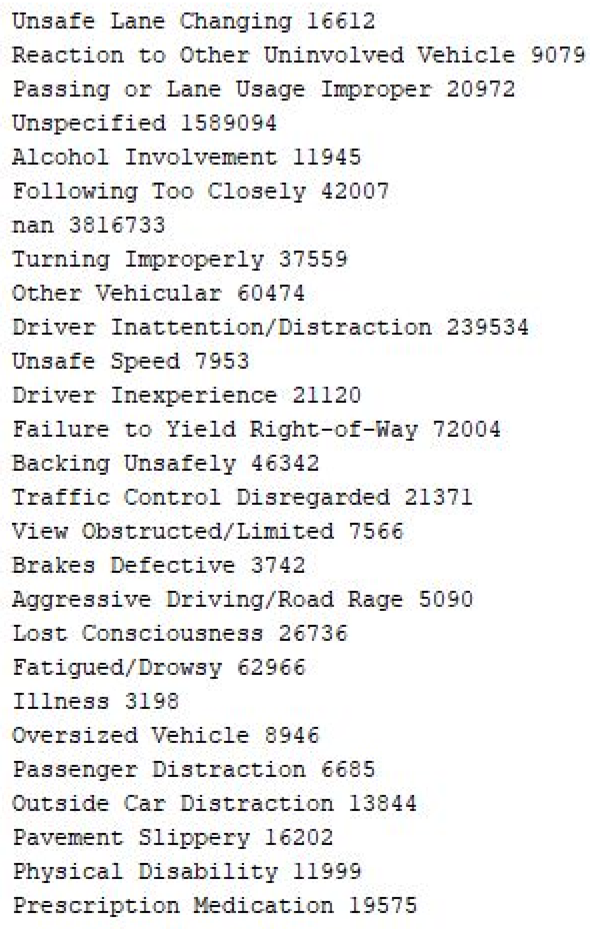
The data clearly shows a large discrepancy between accidents that occur in the AM as opposed to PM; there is nearly twice the amount of accidents that occur in the PM. This could be due to a variety of reasons. For starters, midnight to early morning is usually a period of sleep, so there are not as many pedestrians or drivers on the roads. While drivers may be groggy in the morning, they probably know of that fact and are more alert on the road. On the other hand, noon and evening are prime hours of the day where people are out and about. Visibility is poorer at night, and people are probably in a rush to get home, as well as general disdain for traffic rules by many people throughout the day.



Other than 2012 and 2018 being outliers due to not having a full year of data to work with, the data is pretty consistent across each borough. As one might have guessed, the largest borough by population, Brooklyn, also consistently has the most amount of accidents occur, while Staten Island, which has by far the smallest population, has the fewest number of accidents. The following pie chart visualizes the total percentage of accidents that occur in each borough.

