Tab 1

PGP AIML Feb 24 A - Group 7

**CAPSTONE PROJECT - Interim Report**

NATURAL LANGUAGE PROCESSING

INDUSTRIAL SAFETY - NLP BASED CHATBOT

horizontal line

# Placeholder image

# 

# TEAM DETAILS

| **Mentor**: ROHIT GUPTA | |
| --- | --- |
| NAVEEN JOTHI  MOHIT JAIN RAHUL PATERIA  AKASH KUMAR  SOHIT SINGH  PRATIK PATIL | [naveenjothi040@gmail.com](mailto:naveenjothi040@gmail.com)  [mohit\_jain@yahoo.com](mailto:mohit_jain@yahoo.com)  [rahulpateriya121@gmail.com](mailto:rahulpateriya121@gmail.com)  [akki.vik3@gmail.com](mailto:akki.vik3@gmail.com)  [Singh.sohit46@gmail.com](mailto:Singh.sohit46@gmail.com)  [pppratikpatil723@gmail.com](mailto:pppratikpatil723@gmail.com) |

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

[**TEAM DETAILS 2**](#_gjdgxs)

[**PROBLEM STATEMENT 6**](#_arolcxe0i15c)

[DATA DESCRIPTION 6](#_la5jp5tnimjw)

[**PROJECT OBJECTIVE 6**](#_4sfysv2zp72)

[**EXPLORATORY DATA ANALYSIS 7**](#_dt16u8sqi2rl)

[INITIAL REVIEW 7](#_suygn8zzl2p)

[HANDLING DUPLICATES AND OTHER DATA CLEANING STEPS 8](#_ma0abj7on5s5)

[ANALYSING PATTERNS 11](#_f8v1xtd5ja6q)

[OBSERVATIONS 12](#_a6ya9b49m645)

[VISUALIZATION 12](#_t3sldqejev78)

[UNIVARIATE ANALYSIS 12](#_ic05hbetxeqt)

[KEY INSIGHTS - Univariate Analysis 16](#_9fs7evvfko04)

[1. Geographical Distribution 16](#_hpclbhc5b2d6)

[2. Industry Sectors 16](#_l1vksfu8osro)

[3. Accident Severity 17](#_erqk9rhayz67)

[4. Demographics 17](#_4pohq7uud0g)

[5. Critical Risks 17](#_vll4eku3fj57)

[6. Temporal Trends 18](#_tgn5qh4fa7uy)

[RECOMMENDATIONS 18](#_vxxjkst55j1t)

[BIVARIATE ANALYSIS 18](#_rjw0eopizwhx)

[Correlation between ‘Accident Level’ and ‘Potential Accident Level’ 18](#_23fabmwv834z)

[Correlation interpretation and implications 19](#_5l912kg2twbb)

[Trend of accidents by year-month by ‘Potential Accident Level’ 23](#_ikha6h2qd6my)

[KEY INSIGHTS - Bivariate Analysis 25](#_39vgytmdj8a8)

[Potential Accident Level by Industry Sector 25](#_hog0zfn2m5iq)

[Potential Accident Level by Year 25](#_uh4urh951xau)

[Potential Accident Level by Critical Risk 25](#_98d81kb6d8y9)

[General Observations 25](#_s5krr392umz)

[MULTIVARIATE ANALYSIS 26](#_xgr2r4z9oxcj)

[KEY INSIGHTS - Multivariate Analysis 29](#_e0ldtyu2fx0s)

[Accident Level vs. Potential Accident Level by Industry Sector 29](#_olbgn3kcgt1h)

[Accident Level by Local and Country 29](#_ubdj1ehtta33)

[Critical Risks vs. Accident Level by Gender 29](#_kh74fd76fk85)

[Year-Month Trends of Accident Levels by Industry Sector 29](#_jymvx5m0gajk)

[General Observations 30](#_rtyk482al12l)

[RECOMMENDATIONS 30](#_6wlsrg5vebhy)

[**NLP - PREPROCESSING 31**](#_fyma60c4kb2p)

[TRANSLATION 31](#_d7bp6bylu4qi)

[CLEANUP STEPS 33](#_j1llwp5jlboc)

[OBSERVATIONS 35](#_5t01nb599veu)

[**VISUALIZING THE DATA (NLP) 36**](#_jznqubk3v8js)

[WORD FREQUENCY DISTRIBUTION 36](#_ohyo8ug9322m)

[OBSERVATIONS 36](#_pps5b8mfa3v5)

[DISTRIBUTION OF TEXT LENGTH 37](#_pbw32w4ekb28)

[OBSERVATIONS 38](#_g9kc1s8my78k)

[RECOMMENDATIONS 38](#_7lrabftprexk)

[TOP N MOST COMMON WORDS 39](#_dicepm4wogxh)

[OBSERVATIONS 40](#_ad85t5w43ztk)

[N-GRAMS 41](#_pxhckk19a9rn)

[UNI-GRAMS 41](#_cpfiazmd3ag5)

[BI-GRAMS 42](#_n3tr1xp2nt9j)

[TRI-GRAMS 43](#_uwdsjka06i2h)

[OBSERVATIONS FROM N-GRAMS: 43](#_qfnw40r399wz)

[N-GRAMS BY INDUSTRY SECTORS 44](#_ikpzuxjsege9)

[OBSERVATIONS AND INSIGHTS: 45](#_9b6c4oc3j5ye)

[POS Tagging 45](#_m9k4g3hdsbpj)

[OBSERVATIONS AND INSIGHTS: 46](#_mqm6lp5bultf)

[**EXPORTED TO EXCEL FORMAT 47**](#_xafbwis2st77)

[**DECIDING MODELS AND MODEL BUILDING 47**](#_8dlsdxuoj69k)

[CHECK FOR IMBALANCES 47](#_chv4o1puvnaq)

[OBSERVATIONS AND INSIGHTS 48](#_u54dqmspx6k)

[MERGING POTENTIAL ACCIDENT LEVEL VI WITH V 49](#_nmjpaj4gc0e)

[OBSERVATIONS FROM THE PLOT AND DISTRIBUTION 49](#_td1mq4jpp7rd)

[UPSAMPLING OF TARGET VARIABLE 50](#_d5u37r6u4qd8)

[MODEL BUILDING 51](#_ik0l5zjnsgvv)

[WHAT MODELS ARE USED AND WHY: 52](#_46kfbivmmkdt)

[WHY THIS APPROACH MATTERS: 53](#)

[DIFFERENT CLASSIFIER INSIGHTS 54](#_233q6o8fs3qd)

[Observations and Insights 55](#_1tbrkrfssu2u)

[Visualization Insight: 55](#_li40aj4y9h65)

[Next Steps 55](#_8oo7i7bgq9c5)

[GRID SEARCH FOR NAIVE BAYES/SVM/LOG REG/SVM 56](#_2htg6rajwdrl)

[Insights from Hyperparameter Tuning 57](#_exbfyxrg378z)

[Next Steps 58](#_ioi42ss45o0)

[SYNONYM REPLACEMENT 58](#_3rl3bne62w1j)

[**KEY INSIGHTS AFTER DATA AUGMENTATION 60**](#_v5kfbr798gk4)

[TOP PERFORMERS 60](#_3mg5pr5m6led)

[Consistent Performers 60](#_ol3sf4390n7c)

[Moderate Performers 61](#_76h73x2st4dy)

[Recommendations¶ 61](#_ld6q30aybws2)

# 

# PROBLEM STATEMENT

The database comes from one of the biggest industries in Brazil and in the world. It is an urgent need for industries/companies around the globe to understand why employees still suffer some injuries/accidents in plants. Sometimes they also die in such an environment.

## DATA DESCRIPTION

This database is basically records of accidents from 12 different plants in 03 different countries where every line in the data is an occurrence of an accident.

**Columns description:**

* Data: timestamp or time/date information
* Countries: which country the accident occurred (anonymised)
* Local: the city where the manufacturing plant is located (anonymised)
* 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.

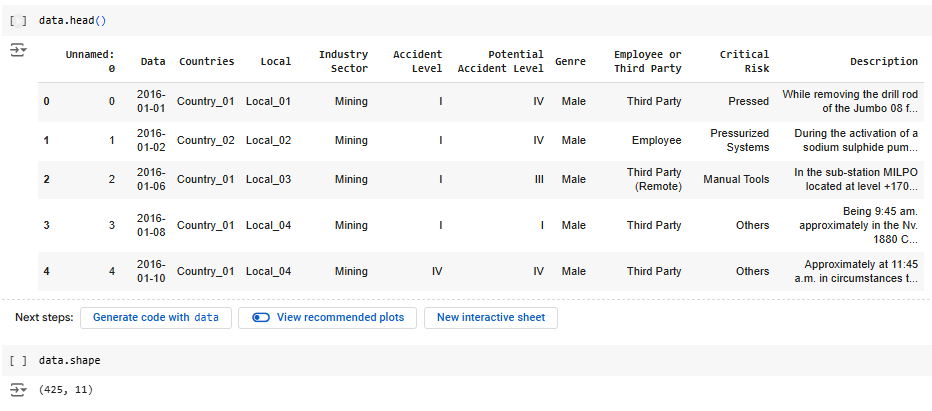
# PROJECT OBJECTIVE

Design a ML/DL based chatbot utility which can help the professionals to highlight the safety risk as per the incident description.

# EXPLORATORY DATA ANALYSIS

## INITIAL REVIEW

After importing the necessary libraries and loading the data, we analyzed the data's structure and significance, taking appropriate actions based on our observations.

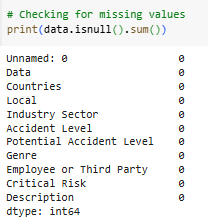


**Insights gained on dataset:**

The dataset is relatively small (425 rows x 11 columns) but contains relevant information.

It is important to note that some components of the dataset were anonymized to conceal the names and locations of the facilities.

**ANALYSING AND HANDLING MISSING/NULL VALUES**



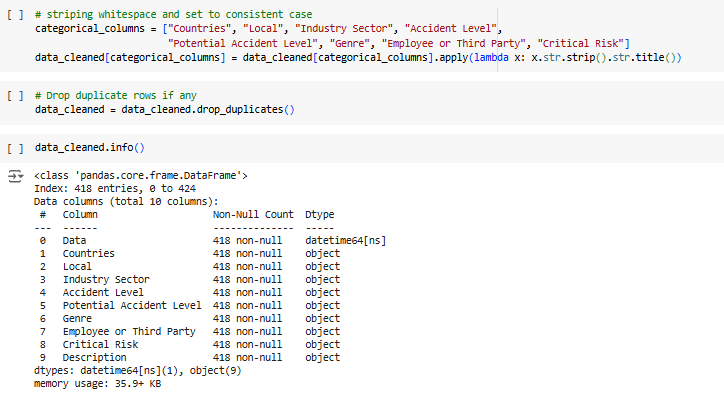


There are no missing values in the dataset. An unnecessary column named "Unnamed" was removed due to the lack of related metadata, as it added no value to the analysis.

## HANDLING DUPLICATES AND OTHER DATA CLEANING STEPS

We observe duplicate records in the given dataset, hence we drop the duplicates.

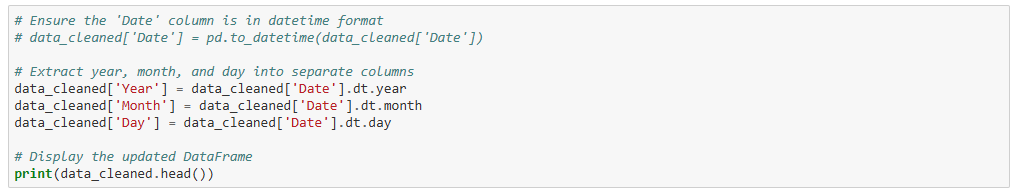
Initial dataset had 425 rows as observed in the previous step, after dropping 7 duplicate rows, we are left with 418 rows.



We noticed that the dataset contains two column names, "Data" and "Genre," in Portuguese. To ensure consistency in the language across the dataset, we are converting these columns to English.

## 

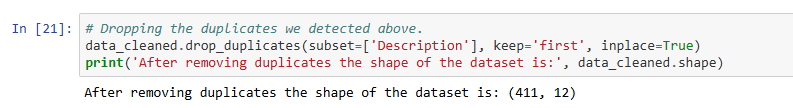
We are extracting the day, month, and year from the date column in order to conduct exploratory data analysis (EDA) in the subsequent steps. This will help us understand the distribution and identify patterns related to the days and months when incidents occurred.



Since we have already extracted the date features into individual columns, the original "Date" feature is no longer necessary and will be dropped.

## 

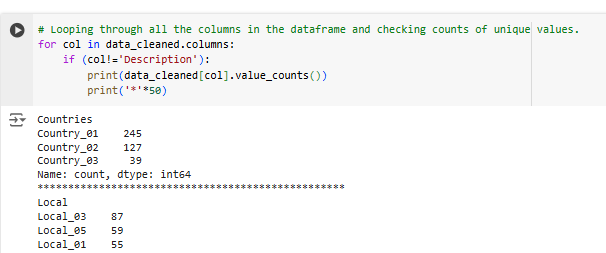
Additionally, we observe that the data frame contains 7 rows where only one or two column values differ, while the "Description" column remains the same across these rows. This inconsistency is logically incorrect, so we have decided to drop these rows as well.

