Exploratory Data Analysis on Collision Records in NYC (July 2012 - November 2018)

Location, Factors and Time Series Perspectives

U.C. Berkeley MIDS, W200: Python for Data Science, Prof. Richard Huntsinger
David Gamez and Ernesto Del Valle
December 13th, 2018
Presentation link (PDF in Repo too)

Introduction

On March 18th, 2018, the first pedestrian fatality associated with Level 3 self-driving technology took place in Tempe, AZ. This accident garnered considerable media attention and raised concerns about the viability of the technology, and in particular whether Uber was being too aggressive in their development with Volvo.

Our project investigates ~1.3mm collision records in New York City (NYC), with the objective of better understanding human error, before autonomous vehicles intervene and potentially ameliorate the issue eventually. We dissect our data set at hand across three dimensions in particular: **[1]** Location, **[2]** Factors and **[3]** Time Series. To guide the reader through our discussion of these, we groups these three factors into a logical flow through four questions:

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

Before jumping right into our questions we present an overview of our main data set and a summary of how we prepared it for exploration.

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

Dataset Description: created on April 28th, 2014, our dataset contains collisions records in NYC spanning from July 2012 through mid-November 2018 with columns that fall under the following categories:

1. *Date and Time*
2. *Location* (borough, zip code, GPS coordinates)
3. *Injuries* (severity on number of pedestrians/cyclists/motorists injured)
4. *Fatalities* (severity on number of pedestrians/cyclists/motorists killed)
5. *Contributing Factors* (when identified)
6. *Vehicles Involved* (when identified)

The NYC Government updates the data every month, as reviewed by the TrafficStat Unit, before posting it on the New York Police Department (NYPD)'s website.

Dataset Structure: there are 1,385,480 entries with 29 columns in our dataset. We specified preferable data types at import with the following results (Table 1):

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

Table 1: Data Type Structure at Import.

The referenced website contains descriptions of these variables under a spreadsheet named "NYPD_Collision_DataDictionary.xlsx". These variables represent the building blocks of our three pillars of exploratory analysis **[1]** Location, **[2]** Factors and **[3]** Time Series.

Exploratory Data Analysis (EDA)

Initial Exploration for Completion and Consistency

After some initial investigation of every variable and sampling of observations, we performed multiple sanity checks on the dataset. Our main findings include:

1. Very few descriptions are available in the "NYPD_Collision_DataDictionary.xlsx" dictionary and the names of the variables do not always correspond to the ones in the dataset, meaning the dictionary is outdated
2. *Borough* had null values on 406,750 observations. However, in 180,389 of these we did get GPS coordinates through *Location*. Whenever *Borough* was missing, *Zip Code* was missing too
3. Not all 61 *Contributing Factors* were unique as 3 in fact had typos and hence were repeated (e.g. "Illness" vs. "Ilnes") or contained case sensitive differences ("Cell Phone (hand-Held)" vs. "Cell Phone (hand-held)" and "Drugs (Illegal)" vs. "Drugs (illegal)"
4. On 44.9% of the observations, *Contributing Factor for Vehicle 1* is "Unspecified"
5. The observations for variable *Vehicle Type Code 1* to *Vehicle Type Code 5* did not follow any procedure and appeared as quite messy. For example, in some observations it contains the brand of the car (e.g. "Ford", "Volkswagen", etc.) and in others it contains the type of vehicle ("Sedan", "Truck", "Fire", etc.). Even worse, the same type of vehicle is often named in several different ways (e.g. 12 different ways to specify "Ambulance" as 'ABULA', 'AM', 'AMABU', 'AMB', 'AMBU', 'AMBUL', 'AMBULANCE', 'AMbul', 'Ambul', 'Ambulance', 'ambu', 'ambul'). This is the reason why there are 459 unique vehicle types in the data set, in such a way that it makes most of this information useless
6. Date and Time were two different variables. We merged them to a new variable, named *Date* with information from both of them under a datetime data type
7. The number of persons injured approximates the sum of pedestrians, cyclists and motorists injured for every observation. Adding the total injuries column and then subtracting components yields a difference of as shown in Figure 1. While the total sum does not completely match the sum of its components, it is quite close considering the 1.3mm data points. One potential reason for this, offered by the aforementioned data documentation file, is that "motorist" includes the owner of a parked vehicle
8. Regarding fatalities, the data matches more closely with difference of 1 only:

```
Total injured: 357594
  [A] Pedestrians injured: 70694
  [B] Cyclists injured: 28562
  [C] Motorists injured: 259286
    [A]+[B]+[C] check total injured: 358542

Total killed: 1620
  [A] Pedestrians killed: 881
  [B] Cyclists killed: 110
  [C] Motorists killed: 628
    [A]+[B]+[C] check total killed: 1619
```

Figure 1: Comparing Injuries and Fatalities Totals Columns with Components.

Data Wrangling and Imputation Process

Given our findings, we wrangled and fixed what possible and also imputed what made sense location-wise. More details:

1. Populating *Zip Code* and from there *Borough* represented the main work of data imputation. We were able to get our *Borough* missing values down from 406,750 or almost a third of our data set to 198,210. This effectively reduced missing data under these measures of location by more than half and brought down null values to almost 10% of our entire data instead. We followed two primary procedures for this:
    a. Local imputation: we queried our dataset and noted that exactly the same GPS coordinates (latitude, longitude) where *Borough* was missing occurred elsewhere with *Borough* and *Zip Code* actually populated. Hence, we implemented an algorithm to identify these and consistently populate these matches. This filled 67,743 cases without having to access any external data source. The algorithm had to be efficient given searches over such a large data frame, so we proceeded to first create the smallest populated subset of interest we could search from and source from there our null *Borough* and *Zip Code* values on same pass
    b. External imputation: for the remainder missing values where we had *Location* (GPS coordinates), but no *Borough* nor *Zip Code* anywhere in the data set, we ran Google's Geolocation API for external extraction of our data. We used Python's 'geopy' module with a GoogleV3 geolocator for this task. There were about 150,102 missing *Zip Code* values with *Location* present, which represented natural candidates for reverse geolocation sending our GPS coordinates to the cloud, extracting the closest address possible, and then reading from there the corresponding *Zip Code* using regex. We actually narrowed down the request to 18,473 unique such cases for efficiency and then distributed locally those that were transmitted successfully and matched a *Borough* (about 95%). The sets of zip codes per borough to evaluate against for each of the 5 boroughs (Bronx, Brooklyn, Manhattan, Queens and Staten Island) were simple to

extract from an external source, so we only focused on reading zip codes (5 digits) from the API response. More details in our Jupyter Notebook. This gave us 140,797 additional *Borough* entries overall

2. We replaced the 3 typos from point #3 in the previous section for every observation in the dataset
3. If the contributing factor was "Unspecified", there was nothing we could do
4. As aforementioned, we transformed *Date* and *Time* into a single *Date* column following a 'datetime' data type, effectively a time stamp
5. Vehicle type did not follow any standard and it is impossible to guess the type of car from the brand or vice versa, so no improvement was feasible there

