AI ML Capstone Project – NLP 1 Interim Report
for
NLP-Based Safety Chatbot of Industrial companies

Project Interim Report prepared by

Dinesh Chauhan

Saravana Kumar Kandasamy

Sujitha Harini

Vojjala Srikar

Date: 18-Aug-2024

Table of Contents

1. Project Overview and Key Insights

1.1 Problem Statement:

In industrial environments, the prevalence of accidents and injuries remains a significant challenge. Despite existing safety measures, incidents continue to occur, sometimes resulting in severe injuries or fatalities. The need to understand the underlying causes of these accidents and to prevent future occurrences is critical.

To address this challenge, the project aims to design and implement an advanced ML/DL-based chatbot. The chatbot will analyze and interpret detailed descriptions of industrial accidents from a comprehensive dataset. This dataset includes records from 12 manufacturing plants across three different countries, encompassing a range of accident severity levels and risk factors.

Objectives:

1. **Develop a Chatbot for Accident Analysis:**
    - Create a chatbot capable of processing and analyzing detailed accident descriptions.
    - Use natural language processing (NLP) and machine learning (ML) techniques to understand and interpret the content of accident records.
2. **Identify and Highlight Safety Risks:**
    - Enable the chatbot to identify significant safety risks based on the analyzed accident descriptions.
    - Provide safety professionals with insights into common risk factors and patterns observed in the data.
3. **Enhance Workplace Safety:**
    - Utilize the chatbot's insights to recommend improvements to safety protocols and procedures.
    - Aim to prevent future accidents by addressing identified risks and implementing recommended changes.

Rationale:

- **Urgency of Safety Improvements:** Industrial accidents pose serious risks to employee well-being and operational efficiency. Effective accident analysis and risk identification are crucial for enhancing workplace safety.
- **Data-Driven Insights:** The dataset provides a rich source of information on various accident-related factors. By analyzing this data, the chatbot can uncover patterns and insights that are not immediately apparent through traditional methods.
- **Advanced Technology:** Leveraging ML and NLP techniques allows for the automation of complex analysis tasks, providing timely and accurate insights that can help in making informed safety decisions.
- **Preventive Measures:** By understanding the root causes of accidents and highlighting potential risks, the chatbot can help in implementing preventive measures, thereby reducing the likelihood of future incidents and improving overall safety standards.

Expected Impact:

The chatbot is expected to significantly enhance the ability of safety professionals to analyze accident data, identify risks, and implement effective safety measures. This will contribute to a safer working environment, reduce the incidence of accidents, and improve overall safety management practices.

1.2 About Data and Our Findings

The dataset used in this project originates from industrial. This dataset is crucial for understanding and addressing industrial safety concerns.

- **Source:** The data is collected from 12 different plants located in three different countries. The source provides comprehensive records of industrial accidents, which are essential for analyzing safety issues.
- **Format:** The data is structured in a tabular format and includes various columns that capture different aspects of each accident. The dataset is available in CSV format, making it easy to work with for data analysis and preprocessing tasks.
- **Content:** Each record in the dataset represents an individual accident, including detailed descriptions and contextual information about the incidents. This includes time and date of the accidents, locations, severity levels, and other relevant details that help in understanding the nature and impact of the accidents.
- **Purpose:** The dataset is intended to be used for developing an ML/DL-based chatbot. This chatbot will analyze the accident descriptions to identify safety risks and provide recommendations to prevent future incidents.

Dataset Attributes and Description

- **Data:** Timestamp or date and time information of the accident.
- **Countries:** Anonymized country identifiers where the accident occurred.
- **Local:** Anonymized city identifiers where the manufacturing plant is located.
- **Industry Sector:** The sector to which the plant belongs.
- **Accident Level:** A severity rating from I to VI, where I denotes a minor accident and VI denotes a very severe accident.
- **Potential Accident Level:** Indicates the possible severity of the accident based on other contributing factors.
- **Genre:** Gender of the person involved (male or female).
- **Employee or Third Party:** Specifies whether the injured person is an employee or a third party.
- **Critical Risk:** Description of the critical risk factors involved in the accident.
- **Description:** A detailed narrative of how the accident occurred.

1.3 Key Findings from Accident Data Analysis

1. **Severity Distribution:** The dataset shows a wide range of accident severity levels, with fewer instances of very severe accidents (Levels V and VI) compared to less severe ones (Levels I and II). This distribution highlights varying degrees of risk across different plants and locations.
2. **High-Risk Industry Sectors:** Certain industry sectors are associated with higher accident rates and more severe incidents. This suggests that specific industries may need more stringent safety protocols and preventive measures.

3. **Geographic Patterns:** Accidents are not evenly distributed across all locations. Some plants or regions experience higher frequencies of accidents, indicating potential location-based risk factors or safety issues.
4. **Temporal Patterns:** The data reveals patterns related to the time of day or shift during which accidents occur. For instance, certain shifts might be more prone to accidents, pointing to potential operational or staffing issues.
5. **Gender Differences:** There are noticeable differences in accident rates and severity between male and female employees. This finding suggests that safety measures may need to be tailored to address gender-specific risks and ensure equitable safety practices.

1.4 Implications

1. **Enhanced Safety Protocols:** The identification of high-risk industry sectors and locations highlights the need for more stringent and tailored safety protocols. Implementing enhanced safety measures in these areas can reduce accident rates and improve overall safety.
2. **Targeted Risk Management:** The temporal patterns and shifts associated with higher accident rates suggest that operational changes or additional safety measures during specific times could mitigate risks. Addressing staffing and operational issues during high-risk periods can help prevent accidents.
3. **Gender-Specific Safety Measures:** Notable differences in accident rates and severity between male and female employees indicate that safety measures may need to be adapted to address gender-specific risks. Developing gender-inclusive safety protocols can ensure equitable protection for all employees.
4. **Predictive and Preventive Strategies:** The chatbot's ability to analyze and interpret accident descriptions allows for predictive insights into potential safety risks. By leveraging these insights, industries can adopt preventive strategies to avoid future accidents and improve overall safety practices.
5. **Data-Driven Decision Making:** The project underscores the value of using data-driven approaches to enhance industrial safety. By continuously analyzing accident data, industries can make informed decisions, adapt safety measures based on real-time information, and drive continuous improvement in safety standards.

2. EDA and Text Preprocessing analysis

This section details the major components of the analysis:

- **EDA - Analysis on Categorical Columns and Numerical Data:**
  - A comprehensive exploration was conducted using univariate, bivariate, numerical analysis, and statistical tests to identify key factors influencing the target variable.
- **NLP Pre-processing on the Description Column:**
  - Various NLP techniques were employed to preprocess the accident descriptions, including n-gram analysis and Word Cloud generation, to extract meaningful terms that could be predictive of the target variable.

2.1 Initial Data Observations

Before delving into the EDA, the dataset was carefully examined to understand its structure and quality. Here are the key observations:

- **Dataset Shape:** The dataset consists of 425 records and 11 columns, indicating a moderately sized dataset suitable for thorough analysis.
- **Data Types:**
    - The columns exhibit a mix of data types:
        - **Numerical:** Unnamed: 0 (int64)
        - **Datetime:** Data (datetime64[ns])
        - **Categorical:** Countries, Local, Industry Sector, Accident Level, Potential Accident Level, Genre, Employee or Third Party, Critical Risk, Description (object)

    This variety of data types suggests that the dataset includes both quantitative and qualitative information, which will be analyzed accordingly in the EDA phase.
- **Missing Values:**
    - The dataset is complete with no missing values across any columns, ensuring that all variables are fully populated. This eliminates the need for imputation or data cleansing and allows for a straightforward analysis.

```
Dataset shape: (425, 11)
Data types:
 Unnamed: 0                         int64
Data                       datetime64[ns]
Countries                          object
Local                              object
Industry Sector                    object
Accident Level                     object
Potential Accident Level           object
Genre                              object
Employee or Third Party            object
Critical Risk                      object
Description                        object
dtype: object
Missing values:
 [0 0 0 0 0 0 0 0 0 0 0]
```

Continuing from the initial observations regarding missing values, the following preprocessing steps were performed to refine and standardize the dataset for analysis:

- **Column Renaming:**
    - **Objective:** To enhance clarity and consistency in the dataset by using more descriptive and standardized column names.
    - **Actions:** The columns were renamed as follows:
        - Data was renamed to Date
        - Industry Sector to Industry_Sector
        - Accident Level to Accident_Level
        - Countries to Country
        - Genre to Gender
        - Potential Accident Level to Potential_Accident_Level
        - Employee or Third Party to Employee_Type
        - Critical Risk to Critical_Risk

These changes improve the readability of the dataset and align the column names with common naming conventions, making subsequent analyses more intuitive.

- **Removal of Unnecessary Columns:**
  - **Objective:** To streamline the dataset by eliminating irrelevant columns that do not contribute to the analysis.
  - **Actions:** The Unnamed column, which likely served as an index or placeholder, was removed from the dataset.

  By excluding this column, the dataset now contains only relevant information, ensuring a more focused and efficient analysis.

2.2 Knowing more about Dataset:

- **Checking for Duplicates in the Description Column:**
  - **Objective:** To ensure the dataset's integrity by identifying and removing any duplicate records, particularly in the Description column, which is crucial for the analysis.
  - **Actions:**
    - The Description column was inspected for duplicate entries.
    - **Outcome:** 7 duplicate rows were found. These duplicates could skew the analysis, particularly in NLP-based processing, by over-representing certain incidents.
- **Dropping Duplicates:**
  - **Objective:** To refine the dataset by eliminating redundant entries, thereby ensuring that each record is unique and represents distinct incidents.
  - **Actions:** The identified 7 duplicate rows were removed from the dataset.
  - **Result:** The dataset was reduced from 425 records to 418 records, with 10 columns remaining after the duplicates were dropped.

```
Found 7 duplicate rows. Dropping them...
New dataset shape after dropping duplicates: (418, 10)
```

Observation on the Description Column:

- **Observation:** Despite the dataset containing 418 records, the Description column has only 411 unique values. This discrepancy indicates that there are still 7 duplicate entries within the Description column, which could potentially distort the analysis if not addressed.
- **Assumption:** Given this observation, it is assumed that these 7 records are duplicates that need to be identified and removed to maintain the dataset's integrity.

	Description
count	418
unique	411
top	During the activity of chuteo of ore in hopper...
freq	2

Inspection of Rows with Duplicate Descriptions:

- **Objective:** To further ensure the quality of the data by closely examining the rows with duplicate values in the Description column before their removal. This step was essential to confirm whether any of the duplicates contained slight variations or additional context that could be relevant to the analysis.
- **Actions:**
  - A detailed inspection was conducted on the rows identified as having duplicate descriptions. This was done using a function specifically designed to filter and display the duplicate entries.
  - The function inspect_description_duplicates(df) was implemented to check for duplicates in the Description column. This function provides a sample of the rows with duplicate descriptions, offering a closer look at these entries.
  - During this inspection, the function checked if the Description column was present in the DataFrame and then identified and displayed the rows with duplicate descriptions.
- **Outcome:**
  - The inspection revealed that there were indeed duplicate descriptions within the dataset. A sample of these rows was displayed to facilitate a manual review.
  - This process ensured that the duplicates were not removed hastily, allowing for an informed decision on whether to retain any entries that might contain unique information despite having similar descriptions.

```
Sample of rows with duplicate descriptions:
                                          Description
37   When starting the activity of removing a coil ...
130  In the geological reconnaissance activity, in ...
143  Project of Vazante that carried out sediment c...
166  At moments when the MAPERU truck of plate F1T ...
261  During the activity of chuteo of ore in hopper...
```

- **Removing Duplicate Rows Based on the `Description` Column:**
  - **Objective:** To ensure that each entry in the dataset is unique by eliminating any redundant rows with duplicate descriptions.
  - **Actions:**
    - A function `remove_description_duplicates(df)` was employed to drop duplicate rows from the `Description` column while retaining the first occurrence of each unique description.
    - This function was applied to the dataset, and the number of removed duplicates was calculated.
  - **Outcome:**
    - The function successfully removed 7 duplicate rows based on the Description column.
    - As a result, the dataset was cleaned and now contains 411 records, each with a unique description.

```
Removed 7 duplicate rows based on 'Description' column.
```

This step finalized the data cleaning process for the Description column, ensuring the dataset is now free from redundant entries and ready for further analysis.

