Parth Khanna

260733741

11/29/2017

Contents:

Outline

Approach

Data Analysis and Findings

* Season
* Time of Day
* Borough
* Reason of Collision
* Type of Commuter

Concluding Remarks

Bibliography and Appendix

**Technical Appendix**

New York City Vehicle Collisions

INSY 336 with Mr. Kartik Krishna Ganju

Detailed Report

# Outline

From the beginning of time, humans have moved from one place to another place. Commuting is ingrained in human behavior and cannot be restricted. This makes ‘Vehicle Collisions’ a deep-rooted problem that is very significant in today’s rapidly progressing world. This problem’s importance is amplified when one realizes the numerous lives that lost, or affected in other ways, because of it.

These collisions, however, can be reduced significantly if proper measures are employed. Given that each collision occurs due to a tandem of reasons, one can infer that distinct factors are associated with it, the right analysis of which may give us some interesting and influential results. What time of the day does Bronx have most number of collisions? How common is ‘Influence of Alcohol’ as a reason for accidents at night? Certain parameters can be analyzed to identify some patterns, thereby contributing to the whole issue.

The problem emerges particularly in the metropolitan cities where work impacts people’s lives, arguably, more than the people living in less-developed cities. For this reason, the New York City Vehicle Collisions dataset proves to be a very good resource for analyzing this issue in the NYC region. The goal of the project is essentially to identify the popular causes of collisions and other important factors that can then be monitored/avoided to reduce vehicle collisions. From the spectrum of these selected factors, the analysis attempts to identify collision prone areas of New York City.

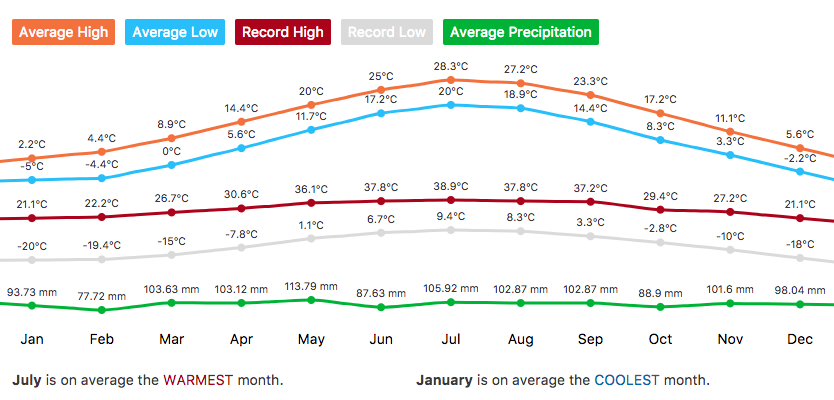
# Approach

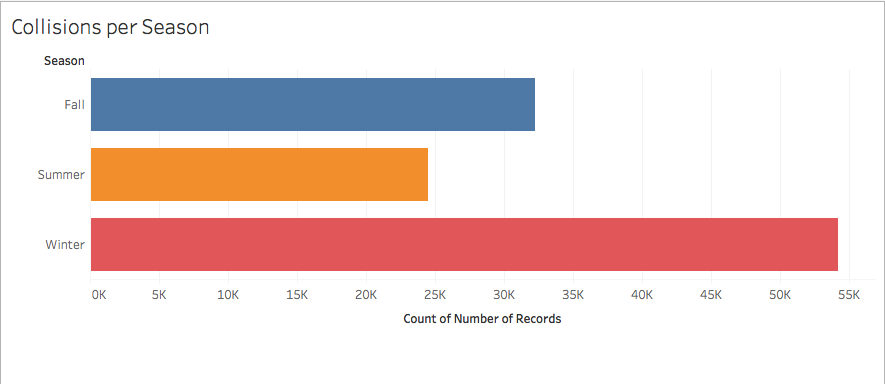
Very succinctly put, I used a ‘Divide and Conquer’ approach throughout my analysis. Since the dataset was extremely large, my approach followed a two-way approach wherein I first explored the subsets of the main dataset taken out based on different parameters and then selected certain factors that showed interesting patterns. Indeed, I had to clean the data initially but besides that, I had to consistently manipulate the dataset throughout my analysis to magnify and observe the hints that I had gotten from the subsets by using visualizations and different R functions.

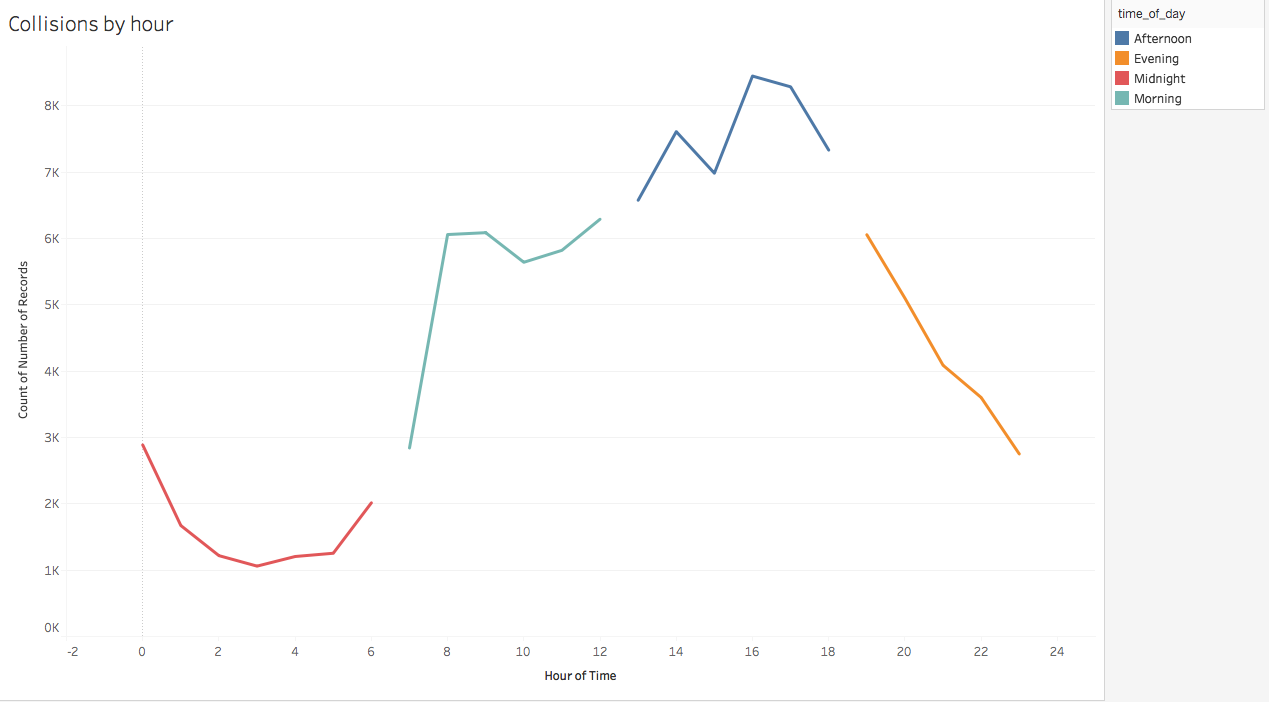
Before and while doing the basic analysis, I brainstormed/realized questions that the project would answer and these became the guiding questions. Since I was doing the project individually, I used GIT (Version-Control) to make sure I don’t lose any important results that I discover. This was done because the entire data analysis took me a long time and several edits of the R script. Losing important results due to various reasons meant inaccurate analysis and the point was to avoid this. You can find the online repository containing the code file [here](https://github.com/parthkhanna150/R-stuff). Besides that, I sought help from resources online (see in Bibliography) to make sure I was on the right track. It is imperative to keep a constant check on your observations with those of the experts in the field. For example, I didn’t assume the periods of seasons in the years of my dataset. Instead, I used an external table of temperatures of NYC that I found online from a reliable source. While most of my analysis was in R-studio and Tableau, I sought Google Maps to identify the areas that I discovered to be accident prone. This was done particularly to give a better understanding of the whereabouts of the region as Google Maps’ visualizations highlight important landmarks of the region.

