Infosys Springboard Internship Program

EDA - US RoadSafe Analytics

- Artificial Intelligence Intern Team

AGENDA

- **Project Statement:** The aim and overview of the accident severity analysis.
- **Project Objectives:** Key aims and focus areas of the analysis.
- **Documentation:** Analytical approach and dataset details.
- **ML Algorithms:** Algorithms application and data-driven predictions.
- **Dashboard View and App Structure:** Key features and interactive visualization.
- **Key Findings and Insights:** Major insights and their implications.
- **Tech Stack Avery:** Tools and technologies used together.
- **Future Scope:** Suggestions for safety and research improvement.
- **Conclusion:** Summary and final thoughts.

PROJECT STATEMENT

The primary goal of this project is to analyze a large dataset of road accidents to uncover trends, patterns, and key factors contributing to accident severity. The project involves performing in-depth exploratory data analysis (EDA) using Python libraries such as Pandas for data manipulation, Matplotlib and Seaborn for statistical visualization, and Streamlit for developing an interactive dashboard to extract meaningful insights that can help improve road safety across the United States.

Business Context:

- Stakeholder presentations
- Academic publication
- Project portfolio
- Grant proposals

PROJECT OBJECTIVES

- Explore and understand real accident data from across the US.
- Clean and prepare the data for accurate analysis.
- Study how time, weather, and location affect accident severity.
- Identify high-risk areas and major accident causes.
- Build an interactive dashboard for easy data visualization.
- Provide insights that can help improve road safety.

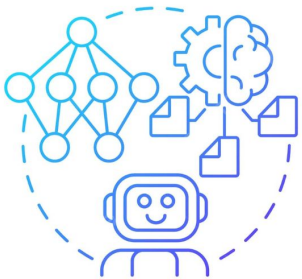
Documentation and ML Algorithms

Dataset

Documentation

Dataset Overview

- Dataset Name : US Accidents (2016 - 2023)
- Geographic Coverage : 49 states of the USA
- Time Period : February 2016 to March 2023
- Total Records : Approximately 7.7 million accident records
- Source : Kaggle



Machine Learning

ML Algorithms

- **DBSCAN** stands for **Density-Based Spatial Clustering of Applications with Noise.**
- It uses **distance (ε)** and **minimum points (minPts)** to find meaningful clusters.
- Geospatial Analysis: Detects clusters of any shape.
- In the US Accidents dataset, it spots accident hotspots and filters out rare cases.

DASHBOARD VIEW AND APP STRUCTURE

DASHBOARD 1

Deploy

Navigation

Go to Section

- Home Dashboard
- Preprocessing
- Univariate Analysis
- Comparative Analysis
- Geospatial Analysis
- Insights & Hypothesis
- Key Findings

US RoadSafe Analytics

Analyze and visualize U.S. road accident trends to improve road safety awareness.

Loaded full dataset. This may take longer.

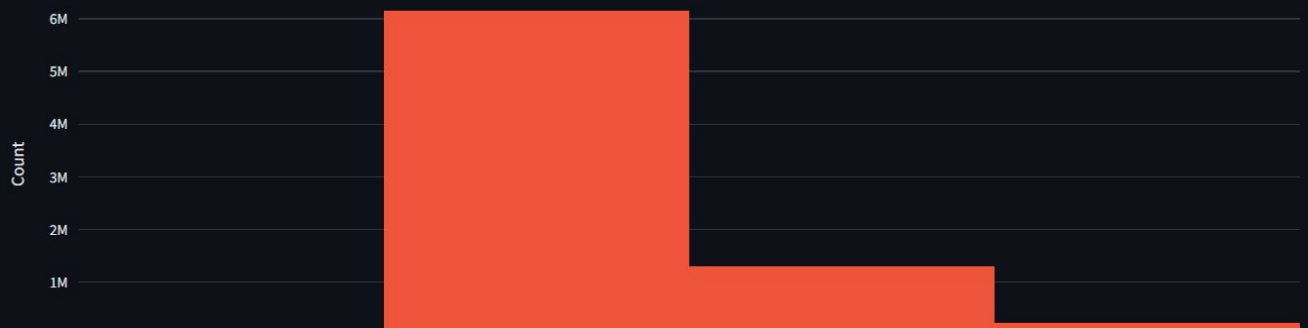
Total Accidents

7728394

Average Severity

2.21

Severity Distribution



DASHBOARD 2

Geospatial Accident Analysis with Hotspot Counts

Deploy

★ Geospatial Analysis shows distribution across locations such as (states and cities) and helps to understand regional variations in accident patterns.