2.3 Visualisation Insights (Univariate analysis):

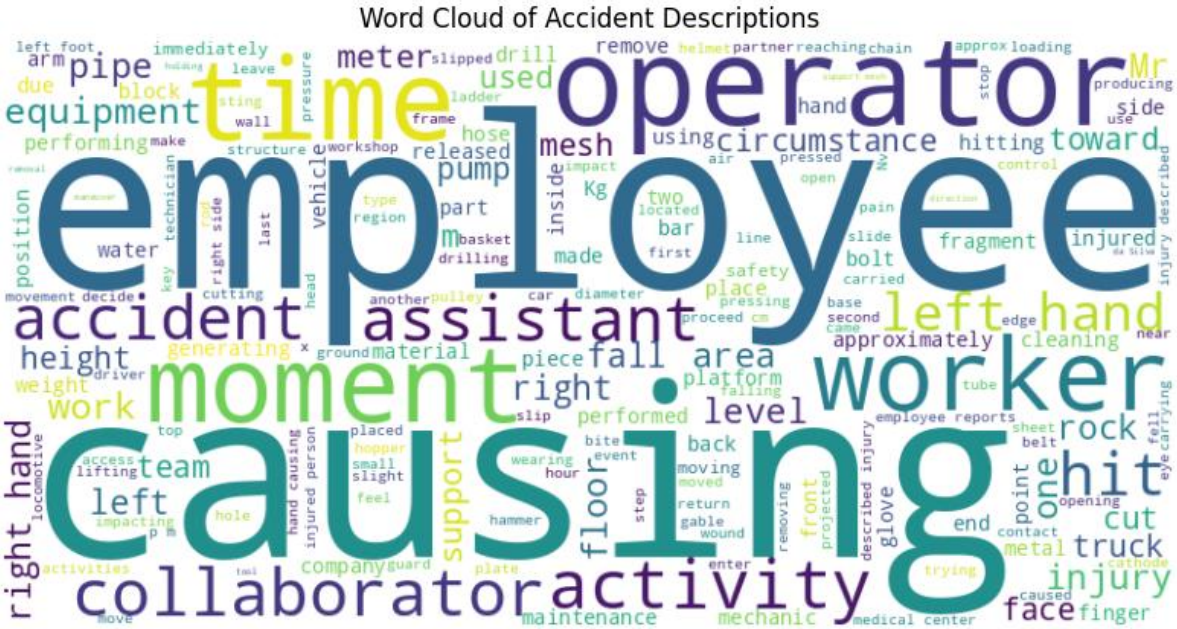
With the dataset pre-processed, the next phase involves detailed exploratory data analysis:

- **Univariate Analysis:**

  Understand the distributions and characteristics of individual variables**.** This involves gaining insights into how each variable behaves on its own, which provides a foundation for further analysis and interpretation.

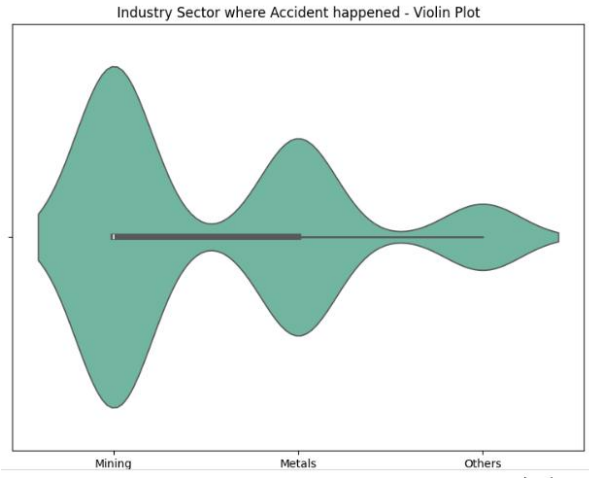
Word Cloud for Accident Descriptions

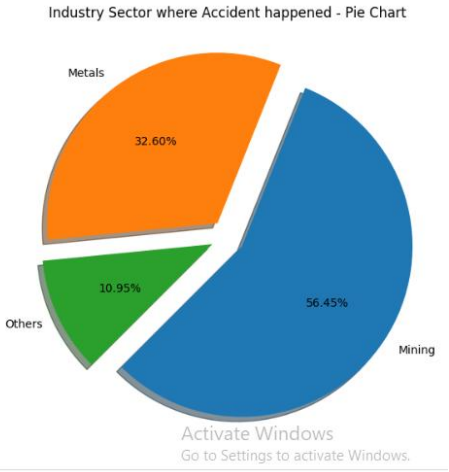
- **Purpose:** To visualize the most common words in the Description column of the dataset.
- **Insight:** The word cloud highlights frequently occurring terms in accident descriptions, helping identify prevalent themes and key issues. This visualization reveals important keywords that may inform further analysis and model development.



Industry Sector

Purpose: To analyze the distribution of accident occurrences across different industry sectors.

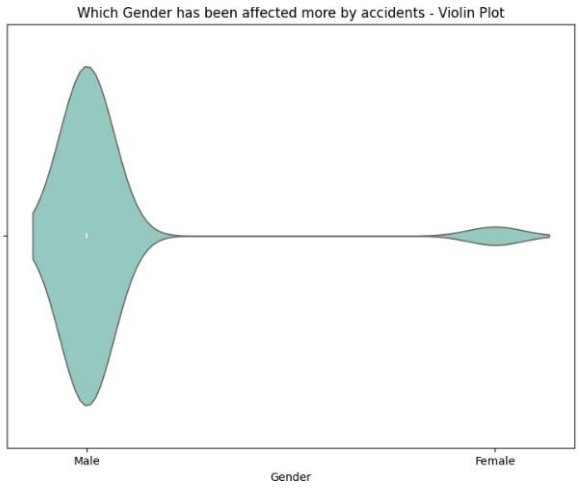


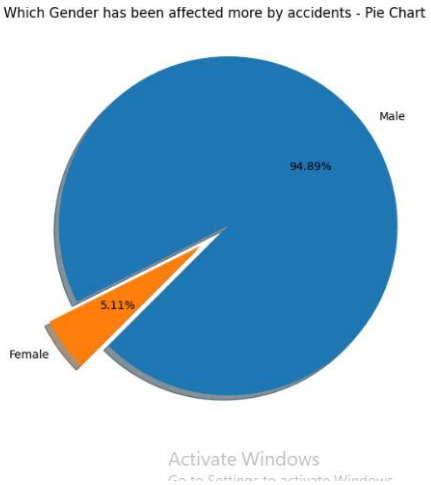


Insight: The univariate plots reveal that the mining industry has a significantly higher frequency of accidents compared to the metal industry. This indicates that jobs in the mining sector are notably riskier than those in the metal sector. This insight can guide targeted safety interventions and risk assessments for different industries.

Gender

Purpose: To examine the distribution of accidents by gender.

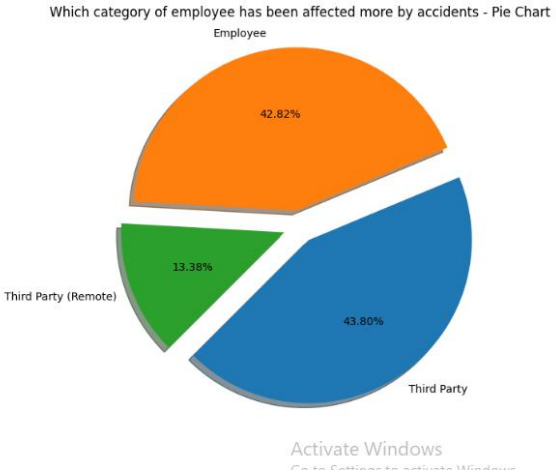




Insight: The analysis indicates a significant bias towards male employees being affected by accidents. This highlights that the dataset predominantly features a higher proportion of male employees, reflecting a need to address gender-specific safety measures and considerations.

Employee Type

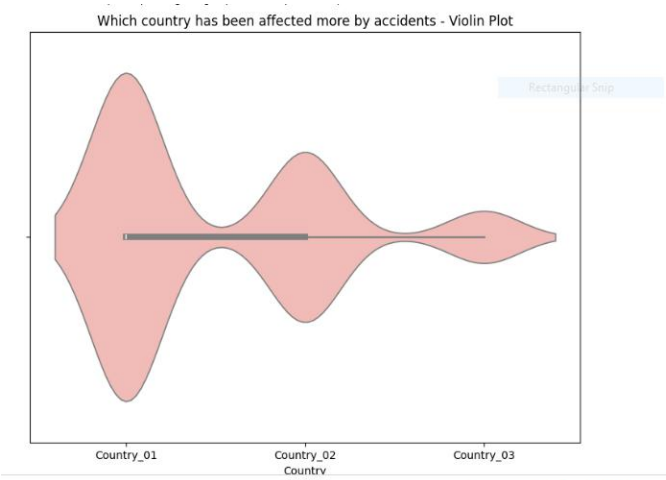
Purpose: To analyze the distribution of accidents across different categories of employees.

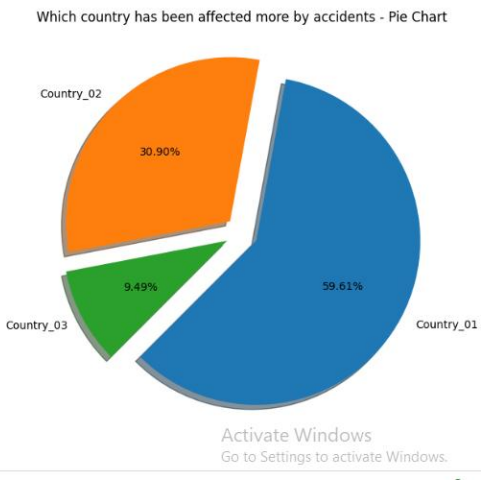


Insight: The analysis shows that the number of direct employees and third-party employees is nearly equal, while the number of third-party remote employees is relatively lower. This suggests that direct and third-party employees are equally impacted by accidents, with fewer incidents affecting third-party remote employees.

Country

Purpose: To evaluate the distribution of accidents across different countries.

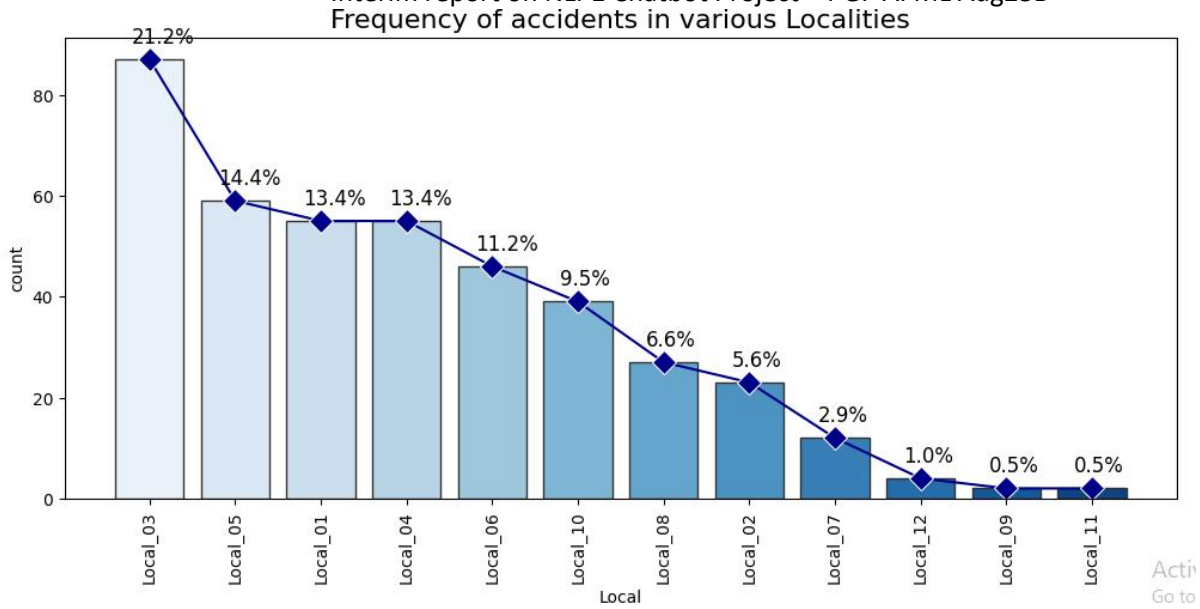




Insight: The analysis indicates that Country 1 has experienced more accidents compared to Country 2, while Country 3 has had fewer incidents. This highlights varying levels of incident frequency across countries, which may guide country-specific safety measures and resource allocation.

Local

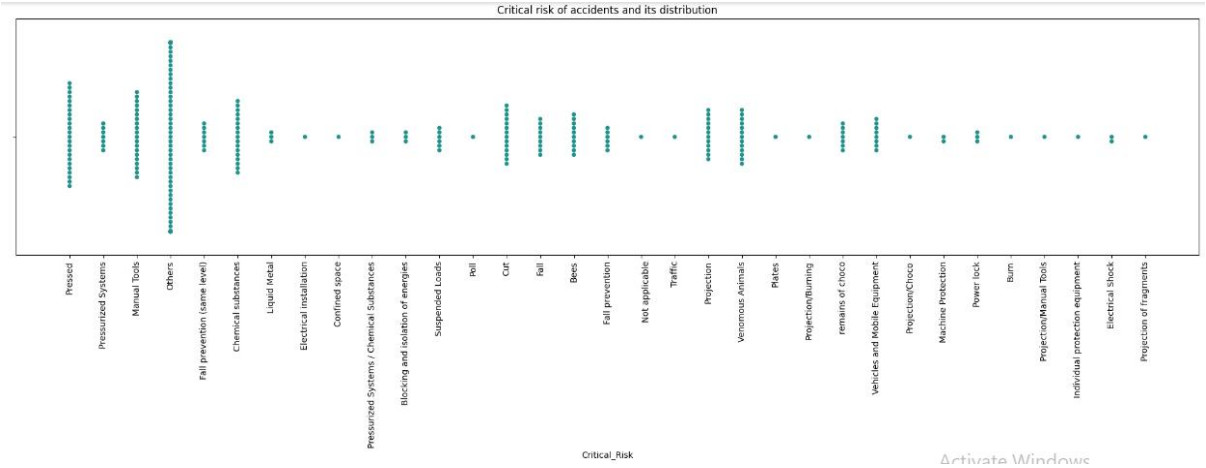
Purpose: To examine the distribution of accidents across different localities.



