**AIML**

**CAPSTONE PROJECT - INTERIM REPORT**

**NLP CHATBOT For Industry Safety**

**Team Mentor**

Rohit Gupta

**Team Members**

Sarita Ghadge

Vikram Rai

Rohan Bhattacharya

Meetu Chandra

# Problem Statement

Major Industries in Brazil are facing multiple workplace hazards leading to multiple accidents, some even leading to death. The aim of this project is to build an AI powered Chatbot for employees and stakeholders to understand the various reason for the accidents. This can help the employees take the necessary precautions and for the stakeholders to identify the root cause for the accidents and put a fix in place.

The Chatbot will be powered by an AI model that has been trained to classify the accident, given a summary of the various scenarios when the accident happened in past.

# Approach

An NLP based AI model will be trained on industry data from **3 countries** and **12 plants.** The data is from industries operating majority in Metal and mining. Machine learning models using different strategies will be developed and compared to identify the best performing model.

The below models will be used and compared.

**SVM classifier** - Useful for tasks with high-dimensional data and clear margins between classes.

**Random Forests** - Work well for structured data representations like TF-IDF or word embeddings.

**Recurrent Neural Networks (RNNs)/LSTM** - Good for sequential data processing.

**Transformers (e.g., BERT, DistilBERT)** - State-of-the-art for modern NLP classification tasks.

# Data Overview

# We analysed data from one of the largest industries in Brazil, which consists of 425 accident records from January 2016 to September 2017 across 3 countries and 12 locations. The dataset includes key information provided in the table below

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Column Name** | **Description about the column** |
| 1 | Unnamed | Index Column  **Data type**: Integer |
| 2 | Data | Date of accident  **Data type**: datetime  **Range**: January 2016 to July 2017 |
| 3 | Countries | which country the accident occurred (anonymised)  **Data Type:** Object  **Unique values**: Country\_01 Country\_02 Country\_03 |
| 4 | Local | The city where the manufacturing plant is located (anonymised)  **Data Type**: Object  **Unique values**: Local\_01 to Local\_12 |
| 5 | Industry sector | Which sector the plant belongs to?  **Data Type**: Object  **Unique values**: Mining, Metas, Others |
| 6 | Accident level | From I to VI, it registers how severe was the accident (I means not severe but VI means very severe)  **Data Type**: Object  **Unique values**: I,II, III, IV, V |
| 7 | Potential Accident Level | Depending on the Accident Level, the database also registers how severe the accident could have been (due to other factors involved in the accident)  **Data Type**: Object  **Unique values**: I,II, III, IV, V, VI |
| 8 | Gender | If the person is male of female  **Data Type**: Object  **Unique values**: Male, Female |
| 9 | Employee or Third Party | if the injured person is an employee or a third party  **Data Type**: Object  **Unique values**: Third Party Employee Third Party (Remote) |
| 10 | Critical Risk | Some description of the risk involved in the accident  **Data Type**: Object  **Unique values**: There are number of unique values like – could be a short description for the accident description. |
| 11 | Description | Detailed description of how the accident happened.  **Data Type**: Object (textual data) |

The below steps were performed

# Exploratory Data Analysis (EDA)

## Initial Analysis:

There were 425 records with 10 columns. None of the columns had missing values. All the columns except for Date and description were categorical.

## Renaming columns and removal from initial analysis

Few of the column names had to be renamed to as per the data they were representing: -

* “Data” was renamed to Date
* “Genre” was renamed to Gender.
* “Unnamed” which was an index columns was dropped.

## Check for Missing values

There were no missing values in the dataset

## Check for Duplicates

* There were 7 records that were duplicate and the duplicate records were dropped. That left us with 418 records.
* Out of the remaining 418 records, 7 accident descriptions are repeated in the data.
* These are accidents which happened at the same time where a group was invovled and there are different records for each person.
* As this corresponds to different records they were not removed

