NYC Motor Vehicle Collisions Data Exploration

Team HAKR, Group 5

Members:

Hao Deng, Yue ‘Alex’ Fu, Ravi Bhagwat, Kim-Cuong Nguyen

Introduction

Background

As the most populated city in the United States, New York City is also the most traffic-congested city in the US. With such a high volume of traffic and the city plan of NYC, there are a lot of vehicle collisions happening on a daily basis. According to statistics collected by the New York City Police Department (NYPD), there were 228,048 accidents reported in 2018; that's 19,000 accidents per month and 1 vehicle accident every 2 minutes. A survey done by AllState Insurance back in 2014 revealed that NYC drivers were 28.8% more likely to be in a vehicle accident than the national average. These statistics, along with the constant complaints of traffic by one of our team members who frequently travels to NYC, intrigued our interest to find out the reasons.

We brainstormed the possibilities of an accident and also did some relevant research. Interestingly, we have found that 'Distracted Driving' tops the list. In order to assess how serious the situation is, our group employs different types of data to have a correct view with a good amount of supporting evidence.

Goal

Our goal is to find out undiscovered consequences or patterns behind the collision data other than common senses (e.g. High collision rate over a rush hour, downtown Manhattan, etc.). We would also give our suggestions and if possible, notify motorists, pedestrians on the road to be careful during certain times of the day and at certain locations.

Data background

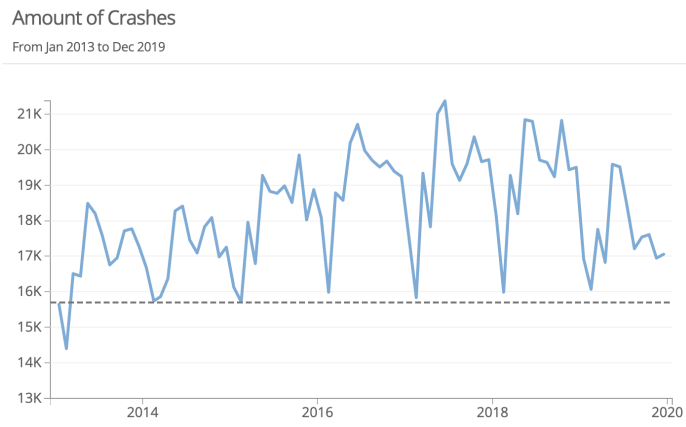
Motor Vehicle Collisions Dataset includes 3 separate datasets from the official website of the U.S. Government's open data - 'Vehicles', 'Crashes', and 'Person'. One row in 'Vehicles' dataset represents a vehicle that involved in a crash; Likewise, one row in 'Crashes' dataset represents an accident event and each row in 'Person' represents personnel (driver, passenger, pedestrian, bicyclist, cyclist, motorcyclist, etc.) involved in a crash.

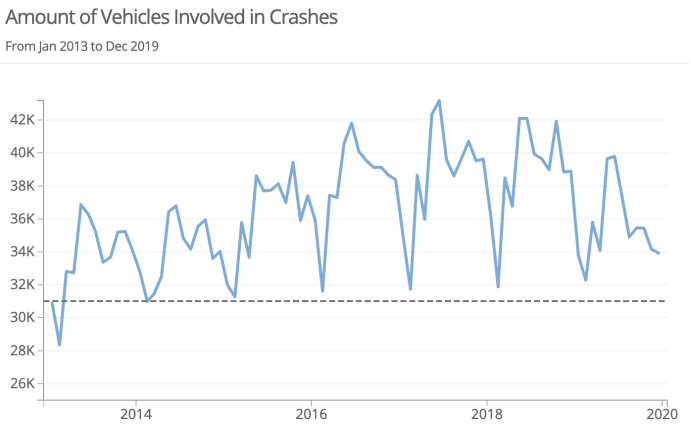
Original metadata contains data from July 2012 to January 2020.

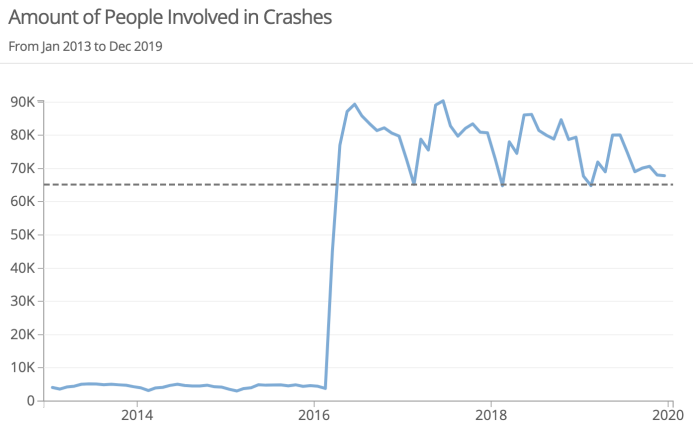
Data Summary

We are focusing on several problems based on the three datasets, but let's look at our dataset in a general way to oversee the overall situation in NYC at first. Due to the limited space and memory function on SAS University Edition, we cannot load the full dataset into SAS. Therefore, in order to get a better overall insight for the dataset, we applied the visualization tool on the official website NYC Open Data for the full version of datasets.

Firstly, by grouping the observations in 'Crashes' by month and year, and counting the distinct values of collision ID, we get the number of crashes in each month from January 2013 to December 2019. Repeating the same process for 'Vehicles' and 'Person', we plotted 3 figures as below:







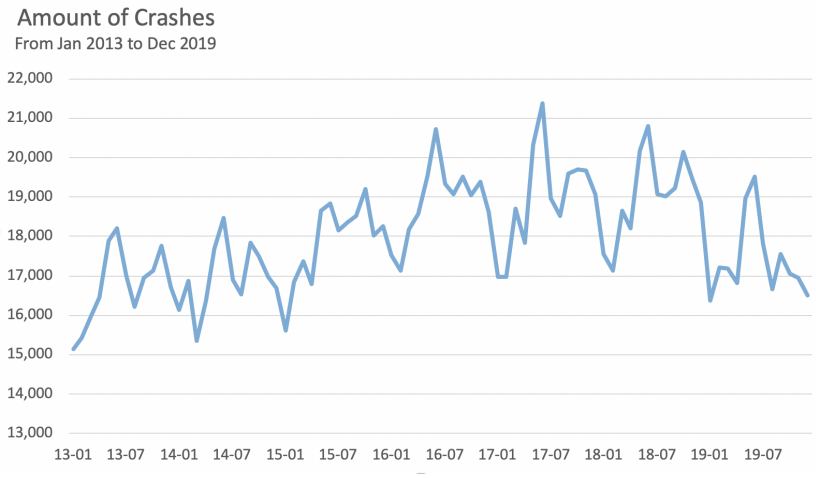
As we can see, the plot of 'Vehicle' is highly similar to the plot of 'Crashes' except the scale of the y-axis, which is the amount.

The number of people involved in crashes was minimal before March 2016, which reminds us that the New York Police Department (NYPD) in March 2016 replaced the Traffic Accident Management System (TAMS) with the new Finest Online Records Management System (FORMS). The update of the recording system might be the main reason for the weird pattern in the

plot of 'Person'. Therefore, we definitely should pay extra attention to the data changes in future analysis.