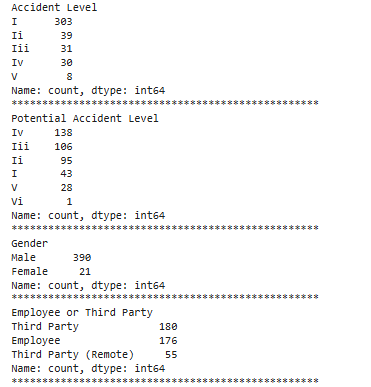


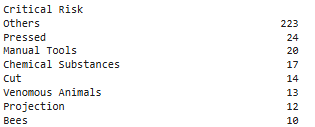
## 

## ANALYSING PATTERNS

Analyzing the occurrence patterns across categorical data within the dataset to identify key trends and insights.









### OBSERVATIONS

* Most of the incidents (over 50%) occurred in Country\_01
* Most of the incidents from Country\_01 has been reported from Local\_03 city
* Most incidents reported are least severe (Level I)
* Most incidents occurred are to male which signifies it's a male dominated industry
* "Others" is the largest category (223 incidents), suggesting the need for clearer risk definitions.

## 

## 

## 

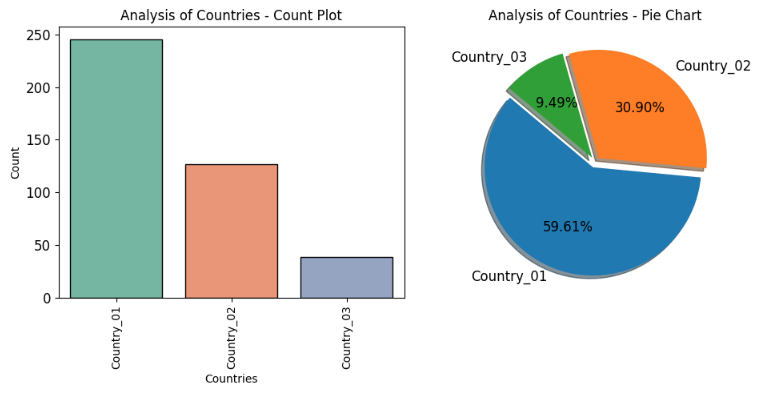
## 

## 

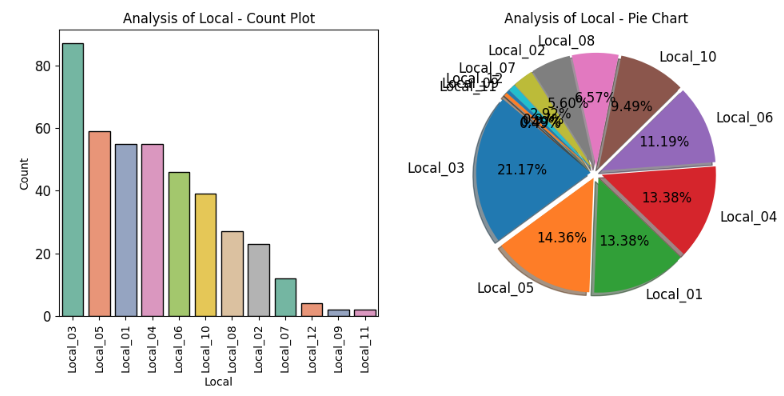
## VISUALIZATION

### UNIVARIATE ANALYSIS

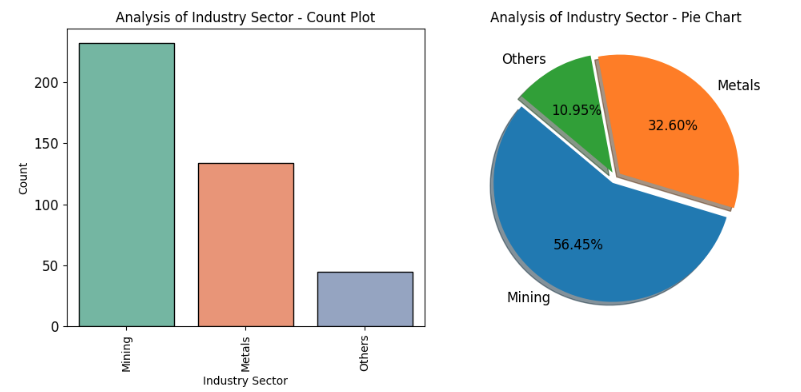
Incident occurrences by countries



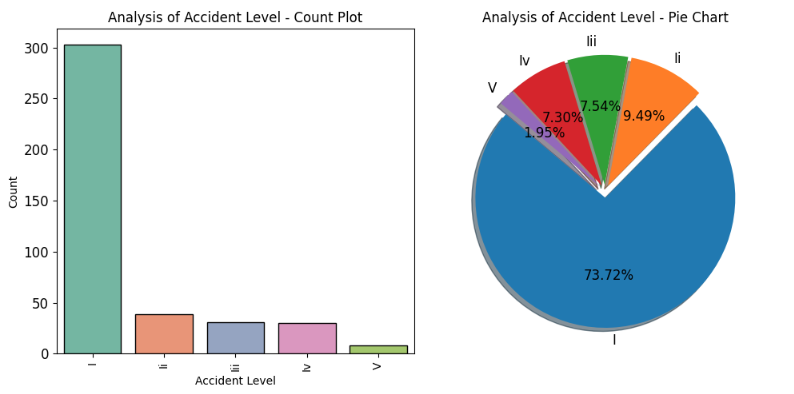
Incident occurrences by local (City)



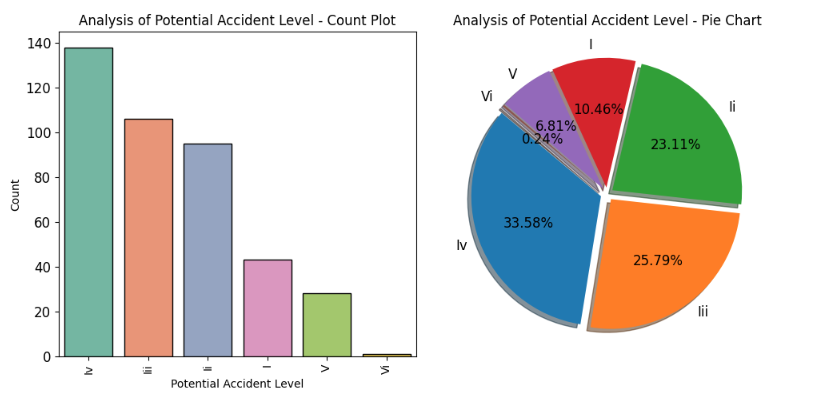
Incident occurrences by industry sector



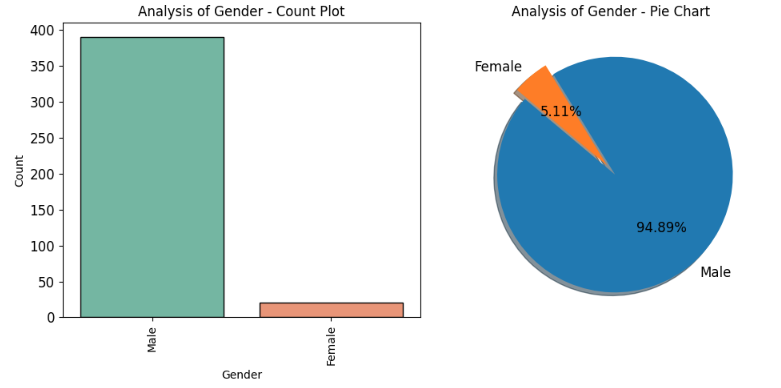
Incident occurrences by accident level



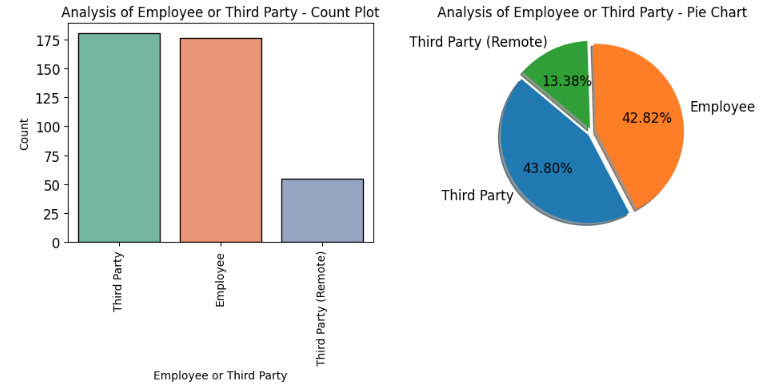
Incident occurrences by potential accident level



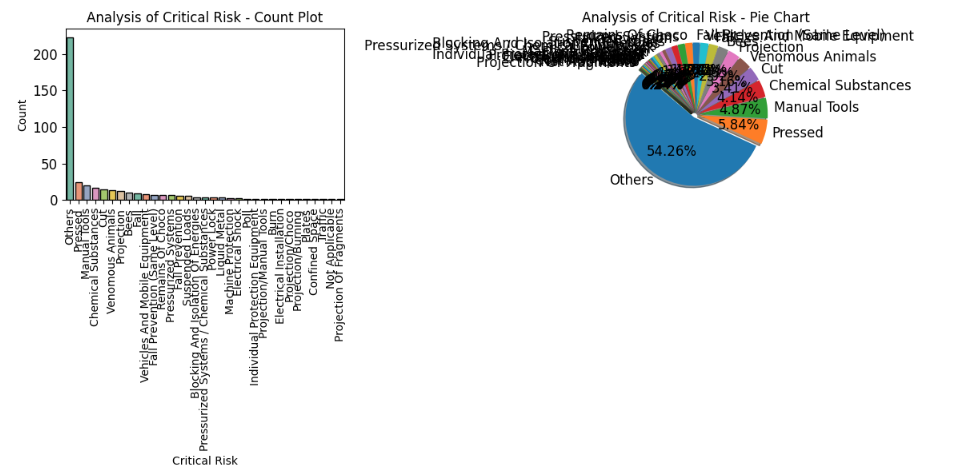
Incident occurrences by gender



Incident occurrences by employment type



Incident occurrences by critical risk



#### KEY INSIGHTS - Univariate Analysis

##### 1. Geographical Distribution

* **Countries**:
  + Country\_01 accounts for a significant majority of incidents (245), comprising over half the dataset.
  + Country\_02 and Country\_03 have substantially fewer incidents (127 and 39, respectively).
* **Local Areas**:
  + Local\_03 stands out with 87 incidents, making it the most accident-prone area.
  + Local\_05, Local\_01, and Local\_04 follow closely with similar counts, suggesting these locations require focused attention.

##### 2. Industry Sectors

* **Mining** is the leading sector with 232 incidents, emphasizing its high-risk nature.
* **Metals** contribute to 134 incidents, indicating it is also a critical area for safety interventions.
* Other sectors account for a relatively small proportion (45).

##### 3. Accident Severity

* **Accident Level**:
  + Level I accidents dominate the dataset (303 incidents), signifying that minor accidents are the most common.
  + Higher-severity accidents (Levels IV and V) are less frequent but still significant for targeted safety measures.
* **Potential Accident Level**:
  + Potential Level IV accidents (138) and Level III (106) highlight areas of latent high-risk scenarios.

##### 4. Demographics

* **Gender**:
  + Males represent a staggering 95% of incidents, suggesting male-dominated roles in these industries may face greater exposure to risks.
* **Employee vs. Third Party**:
  + Third-party individuals (including remote third parties) are involved in more incidents (235) compared to employees (176), indicating external personnel face considerable safety challenges.

##### 5. Critical Risks

* **Top Risk Factors**:
  + "Others" (223 incidents) may require further investigation to identify underlying risk contributors.
  + Pressed (24), Manual Tools (20), and Chemical Substances (17) are notable specific risks.

##### 6. Temporal Trends

* **Yearly Data**:
  + A majority of accidents occurred in 2016 (278 incidents), compared to 2017 (133), indicating a declining trend.
* **Monthly Data**:
  + February (60) and April (51) saw the highest number of incidents, suggesting seasonality effects or operational peaks.

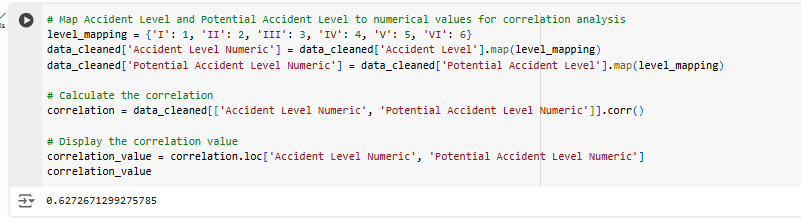
#### RECOMMENDATIONS

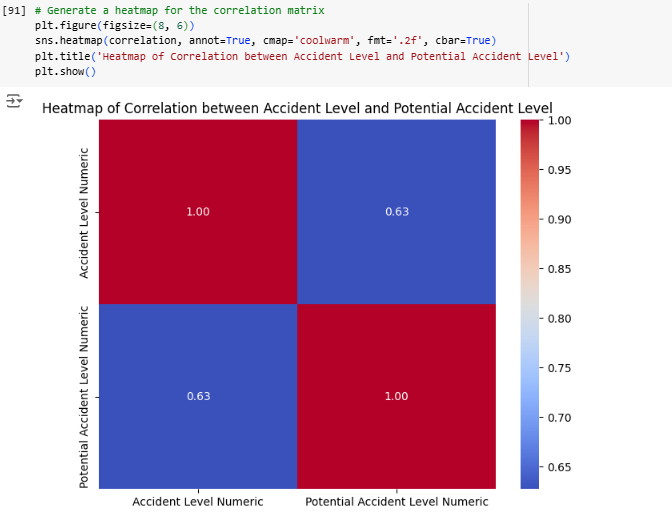
* Focus safety interventions on Country\_01 and Local\_03.
* Address critical risks like Pressed, Manual Tools, and Chemical Substances.
* Enhance safety measures in Mining and Metals industries.
* Develop targeted safety programs for males and third-party workers.
* Investigate incident spikes in February and April for potential seasonal or operational causes.most incidents (278).