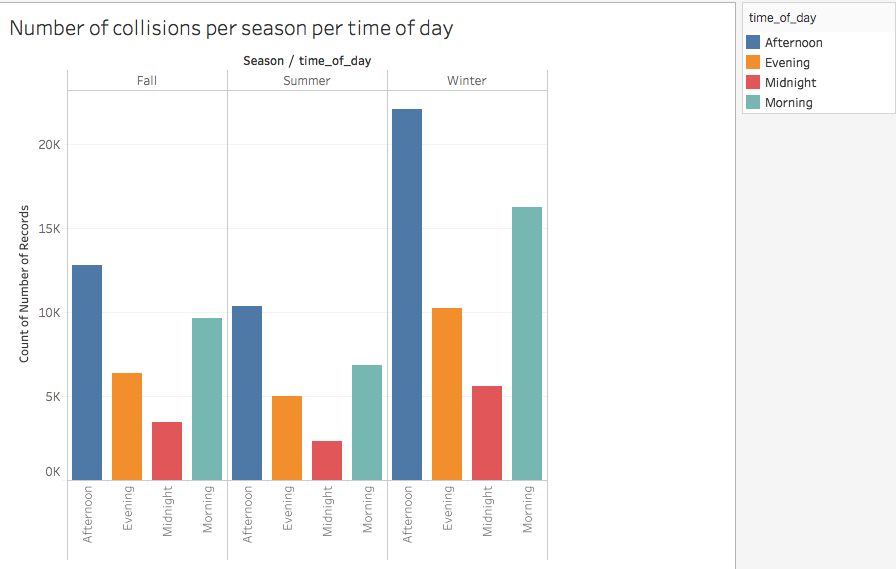
# Data Analysis and Findings

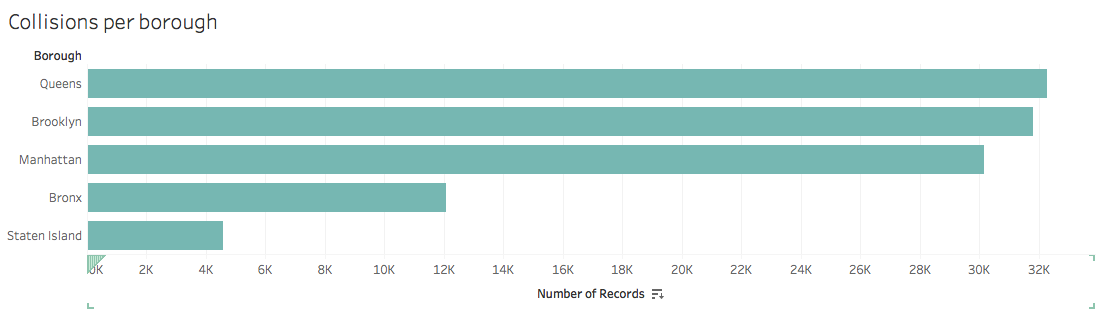
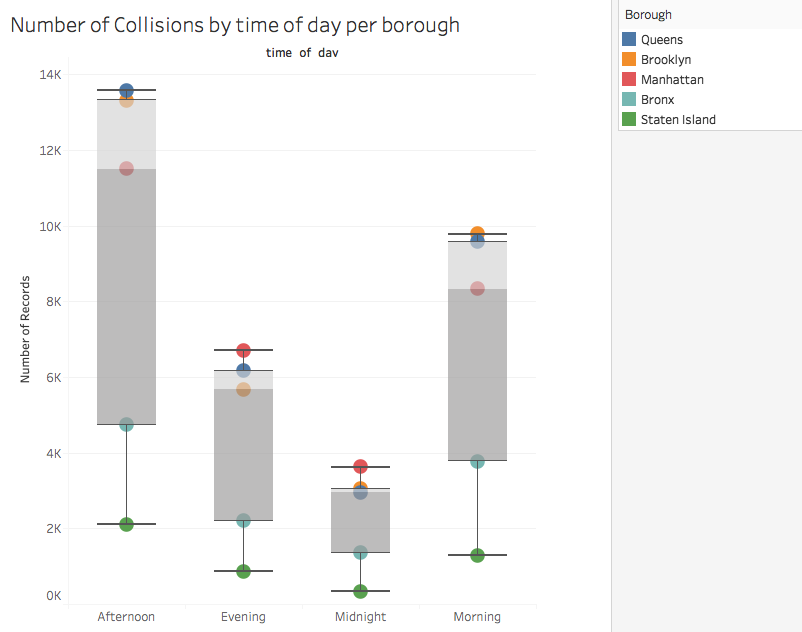
Several dataframes were subsetted out, in order to figure out angles of exploring different patterns in the main dataset. While many didn’t help out significantly, there were some that led to some very interesting results. For example, the number of people involved in pedestrian collisions was seen to be greater than the number of people involved in Motorist collisions. However, when I removed the records where no one was injured/killed, it was seen that the number of motorists involved is twice that of pedestrians involved! The analysis such as this, but deeper has been explained in detail below.

To analyze patterns by season, my first step was to observe the ‘number of collisions per season’. I recognized only three seasons even though I was keen to consider the ‘Monsoon’ season’s impact because I learnt that NYC didn’t have a well-defined Monsoon season, as can be seen from the Avg. Precipitation (Green line) in the picture below.

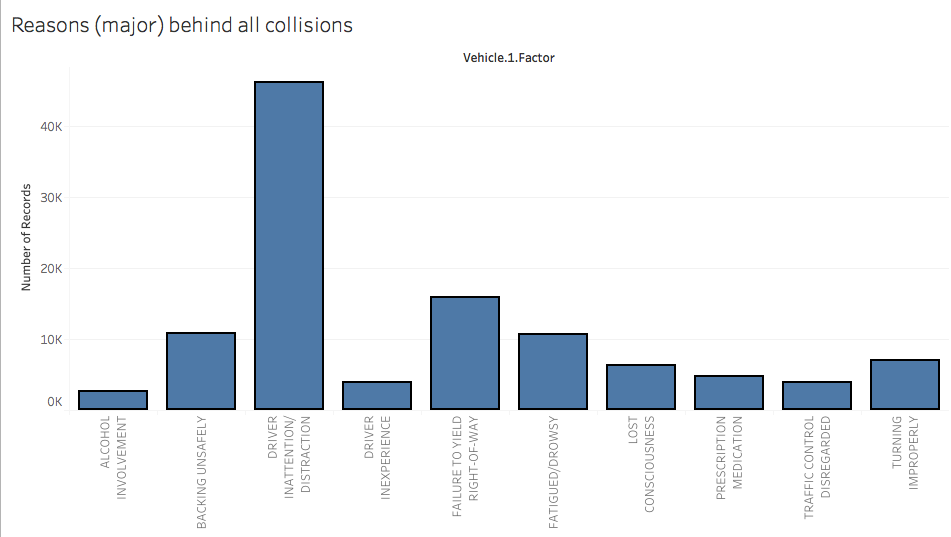
I noticed some peculiar behavior of the data in that the winter collisions were more than twice of the summer collisions. This was drastic for 5 months of Winter and 4 months of Summer. I read online about this issue and it indeed was true that winter had more collisions than summer. But due to the unusually high discrepancy, I decided not to go very deep in analyzing the dataset from the lens of ‘Season’.

