Group Project Proposal

Group 3

Group Members:

Stephen Ling (jling9)	jling9@wisc.edu
Hongtao Zhang (hzhang784)	hzhang784@wisc.edu
Wenxi Yang(wyang235)	wyang235@wisc.edu
Chenying Mi(cmi)	cmi@wisc.edu

Description of Dataset:

Our dataset is about New York City (NYC) Traffic Accidents uploaded by "mysar" on *Kaggle* **(**URL**:** https://www.kaggle.com/mysarahmadbhat/nyc-traffic-accidents). The dataset contains 74881 observations of 29 variables about motor vehicle collisions reported by NYC Police Department (NYPD) from January to August in 2020. Each accident (observation) is described through variables around 5 categories: time of the accident, location of the accident, number of injuries, vehicles involved in the accident, and identification code of the accident.

Statistical Questions:

1. **What is the impact extent for each variable on the accidents? (How many variables affect the accident)?**
   Here are some aspects we would like to look into for this question:
   a. Patterns of motor traffic accidents are according to factors such as time (possibly rush hours), location (most accidental place), and vehicle type respectively
   b. Finding a variable in the dataset contributes to the traffic accidents most.
   c. Find the severity or frequency of traffic accidents in each district of NYC.
   d. Finding a contributing factor from vehicles contributes to the traffic accidents the most.
   e. Checking if there is a statistical difference between accidents that happen in summer and winter.
   f. … (Some Other Possible Plans/Guidance for this question)

2. **What is the correlation between variables of accidents? (Is one contributing factor independent from another?)**

In general, we would like to build a model, where we can predict the probability of the accident given the location, time, and vehicle type.

Description of Variables (29 variables):

1. ***Crash Date***: the date when a collision happened in form (yyyy-mm-dd).
2. ***Crash Time***: the time when a collision happened (precised to second).
3. ***Borough***: the borough where a collision happened.
4. ***Zip Code***: the zipcode of area where a collision happened.

5. ***Latitude***: the latitude of the place where a collision happened.
6. ***Longitude***: the longitude of the place where a collision happened.
7. ***Location***: points in form of (latitude, longitude) where a collision happened.
8. ***On Street Name***: the name of street where a collision happened.
9. ***Cross Street Name***: the street name that intersects with the main street that accident happens.
10. ***Off Street Name***: the closest street that near the area a collision happened.
11. ***Number of Persons Injured***: Number of person injured in the accident.
12. ***Number of Persons Killed***: the number of person killed in the accident.
13. ***Number of Pedestrians Injured***: the number of pedestrians injured in the accident.
14. ***Number of Pedestrians Killed***: the number of pedestrians killed in the accident.
15. ***Number of Cyclist Injured***: the number of cyclists injured in the accident.
16. ***Number of Cyclist Killed***: the number of cyclists killed in the accident.
17. ***Number of Motorist Injured***: the number of motorist injured in the accident.
18. ***Number of Motorist Killed***: the number of motorist killed in the accident.
19. ***Contributing Factor Vehicle 1~Vehicle 5***: [5 variables] how the first~fifth vehicle contributes to the accident, some could be NA since not all accidents involve 5 vehicles.
20. ***Collision ID***: the unique index to each collision.
21. ***Vehicle Type Code 1~ Code 5***: [5 variables] the type of the first~fifth vehicle in the accident, some could be NA since not all accidents involve 5 vehicles.

Statistical/Data Science Methods

- We intend to use Clustering and PCA to do the visualization of our data at the beginning of the analysis to find out if there are some patterns in our data (Statistical Question 1.a and 1.b). We also plan to map the frequency of traffic accidents in each district of NYC combined with the NYC geographic map (Statistical Question 1.c).
- After a series of visualizations, Chi-square test will be used to examine the correlation between classified variables and the accident rate (Statistical Question 1.d). Furthermore, we plan to use Monte Carlo Simulation to test if there is statistical difference between accidents that happen in summer and winter (Statistical Question 1.e).
- Additionally, We plan to create a model for accident severity to indicate the safety level by using PCA to extract the main contributing factor and then using Linear/Logistic Regression to get the influencing score of each attribute (Statistical Question 1).
- In the end, we plan to use the PCA & Chi-Square test to find correlation between each variable in the dataset, and give possible explanations based on the result we obtain (Statistical Question 2).

In short, the Statistical/Data Science we plan to use are:PCA, Clustering, Mapping, Chi-square test, Monte Carlo Simulation, Linear/Logistic Regression.