### 

### BIVARIATE ANALYSIS

#### Correlation between ‘Accident Level’ and ‘Potential Accident Level’



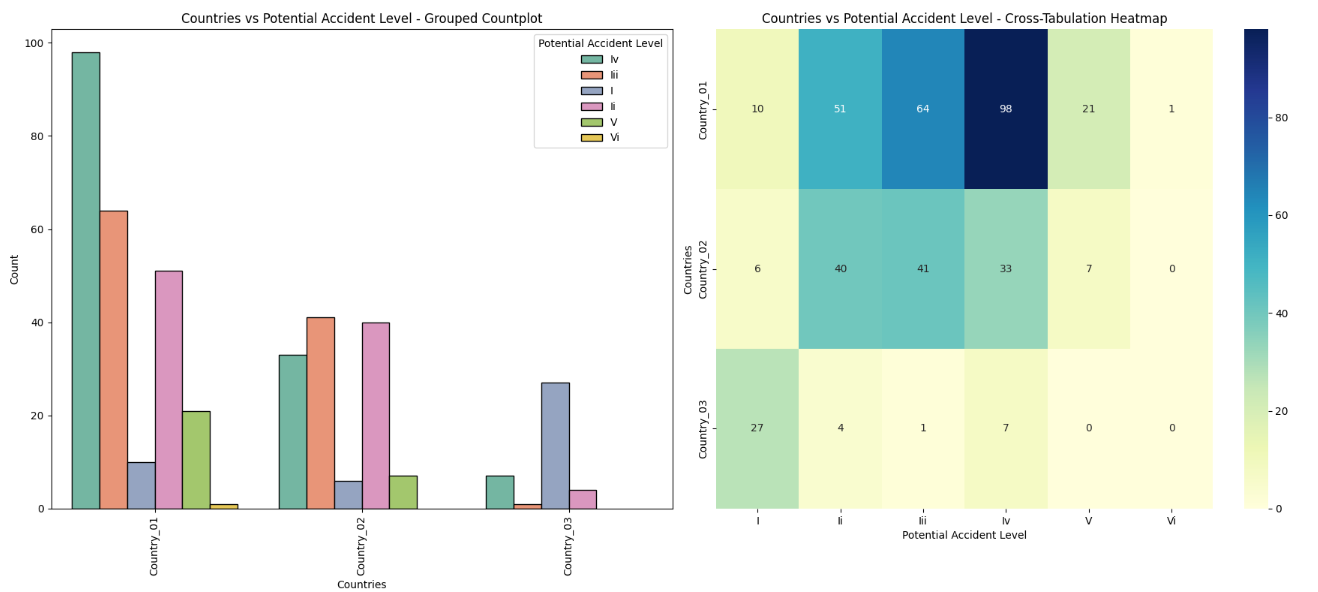


A correlation of **0.627** between Accident Level and Potential Accident Level indicates a **moderately strong positive relationship** between these two variables. Here's what this means in context:

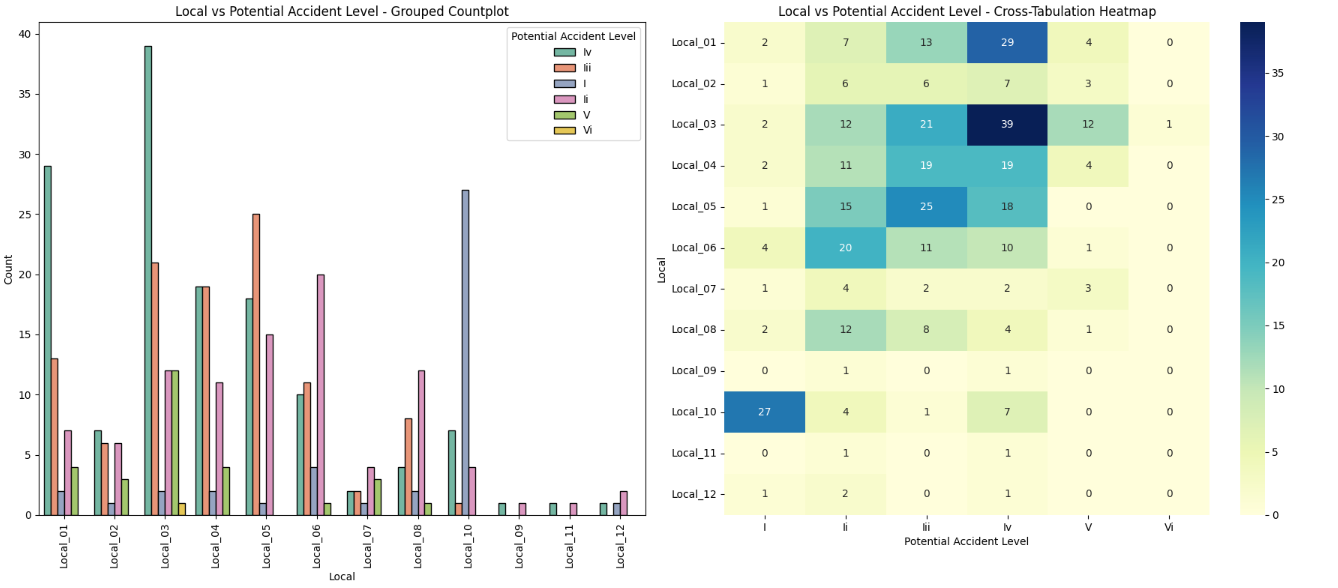
#### Correlation interpretation and implications

1. **Positive Relationship**:
   * As the **Accident Level** increases (indicating more severe actual accidents), the **Potential Accident Level** also tends to increase (indicating higher potential severity of the accident).
2. **Moderately Strong Correlation**:
   * The value of **0.627** suggests that while there is a significant relationship, it is not perfect. Other factors may also influence the Potential Accident Level, apart from the Accident Level.
3. **Practical Implication**:
   * This correlation implies that high-severity accidents are often associated with high potential risks, indicating a need to prioritize preventive measures for incidents with high potential severity to avoid severe actual outcomes.

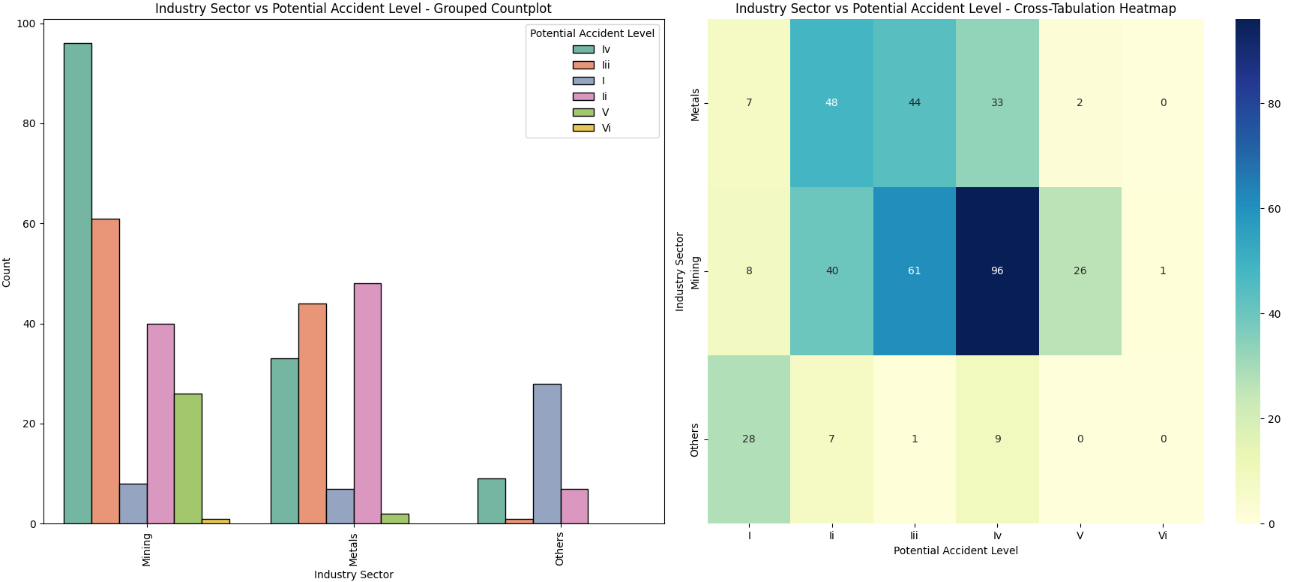
Selecting the ‘**Potential Accident Level**’ as the **target variable** performing the bivariate analysis against all the other categorical columns.

‘Potential Accident Level’ by ‘Countries’:

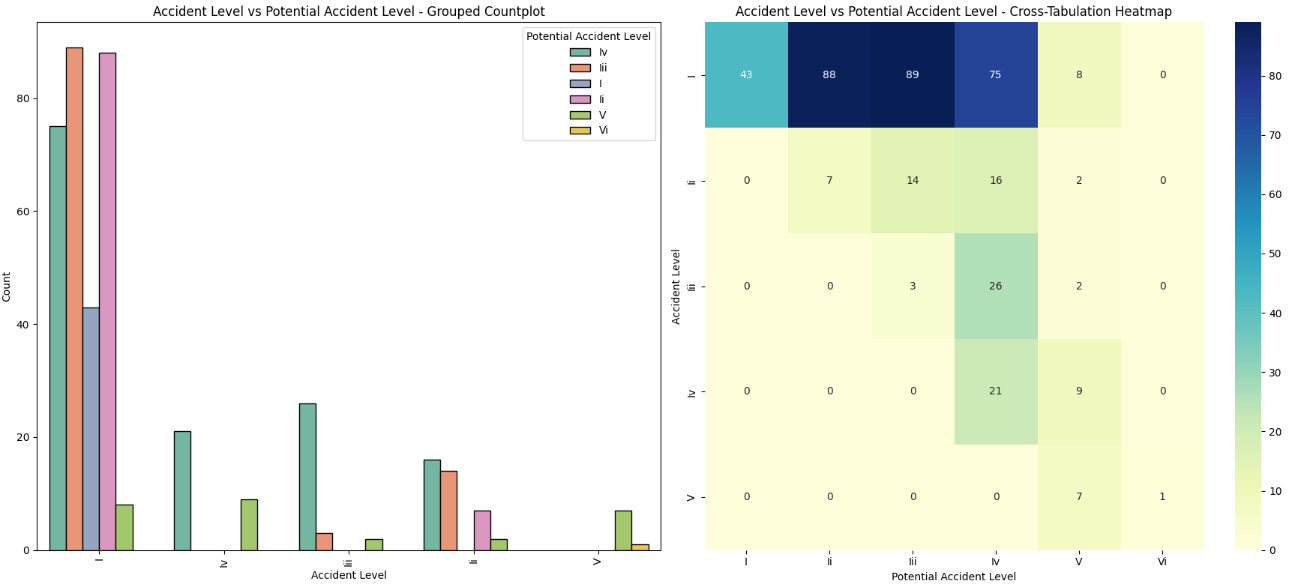
‘Potential Accident Level’ by ‘Local’ (City):



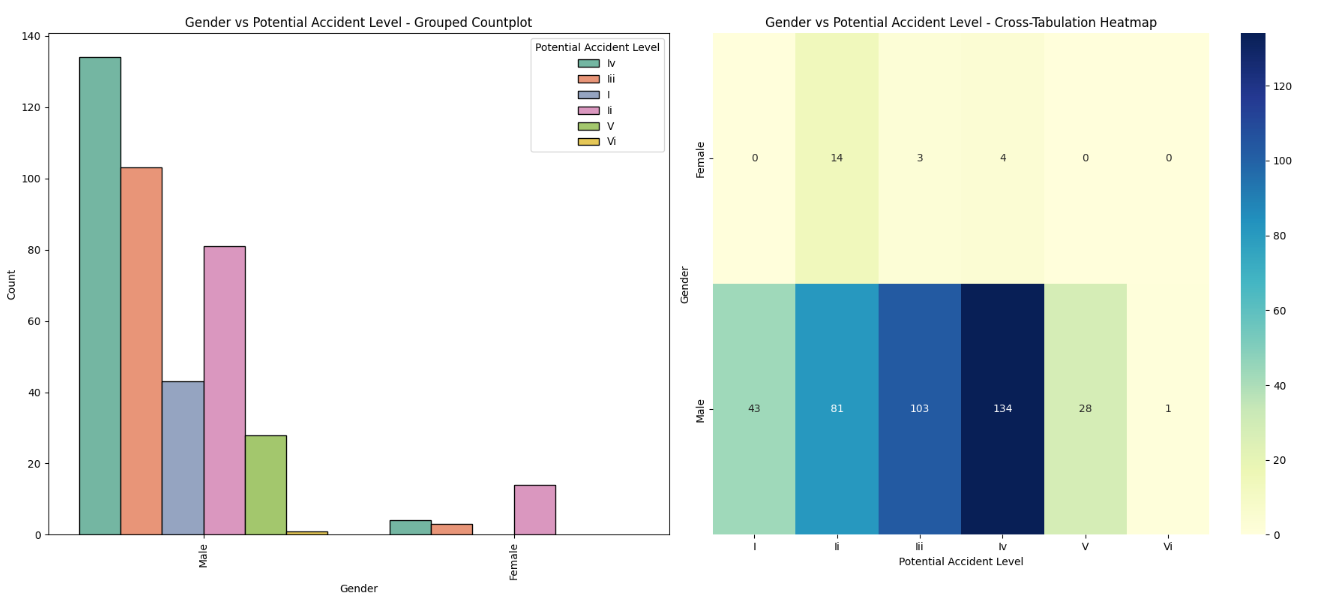
‘Potential Accident Level’ by ‘Industry Sector’:



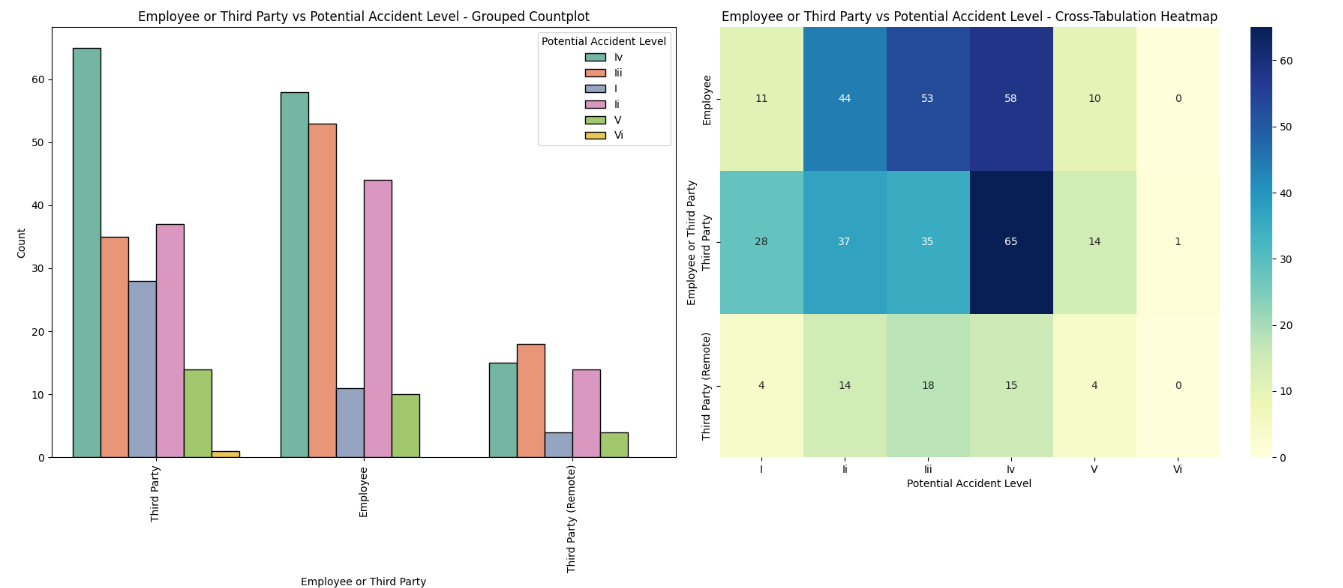
‘Potential Accident Level’ by ‘Accident Level’:



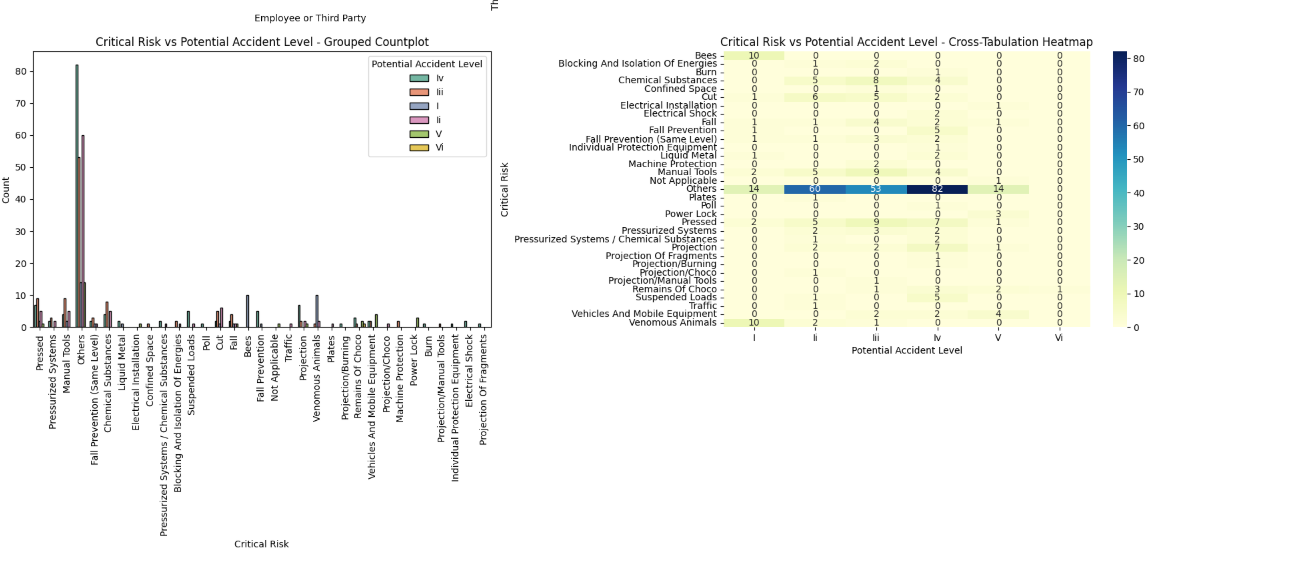
‘Potential Accident Level’ by ‘Gender’:



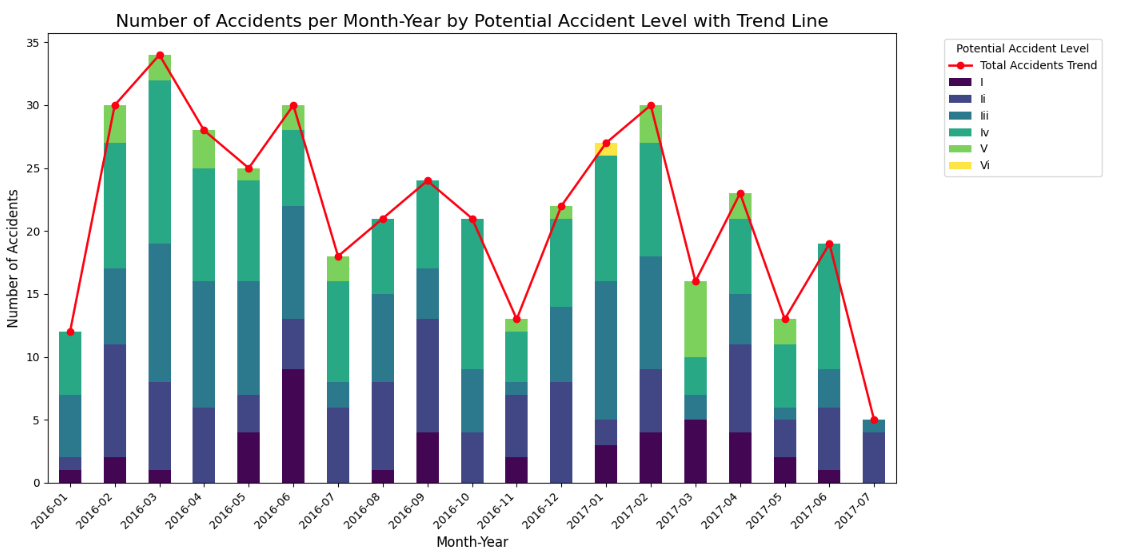
‘Potential Accident Level’ by Employment Type:



‘Potential Accident Level’ by ‘Critical Risk’:



#### Trend of accidents by year-month by ‘Potential Accident Level’



**Level IV Dominance**: Level IV potential accidents consistently dominate across all months, highlighting the need to focus on high-severity risk mitigation.

**Fluctuating Monthly Accidents**: Peaks in total accidents occur in February-March 2016 and January-February 2017, suggesting periodic or seasonal factors affecting incident rates.

**Decline in Mid-2017**: A clear decline in total accidents is observed in mid-2017, potentially reflecting effective safety measures or changes in activity levels.

**Low but Present Extreme Risks**: Levels V and VI (extreme risks) are rare but appear sporadically, underscoring the importance of preparedness for severe incidents.

**Consistent Moderate Risks**: Lower-potential severity levels (I, II, III) remain consistently present, indicating the need for continuous monitoring and interventions for moderate risks.

#### KEY INSIGHTS - Bivariate Analysis

##### Potential Accident Level by Industry Sector

* **Level IV** potential accidents dominate across all industries (141 incidents), followed by **Level III** (106) and **Level II** (95).
* **Mining** sees the highest count of potential accidents at all levels, reflecting its inherent high-risk nature.
* **Metals** and **Others** have relatively fewer incidents but still contribute to higher-level potential accidents.

##### Potential Accident Level by Year

* **2016** recorded the majority of higher potential severity levels, with Level IV being the most frequent.
* A slight decline in high-severity potential incidents is observed in 2017.

##### Potential Accident Level by Critical Risk

* **Level IV** potential accidents are spread across various risks, with "Others" being the most common, signalling ambiguous or uncategorised hazards.
* Risks like **Pressed** and **Manual Tools** are associated with higher potential severity levels (III and IV).
* **Chemical Substances** frequently appear at moderate potential levels (II and III).

##### General Observations

* **Mining** industry and **2016** have the most significant contributions to higher potential accident levels.
* The need for clearer categorisation of critical risks, especially those labelled as "Others," is evident.
* High-severity risks demand targeted safety interventions to reduce potential impact.

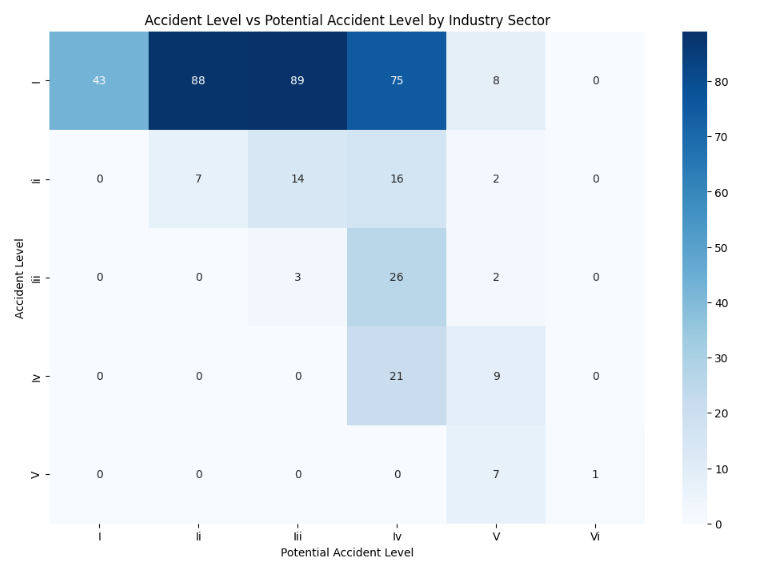
### 

### 

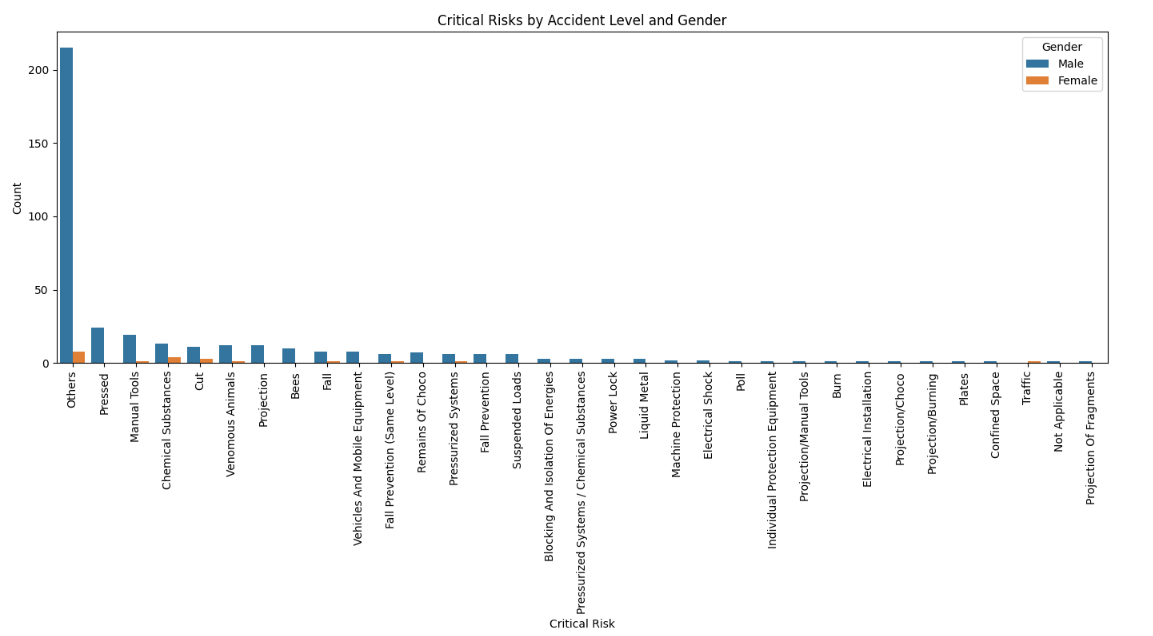
### 

### MULTIVARIATE ANALYSIS

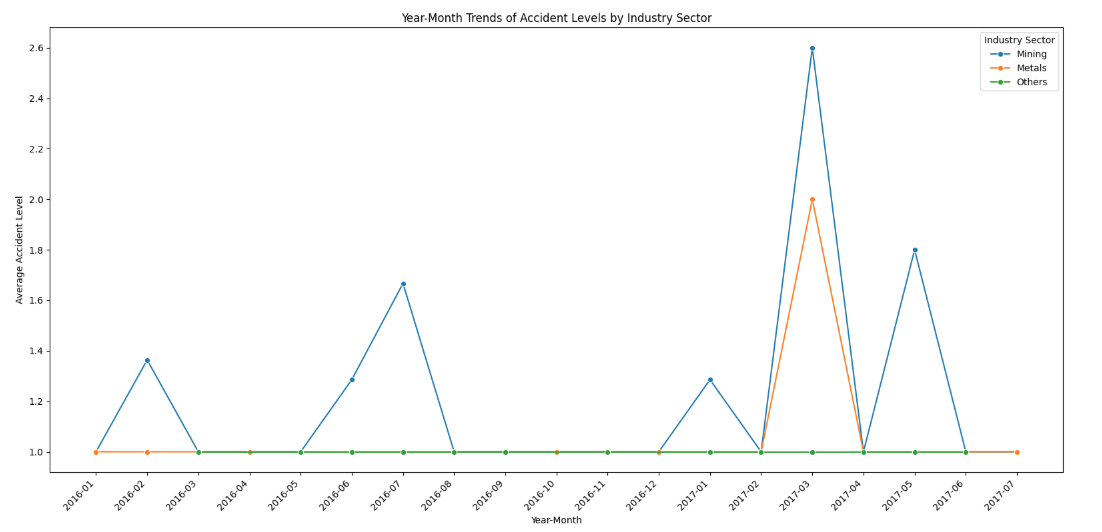
Accident Level vs Potential Accident Level by Industry Sector

