SafetyNet Project Report: Phase I - Data Cleaning and Exploratory Data Analysis (EDA)

Nishant Singh, Roll no. 230706

Week 1 Progress: December 10, 2025

Abstract

This report documents the initial phase (Week 1) of the SafetyNet project, focusing on preparing and exploring a large dataset of workplace incidents. The raw data, consisting of more than 100,000 rows and 28 columns, was systematically cleaned by removing irrelevant identifiers, redundant geographical data, and non-informative text fields. A robust Exploratory Data Analysis (EDA) was then performed to uncover key patterns, including hospitalization rates, geographical incident hotspots (top states and cities), and the most frequent injury types, body parts, and incident events. Key findings include the proportion of amputations across different body parts and the temporal trend of incidents over the recorded period.

Contents

1 Introduction

The SafetyNet project aims to analyze a large dataset of employee incidents to identify risk factors and common trends, ultimately informing safety policy improvements. This Phase I report details the crucial steps of data preparation and initial exploratory analysis.
The initial dataset contained records of accidents, including details about the employee's living area, the part of the body injured, and the mechanism of injury.

2 Data Cleaning and Preprocessing

The raw dataset was substantial, featuring **28 columns** and **more than 100,000 rows**. The data cleaning process was critical to reduce noise and enhance the quality of the subsequent analysis.

2.1 Column Selection and Removal

Several columns were identified as having low analytical value or containing redundant/sensitive information, and were consequently dropped:

- **Identifiers/Administrative Data**: `ID, UPA, Inspection, FederalState`
- **Detailed Geographical Data**: `Address2, Latitude, Longitude, Zip`
- **Redundant/High-Cardinality Text**: `Employer, Final Narrative, Secondary Source, Secondary Source Title`
- **Target Duplicates**: The column related to 'loss of eye' was removed as it contained no positive cases.

The final set of columns retained were those essential for understanding the **what, where, how, and severity** of the injury (e.g., body part, event, location, hospitalization status, year, month).

2.2 Data Integrity Checks

Following column removal, standard data integrity steps were performed:

- **Duplicate Removal**: A dropdown function was applied to remove any duplicate rows, ensuring each record represents a unique incident.
- **Missing Value Handling**: Any rows containing Null/NaN values across the remaining columns were dropped to ensure the integrity of the EDA calculations.

3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted on the cleaned dataset to uncover patterns and distribution characteristics. The key analytical areas are detailed below.

3.1 Incident Frequency and Severity

3.1.1 Hospitalization Analysis

A direct count was performed to determine the total number of individuals hospitalized versus those who were not.

3.1.2 Top Incident Characteristics

The following analyses focused on the most frequent categories:

- **Top 10 Injury Nature** (What is the injury): [List the top 10 injury natures (e.g., Fractures, Sprains, Burns)].
- **Top 10 Incident Events** (How did the injury occur): [List the top 10 incident events (e.g., Falls, Struck by Object, Overexertion)].
- **Top 10 Injured Body Parts**: [List the top 10 most injured body parts (e.g., Hand, Back, Knee)].

3.2 Geographical Hotspots

The location of incidents was analyzed to identify high-risk areas:

- **Top 10 States for Incidents**: A count of incidents per state revealed the primary geographical areas of concern.
- **Top 10 Cities for Incidents**: Further granularity was achieved by identifying the top cities contributing to the incident count.

3.3 Rate-Based Analysis (Proportions)

Calculations involving the mean of binary (0/1) columns were used to determine rates or proportions:

3.3.1 Amputation Rate by Body Part

The mean of the binary `Amputation` column, grouped by `Part of Body Title`, was calculated. This value represents the proportion of injuries resulting in an amputation for that specific body part.

3.3.2 Hospitalization Rate by Event

The mean of the binary `Hospitalized` column, grouped by `EventTitle`, was calculated. This indicates which incident types have the highest risk of leading to hospitalization.

Table 1: Mean Amputation Rate Grouped by Body Part

Part of Body Title	**Incident Count**	**Amputation Rate (Mean)**
Finger(s)	around 25k	0.9-1
Hand (not otherwise classified)	around 26k	0.8-0.9
...	...	...

3.4 Temporal Analysis

3.4.1 Monthly Trend

The number of incidents was counted by month and year using the `Month` period column. This data was used to plot the time series line graph, illustrating the overall monthly trend of incidents over the years covered by the dataset.

3.4.2 Pair Plot Analysis

A pair plot was generated to visually explore the bivariate relationships and distributions between the key numerical severity and time variables: `Hospitalized`, `Amputation`, and `Year`. This helps in identifying potential correlations or clustering patterns between these variables.

4 Conclusion and Next Steps

The data cleaning process successfully transformed the raw data into a manageable and reliable format, discarding over a dozen non-critical columns and handling missing/duplicate values. The EDA successfully highlighted key risk areas, including the states and cities with the highest incident counts, the most frequent injury mechanisms, and the rate of severe outcomes (hospitalization and amputation) linked to specific injury characteristics.