Road Accidents Analysis

# Project Overview

This project aims to analyze trends in car accidents in the United States and evaluate their impact on traffic. By examining various factors such as accident frequency, location, time of day, and severity, we seek to identify trends and implement solutions aimed at reducing accidents and minimizing their impact on traffic flow. This research contributes to enhancing road safety and improving the overall efficiency and reliability of transportation networks.

# Table of Contents

1. Introduction

2. Data Collection and Data Sources

3. Data Cleaning and Exploratory Data Analysis

4. Visualization

5. Database Type Selection

6. MySQL Queries and ERD

7. API

8. Web Scraping

9. Machine Learning

10. GDPR Compliance

# Introduction

Studying the impact of car accidents on traffic in the United States is crucial for understanding the dynamics of roadway safety, traffic flow, and urban planning. Analyzing these incidents provides valuable insights into the causes and consequences of traffic disruptions, allowing for the development of more effective traffic management strategies and safety measures.

# Data Collection and Data Sources

- Flat Files: The primary dataset was sourced from Kaggle under the name "US Accidents (2016 – 2023)". For MySQL operations, a lighter version focused on March 2023 was used.

2016-2023: <https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents>

March 23: https://drive.google.com/file/d/1U3u8QYzLjnEaSurtZfSAS\_oh9AT2Mn8X/edit

- API: An API was created to facilitate data access.

- Web Scraping: Additional data on population and state names were scraped from Wikipedia to scale high occurrences of accidents.

# Data Cleaning and Exploratory Data Analysis

Data cleaning and exploratory data analysis were conducted on the flat file from Kaggle. Key steps included checking data types, handling null values, and removing duplicates. Visualization techniques were then applied to present the main findings.

# Visualization

Visualization of the data included:

- Top ten accidents by state (graph and map)

- Various other charts to show accident trends and distributions

# Database Type Selection

MySQL was chosen due to its superior performance, extensive community support, flexibility, and proven reliability. It is especially efficient in read-heavy operations, making it suitable for our data needs.

# MySQL Queries and ERD

The project involved creating a database in MySQL, populating it with the cleaned data, and generating an Entity-Relationship Diagram (ERD). Various SQL queries were run to extract and manipulate data as needed.

# API

An API was developed to expose the data. For example, one endpoint returns the top 20 cities with the highest number of accidents.

# Web Scraping

Web scraping was used to obtain full state names and additional demographic data from Wikipedia. This information was crucial for performing more detailed analysis, such as calculating accidents per capita.

# Machine Learning

Machine learning was applied to predict the impact of road accidents on traffic. The project demonstrates the necessity of machine learning in automating complex tasks, improving accuracy, and uncovering insights from large datasets.

# GDPR Compliance

The project strictly adhered to GDPR guidelines, ensuring that no personal information was collected or used. All data was anonymized, and technical safeguards were employed to protect privacy.

# How to Run the Project

1. Clone the repository:

git clone <repository-url>

2. Navigate to the project directory:

cd road-accidents-analysis

3. Set up a virtual environment:

python -m venv env

4. Activate the virtual environment:

- On Windows:

.\env\Scripts\activate

- On macOS and Linux:

source env/bin/activate

5. Install the required packages:

pip install -r requirements.txt

6. Run the Jupyter notebooks:

jupyter notebook

7. Execute the notebooks: Open Final\_project.ipynb and wiki\_web\_scraping.ipynb and run all cells to reproduce the analysis.