Miles Levine, Salih Awel

1. Brief Description of Data Collection (3 Points)

Provide a concise description of how you collected the data for your project. Your description should include answers to key questions related to data provenance:

- **Who collected it?**
  Clearly state who was responsible for collecting the data (e.g., you, a research group, a data repository, etc.).

The data was originally collected by the New York City Police Department (NYPD) and is provided publicly through the New York City Open Data portal. We accessed it directly from this open-source repository.

- **How was the data collected?**
  - If you **downloaded a dataset** from an open-source data repository (e.g., Kaggle, UCI Machine Learning Repository), specify the platform and provide details on how the repository originally collected the data.

  - If you collected the data yourself through any means (e.g., **API**, conducting a **survey**, or **web scraping**), describe the **collection process step-by-step**. This should include the tools or platforms used, any challenges faced, and how you addressed them.
  - If your data collection involved a **survey**, provide details on how participants were recruited and the process followed for data collection.

The dataset, titled "Motor Vehicle Collisions - Crashes," was downloaded from the NYC Open Data portal, as a CSV file. The NYPD collects this data by logging reported motor vehicle collisions in New York City, capturing details like location, contributing factors, and outcomes (injuries, fatalities, property damage) since April 28, 2014. This repository updates continually, making it a comprehensive and current resource.

- **Were there any modifications along the way? By whom?**
  Mention if any changes were made during the data collection phase (e.g., filtering or cleaning data, adjustments to the collection process). Specify who made these changes and why they were necessary.

For this project, our original dataset contains 29 columns and 2.13 million rows where every row is a motor vehicle collision. Our dataset was also created April 28, 2014 and is still being updated currently. After careful considerations, we intentionally reduced it to approximately 10,000 random entries. This was done by the project team to manage the dataset size while retaining significant detail for meaningful analysis. I downloaded the full csv file, then uploaded it to google colab. I ran a script that Randomly sampled 10,000 rows and saved the sampled data to a new CSV file. By randomly selecting 10,000 entries, we can be representative of the whole dataset and also still be able to make significant findings through this reduced amount. The initial cleaning involved handling missing fields, normalizing text (e.g., street names and contributing factors), and formatting dates and times for consistency.

2. Data Upload to Cloud (If Applicable) (1 Point)

- If you collected the data yourself, upload the dataset to a cloud storage service (e.g., Google Drive, Box, Dropbox). Ensure that the file is shared with a **view-only link**. Include this link in your document.
- If you used an open-source dataset, you **do not** need to upload the file; just reference the original source.

https://drive.google.com/file/d/1tas5Rq7i8O_z5iWF06VIjR_98ym92ugU/view?usp=sharing

3. Basic Description of the Data (1 Point)

Number of Samples (Rows) and Variables

The analyzed dataset contains 10,000 rows of traffic accident records with 29 different variables of data for each accident. Each variable describes various attributes of each collision: location, time of accident, casualties, and vehicle data.

Geographical Distribution

Accidents are recorded across the five boroughs of New York City as follows:

- - Brooklyn: 2,144 accidents (21.4%)
- - Queens: 1,828 accidents (18.3%)

- - Manhattan: 1,217 accidents (12.2%)
- - Bronx: 1,103 accidents (11.0%)
- - Staten Island: 240 accidents (2.4%)

-This distribution indicates that nearly 40% of all reported accidents in our sample are in Brooklyn and Queens.

Casualty Statistics

Casualty patterns as extracted from the data are listed below:

- - Total number of injuries: 3,318
- - Total number of fatalities: 19
- - Average injuries per accident: 0.33
- - Maximum injuries in a single accident: 11

These statistics show that most accidents do not cause injury, but the incidents can sometimes be serious.

Temporal Patterns

Accidents by day of the week:

- Friday: 1,574 accidents (15.7%)
- Wednesday: 1,479 accidents (14.8%)
- Tuesday: 1,478 accidents (14.8%)
- Thursday: 1,473 accidents (14.7%)
- Monday: 1,426 accidents (14.3%)
- Saturday: 1,376 accidents (13.8%)
- Sunday: 1,194 accidents (11.9%)

As shown from the above distribution, accidents are slightly higher during the weekdays, especially on Fridays, and lower during weekends.

Contributing Factors

The top five contributing causes of accidents:

- Driver Inattention/Distraction: 2,486 cases (24.9%)
- Unspecified: 2,359 cases (23.6%)
- Following Too Closely: 846 cases (8.5%)
- Failure to Yield Right-of-Way: 688 cases (6.9%)
- Passing or Lane Usage Improper: 435 cases (4.3%)

-Inattention/distraction of drivers is the leading known cause of accidents.

Vehicle Types Involved

The most common vehicle types in accidents:

- Sedan: 4,714 vehicles (47.1%)
- Station Wagon/Sport Utility Vehicle: 3,580 vehicles (35.8%)
- Taxi: 392 vehicles (3.9%)
- Pick-up Truck: 287 vehicles (2.9%)
- Box Truck: 168 vehicles (1.7%)

Sedans and SUVs dominate in almost 83% of the cases.