Industrial safety NLP based Chatbot

Interim report group 4

Team mentor:

Sumit Kumar

Team members:

1. Ajith Chandran
2. Devina
3. Dev
4. Latha
5. Soumita

**Table of Contents**

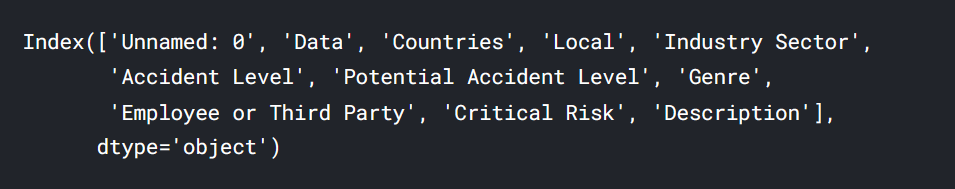
1. Overview
   * Project Detail
2. Column’s description
3. Import libraries
4. Load the dataset
5. Pre-processing
   * NLP Pre-processing
6. [EDA](https://www.kaggle.com/mysh7989/industrial-accident-causal-analysis#5.-EDA)
   * Univariate Analysis
   * Multivariate Analysis
   * NLP Analysis
7. Modeling
   * Feature Engineering
   * Case1 : Accident Level
   * Case2 : Potential Accident Level
8. Conclusion
   * Task Submission
9. References

# 1. Overview

## Project Detail

In [this dataset](https://www.kaggle.com/ihmstefanini/industrial-safety-and-health-analytics-database), the information about accidents in 12 manufacturing plants in 3 countries are given. We need to use this dataset to understand why accidents occur, and discover clues to reduce tragedy and accidents.

## Import Dataset and data columns



Dataset columns are below:

* **Data** : timestamp or time/date information
* **Countries** : which country the accident occurred (**anonymized**)
* **Local** : the city where the manufacturing plant is located (**anonymized**)
* **Industry sector** : which sector the plant belongs to
* **Accident level** : from I to VI, it registers how severe was the accident (I means not severe but VI means very severe)
* **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)
* **Genre** : if the person is male of female
* **Employee or Third Party** : if the injured person is an employee or a third party
* **Critical Risk** : some description of the risk involved in the accident
* **Description** : Detailed description of how the accident happened

The dataset is in csv format. Basic exploration of the data is as below

**Shape of the dataset**

****

### **Rename columns**

The field Unnamed: 0", is dropped and columns are renamed following

* ‘Data':'Date',
* 'Countries':'Country',
* 'Genre':'Gender',
* 'Employee or Third Party':'Employee type'