## Unique values of the categorical columns

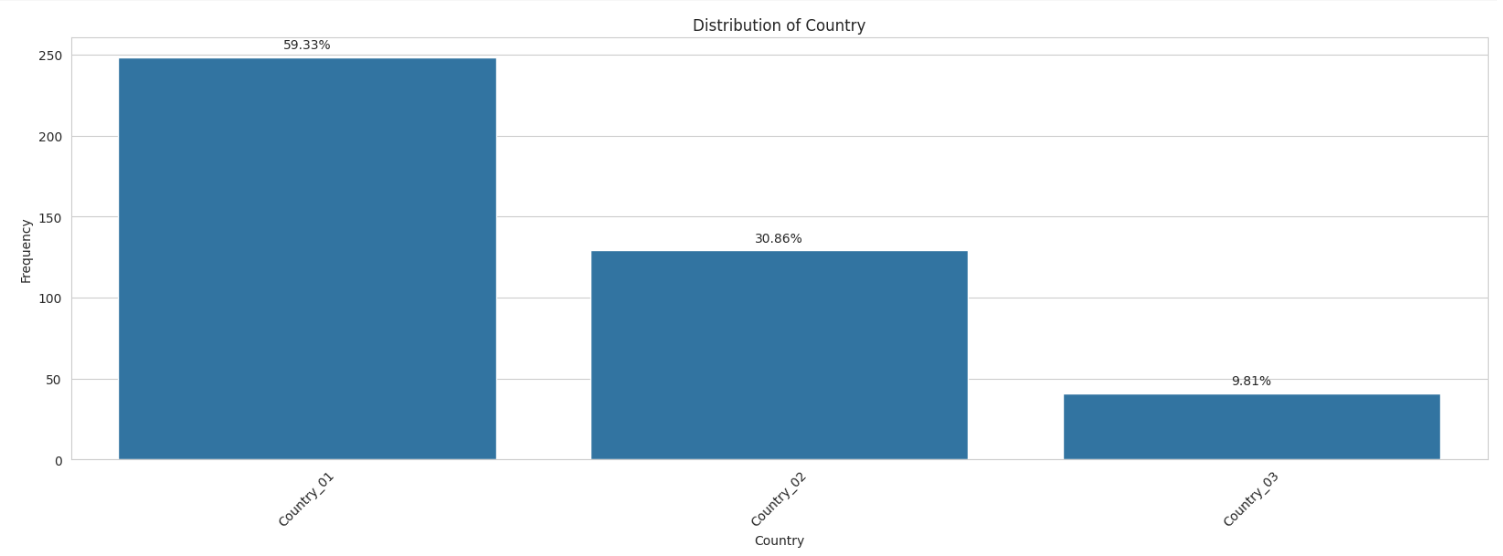
Unique values for the categorical columns were identified (The values are available in the above table).

## Univariate Analysis

Count plots were used to understand the spread of data across each feature.