Insight: The analysis reveals that Local_03 has reported the highest number of accidents, accounting for approximately 21% of all incidents. This is followed by Local_05, Local_01, and other localities. The results highlight the localities with the highest accident frequencies, suggesting a need for targeted safety measures and further investigation into the factors contributing to higher incident rates in these areas.

Critical Risk

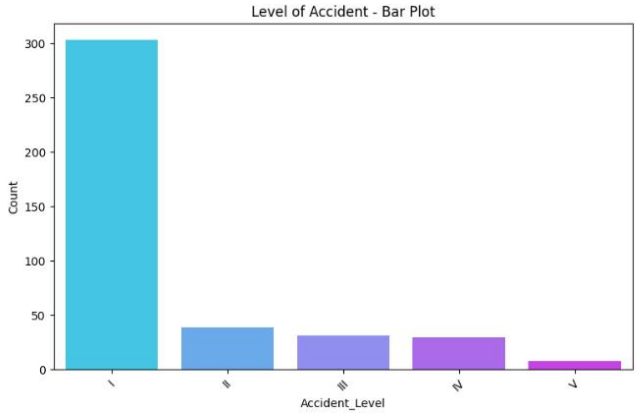
Purpose: To visualize the distribution of different critical risk categories associated with accidents.

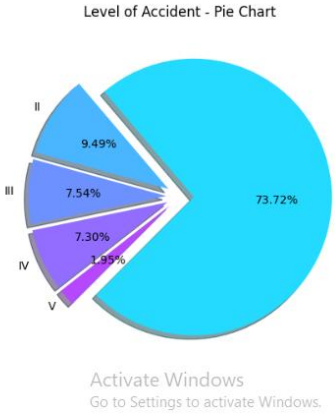


Insight: The analysis shows that nearly 50% of the dataset's critical risks are classified as 'Others,' highlighting a need for more precise risk classification. The remaining risks are categorized into Pressed, Manual tools, Chemical substances, Cut, among others. This distribution indicates the predominant types of risks and suggests areas for improving risk categorization and management.

Accident Level

Purpose: To analyze the distribution of accidents across different severity levels.

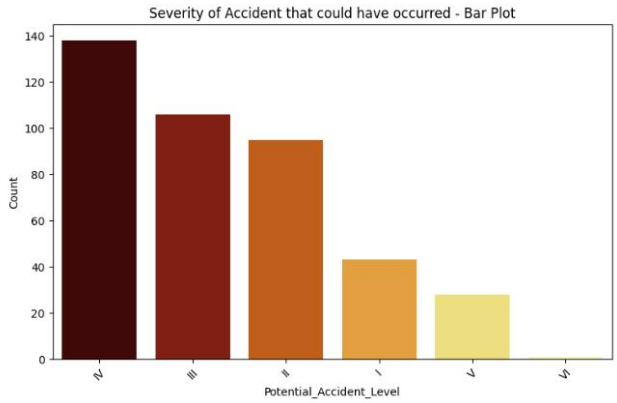


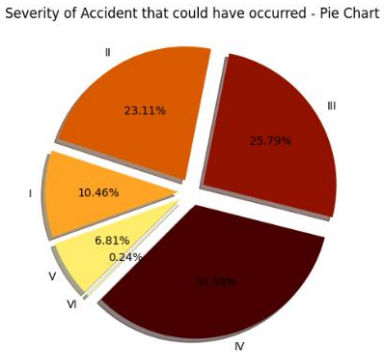


Insight: The analysis shows that Level I, representing minor severity, is the most frequent category, often involving minor issues such as forgetting PPE or dropping a tool. In contrast, Level V denotes extreme severity. This distribution provides insight into the severity of accidents and can help prioritize safety measures and interventions based on the severity levels.

Potential Accident Level

Purpose: To analyze the distribution of potential accident severity levels.





Insight: The analysis reveals that Potential Accident Level IV, indicating moderate severity, has the highest count. This suggests that many accidents have a potential for moderate severity based on additional factors. Further examination is needed to understand the correlation between Potential Accident Level and Accident Level, as well as its relation to the industry sector, to better assess risk factors and improve safety measures.

2.4 Visualisation Insights on Bivariate, Numerical and Statistical Analysis:

Bivariate Analysis:

To investigate how the Potential_Accident_Level interacts with other variables in the dataset, such as industry sector, location, and employee type, to identify factors that influence the severity of potential accidents. This analysis aims to uncover patterns and relationships that can inform targeted safety measures and improve risk assessment strategies.

Numerical Analysis:

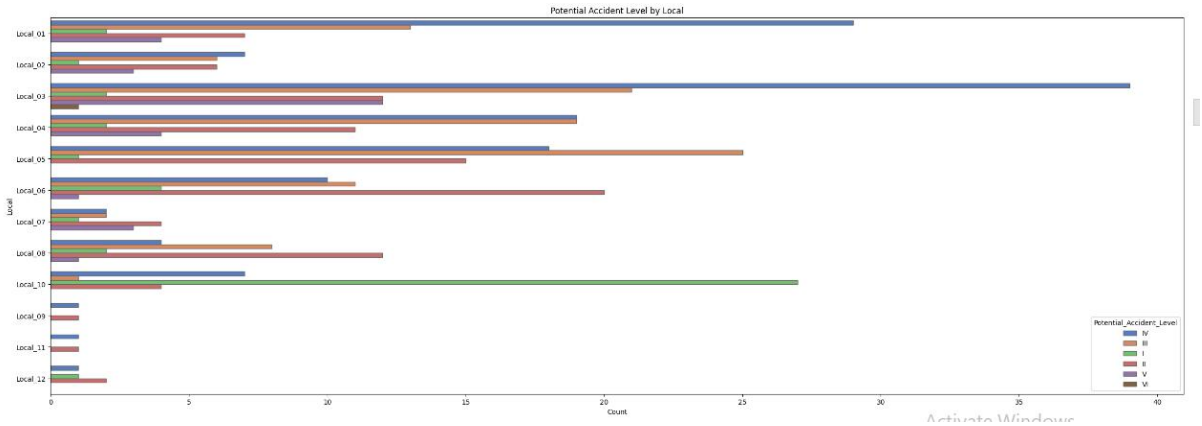
To analyze the numerical data related to the Potential_Accident_Level and other variables. This includes computing summary statistics such as means, medians, and variances to understand the central tendencies and dispersion of the data.

Statistical Tests:

To assess the significance of relationships between Potential_Accident_Level and other categorical variables using statistical tests.

Local vs. Potential Accident Level

To analyze the distribution of Potential_Accident_Level across different localities.



Insight:

Local_03: This location, situated in Country_01, has the highest frequency of accidents, including the most severe cases. It is identified as the riskiest area, indicating a critical need for targeted safety improvements and enhanced risk management strategies.

Local_10: This locality reports many minor accidents (Level I) but fewer severe incidents. It suggests a focus on preventive measures for minor issues may be beneficial.

Local_05 and Local_06: Both locations have high accident counts, with most incidents falling into moderate severity levels. These areas may benefit from focused safety interventions to address moderate risks.

Local_09, Local_11, and Local_12: These locations report very few accidents, indicating generally safer conditions compared to others.

Potential_Accident_Level	I	II	III	IV	V	VI	All
Local							
Local_01	2	7	13	29	4	0	55
Local_02	1	6	6	7	3	0	23
Local_03	2	12	21	39	12	1	87
Local_04	2	11	19	19	4	0	55
Local_05	1	15	25	18	0	0	59
Local_06	4	20	11	10	1	0	46
Local_07	1	4	2	2	3	0	12
Local_08	2	12	8	4	1	0	27
Local_09	0	1	0	1	0	0	2
Local_10	27	4	1	7	0	0	39
Local_11	0	1	0	1	0	0	2
Local_12	1	2	0	1	0	0	4
All	43	95	106	138	28	1	411

Numerical Insights :

- The numerical analysis corroborates the findings from the bivariate analysis, highlighting that Local_03 has the highest number of severe accidents, reinforcing its designation as the riskiest area. Other locations like Local_10, Local_05, and Local_06 show varying levels of risk, with some indicating the need for specific safety measures based on the severity of incidents.

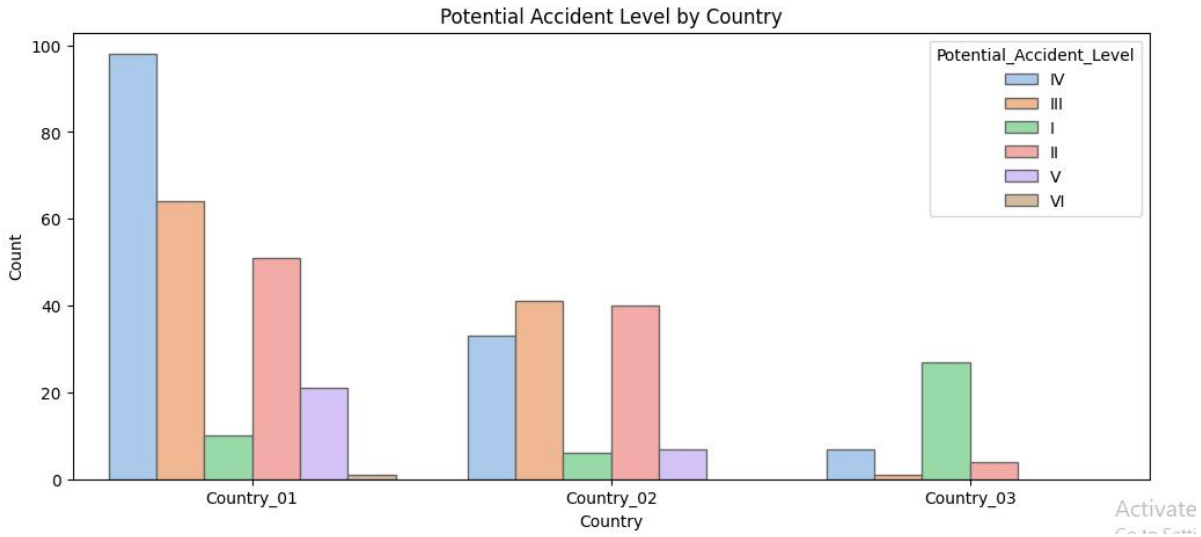
```
Chi-square Statistic: 235.36221761186388, p-value: 3.594507833409111e-24
```

Statistical Insights :

- **Chi-square Statistic:** 235.36
- **p-value:** 3.59e-24
- **Insight:** The extremely low p-value indicates a significant association between Potential_Accident_Level and the variables being analyzed. This suggests strong evidence of relationships between accident severity levels and the factors studied, underscoring the importance of these relationships in shaping risk management strategies.

Country vs. Potential Accident Level

To explore how the Potential_Accident_Level varies across different countries. This analysis aims to identify country-specific patterns in accident severity and frequency, which can guide targeted safety measures and resource allocation.



Observation:

Country_01 has the highest concentration of severe accidents (Levels IV and V), indicating it is the most critical area for severe incidents. **Country_02** shows moderate severity distribution, while **Country_03** reports fewer and less severe accidents.

Potential_Accident_Level	I	II	III	IV	V	VI	All
Country							
Country_01	10	51	64	98	21	1	245
Country_02	6	40	41	33	7	0	127
Country_03	27	4	1	7	0	0	39
All	43	95	106	138	28	1	411

Numerical Observations :

Country_01: Most severe accidents, especially at Level IV, making it the riskiest location.
Country_02: Moderate accident counts with fewer severe incidents compared to Country_01.

Country_03: Lowest accident numbers, mostly minor, suggesting it is relatively safer.