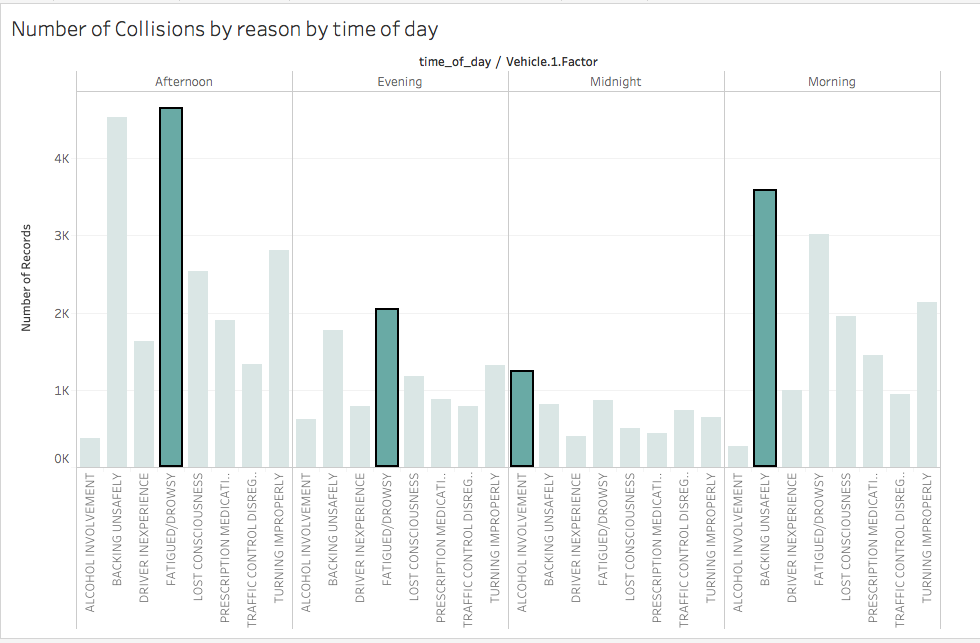
Next, I wanted to see how collisions are affected at different times during a day. So, after getting the hours and categorizing them into morning (6am-12pm, afternoon (12pm-6pm), evening (6pm-12am) and midnight (12am-6am), I plotted them and realized that the most collisions occurred during afternoon. This was intuitive and the result is shown below.

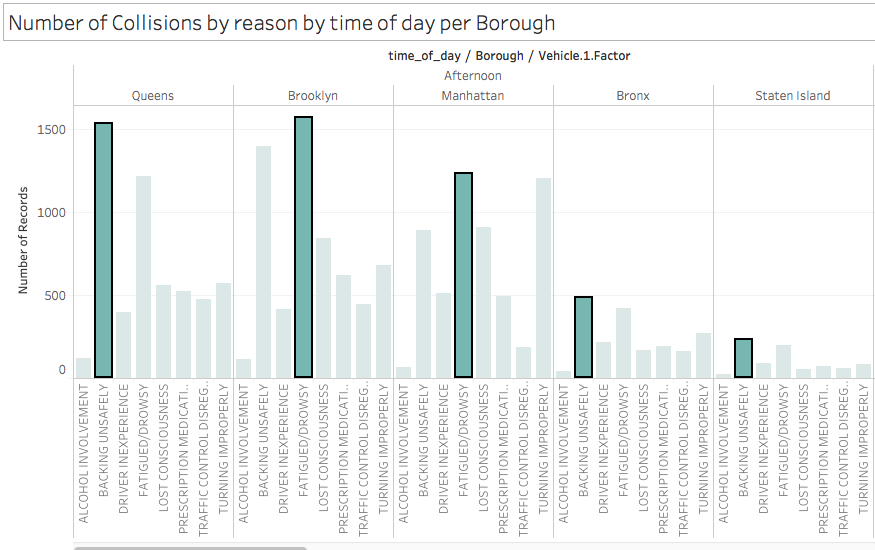
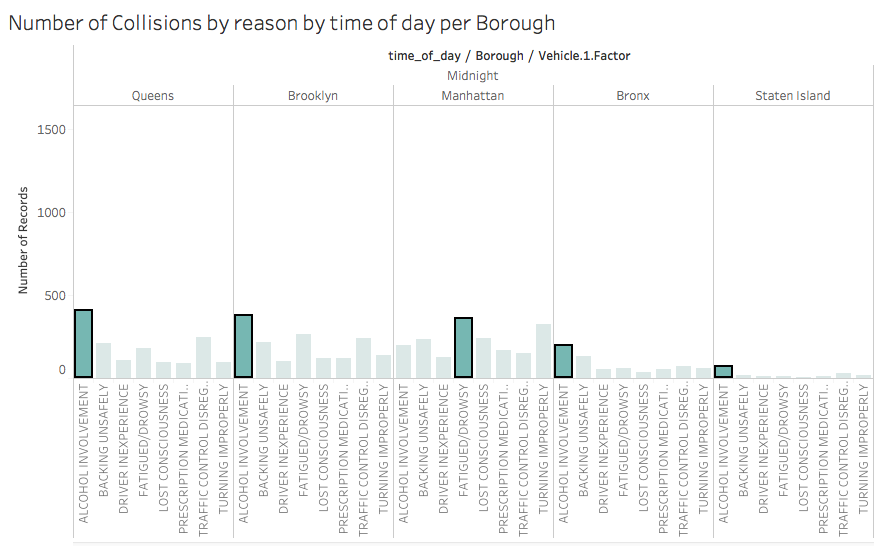
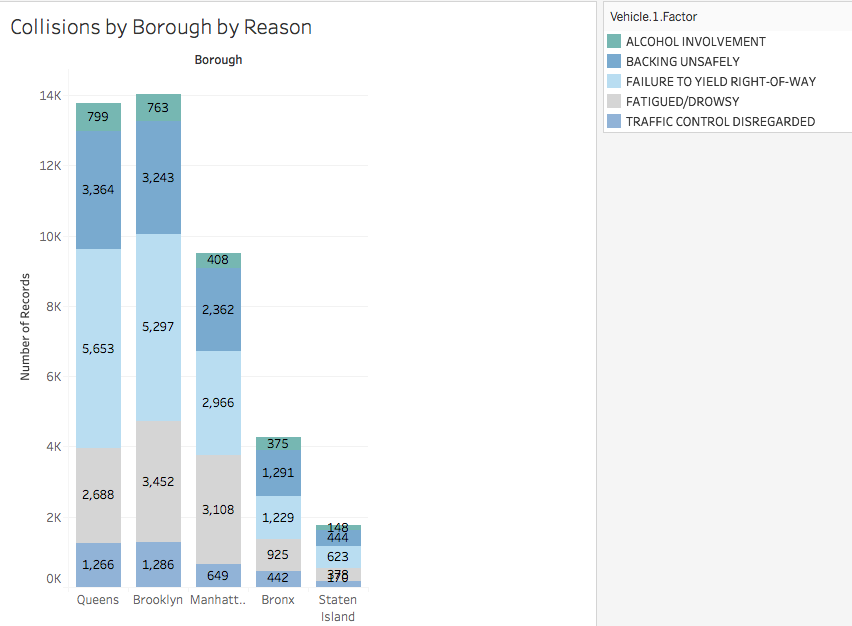
Instead of completely disregarding ‘Season’ as an important factor, I wanted to atleast see how the number of collisions was influenced by different times of day in different seasons. So, I plotted a graph only to realize that each of the three seasons had very similar (not identical- if you look closely) pattern of collisions during different times of the day. Note that the graph isn’t proportionate, in that winter has more records as previously discussed, yet the pattern is definite. Because of the nature of this pattern, I decided to leave out season as a factor in my further analysis.

Next, I looked at NYC geographically and divided the city into its 5 boroughs to analyze closely. First, I looked at the number of collisions per borough. This ranked Queens, followed by Brooklyn and Manhattan, followed (very behind) by Bronx and Staten Island. An interesting insight is that according the Census of 2016, Brooklyn is more populous than Queens which reflects its ‘slightly’ better urban planning and policies. Besides this observation, the number of collisions closely correlated with the populations (see in appendix) of each borough. Note that the image above accounts for ‘all’ collisions and not necessarily just the ones where people got injured/killed i.e. there exist collisions where 0 people got injured/died.