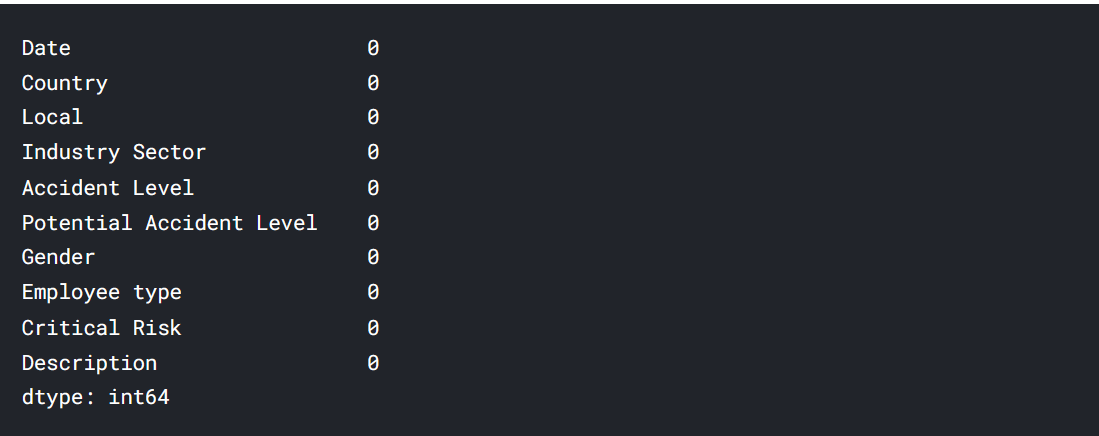
### **Display datatypes**

### 

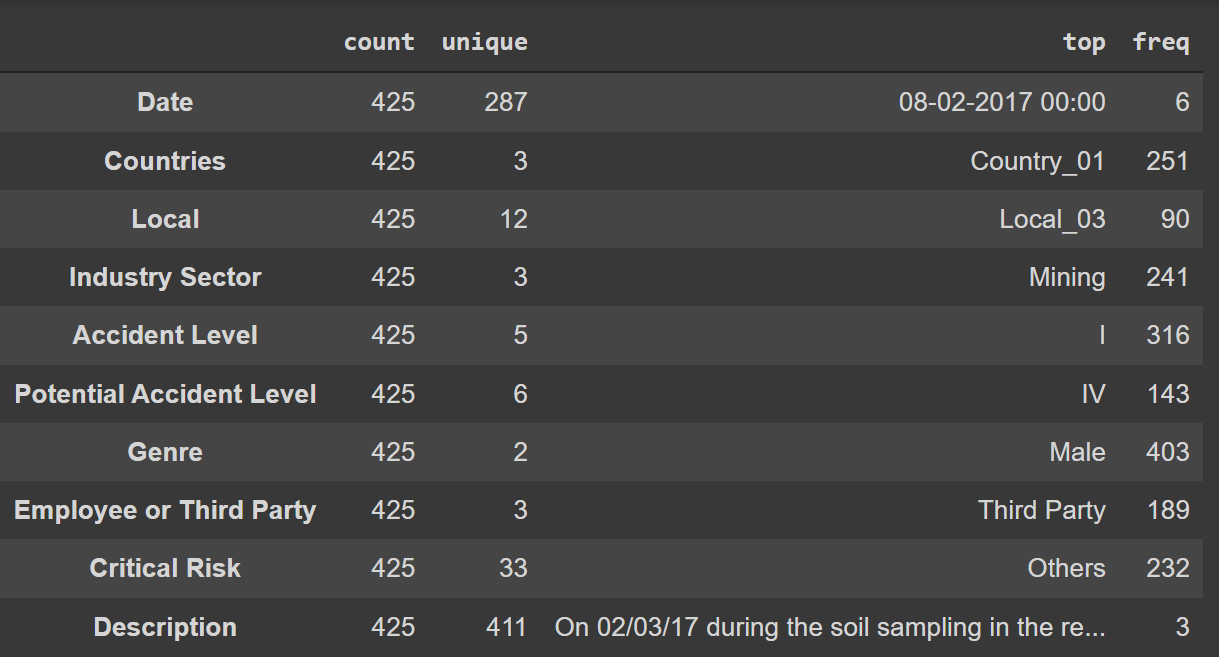
### Here, we can see that all the columns of the dataset are of "object" datatype. Coming to the type of data present in each column, we can see that there is a column "Date", which means it holds time series data. All other columns except "Description" are of categorical datatype.

### **Check for null values**

There are no null values present



Dataset Description



From the above table, we can infer the below:

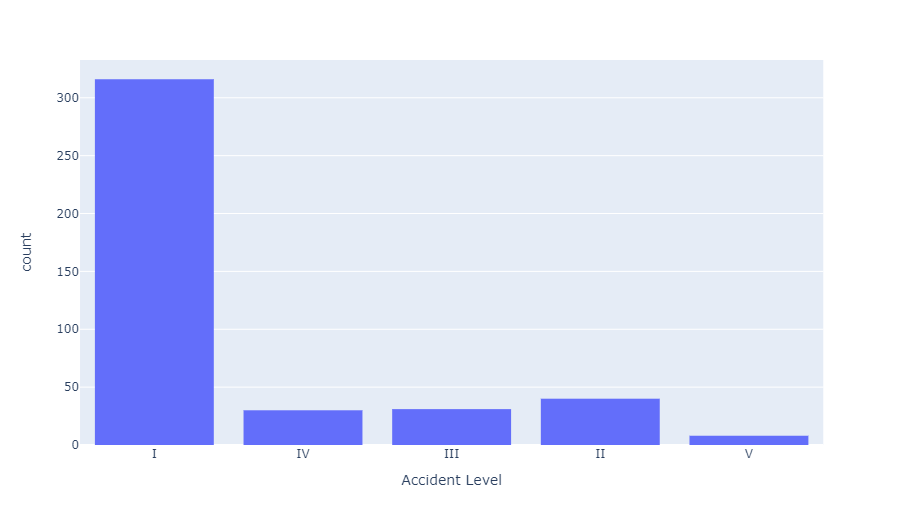
1. This dataset contains accident data of 3 countries, out of which Country1 has the most number of accidents.
2. The data is collected from 3 types of industry sectors.Local\_3 has the most number of accidents.
3. There are 5 major accident levels in which this dataset has been classified.316 accidents are of accident level 1, making it the most frequent accident type. This also means that the data is not distributed evenly.
4. The data is a consolidation of accidents faced by employees as well as third party vendors and others. Third party employees have faced the most number of accidents according to this dataset.
5. 403 male employees have been reported to have accidents, which mean the distribution of data in this case is also not evenly balanced.
6. 33 different types of critical risks have been identified in the dataset.

The Categorical Variables that can be encoded to Numerical Values from the dataset

1. Local
2. Accident Level
3. Potential Accident Level

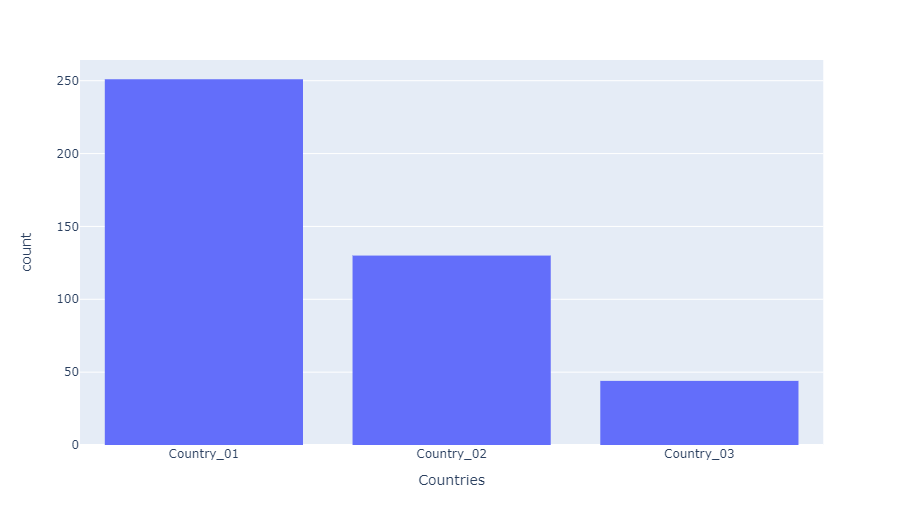
Univariate Analysis:

Checking the distribution of data based on accident levels



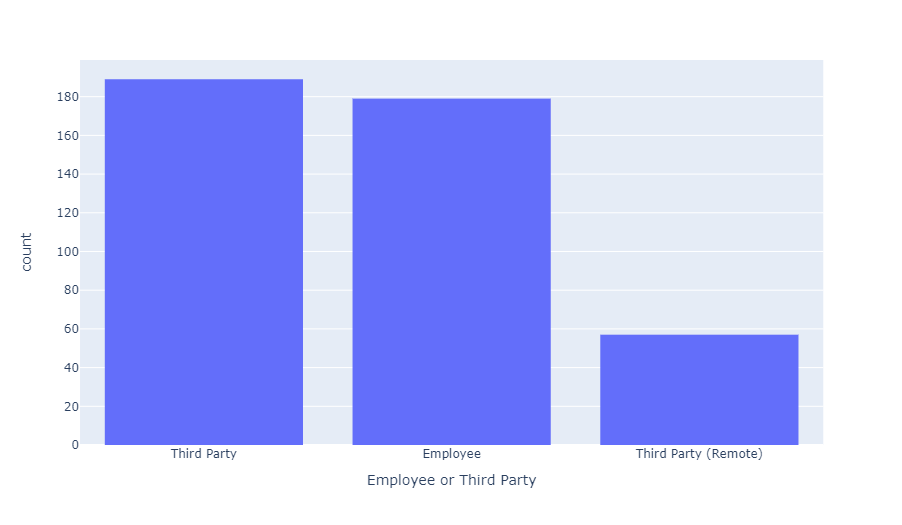
 The distribution of Accident Levels is highly imbalanced in the dataset

**Distribution of the data based on country wise**

****

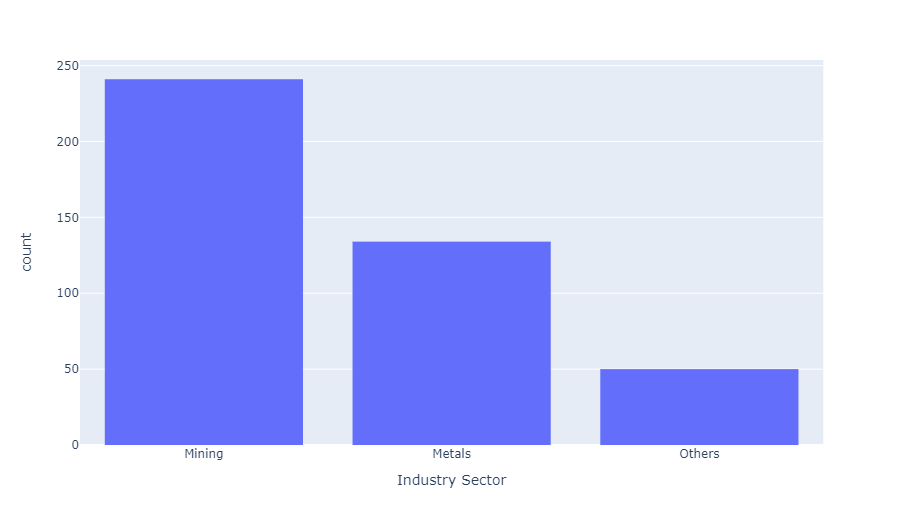
"Country\_01" has the most number of accident cases

**Distribution of accidents by Employee Types:**

****

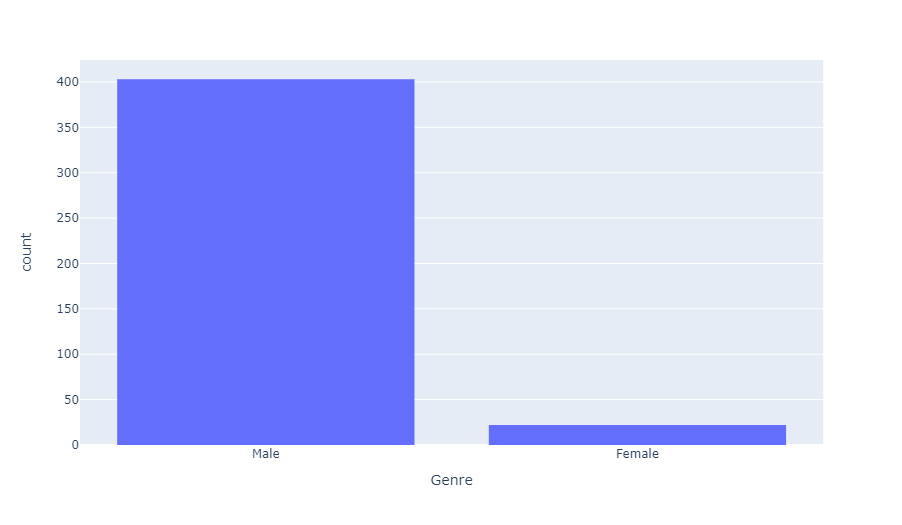
From the graph it is very clear that accidents have happened in almost equal proportions among permanent employees or third party contractors, with thrid party contractors a bit on the higher side.