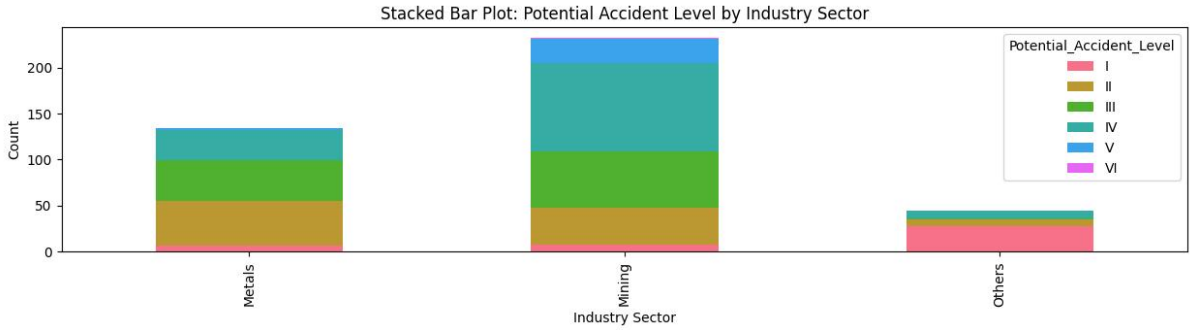
```
Chi-square Statistic: 172.51781076583225, p-value: 8.348502170052156e-32
```

Statistical Observations :

Chi-square Statistic: 172.52, **p-value:** 8.35e-32

Observation: Significant association between country and accident severity, confirming that severity levels vary significantly across countries. This highlights the need for tailored safety measures in different countries.

Industry Sector vs. Potential Accident Level



Observation:

- The Mining Industry has experienced the most severe accidents, with the highest corresponding potential accident levels.
- The Metal industry follows, with other industries having fewer severe incidents.
- The Mining sector shows a higher rate of Level IV accidents, slightly surpassing Level II and III incidents in severity.

Potential_Accident_Level	I	II	III	IV	V	VI	All
Industry_Sector							
Metals	7	48	44	33	2	0	134
Mining	8	40	61	96	26	1	232
Others	28	7	1	9	0	0	45
All	43	95	106	138	28	1	411

Numerical Observations :

Mining is the top priority for safety improvements due to its high overall and severe accident counts.

Metals should enhance safety measures to reduce moderate-level accidents.

Others appears safer but should maintain safety practices to avoid potential future issues.

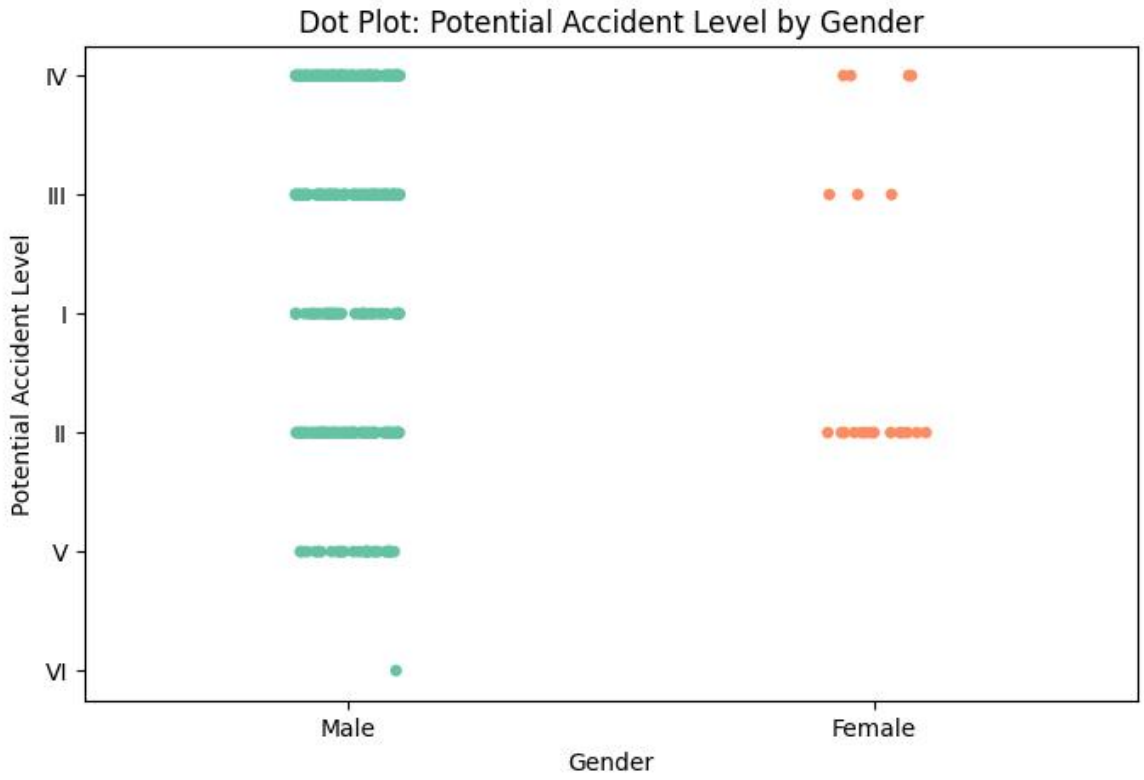
```
Chi-square Statistic: 181.73738344191983, p-value: 1.0209410384588079e-33
```

Statistical Observations :

The extremely low p-value indicates a statistically significant difference in accident severity levels across industry sectors.

This suggests that the distribution of accident severity varies considerably between different industry sectors, highlighting the need for sector-specific safety strategies and interventions.

Gender vs. Potential Accident Level



Males are predominantly involved in more severe accidents, while females are more frequently associated with less severe accidents, particularly at Level II.

Potential_Accident_Level	I	II	III	IV	V	VI	All
Gender							
Female	0	14	3	4	0	0	21
Male	43	81	103	134	28	1	390
All	43	95	106	138	28	1	411

Numerical Observations :

Male employees are at a higher risk for accidents and more severe accident levels compared to females.

Female employees have fewer accidents and lower severity, suggesting a need for targeted interventions to address higher accident rates among males.

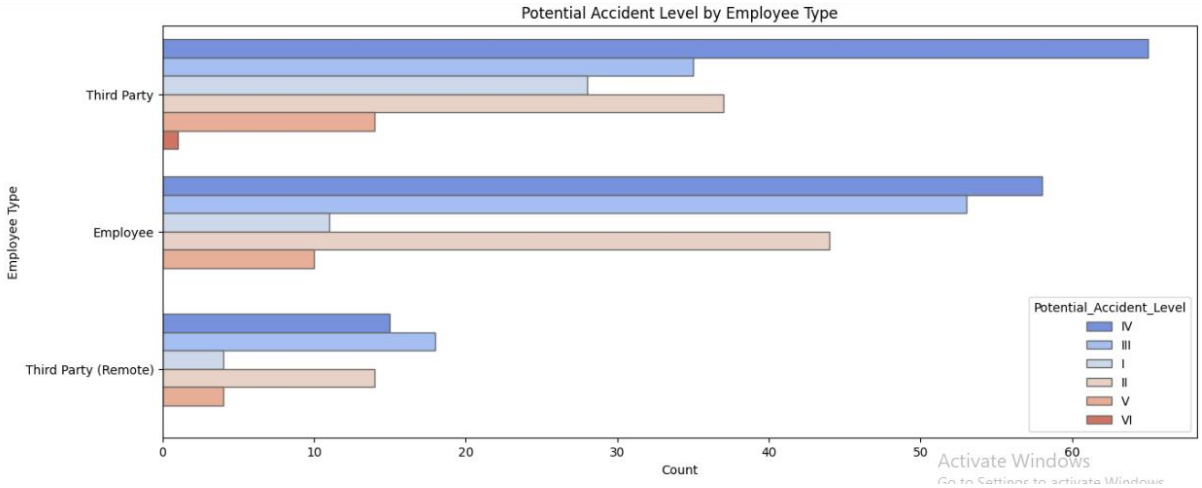
```
Chi-square Statistic: 24.56496792084331, p-value: 0.0001690321312212402
```

Statistical Observations :

The very low p-value indicates a statistically significant difference in accident severity levels between genders.

This suggests that gender significantly influences the distribution of accident severity, with males experiencing more severe accidents compared to females.

Employee Type vs. Potential Accident Level



- Third-party employees are more frequently involved in accidents.
- Besides Accident Level I, employees are also encountering severe accidents, particularly at Accident Level IV, in the industry.

Potential_Accident_Level	I	II	III	IV	V	VI	All
Employee_Type							
Employee	11	44	53	58	10	0	176
Third Party	28	37	35	65	14	1	180
Third Party (Remote)	4	14	18	15	4	0	55
All	43	95	106	138	28	1	411

Numerical Observations :

Third Party is the highest priority for safety interventions due to the high total number of accidents and higher severity levels.

Employee experiences a moderate number of accidents with fewer severe incidents, suggesting a need for focused safety measures to address moderate-risk areas.

Third Party (Remote) appears safer but should still be monitored to maintain safety standards.

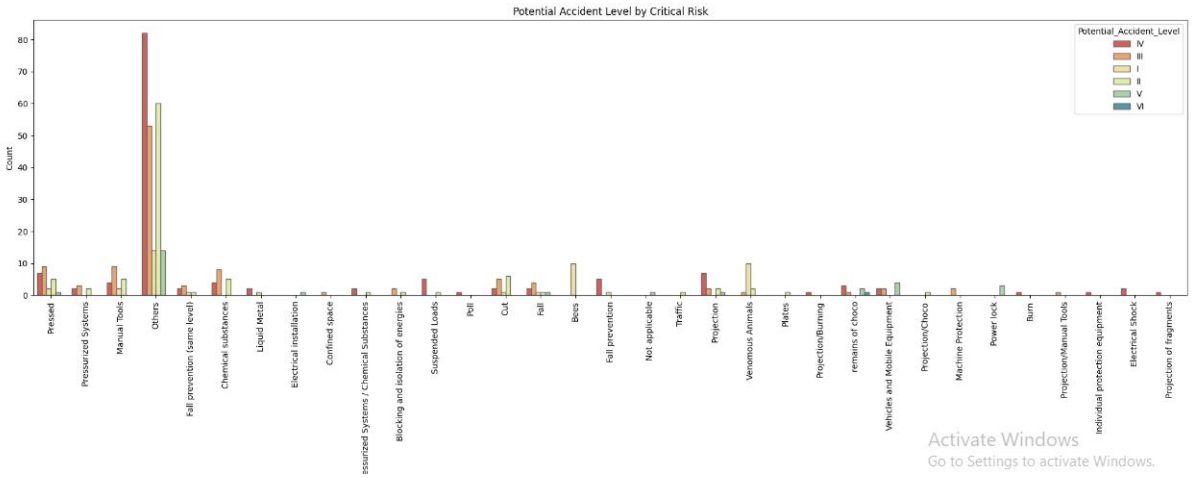
```
Chi-square Statistic: 16.89845575822984, p-value: 0.07664131046907577
```

Statistical Observations :

The p-value of 0.077 is slightly above the conventional significance level of 0.05. This means that the result is not statistically significant at the 5% level.

It implies that there is not strong enough evidence to reject the null hypothesis, suggesting that the differences in accident severity across Employee_Type categories are not statistically significant.

Critical Risk vs. Potential Accident Level



The "Others" category in Critical Risk shows a higher concentration of incidents with severe Potential Accident Levels, indicating a significant risk associated with these unspecified or miscellaneous factors.

Potential_Accident_Level	I	II	III	IV	V	VI	All
Critical_Risk							
\nNot applicable	0	0	0	0	1	0	1
Bees	10	0	0	0	0	0	10
Blocking and isolation of energies	0	1	2	0	0	0	3
Burn	0	0	0	1	0	0	1
Chemical substances	0	5	8	4	0	0	17
Confined space	0	0	1	0	0	0	1
Cut	1	6	5	2	0	0	14
Electrical Shock	0	0	0	2	0	0	2
Electrical installation	0	0	0	0	1	0	1
Fall	1	1	4	2	1	0	9
Fall prevention	1	0	0	5	0	0	6
Fall prevention (same level)	1	1	3	2	0	0	7
Individual protection equipment	0	0	0	1	0	0	1
Liquid Metal	1	0	0	2	0	0	3
Machine Protection	0	0	2	0	0	0	2
Manual Tools	2	5	9	4	0	0	20
Others	14	60	53	82	14	0	223

Plates	0	1	0	0	0	0	1
Poll	0	0	0	1	0	0	1
Power lock	0	0	0	0	3	0	3
Pressed	2	5	9	7	1	0	24
Pressurized Systems	0	2	3	2	0	0	7
Pressurized Systems / Chemical Substances	0	1	0	2	0	0	3
Projection	0	2	2	7	1	0	12
Projection of fragments	0	0	0	1	0	0	1
Projection/Burning	0	0	0	1	0	0	1
Projection/Choco	0	1	0	0	0	0	1
Projection/Manual Tools	0	0	1	0	0	0	1
Suspended Loads	0	1	0	5	0	0	6
Traffic	0	1	0	0	0	0	1
Vehicles and Mobile Equipment	0	0	2	2	4	0	8
Venomous Animals	10	2	1	0	0	0	13
remains of choco	0	0	1	3	2	1	7
All	43	95	106	138	28	1	411

Numerical Observations :

High-risk areas include categories like Others and Manual Tools, which contribute significantly to accident severity.

Low-risk areas are generally those with fewer severe incidents, such as Bees and Burn.