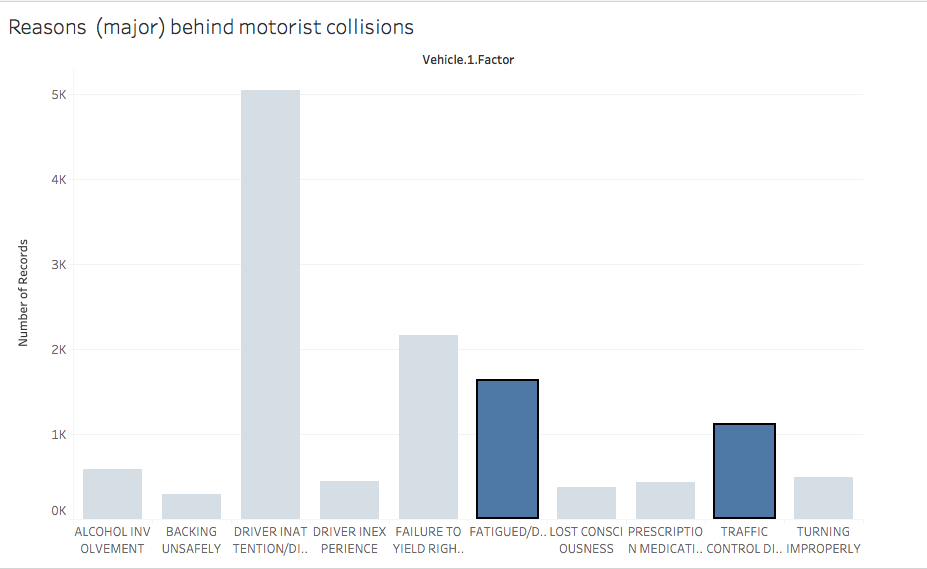
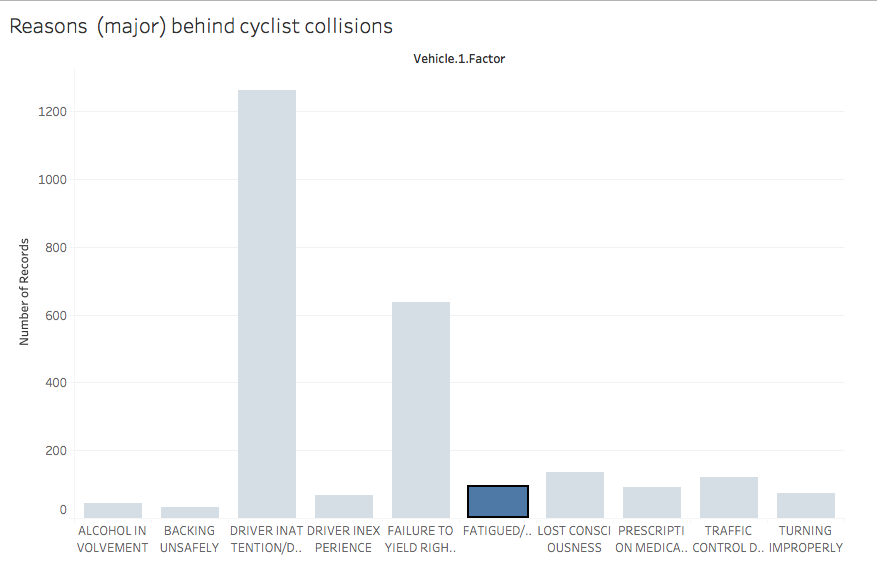
Thereafter, I wanted to analyze which borough has most collisions during different times of day. When I plotted that, I discovered an interesting result that showed Manhattan distinctly on top for both evening and midnight while Brooklyn and Queens almost at par during morning and afternoon. I inferred that this reflected how people commuted to Manhattan in the evening after work for some recreational time or the popular nightlife in Manhattan compared with other areas.

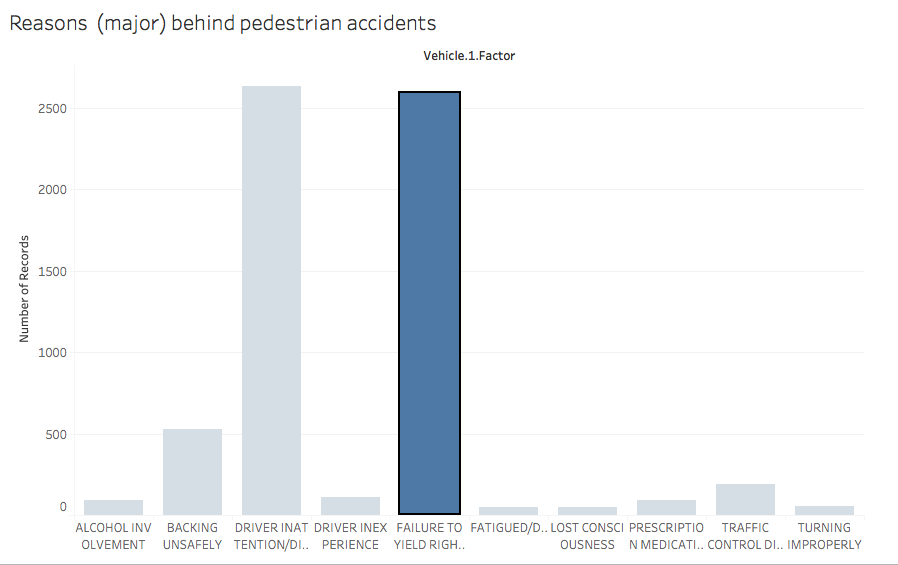
After this result, I wanted to indulge into the repetitive reasons that have led to accidents. So, after cleaning the data and plotting the identified popular reasons, I obtained the histogram given below. Note that I didn’t account for the records that had NULL or ‘UNSPECIFIED’ as their reason. ‘UNSPECIFIED’ was the most frequent reason but it wasn’t fruitful for my analysis so I ignored that. ‘Driver Inattention’ was the most common reason for collisions, followed by ‘Failure to yield right-of-way’, ‘Backing unsafely’ and ‘Fatigued’. This discovery, too, was intuitive to some extent but I wanted to analyze this trend further.

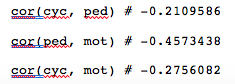
Therefore, I decided to break it down and analyze it together with ‘time of day’ parameter and the result I obtained was even more revealing. The histogram above shows the popular (the top reason is highlighted) reasons that have caused collisions during different times of day. We see that while ‘Fatigue’ is distinctly the winner during afternoon and evening, ‘Alcohol Involvement’ and ‘Backing Unsafely’ become top reasons for collisions during midnight and morning respectively. This result supports the common routines of most people in that they are ‘sleepy’ in the morning and so prone to backing not vigilantly, tired in the middle of the day due to work and then relax later by resorting to things like liquor.

Since the analysis along the lines of ‘reasons for collisions’ seemed satisfying yet interesting, I decided to go even further with it. Next, I introduced ‘Boroughs’ as a parameter into the above equation to find out the popular reasons for collisions during different times of day in different boroughs. The following two histograms for afternoon and midnight (the rest are in Appendix). During afternoon, we see that Brooklyn and Manhattan have ‘Fatigue’ as their main reason for collisions whereas the others have ‘Backing Unsafely’. While ‘Fatigue’ is more of a personal reason, ‘Backing Unsafely’ reflects an aspect that the policy makers or urban planning organizations of Queens, Bronx and Staten Island need to introspect to reduce collisions. Similarly, for midnight, all except Manhattan have ‘Alcohol Involvement’ as a top reason which is again intuitive but something that needs to be looked at by the policy makers of Queens, Brooklyn, Staten Island and Bronx. This also reflects the strict law and order for consumption of Alcohol at night in Manhattan. The evening and morning histograms are included in the appendix and they too show insights such as these. We see that in the morning the main reason happens to be ‘Backing unsafely’ and in the evening, it happens to be ‘Fatigued’. One insight worth mentioning is that Manhattan has ‘Turning Improperly’ as another very common reason in the evenings and this, too, is something the policy makes should look into.