**Distribution of accidents as per industry sector.**

****

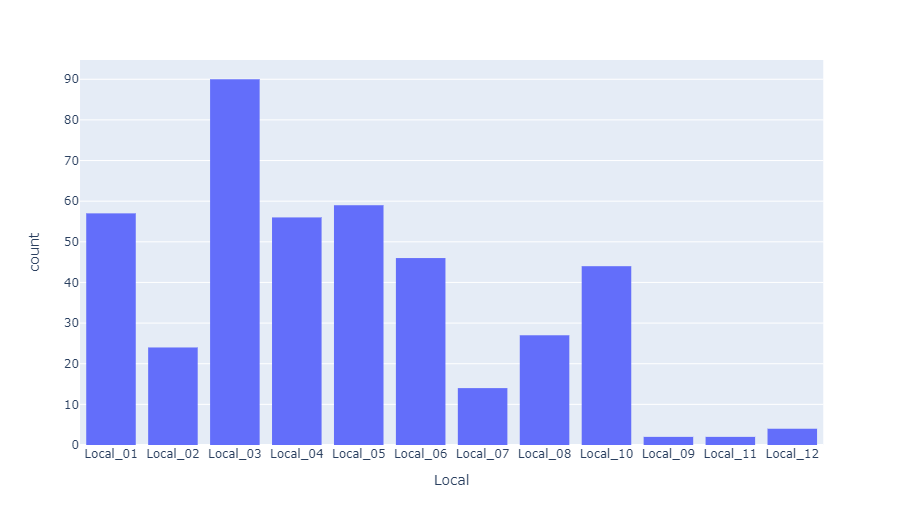
Majority of the accidents have happened in the mining sector, followed by metal industry and other type of industries.

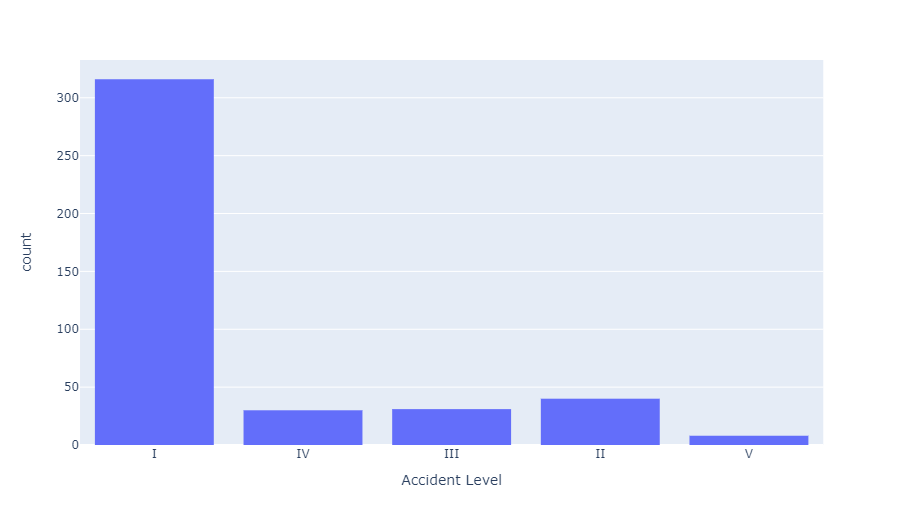
**Distribution of accidents as per Gender**

****

The distribution of accidents is imbalanced when checked by "Genre". The count of accidents in males is way higher than that in females.