Select geography level

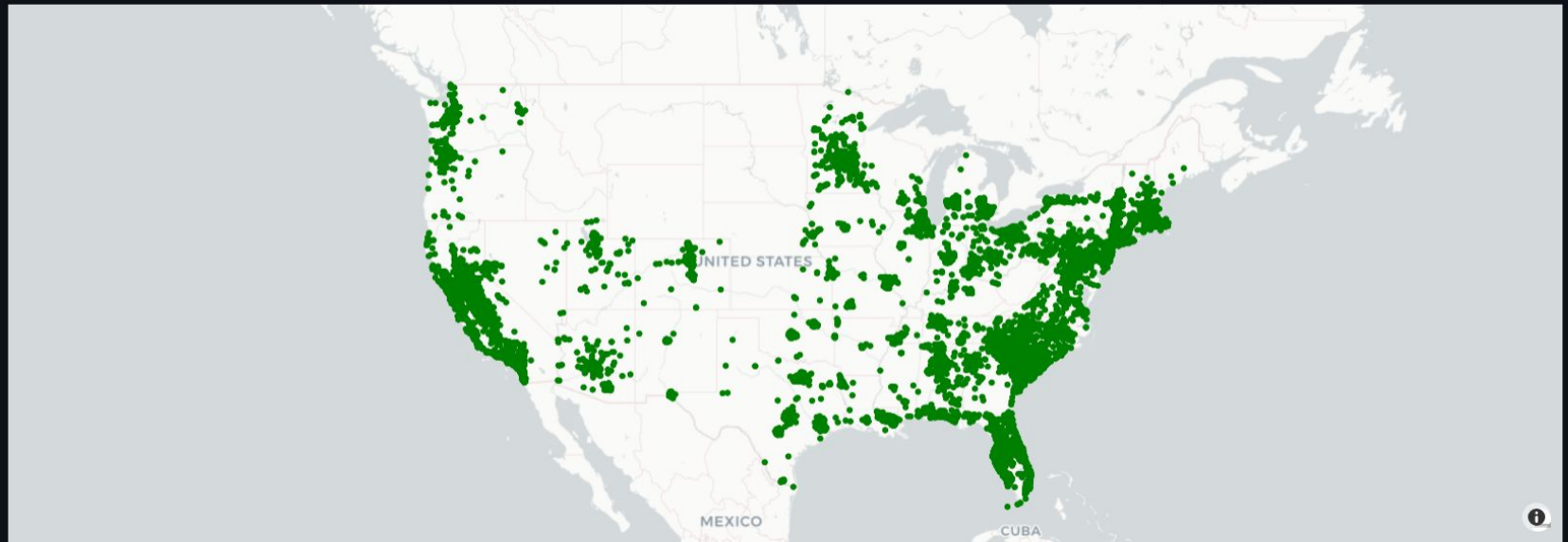
- (•) Country
- ( ) State
- ( ) City

Select visualization type

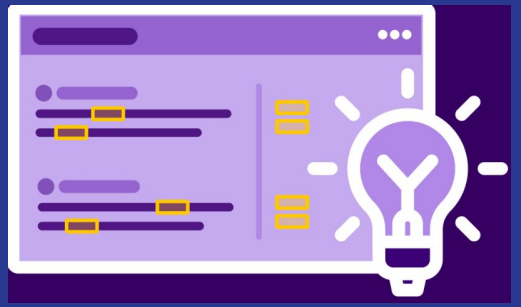
- (•) Point Map
- ( ) Hotspot Density

Select Severity Level

1



Key Findings and Insights



DASHBOARD 3

Deploy

Key Findings & Summary Dashboard

Summary Metrics

Total Accidents Analyzed
6,728,423

Peak Accident Hour
7

High Severity Accidents (Severity≥3)
1,428,830

★ Findings :-

1. Large Dataset Coverage - Over 6.7 million U.S. accidents analyzed for nationwide insights.

2. Peak Accident Time - Most accidents happen around 7 AM during rush hours.

3. High-Risk States and Cities - California, Florida, and Texas with Houston, Miami and Los Angeles report the most accidents.

4. Road & Weather Impact - Clear and rainy weather, along with junctions and crossings, see more accidents.

Top 5 Accident-Prone States

Top 5 Accident-Prone Cities

Example Insight : Accident Frequency by Hour of Day

Tech Stack Avery

Data Layer

Dataset Source

- US Accidents Dataset (Kaggle / US DOT)
- CSV format with ~7 million records

Processing Layer

Core Analysis

- Python
- Jupyter Notebook

Visualization Libraries

- Matplotlib
- Seaborn
- Plotly

Visualization Layer

Framework

- Streamlit

Features

- Real-time filtering of accident data
- Visualization of hotspots, weather impact, and severity
- Intuitive and user-friendly interface

Dataset

Processing

Documentation & Reporting Layer

FUTURE SCOPE

1. Real-Time Accident Monitoring - Integrating live traffic, weather, and visibility data can help in identifying accident-prone areas instantly.

2. Region-Specific Safety Insights - Expanding the model to analyze regional driving patterns can support targeted road-safety planning.

3. Predictive Severity Modeling - Enhancing the system to forecast accident severity based on evolving environmental and road conditions.

4. Public Awareness and Policy Support - Using insights to design awareness tools and assist policymakers helps in improving safety regulations.

CONCLUSION

This project analyzed large-scale U.S. accident data to uncover key factors influencing accident severity, such as weather, visibility, and time of occurrence. The developed interactive dashboard makes complex data easy to interpret, helping identify high-risk patterns and areas. Overall, the study highlights how data-driven insights can support **smarter decisions** and **contribute to improving road safety.**

THANK YOU