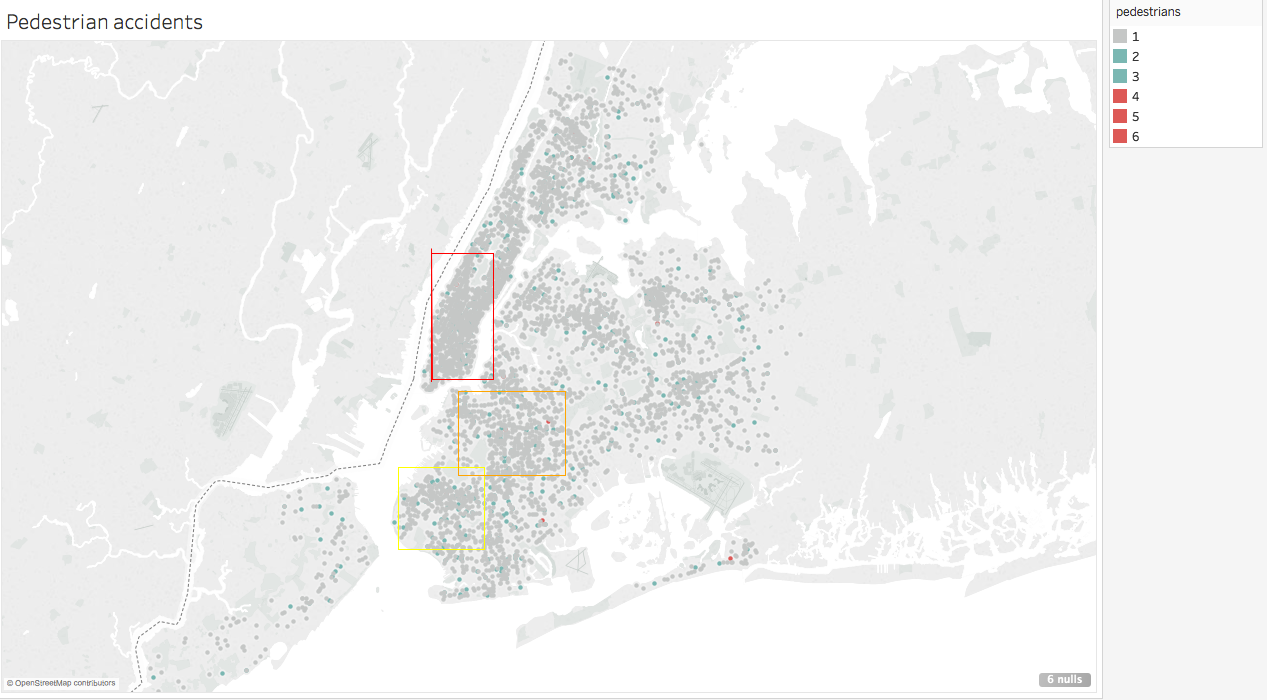
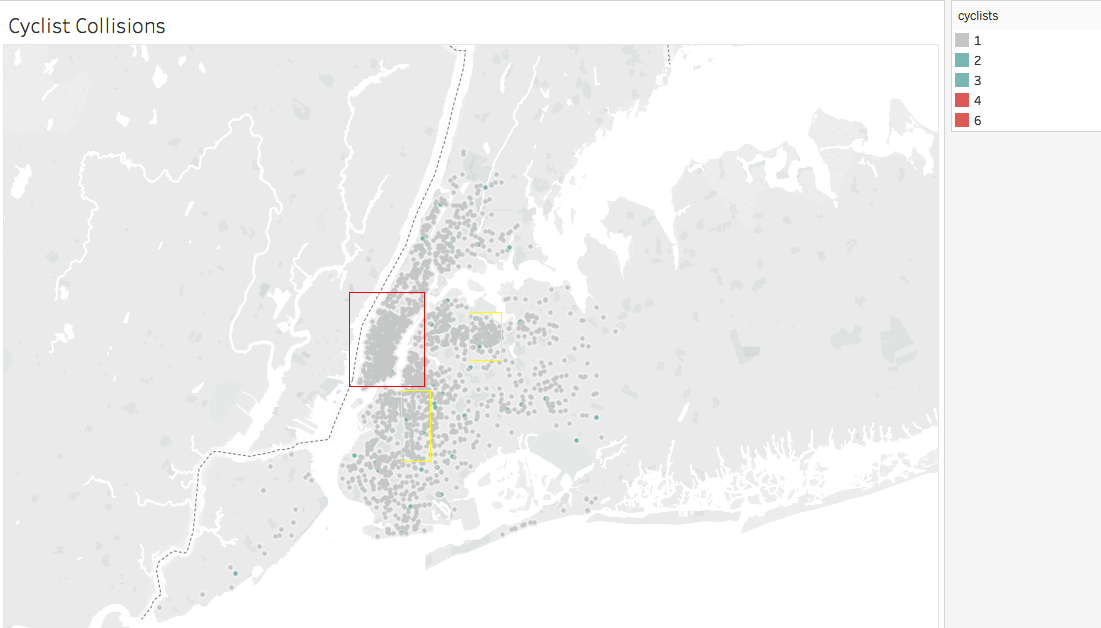
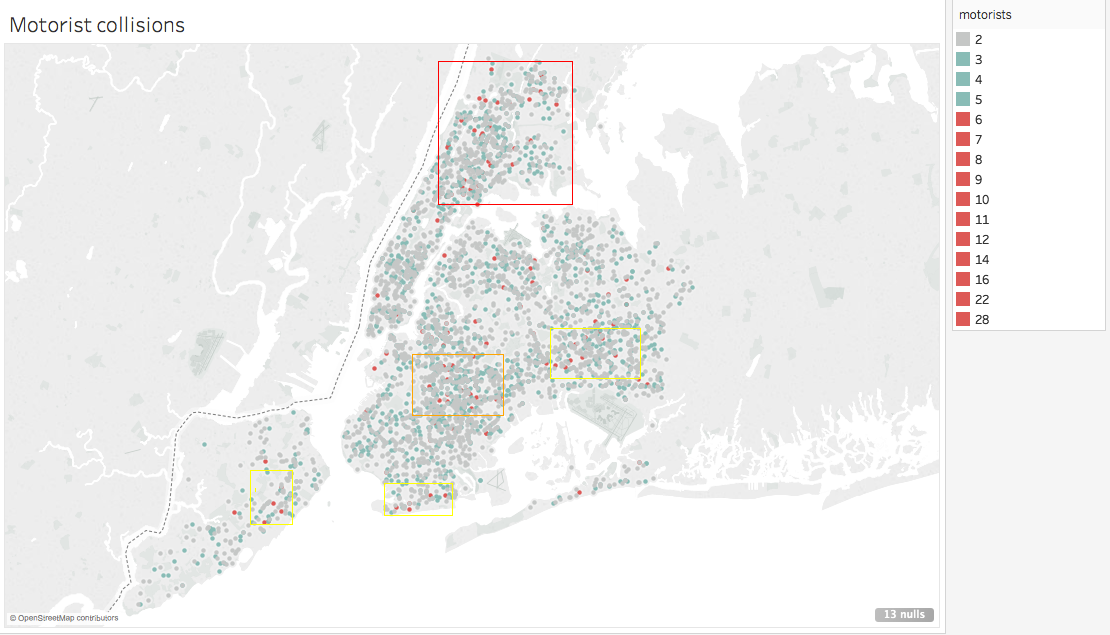
Next, I decided to explore the dataset from the lens of ‘type of commuter’ i.e. cyclists, motorists and pedestrians. I decided to put this parameter alongside the ‘reasons for collisions’ that I had been exploring. Interestingly, the most common reason for motorist collisions was ‘Fatigue’. This makes sense if you assume that motorists constitute mostly the people who belong to the working age group and so will have an average age higher than that of the pedestrians and cyclists, both of which most certainly include children and younger people. This continues to bode well when we see the low frequency of accidents of cyclists and pedestrians due to ‘Fatigue’. However, this is also something that needs further analysis as it may be that since cycling and walking in fact promote a healthy lifestyle, people tend to be less fatigued. The given data was insufficient so I couldn’t go forward in this direction. The three histograms below show these observations. Another very interesting discovery was that ‘Failure to yield right-of-way’ formed the top reason for pedestrian collisions. This is highly intuitive and avoiding these collisions is something only the commuters can do i.e. the Policy Makers can’t make a rule regarding this.





Following this, I wanted to analyze the ‘nature of people’ involved in the collisions i.e. what were the likelihoods of having two different kinds of commuters involved in a collision with each other. To obtain this result, I had to first remove all the records where there was no one was injured or killed. Next, I leveraged R language’s function for correlation on the required vectors to discover the likelihood. The above image below shows the results. The results weren’t very definite in that no correlation was surprisingly high. But they all involved an inverse relationship (negative correlation). This happens when one variable often appears higher when the other appears lower and vice-versa. Motorists and Pedestrians have more than double the correlation than that of Pedestrians and Cyclists. This, to some extent, is intuitive because the number of motorists injured/killed in a collision with pedestrian(s) will be most likely smaller than the number of cyclists injured/killed in a collision with pedestrian(s).