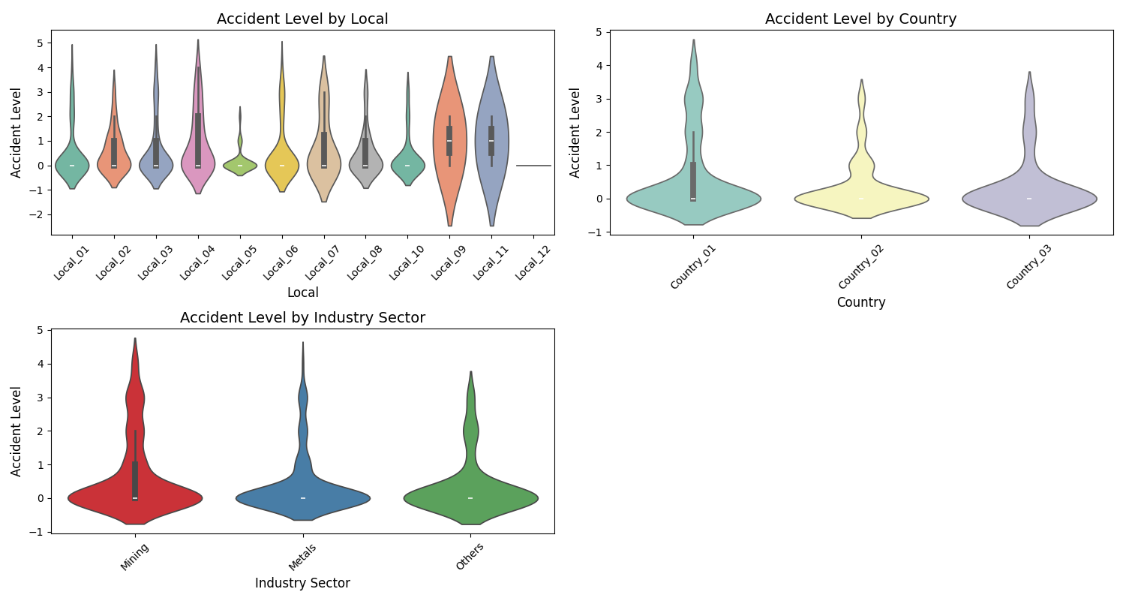


Critical Risks by Accident Level and Gender

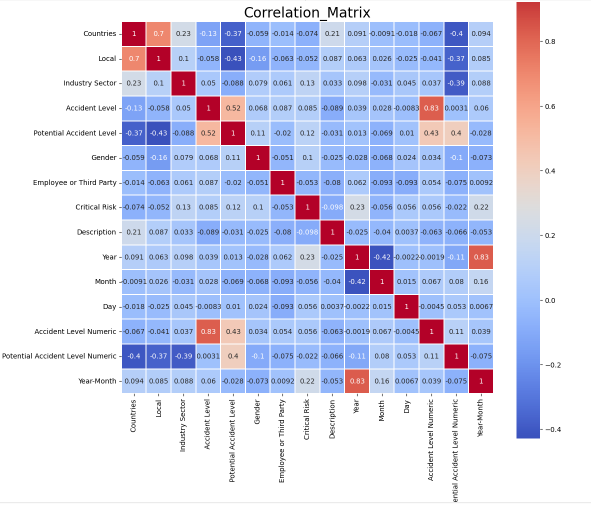


Year-Month Trends of Accident Levels by Industry Sector





Correlation Matrix



#### KEY INSIGHTS - Multivariate Analysis

##### Accident Level vs. Potential Accident Level by Industry Sector

* The violin plot reveals a broader spread of accident levels in the Mining sector, with a skew toward more severe incidents. This highlights the variability and the potential for high-severity accidents within this sector.
* Metals demonstrate a narrower distribution of accident levels, indicating fewer high-severity incidents and suggesting better control mechanisms.
* The "Others" category shows moderate variability, with some potential outliers indicating unusual incidents requiring further investigation.

##### Accident Level by Local and Country

* Significant variability is observed across different locales, with some exhibiting tightly concentrated distributions around lower accident levels and others showing broad distributions with higher frequencies of severe incidents.
* At the country level, accident distributions are more consistent, with fewer extreme values, indicating country-specific safety standards may play a role in moderating risk variability.

##### Critical Risks vs. Accident Level by Gender

* The violin plot analysis confirms that males are more frequently involved in severe incidents, as the spread of accident levels for males is broader and skewed toward higher severities.
* For females, the distribution of accident levels is narrower, suggesting fewer incidents overall, though certain risk types like "Others" still require investigation to ensure equitable safety measures.

##### Year-Month Trends of Accident Levels by Industry Sector

* Temporal trends corroborated by the violin plot indicate that accident levels in the Mining sector are seasonally influenced, with higher peaks during specific months. The broad variability during these peaks suggests fluctuating risk factors, possibly linked to operational cycles or environmental conditions.
* Metals exhibit steadier trends with a narrow spread, indicating a more consistent risk profile.

##### General Observations

* The Mining sector stands out for its high variability in accident levels, with a wider range of severity compared to other sectors. This reinforces its designation as a high-risk industry.
* The "Others" risk category displays broader and often ambiguous distributions, masking the underlying hazards that need clearer categorization.
* Gender-specific differences in accident distributions emphasize the need for tailored safety interventions to address distinct risk exposures for males and females.

#### RECOMMENDATIONS

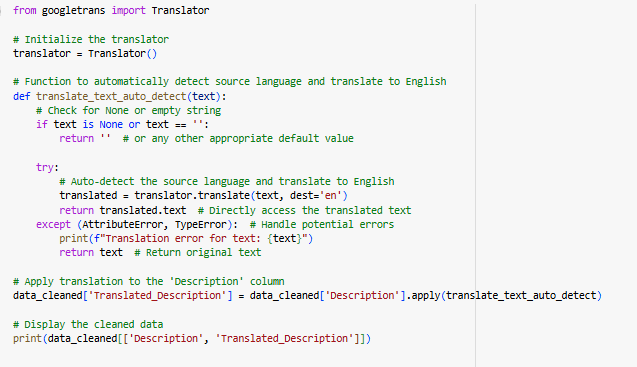
* **Enhance Safety Measures in Mining**Address the high variability and severity of accidents in the Mining sector by implementing advanced risk management tools, particularly during seasonal peaks identified in the temporal trends.
* **Investigate and Categorize "Others" Risks**Refine the "Others" risk category to uncover specific hazards, enabling more targeted safety measures and reducing ambiguity in accident reporting.
* **Implement Gender-Specific Safety Programs**Tailor training and safety protocols for male workers exposed to high-risk activities (e.g., manual tools, pressing equipment) and ensure equitable measures for female workers to address their specific risk profiles.
* **Standardize Local Safety Practices**Reduce inconsistencies between locals with varying accident distributions by standardizing safety practices and addressing outliers in high-risk locations.
* **Strengthen Seasonal Safety Protocols**Introduce heightened safety measures in Mining during peak accident months, with resources allocated to manage higher severity incidents effectively.
* **Continuous Monitoring and Data-Driven Adjustments**Regularly analyze accident distributions to track shifts in central tendencies, variability, and outliers, ensuring safety strategies remain responsive to evolving risks.
* **Refine Safety in Metals Sector**Conduct regular safety audits in the Metals sector to maintain its consistent risk profile and investigate any emerging outliers to prevent escalation.
* **Promote Consistent Risk Management Across Countries**Leverage insights from the relatively uniform accident distributions at the country level to reinforce effective safety standards across all regions.

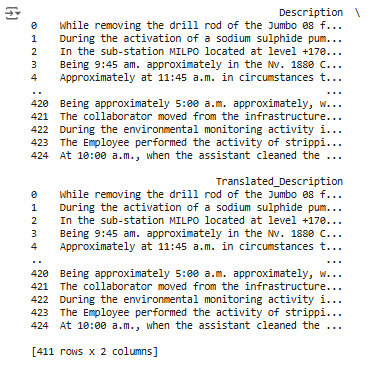
# NLP - PREPROCESSING

Before proceeding with model building, it is essential to pre-process the description column in the dataset to ensure the text is clean and structured for analysis. Key pre-processing steps include text cleaning, tokenization, stopword removal, stemming or lemmatization, handling punctuation, translations, and spell checks. The specific steps chosen for this dataset are detailed in the following section.

## TRANSLATION

Given that the dataset primarily originates from Brazil and may also include data from other regions, we opted to perform translation from various languages to English. This ensures consistency and standardization, making the data more suitable for analysis and model training.





### 

### 

## CLEANUP STEPS

**Handle Missing Text**: If the input text is empty or None, return an empty string.

**Convert to Lowercase**: The text is converted to lowercase to standardize it.

**Tokenization**: The text is split into individual words (tokens).

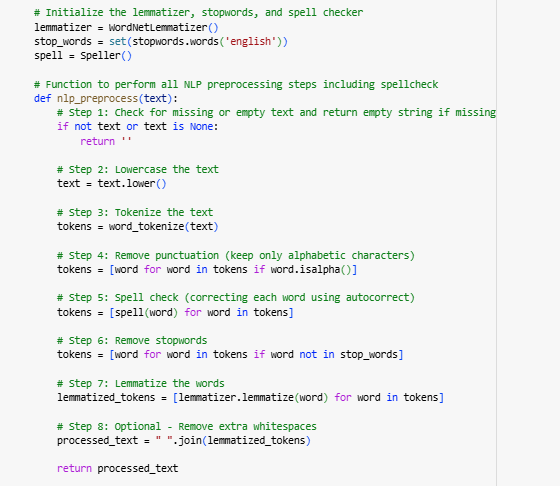
**Remove Punctuation**: Only alphabetic characters are retained, removing any punctuation.

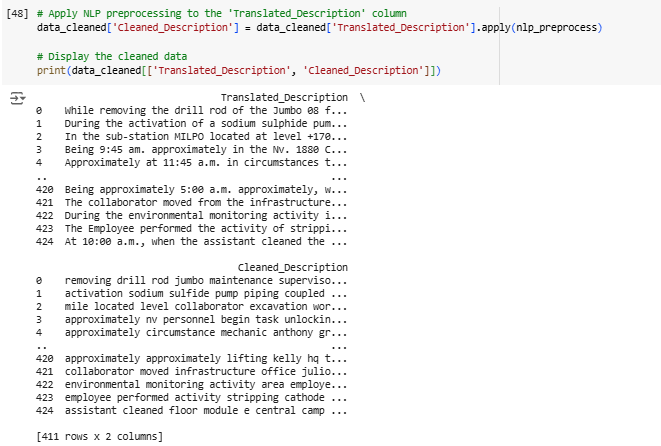
**Spell Checking**: Each word is spell-checked and corrected using the autocorrect function.

**Remove Stopwords**: Common stopwords (e.g., "the", "is") are removed to focus on important words.

**Lemmatization**: Words are reduced to their base form (e.g., "running" to "run"). We choose lemmatization over stemming as it gives better accuracy for model training.

**Whitespace Removal**: Extra whitespaces are removed, and the tokens are joined back into a clean string.



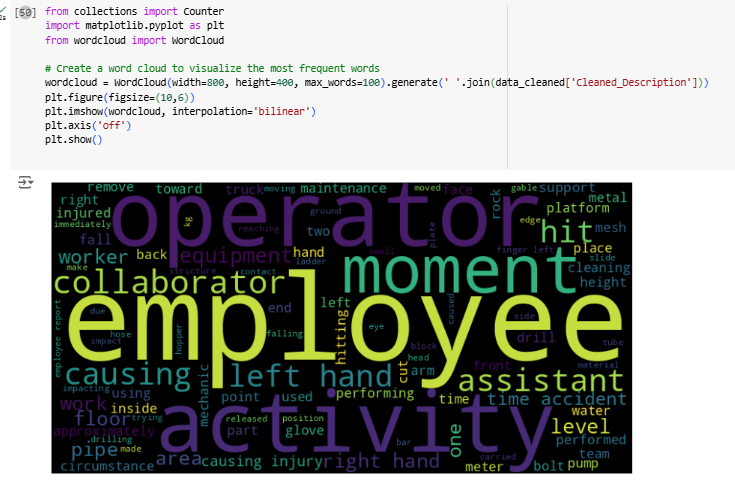


## OBSERVATIONS

* The cleaned descriptions show that functional words like "the," "a," and "in" have been removed, making the text more concise and focused.
* Text is converted to lowercase, ensuring uniformity and treating capitalized and non-capitalized words as the same.
* Punctuation marks have been eliminated, simplifying the text for easier processing in NLP tasks.
* Non-alphabetic characters are removed, and words are tokenized, keeping only the relevant terms.
* Spelling errors are corrected, such as changing "sulphide" to "sulfide," ensuring consistent terminology.
* Lemmatization has reduced words to their base forms, improving consistency and preparing the text for further analysis.