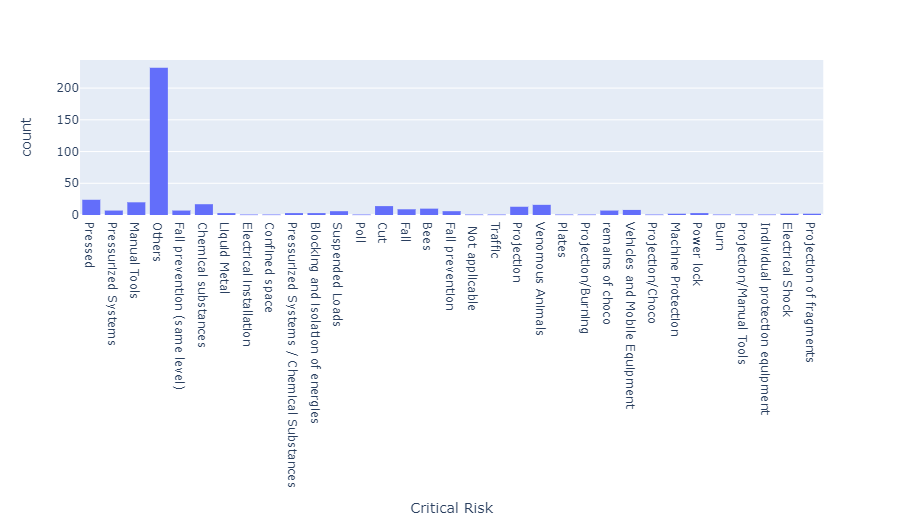
**Distribution by Locals.**

****

****

**Observation from the above graph**

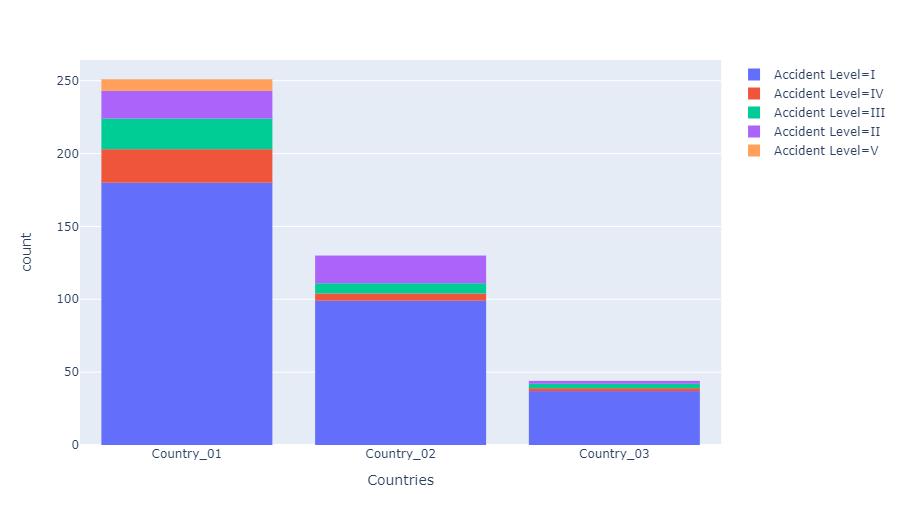
 Most of the people have met with accident having level 1

****

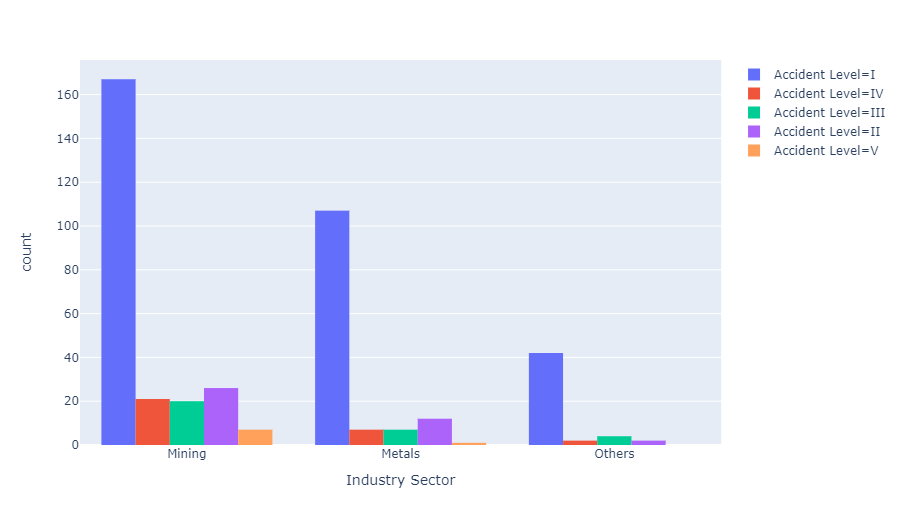
We can see from the graph that the Critical risk category "Others" have the most number of accidents. This means we are not clear about the exact risk factor associated with accidents in this dataset.