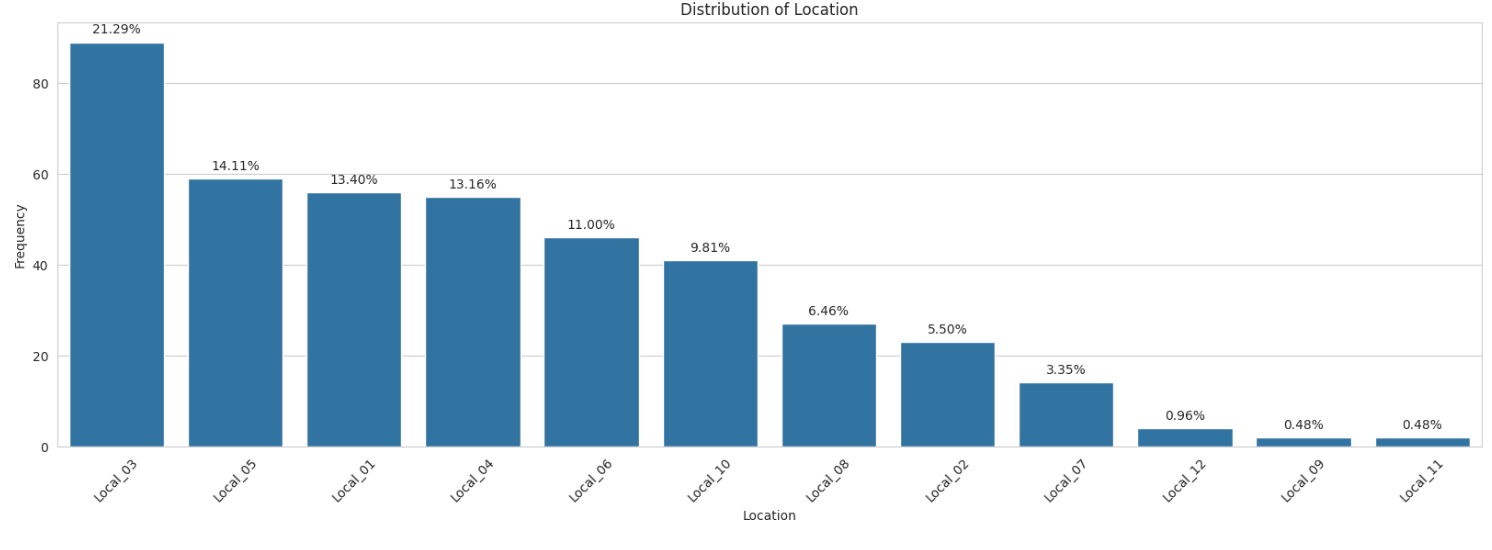
### Country

There is data for 3 countries, 60% of the data is for Country\_01, 31% for countrty\_02 and only 10% for country\_03. It is possible country\_01 is more prone to accidents.



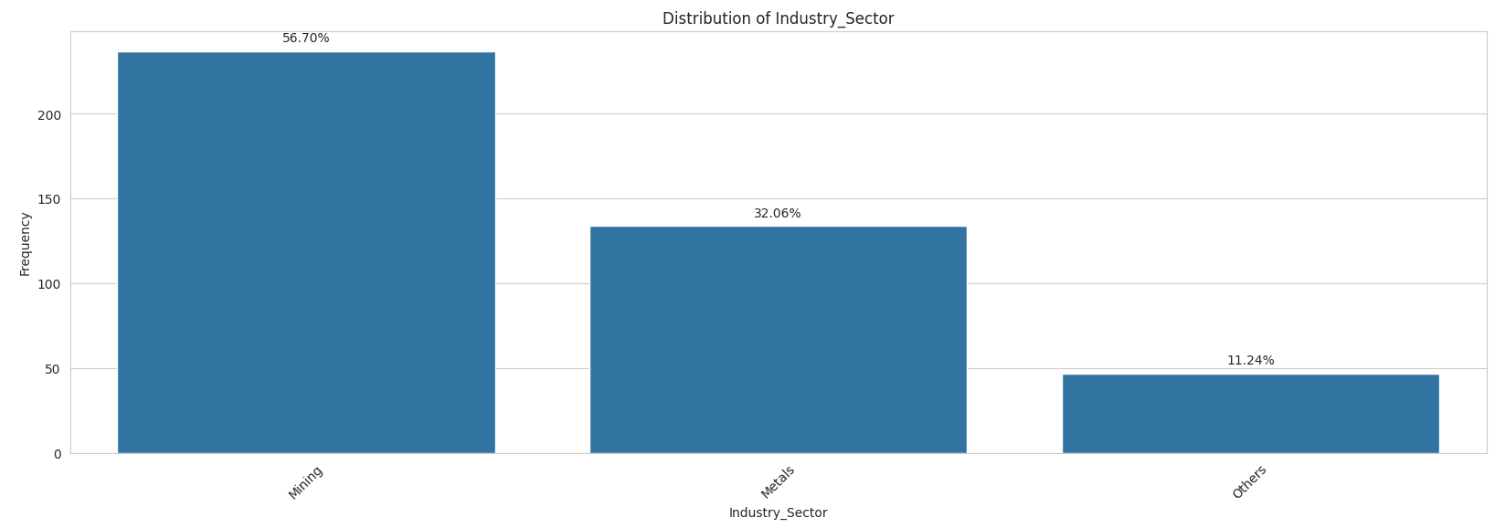
### Location

There is data for 12 locations across 3 countries. Local\_03 has seen the maximum number of accidents, which is around 20% of all the accident cases recorded.