Post this discovery, I wanted identify the geographic areas that were unsafe for different types of commuters. The next three images are of NYC marking extremely severe accidents in red, moderately severe accidents in green and small accidents in gray. I could have only included highly sever accidents and thus identify some areas for different commuters but I didn’t want to ignore the less severe collisions simply because those were the more frequent ones and ignoring those could lead to an inaccurate inference. We see that Bronx is largely the area that is most unsafe for Motorists while Manhattan is the most unsafe for both Cyclists and Pedestrians. The area in Brooklyn that is unsafe is roughly the same for all kinds of commuters. This is important insights for policy-makers when making decisions regarding traffic control and hazard control.



# Concluding remarks

Based on all the observations I concluded that there are two sides - the commuters (basically everyone!) and the Policy Makers or the Municipal Authorities - that need to be addressed with the ‘key takeaways’.

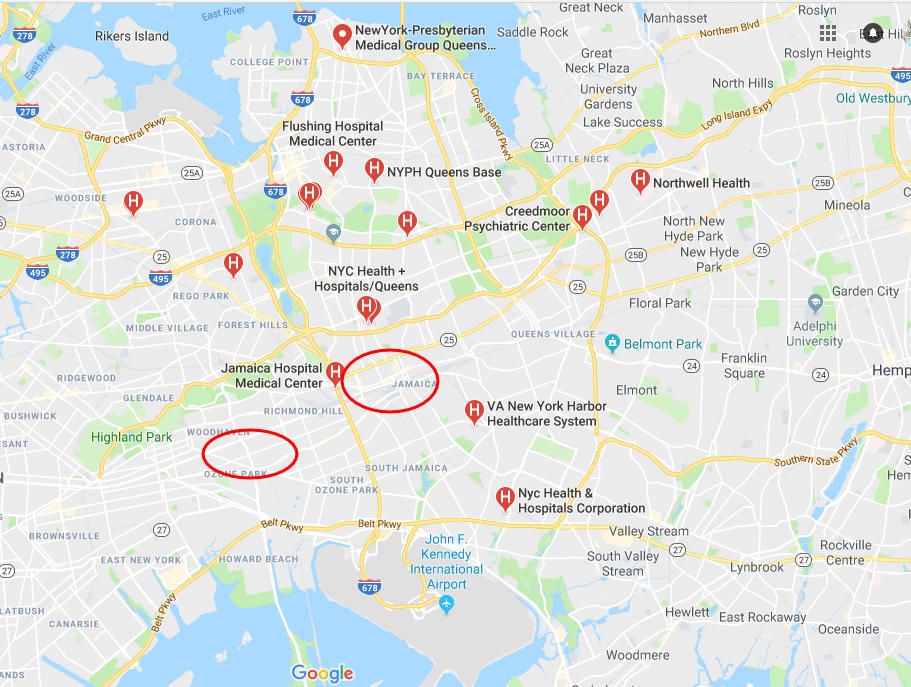
Municipal Authorities

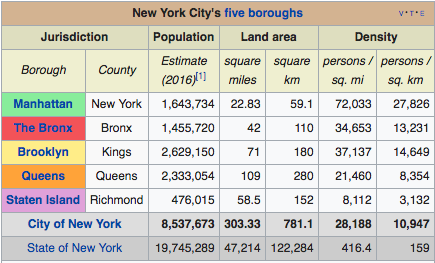
1. ‘Backing Unsafely’ is one of the major reasons for collisions and is the most impactful during mornings. The authorities should make rules guarding this and potentially, in long-term, restructure specific regions where it is difficult to ‘back’.
2. Another interesting insight that should be introspected is that the accident prone areas are close to the borders of the boroughs. We see this in both Bronx-Manhattan and Queens-Brooklyn borders.
3. Locations of hospitals should be noted with respect to the accident-prone areas and where needed hospitals should be constructed, and/or speedy emergency services provided. For example, the region around *Ozone Park* is accident prone but there’s no hospital close to it.
4. Similarly, Police Stations should also be strategically located with respect to accident-prone areas. In case of absence, police teams should be positioned here during collision prone times. For example, the street around *Saint Mary’s Park* are accident-prone but there’s no Police station nearby.
5. Cycling collisions aren’t affected by ‘fatigue’ and so the authorities should encourage public bike transport during the afternoon by reducing prices in these hours.
6. ‘Alcohol Involvement’ is a menace during midnight hours. Such collisions can be reduced by employing stricter laws in all boroughs other than Manhattan.

Commuters

1. Afternoon time is the most unsafe time for commuting and it has been observed that ‘Fatigue’ forms the main reason of collisions. Therefore, commuters are advised to take a short break before commuting and be vigilant during these hours.
2. During midnight, commuters in boroughs other than Manhattan should keep in mind that there is a high likelihood of some people driving under influence of alcohol.
3. Cycling is encouraged as the cyclist data suggests a lower number of collisions due to ‘Fatigue’, which is one of the main reasons of collisions for motorists.
4. In general, everyone should have an idea of the accident-prone areas as marked on the Google Maps (in the Appendix).
5. Although obvious, as the correlation showed, Pedestrians should be more conscious about Motorists than Cyclists when crossing roads.

The accident-prone areas are identified below (per borough). I’ve also marked the police stations and hospitals in the figures to get some perspective of which of the identified areas are relatively more dangerous. All such images are included in the appendix.

While New York City remains to be the corporate capital of the world, every life is valuable and lives lost in pursuit of this work in vehicle collisions will hopefully be diminished gradually.

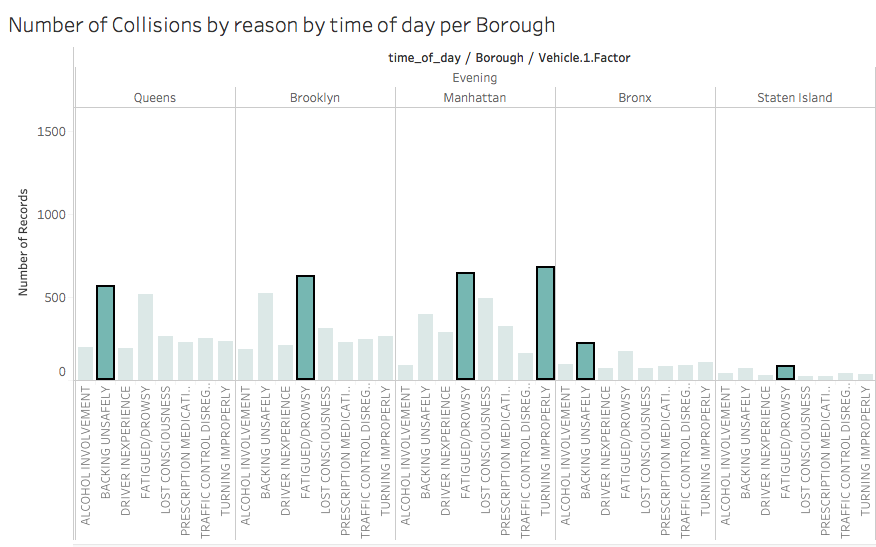
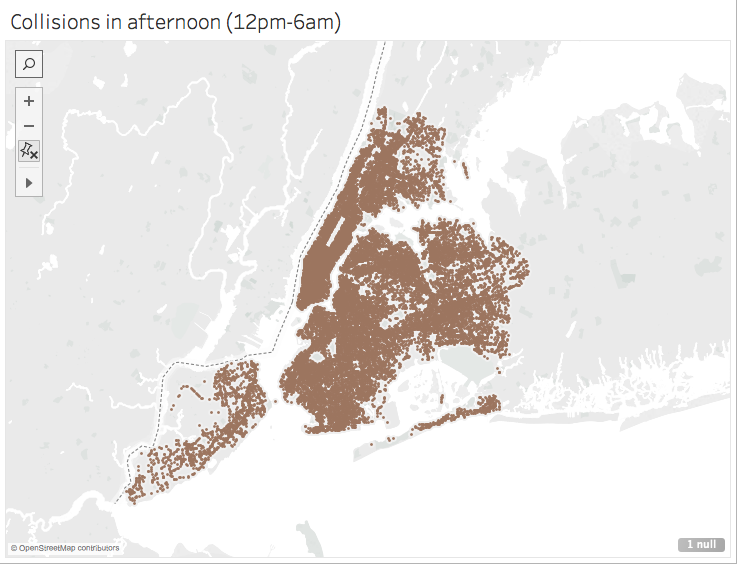


