Statistical Analysis Plan

Group 12: Divya Sharma (ds655), Dhaval Potdar(dsp50), Jiayi Zhou (jz456), Jiechen Li (jl1254)

1. Data Overview

Dataset: The data-set has data on Traffic Crashes in Chicago from 2015 onwards, and contains details of each crash, including:

- `Location`: The latitude and longitude of the crash.
- `Date and time`: The date and time of the crash.
- `Injuries`: The number and type of injuries that occurred.
- `Damage`: The estimated damage cost of the crash.
- `Crash type`: The details about the type of crash - hit and run, no right of way, intersection related, causes etc.
- `Conditions`: The details about the weather, lighting, traffic, traffic control devices, and roads etc.

This data can be used to identify the probabilities of crashes happening, and their eventual consequences, given the surrounding situations such as time of day, weather, road conditions, traffic conditions, lighting, and roadway surface etc.

Source: The dataset was obtained from an online repository, though the specific origin or institution responsible for its collection and maintenance was not specified. The dataset is selected from City of Chicago's official page.

Sample Size: The dataset contains a total of 769,100 observations.

Number of Variables: There are 49 variables in the dataset.

Collection Method: The data is collected from traffic crash reports in Chicago. These reports can be sourced from various entities such as traffic police, surveillance systems, or reporting by involved parties. The exact method of data collection, however, has not been provided.

Unit of Analysis: Each row in the dataset represents a unique traffic crash incident in Chicago. The details of the crash, including specifics like location, time, type of crash, conditions at the time of the crash, and the aftermath (like injuries and damages), are captured across the 49 variables.

Research Questions:

1. Given that there is a crash, based on the location, time, crash type, and surrounding road and traffic conditions, how much damage is to be expected (<=1.5K$ or >1.5K$)?

- **Outcome Variable** - Damage (<=1.5K$ or >1.5K$) - categorical variable
- **Input/Explanatory Variables** - Location, date, time of the day, crash type, and surrounding conditions

2. How long it takes in minutes for police to be notified upon accident, given the nature of the accident?

- **Outcome Variable** - Police Notification Time (in minutes) - continuous variable
- **Input/Explanatory Variables** - Location, date, time of the day, crash type, and surrounding conditions.

2. Modeling

2.1 Model Types - For our first research question, which is a binary classification problem, we will utilize a Logistic Regression model. In addressing our second research question, where the outcome is a continuous variable, we will employ a Linear Regression model.

2.2 Inference or Prediction - Both research questions pertain to prediction. Consequently, we can easily transform any variable if necessary. Regarding variable selection, we have the flexibility to include any desired variables in the preliminary selection process. To evaluate models, we can employ cross-validation for model comparison. For logistic regression, the confusion matrix and ROC curve can serve as valuable assessment tools. The importance of interpreting coefficients is not as critical as it is for inference-focused problems.

2.3 Variables - For both research questions, we will incorporate the following variables: location (distance to downtown, calculated based on latitude and longitude variables in the dataset), posted speed limit, crash type, and various traffic conditions, including weather, lighting, traffic control devices, and road conditions.

2.4 Interaction Term - We are particularly interested in the relationship between distance to downtown and the posted speed limit. Therefore, we will introduce an interaction term involving distance to downtown and the posted speed limit.

3. Potential Challenges

3.1 Missing Original Source - We were able to locate the data dictionary for the same.

3.2 Missing Values - Our initial approach will involve removing the rows containing NAs. This is a viable solution, as we have sufficient data in both classes for the classification aspect of this assignment. Our secondary approach will be to utilize standard imputation techniques such as mean imputation.

3.3 Outliers and Unreasonable Values - Our strategy involves referencing the dataset's source to determine if there is a plausible explanation for the outliers. If none can be found, we will proceed to remove these erroneous values.

3.4 Categorical Variables have High Imbalance - In the first iteration of model building, we will incorporate these variables. Depending on their standard error, we may consider grouping them into fewer classes, among other options.

3.5 Majorly categorical variables - We will create a continuous "distance to downtown" variable, calculated based on the latitude and longitude variables in the dataset.