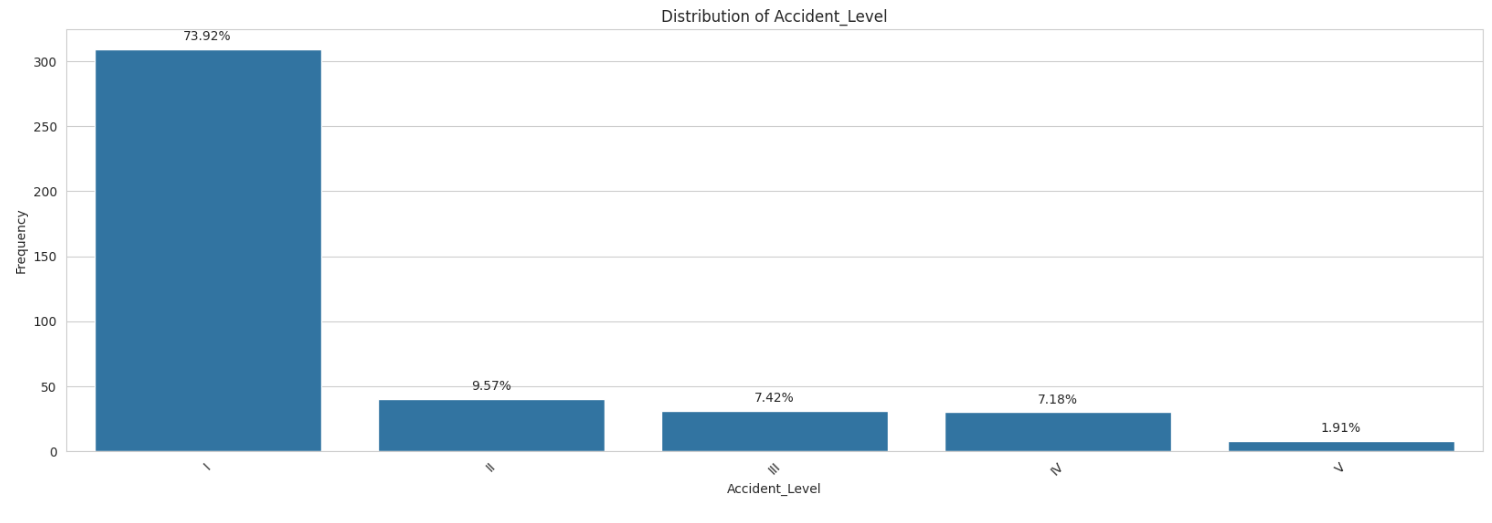
### Industry Sector

Mining sector has the most accident cases than any other sector. Thus, we can say that jobs in the mining industry sector are riskier than metal or any other sector.



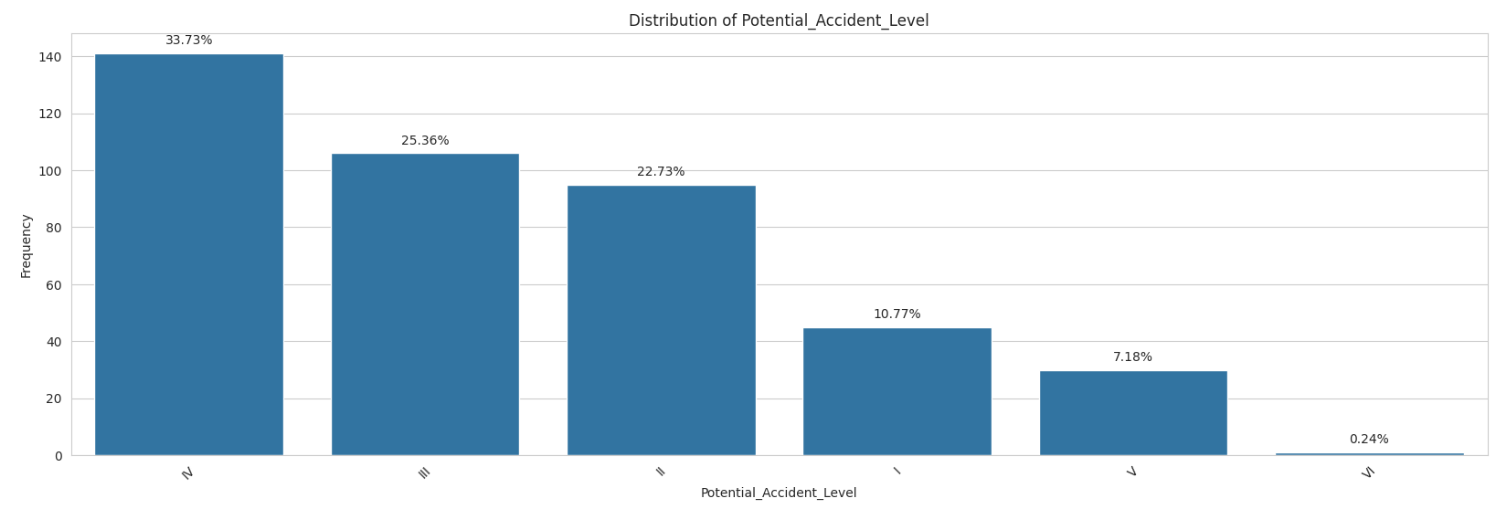
### Accident Levels

Accident levels are mostly of Level 1 with 74% of the data, followed by 9.57% of level 2 and ~7% for levels 3 and 4 and ~2% for level 5. There have been no accidents of level 6 which is the highest level.



