Group Members: Miles Levine, Salih Awel

This assignment is the first deliverable for your final project, where you will submit a detailed project proposal. The goal of this proposal is to describe the data analysis pipeline you will implement over the course of the project. You will outline the problem you intend to solve, define your research question, and describe how you plan to collect, clean, analyze, and present your data. This is an initial plan, so adjustments can be made as the project progresses.

Problem Description (100-200 words)

The problem this project aims to address is the high frequency of traffic accidents in New York City, which often result in injury, fatality, or significant damage. Traffic accidents not only pose serious safety risks but also cause economic losses and disruptions to commuting and city life. In a densely populated urban area like New York City, identifying accident-prone locations is critical for improving traffic safety and reducing the impacts of accidents. The purpose of this project is to use the NYPD's motor vehicle collision data to identify traffic accident-prone locations in New York City. By pinpointing locations with the highest frequency of accidents, this project will help city planners, law enforcement, and policymakers implement targeted safety interventions. Increased traffic monitoring, enhanced road infrastructure, and public awareness campaigns promoting road safety are all possible remedies.

Research Question(s)

- Where are the most frequent traffic accident-prone locations in New York City, and what factors contribute to the occurrence of accidents in these locations?
- Which contributing factors are most commonly reported in accidents occurring at identified accident-prone locations?

Data Source(s)

For our project, we will be using publicly available third-party datasets provided by the New York City Police Department (NYPD) through the New York City Open Data portal. The dataset titled Motor Vehicle Collisions - Crashes can be accessed via the following link: https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes.

The dataset can be accessed and downloaded directly from the portal as a CSV file or through an API. For this project, we will download the dataset and import it into a local environment where we will perform the data cleaning, analysis, and visualization steps. By using this dataset, we will be able to analyze traffic accident patterns across New York City and identify accident-prone locations.

Data Collection Plan

Since the Motor Vehicle Collisions - Crashes dataset provided by the NYPD is already publicly available and comprehensive, the target data size will be determined later in our data collection deliverable. The dataset includes records of traffic accidents in New York City since July 1, 2012, and is continually updated. The dataset currently contains over 2 million records and continues to grow as new incidents are reported. Each row represents an individual motor vehicle collision that meets the reporting criteria. Since this dataset is so large by nature, we plan on reducing the number of rows to around ten thousand of the most recent data entries. Although we are reducing this dataset to a more manageable level, we can still make sure to gather sufficient detail for meaningful analysis. The dataset covers over 12 years, from July 1, 2012, to the present, providing an extensive range of data for analysis. The dataset includes 29 fields, which provide detailed information on the crash, such as location, time, contributing factors, and outcomes (injuries, fatalities, property damage).

Initial Data Cleaning and Wrangling Plan

Initial cleaning will include analyzing and adjusting missing or incomplete fields, particularly in columns like location, contributing factors, or weather conditions. Text fields such as street names and contributing factors might also need to be cleaned due to the possibility of inconsistent spelling or abbreviations. We will standardize these entries to avoid duplication or misclassification. Dates and times may need to be formatted in a standard consistently for analysis. We will ensure that the date and time are in appropriate formats for sorting and filtering. If necessary, we will also normalize the columns to ensure consistency in any statistical analysis.

Data Analysis Plan

First, our analysis will start with extensive data cleaning to ensure quality. We'll remove duplicate accident reports, deal with missing information by either removing incomplete entries or filling in gaps where possible, and standardize the format of dates and times. Later on, we'll analyze the cleaned data extensively: examining the frequency of accidents in different parts of NYC and temporal patterns such as time of the day, day of the week, and month of the year. We are also going to investigate common factors associated with accidents, such as weather and road conditions.

We will visualize accident hotspots using heat maps of NYC and apply techniques such as clustering. We'd like to find out if there are patterns in the occurrence of accidents, such as certain types of accidents occurring in specific areas, or if accident timing follows a pattern. We will run statistical tests to determine which factors contribute significantly to the occurrence of accidents. Using this data, we'll try to predict possible locations and timings of future accidents.

Medium for Displaying Results

To accomplish this, we'll be using many different visualization techniques. We'll create a number of charts and graphs in matplotlib. For example, we can create bar charts - one for accident counts based on borough or based on the hour of day. Line graphs could show how accidents have changed over time.

We will also create some histograms to visualize the distribution of accidents concerning different categories. To understand the relationship between different factors, we may plot scatter plots. All these visualizations will be done using Python; in particular, a Python library called matplotlib.

We will also produce a written report summarizing our findings in addition to these plots. This report will incorporate the most important visualizations and also a proper description of what is shown. We will structure this report using Jupyter Notebook, which allows us to merge code, outputs, and written analysis into one document.

This approach will enable us to provide a very clear and attractive presentation of our data analysis results, making it easier for readers to understand the patterns and insights we've drawn from the NYC traffic accident data.