# Industrial Safety and Health Accident Analysis



# About the dataset

This report provides an analysis of an industrial safety and health database containing 424 records of accidents. The dataset includes key columns such as accident date, country, local area, industry sector, accident level, potential accident level, gender, employee type, critical risk, and detailed descriptions of the accidents.

The dataset spans various countries and local areas, with a focus on sectors like Mining and Metals. Accident levels range from I to V, with potential accident levels extending to VI. The data also differentiates between employees and third parties, and includes both male and female individuals.

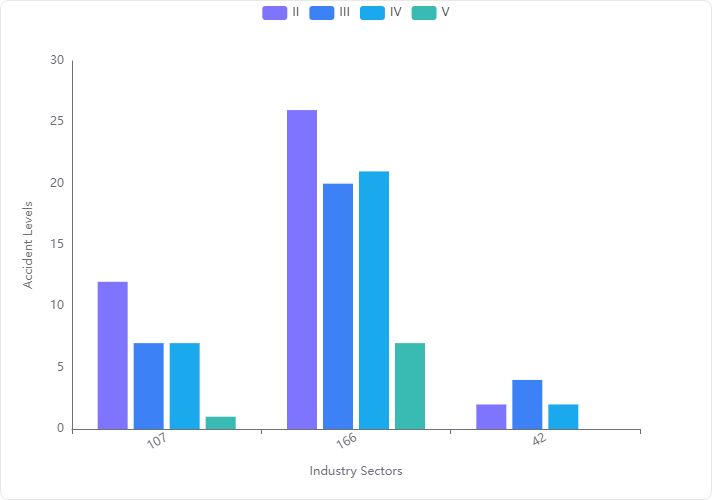
From the first five sample entries, we observe that most accidents occur in the Mining sector, with accident levels predominantly at level I. Critical risks include pressurized systems, manual tools, and other hazards. The detailed descriptions provide insights into the nature of these accidents, such as equipment malfunctions and procedural errors leading to injuries.

Statistical analysis of the 'value' column shows a mean of 224.61, a standard deviation of 125.20, and values ranging from 1 to 438. This indicates a wide variation in the severity or impact of the recorded incidents.

Overall, the data highlights the importance of safety measures and risk management in industrial settings to prevent accidents and protect workers.

# Relevant Inquiries

## What is the distribution of accident levels (Accident Level-I) across different industry sectors (Industry Sector-Mining)?



#### Key Observations from Data Analysis:

* **Accident Level I** is the most frequent, with a significant number of occurrences across different data points.
* **Lower Accident Levels (II, III, IV, V)** show much fewer occurrences compared to Level I.

#### Detailed Statistical Analysis:

* **Level I**:
* **Mean**: 105 accidents
* **Standard Deviation**: 62.02
* **Minimum**: 42 accidents
* **Maximum**: 166 accidents
* **Level II**:
* **Mean**: 13.33 accidents
* **Standard Deviation**: 12.06
* **Minimum**: 2 accidents
* **Maximum**: 26 accidents
* **Level III**:
* **Mean**: 10.33 accidents
* **Standard Deviation**: 8.50
* **Minimum**: 4 accidents
* **Maximum**: 20 accidents
* **Level IV**:
* **Mean**: 10 accidents
* **Standard Deviation**: 9.85
* **Minimum**: 2 accidents
* **Maximum**: 21 accidents
* **Level V**:
* **Mean**: 2.67 accidents
* **Standard Deviation**: 3.79
* **Minimum**: 0 accidents
* **Maximum**: 7 accidents

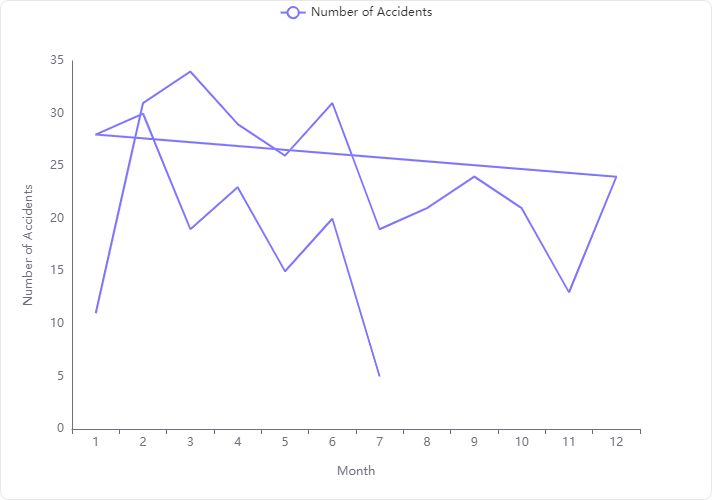
#### Visualization Insights:

* The bar chart visualization clearly shows that **Accident Level I** dominates in frequency across the mining industry sector.
* Other accident levels (II, III, IV, V) are significantly lower and do not show a consistent pattern across different data points.