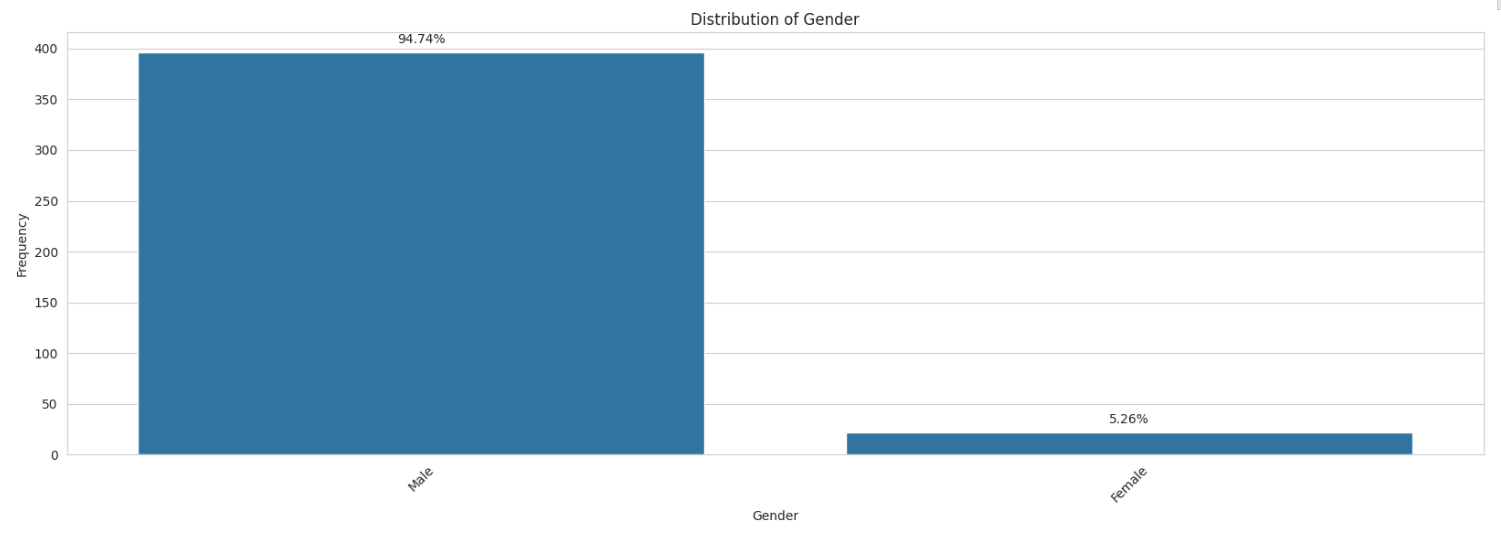
### Potential Accident Level

Potential accident level indicates how severe the accident would have been due to other factors involved in the accidents. As per the graph, level IV has the highest count, which corresponds to moderate severity of accidents, followed by 25.3% of level 3. Also 0.24% chances of the most severe level 6.