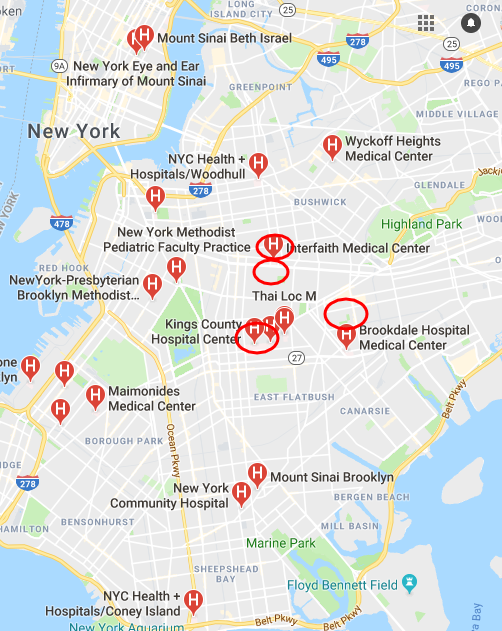
# Bibliography and Appendix

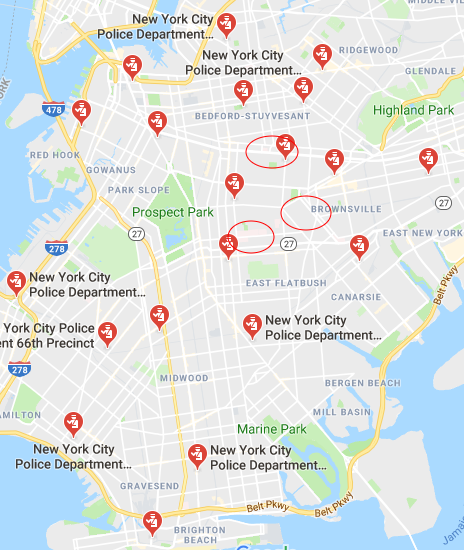
<http://www.syracuse.com/weather/index.ssf/2015/03/traffic_deaths_snow_sleet_upstate_new_york_united_states.html>

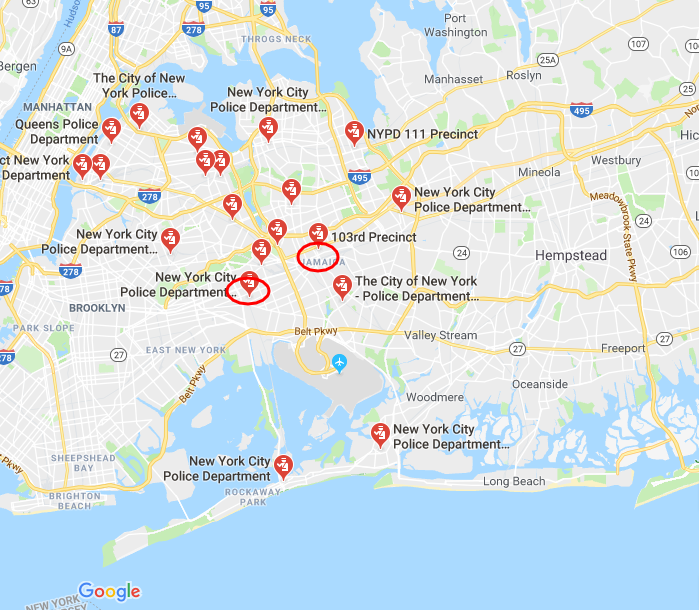
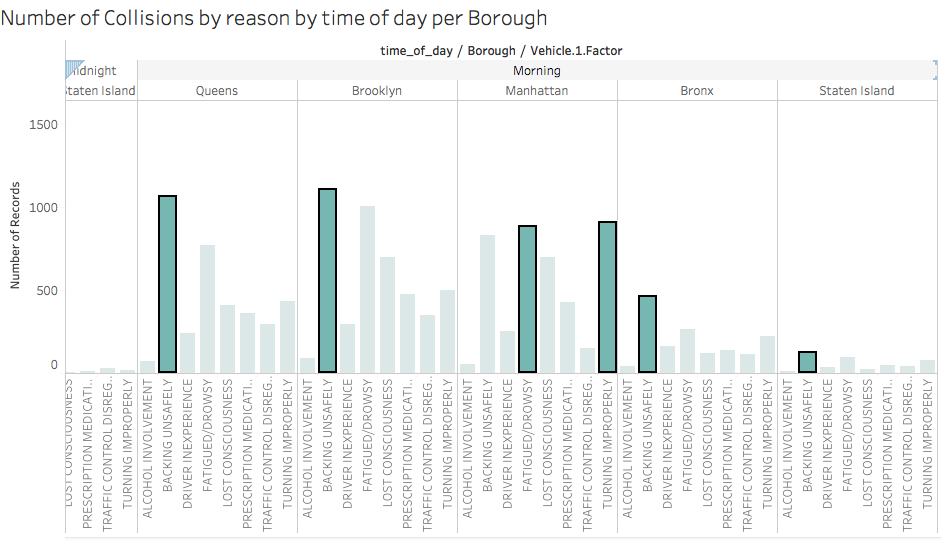
Temperatures table: <https://www.nyc.com/visitor_guide/weather_facts.75835/>

Precipitation graph: <https://weather.com/en-CA/weather/monthly/l/USNY0996:1:US>









Technical Appendix

# use of setwd() to get to the working directory

# reads from the csv file, stores into a dataframe called data\_set

data\_set = read.csv("NYC-vehicle-collisions.csv")

# data cleaning by removing the unnecessary variables, faulty records etc

# removes the two variables that were not needed i.e. ID (unique) and Code (all same)

# basic statements without any real analysis

data\_set = data\_set[,-1]

data\_set = data\_set[,-7]

# removes first three rows because the ‘dates’ of these were in a wrong format that didn’t

# make sense – this was perhaps due to a change of system of recording collisions. Since this

# discrepancy was negligible (3 records out of ~500,000), they were removed for the sake of uniformity

data\_set = data\_set[-(1:3),]

# removes variables by their name that were unimportant

data\_set = data\_set[ , -which(names(data\_set) %in% c("OFF.STREET.NAME","VEHICLE.2.FACTOR","VEHICLE.3.FACTOR", "VEHICLE.4.FACTOR", "VEHICLE.5.FACTOR","VEHICLE.3.TYPE", "VEHICLE.4.TYPE", "VEHICLE.5.TYPE"))]

# made the ‘dplyr’ library available which was pre-installed

library(dplyr)

# used the ‘filter’ method of dplyr to remove all the records that didn’t have a borough specified

# previously used the Excel function to convert Queens to 1, Brooklyn to 2 and so on…

main\_boroughs = dplyr::filter(data\_set, BOROUGH==1 | BOROUGH==2 | BOROUGH==3 | BOROUGH==4 | BOROUGH==5)

# use of dplyr’s ‘mutate’ and the ‘gsub’ method to add another column that only contains the hour

# Eg: 11:45 gives me 11 and 0:15 gives me 0. The dataset already followed a 24-hour time

main\_boroughs = dplyr::mutate(main\_boroughs, hour = gsub("^(.\*?):.\*", "\\1", main\_boroughs$TIME))

# Similarly used ‘substr’ along with ‘mutate’ to substring out (fetch) the month from the date string

# and store in a different variable/column ‘month’

main\_boroughs = dplyr::mutate(main\_boroughs, month = substr(main\_boroughs$DATE, 1, 2))

# Since the hour was a string, it needed to be converted to a number for comparison so:

main\_boroughs$hour = as.numeric(main\_boroughs$hour)

# Exploring the patterns of the large dataset through subsets based on various factors

midnight = dplyr::filter(main\_boroughs, hour >= 1 & hour <= 6)

morning = dplyr::filter(main\_boroughs, hour > 6 & hour <= 12)

afternoon = dplyr::filter(main\_boroughs, hour > 12 & hour <= 18)

# evening = dplyr::filter(main\_boroughs, hour >= 19 & hour <=0)

# didn't work cause the numbers don't make sense, instead:

evening = dplyr::filter(main\_boroughs, !(hour >= 1 & hour <= 18))

