RoadSafe Analytics

Road Accidents – Exploratory Data Analysis and Insights

1 Project Overview

RoadSafe Analytics is a data science project aimed at analyzing large-scale road accident data to identify patterns, trends, and risk factors associated with accident occurrence and severity. The project applies exploratory data analysis (EDA), statistical hypothesis testing, and visual analytics to derive meaningful insights that can support road safety planning and decision-making.

2 Dataset Description

The dataset consists of road accident records containing information related to accident severity, time and date of occurrence, geographic location (latitude, longitude, state, city), weather conditions, visibility, and various road and traffic features such as junctions and traffic signals.

3 Tools and Technologies

- Programming Language: Python
- Libraries: pandas, numpy, matplotlib, seaborn
- Visualization: Matplotlib, Seaborn
- Dashboard Development: Streamlit
- Environment: Jupyter Notebook
- Version Control: Git and GitHub

4 Week-wise Work Documentation

Week 1: Dataset Understanding and Initial Exploration

Objectives

- Understand dataset structure and schema.
- Inspect data quality.

Work Done

- Loaded the dataset and examined its shape and structure.
- Inspected column names, data types, and sample records.
- Identified missing values across columns.
- Generated basic descriptive statistics.

Outcome

A clear understanding of dataset composition and identification of columns requiring preprocessing.

Week 2: Data Cleaning and Preprocessing

Objectives

- Prepare the dataset for analysis.

Work Done

- Converted timestamp columns to proper datetime format.
- Created new features such as Hour, Weekday, and Month.
- Dropped columns with excessive missing values.
- Imputed missing values using median and mode.
- Removed duplicate records.

Outcome

A clean and structured dataset ready for exploratory analysis.

Week 3: Univariate and Temporal Analysis

Objectives

- Analyze individual variables and time-based patterns.

Work Done

- Analyzed accident severity distribution.
- Studied accident frequency by hour, day of week, and month.
- Used bar charts, histograms, and pie charts for visualization.

Outcome

Identification of peak accident hours and strong temporal patterns related to traffic flow.

Week 4: Bivariate and Multivariate Analysis

Objectives

- Explore relationships between accident severity and contributing factors.

Work Done

- Analyzed severity against visibility, weather conditions, road surface conditions, and traffic congestion.
- Created correlation heatmaps, boxplots, and pair plots.
- Identified limitations due to severity class imbalance.

Outcome

Observed weak relationships between severity and environmental factors, largely influenced by data imbalance.

Week 5: Geospatial Analysis

Objectives

- Identify accident hotspots.

Work Done

- Visualized accident locations using scatter plots and density maps.
- Identified top five accident-prone states and cities.

Outcome

Clear geographic clustering of accidents in urban and high-traffic regions.

Week 6: Insight Extraction and Hypothesis Testing

Objectives

- Answer key analytical questions using statistical testing.

Work Done

- Defined null and alternative hypotheses.
- Applied chi-square tests, t-tests, and Pearson correlation tests.
- Analyzed p-values and provided conclusions.

Outcome

Time of day significantly affects accident frequency, while severity shows limited dependence on weather and visibility due to class imbalance.

Week 7: Dashboard Development

Objectives

- Build an interactive visualization platform.

Work Done

- Developed an interactive Streamlit dashboard.
- Visualized accident trends, hotspots, and state-wise statistics.
- Enabled user interaction through filters.

Outcome

Transformation of static EDA results into an interactive analytical dashboard.

5 Key Insights

- Accident frequency peaks during rush hours.
- Urban and highly populated regions experience more accidents.
- Weather and visibility alone do not strongly influence accident severity.

6 Limitations

- Highly imbalanced severity distribution.
- Lack of direct road surface condition data.
- Absence of traffic speed and volume information.