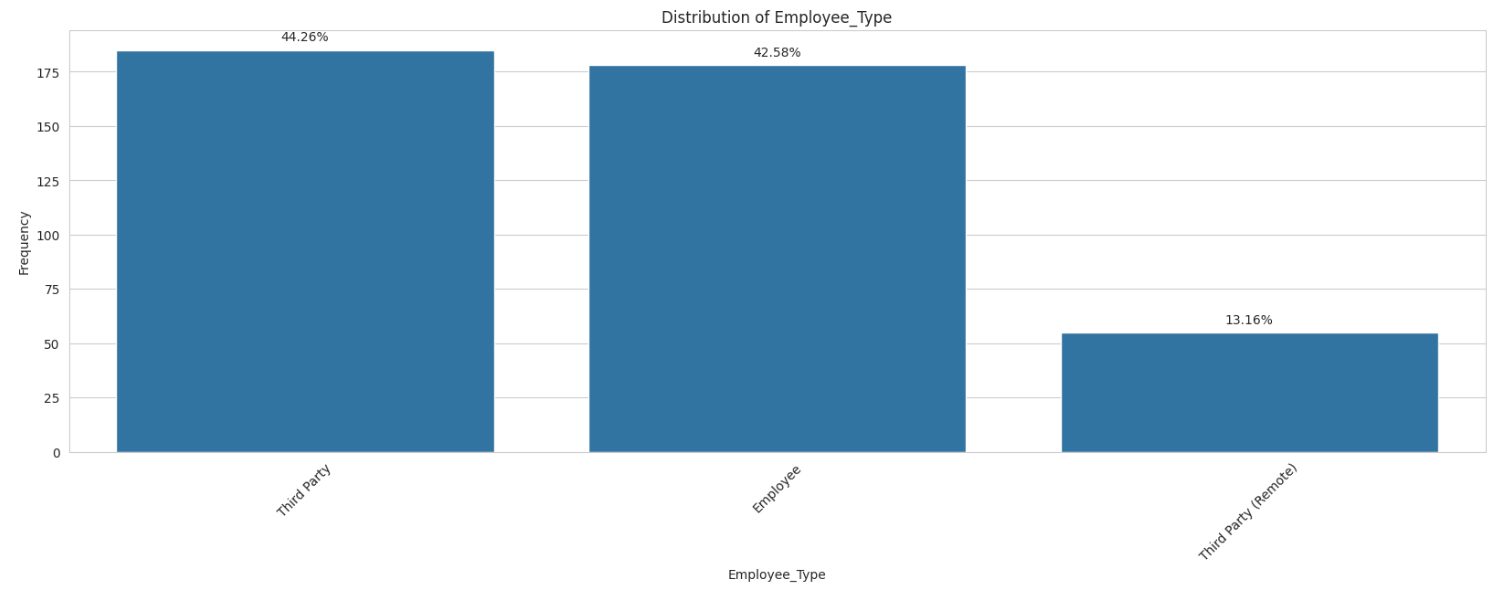
### Gender

Dataset is more biased towards male employees; this is possible because Mining and Metal industries are more male dominant.



