Project Title:

Reveal Crash Severity: A comprehensive Analysis of Location, Weather, and Time Factors.

Author: Md Junayed Miah (23120467)

Question: What factors contribute to the severity of motor vehicle crashes?

Objective:

Use features like crash location, time weather conditions, and road type to predict the seriousness of crashes. (e.g. minor, serious, fatal). And explain how analyzing crash data helps reduce accidents, inform policymakers, and improve public safety.

Data Sources:

- **Data Source-1:** Montgomery County Crash Data Source:

  The dataet provides me the detailes information about motor vehicle crashes in Montgomery County, including crash date, route type, road conditions, weather, and driver at fault. It offers a lot of details about the factors those are related with crash severity.

  - **Url:** Crash Reporting - Drivers Data
  - **Content:**

    Covers details about crashes, including date, weather, road conditions and locations. There are some important fields like Crash Date/Time, Weather, Road Name, Speed Limit etc.It also allows a deep dive into crash patterns in a localized region, facilitating targeted analysis. It also contains multiple dimensions of a crash, enabling thorough EDA.

  - **License:** Open Data Common Public Domain Dedication and License (ODC PDDL).
    - **Obligations:** Proper attribution and ensuring transparency in usage.
    - **Link:** License details on Montgomery County Open Data Portal.

- **Data Source-2:** New York City Vehicle Collision Data Source:

  This dataset offers comprehensive data on motor vehicle collisions in New York City, including details about crash dates, times boroughs, injured or killed persons, and vehicle types invloved.

- **Url:** NYC Vehicle Collision Data
- **Content:**

  Provides detailed records of motor vehicle collisions, including location, time, injuries, and fatalities. There are some major fields like Crash Date, Borough, Number of Persons injured, Latitude etc.

- **License:** Available through NYC OpenData, aligning with open-data principles, ensuring it can be used without legal constraints.
  - **Obligations:** Use data responsibly and attribute the source.
  - **Link:** NYC Open Data terms and conditions.

Data Pipeline:

High-Level Overview:

- **Purpose:** Automate the data extraction, cleaning, transformation, and storage process.
- **Technologies:** Python (Pandas, SQLite), Bash scripts.

Transformation and Cleaning steps:

- **Data Extraction:**
  - Direct download from the provided URLs using python. Then, ensures access to the latest data without manual intervention.
- **Data Cleaning:** To ensures consistency and prevents issues during analysis.
  - Drop Missing Values: Criticle fields like location, date/time.
  - Standardize Columns Names: Convert to lowercase, replace spaces with underscores.
- **Transformation:** To focuses on essential data for predictive analysis.
  - Field Selection: Choose relevant columns (e.g. crash time, weather, severity indicators).
  - Data Type Conversion: Ensure date/time fields are in a consistent format (datetime64).
- **Data Storage:** Lightweight, easily queryable, and itegrates well with Python for analysis.
  - Format: SQLite databases(.db files).

Challenges and Solutions:

- **Inconsistent Column Names:**
  - **solution:** Unified them using Pandas' .str.lower() and .str.replace() methods.
- **Missing Data:**
  - **Solution:** Removed rows with missing critical values and imputed non0critical missing data.
- **Large Dataset Size:**

- **Solution:** Processed in chunks and stored as databases to optimize performance.

Results and Limitations:

- **Output:**
  My data pipeline processes two datasets related to motor vehicle accidents: one from Montgomery County, MD, and the other form New York City. After Cleaning and transforming the data, the output consists of two main datasets stored.

  - **Format:** Two SQLite databases (crash_reporting.db, vehicle_collision.db).

  - **Structure-1:** Montgomery County Data:
    - **Columns:** crash_date/time, route_type, road_name, collision_type, weather, surface_condition, light, latitude, longitude, etc.
    - **Data Types:** Dates (datetime64), strings (e.g., collision_type), and numerical values (e.g., speed_limit, latitude).

  - **Structure-2:** NYC Vehicle Collision Data:
    - **Columns**: crash_date, crash_time, borough, number_of_persons_injured, number_of_persons_killed, latitude, longitude, etc.
    - **Data Types:** Dates (datetime64), strings (borough), and integers (number_of_persons_injured).

- **Data Quality Dimensions:**

  The output data is stored in SQLite databases and CSV files with standardized, cleand, and structured fields, ensuring accuracy through critical field validation, completeness by removing or filling missing values, and consistency via uniform column names and date formats.

- **Limitations:**

  - **Data Bias:** Underreporting of minor incidents. Like not all motor vehicle crashes are reported, especially minor incidents where no police report is filed. This can lead to an incomplete dataset that may skew the severity distribution toward more serious accidents.

  - **Regional Differences:** Variations in reporting standards between Montgomery County and NYC.

  - **Timeliness and Data Recency:**
    Data might not be updated in real-time and could lag behind current trends, especially with manually reported incidents. Outdated information can reduce the accuracy of predictive models and limit the relevance of findings for current traffic management or policy decisions.