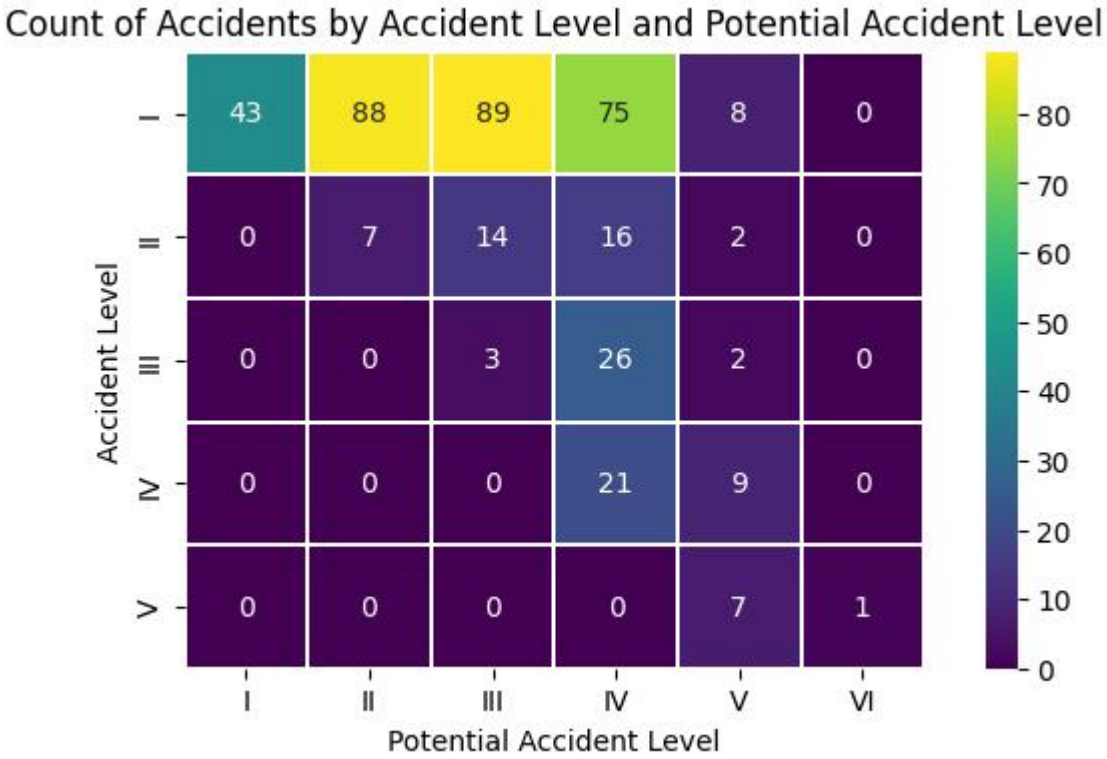
Monitoring and safety interventions should prioritize high-severity categories while ensuring that low-risk factors are not neglected.

```
Chi-square Statistic: 402.803215742986, p-value: 6.543004501368816e-23
```

Statistical Observations :

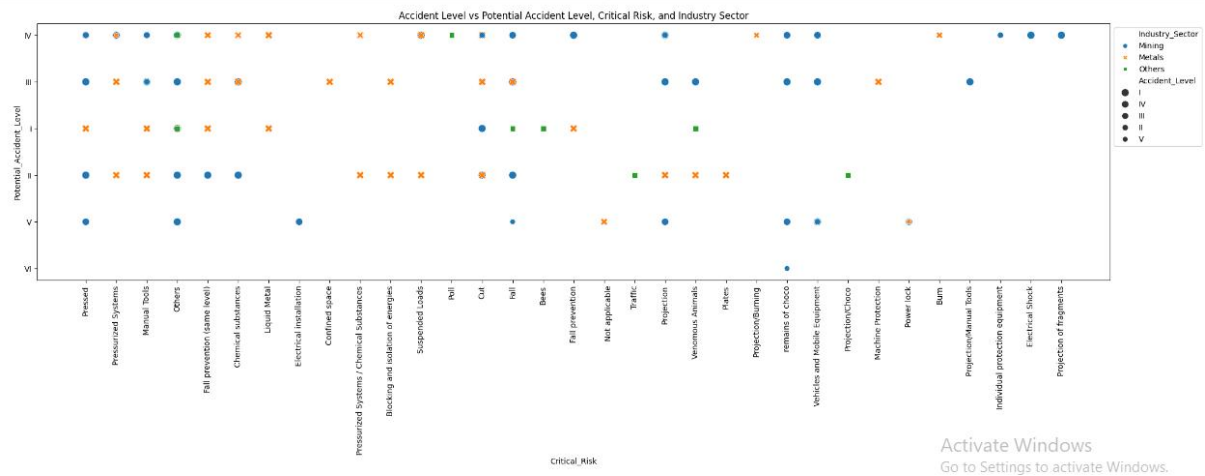
The p-value is extremely low (much less than 0.05), indicating that the result is highly statistically significant.

This suggests a strong association between Critical_Risk and Potential_Accident_Level, meaning that the severity of accidents is significantly influenced by the critical risk factor.



The heatmap reveals a moderate correlation between Accident Level and Potential Accident Level. This indicates that while there is some alignment between the severity of accidents and their potential severity, the relationship is not strongly defined. This insight suggests that while potential accident severity can provide some indication of actual accident severity, other factors may also influence the outcome.

Accident Level vs Potential Accident Level, Critical Risk, and Industry Sector



Observations:

Mining Industry: The Mining Industry has experienced the highest number of severe accidents, with corresponding high potential accident levels.

Accident Distribution: The majority of accidents come from the Mining Industry, primarily of moderate severity, followed by the Metal industry and other sectors.

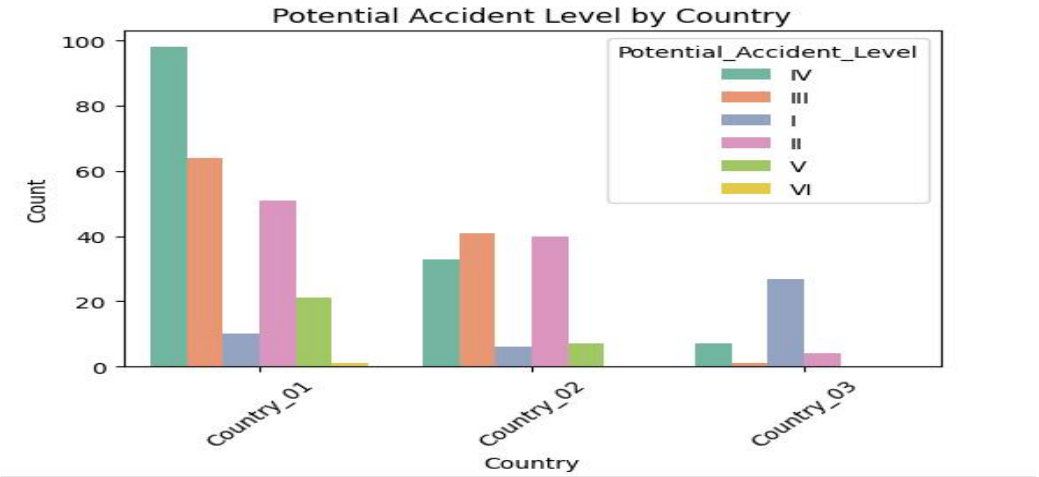
Accident Levels: The severity of accidents (Accident Level) closely matches the potential severity (Potential Accident Level).

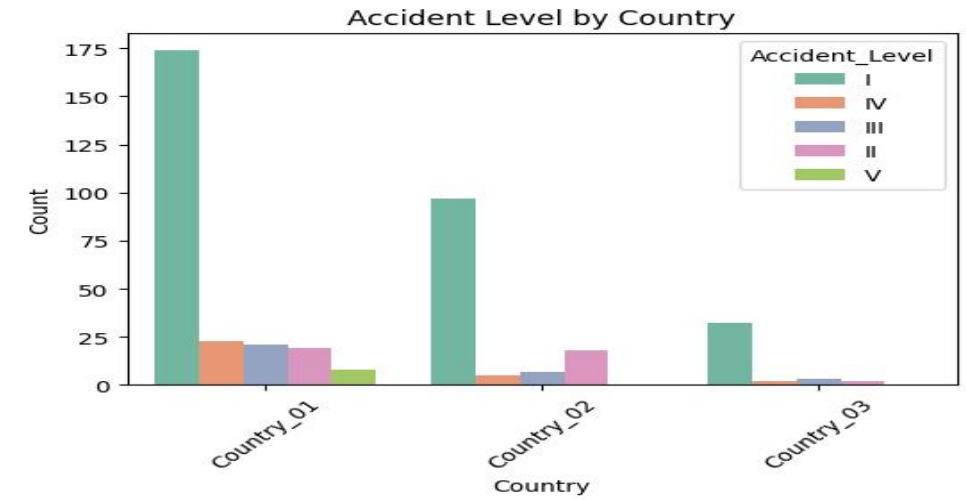
Critical Risks: Some critical risks, despite being less severe, have recorded a significant number of accidents.

Exploring the association of all the variables with regards to Accident level and Potential Accident Levels

The analysis aims to identify how various factors relate to Accident_Level and Potential_Accident_Level. This helps in:

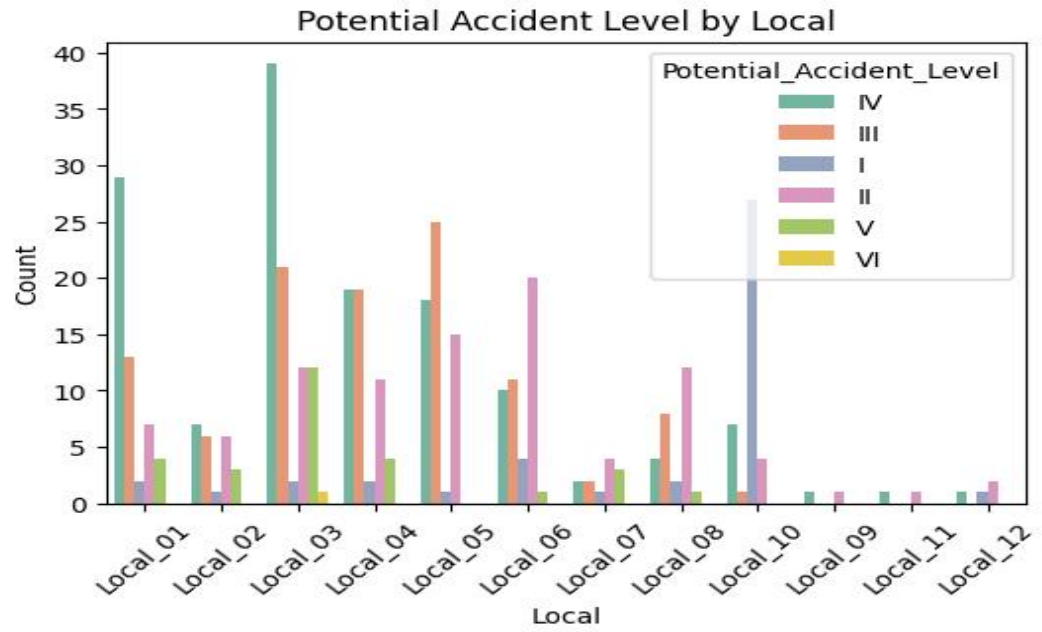
- **Identifying Key Influences:** Determining factors that impact accident severity.
- **Improving Risk Assessment:** Enhancing prediction of high-risk conditions.
- **Targeting Interventions:** Developing focused safety measures based on identified patterns.

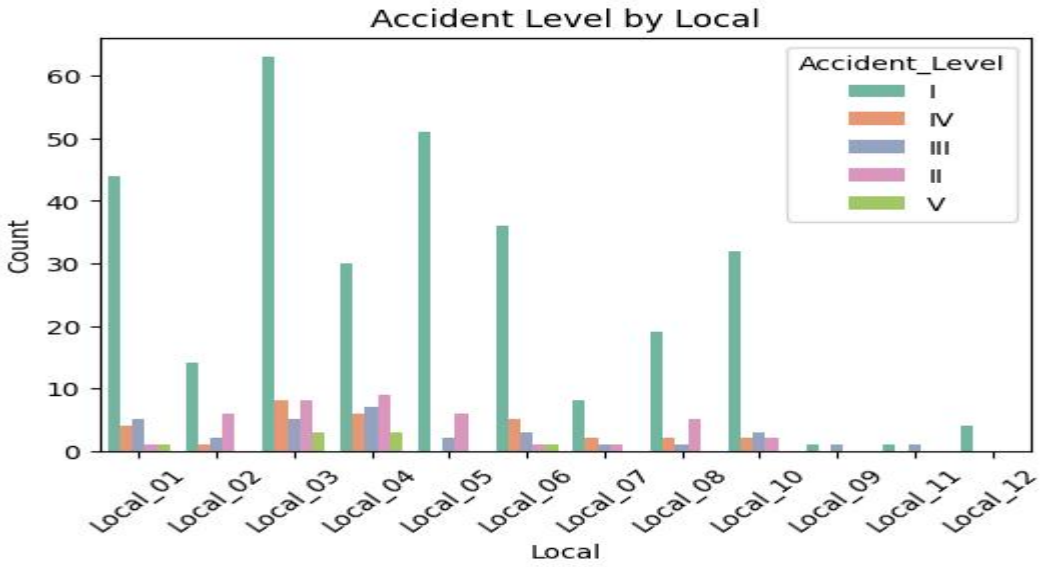




Inferences:

● All countries have significantly higher Severity - I (Lowest severe) counts, whereas if you see potential accident levels, it is leaning more towards Severity Level III and IV. ● This again validates our observation that the potential accident level is a much more valid target.