Looking into all the three plots, there is an interesting pattern for each year. A significant drop in February of every year. Here are two reasons that come from our analysis. One is the number of crashes, vehicles, and people involved in crashes truly decrease in February because of the season factor. The other reason is the number of days in February in the least among the whole year. 2 or 3 days means 6.67% or 9.68% differences, which are enough to impact the total number of crashes in the whole month.

Therefore, by scaling the days of all months to 30 days, the number of crashes for each month is isolated with the affection of the number of the days. We got the plot of 'Crashes' as the figure below:



Overall, the tendency of the data from 2013 to 2017 was increasing and reached the maximum in 2017. Starting in 2018, the overall level of the number of crashes started dropping. It is good news for NYPD since its goal for updating the system is to analyze the characteristics of crashes to prevent more crashes. We believe there is a delay and that's why the data dropped since 2018 yet not drop instantly in 2016.

By looking at the details, it's easy to find out that every year the number of crashes goes to the peak in June and gets to another bounce in around October or November. Moreover, there are three troughs in January, April, and August. The pattern might be related to the vacations and popular seasonal visiting to NYC. For example, June is the start of summer vacation, which means it is a good

time for parents and children to visit the most popular city in the world. In January, the weather is cold and tourists are more likely to travel somewhere warm instead of visiting NYC.

For future analysis, due to the limitation of SAS UE, we are going to split the three datasets into several subsets according to the year of the data and focus on the subset in 2019. Because the three datasets have a huge volume of observations; for example, almost 4 million for dataset 'Vehicles', we could not use either Excel or SAS to process the datasets. Thus, the team decided to use MS Access which is quite efficient in processing huge datasets and ran the built-in query to split the datasets by year.

Limitations

- **Limitation on the Size of datasets**

As mentioned above, Motor Collision Datasets contain data from 2014 to up to date. Our team downloaded datasets as JSON format. There were three datasets which were 'Crashes', 'Person', and 'Vehicles'. The sizes of the three files were 740.8MB, 1.38GB, and 1.27GB, respectively. Unfortunately, SAS University Edition was not capable of loading the size of the full dataset as large as three data files. Temporary memory was already full before loading the data successfully and SAS cannot perform any other tasks due to such limitation. Our team also attempted to explore datasets in CSV format with Excel and RStudio.

In Excel, we experienced an unknown reason of data loss while we were exploring the Vehicles dataset. Observations from June 2018 to August 2019 were missing directly from the original dataset without any interventions from the team member. Completeness of the dataset was proved after we broke the dataset down by year through RStudio.

Eventually, we were able to explore datasets without losing data in RStudio. With over 1GB size of datasets, we decided to narrow down by year due to overcapacity of laptops while running dataset summary in RStudio.

- **Data type**

In all datasets, numeric variables, such as *Collision ID*, *Numbers of people injured*, *Age*, *date*, etc., are available to use without turning into other data formats. However, character variables, such as *Contributing factors*, *Points of impact*, *Position of pre-crash*, etc., needed further attention to be transformed into radical variables or binary variables for regression analysis. Downsizing of character variables by turning them into numeric variables is an advantage for our team to load the dataset efficiently.

- **Inconsistent text input**

Among three datasets, variables like *Vehicle type*, *Person Sex, Contributing factors*, and *Position of pre-crash* were facing the problem of inconsistent text input. In *Vehicle type*, we found out there were over 400 types of vehicles existing overall observations. Some text inputs were the full name of the vehicle type, yet some names were abbreviations. For example, Police officers may put “Ambu” as standing for “Ambulance”, “Pick-up” for “Pick-up Truck”, etc. Miscellaneous and inconsistent text input names were giving us a hard time to filter or count the same type of variable with catching all of them.

- **Creating a prediction model**

It is less possible to create a collision prediction model by applying variables in all datasets. A vehicle collision contains many contributing factors including multiple factors from the driver, weather condition, traffic condition, vehicle condition, etc. Many factors are missing from the dataset. And the format of some factors is needed to transform even if they are recorded in the dataset. Driver information is a critical factor to predict vehicle collision. The driver’s driving behavior, personality, and emotion are missing from the dataset and we can’t predict who is going to drive. Current data is not capable of predicting vehicle collisions that we could only analyze and conclude hazardous locations, high-frequency timing, and other related collision patterns.

Focus & Findings

Problem Statement 1:

Is there any seasonality effect (time during a day, days during a week, months during a year) on collisions among each kind of vehicles (buses, bikes, taxis, passenger cars, etc.)?

Data Processing:

In order to address this problem, the team utilized only 'Vehicles' dataset because this dataset could provide enough information on *Crash Date, Crash Time* and *Vehicle types*.

Crash time & date: Based on the *Crash Date*, we ran the code to find weekdays and months. We reformatted the time to 24-hour format with the value of hour only instead of the hour and minute.

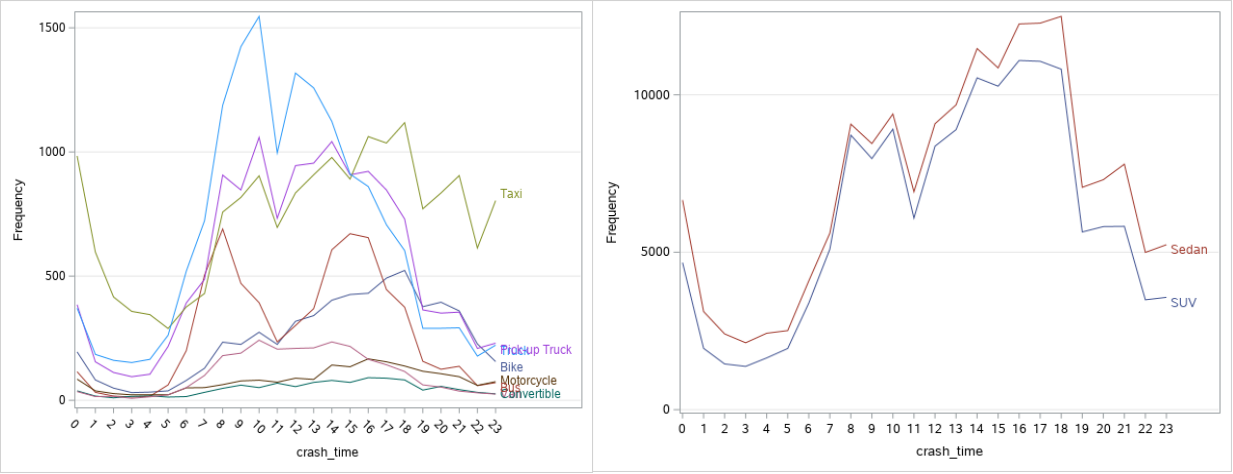
Vehicle types: this variable has lots of noise with over 500 types out of more than 400,000 observations, most of which have a frequency of 1 only. The reason is that the input value for vehicle types has an inconsistent format and it seems that police officers could input random values. At a quick glance, the values with a frequency lower than 100 do not provide valuable meanings. Thus, our team first created a subset using only *Vehicle Types* with a frequency of more than 100, reducing the number of vehicle types to 35 and the number of observations by about 30,000. Then we re-classified vehicle types into more general groups; for example, we grouped 2-door sedans and 4-door sedans into sedans. We decided to analyze 10 types with the highest crash frequency considering the low crash numbers involving other types.

Results:

All 10 vehicle types are taken into consideration include bikes, buses, convertibles, motorcycles, pick-up trucks, sedans, SUVs, taxis, trucks, and vans.