# Similarly, month is converted from a string character to a number capable of being compared

main\_boroughs$month = as.numeric(main\_boroughs$month)

# Extra dataset (very small) that I found online just to make sure the months of different

# seasons from 2015 – 2017. Link: <https://www.nyc.com/visitor_guide/weather_facts.75835/>

# Reads the csv file that I saved in the same directory

temperatures = read.csv("NYC-temperatures.csv")

# In order to join the two datasets, I changed the name of the variable ‘month’ to ‘Month’ in

# our main dataset. Next I used dplyr’s inner\_join method to join the two datasets and called our

# new dataset ‘main\_boroughs\_temp’

names(main\_boroughs)[names(main\_boroughs) == "month"] = "Month"

main\_boroughs\_temp = dplyr::inner\_join(main\_boroughs, temperatures, by = "Month")

# Use of mutate to add a variable ‘time of day’ that takes in values based on the hour column

# used ‘ifelse’ to write the conditions

main\_boroughs\_temp = dplyr::mutate(main\_boroughs\_temp, time\_of\_day = ifelse(hour %in% 0:6, "Midnight",

ifelse(hour %in% 6:12, "Morning",

ifelse(hour %in% 12:18, "Afternoon","Evening"))))

# Cleaning data from the new bigger dataset – Avg Precipitation is roughly the same the whole year

# And the analysis doesn’t really use the record-lows and highs of temperatures

main\_boroughs\_temp = main\_boroughs\_temp[ , -which(names(main\_boroughs\_temp) %in% c("Avg..Precip.","Record.Low", "Record.High"))]

# use of mutate + ifelse to add another variable called season and fill it up based on ‘Month’s’ values

main\_boroughs\_temp = dplyr::mutate(main\_boroughs\_temp, season = ifelse(Month %in% 5:8, "Summer",

ifelse(Month %in% 9:11, "Fall","Winter")))

# At this point, I used tableau to see the most common reasons for collision and

# cleaned my main dataset further to eliminate the insignificant reasons

main\_boroughs\_temp = dplyr::filter(main\_boroughs\_temp,

VEHICLE.1.FACTOR == 'ALCOHOL INVOLVEMENT' |

VEHICLE.1.FACTOR == 'BACKING UNSAFELY' |

VEHICLE.1.FACTOR == 'DRIVER INEXPERIENCE' |

VEHICLE.1.FACTOR == 'DRIVER INATTENTION/DISTRACTION' |

VEHICLE.1.FACTOR == 'TURNING IMPROPERLY' |

VEHICLE.1.FACTOR == 'TRAFFIC CONTROL DISREGARDED' |

VEHICLE.1.FACTOR == 'FAILURE TO YIELD RIGHT-OF-WAY' |

VEHICLE.1.FACTOR == 'FATIGUED/DROWSY' |

VEHICLE.1.FACTOR == 'PRESCRIPTION MEDICATION' |

VEHICLE.1.FACTOR == 'LOST CONSCIOUSNESS')

# Explore by subsetting the dataset into temporary dataframes, this time based on ‘Reason’

# for collision. Use of ‘grep’ to get the key phrase used in the ‘Reason’ column

# (VEHICLE.1.FACTOR). The key phrase forms the basis of a temporary subset.

# ‘Inattention’ and ‘Distraction’ were used randomly (even together) for different records

# since both refer to one reason, the subset dataframe was obtained as follows:

factor\_inattention = main\_boroughs\_temp[ grep("INATTENTION", main\_boroughs\_temp$VEHICLE.1.FACTOR),]

tmp\_distract = main\_boroughs\_temp[ grep("DISTRACTION", main\_boroughs\_temp$VEHICLE.1.FACTOR),]

factor\_inattention = dplyr::union(tmp\_distract, factor\_inattention)

factor\_inattention = dplyr::distinct(factor\_inattention)

# Continued exploring by making subsets for other major reasons

factor\_alcohol = main\_boroughs\_temp[ grep("ALCOHOL", main\_boroughs\_temp$VEHICLE.1.FACTOR),]

factor\_backing = main\_boroughs\_temp[ grep("BACKING", main\_boroughs\_temp$VEHICLE.1.FACTOR),]

factor\_inexperience = main\_boroughs\_temp[ grep("INEXPERIENCE", main\_boroughs\_temp$VEHICLE.1.FACTOR),]

factor\_turning = main\_boroughs\_temp[ grep("TURNING", main\_boroughs\_temp$VEHICLE.1.FACTOR),]

factor\_drowsy = main\_boroughs\_temp[ grep("DROWSY", main\_boroughs\_temp$VEHICLE.1.FACTOR),]

factor\_trafficControl = main\_boroughs\_temp[ grep("TRAFFIC CONTROL", main\_boroughs\_temp$VEHICLE.1.FACTOR),]

factor\_medication = main\_boroughs\_temp[ grep("MEDICATION", main\_boroughs\_temp$VEHICLE.1.FACTOR),]

factor\_consciousness = main\_boroughs\_temp[ grep("CONSCIOUSNESS", main\_boroughs\_temp$VEHICLE.1.FACTOR),]

factor\_yield = main\_boroughs\_temp[ grep("FAILURE TO YIELD", main\_boroughs\_temp$VEHICLE.1.FACTOR),]

# Next, I explored the dataset by commuter type and the severity in terms of injured/killed persons

cyclists = dplyr::filter(main\_boroughs\_temp, `CYCLISTS.KILLED` != 0 | `CYCLISTS.INJURED` != 0)

cyclists\_injured = dplyr::filter(main\_boroughs\_temp, `CYCLISTS.INJURED` != 0)

cyclists\_killed = dplyr::filter(main\_boroughs\_temp, `CYCLISTS.KILLED` != 0)

motorists = dplyr::filter(main\_boroughs\_temp, `MOTORISTS.KILLED` != 0 | `MOTORISTS.INJURED` != 0)

motorists\_injured = dplyr::filter(main\_boroughs\_temp, `MOTORISTS.INJURED` != 0)

motorists\_killed = dplyr::filter(main\_boroughs\_temp, `MOTORISTS.KILLED` != 0)

pedestrians = dplyr::filter(main\_boroughs\_temp, `PEDESTRIANS.KILLED` != 0 | `PEDESTRIANS.INJURED` != 0)

pedestrians\_injured = dplyr::filter(main\_boroughs\_temp, `PEDESTRIANS.INJURED` != 0)

pedestrians\_killed = dplyr::filter(main\_boroughs\_temp, `PEDESTRIANS.KILLED` != 0)

# Made new variables that accounted for the sum of persons injured and killed per type of commute

# This was done to allow for simplified visualisations later.

main\_boroughs\_temp = dplyr::mutate(main\_boroughs\_temp, cyclists = CYCLISTS.INJURED+CYCLISTS.KILLED)

main\_boroughs\_temp = dplyr::mutate(main\_boroughs\_temp, motorists = MOTORISTS.INJURED+MOTORISTS.KILLED)

main\_boroughs\_temp = dplyr::mutate(main\_boroughs\_temp, pedestrians = PEDESTRIANS.INJURED+PEDESTRIANS.KILLED)

# At this point, many visualisations were made using Tableau

# Cleaning the dataset one last time to get only the records where *atleast 1* person got injured

# for some other results.

main\_boroughs\_temp = dplyr::filter(main\_boroughs\_temp, pedestrians!=0 | cyclists!=0 | motorists !=0)

# Observed some correlations and their result was quite intuitive with all of them

# being inversely related.

# be ignored. All of them have inverse (slight) relationships

cyc = c(main\_boroughs\_temp$cyclists)

ped = c(main\_boroughs\_temp$pedestrians)

mot = c(main\_boroughs\_temp$motorists)

cor(cyc, ped) # -0.2109586

cor(ped, mot) # -0.4573438

cor(cyc, mot) # -0.2756082

# Tried using ggplot but discovered that Tableau was more

# user-friendly so used that for most of my visualizations

# library(ggplot2)

# temp\_plot\_ = ggplot(factor\_distraction, aes(VEHICLE.1.FACTOR, PERSONS.KILLED))

# temp\_plot\_ + geom\_point()