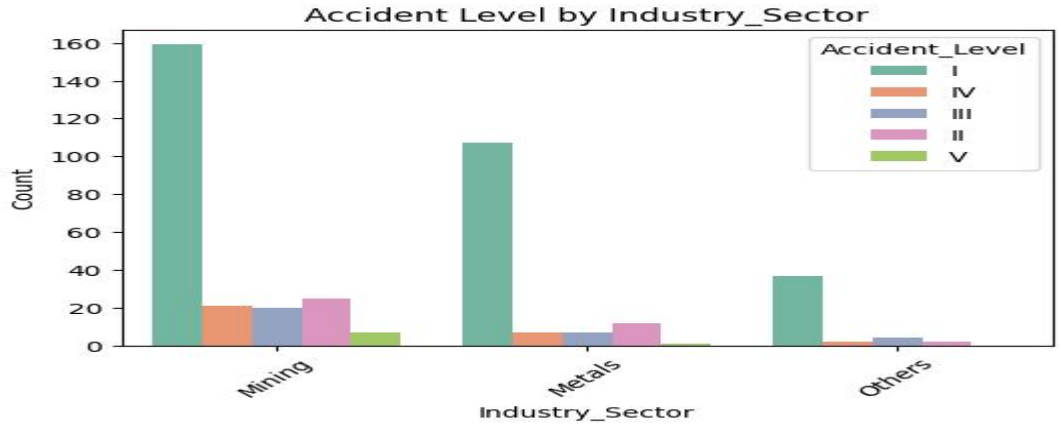


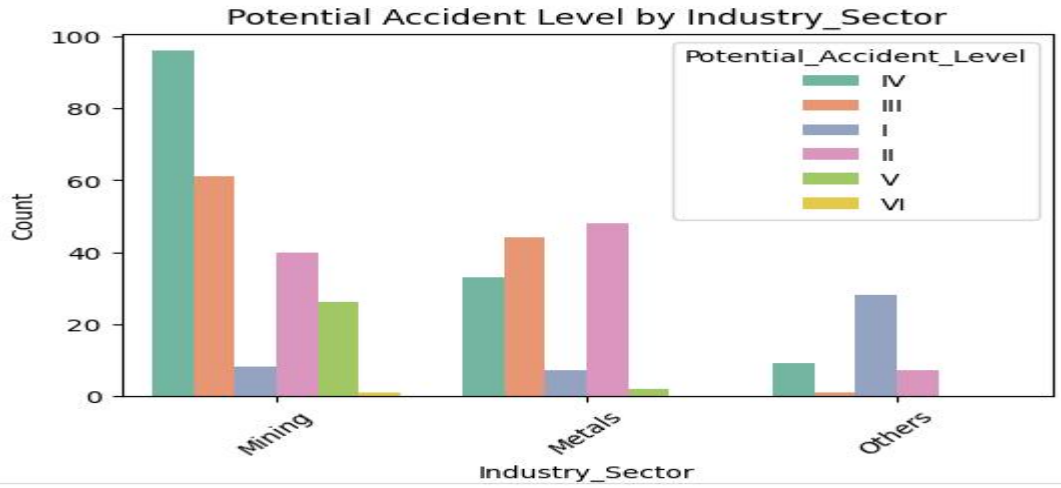
Inferences:

● Similar inferences as above can be observed and derived.

● The local_09 and local_11 has an equal number of recorded III incidents. This is also changed in potential accident levels to IV.

● Local_10 seems to be incurring the maximum least severe incidents. Are they imparting better practices or are their training programs better or are their work conditions safe? We do not have that much data to validate or observe such outlier behaviour.

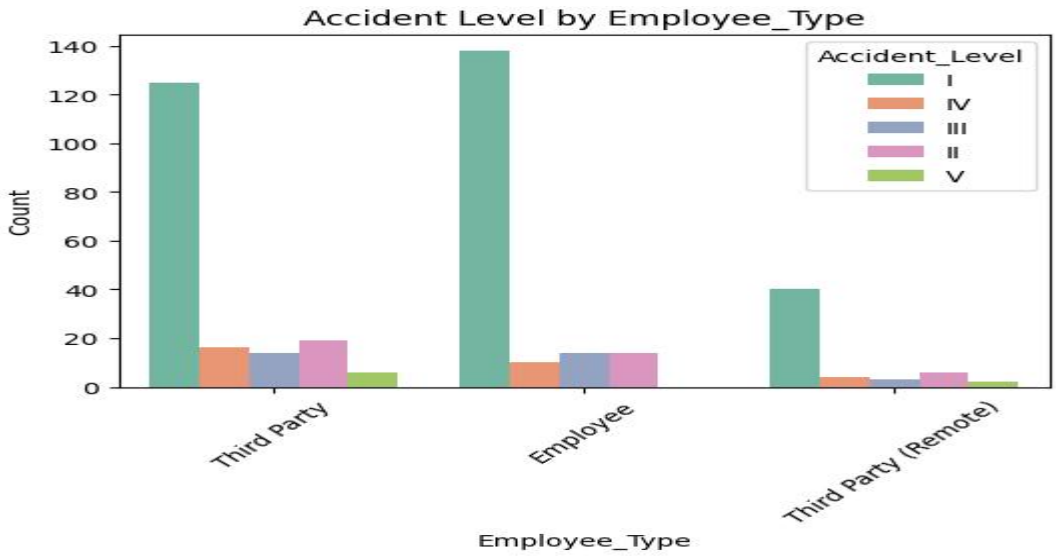


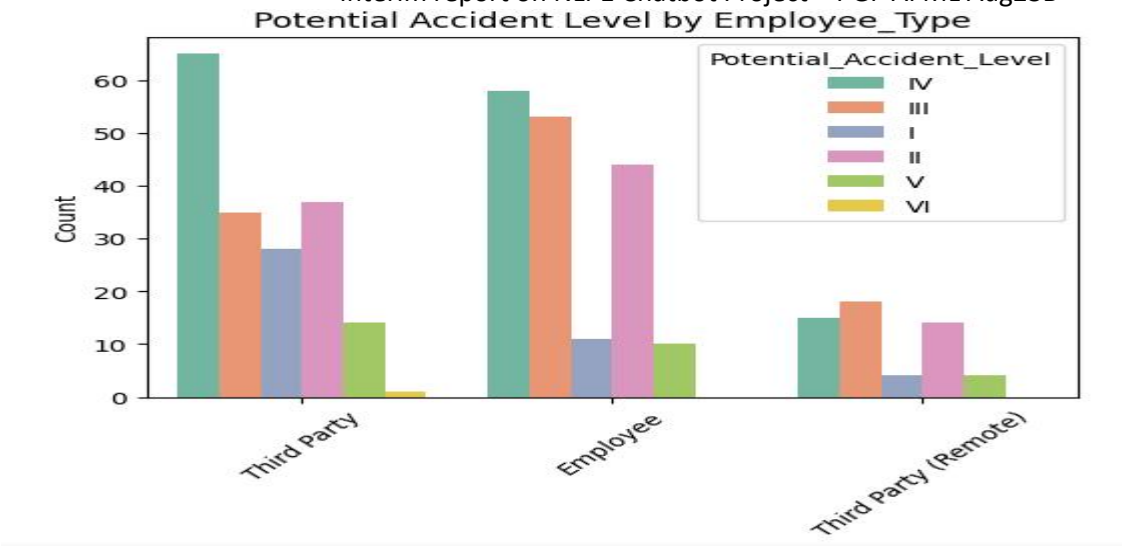


Inferences:

● The potential Accident levels III and IV are higher in the Mining area whereas in Metals, Levels II and III are higher.

● Other Industries do not encounter many higher severe accidents than mining and metals. Mostly (60%) it is all Least severity ones.



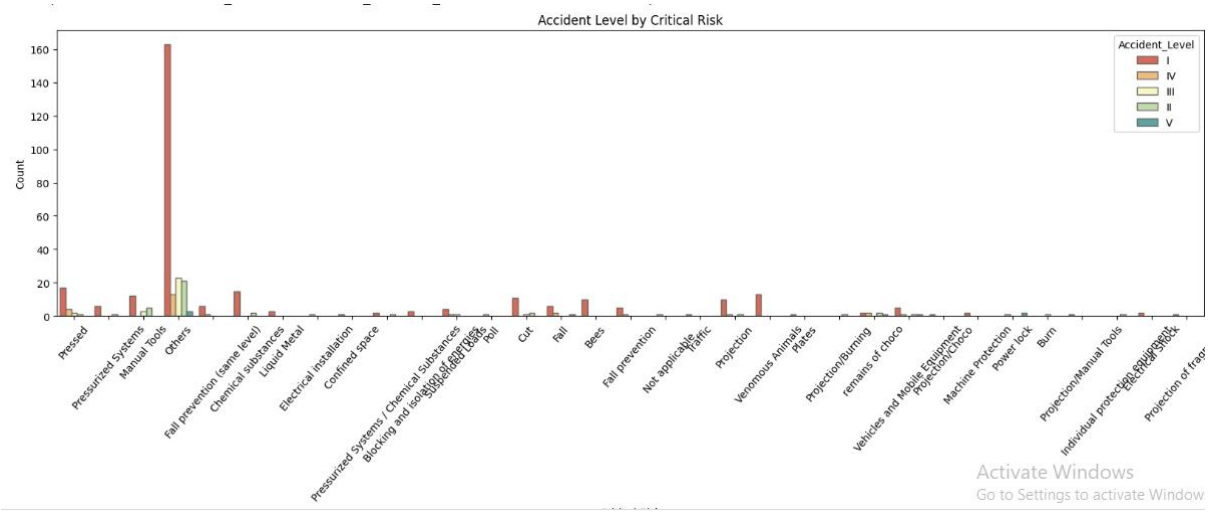


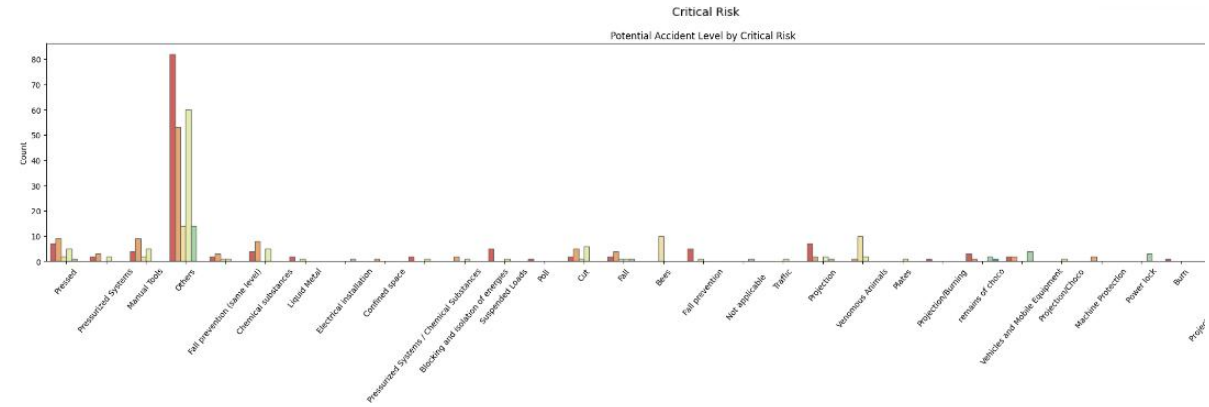
Inferences:

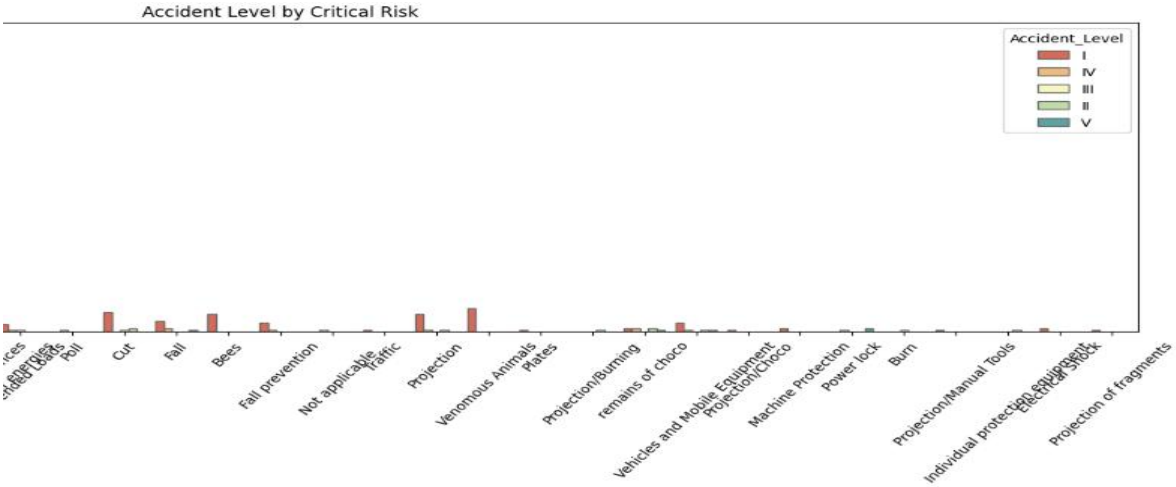
● Potential Accident levels are almost equally distributed for severity level II, III and IV for both third party contractors and employees. The remote third party also encounters more higher severity incidents.