After this process, we obtained a final DataFrame object with the following information shown in Table 2:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1385480 entries, 0 to 1385479
Data columns (total 28 columns):
date                            1385480 non-null datetime64[ns]
borough                         1187269 non-null object
zip_code                        1192554 non-null object
latitude                        1159119 non-null float64
longitude                       1159119 non-null float64
location                        1159119 non-null object
on_street_name                  1106500 non-null object
cross_street_name               997195 non-null object
off_street_name                 230503 non-null object
number_of_persons_injured       1385480 non-null int64
number_of_persons_killed        1385480 non-null int64
number_of_pedestrians_injured   1385480 non-null int64
number_of_pedestrians_killed    1385480 non-null int64
number_of_cyclist_injured       1385480 non-null int64
number_of_cyclist_killed        1385480 non-null int64
number_of_motorist_injured      1385480 non-null int64
number_of_motorist_killed       1385480 non-null int64
contributing_factor_vehicle_1   1378099 non-null object
contributing_factor_vehicle_2   1185709 non-null object
contributing_factor_vehicle_3   89215 non-null object
contributing_factor_vehicle_4   19108 non-null object
contributing_factor_vehicle_5   4755 non-null object
unique_key                      1385480 non-null object
vehicle_type_code_1             1374445 non-null category
vehicle_type_code_2             1164311 non-null category
vehicle_type_code_3             118974 non-null category
vehicle_type_code_4             46442 non-null category
vehicle_type_code_5             9838 non-null category
dtypes: category(5), datetime64[ns](1), float64(2), int64(8), object(12)
memory usage: 252.4+ MB
```

Table 2: Data Structure post-Wrangling.

Answering Exploratory Questions

The following questions drove our exploratory analysis together with the presentation delivered. Our submitted Jupyter Notebook follows the same structure so more details per question can be found there. Through these we attempt to dissect our available data into understanding it from **[1]** Location, **[2]** Factors and **[3]** Time Series perspectives:

Question 1

Location - What *Borough* carries the largest collisions, injuries and fatalities? How about normalized by population?

We looked for:

- 1.1 Absolute count of total collisions/injuries/fatalities per borough and per injured/fatality type (pedestrian/cyclist/motorist)
- 1.2 Normalized count of total collisions/injuries/fatalities per borough and per injured/fatality type (pedestrian/cyclist/motorist)

Q1.1 Absolute count of total collisions/injuries/fatalities per borough and per injured/fatality type (pedestrian/cyclist/motorist):

We first looked at absolute figures for total collisions, injuries and fatalities per borough in search for bigger picture structure in our data and potential initial stories to identify:

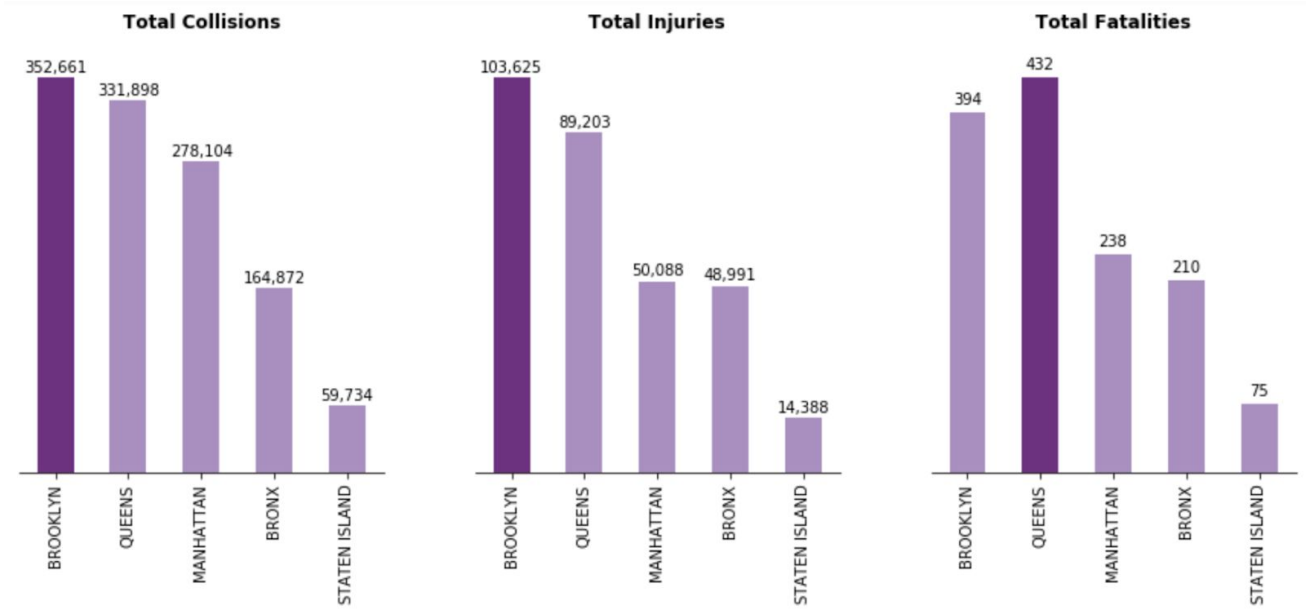


Figure 2: Absolute Total Collisions, Injuries and Fatalities per Borough.

