W200 Project 2 Proposal: Exploring Collision Data in NYC for Prevention

Team Members: David Gamez and Ernesto Del Valle

Section // Team: Prof. Richard Huntsinger Section 6 (Thursday, 630pm PST) // Team 2

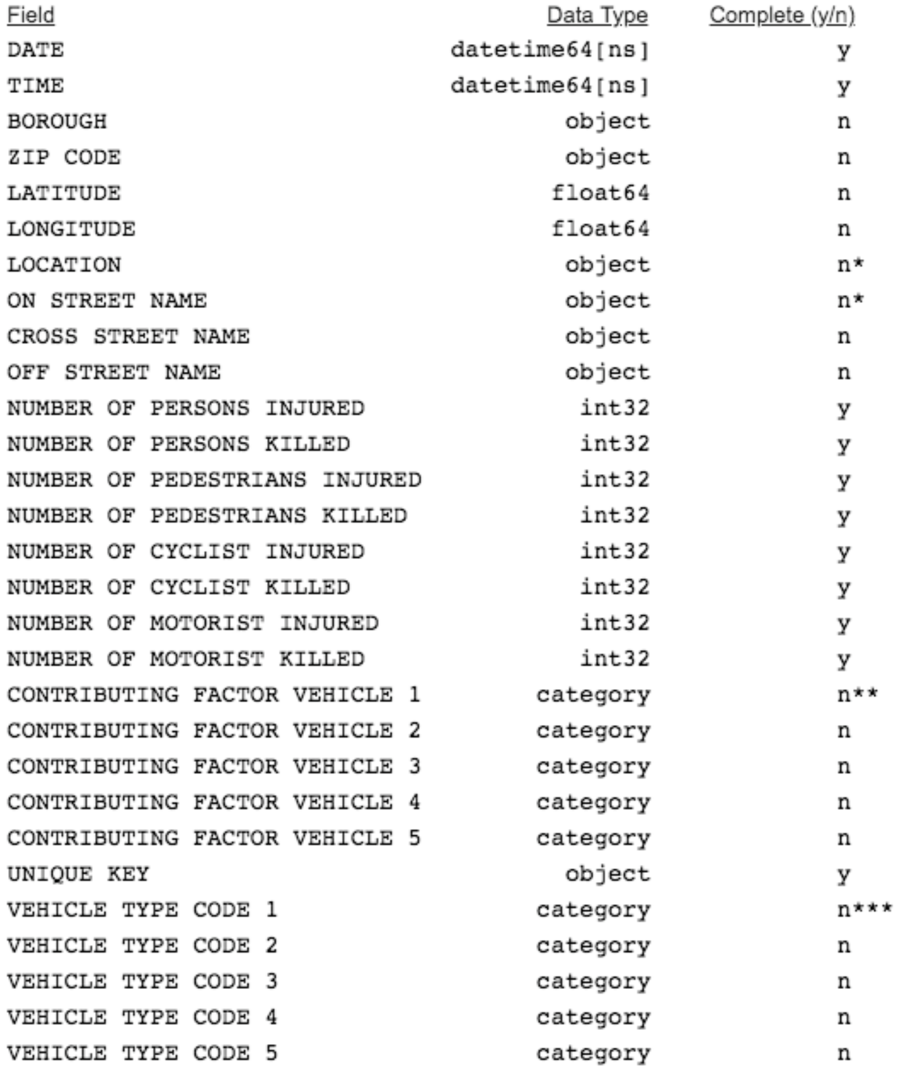
Repository Link: https://github.com/UCB-INFO-PYTHON/W200_F18_Pr2_S6_team2REPO

Data Set Link: https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95

Data Set Description: collisions records in NYC broken down by:

- (1) *Date and Time*
- (2) *Location*
- (3) *Injury types (severity)*
- (4) *Contributing factors* (when identified)
- (5) *Vehicles involved* (when identified)

Data Set Structure: collisions since March 2014, mapping to 1,385,480 entries with 29 columns:



Field	Data Type	Complete (y/n)
DATE	datetime64[ns]	y
TIME	datetime64[ns]	y
BOROUGH	object	n
ZIP CODE	object	n
LATITUDE	float64	n
LONGITUDE	float64	n
LOCATION	object	n*
ON STREET NAME	object	n*
CROSS STREET NAME	object	n
OFF STREET NAME	object	n
NUMBER OF PERSONS INJURED	int32	y
NUMBER OF PERSONS KILLED	int32	y
NUMBER OF PEDESTRIANS INJURED	int32	y
NUMBER OF PEDESTRIANS KILLED	int32	y
NUMBER OF CYCLIST INJURED	int32	y
NUMBER OF CYCLIST KILLED	int32	y
NUMBER OF MOTORIST INJURED	int32	y
NUMBER OF MOTORIST KILLED	int32	y
CONTRIBUTING FACTOR VEHICLE 1	category	n**
CONTRIBUTING FACTOR VEHICLE 2	category	n
CONTRIBUTING FACTOR VEHICLE 3	category	n
CONTRIBUTING FACTOR VEHICLE 4	category	n
CONTRIBUTING FACTOR VEHICLE 5	category	n
UNIQUE KEY	object	y
VEHICLE TYPE CODE 1	category	n***
VEHICLE TYPE CODE 2	category	n
VEHICLE TYPE CODE 3	category	n
VEHICLE TYPE CODE 4	category	n
VEHICLE TYPE CODE 5	category	n

* Whenever 'ON STREET'/'CROSS STREET' blank, we have 'LOCATION' or 'OFF STREET'. Hence, there is always some indication of (1) *Location*. Also 'LOCATION' follows a "('LATITUDE', 'LONGITUDE')" format.