# VISUALIZING THE DATA (NLP)

## WORD FREQUENCY DISTRIBUTION



### OBSERVATIONS

**Dominant Keywords**:

* Words like "employee," "activity," "moment," "operator," and "assistant" appear prominently, indicating frequent mentions in the dataset.
* These terms likely reflect the primary subjects or roles associated with incidents or actions.

**Focus on Actions and Context**:

* Words such as "causing," "hitting," "injury," and "performing" suggest the dataset is related to workplace incidents, with emphasis on actions leading to specific outcomes.
* Context-related terms like "left hand," "equipment," and "floor" highlight common areas or objects associated with these incidents.

**Potential Risk Areas**:

* Terms like "injury," "drill," "equipment," and "cleaning" point toward tasks or tools potentially associated with workplace hazards.
* The frequent mention of "causing" and "accident" suggests an analysis of causes and effects within the dataset.

**Relevance of Specific Body Parts**:

* "Left hand" and "right hand" are highlighted, suggesting a focus on injuries involving hands, which may represent a high-risk area in the workplace.

## 

## 

## 

## 

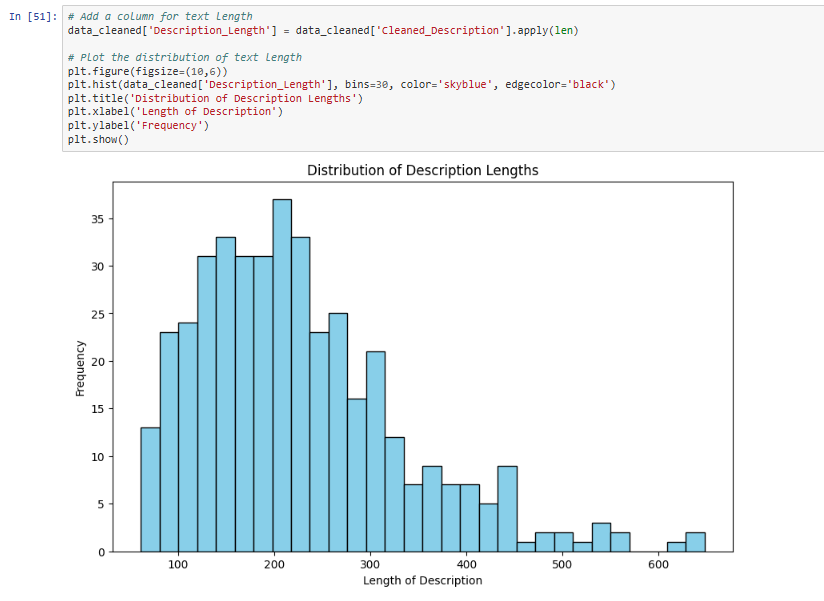
## 

## 

## 

## 

## DISTRIBUTION OF TEXT LENGTH



### OBSERVATIONS

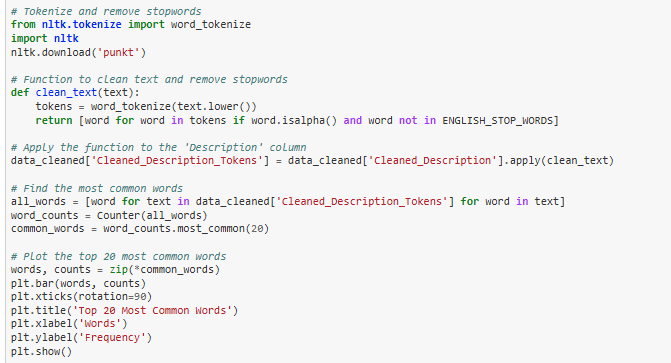
* **Right-Skewed Distribution**: Most descriptions are short, with fewer being much longer.
* **Frequent Length Range**: Descriptions mostly fall between 100-250 characters, peaking around 150-200.
* **Long Descriptions**: A small number exceed 300 characters, with some surpassing 600, likely indicating more complex incidents.
* **Mode**: The most common description length is slightly under 200 characters.

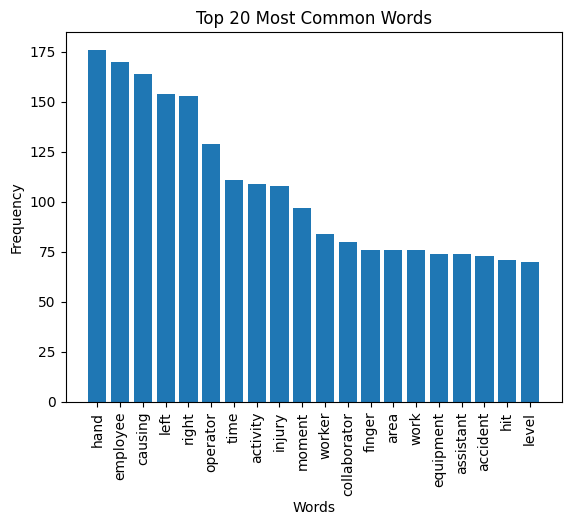
### 

### RECOMMENDATIONS

* **Modeling Considerations**: The prevalence of short descriptions may bias NLP models, requiring strategies for truncation or padding.
* **Text Augmentation**: Augmenting long descriptions or summarizing lengthy ones can address the imbalance.
* **Data Quality Check**: Review long outliers for redundancy or over-detailing and trim for consistency.
* **Segmentation by Length**: Segment descriptions by length to analyze any correlation with accident severity.

## TOP N MOST COMMON WORDS



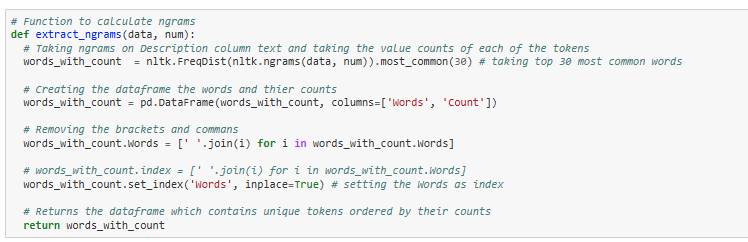


### OBSERVATIONS

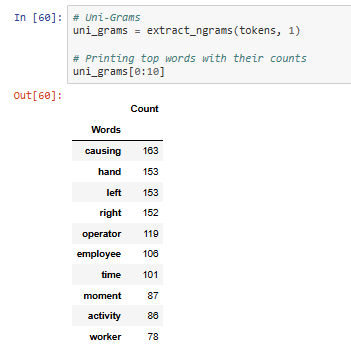
* **Frequent Hand Injuries** – "Hand" and "finger" are the most common words, indicating frequent hand-related incidents.
* **Employee-Centered** – Words like "employee," "operator," and "worker" highlight workplace accidents involving staff.
* **Injury and Direction** – Terms like "injury," "left," and "right" suggest incidents often specify body parts and sides.
* **Task and Time Focus** – "Activity," "moment," and "time" indicate descriptions detail when and during which tasks accidents occur.
* **Equipment and Area** – The presence of "equipment" and "area" points to machinery and specific work zones as common incident locations.

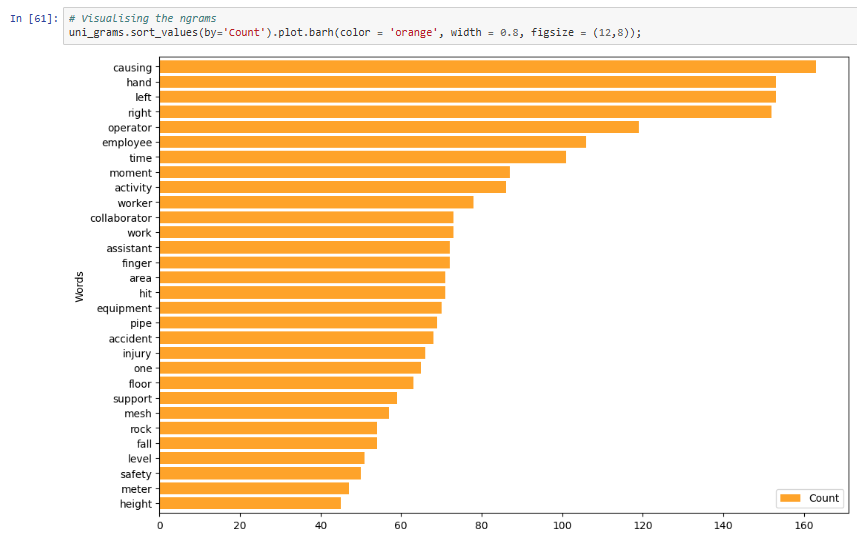
## 

## N-GRAMS



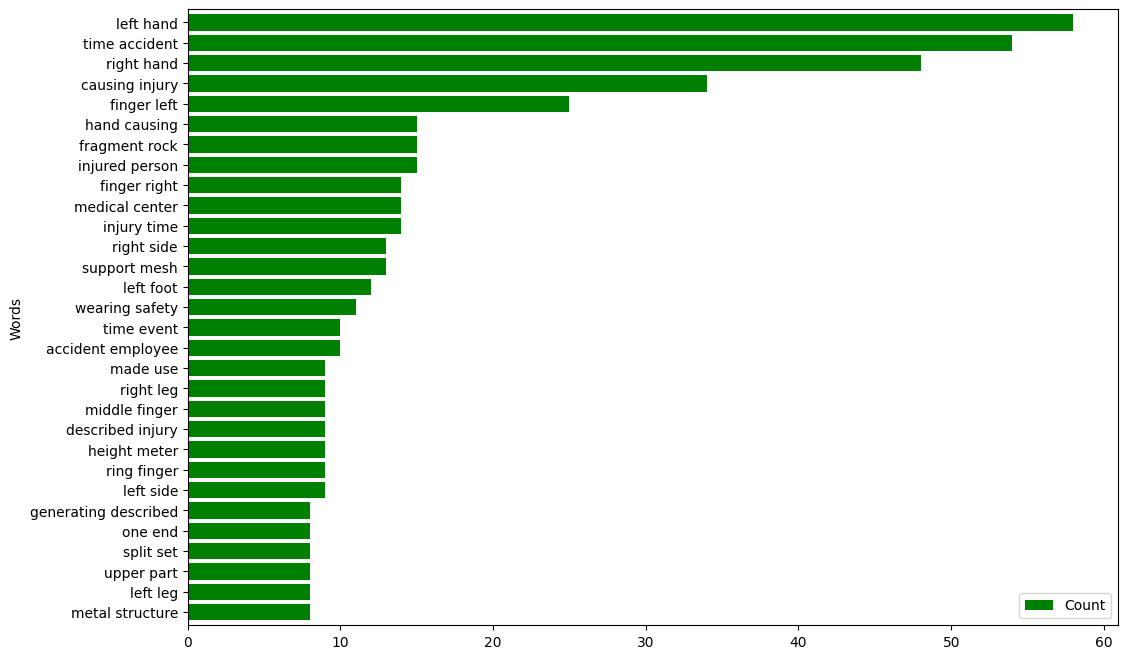
### UNI-GRAMS



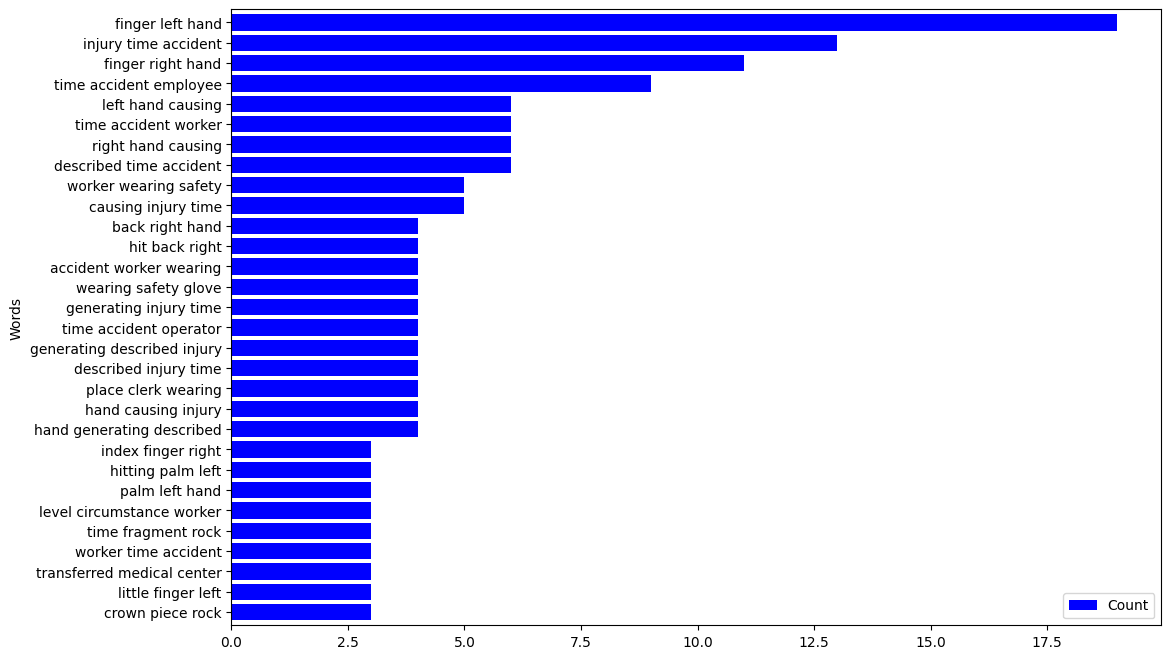


Similarly…

### BI-GRAMS



### TRI-GRAMS



### OBSERVATIONS FROM N-GRAMS:

**Focus on Hands and Injuries**

* Across all n-grams, terms like **"hand," "left hand," "finger left hand"** dominant, highlighting frequent **hand-related injuries**. This suggests a focus on **manual work hazards**.