● One thing can be inferred is employee type is important to be recorded but incidents occur across various types. There aren't specific observations

Critical risk factors







Inferences:

● Many not applicable categories have had potentially severe accidents than what was recorded. This may be derived from text processing which we will analyse in the below sections.

● Bees attack typically suggest its a lower grade severity

● Burns category - can be potentially higher severe ones than what was recorded. The severity of burn degree may be derived from text analysis.

● Some Chemical Substances can lead to higher severe accidents in potential accident levels.

● Electrical shock - higher severe incidents. The recorded severity levels are lowest. It certainly requires an update to SOP or better training programs.

● Confined Spaces, Cut or a Fall category is generally recorded as low severe apart from few serious cases whereas in reality, this can be a potentially higher severe one

● The Liquid Metal category has potentially higher severe category ones than the recorded ones.

● The Poll category has been recorded and observed in the right way. Is it that the systems or SOPs are better maintained for this type of incident?

● Powerlock category is typically observed as potentially the highest severe incident type. Even some of the recorded observations validate that. As in the above point, are the systems or SOPs better maintained for this type of incident?

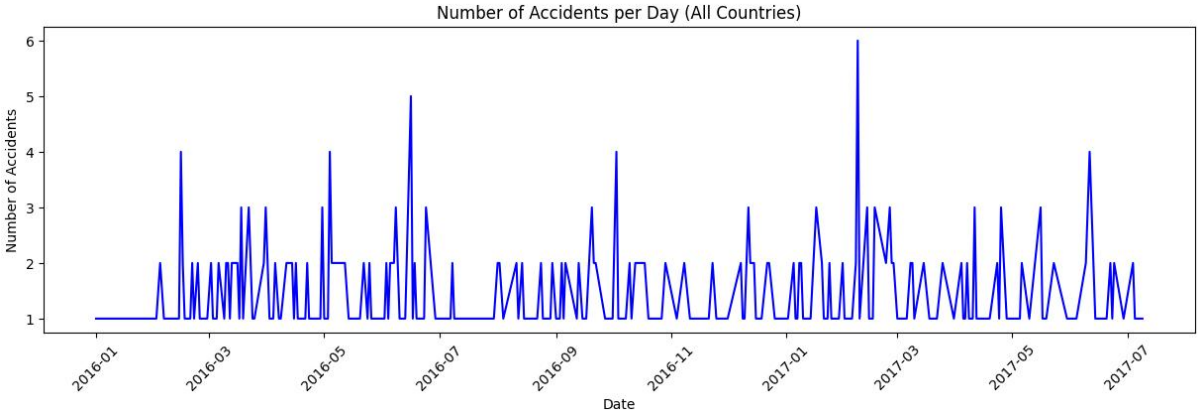
● Electrical installation - again a higher severe category was observed than what was recorded. Even in recorded cases, it was all severity - 4 ones. Maybe a little better training or updating SOP may rectify this error. It is a possible conjecture.

● Venomous animals can have slightly higher severe categories. This need to be observed from text processing

● Suspended loads also are treated as higher severe incidents than what was recorded. Again few changes to SOP and text analysis can lead us to right accident level determination. This critical risk factor parameter is really important. It tells some important observations with regards to either limitations in training program/ SOPs or the analysis at first level recording information was poor.

This can be a potential case to be automated using text analysis and running through our models to have better prediction of accident levels. Once the accident levels are identified closer to the right level, the treatment towards it can be addressed. Lot of data analysis to such levels can potentially save the people working in such an environment.

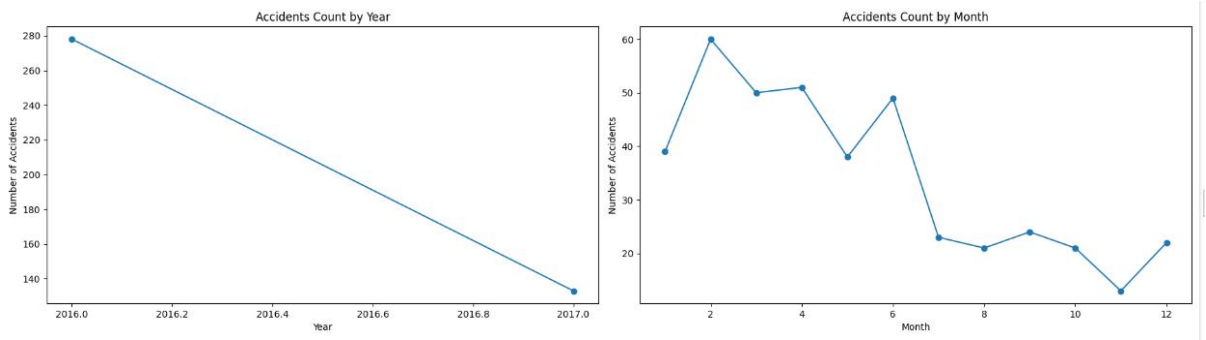
2.5 Temporal Trends Analysis:

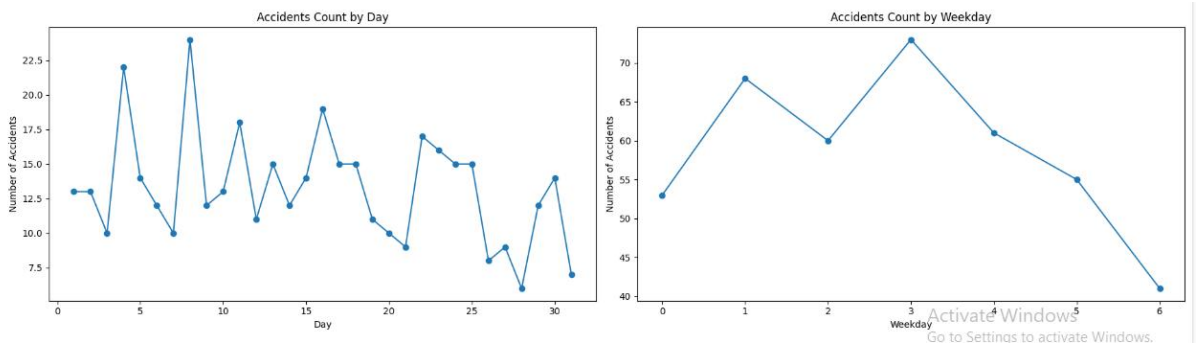


Inferences

1. **Recurring Peaks:** The data shows multiple peaks in accident frequency each year, indicating potential seasonal or cyclical patterns.
2. **Notable Spike:** A significant spike in accidents occurred in February 2017, suggesting a period of heightened risk.
3. **Safety Implications:** The observed trends highlight the need for targeted safety measures during high-risk periods.

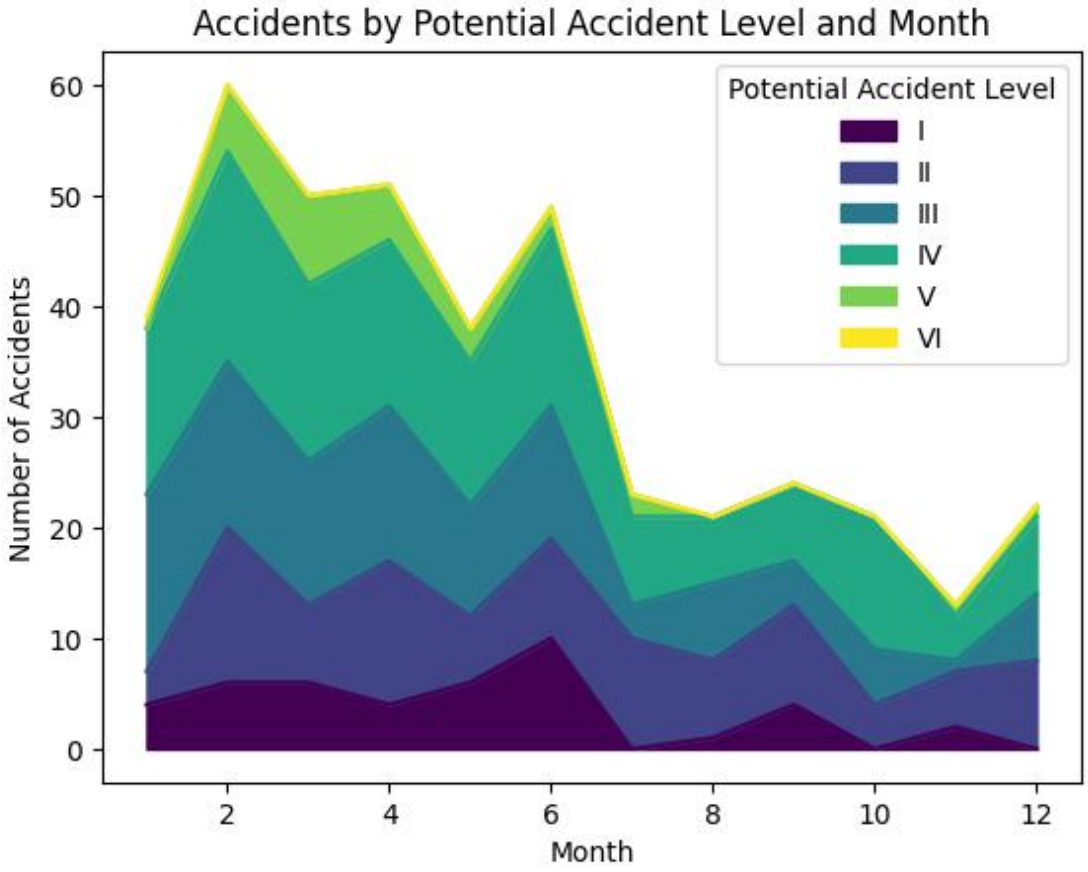
Accidents by Year, Month, Day, and Weekday





Inferences

1. **Yearly Comparison:** There were more accidents in 2016 compared to 2017. However, it's important to note that 2017 data only covers the first seven months.
2. **Seasonal Decline:** Accidents tend to decrease towards the latter part of the year and month, indicating possible seasonal patterns.
3. **Weekly Trend:** The number of accidents increases midweek, peaking around Wednesday, and then declines, suggesting midweek may be a higher-risk period for incidents.



Inferences

- Non-severe accidents (Levels I, II) decrease throughout the year.
- Severe accidents (Levels III, IV) stay consistent over time.
- High-severity accidents (Level V) occur more in the early months.

- Overall, accidents are more frequent at the start of the year and decrease towards the end.

Despite these observations, we found no significant patterns or correlations between accident severity and specific dates, months, or weekdays. The temporal aspects of the data did not show meaningful variations in severity that would impact our primary focus. Therefore, while the date information is useful for recording purposes, it does not significantly contribute to understanding accident severity. As a result, we have decided to exclude the date column from further core analysis, concentrating instead on variables that directly influence accident severity.

2.5 Conclusion of EDA:

● EDA analysis has thrown some important decisions on our parameters along with potential accident levels. This will help us in building our chatbot utility to see which parameter will be significant and should be asked.

● We concluded that Potential accident level is our target variable.

● Like we said above, “Date” does not add much value but can be used for only recording purposes.

● Gender can be important parameter although it has data imbalances.

● Higher potential accident level is occurring more in Country_01. Statistically significant difference between countries.

● Statistically significant difference between various locations or locals.

● Higher potential accident level is more in mining, intermediate in metals sector compared to other sectors. Statistically significant difference between industry sectors.

● All potential accident levels were more in males. Statistically significant difference between gender.

● Having observed many of these and while these parameters are important, these alone cannot predict the potential accident level. They are still categorical and there is no sure way of using such categorical variables to predict the accident levels.

● Hence, we require text processing or analysis on the description to enhance our prediction of potential accident levels. This will be taken in the next section.

● **Why and how EDA is important for our Chatbot utility: Chatbot will be the final output that would be visible to end users.** **All these statistical analysis shows an important aspect for our Chatbot utility. This will aid in designing our chatbot questions or GUI questions to take input on all these parameters that can potentially impact or predict the potential accident levels.**

3.0 Text Data Pre-processing: NLPTechniques

Data preprocessing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a dataset and make it as meaningful data .