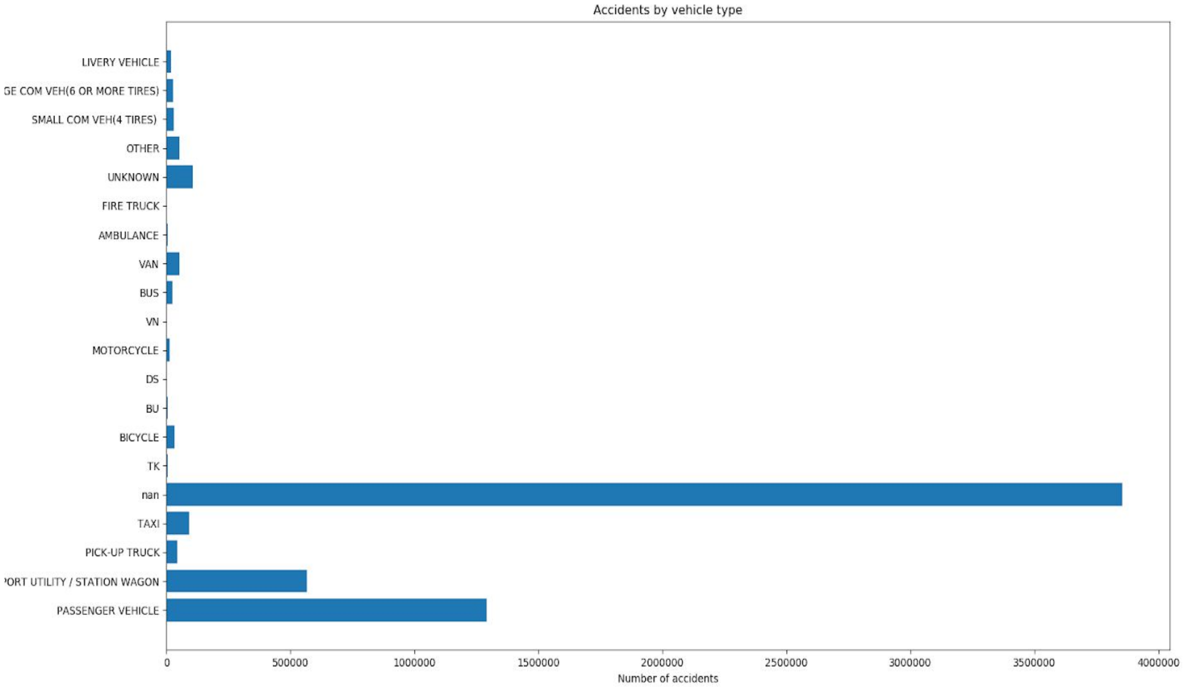


This pie chart compared to the previous bar graph demonstrates the disparity in accident occurrences. Interestingly, Manhattan has even less accidents than Queens, despite having much more people squeezed into each square mile, which shows that one of the more important factors in having a lot of accidents is the number of people that live in each borough.



```
Unsafe Lane Changing 16612
Reaction to Other Uninvolved Vehicle 9079
Passing or Lane Usage Improper 20972
Unspecified 1589094
Alcohol Involvement 11945
Following Too Closely 42007
nan 3816733
Turning Improperly 37559
Other Vehicular 60474
Driver Inattention/Distraction 239534
Unsafe Speed 7953
Driver Inexperience 21120
Failure to Yield Right-of-Way 72004
Backing Unsafely 46342
Traffic Control Disregarded 21371
View Obstructed/Limited 7566
Brakes Defective 3742
Aggressive Driving/Road Rage 5090
Lost Consciousness 26736
Fatigued/Drowsy 62966
Illness 3198
Oversized Vehicle 8946
Passenger Distraction 6685
Outside Car Distraction 13844
Pavement Slippery 16202
Physical Disability 11999
Prescription Medication 19575
```

Unsurprisingly, the most common causes of an accident are when a driver is inattentive or has poor driving skills. The other big reason seems to be related to health, such as drowsy drivers or having poor mental health and losing consciousness. A good segue could be to research and try to prevent drivers from taking the road when they are not alert enough to drive.



Almost all accidents seem to occur in a passenger vehicle, or a vehicle that is transporting someone. This coupled with the realization that most accidents are caused by negligence or health issues seem to point to some sort of correlation between civilian drivers and drowsiness or other health issues, which could be a good research direction. On the other hand, there is a surprisingly large amount of accidents where the vehicle type is not available, but I suspect that is due to my own lack of properly cleaning the data to remove accidents where not all vehicles are involved.

Final Project Report

As I expected, the algorithms I ended up using in order to visualize the data were clustering, k-NN, and k-Means. I faced a lot of challenges and difficulties along the way, mostly stemming from my inexperience using the matplotlib and pandas libraries of

Python. I spent the most amount of time trying to learn how to properly code the graphs so that they would read in the data I had gathered correctly. I also had trouble sometimes making the font fit properly on the graph, or having the font show big enough so as to be legible. My next biggest problem was figuring out the proper way to gather the data given the massive dataset we were given. I realized I had to clean the data by stripping whitespace from the column headers, as well as grouping what I thought were the most important factors together each time I wanted to test something new. The algorithm that worked the best for all of this was clustering, as it was the simplest way to visualize the data. It easily let you group together accidents in terms of month or borough or etc. I thought that this was a supremely interesting and fun project to work on. It showed that I had learned a lot from this class in terms of gathering data and learning how to visualize and analyze said data. From what I have seen in the data, government officials and business savvy people looking to make a buck alike might want to think of ways to prevent drowsy driving, as well as regulate the types of drivers that are on the road (so we can lower the amount of accidents that occur from negligent drivers!).