**Employee and Worker-Centric**

* Words like **"employee," "worker," "collaborator"** frequently appear, indicating that **workplace incidents involving staff** are common. Safety measures should target employee protection.

**Directional and Specific Injuries**

* Phrases like **"left hand," "right hand," "finger left"** suggest that incident reports often **specify the body side and part** affected. This can inform **ergonomic and safety gear improvements**.

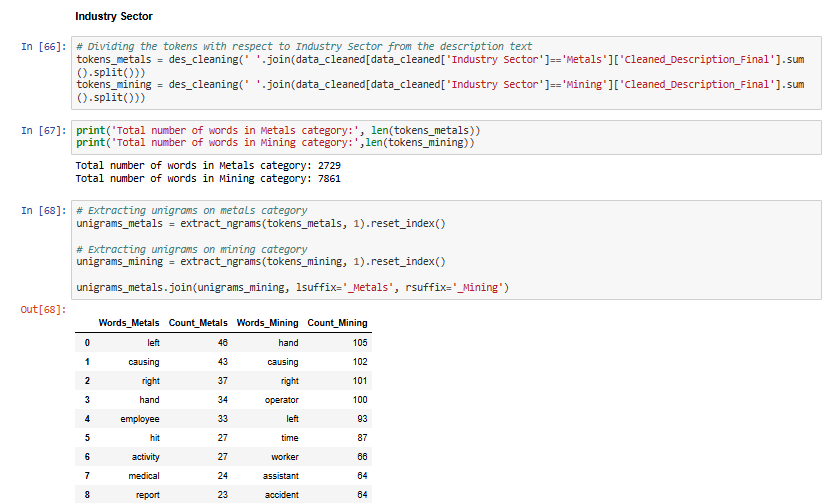
**Frequent Causes and Activities**

* Bigrams and trigrams like **"causing injury," "time accident," "described injury time"** reflect a focus on **causal factors** and the **timing of incidents**. This indicates **incident reporting emphasizes root causes and accident timelines**.

**Equipment and Environmental Factors**

* Unigrams such as **"equipment, pipe, area"** and phrases like **"fragment rock"** suggest that incidents often involve **machinery, tools, or environmental hazards**. Equipment maintenance and area monitoring are essential.

### N-GRAMS BY INDUSTRY SECTORS

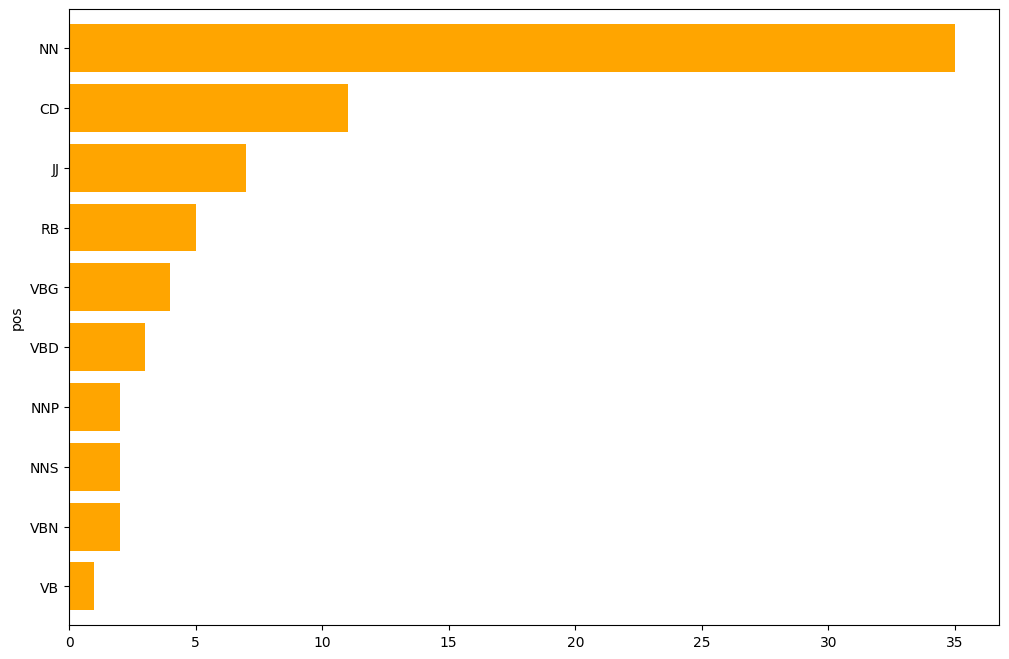


#### OBSERVATIONS AND INSIGHTS:

* **Hand Injuries Dominate** – "Hand" and "left/right" are top terms, indicating frequent **hand-related incidents** in both sectors.
* **Focus on Causes** – "Causing" and "time/moment" highlight emphasis on **incident causes and timing**.
* **Worker-Centered** – "Employee, operator, worker" show **human factors** are key in accidents.
* **Equipment Risks** – Mining involves "equipment, pipe, rock," while metals show risks with **"hose, pump, acid"**.
* **Injury and Safety** – "Finger, injury, fall, cut" reflect focus on **personal injury and safety measures**.

## POS Tagging





### OBSERVATIONS AND INSIGHTS:

* **Nouns (NN) Dominate** – Focus on **objects and entities** like "hand" and "worker."
* **Frequent Numbers (CD)** – Highlights **measurements and quantities** in reports.
* **Descriptive Adjectives (JJ)** – Emphasizes **qualities** of objects/incidents.
* **Actions and Adverbs (RB, VBG, VBD)** – Reflects **detailed event descriptions**.
* **Proper and Plural Nouns (NNP, NNS)** – References to **specific items and multiple objects**.

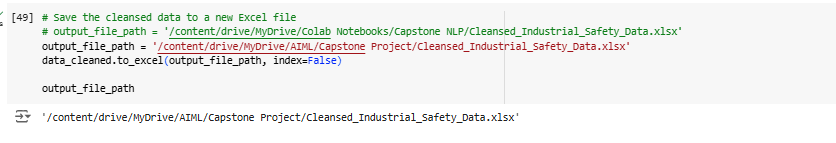
# 

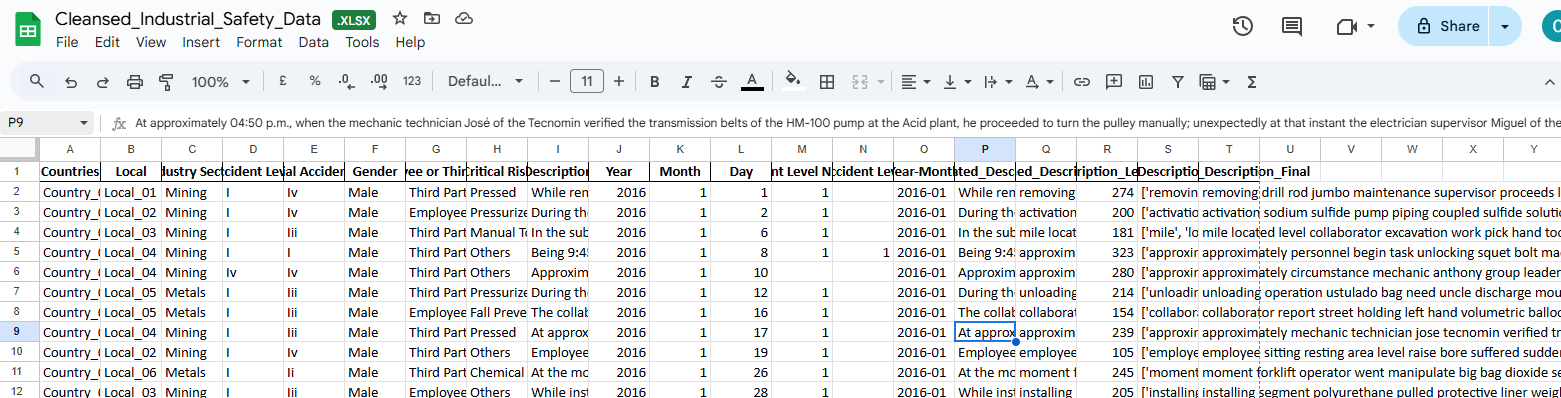
# 

# 

# 

# EXPORTED TO EXCEL FORMAT





# DECIDING MODELS AND MODEL BUILDING

## CHECK FOR IMBALANCES

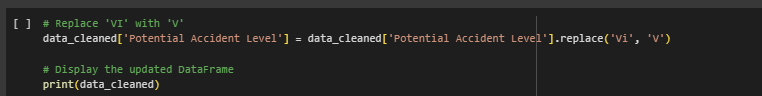
We examined the data for imbalance and analyzed the distribution of the target variable along with the visualization:



### OBSERVATIONS AND INSIGHTS

* We merged Class VI with Class V, aimed at reducing class imbalance or simplifying the classification by combining rare categories.
* The target variable is Potential Accident Level, and its distribution has been visualized.
* A count plot shows the following class distribution:
  + **IV:** 138
  + **III:** 106
  + **II:** 95
  + **I:** 43
  + **V (merged with VI):** 29
* There is a noticeable imbalance in the data. Class IV has the highest number of incidents, while Class V (formerly VI) has the fewest.

### MERGING POTENTIAL ACCIDENT LEVEL VI WITH V





### 

### OBSERVATIONS FROM THE PLOT AND DISTRIBUTION

* **Class Imbalance:**
  + The imbalance might affect model performance, especially for underrepresented classes (Class V). Consider techniques like oversampling (SMOTE) or class-weight adjustments during model training.
* **Dominance of Certain Classes:**
  + Classes IV and III dominate, while Class I and V appear less frequently. This might influence how well the model predicts rare but critical events.
* **Possible Data Issues:**
  + If Class V represents severe incidents, the lower frequency could imply underreporting or less frequent severe accidents.

## UPSAMPLING OF TARGET VARIABLE

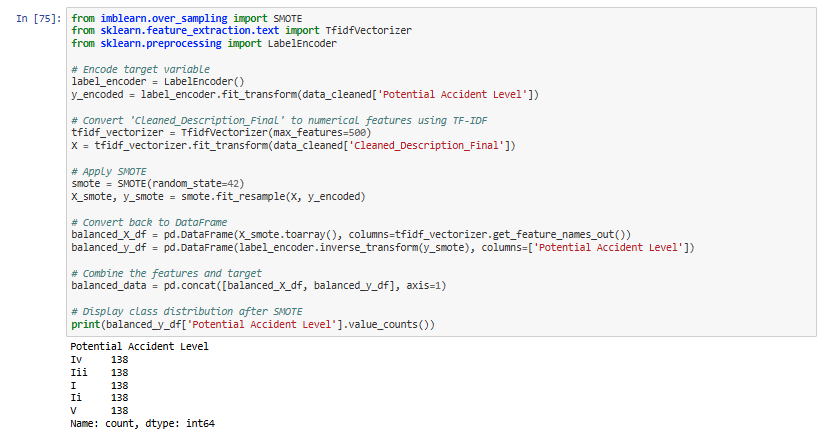
* **Before SMOTE**: The target class distribution was skewed, with class IV having the most samples (138) and class V having the least (29).
* **Problem**: Imbalanced datasets can lead to:
* **Bias** towards majority classes.
* **Poor recall** for minority classes.
* **Inaccurate models** that perform well overall but fail to predict minority class instances correctly.

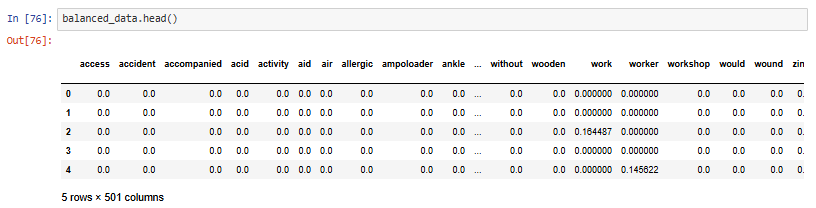
**Solution**: SMOTE addresses this by **upsampling** the minority classes to match the majority, reducing class imbalance.

* The combination of **SMOTE and TF-IDF** creates a robust pipeline for handling text-based classification problems in safety-critical domains.

**Applying SMOTE:**

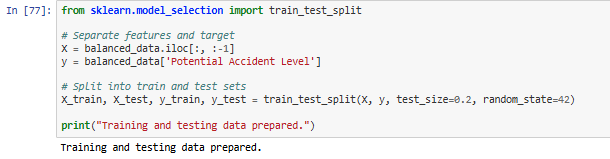
* SMOTE generates synthetic samples for the minority classes by creating new instances that are interpolations between existing samples.
* This ensures that all classes (I, II, III, IV, V) have **138** samples each, balancing the dataset.





## MODEL BUILDING

We are splitting the test train to a 20:80 ratio, with potential accident level as the target variable.



### WHAT MODELS ARE USED AND WHY:

* **Diversity of Models:**
  + Simple models like **Logistic Regression and Naive Bayes** for quick baselines.
  + More complex models like **Random Forest, SVM, and Boosting algorithms (XGBoost, LightGBM)** for higher accuracy potential.
* **Parameterized Models:**
  + Some models are customized (e.g., max\_depth, n\_estimators), ensuring the comparison involves tuned models rather than default parameters.
* **Random State:** Consistency is maintained by fixing random seeds across models, ensuring reproducibility.
* **Automated Model Evaluation:** Testing multiple models with minimal code.
* **Fault Tolerance:** If any model encounters an error (e.g., incompatible input shapes), the loop continues without breaking.
* **Efficiency:** This process eliminates the need to manually write separate code for each model, streamlining experimentation.