One of the common steps in NLP data preprocessing is eliminating stop words, which are words that do not carry much meaning or information, such as "the", "a", "and", "of", and so on.

hence here we have created a set of English stopwords using the NLTK stopwords corpus and removed those from the data.

We have converted all the texts to lowercase for simplicity which helps with consistency of the output.

Same way we have removed all the special characers , Digits , am , pm as these data does not have much meaning to predict the output.

Finally applied stemming which reduces the number of unique words that need to be processed by an algorithm, which can improve its performance. Additionally, it can also make the algorithm run faster and more efficiently.

head results of description field after running text processing

```
0    remov drill rod jumbo mainten supervisor proce...
1    activ sodium sulphid pump pipe uncoupl sulfid ...
2    substat milpo locat level collabor excav work ...
3    approxim nv cx ob personnel begin task unlock ...
4    approxim circumst mechan anthoni group leader ...
Name: cleaned_description, dtype: object
```

4.0 Model Selection and Performance

In this project, we aimed to identify the best-performing model for classifying imbalanced datasets by experimenting with a variety of machine learning algorithms. The models selected for this task included Logistic Regression, Random Forest, Decision Tree, SVM, K-Nearest Neighbors, Gradient Boosting, Naive Bayes, AdaBoost, XGBoost, and LightGBM. We used two text vectorization techniques—TF-IDF and GloVe—to convert the textual data into numerical features. To address class imbalance in the dataset, we employed SMOTE (Synthetic Minority Over-sampling Technique) and applied hyperparameter tuning using RandomizedSearchCV to enhance model performance.

4.1 Model Evaluation based on TF-IDF word tokenizer:

abb	abdomen	...	yields	yolk	young	zaf	zamac	zero	zinc	zinco	zn	zone
0.0	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	...	0.0	0.0	0.0	0.197252	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
0.0	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0

We evaluated each model using a combination of baseline models and models with fine-tuned parameters

Model Performance Evaluation

Combination 1: Using SMOTE with Baseline Models

- **Logistic Regression**: 0.4819
- **Random Forest**: 0.4337
- **SVM**: 0.4819
- **K-Nearest Neighbors**: 0.3614
- **Gradient Boosting**: 0.4217
- **Decision Tree**: 0.4458
- **Naive Bayes**: 0.3855
- **AdaBoost**: 0.2771
- **XGBoost**: 0.4217
- **LightGBM**: 0.4458
- **Extra Trees**: 0.5181

Combination 2: Using SMOTE on Best Parameter Models

- **Logistic Regression**
  Best Parameters: {'solver': 'newton-cg', 'penalty': 'l2', 'max_iter': 200, 'C': 1}
  Accuracy after tuning: 0.5422
- **SVC**
  Best Parameters: {'kernel': 'linear', 'gamma': 1, 'C': 1}
  Accuracy after tuning: 0.5422
- **Random Forest Classifier**
  Best Parameters: {'n_estimators': 100, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_depth': None, 'bootstrap': False}
  Accuracy after tuning: 0.4217

- **Decision Tree Classifier**
  Best Parameters: {'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': None, 'max_depth': 20, 'criterion': 'entropy'}
  Accuracy after tuning: 0.3373
- **K-Nearest Neighbors**
  Best Parameters: {'weights': 'distance', 'n_neighbors': 9, 'metric': 'manhattan'}
  Accuracy after tuning: 0.3494
- **Gradient Boosting**
  Best Parameters: {'n_estimators': 200, 'max_depth': 3, 'learning_rate': 1}
  Accuracy after tuning: 0.4337
- **AdaBoost**
  Best Parameters: {'n_estimators': 100, 'learning_rate': 0.01}
  Accuracy after tuning: 0.3976
- **XGBoost**
  Best Parameters: {'subsample': 0.9, 'n_estimators': 300, 'max_depth': 4, 'learning_rate': 0.2, 'colsample_bytree': 0.9}
  Accuracy after tuning: 0.3855
- **LightGBM**
  Best Parameters: {'num_leaves': 63, 'n_estimators': 300, 'max_depth': 20, 'learning_rate': 0.2, 'boosting_type': 'dart'}
  Accuracy after tuning: 0.4578
- **Extra Trees**
  Best Parameters: {'n_estimators': 200, 'min_samples_split': 10, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': 30, 'bootstrap': True}
  Accuracy after tuning: 0.5663

Combination 3: Without SMOTE with Baseline Models

- **Logistic Regression**: 0.5542
- **Random Forest**: 0.4578
- **SVM**: 0.4096
- **K-Nearest Neighbors**: 0.3735
- **Gradient Boosting**: 0.4819
- **Decision Tree**: 0.4578
- **Naive Bayes**: 0.3855
- **AdaBoost**: 0.3614
- **XGBoost**: 0.3976
- **LightGBM**: 0.4578
- **Extra Trees**: 0.5060

Combination 4: Without SMOTE on Best Parameter Models

- **Logistic Regression**
  Best Parameters: {'solver': 'newton-cg', 'penalty': 'l2', 'max_iter': 200, 'C': 1}
  Accuracy after tuning: 0.5422
- **Random Forest**
  Best Parameters: {'n_estimators': 100, 'min_samples_split': 10, 'min_samples_leaf': 2, 'max_depth': 20, 'bootstrap': False}
  Accuracy after tuning: 0.4217
- **Decision Tree**
  Best Parameters: {'min_samples_split': 5, 'min_samples_leaf': 8, 'max_depth': 10, 'criterion': 'gini'}
  Accuracy after tuning: 0.4337
- **SVC**
  Best Parameters: {'kernel': 'linear', 'gamma': 1, 'C': 1}
  Accuracy after tuning: 0.5301
- **K-Nearest Neighbors**
  Best Parameters: {'weights': 'distance', 'n_neighbors': 11, 'metric': 'euclidean'}
  Accuracy after tuning: 0.4337
- **Gradient Boosting**
  Best Parameters: {'subsample': 0.8, 'n_estimators': 200, 'max_depth': 3, 'learning_rate': 0.1, 'criterion': 'friedman_mse'}
  Accuracy after tuning: 0.4096
- **AdaBoost**
  Best Parameters: {'n_estimators': 50, 'learning_rate': 0.01}
  Accuracy after tuning: 0.3735
- **XGBoost**
  Best Parameters: {'subsample': 1.0, 'n_estimators': 150, 'max_depth': 3, 'learning_rate': 0.2, 'colsample_bytree': 1.0}
  Accuracy after tuning: 0.4578
- **LightGBM**
  Best Parameters: {'subsample': 0.9, 'num_leaves': 63, 'n_estimators': 200, 'max_depth': 20, 'learning_rate': 0.1, 'colsample_bytree': 0.8}
  Accuracy after tuning: 0.4337
- **Extra Trees**
  Best Parameters: {'n_estimators': 100, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_depth': None, 'bootstrap': False}
  Accuracy after tuning: 0.5060

Top 5 accuracies we acquired using TF-IDF vectorization :

Extra Trees (Combination 2: Using SMOTE on Best Parameter Models): 0.5663

Logistic Regression (Combination 3: Without SMOTE with Baseline Models): 0.5542

SVC (Combination 4: Without SMOTE on Best Parameter Models): 0.5301

Logistic Regression (Combination 2: Using SMOTE on Best Parameter Models): 0.5422

SVC (Combination 2: Using SMOTE on Best Parameter Models): 0.5422

4.2 Model Training with GLOVE Embeddings

We also tried the GLOVE embeddings (https://nlp.stanford.edu/projects/glove/) 6B tokens with 50 dimensional tensor.

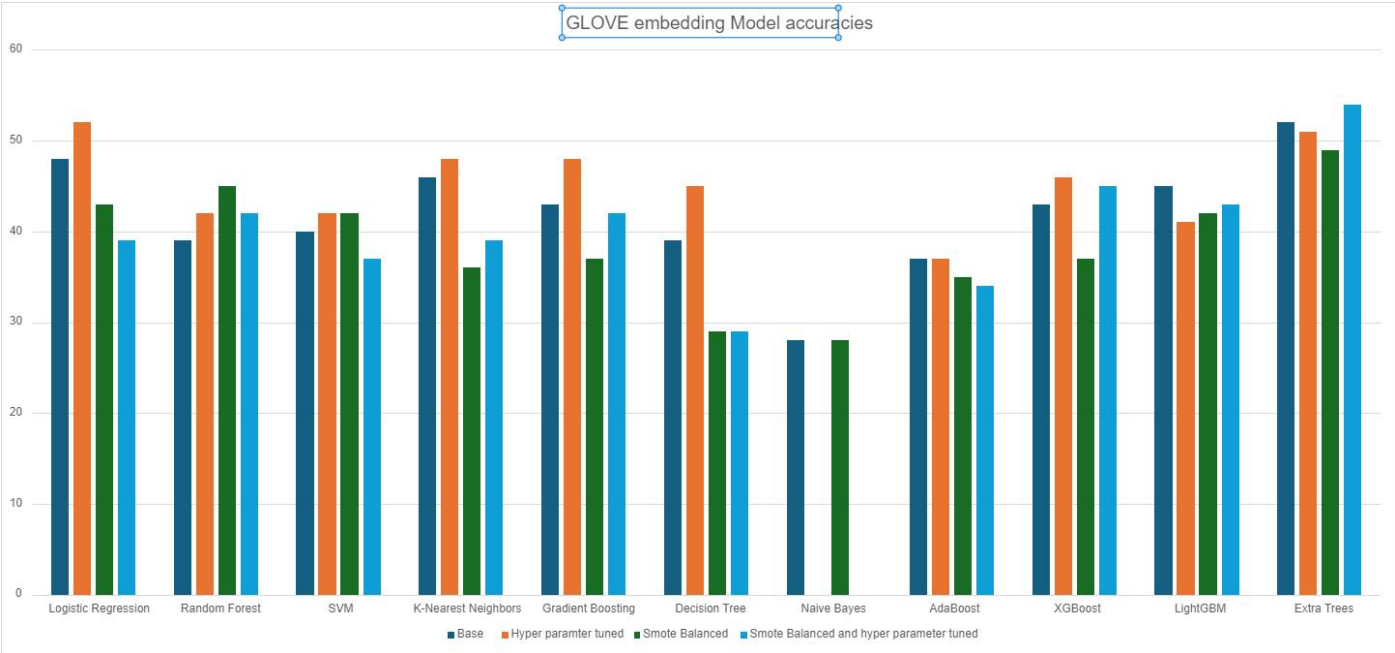
Here is snapshot of features created:

	Feature 0	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	...	Feature 47
0	-0.077208	-0.017825	0.217730	-0.138790	0.082557	-0.012006	-0.059623	0.043636	0.024376	-0.250302	...	0.013827
1	0.271806	0.274290	0.320448	-0.206786	0.053210	0.363607	0.094786	-0.175129	0.162497	0.321926	...	-0.242819
2	0.054804	-0.105934	0.211705	-0.163198	0.265520	0.088124	-0.313619	-0.044364	0.109325	-0.049914	...	-0.372288
3	0.164808	-0.093788	0.281989	0.042357	0.196371	0.096150	-0.033387	0.009267	0.075984	-0.193153	...	0.002792
4	-0.077230	0.133346	0.468211	-0.139532	0.158050	0.321304	0.074443	-0.156500	0.108430	-0.129394	...	-0.389431
...	...	...	...	...	...	...	...	...	...	...	...	...
406	-0.019988	-0.152403	0.372265	-0.147685	0.451526	0.208577	-0.002980	-0.087166	-0.003916	-0.317980	...	-0.335043
407	0.015563	-0.106882	0.122386	-0.107212	0.170532	-0.172028	-0.109121	-0.081224	-0.148228	0.014816	...	-0.473616
408	0.425928	-0.063157	0.123614	0.008376	0.058975	0.026745	-0.219314	0.065336	0.124567	-0.232563	...	-0.166742
409	0.074327	-0.139056	0.263140	-0.273395	0.116468	0.168534	-0.080029	0.247649	-0.115975	-0.014758	...	-0.117455
410	0.166996	-0.054175	0.337276	-0.252623	0.144839	-0.259541	-0.211929	0.215639	0.170826	-0.244574	...	-0.405413

Here are the results of the accuracies obtained when ML models trained with GLOVE features:

	Base	Hyper parameter tuned	Smote Balanced	Smote Balanced and hyper parameter tuned
Logistic Regression	48	52	43	39
Random Forest	39	42	45	42
SVM	40	42	42	37
K-Nearest Neighbors	46	48	36	39
Gradient Boosting	43	48	37	42
Decision Tree	39	45	29	29
Naive Bayes	28		28	
AdaBoost	37	37	35	34
XGBoost	43	46	37	45

LightGBM	45	41	42	43
Extra Trees	52	51	49	54



4.3 Conclusion:

From our analysis we can say we have received best model accuracy with Extra Tree Classifier which have us the max accuracy of 56% with TF-IDF tokenizer and 54 % accuracy with Glove vector embeddings.