We plotted crash frequencies of these 10 vehicle types in accordance with crash weekdays, crash months, and crash time for further analysis. There was a big gap between the number of accidents involving sedans & SUVs and other vehicles; thus, we plotted the charts for sedans and the 9 remaining types separately.

Crash time:



According to the figures above, the number of crashes involving all the 10 vehicles is high between 7 am to 6 pm. Because this is the time when everyone goes to school or to work and thus, there is a high volume of traffic on the road. The most common contributing factors to accidents are driver distraction, passing/following too closely, and improper lane usage. The crash patterns among different vehicle types are as below:

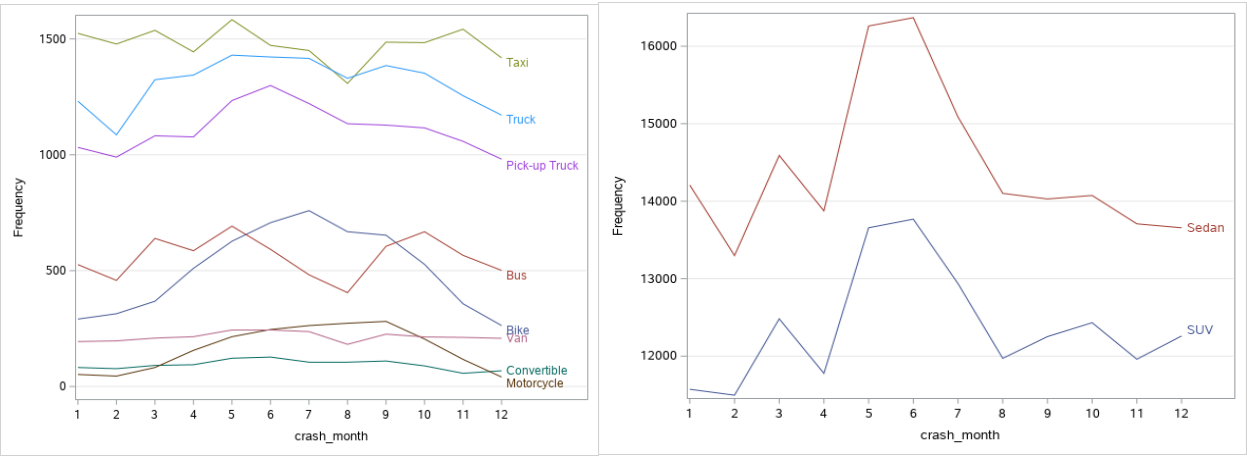
- Convertibles and Motorcycles shared similar patterns. The number of crashes, related to ambulance, convertibles, and motorcycles, increased during the daytime but there was no sudden rise in any specific time.
- Pickup trucks & Vans had a lot more accidents during daytime than nighttime, but the number of crashes among different hours was similar and there was no peak at any specific time.
- Bikes, SUVs, and Sedans had crashes that occur mostly in afternoon rush hours, between 5 and 6 pm.
- Bus crashes happen most frequently around 8 in the morning, and between 3 and 4 in the afternoon. Bus crashes happened more before the afternoon rush hours. Presumably, school buses often drop students off between 3 and 4 in the afternoon. Also, concerning tourist buses, they tend to drop off tourists and finish the tours before 4 or 5 to avoid rush hours. Plus, during rush hours, buses usually have their own lanes, and thus, can better avoid getting into crashes with other vehicles.
- Taxis: most of the accidents related to taxis happened in the afternoon rush hour – 4 pm and at midnight. The accidents at midnight could be explained by the fact that people have more tendency to call cabs when they need to go out at midnight for safety reasons. Also, taxi shift change in

NYC traditionally occurs between 4 and 5 pm, so drivers might feel fatigued at midnight and were more likely to cause accidents. In addition, taxis could get higher demands at midnight; taxi drivers might drive faster and carelessly in order to deliver more passengers as many as possible for better earnings.

- Trucks had more accidents that occur in the morning than in the afternoon. Trucks here include different kinds of trucks, mainly garbage trucks and semi-trucks. Morning hours are the main times when these trucks participate in traffic. In order to reduce the number of truck crashes, we recommend that delivery trucks/garbage trucks can shift their operating hours into night hours instead and leave the city before rush hours. Trucks are big and have many blind spots while rush hours are the time with significantly high volumes of vehicles on the road. The fact that fewer trucks on the road can help reduce the number of crashes in general. However, this recommendation also brings some disadvantages. Some deliveries can be made during the night, but some must be completed during the daytime instead. Also, if night-time shifts are performed, higher pay must be made to the people of concern, which would put higher potential financial stress on related businesses.

We could conclude that rush hour is a particularly vulnerable time for all drivers to be on the road with any kind of vehicles and afternoon rush hours are generally worse than in the morning.

Crash month:

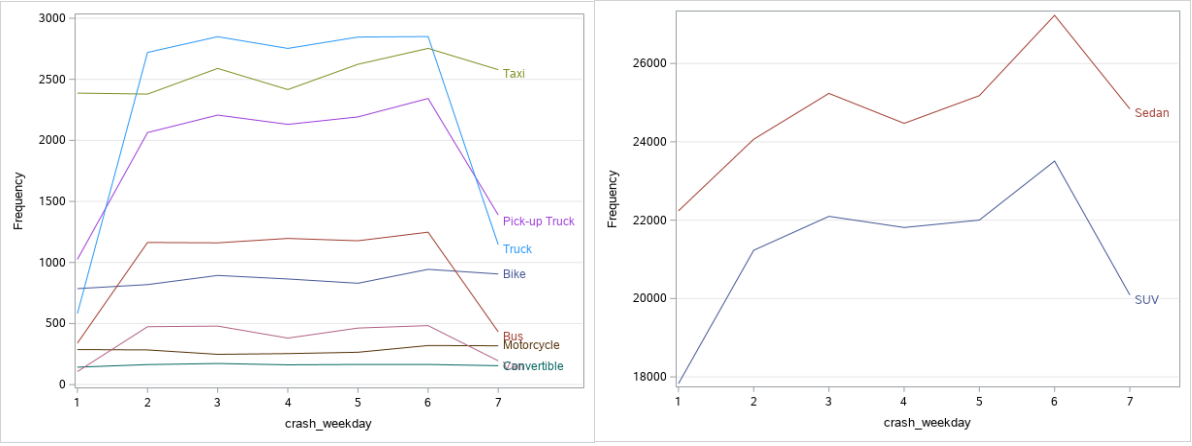


Buses and Taxis shared similar patterns, in which the number of crashes is related to these two vehicles dropped in August. Both buses and taxis are shared vehicles. Convertibles and vans have accidents at a steady rate around the year.

All remaining vehicle types had more accidents occur between April and October. The higher number of crashes by bikes and motorbikes during this period results from the fact that people tend to use the vehicles more during summer and fall because of the nice weather; Winter is not ideal for cyclists and motorcyclists. Concerning trucks, pickup trucks, SUVs, and sedans, we assume that drivers tend to be more careless during the period from April to October when weather conditions are more favorable for driving.

Interestingly, winter is supposed to be a better time for drivers.

Crash weekday:



Bikes, motorbikes, and convertibles had crash frequency similar throughout the week.

Trucks, Pickup trucks, Buses, and Vans had more accidents during weekdays than weekends and the number of crashes per weekday was similar. This is because there are less of those vehicle types on the road on weekends.