Our initial view from Figure 2 pointed towards a strong mapping to the size of the boroughs themselves, considering Brooklyn has the largest extension and population of any borough while Staten Island the smallest. Queens fatalities did appear to stand out in spite of the large size of the borough. Our initial hypothesis after seeing this related to the highways in the borough, in particular those leading to and stemming from JFK Airport. This would require further analysis of course though.

Inspecting injuries and fatalities further by type of individual involved, whether pedestrians, cyclists or motorists led to another interesting tidbit in spite of first focusing on absolute numbers (Figure 3). Manhattan stood out to us for cyclist injuries, which made intuitive sense considering the city is not really built for bikers. Further inspection of motorists confirmed fatalities in Queens could be related to highways.

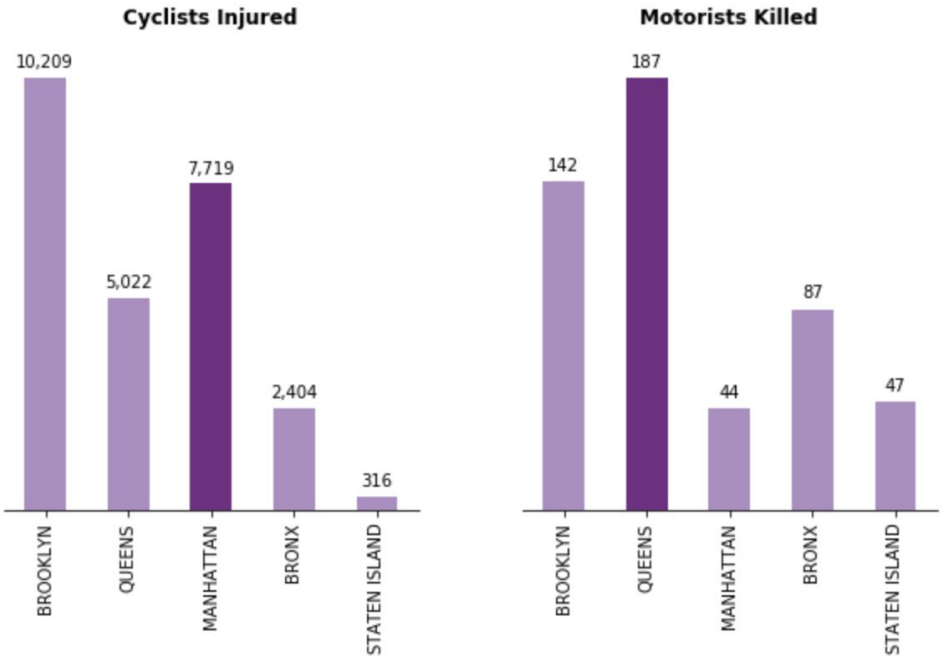


Figure 3: Absolute Cyclist Injuries and Motorist Fatalities per Borough.

Of course the big caveat with the exploration shown through Figures 2 and 3 is the lack of normalization by population. Hence, we wanted ahead and extract population numbers (from here) to perform the same analysis, but on a normalize basis.

Q1.2 Normalized count of total collisions/injuries/fatalities per borough and per injured/fatality type (pedestrian/cyclist/motorist):

When inspecting our dataset on a normalized basis per 100k inhabitants per borough, our understanding of the riskiest boroughs shifted as follows. For the sake of brevity we focus on injuries and fatalities, where the main takeaways came from:

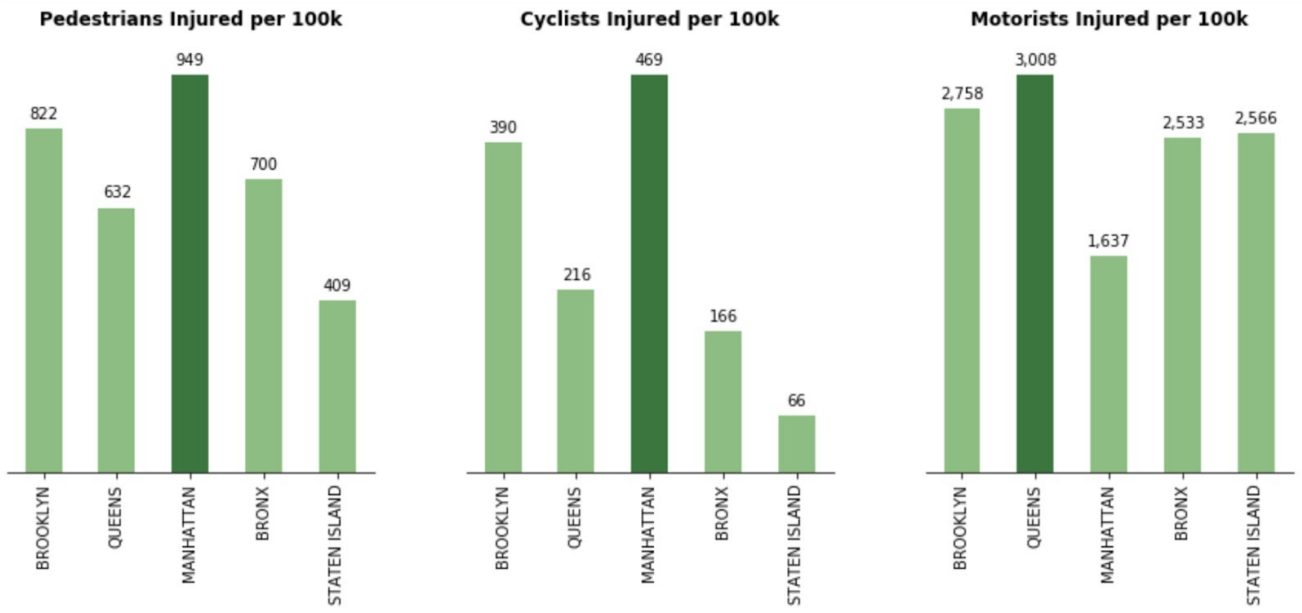


Figure 3: Normalized Injuries per 100k Inhabitants per Borough.

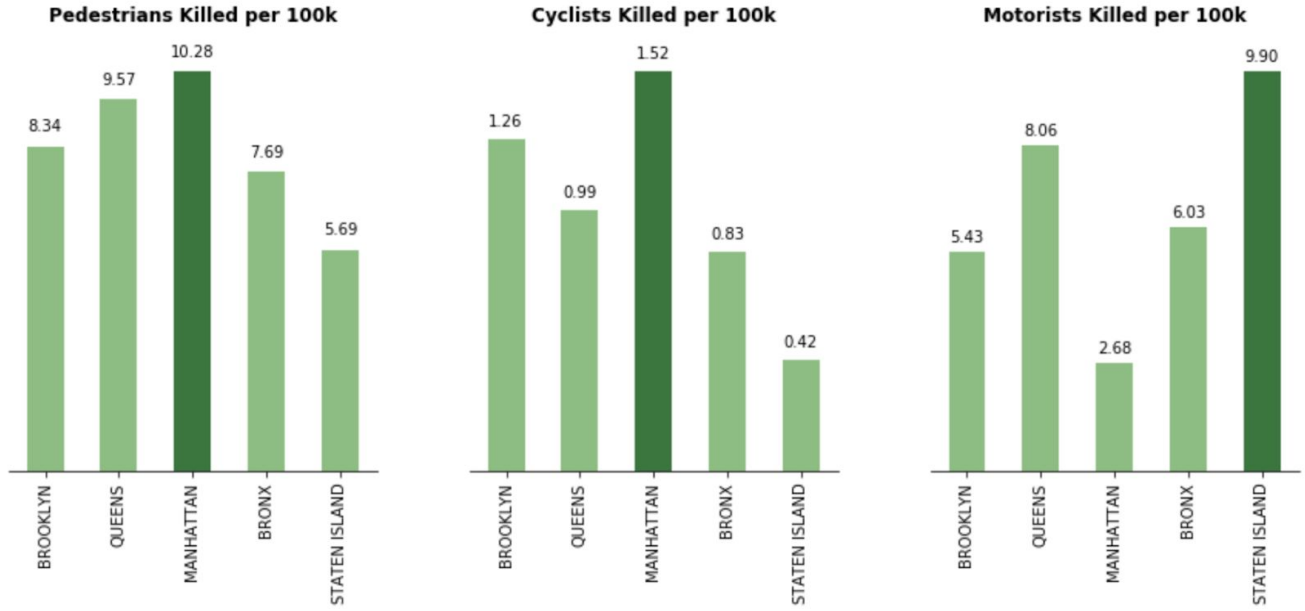


Figure 4: Normalized Fatalities per 100k Inhabitants per Borough.

Note how on a normalized basis, Manhattan rose above any other borough as the most dangerous for both pedestrians and cyclists. On the other hand, Figures 3 and 4 show Queens remains dangerous for motorists in terms of both injuries and fatalities. Regarding motorist fatalities, it was interesting Staten Island came on top, but considering the smaller population relative to other boroughs and that it still has highways that connect it, this result appeared reasonable.

Main takeaway from Question 1:

Manhattan appeared as most dangerous for pedestrians and cyclists, while overall Queens remained as such for motorists.

Question 2

Location - What are the top zip codes or even particular crossings to be aware of?

We looked for:

- 2.1 Top 10 zip codes based on count of total collisions/injuries/fatalities.
- 2.2 Top 10 zip codes based on count of pedestrian injuries/fatalities.
- 2.3 Top 10 zip codes based on count of cyclist injuries/fatalities.
- 2.4 Any maps illustrating overall distribution of collisions or showing these top zip codes graphically.

Q2.1 Top 10 zip codes based on count of total collisions, injuries and fatalities:

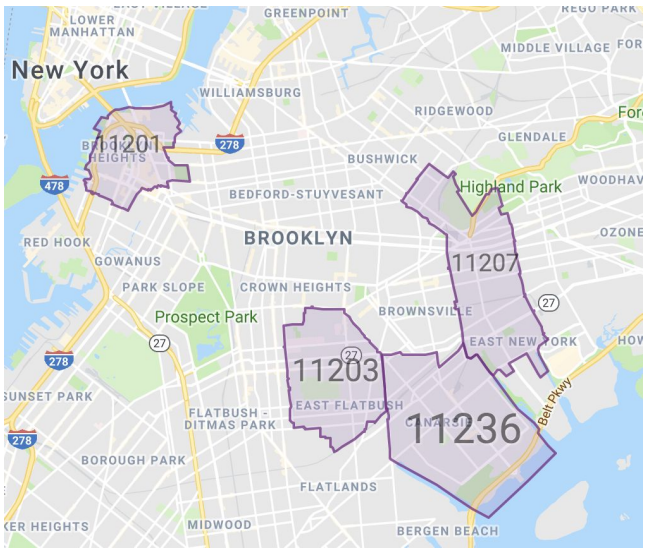
In order to answer this question, instead of grouping by borough as we did above, we proceeded to group by zip code. We also made sure all entries were within official GPS coordinates for integrity of the analysis and any location-driven graphics (extracted official coordinates from here). Sorting by total collisions, total injuries and total fatalities, respectively, we identified the following zip codes as most problematic:

```
Sorting by total collisions:  Sorting by total injuries:  Sorting by total fatalities:
zip_code                      zip_code                    zip_code
Total:    1192554             Total:    308148            Total:    1364
11207       19792             11207       7276            11236       30
11201       17831             11203       5207            11207       24
11101       16899             11236       4879            11354       22
10016       14376             11208       4506            11434       21
10022       14232             11434       4393            11432       18
11234       14195             11212       4210            11229       18
10019       14092             11226       4128            11101       18
10036       13376             11234       3865            10025       17
11385       13374             11101       3707            10029       17
11236       13316             11233       3676            11206       17
```

Figure 5: Top 10 Zip Codes for Total Collisions, Injuries and Fatalities.

Based on Figure 5, we observe the main zip code locations for collisions, injuries and fatalities in Brooklyn. Brooklyn zip codes generally have the form '112**', which pointed our attention towards zip codes 11207, 11236, 11201 and 11203 in particular. Codes of the form '113**' and '114**' belong to Queens. For this borough, fatalities at zip codes 11354, 11434 (injuries too) and 11432 rank high when keeping in context the population adjustment from before. The following map as Figure 6 highlights these for Brooklyn on the left side and Queens on the right side. Notice how these zip codes tend to concentrate around the highway system of these boroughs, and how the Belt Parkway on the southern side of both Brooklyn and Queens:

Brooklyn:



Queens:

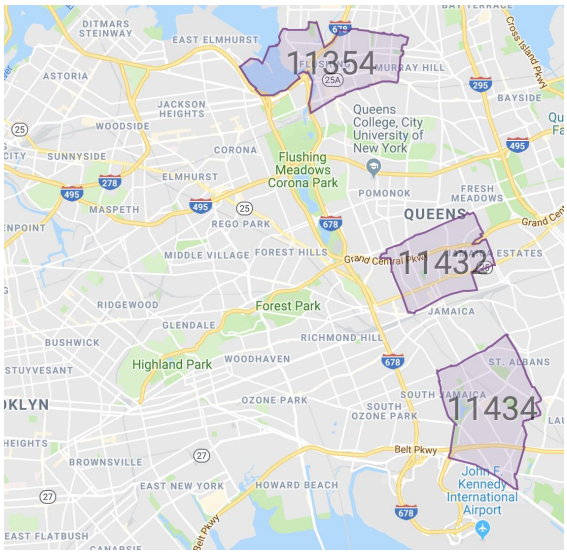


Figure 6: Top Zip Codes for Overall Collisions, Injuries and Fatalities in Brooklyn and Queens, respectively.

The highway-driven suspicion for motorists and in particular within Queens we surmised in Question 1 appears more evident now. We move on to pedestrians and then cyclists in the following sections.

Q2.2 Top 10 zip codes based on count of pedestrian injuries and fatalities:

Performing a similar analysis with the same dataset, just sorting by pedestrian injuries and fatalities we obtain the results shown in Figure 7:

```
Sorting by pedestrian injuries:
zip_code
Total:    63883
11226      1056
11207      1000
10016       917
11212       915
11213       837
11220       837
11203       826
10002       772
11208       743
10467       736
```

```
Sorting by pedestrian fatalities:
zip_code
Total:    754
11354      18
11236      15
11207      14
10025      14
11372      13
11229      13
10002      12
11214      12
10022      11
11235      11
```

Figure 7: Top 10 Zip Codes for Pedestrian Injuries and Fatalities, respectively.

Notice how focusing on pedestrians, zip codes belonging to Manhattan start showing up. These zip codes have the form '100**', highlighting 10016, 10002 (twice for injuries and fatalities), 10025 and 10022. We also get a couple of incremental zip codes for Brooklyn as 11226, 11212/3 and 11220 were not previously highlighted. On the other hand, 11354 leads Queens again, but 11372 also comes into the mix. Figure 8

shows Manhattan zip codes on the left and these incremental zip codes for Brooklyn and Queens on the right:

Manhattan:

Incremental Brooklyn and Queens:

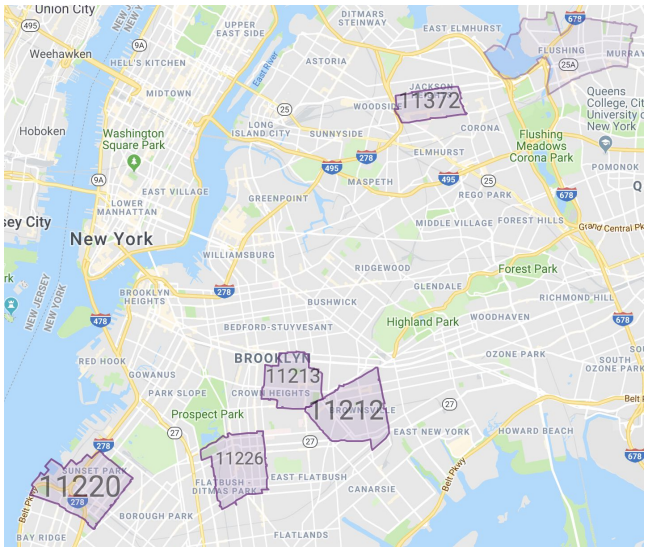


Figure 8: Top Zip Codes for Pedestrian Injuries and Fatalities in Manhattan and Brooklyn/Queens, respectively.

Overall, pedestrian fatalities seem tied to highways too even with zip code 10025 within Manhattan. However, when it comes to injuries, downtown Manhattan appears as the most problematic around 10002 and 10016 together with the heart of Brooklyn as represented by 11212 and 11213 on the right.

Q2.3 Top 10 zip codes based on count of cyclist injuries and fatalities:

Next we focus on cyclists in the same way we did for pedestrians. Figure 9 shows our query for top 10 zip codes for cyclist injuries and fatalities, respectively:

```
Sorting by cyclist injuries:
zip_code
Total:    25683
11206       636
11211       612
10002       531
11368       466
10003       459
11226       424
11221       402
11205       400
10016       396
11201       393
```

```
Sorting by cyclist fatalities:
zip_code
Total:    96
10029      4
11378      3
11223      3
10001      3
11221      3
11101      3
10013      2
10035      2
11218      2
10461      2
```

Figure 9: Top 10 Zip Codes for Cyclist Injuries and Fatalities, respectively.

Note how 10002 again ranks high for Manhattan, but also how we get 11206 and 11211 as incremental zip codes for injuries in Brooklyn not previously highlighted. In terms of fatalities, 10029 in Manhattan tops the list followed by a blend of others in

Queens 11378, Brooklyn 11223 and Manhattan again 10001. We consider it best to visualize top 5 zip codes for injuries on the left and fatalities on the right this time:

Cyclist Injuries:

Cyclist Fatalities:

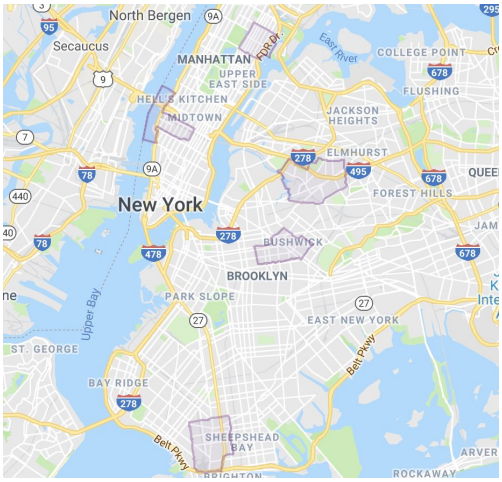


Figure 10: Top 5 Zip Codes for Cyclist Injuries (left) and Fatalities (right).

Figure 10 shows how fatalities remain tied to highways even for cyclists. This would appear to suggest speed as a driver of fatalities instead of injuries, which concentrate more on densely populated areas not as integrated with the highways system.

Q2.4 Any maps illustrating overall distribution of collisions or showing these top zip codes graphically:

In analyzing this data, and given the specific GPS locations we have for the location subset of about 1.19mm data points, we were wondering if we could map all collisions simultaneously. Figure 11 shows such mapping into a latitude/longitude scatterplot:

Figure 11: Visualizing Entire Collisions Dataset at Once.

Notice how a map of anywhere a motor vehicle can circulate resulted. What started as an exploration endeavor ended providing a clear testament to the structure and reliability of the underlying data. Honestly we were surprised location registrations were this granular, precise and consistent over the years. We would also like to highlight Broadway avenue in particular as collision-prone. This is the diagonal that crosses Manhattan from Northwest to Southeast and shows as clearly bolder due to higher collision frequency. We verified this through actual queries on the street data, as shown in Figure 12:

On Street - Total Collisions:

Collision On Street:	
BROADWAY	12222
ATLANTIC AVENUE	10925
3 AVENUE	8698
NORTHERN BOULEVARD	8078

Cross Street - Total Collisions:

Collision Cross Street:	
3 AVENUE	9668
BROADWAY	9506
2 AVENUE	8833
5 AVENUE	7317

Figure 12: Broadway… Not Broad Enough?.

Main takeaway from Question 2:

Highways appear as the main driver of fatalities across all categories and boroughs. For cyclist and pedestrian injuries, downtown Manhattan appears problematic in particular. Regarding overall collisions, not necessarily fatal ones, Broadway stands out above any other street or avenue.

Question 3

Contributing Factor - What are the most commonly identified as contributing to/involved in an accident?

We looked for:

- 3.1 The most frequent contributing factor in terms of number (count) of collisions and in terms of number of persons injured (sum).
- 3.2 The most important contributing factor in terms of severity (highest ratio of injured per collision).
- 3.3 The most frequent contributing factor in terms of severity (highest ratio of killed per collision).

Q3.1a The most frequent contributing factor in terms of number (count) of collisions:

Analysing the most contributing factor (Table 3), in this case for vehicle 1, we found the following values, also represented in a bar chart on Figure 13 below adjusting for "Unspecified":

Contributing factor	Number of collisions
Unspecified	622353
Driver Inattention/Distraction	221340
Failure to Yield Right-of-Way	70321
Following Too Closely	50263
Fatigued/Drowsy	48697
Backing Unsafely	46828

Table 3: Top 6 Contributing Factors by Number of Collisions.

As shown in Table 3, 622,353 out of 1,385,480 registered collisions do not have a specified contributing factor, that is a 44,92%, which means there is room for improvement: If the authorities care about reducing the number of accidents they need to make a bigger effort on registering the causes.

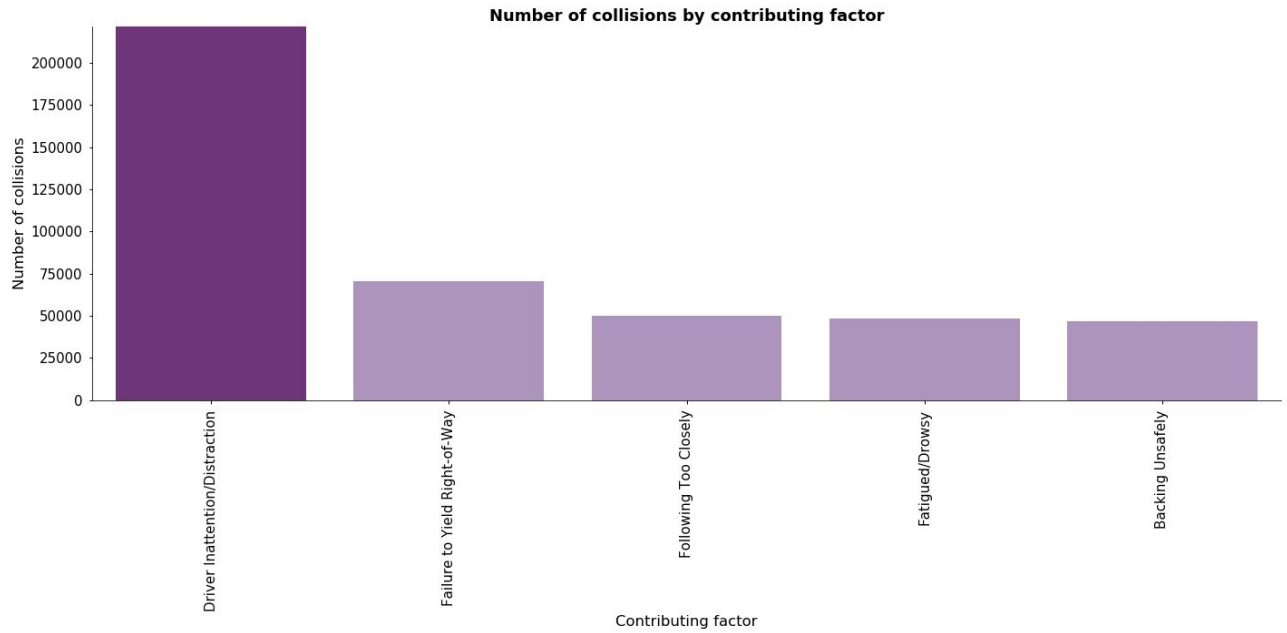


Figure 13: Top 5 Contributing Factors by Number of Collisions, "Unspecified"-adjusted.

The first registered contributing factor is "Driver Inattention/Distraction" with a 15.98%, which is 3 times the third one "Failure to Yield Right-of-Way", as shown in Figure 13.

Q3.1b The most frequent contributing factor in terms of number of injured people:

The first registered contributing factor, now in terms of number of persons injured, is also "Driver Inattention/Distraction", again followed by "Failure to Yield Right-of-Way". See Table 4 below.