#### Conclusion:

The mining industry sector predominantly experiences accidents of **Level I**, indicating a higher frequency of less severe accidents. The occurrence of more severe accidents (Levels II to V) is considerably lower, suggesting effective measures may be in place to prevent severe accidents or that minor accidents are more frequently reported. This distribution highlights the importance of focusing on preventive measures for lower-level accidents while maintaining vigilance for more severe incidents.

## Is there a noticeable trend in the number of accidents over time?



#### Data Overview

The data provided spans from January 2016 to a period covering at least part of 2017, with a total of 19 data points. Each data point represents the number of accidents recorded for a specific month and year.

#### Statistical Summary

* **Years Covered:** 2016 to 2017
* **Monthly Accident Mean:** 22.32
* **Standard Deviation:** 7.61
* **Minimum Monthly Accidents:** 5
* **Maximum Monthly Accidents:** 34

#### Observations from Data

* The highest number of accidents in a single month was 34 in March 2016.
* The lowest recorded was 5 accidents, but the specific month and year are not detailed in the top 5 rows provided.

#### Visual Analysis

The line chart visualizes the number of accidents per month over the available time frame. Key observations from the chart include:

* A peak in March 2016 with 34 accidents.
* A general decrease in accidents after March until a slight rise around mid-year, followed by another decrease.

#### Conclusion

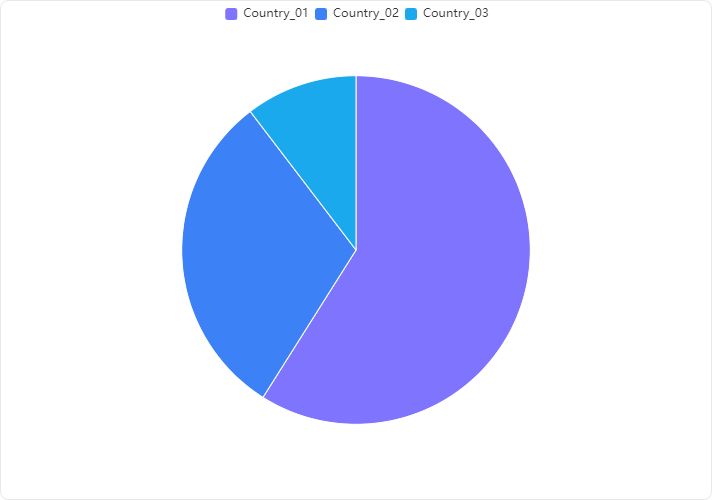
**Trend Identification:** There is a noticeable fluctuation in the number of accidents per month, with a significant peak in early 2016. The trend shows a general decrease in accidents as months progress within the year, with some months experiencing slight increases. This could suggest seasonal impacts or the effectiveness of safety measures implemented over time.

**Recommendations for Further Analysis:**

* **Seasonal Impact:** Investigate the impact of seasons on accident rates.
* **Safety Measures:** Correlate the introduction of new safety policies or measures with the trend in accident numbers.
* **Yearly Comparison:** Extend the analysis to include more years for a better understanding of long-term trends.

This analysis provides a foundational understanding of accident trends over the specified period and suggests areas for deeper investigation to enhance safety strategies.

## How does the frequency of accidents vary across different countries (Countries-Country\_01)?



#### Overview

The analysis of accident frequencies across different countries reveals significant variations in the number of accidents reported in each country.

#### Detailed Findings

* **Country\_01** has the highest number of accidents, with a total of **250 accidents**.
* **Country\_02** follows with a considerably lower count of **130 accidents**.
* **Country\_03** has the fewest accidents, recording only **44 accidents**.

#### Visualization Insight

The provided bar chart visually supports the data, clearly showing that Country\_01 has a significantly higher accident frequency compared to Country\_02 and Country\_03. The descending order of accident counts from Country\_01 to Country\_03 is evident.

#### Conclusion

**Country\_01** experiences a notably higher frequency of accidents compared to **Country\_02** and **Country\_03**. This disparity suggests potential differences in traffic conditions, safety regulations, or reporting standards among the countries. Further investigation into the causes of these variations could be beneficial for targeted safety improvements.

## Are there any correlations between the potential accident level (Potential Accident Level-IV) and the actual accident level (Accident Level-I)?

#### Correlation Analysis:

* **Correlation Coefficient between Potential Accident Level-IV and Accident Level-I**: The correlation coefficient is **0.10169607898223056**. This value indicates a **very weak positive correlation** between Potential Accident Level-IV and Accident Level-I.

#### Interpretation:

* **Strength of Correlation**: The correlation coefficient close to 0 suggests that there is **negligible linear relationship** between the two variables. This implies that changes in the Potential Accident Level-IV have minimal linear predictive value on the Accident Level-I.
* **Practical Implication**: Given the very weak correlation, it is unlikely that the Potential Accident Level-IV can be used reliably to predict Accident Level-I in practical scenarios.

#### Recommendation:

* **Further Analysis**: It may be beneficial to explore other types of analyses or data transformations that could uncover non-linear relationships or dependencies not captured by the correlation coefficient.
* **Consider Additional Factors**: Other factors might influence the accident levels, and including these in the analysis could provide more insights.