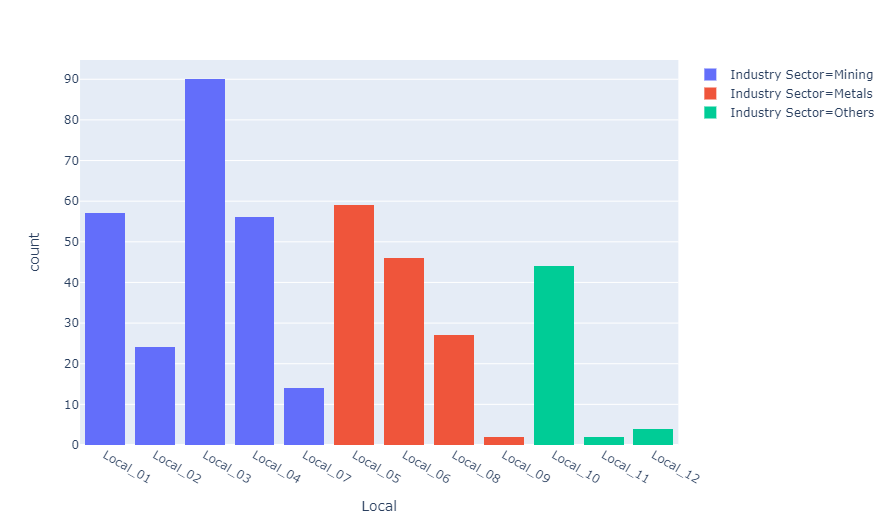
**Bivariate Analysis**

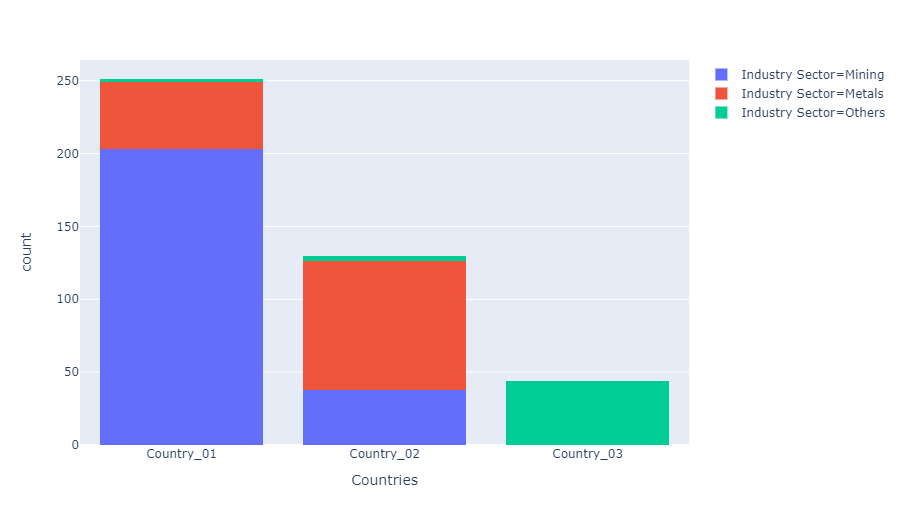
Distribution of different accident levels occurred per country

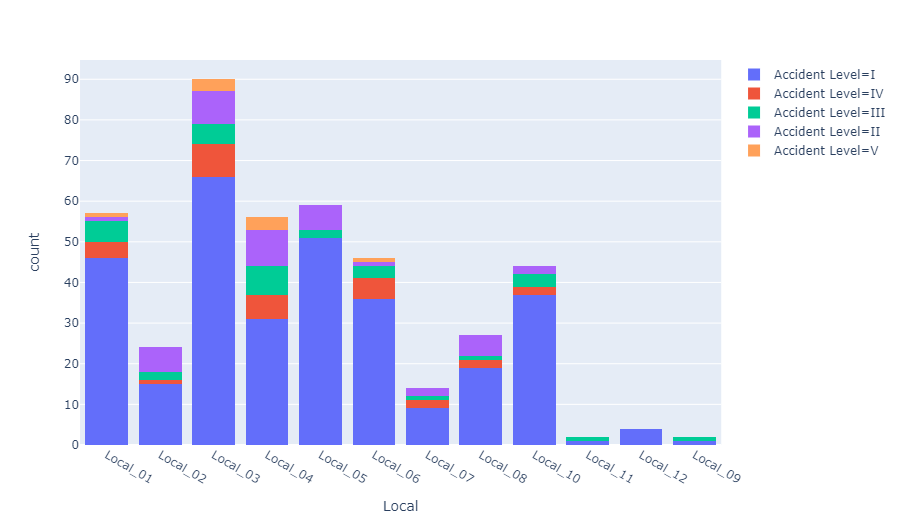


1. Majority of the accident Level I accidents has occured only in Country I.
2. Maximum number of accidents in all countries are mainly of type Accident Level I.
3. Country\_01 has had accidents of all Accident types, making it the riskiest place as per the dataset.

The most number of accidents have occurred in the Mining Industry in Country 1 so far, followed by the metal industry, also in Country 1.

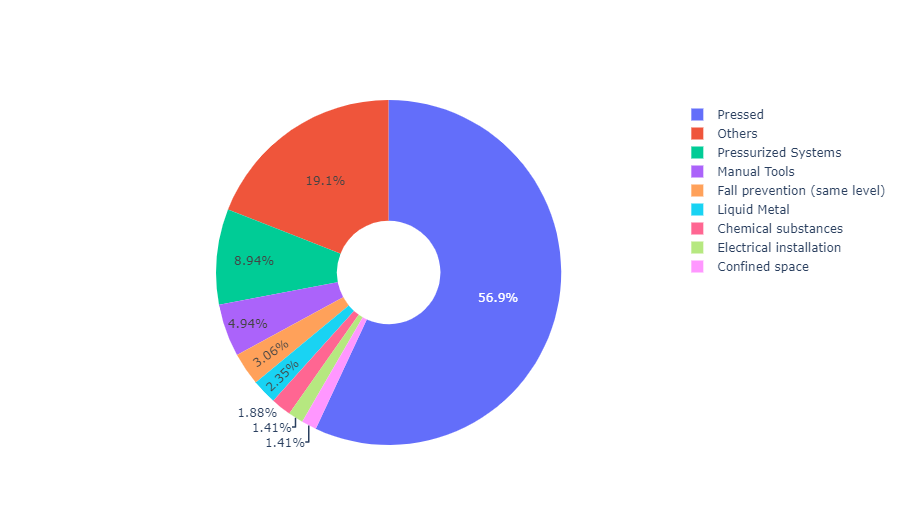






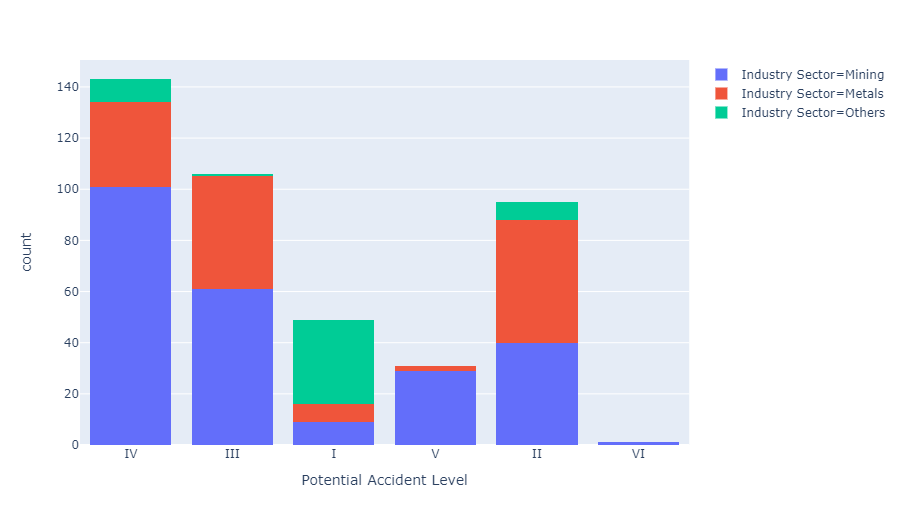
Observations from the above graphs:

1. Local 01,Local 02,Local 03,Local 04,Local 07 all have plants belonging to the Mining Sector and they have had the most number of accidents.
2. Other industry sectors have had the least number of accidents.
3. Local 09 and Local\_11 seems to be the safest cities, with only 2 accidents, even though it has plants belonging to the Metal sector.

****

**Observations from the above graph**

1. Local 04 and Local\_03 have the highest number of unknown risks i.e. of type "Other".
2. We can see that Local 06 has various kind of risks present, which means it could have more of metal plants there.
3. We can also see that the next critical risks apart from Other are "Power Lock","Fall/Fall Prevention" and "Vehicle and mobile equipment".

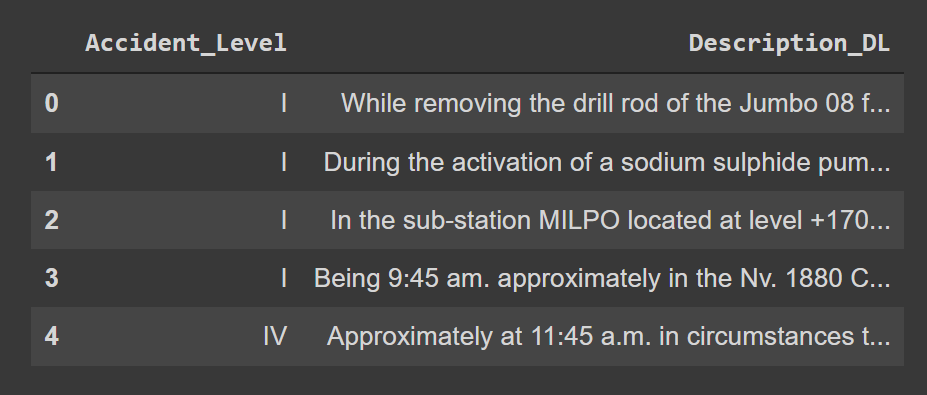


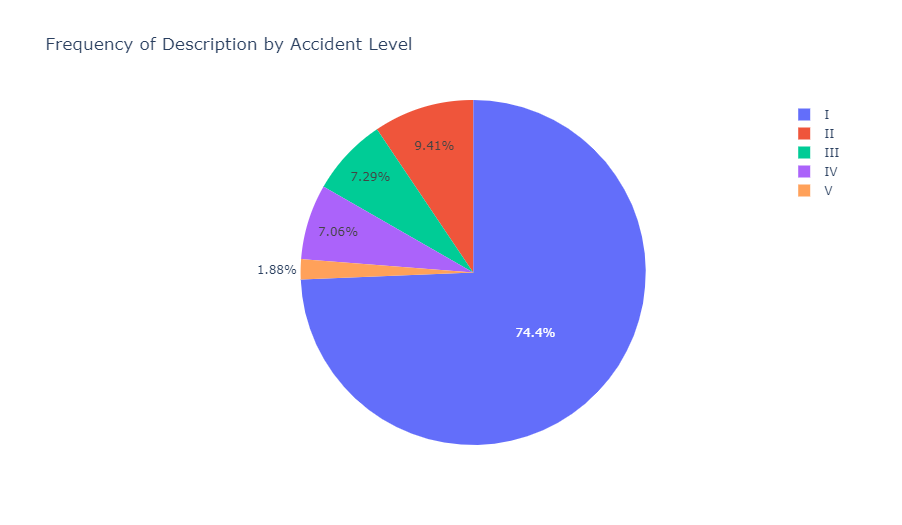
**Observations from the above graph**

Here it is clearly visible that in the mining industry, third party employees have met with the maximum number of accidents as compared to the metal industry where their employees have met with the highest number of accidents.

1. Major number of accidents have occured in the Potential Accident Level 3 category.
2. Potential Accident Level 5 is least in the mining industry.

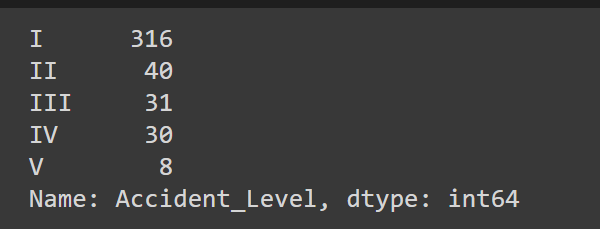
**Data Augmentation**

A dataset is created using only the class variable "Accident Level" and Description column.  


****

Description column is imbalanced in the dataset. Most of the description is present only for Accident Level I(0)

Checking the exact counts of Descriptions per Accident level.