Contributing factor	Number of persons injured
Unspecified	146862
Driver Inattention/Distraction	58377
Failure to Yield Right-of-Way	29187
Following Too Closely	15858
Fatigued/Drowsy	12605
Traffic Control Disregarded	11773

Table 4: Top 6 Contributing Factors by Number of Persons Injured.

In this case, it is interesting to notice that "backing unsafely", which was the 5th registered contributing factor in terms of number of accidents does not appear in this shot list as it is the 11th factor in terms of number of persons injured.

It is the opposite situation of "traffic control disregarded" which is not so frequent (10th) but causes more persons injured (5th).

Q3.2 The most important contributing factor in terms of severity for persons injured (highest ratio of persons injured per collision):

Analysing the most important contributing factor in terms of severity, as measured by ratio of persons injured per collision, was "Listening/Using Headphones" with a value of 1.5. However, there were only 2 observations with this factor and 3 persons injured.

Hence, we filtered those contributing factors with a total number of persons injured which would represent at least the 100th percentile, that is 3,575 persons injured in the whole period. For each one of them, we calculated the total number of persons injured, the number of collisions and the severity, ordering the results by severity. The result is shown in Table 5 below:

Contributing factor	Persons injured	Number of collisions	Severity ratio
Physical Disability	5809	9426	0.616274
Traffic Control Disregarded	11773	19384	0.607357
Unsafe Speed	4984	8371	0.595389
Alcohol Involvement	5501	11914	0.461726
Failure to Yield Right-of-Way	29187	70321	0.415054
Following Too Closely	15858	50263	0.315500
Pavement Slippery	3951	13618	0.290131
Driver Inattention/Distraction	58377	221340	0.263744
Fatigued/Drowsy	12605	48697	0.258846
Driver Inexperience	4397	18497	0.237714
Unspecified	146862	622353	0.235979
Other Vehicular	8072	41054	0.196619
Turning Improperly	5151	33967	0.151647
Backing Unsafely	4732	46828	0.101051

Table 5: Top 14 Contributing Factors in Terms of Severity for Persons Injured.

What is clear from this data summary is that human is the main responsible for almost all these contributing factor, and that the top 4 can associated with illegal actions, not just human errors.

Q3.3 The most important contributing factor in terms of severity for persons killed (highest ratio of persons killed per collision):

Repeating the process described in the section above for persons killed, filtering the 100th percentile -at least 16 persons killed by that factor in the whole period- we got a list of contributing factors. The 4 most severe ratio are 2 with contributing factors related with distractions and 2 related with illegal actions: unsafe speed and traffic control disregarded, the last one being also meaningful in terms of number of persons killed (around 10% of total fatalities).

Additionally, we calculated the distribution of number of persons killed by accident and merged it with information described above, getting a very descriptive dataframe. See Table 6 below, with the data ordered by severity ratio:

Contributing factor	Persons killed	Number of collisions	Severity ratio	0	1	2	3	4	5	8
Unsafe Speed	80	8371	0.009557	8298	67	5	1	0	0	0
Passenger Distraction	53	5845	0.009068	5792	53	0	0	0	0	0
Pedestrian/Bicyclist/Other Pedestrian Error/Confusion	29	3382	0.008575	3353	29	0	0	0	0	0
Traffic Control Disregarded	158	19384	0.008151	19245	126	9	3	0	1	0
Illness	19	2780	0.006835	2762	17	1	0	0	0	0
Alcohol Involvement	48	11914	0.004029	11868	44	2	0	0	0	0
Physical Disability	30	9426	0.003183	9396	30	0	0	0	0	0
Failure to Yield Right-of-Way	129	70321	0.001834	70193	127	1	0	0	0	0
Driver Inexperience	22	18497	0.001189	18478	18	0	0	1	0	0
Unspecified	685	622353	0.001101	621687	649	15	2	0	0	0
Lost Consciousness	19	20410	0.000931	20391	19	0	0	0	0	0
Driver Inattention/Distraction	178	221340	0.000804	221167	169	3	1	0	0	0
Backing Unsafely	22	46828	0.000470	46806	22	0	0	0	0	0
Other Vehicular	19	41054	0.000463	41042	11	0	0	0	0	1
Following Too Closely	20	50263	0.000398	50244	18	1	0	0	0	0

Table 6: Top 14 Contributing Factors by Severity on Fatalities (Highest Ratio of Persons Killed per Collision).

This information allows to get a general overview of collisions with fatalities in one data table only. Observe that:

- As a curiosity, there have been no collisions with 6 or 7 persons killed and only one with 8
- "Unsafe speed" and "traffic control disregarded" have more observations with higher number of fatalities

Question 4

Time Series - From a time series perspective, is there seasonality in overall injuries/fatalities? How about seasonality per type of injured subject (pedestrian/cyclist/motorist)?

