College of Professional Studies

Northeastern University San Jose

MPS Analytics

Course: ALY6110 – Data Management and Big Data

Assignment:

Module 2 Dataset Selection

Submitted on:
June 7th, 2023

Submitted to: **Submitted by:**
Professor: BEHZAD AHMADI   ARCHIT BARUA
                                    BHAGYASHRI KADAM
                                    NIKSHITA RANGANATHAN

PROBLEM DEFINITION

The problem of traffic accidents in the United States is a significant concern that has been escalating over time. Researchers, law enforcement agencies, government bodies, and related organizations have been actively working to address this issue. The number of traffic accident fatalities each year in the US is alarmingly high, reaching nearly 1.25 million deaths.

This problem affects all states across the country, and the frequency of accidents has been on the rise. The increasing number of accidents poses a threat to public safety, necessitating urgent action to find effective solutions.

To gain valuable insights and address this problem, a dataset related to US traffic accidents would be analyzed. The analysis aims to identify meaningful patterns and provide actionable analysis that can aid in curbing the issue and reducing the number of accidents and fatalities on the nation's roadways.

INTRODUCTION

About the Dataset:

The US Accidents dataset contains information about countrywide traffic accidents that occurred in the United States from February 2016 to March 2023. Each record includes details about the accident, such as the date, time, location, weather conditions etc.

The dataset also includes information about the road conditions, such as the type of road, the number of lanes, and the presence of traffic lights or stop signs. This dataset is a valuable resource for businesses and organizations that are interested in improving road safety.

The dataset has around **7.7M records and 47 columns**.

Data source: The dataset has been sourced from Kaggle:

https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

The data is continuously being collected from February 2016, using several data providers, including multiple APIs that provide streaming traffic event data.

Objective/Business Question :

1. What are the key factors contributing to traffic accidents in a specific city/state ?
2. What the most effective ways to mitigate these accidents?
3. What are the underlying relationships between the attributes, especially Severity ?
4. Which are the top states that have the highest number of accidents?
5. How does the weather has an impact on the accidents?

Data Mining Application

6. How time of the year or week contributes more towards the accident?

Inspiration/Applications of Dataset :

The US-Accidents dataset can be applied to a wide range of applications, including the prediction of car accidents in real time, the analysis of accident hotspots, casualty analysis, the extraction of cause-and-effect rules to forecast car accidents, and the study of how precipitation or other environmental factors affect the likelihood of accidents occurring.

Group Members :

- Archit Barua
- Bhagyashri Kadam
- Nikshita Ranganathan

Data Mining Application