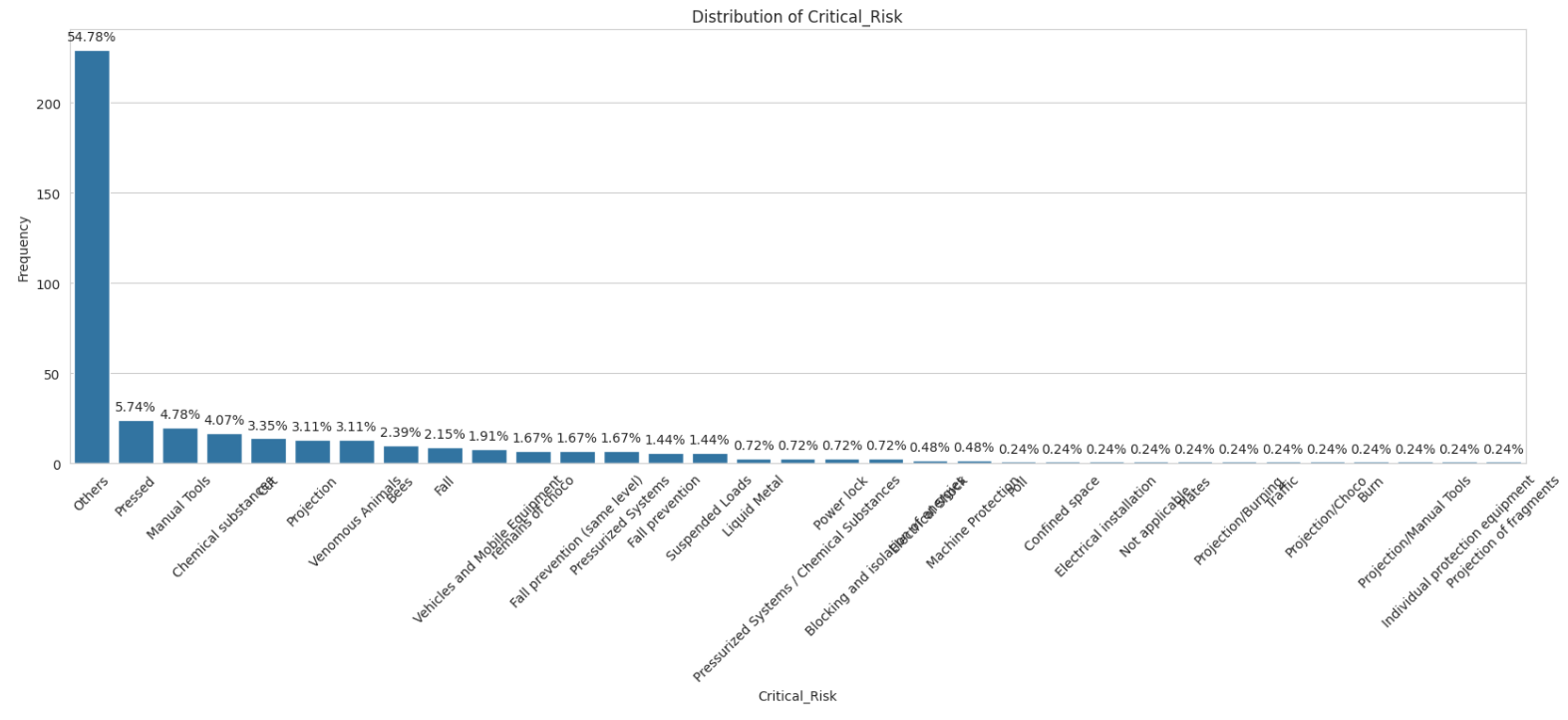
### Employee Type

Total number of internal employees and Third-Party employees is more or less the same. But, we can also see that Third party remote employees are comparatively less in number.



### Critical Risks

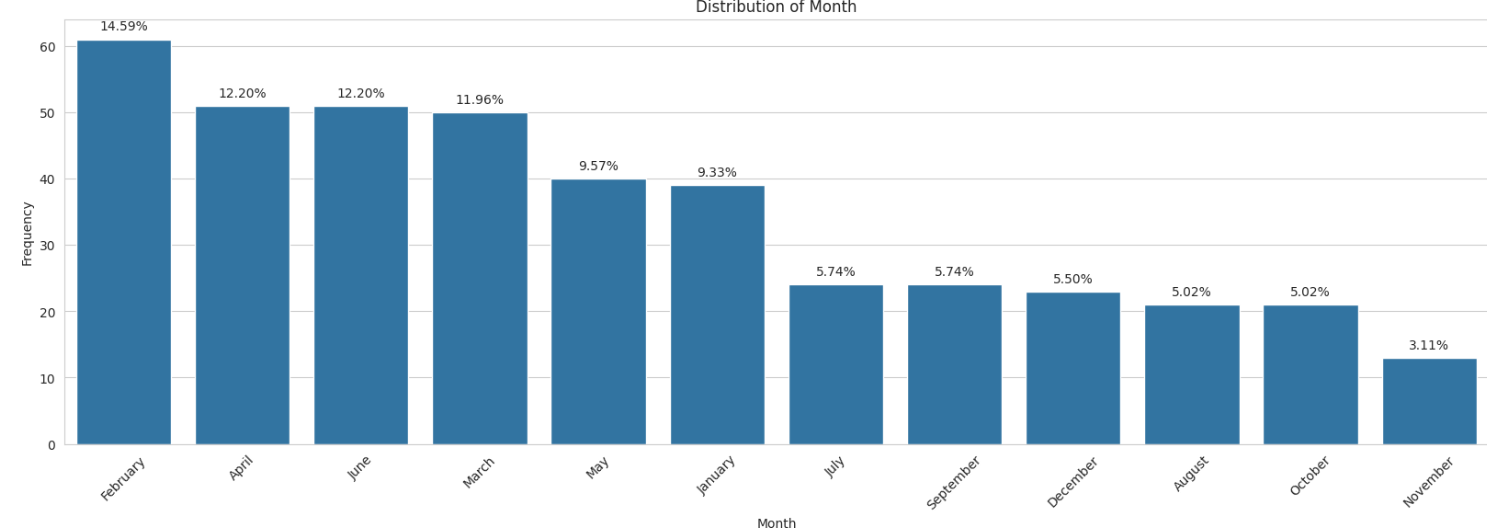
Most of the Critical Risks are classified as 'Others'. It holds around 55% of the total Critical Risks. It is followed by Pressed, Manual tools, Chemical substances, etc.