We looked again to the data from a time series perspective, trying to answer the questions above. The analysis was done for all the types of injured subject and they are available in the Jupyter notebook. Only a part of it is shown next:

- 4.1 Is there seasonality in overall injuries/fatalities?
- 4.2 How about seasonality for cyclist?

Q4.1 Is there seasonality in overall injuries and fatalities?

Plotting the overall persons injured over time, grouped by day in the period, we got the following, Figure 14:

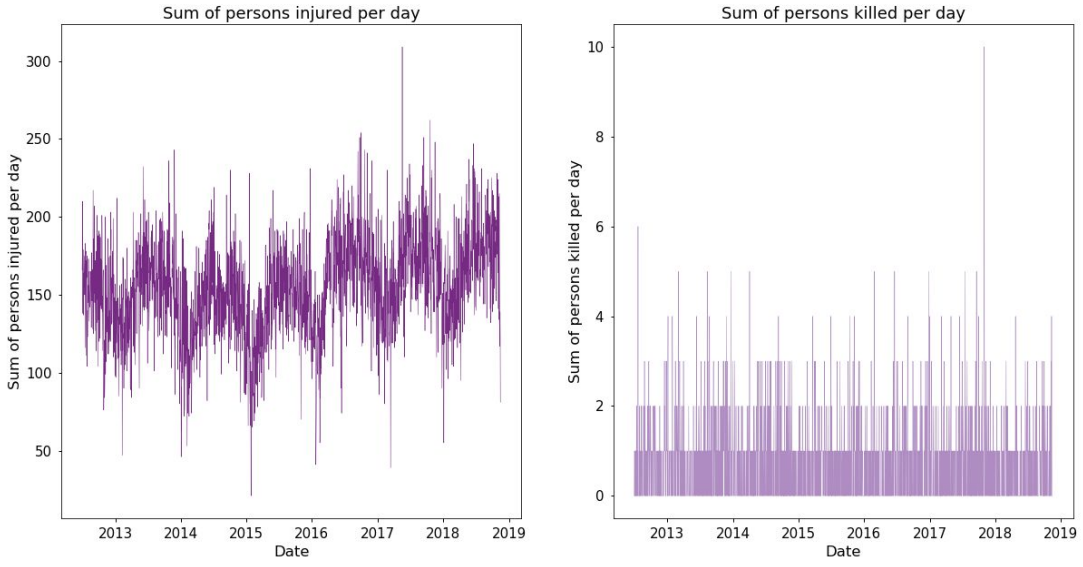


Figure 14: Time Series of Daily Persons Injured (left) and Killed (right).

Injuries, on the left: there seems to be some seasonality and some tendency too.
Fatalities, on the right: impossible to guess from this plot.

We calculated the tendencies as 365 days rolling averages, to eliminate the effect of

seasonality. You can see, Figure 15 below, there has been a slight increasing tendency in the number of persons injured in the last years and the opposite for persons killed.

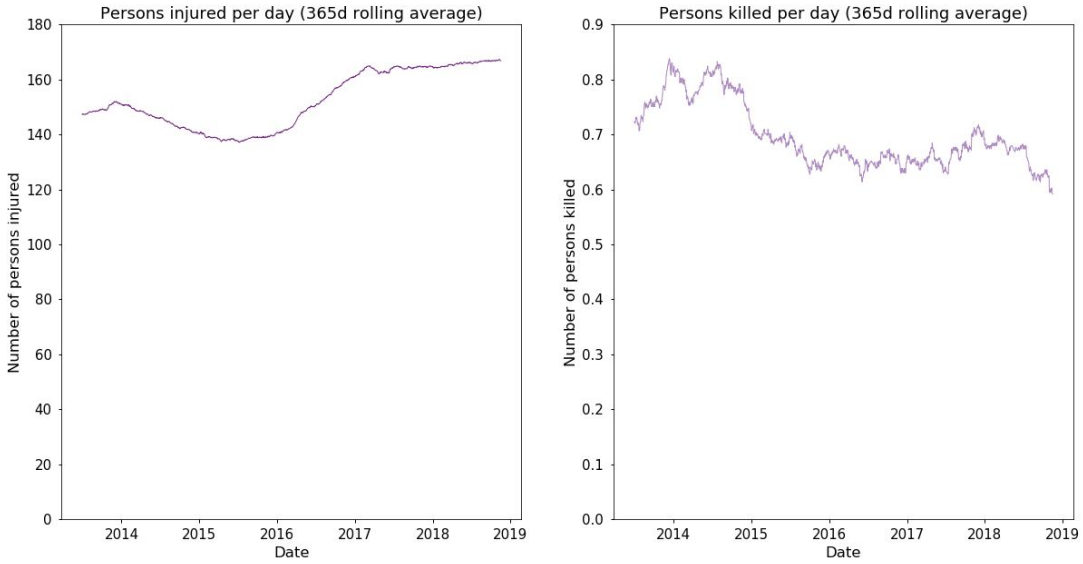


Figure 15: Tendency of Daily Persons Injured (left) and Killed (right) as a 365-day Rolling Average.

Eliminating the tendency from the evolution of the number persons injured and plotting the 30 days rolling average on top, we could clearly see there is a seasonality decrement in winter. There seems to be some seasonality too in the rest of the year, but it is not so clear. See Figure 16 below.

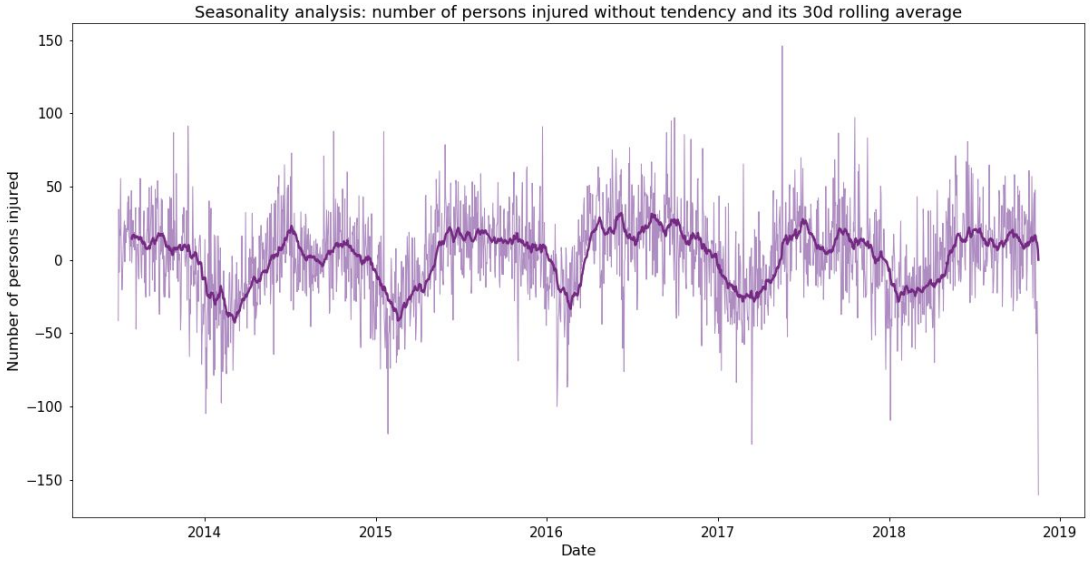


Figure 16: Seasonality Analysis as Number of Persons Injured Without Tendency and 30-day Rolling Average.

So we decided to calculate the autocorrelation to have a definitive and statistical understanding of the issue.

If time series is random, such autocorrelations should be near zero for any and all time-lag separations. If time series is non-random then one or more of the autocorrelations will be significantly non-zero. The horizontal lines displayed in the plot correspond to 95% and 99% confidence bands. See Figure 17 below.

In relation to the number of persons injured, there is a clear positive autocorrelation with a lag of 365d and a not so clear negative one with a lag of 180 days.

For the number of persons killed, the autocorrelation value shown in Figure 17 is too close to 0 to be meaningful.

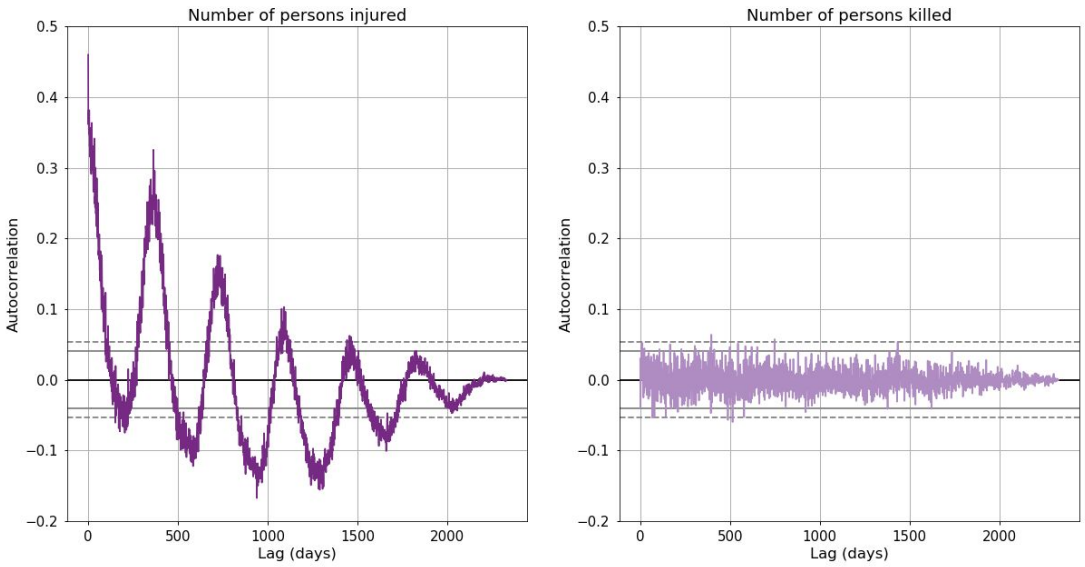


Figure 17: Autocorrelation Analysis of Persons Injured (left) and Killed (right).

Q4.2 Is there seasonality in overall injuries and fatalities?

We decided to focus on persons injured only because of the findings in the previous section. We divided it in the three subgroups: pedestrians, cyclists and motorists. See Figure 18 below.

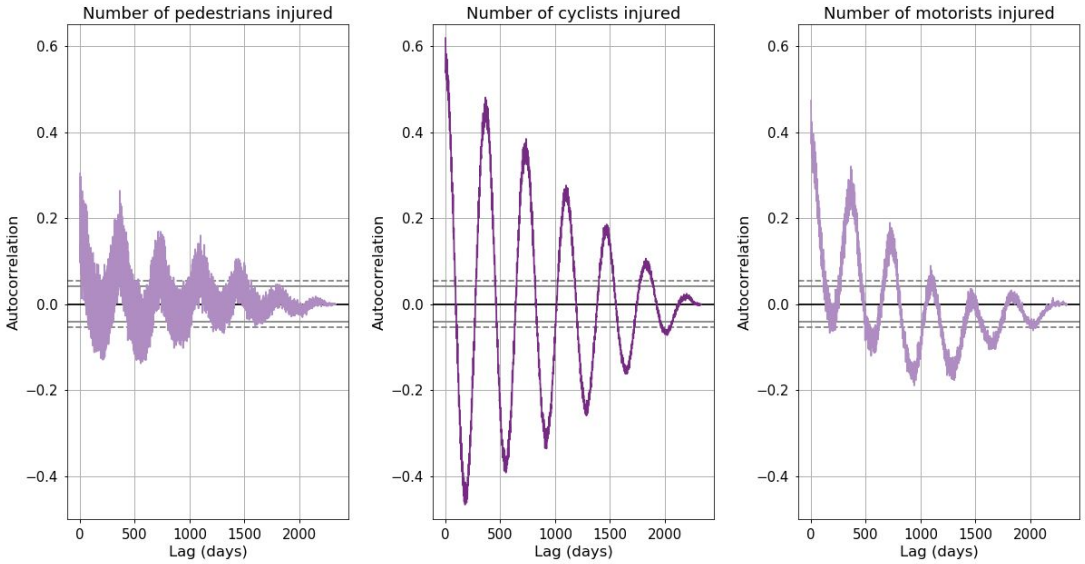


Figure 18: Autocorrelation Analysis of Injuries per Subset: Pedestrians (left), Cyclists (center) and Motorists (right).

From this information, it very straightforward to realize that the cyclists -the one in the middle-is the one with the highest autocorrelation both in summer and winter, most probably caused by the use of bicycles due to the weather conditions along the year.

Potential recommendations

Populating *Zip Code* and from there *Borough* was a must to get our *Borough* missing values down from 406,750 or almost a third of our data set to 198,210. NYC Open Data should consider providing this information to all users.

There are 3 different typos in the factor of contribution, affecting the analysis of the data. Data should be cleansed before uploading it to NYC Open Data.

It is important to mention that the factor of contribution in the original dataset is unspecified in about half of the collisions, the reason being the police officer does not know yet the reason when the accident takes place. If data should be collected afterwards and stored in the database in order to better analyze, know and solve the reasons for the collisions.

The data in vehicle type variables is absolutely useless because it does not follow any procedure and it is quite messy, as explained in the EDA section. Categories for this variable should be presented to the police officer from a fixed number of possibilities.

Conclusion

- Without establishing inferential causality, this exploratory analysis has provided areas to focus on further. [Highways overall] [Cyclists in x season and in particular in Manhattan] [Queens highways in particular too].

- Location-wise Manhattan, and in particular Broadway Avenue, stands out as accident-prone for pedestrians and cyclists. For motorists, highways in Brooklyn and Queens remain the most dangerous.

- The highest contributing factor in terms of number of accidents is "driver inattention/distraction". In terms of severity, "unsafe speed" is the main one.

- There is seasonality for injuries, specially in the case of cyclists. Authorities could reinforce safety messages just before the good weather season.