Sedan and taxis shared similar patterns when the number of crashes they involved were high between Tuesday and Saturday and peaked on Friday. This can be attributed to higher demand for taxis and more usage of sedans on Fridays and Saturdays when people tend to hang out more. When we were looking more into the contributing factors to crashes on these two days of taxis and sedans, 'failure to yield right-of-way' is the factor that accounted for a significant percentage beside the earlier mentioned common factors.

Not surprisingly, Friday is the most dangerous to drive while Sunday is a good day for driving.

Problem Statement 2:

Are there any additional factors that lead to a multi-vehicle collision?

Background:

In daily life, there are always some serious accidents that involve several cars, which costs a great loss and even takes a lot of lives. Therefore, exploring multi-vehicle collisions is meaningful.

Data Processing:

We are going to use 'Crashes' and 'Vehicles' to find the multi-vehicle collision. For the dataset 'Crash', we import it from the JSON file, drop the error variables, which are created by SAS because of the technical issue, adjust the datatype, and rename the variables.

A new variable: *Vehicle_num*

To get the number of vehicles involved in each crash, we are going to use the variables *Vehicle Type Code* 1 to 5, and *Contributing Factor Vehicle* 1 to 5 in 'Crashes'.

In fact, either *Vehicle Type Code* or *Contributing Factor Vehicle* could give us the number of vehicles, but considering the missing values and inconsistent between the two kinds of variables we are going to choose the larger value of the number of vehicles calculated by the two kinds of variables. Therefore, we created 10 more variables to check the 10 original variables respectively. If there is content in the cell, we put 1 in the related variables. if there is blank, we put zero in the related variables. Then we add the variables to two more columns to show the number of vehicles counted by *Vehicle Type Code* 1 to 5 and *Contributing Factor Vehicle* 1 to 5. Then create an additional variable as *Vehicle_num,* we use the function 'max' to get the larger one as the final number of vehicles involved in one collision. Finally, drop all original variables and additional variables, and leave *Vehicle_num*.

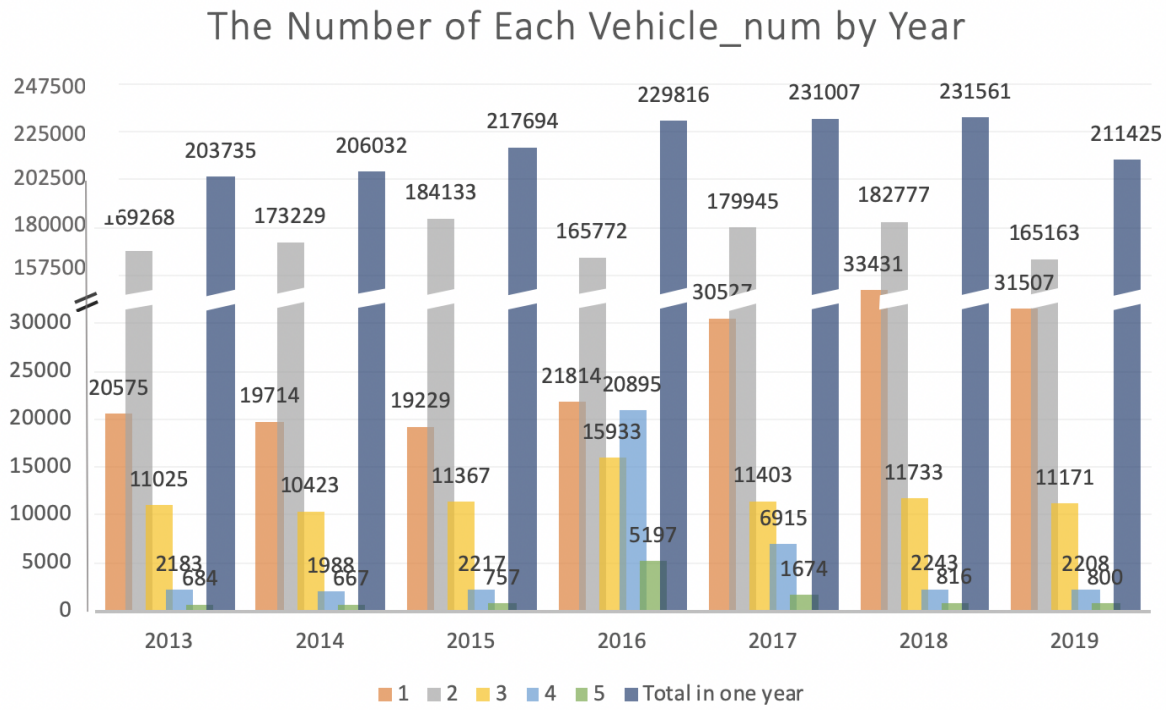
A new table: *Multi-vehicle Collision*

Since the dataset 'Vehicle' is too large to load in SAS, we decide to focus on the cases in 2019. Using the subset 'Vehicle-2019' as CSV version, we did inner join between 'Crashes', where the *Vehicle_num* is larger than 2 (defined as Multi-vehicle collision), and 'Vehicle-2019' on *Collision ID*, and

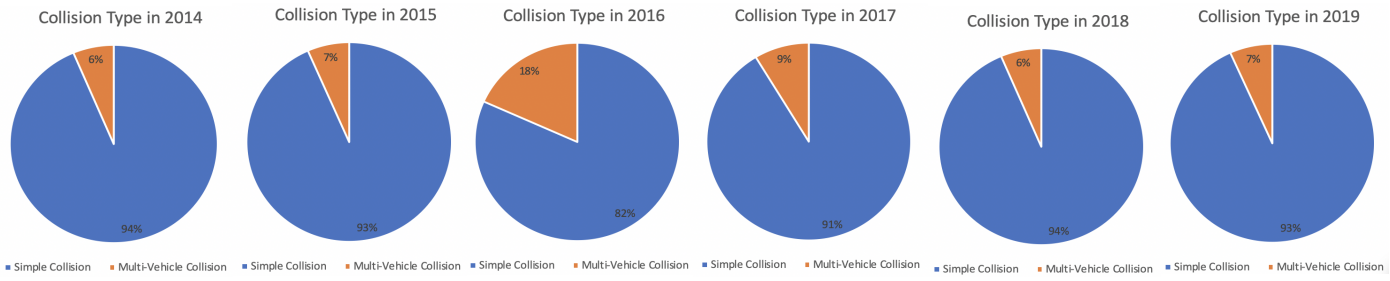
keep the column *Collision ID, Vehicle_num, time, Borough, Travel Direction, Vehicle Occupants, Driver Sex, Pre_crash, ppl_injured, ppl_killed.*

Results and Insights:

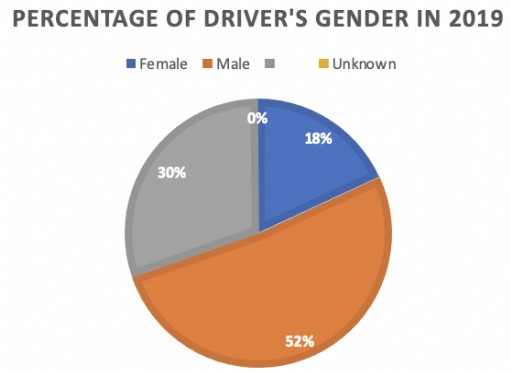
Due to the limited space and memory function of SAS UE, we view the 'Crash' dataset by applying 'where' clause and 'year' function to filter the data for the specific year and *Vehicle_num*, and record the counts of data in Excel to get a plot as below:



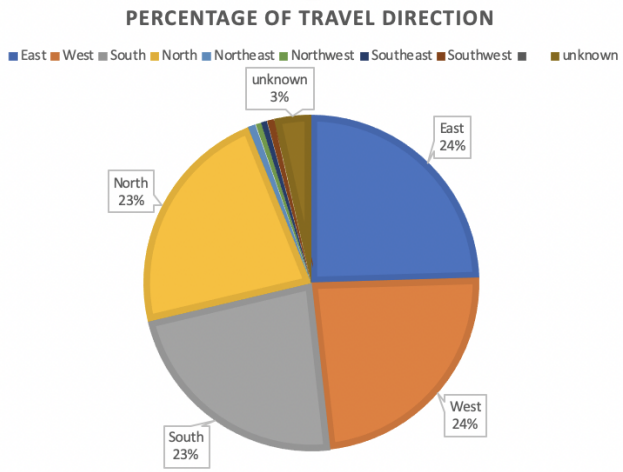
The overall tendency aligns with what we mentioned before in the data summary. It is much more common when only 1 or 2 cars are involved in a crash. There is a drop in 2016 for *Vehicle_num* equals 2, while the number of crashes for *Vehicle_num* larger than 2 was bumped out, which can be observed more clearly in the set of plots by just looking at the orange areas.



In multi-vehicle collision cases, we can see that except for the unknown gender, Male is almost three times than Female as the driver of the involved vehicles. For some reason, the unknown data is quite a lot.

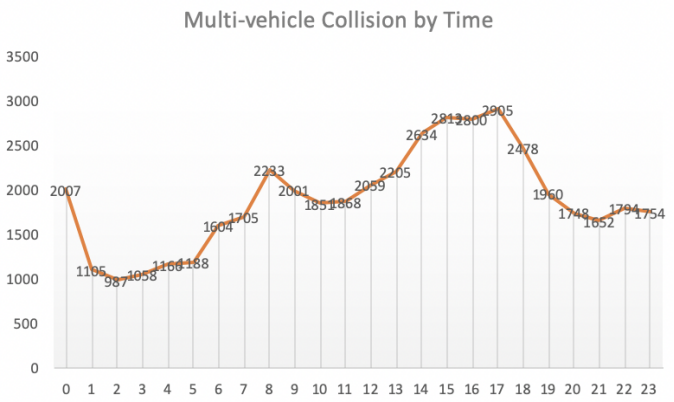


In the plot 'percentage of travel direction', we can see that the four main directions have evenly separated shares of the total. Therefore, the direction of traveling has nothing to do with the multi-vehicle collisions.



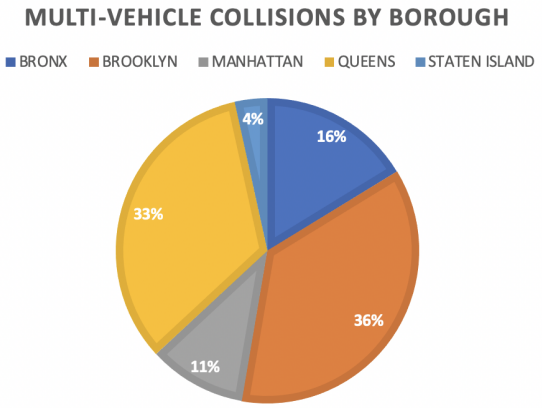
According to the plot 'Multi-vehicle Collision', we can see the multi-vehicle collisions are more common in the afternoon from 14 pm to 18 pm. Also, there is a little bump at 8 am.

Therefore, we got one tip that multi-vehicle collisions tend to happen in the rush



hour, especially in the afternoon. We can make an assumption that people get tired after an entire day of work, and it is hard to concentrate on driving on the way home.

By grouping the data based on the borough, we can easily find out that Queens and Brooklyn are the two areas having more frequent multi-vehicle collisions.



Problem Statement 3:

Is there any collision related to vehicle driving direction, geographical location and time of the day? (Sunlight effect to rush hour driving)

Background:

This problem statement is mainly targeted to find out whether sunrise or sunset would cause vehicle collision. However, there is no weather condition by day and time due to the limitation of our datasets. According to BestPlace.net as the figure below, New York City has higher average sunny days compared to the United States average. We assume the sun is not blocked by clouds in this case.

Climate Averages

	New York, New York	United States
Rainfall	46.6 in.	38.1 in.
Snowfall	25.3 in.	27.8 in.
Precipitation	118.8 days	106.2 days
Sunny	224 days	205 days
Avg. July High	84.2°	85.8°
Avg. Jan. Low	26.1°	21.7°
Comfort Index (higher=better)	7.3	7
UV Index	3.8	4.3
Elevation	33 ft.	2443 ft.

Data Processing:

In this problem statement, we are focusing on accidents related to geographical locations, the direction of the car and the time of the accident. We selected 'Vehicles' and 'Crashes' dataset with **Manhattan Borough** as our major focus datasets as Manhattan has the island shape directly facing west and east for both sides. So, it is easier to observe from the analysis whether sunlight would have a correlation to car accidents. Within 'Crashes' we kept 4 variables, *Crash date*, *Crash time*, *Latitude* and

Longitude. Then in the 'Vehicles' dataset, we kept only one variable and removed all others, *Direction*, which is the car heading direction when a car accident occurs.

New Variables: *Transform of Direction*

To combine two datasets, sort by *Collision ID* was performed prior to the merge. After merging into one dataset, we need to turn *Direction* into numeric variables since it was character variable before, and a numeric variable was needed for regression & correlation analysis. *Direction* then was turned as "0 = East, 1 = Northeast, 2 = North, 3 = Northwest, 4 = West, 5 = Southwest, 6 = South, 7 = Southeast and 8 = Unknown". Yet, turning *Direction* into radical variables was not helpful in regression & correlation analysis. To further break down the *Direction* for analysis, we then created 9 new columns in the merged dataset and turned *Direction* into binary variables.

Results and Insights:

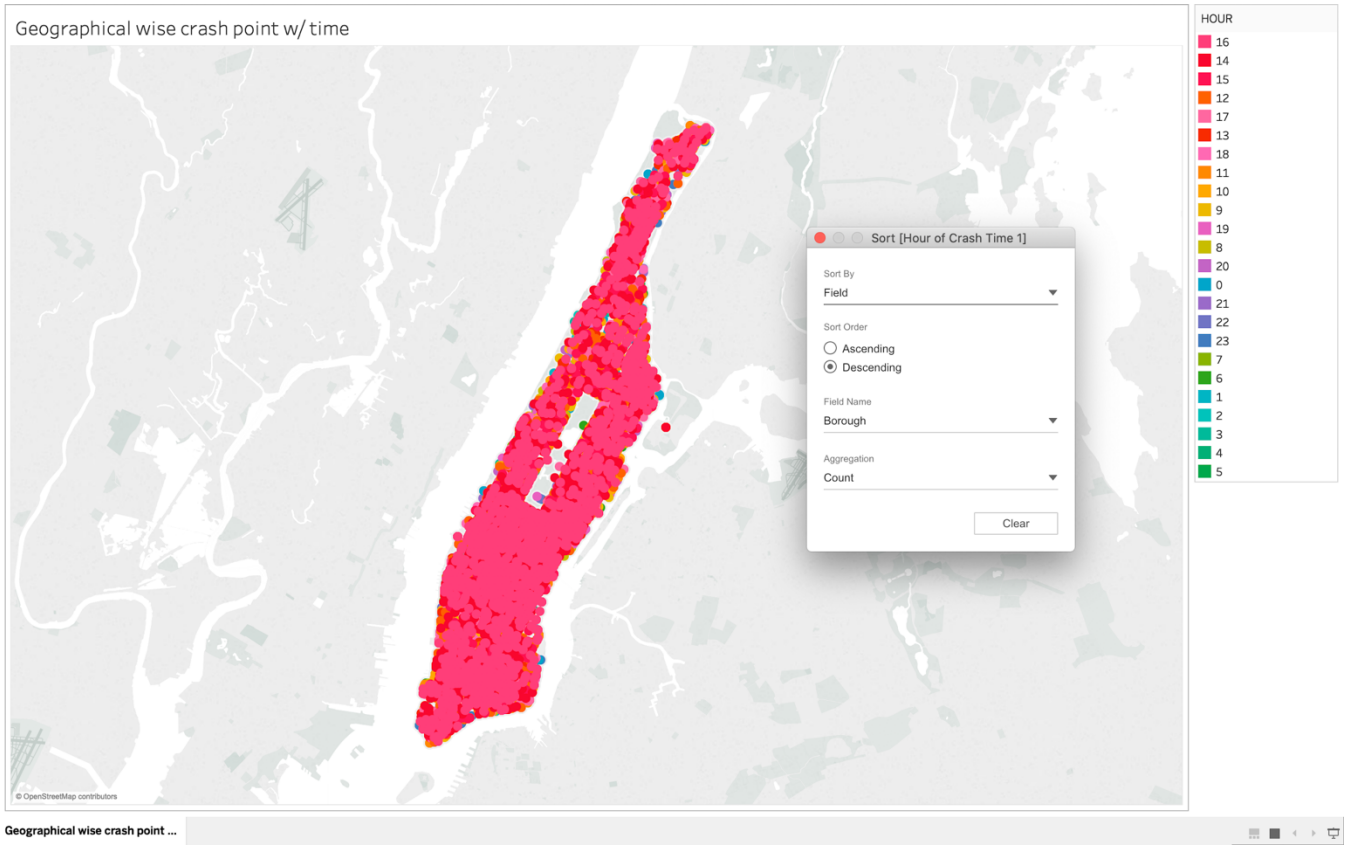
'PROC CORR' was performed for the merged dataset in SAS. As the figure is shown below, there was no evidence we can prove the direction of the car, the side of location by different time would cause a car accident. All variables have a correlation near 0.

Pearson Correlation Coefficients, N = 48286
Prob > |r| under H0: Rho=0

	CRASH_DATE	CRASH_TIME_1	LATITUDE	LONGITUDE	COLLISION_ID	Direction	East	Northeast	North	Northwest	West	Southwest	South	Southeast	Unknown
CRASH_DATE	1.00000	0.02249 <.0001	0.00614 0.1776	-0.00615 0.1763	0.99334 <.0001	0.00110 0.8091	0.00375 0.4098	-0.00286 0.5297	-0.00076 0.8679	-0.00471 0.3006	0.00657 0.1486	-0.00677 0.1369	-0.00361 0.4278	-0.00406 0.3728	-0.00220 0.6289
CRASH_TIME_1	0.02249 <.0001	1.00000	-0.00671 0.1403	0.00653 0.1511	0.02473 <.0001	-0.00338 0.4571	-0.00661 0.1466	-0.00631 0.1655	0.00974 0.0323	-0.00555 0.2230	-0.00829 0.0687	0.00247 0.5869	0.01771 <.0001	-0.00312 0.4931	-0.02628 <.0001
LATITUDE	0.00614 0.1776	-0.00671 0.1403	1.00000	-0.99956 <.0001	0.00676 0.1372	0.00110 0.8087	0.00032 0.9447	-0.00831 0.0678	-0.01218 0.0074	0.00438 0.3356	0.01402 0.0021	-0.00436 0.3380	-0.00057 0.9004	0.00457 0.3153	-0.00077 0.8664
LONGITUDE	-0.00615 0.1763	0.00653 0.1511	-0.99956 <.0001	1.00000	-0.00672 0.1396	-0.00106 0.8162	0.00007 0.9883	0.00805 0.0770	0.01225 0.0071	-0.00477 0.2943	-0.01414 0.0019	0.00395 0.3857	0.00046 0.9192	-0.00475 0.2965	0.00115 0.7999
COLLISION_ID	0.99334 <.0001	0.02473 <.0001	0.00676 0.1372	-0.00672 0.1396	1.00000	0.00053 0.9078	0.00447 0.3259	-0.00299 0.5119	-0.00021 0.9640	-0.00571 0.2099	0.00614 0.1772	-0.00675 0.1378	-0.00460 0.3124	-0.00438 0.3353	-0.00061 0.8939
Direction	0.00110 0.8091	-0.00338 0.4571	0.00110 0.8087	-0.00106 0.8162	0.00053 0.9078	1.00000	-0.09757 <.0001	-0.01557 0.0006	-0.04067 <.0001	-0.00219 0.6302	0.01929 <.0001	0.01079 0.0177	0.07909 <.0001	0.02355 <.0001	0.03809 <.0001
East	0.00375 0.4098	-0.00661 0.1466	0.00032 0.9447	0.00007 0.9883	0.00447 0.3259	-0.09757 <.0001	1.00000	-0.06591 <.0001	-0.30011 <.0001	-0.06309 <.0001	-0.29203 <.0001	-0.06468 <.0001	-0.29550 <.0001	-0.06392 <.0001	-0.08118 <.0001
Northeast	-0.00286 0.5297	-0.00631 0.1655	-0.00831 0.0678	0.00805 0.0770	-0.00299 0.5119	-0.01557 0.0006	-0.06591 <.0001	1.00000	-0.06834 <.0001	-0.01437 0.0016	-0.06650 <.0001	-0.01473 0.0012	-0.06729 <.0001	-0.01455 0.0014	-0.01849 <.0001
North	-0.00076 0.8679	0.00974 0.0323	-0.01218 0.0074	0.01225 0.0071	-0.00021 0.9640	-0.04067 <.0001	-0.30011 <.0001	-0.06834 <.0001	1.00000	-0.06541 <.0001	-0.30278 <.0001	-0.06706 <.0001	-0.30638 <.0001	-0.06627 <.0001	-0.08417 <.0001
Northwest	-0.00471 0.3006	-0.00555 0.2230	0.00438 0.3356	-0.00477 0.2943	-0.00571 0.2099	-0.00219 0.6302	-0.06309 <.0001	-0.01437 0.0016	-0.06541 <.0001	1.00000	-0.06365 <.0001	-0.01410 0.0019	-0.06441 <.0001	-0.01393 0.0022	-0.01770 0.0001
West	0.00657 0.1486	-0.00829 0.0687	0.01402 0.0021	-0.01414 0.0019	0.00614 0.1772	0.01929 <.0001	-0.29203 <.0001	-0.06650 <.0001	-0.30278 <.0001	-0.06365 <.0001	1.00000	-0.06526 <.0001	-0.29814 <.0001	-0.06449 <.0001	-0.08191 <.0001
Southwest	-0.00677 0.1369	0.00247 0.5869	-0.00436 0.3380	0.00395 0.3857	-0.00675 0.1378	0.01079 0.0177	-0.06468 <.0001	-0.01473 0.0012	-0.06706 <.0001	-0.01410 0.0019	-0.06526 <.0001	1.00000	-0.06604 <.0001	-0.01428 0.0017	-0.01814 <.0001
South	-0.00361 0.4278	0.01771 <.0001	-0.00057 0.9004	0.00046 0.9192	-0.00460 0.3124	0.07909 <.0001	-0.29550 <.0001	-0.06729 <.0001	-0.30638 <.0001	-0.06441 <.0001	-0.29814 <.0001	-0.06604 <.0001	1.00000	-0.06525 <.0001	-0.08288 <.0001
Southeast	-0.00406 0.3728	-0.00312 0.4931	0.00457 0.3153	-0.00475 0.2965	-0.00438 0.3353	0.02355 <.0001	-0.06392 <.0001	-0.01455 0.0014	-0.06627 <.0001	-0.01393 0.0022	-0.06449 <.0001	-0.01428 0.0017	-0.06525 <.0001	1.00000	-0.01793 <.0001
Unknown	-0.00220 0.6289	-0.02628 <.0001	-0.00077 0.8664	0.00115 0.7999	-0.00061 0.8939	0.03809 <.0001	-0.08118 <.0001	-0.01849 <.0001	-0.08417 <.0001	-0.01770 0.0001	-0.08191 <.0001	-0.01814 <.0001	-0.08288 <.0001	-0.01793 <.0001	1.00000