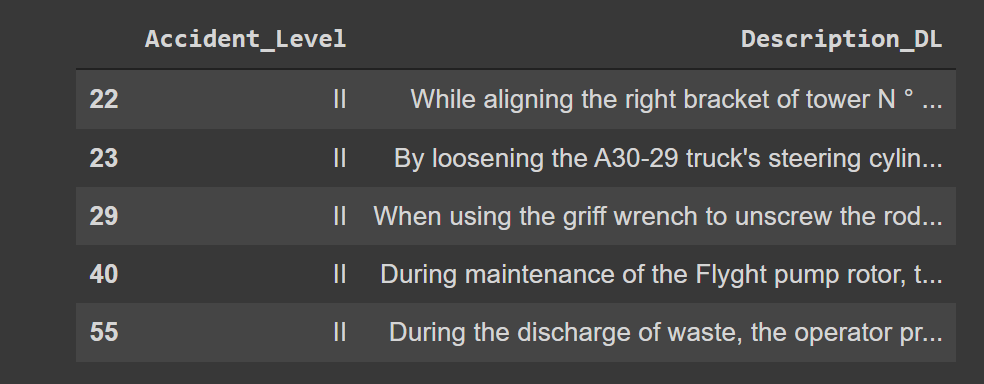


Trying different data augmentation techniques so that the data is balanced properly before it is passed into the dataset.

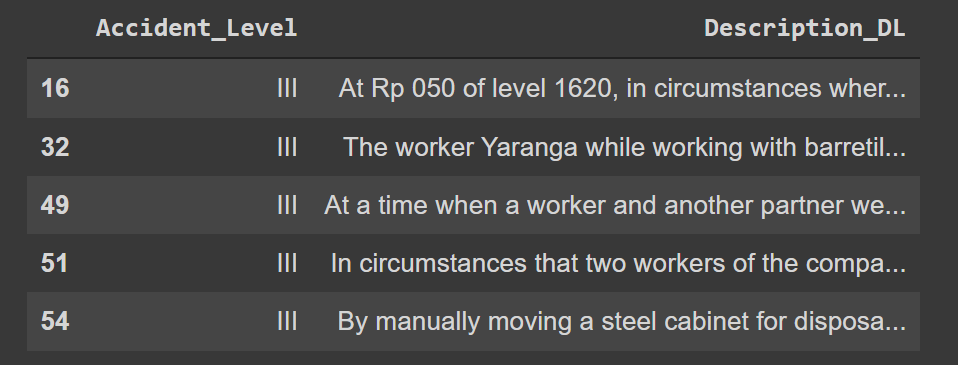
1. SImple upsampling

Using EDA let us perform data augmentation

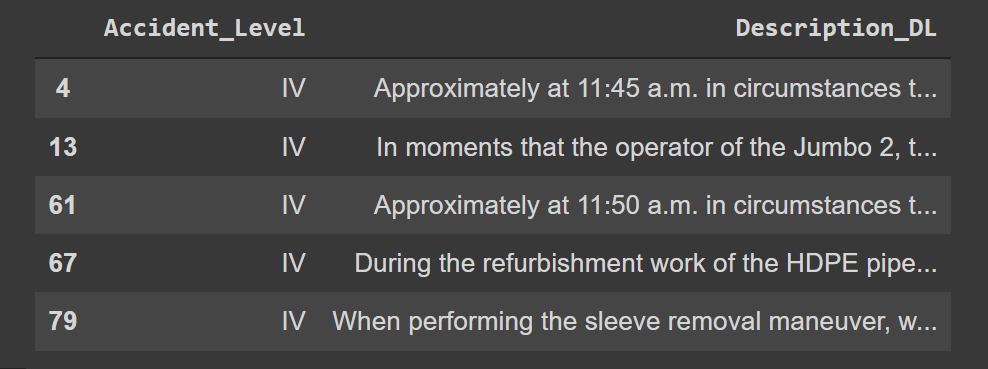
Data Augmentation for the accident **level 'II'**

****

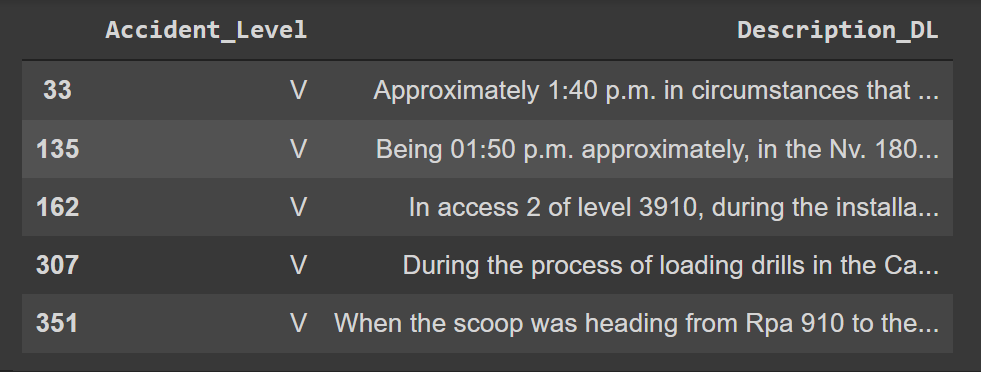
Data Augmentation for the accident **level 'III'**



Data Augmentation for the accident **level 'IV**

**'**

Data Augmentation for the accident **level 'V**

****