**USA Car Accidents Data Processing Documentation**

**Dataset Overview**

The dataset, titled **"US\_Accidents\_March23.csv"**, contains information about car accidents across the United States up to **March 2023**. The dataset provides details such as the location of accidents (city, state), accident severity, start and end times, and other related features.

**Objective**

The goal of this project is to:

1. **Clean and preprocess** the dataset to ensure it's ready for analysis.
2. **Sample** a subset of the data (10,000 rows) for quick exploration and analysis.
3. Ensure **datetime consistency** for time-related columns (Start\_Time and End\_Time).

**Steps Undertaken**

**1. Data Loading**

The dataset was loaded into a pandas DataFrame using the following code:

import pandas as pd file\_path = r"C:\Users\vange\Desktop\github project\archive\US\_Accidents\_March23.csv" df = pd.read\_csv(file\_path)

**2. Cleaning Missing Values**

To ensure the data used for analysis is clean:

* Rows with missing (NA) values were dropped using:

df\_no\_na = df.dropna()

* This step ensures only complete rows are considered for sampling and further analysis.

**3. Sampling 10,000 Rows**

* From the cleaned dataset (df\_no\_na), a random sample of **10,000 rows** was selected to enable faster analysis:

cleaned\_accidents = df\_no\_na.sample(n=10000, random\_state=42)

* **random\_state=42** was used to ensure reproducibility of the sampling process.

**4. Datetime Conversion**

* The dataset includes Start\_Time and End\_Time columns representing the start and end times of accidents. These columns were converted into datetime format for consistency and ease of use in time-based analyses:

# Convert Start\_Time and End\_Time to datetime format

cleaned\_accidents['Start\_Time'] = pd.to\_datetime(cleaned\_accidents['Start\_Time']) cleaned\_accidents['End\_Time'] = pd.to\_datetime(cleaned\_accidents['End\_Time'])

This step ensures proper handling of date and time values for tasks like filtering, aggregating, or analyzing trends over time.

**5. Exporting the Cleaned Sample**

The cleaned and sampled dataset was exported as a CSV file for future analysis:

output\_path = r"C:\Users\vange\Desktop\github project\cleaned\_accidents.csv" cleaned\_accidents.to\_csv(output\_path, index=False)

The cleaned dataset is saved at: **C:\Users\vange\Desktop\github project\cleaned\_accidents.csv**

**6. Team Members**

**This project was a collaborative effort by the following team members:**

Georgios Birmpakos

Vasileios Katsikas

Evangelos Diaskoufis

**7. Questionnaire Highlights**

**The questionnaire aims to explore and analyze various aspects of car accidents across the dataset. Below are a few key details:**

**A. Accidents Per State**

1. **Accident Count by State:**
   * **The number of accidents per state is calculated to identify states with the highest and lowest accident frequencies.**
   * **The top and bottom states will be highlighted to understand accident trends.**

**B. Most Prone State for Accidents**

1. **Top State for Accidents:**
   * **Analysis focuses on identifying the state with the highest number of accidents and examining whether there’s a significant gap between it and others.**

**C. Accidents Per Month**

1. **Seasonal Trends:**
   * **Accident counts are analyzed by month to uncover peak accident periods and explore how seasonality impacts accident trends.**

**D. Accidents at Junctions, Stops, and Traffic Signals**

1. **Accidents Near Specific Locations:**
   * **A breakdown of accidents near:**
     + **Junctions**
     + **Stops**
     + **Traffic signals**
   * **The percentage of total accidents occurring at these locations will also be calculated.**
2. **Next Steps**
3. **Perform Exploratory Data Analysis (EDA) to uncover trends and patterns.**
4. **Analyze Accident Severity:**
   * **Examine accident severity distribution across cities and states.**
   * **Investigate the most dangerous areas based on severity.**
5. **Temporal and Spatial Trends:**
   * **Identify geographical hotspots for accidents.**
   * **Study time-based trends, such as peak accident hours or seasonal variations.**

**Summary**

This documentation provides a detailed outline of the steps taken to clean and preprocess the **USA car accidents dataset up to March 2023.** It also highlights key questions addressed in the analysis. The cleaned and sampled data is ready for further analysis and visualization.