** There is generally at least one 'CONTRIBUTING FACTOR VEHICLE'.

*** Also at least one 'VEHICLE TYPE CODE' present, with frequency matching that of ''CONTRIBUTING FACTOR VEHICLE'.

Questions for Exploration:

Some of the questions we expect to focus on include:

- What borough carries the largest injuries/fatalities? Normalized by population?
- What are the top zip codes or even particular crossings to be aware of? Can policy/adjustments on root causes (from contributing factor descriptions) be implemented on these?
- On the opposite end, which are safest? Why? Can we learn something from them?
- Contributing factor and vehicle-wise, what are the most commonly identified as contributing to/involved in an accident?
- From a time series perspective, is there seasonality in overall injuries/fatalities? How about seasonality per type of injured subject (pedestrian/cyclist/motorist)? At a factor level (e.g. are mobile phones increasingly prevalent as a reason)?

Overarching theme/objective:

- Can we understand the structure of collisions in NYC to better prevent these either through policy or awareness?

Main Variables Involved and Supplemental Data (if any):

In order to effectively explore the above questions we would prioritize the following variables:

- DATE/TIME => For any seasonality analysis
- BOROUGH => As a more aggregated measure of location
- ZIP CODE => As the most reasonably standardized measure of location at a lower level
- CROSS STREET NAME => Digging further into particular zip codes if applicable
- Total Persons Injured (sum of columns) => Measuring total severity of collisions
- Total Persons Killed (sum of columns) => Measuring total severity of collisions
- CONTRIBUTING FACTOR VEHICLE 1 => Analyzing main reason for collision
- VEHICLE TYPE CODE 1 => Analyzing vehicle type involved

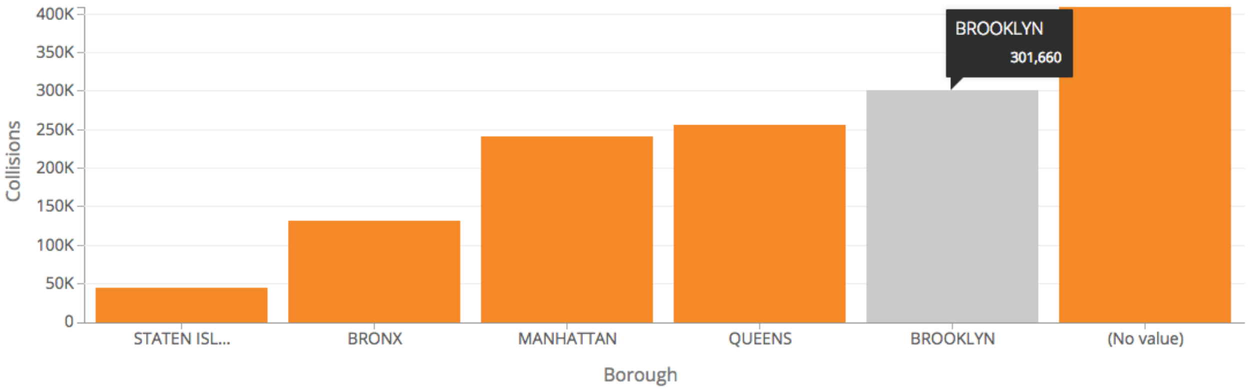
As it pertains to supplemental data, we don't expect to bring in much additional. Maybe very basic ones like population per borough (demographic) in order to normalize some of the statistics

Planned Structure for Final Report:

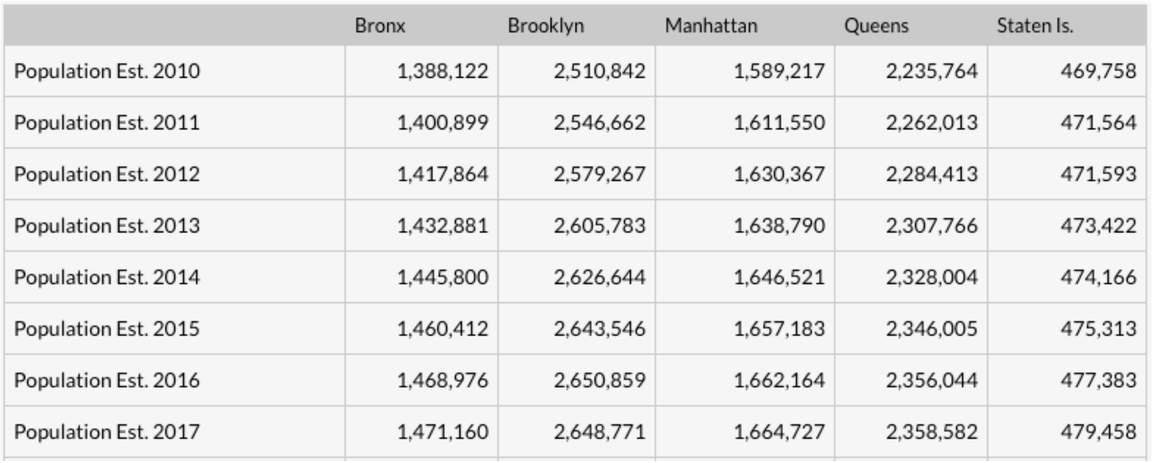
- Introduction
- Exploratory Data Analysis:
  - [1] Frequency of collisions at different measures of location (borough > zip code > cross street)
  - [2] Measuring severity and relationship to location-frequency from [1]
  - [3] Identifying potential root causes/particular vehicle types involved where accidents were most severe/frequent per [1] and [2]
  - [4] Seasonality as part of partial cause, time/subject/factor-wise
- Potential Recommendations
- Conclusion

Some Initial Plots:

- Collisions per BOROUGH:

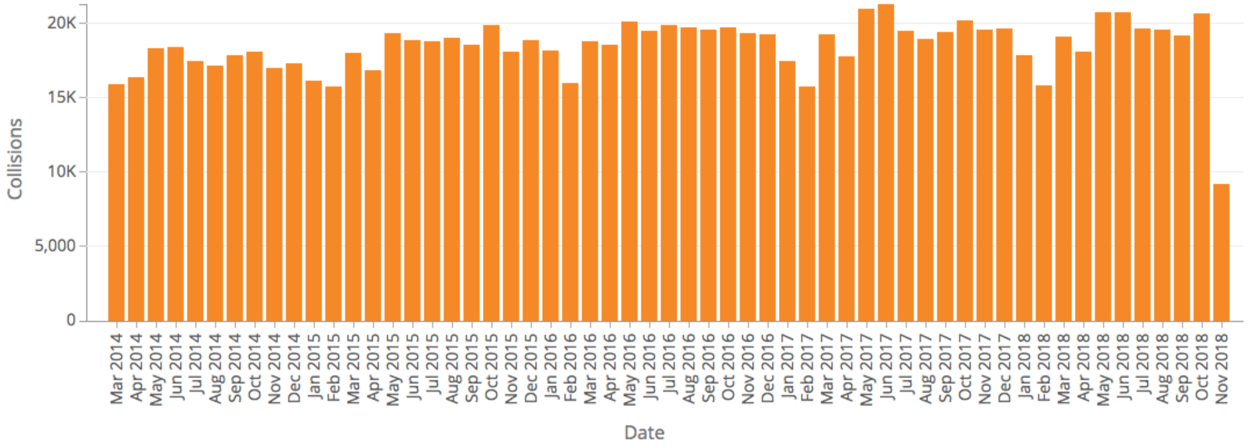


- If wanted to normalize by Population (link) as simple supplemental data (Brooklyn > Queens > Manhattan > Bronx > Staten Island), we can bring in the following. On a per capita basis Manhattan actually on top:



	Bronx	Brooklyn	Manhattan	Queens	Staten Is.
Population Est. 2010	1,388,122	2,510,842	1,589,217	2,235,764	469,758
Population Est. 2011	1,400,899	2,546,662	1,611,550	2,262,013	471,564
Population Est. 2012	1,417,864	2,579,267	1,630,367	2,284,413	471,593
Population Est. 2013	1,432,881	2,605,783	1,638,790	2,307,766	473,422
Population Est. 2014	1,445,800	2,626,644	1,646,521	2,328,004	474,166
Population Est. 2015	1,460,412	2,643,546	1,657,183	2,346,005	475,313
Population Est. 2016	1,468,976	2,650,859	1,662,164	2,356,044	477,383
Population Est. 2017	1,471,160	2,648,771	1,664,727	2,358,582	479,458

- Gaging seasonality through Collisions per DATE (by month):


20K
15K
10K
Collisions
Mar 2014
Apr 2014
May 2014
Jun 2014
Jul 2014
Aug 2014
Sep 2014
Oct 2014
Nov 2014
Dec 2014
Jan 2015
Feb 2015
Mar 2015
Apr 2015
May 2015
Jun 2015
Jul 2015
Aug 2015
Sep 2015
Oct 2015
Nov 2015
Dec 2015
Jan 2016
Feb 2016
Mar 2016
Apr 2016
May 2016
Jun 2016
Jul 2016
Aug 2016
Sep 2016
Oct 2016
Nov 2016
Dec 2016
Jan 2017
Feb 2017
Mar 2017
Apr 2017
May 2017
Jun 2017
Jul 2017
Aug 2017
Sep 2017
Oct 2017
Nov 2017
Dec 2017
Jan 2018
Feb 2018
Mar 2018
Apr 2018
May 2018
Jun 2018
Jul 2018
Aug 2018
Sep 2018
Oct 2018
Nov 2018
Date