In order to further explore the dataset and find out patterns, data visualization software, Tableau, was used to show the trend of accidents by time and location as shown in Figure. Each geographical

point represented a car crash by coordinate and the color of the point represented the hourly time the accident happens. A descending field sort was also performed in this case. As a result, we could see the time range from 12 – 16 had the top accident numbers over other hours. The pattern of accidents was not traceable since they are scattered around every direction and side of Manhattan. Even though the analysis failed to prove the relationship between geographical location and time, we noticed that the top number of accidents happens at around the time of 4 P.M. In other words, citizens may not pay enough attention due to fatigue after working throughout the day, which was one of the main contributing factors to accidents. A new direction of research was found based on the current findings.



Problem Statement 4:

To explore which person type is affected the most on the grounds of injury. What time of the day where the people should be extra careful about their surroundings? Which borough is dangerous for each person type?

Background:

A person's life is the most precious thing; hence the 'Persons' dataset was important to us. It could give us meaningful insights to provide suggestions and help reduce the injury to any person.

Data Processing:

To address the problem statements, we will be primarily using the 'Person' dataset because this dataset provides information about every person involved in the crash. Variables like *Person Type, Person Injury, Person Age, Person Sex* gives us valuable information. We will also merge the 'Persons' dataset with the 'Crashes' dataset to address some problems.

For processing the dataset, we removed certain variables like *Unique ID, Person ID, Ejection, Emotional Status, Position in Vehicle, Safety Equipment, Ped Location, Ped Action, Complaint, Contributing Factor 1, Contributing Factor 2.* The reason for removing so many variables from the dataset was primarily missing values. Others were removed for the purpose of simplifying the dataset or we can say reducing the number of variables without affecting the dataset.

Person Sex: This variable has a lot of noise and there were approximately 40 different sex types like Animal Action, Drug Experience, Pedestrian, Safe Speed, etc. The reason for this inconsistent data could be the police office or the system allowing them to input text value instead of providing them with a dropdown. We removed unnecessary values and the whole observation as well to get a clean dataset. We kept only M, F and U values which means Male, Female and Unspecified respectively.

After processing the dataset, we are left with 8 variables and around 668k observations.

Results and Insights:

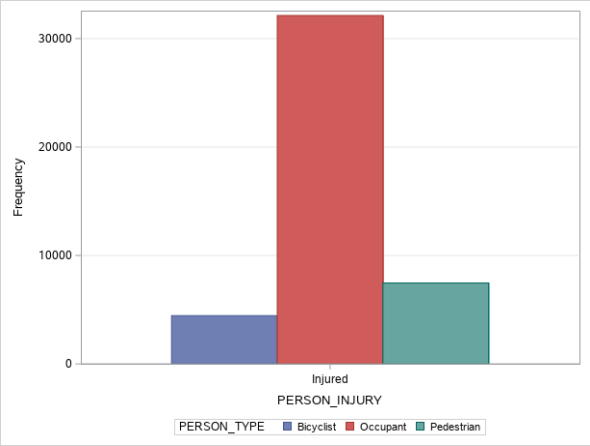
We wanted to first determine what percentage of people fall under each category. The pie chart below shows that 95.3% of the observations are related to occupants. The rest 4.7% is divided into

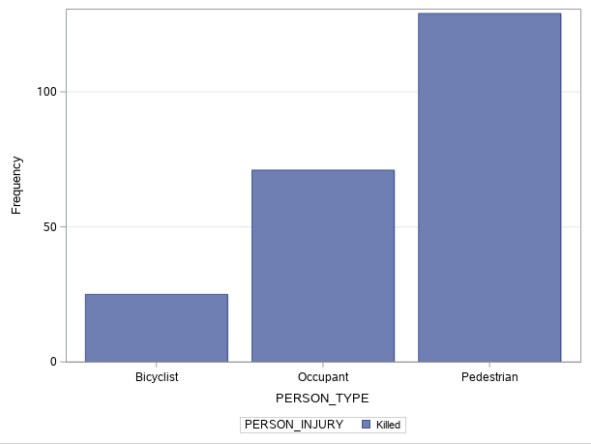
pedestrians and bicyclists. Our goal was to find out who among the three *Person Types* are affected the most. We separated the variable Person Injury into two values – Injured and Killed.

For the people who got injured, below is the frequency.

PERSON_TYPE				
PERSON_TYPE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Bicyclist	4401	10.09	4401	10.09
Occupant	31795	72.93	36196	83.02
Pedestrian	7402	16.98	43598	100.00

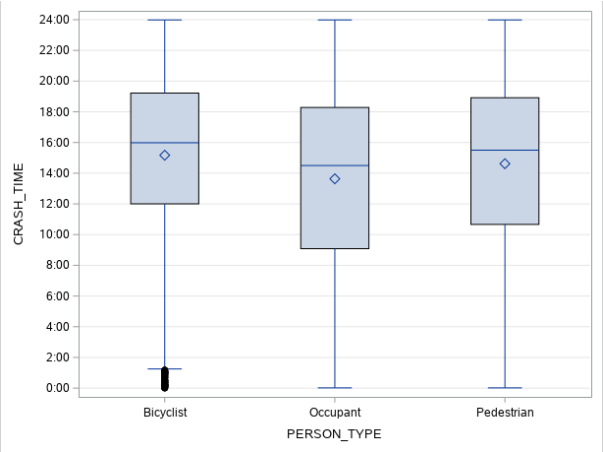
It is obvious that the number of occupants injured will be high but considering the percentage of bicyclists and pedestrians, their frequency of injury is also very high of all.