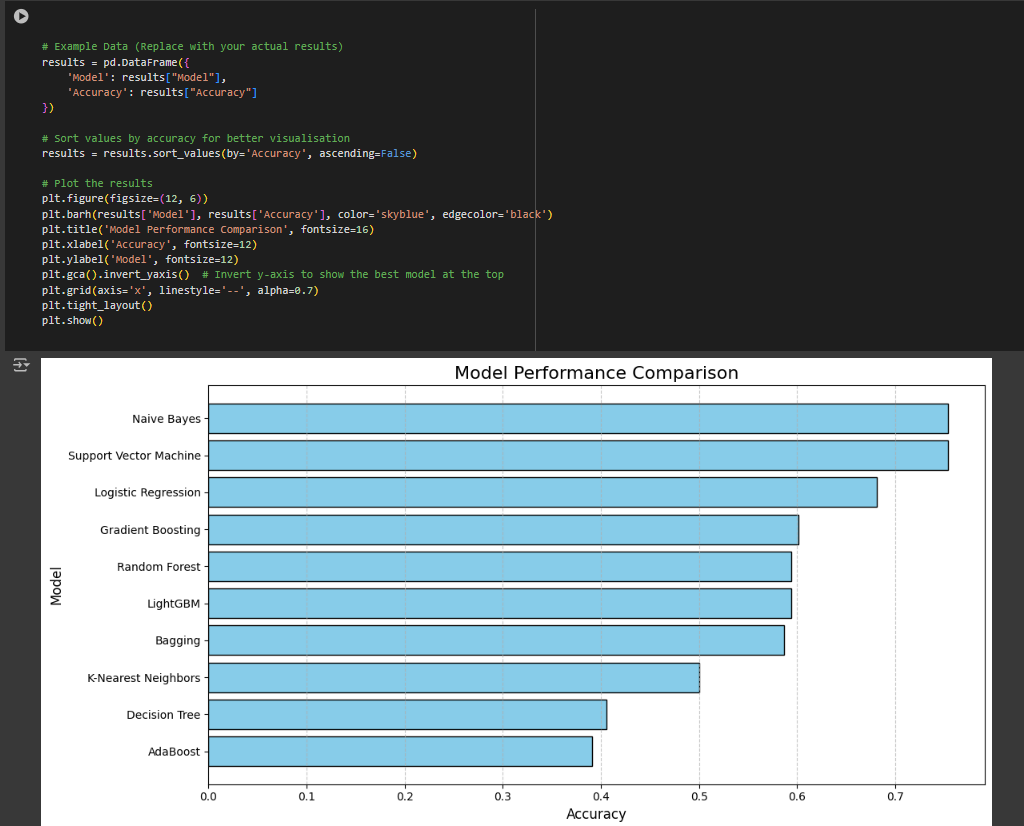
****

### WHY THIS APPROACH MATTERS:

1. Model Comparison:
   * By testing multiple models, this function identifies the most effective one for the given dataset.
   * It enables benchmarking and avoids over-reliance on a single algorithm.
2. Time Efficiency:
   * Running multiple models at once accelerates the workflow, which is crucial during initial exploratory phases.
   * Tuning can then focus on the top-performing models, saving time on less promising ones.
3. Error Handling:
   * Graceful exception handling prevents workflow disruption, ensuring maximum model coverage even if some fail.
4. Ensemble & Boosting Inclusion:
   * Including ensemble models (Bagging, Random Forest) and boosting methods (XGBoost, LightGBM) improves the chances of capturing complex data patterns, often leading to higher performance.

### DIFFERENT CLASSIFIER INSIGHTS

* **Model Comparison**: Visual representation simplifies the comparison, allowing stakeholders to quickly understand performance differences.
* **Efficient Decision Making**: Helps data scientists focus on refining the top-performing models, saving time on weaker models.
* **Explainability**: The visual is easy to interpret, making it suitable for presentations or reports to non-technical stakeholders.

****

#### Observations and Insights

1. **Top-Performing Models:**
   * Naive Bayes and SVM outperform other models, suggesting the dataset may favor algorithms that work well with smaller datasets or linearly separable data.
   * Logistic Regression also performs well, indicating a potential linear relationship between features and the target variable.
2. **Boosting Models:**
   * Gradient Boosting and LightGBM perform moderately well, highlighting their potential in handling complex data patterns.
3. **Poor Performers:**
   * Decision Tree and AdaBoost rank lower, suggesting overfitting or suboptimal parameter tuning.

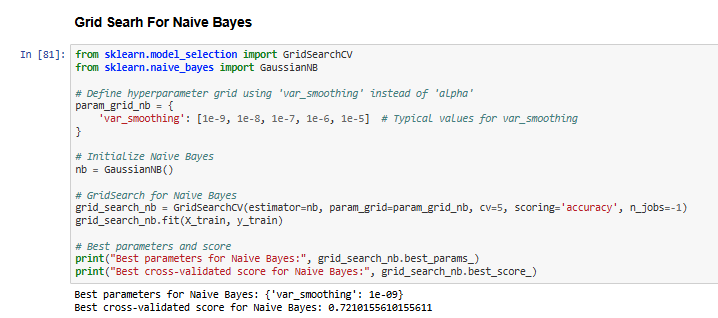
#### Visualization Insight:

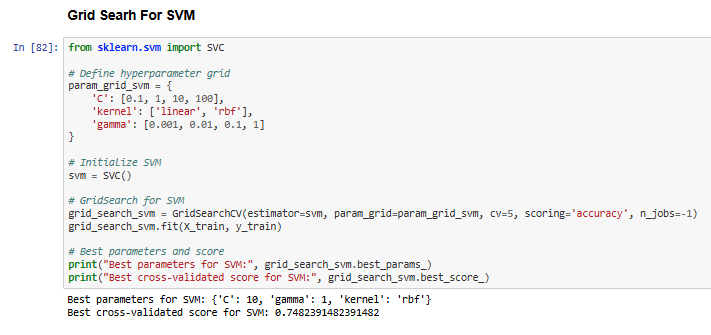
* The **Naive Bayes** model has the highest accuracy, followed by:
  + **Support Vector Machine (SVM)**
  + **Logistic Regression**
  + **Gradient Boosting**
  + **Random Forest**
* **AdaBoost** has the lowest performance, indicating it may not have been effective for the given dataset.

#### Next Steps

* Perform hyperparameter tuning for the top models.
* Explore advanced models like BERT or RoBERTa for text classification.
* Visualise confusion matrices and ROC curves to understand model behaviour better.

## GRID SEARCH FOR NAIVE BAYES/SVM/LOG REG/SVM







### Insights from Hyperparameter Tuning

1. Naive Bayes

* **Best Parameters**: {'var\_smoothing': 1e-08}
  + The smoothing parameter (var\_smoothing) helps manage variance in Gaussian Naive Bayes. A smaller value improves class separation for this dataset.
* **Best Cross-Validated Score**: **70.65%**
  + Naive Bayes performed well but slightly lagged behind other models after tuning.

2. Support Vector Machine (SVM)

* **Best Parameters**: {'C': 10, 'gamma': 1, 'kernel': 'rbf'}
  + **C**: Higher regularization strength allows better fitting to the data.
  + **gamma**: A value of 1 ensures the model captures complex relationships in the data.
  + **kernel**: The rbf kernel effectively handles the non-linear nature of the problem.
* **Best Cross-Validated Score**: **73.74%**
  + SVM achieved the highest score, making it the top-performing model for this dataset.

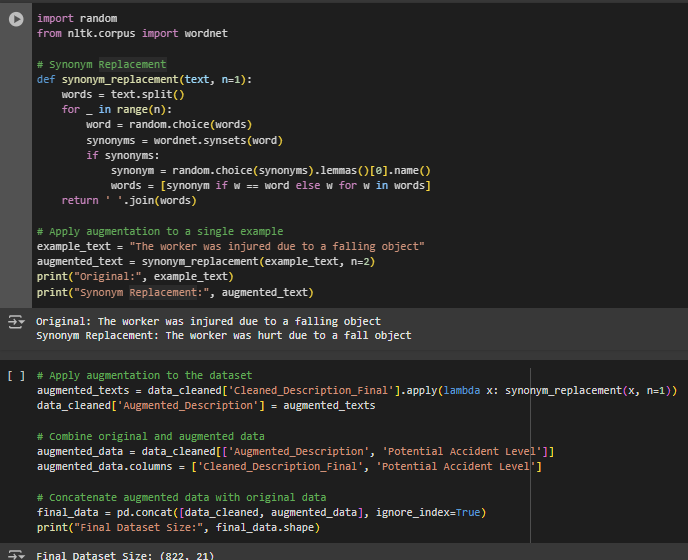
3. Logistic Regression

* **Best Parameters**: {'C': 10, 'solver': 'lbfgs'}
  + **C**: A larger value indicates less regularization, which helps the model fit the data better.
  + **solver**: The lbfgs optimisation algorithm is well-suited for multi-class problems.
* **Best Cross-Validated Score**: **70.12%**
  + Logistic Regression remains a reliable baseline model but underperforms compared to SVM.

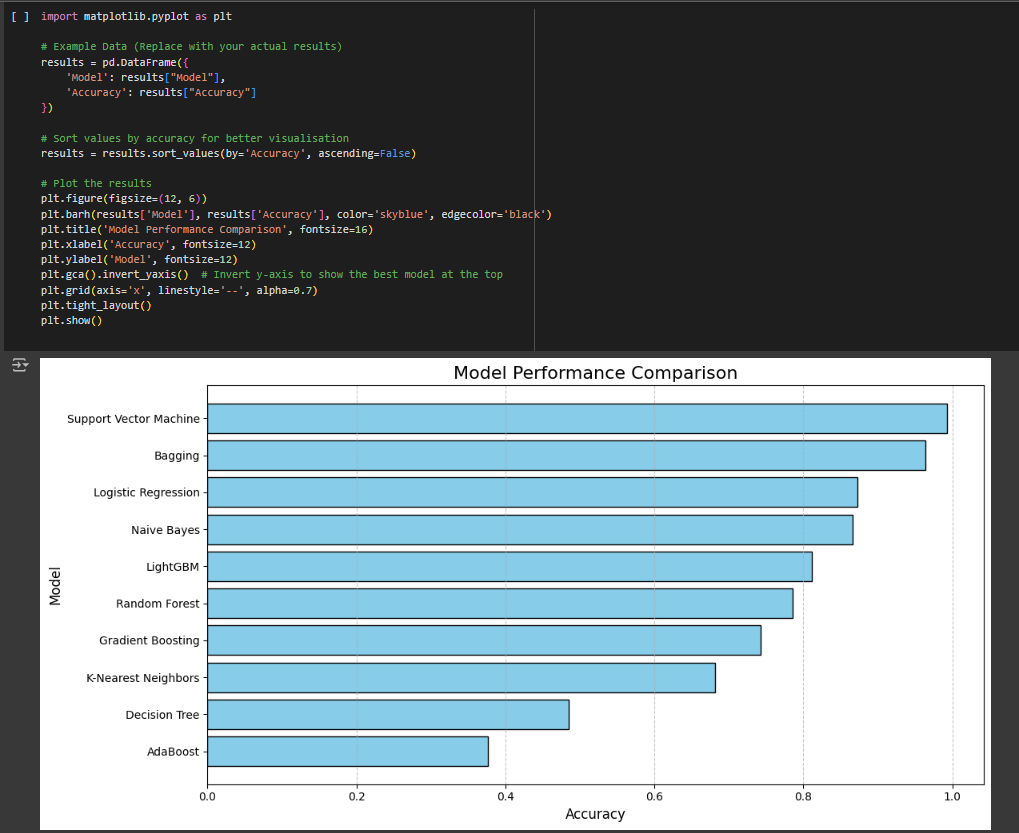
### Next Steps

Since Our accuracy is low we are planning for data augmentation using synonym replacement

## SYNONYM REPLACEMENT



1. **Improved Model Generalization:**
   * Augmenting text data increases the variety of training samples, preventing overfitting and making the model more robust to unseen data.
2. **Data Scarcity Solution:**
   * For **small datasets**, augmentation provides a way to synthetically increase data size without collecting new samples, which is cost-effective.
3. **Better Performance for NLP Tasks:**
   * Text classifiers often benefit from more diverse data, especially in tasks like **accident classification** or **incident reporting** (as in this dataset).



# KEY INSIGHTS AFTER DATA AUGMENTATION

## TOP PERFORMERS

* Support Vector Machine (SVM): Achieved the highest accuracy (98.91%), leveraging the augmented dataset effectively.
* Bagging: Strong performance with 95.29%, indicating ensemble methods benefit from diverse data.

## Consistent Performers

* Naive Bayes: Improved to 88.77%, maintaining adaptability to text data.
* Logistic Regression: Reliable at 86.23%, showing robustness even with augmented data.

## Moderate Performers

* LightGBM: Improved to 83.70%, but still behind top models.
* Random Forest: Scored 77.17%, reflecting its ensemble power but less adaptability to sparse data.

## Recommendations[¶](https://htmtopdf.herokuapp.com/ipynbviewer/temp/d4ebad3699bcc00db467ee1ddee8dbb2/Capstone_NLP_Group7_Final_File.html?t=1734882824647#Recommendations)

* Deploy SVM as the primary model with Bagging as a backup.
* Validate top models on real-world data to ensure robustness.