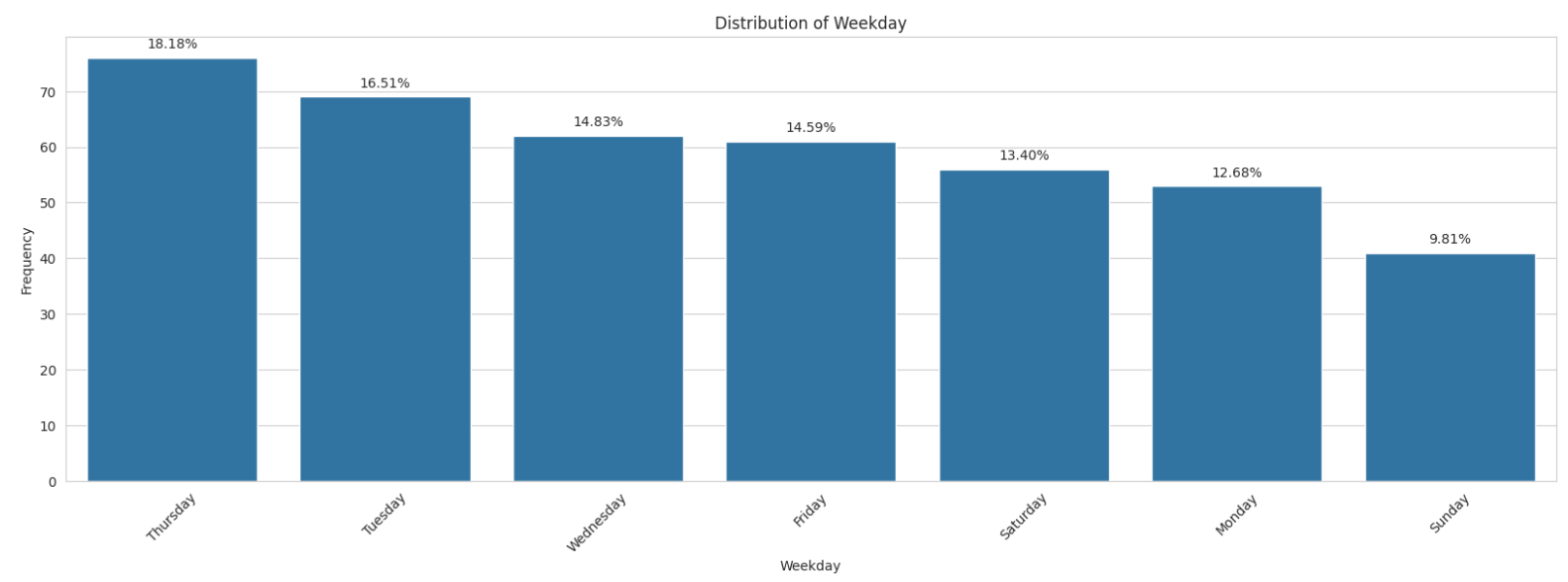
### Date

The Date column was split into Year, Month, Date, Day of the week to check if there is any pattern.

1. Data was available of 2 years 2016 and 2017.
2. Year and Day column did not seem to any relevance for the analysis and were dropped.
3. Accidents were more frequent in the first half of the year with maximum accidents in February at 14.59% followed April, June and March. November recorded the least number of potential accidents.



1. Thursdays are more prone to accidents followed by Tuesday. Sunday has the least



Bivariate analysis was then performed to see the patterns and spread of each feature with the Potential accident levels

# 

# 

# Solution

1. Summary of the Approach to EDA and Pre-processing Include any insightful visualization you have teased out of the data. If youve identified particularly meaningful features, interactions or summary data, share them and explain what you noticed. Visual displays are powerful when used well, so think carefully about what information the display conveys.
2. Data is skewed
3. Potential accident level – why?
4. Dropped junk columns
5. Divided week-day and – more on weekdays /start of the year
6. Deciding Models and Model Building Based on the nature of the problem, decide what algorithms will be suitable and why? Experiment with different algorithms and get the performance of each algorithm.
7. How to improve your model performance? What are the approaches you can take to improve your model? Can you do some feature selection, data manipulation and model improvements. Provide your code and as much as visualizations you can share to describe what you have done so far.

augmentation – description and minority features

identified word cloud – plan to use in data upsampling

hyper parameter tuning/ we used 3 levles – word2Vec/glove/transformer

scope for more

and will further be finetune