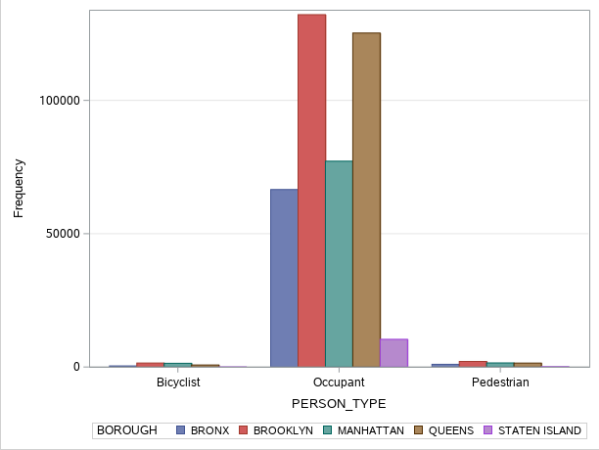
The graph on the left shows the number of each *Person Type* getting injured in the crash. The graph on the right shows the number of each *Person Type* killed during the crashes. Interestingly, we see that even though Occupants were the most injured, Pedestrians were the most who got killed. This is probably due to the fact that occupants are protected inside a car, while pedestrians have no such protection and tend to get heavily injured or die in a crash.

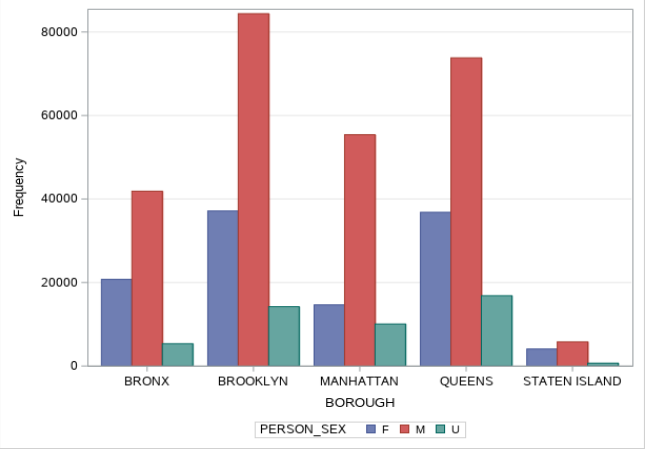
Next, to understand what time of the day is dangerous for each *Person type*, we plotted a box plot to analyze it.



For bicyclists, noon to around 7 in the evening comes out to be the most dangerous. We assumed that in NYC there are a lot of food delivery personnel who deliver food on bicycles that could be the cause of the spike at that time. People who cycle in the morning have fewer chances of getting into a crash as there is less traffic. For occupants, it definitely the rush hour starting from 9 AM to 6 PM in the evening. For Pedestrians, the risky time starts after 10 AM to 7 PM and this could be because they are crossing the roads to reach their destination using various modes of transportation, specifically public transportation like the subway. The area outside a subway station gets crowded when a train leaves and people then cross the road to reach their workplace may get into a crash. In the real-world scenario, in NYC, people crossroads from anywhere and not just the crosswalks, that could be another factor why pedestrians get into a crash.

To find out which *Borough* is the risk for each *Person type*, we had to merge two datasets - 'Persons' and 'Crashes' using the primary key *Collision ID*. Also, we did inner join while joining the tables as we will not have any missing values.





From the bar graph, we can clearly see that Brooklyn *Borough* is the most dangerous for all of the *Person Type*. This is interesting as we always thought that Manhattan *Borough* would be at the top in crashes, however, we realized that people come to Manhattan mostly for work purposes and live in the adjacent Boroughs. We can also say that Staten Island is the safest place to drive, walk or cycle for all the Person Type. The trend is similar to all the Person Type and Brooklyn and Queens.

In the second graph, in each of the *Borough,* the percentage of males to females in a crash is significantly higher. For Manhattan, the difference is almost three times, and we should keep a check on that with the help of authorities.

Final Thoughts

The project gave us an opportunity to play around with SAS and apply the learnings from class. In the initial stages of the work, one of our major hurdles was the size of the datasets as SAS University Edition was unable to handle the size of it. We also had to switch from JSON format to XLS format (for the majority of our analysis). In order to split the datasets, we imported .csv files into MS Access to process and split the dataset for moving ahead with the work and start the analysis. We used Access for its convenience and its capability to deal with huge datasets.

Analyzing the processed version datasets, gave us some useful insights like the seasonality effect on crashes, multi-vehicle collision factors, vehicle direction effect on crashes, person type involved in crashes and risky boroughs by type. These being rich datasets, we tried to address specific questions which we wanted to address. A lot more could be done in future which will surely help the authorities reducing the number of crashes and losses associated with it.

References

Climate in New York, NY, Bestplace.net, extracted on Mar 10th, 2020
https://www.bestplaces.net/climate/city/new_york/new_york

Motor Vehicle Collisions – Person, NYC Open Data
https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Person/f55k-p6yu

Motor Vehicle Collisions – Crashes, NYC Open Data
https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

Motor Vehicle Collisions – Vehicles, NYC Open Data
https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Vehicles/bm4k-52h4

Motor Vehicle Collisions – Vehicles, Data.Gov
https://catalog.data.gov/dataset/motor-vehicle-collisions-vehicles

Motor Vehicle Collisions – Person, Data.Gov
https://catalog.data.gov/dataset/motor-vehicle-collisions-person

Motor Vehicle Collisions – Crashes, Data.Gov
https://catalog.data.gov/dataset/nypd-motor-vehicle-collisions-07420

The New York City Pedestrian Safety Study & Action Plan
https://www1.nyc.gov/html/dot/html/pedestrians/pedsafetyreport.shtml

Appendix

Access code to split the dataset.

The code is similar across different datasets, so we just pasted 1 as an example.

```
SELECT Vehicles.UNIQUE_ID, Vehicles.COLLISION_ID, Vehicles.CRASH_DATE, Vehicles.CRASH_TIME,
Vehicles.VEHICLE_ID, Vehicles.STATE_REGISTRATION, Vehicles.VEHICLE_TYPE,
Vehicles.VEHICLE_MAKE, Vehicles.VEHICLE_MODEL, Vehicles.VEHICLE_YEAR, Vehicles.DRIVER_SEX,
Vehicles.DRIVER_LICENSE_STATUS, Vehicles.DRIVER_LICENSE_JURISDICTION, Vehicles.PRE_CRASH,
Vehicles.POINT_OF_IMPACT, Vehicles.VEHICLE_DAMAGE_1, Vehicles.VEHICLE_DAMAGE_2,
Vehicles.VEHICLE_DAMAGE_3, Vehicles.PUBLIC_PROPERTY_DAMAGE,
Vehicles.PUBLIC_PROPERTY_DAMAGE_TYPE, Vehicles.CONTRIBUTING_FACTOR_1,
Vehicles.CONTRIBUTING_FACTOR_2, Vehicles.TRAVEL_DIRECTION, Vehicles.VEHICLE_DAMAGE
FROM Vehicles
WHERE (((Vehicles.CRASH_DATE)<=#12/31/2019# And (Vehicles.CRASH_DATE)>=#1/1/2019#))
ORDER BY Vehicles.CRASH_DATE;
```

R code to clean missing value of the dataset for Problem Statement 3.

This code is performed in RStudio due to the limited function of SAS UE.

```
 setwd("~/Desktop/MSBA/MIS633/SAS group project")
datac <- read.csv("CRASHMERGE.csv")
summary(datac)
x <- na.omit(datac)
summary(x)
x<-x[x$BOROUGH == "MANHATTAN", ]
#Try to visualize Dataset, result is not optimistic
plot(x[,c(4,5)],col=datac[,8],xlim=c(40.5,41),ylim=c(-74.3,-73.5))
#Save the updated dataset
write.csv(x,"Crashmerge2.csv")
```

*Note: Code used in SAS is in PDF format file attached with this report submission.