Team HAKR, Group 5

(**H**ao Deng, **A**lex Fu, **K**im-Cuong Nguyen, **R**avi Bhagwat)

Project Problem & Data Identification

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 back in 2014 by AllState Insurance reveals that NYC drivers are 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, have intrigued our interest in figuring out the reasons.

We brainstormed the possibilities of an accident occurrence and also did some research about it. Interestingly, we have found that 'Distracted Driving' tops the list. In order to assess how serious the situation is, our group employs some types of data to have a correct view with a good amount of supporting evidence. In addition, our goal is to find out insights by performing analysis on different datasets so that we can suggest several preventive measures to drivers and local government. We aim to create awareness and eliminate the possibility of crashes by making an investigation into a number of areas; for example, the main causes of vehicle collisions, the most dangerous times on the road, the most dangerous regions in NYC, the changing amount of collisions, etc.

Goal

Our objective is to find out the accident patterns in NYC throughout datasets and then, come up with possible solutions to improve the current collision situation in the city.

In order to visualize the current picture of daily accidents in NYC, we are trying to address several questions, including:

- What are the top fatal areas in the city (location and time) where there is a high likelihood for vehicles/pedestrians to collide with injuries or deaths?
- What are seasonality patterns of collisions with injuries or deaths? Some smaller questions include: Are collisions more severe during the winter-time? Do people drive more carelessly during the weekend? Do morning rush hour accidents lead to more injuries or deaths than the afternoon ones?
- 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)?
- What are the top contributing factors that lead to each collision type - collisions with injured/killed pedestrians/cyclists/motorists/etc.?
- What are the seasonality patterns of contributing factors to collisions? Are alcohol and lack of sleep most common among collisions during midnight time? Do texting and cell phones cause more accidents during rush hours?

Based on the analysis result, our team will propose some recommendations that hopefully can reduce the severity of collisions to happen in NYC.

Team HAKR, Group 5

(**H**ao Deng, **A**lex Fu, **K**im-Cuong Nguyen, **R**avi Bhagwat) MIS 633-675

Introduction of Datasets – Motor Vehicle Collisions

We found some interesting datasets from the US Government open data website, data.gov. There are 3 different dataset which we are planning to merge and analyze hidden linkages and insights. The time range which the datasets addresses is from 2012 - 2019. However, NYPD started recording the collision details electronically only after April 2016, so, in order to have consistent data we will be selecting the data after April 2016.

- Motor Vehicle collisions (Vehicles): contains details of vehicles involved in collisions.
- Motor Vehicle collisions (Person): contains details for people involved in crashes.
- Motor Vehicle collisions (Crashes): contains details on each crash event.

We will load three datasets into SAS studio and analyze hidden linkages between datasets by using cross contrast method after cleaning & reconstructing datasets.

How do the data connect with the problems?

Collision ID is the key to connect the three datasets, by which way we can get a wider variety of details on each crash for further analysis. The data are categorized as below (the table of full list of variables is put in appendix attached to this statement):

- Crash ID
- Time (date & time)
- Location (in which borough, which zip code, at which intersection, etc.)
- People involved (sex, age, driver's license status, etc.)
- Vehicle (model, type, state registration, year of make, etc.)
- Pre-crash situation (traveling direction, vehicle status, etc.)
- Impact - Crash damage (people injured/killed, vehicle damage, etc.)
- Contributing factors (Distracted by phone, Sleepy, etc.)

By some certain analysis methodologies (that are not yet to be specifically mentioned within the scope of this data & problem statement), we will employ the data to address our questions as below:

- Connect time & location data with impact data to identify top fatal areas.
- Visualize the frequencies of collisions in accordance with time, weekdays and months to figure out seasonality patterns.
- Incorporate vehicle data into the above-identified seasonality to get the collision seasonality of each kind of vehicle.
- Classify contributing factors based on data on collision consequences to learn about top collision contributing factors per collision type.
- Plot contributing factors with time data to find out the seasonality of the common causes of collisions in NYC.

Preprocessing pipeline

- Cleanup the datasets.
- Identify critical variables and remove insignificant variables among three datasets.

- ➢ Adjust data types before cross-contrasting/merging datasets.
- ➢ Apply programming techniques and functions.
- ➢ Analyze and visualize the data to find patterns and provide insights.

Ultimate Goal:

As we explore and analyze the data, we may uncover some interesting traffic patterns that will have higher possibilities lead to an accident. Insights behind traffic patterns could be disclosed by using statistical methodologies combined with SAS Studio. Within SAS, we could use the function 'where' to filter, 'proc freq' to get the amount of each group. The output might consist of pie charts, histograms, lines, etc. Then, we can conclude what insights would bring behind our output.

By a simple guess, the prediction model may like:

Possibility to get into a crash = Y~ vehicle year + driver age + driver sex + crash time

A "Traffic accident prediction" model would be our ultimate goal after digging deep into datasets so that we can warn NYPD and drivers to be extra cautious in certain time and location.

Team HAKR, Group 5
(**H**ao Deng, **A**lex Fu, **K**im-Cuong Nguyen, **R**avi Bhagwat) MIS 633-675

APPENDIX

Category	Crash Dataset	Vehicle Dataset	Person Dataset
ID	Collision_id	Collision_id Unique_id	Collision_id Unique_id
Time	Crash Date Crash Time	Crash Date Crash Time	Crash Date Crash Time
Location	Borough Zip Code Latitude Location Longitude On Street Name Off Street Name Cross Street Name		
Person		Driver_sex Driver_license_status Driver_license_jurisdiction Vehicle_occupants	Person_id Person_sex Person_age Person_type Ped_role
Vehicle	Vehicle Type Code 1 Vehicle Type Code 2 Vehicle Type Code 3 Vehicle Type Code 4 Vehicle Type Code 5	Vehicle_id Vehicle_model Vehicle_type Vehicle_make Vehicle_year State_registration	Vehicle_id
Pre-cash		Pre_crash Travel_direction	Ped_location Ped_action Safety_equipment Position_in_vehicle
Impact	Number Of Cyclist Injured Number Of Cyclist Killed Number Of Motorist Injured Number Of Motorist Killed Number Of Pedestrians Injured Number Of Pedestrians Killed Number Of Persons Injured	Point_of_impact Vehicle_damage Vehicle_damage_1 Vehicle_damage_2 Vehicle_damage_3 Public_property_damage Public_property_damage_type	Emotional_status Bodily_injury Complaint Person_injury Ejection
Contributing Factors	Contributing Factor Vehicle 1 Contributing Factor Vehicle 2 Contributing Factor Vehicle 3 Contributing Factor Vehicle 4 Contributing Factor Vehicle 5	Contributing_factor_1 Contributing_factor_2	Contributing_factor_1 Contributing_factor_2

Note: Included are all the variables collected from the data source. More data preprocessing will be made so that the team can decide which variables can add significant value to the analysis

Team HAKR, Group 5

(**H**ao Deng, **A**lex Fu, **K**im-Cuong Nguyen, **R**avi Bhagwat) MIS 633-675

References:

https://www.dandalaw.com/are-car-accidents-common-in-new-york-city/
https://catalog.data.gov/dataset/nypd-motor-vehicle-collisions-07420
https://catalog.data.gov/dataset/motor-vehicle-collisions-person
https://catalog.data.gov/dataset/motor-vehicle-collisions-vehicles