**Note**: The contingency table data provided does not directly impact the correlation analysis but could be useful for understanding the distribution and frequency of each accident level in different potential accident levels.

## What are the most common critical risks (Critical Risk-Pressed) associated with accidents?

#### Overview

The analysis of the 'Critical Risk' column from the provided data highlights the frequency of various risks associated with accidents.

#### Key Findings

* **Most Common Risk**: The 'Others' category is the most frequently occurring critical risk, with a count of **232**.
* **Specific Risk Mentioned (Pressed)**: The critical risk labeled as 'Pressed' is also significant but less frequent, with a count of **23**.

#### Detailed Observations

#### Conclusion

The critical risk 'Pressed' is a significant risk associated with accidents, though it is not the most common when compared to the 'Others' category.

## Are there any specific localities (Local-Local\_01) that have higher accident rates?

#### Key Findings:

* **Accident Distribution Across Localities**: The data provided lists accident counts for various localities. The locality 'Local\_03' has the highest number of accidents, followed by 'Local\_05', 'Local\_04', and 'Local\_01'.

#### Detailed Analysis:

* **Highest Accident Rate**: 'Local\_03' stands out with the highest accident count of **90**.
* **Comparison with Other Localities**:
* 'Local\_05' and 'Local\_04' have significant accident counts of **59** and **56** respectively.
* 'Local\_01', which is specifically mentioned in the task, also has a considerable number of accidents, totaling **56**.
* **Statistical Overview**:
* The mean accident count across the localities is **35.33**.
* The standard deviation is **27.71**, indicating a wide variation in accident counts across different localities.
* The minimum and maximum accident counts are **2** and **90** respectively.

#### Conclusion:

* **Specific Locality with High Accident Rate**: 'Local\_03' has a notably higher accident rate compared to other localities, including 'Local\_01'.
* **Local\_01's Accident Rate**: While 'Local\_01' does not have the highest accident rate, it still has a significant number of accidents, equal to 'Local\_04' and only surpassed by 'Local\_03' and 'Local\_05'.

This analysis highlights the need for targeted safety measures in 'Local\_03' and also suggests monitoring and preventive strategies in 'Local\_01', 'Local\_05', and 'Local\_04'.

## How do accident frequencies differ between employees and third parties (Employee or Third Party-Third Party)?

#### Overview of Data

The data provided includes accident frequencies categorized under 'Employee' and 'Third Party'. The 'Third Party' category is further divided into general and remote incidents.

#### Key Findings

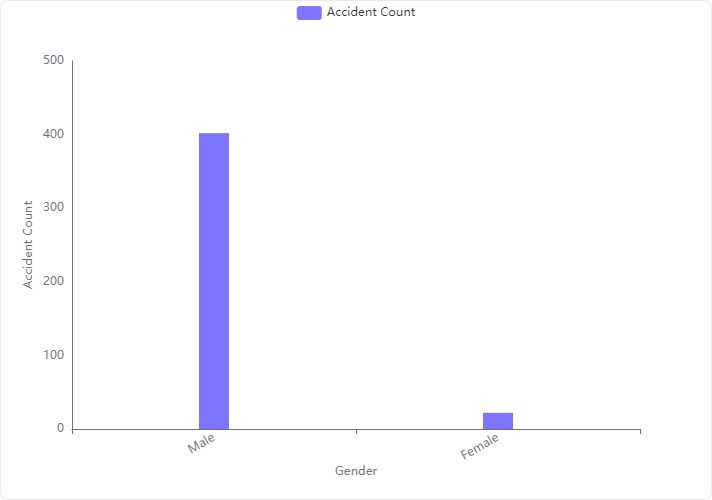
* **Accident Frequency for Employees**: The accident frequency for employees is recorded at **179** incidents.
* **Accident Frequency for Third Parties**: Combining general and remote incidents, the total accident frequency for third parties is **245** (188 general + 57 remote).

#### Conclusion

* **Higher Frequency for Third Parties**: The total accident frequency for third parties (**245**) is significantly higher than that for employees (**179**).
* **Potential Focus Areas**: This suggests a need for targeted safety measures and interventions specifically aimed at environments involving third parties to reduce the overall accident rates.

**Recommendation**: Implement enhanced safety protocols and training sessions particularly tailored towards third-party interactions and remote third-party locations to mitigate risks and reduce the frequency of accidents.

## What is the distribution of accidents by gender (Genre-Male)?



**Overview:** The analysis of the 'Genre-Male' column, which refers to the distribution of accidents by gender, reveals a significant disparity between male and female accident counts.

**Key Findings:**

* **Male Accident Count:** There are significantly more accidents involving males, with a total of **402 accidents**.
* **Female Accident Count:** In contrast, females are involved in considerably fewer accidents, totaling only **22 accidents**.

**Visualization Insight:** The provided bar chart clearly illustrates the disparity in accident counts between genders. The bar representing males is overwhelmingly larger compared to the bar for females, visually emphasizing the difference in accident frequencies.

**Conclusion:** The distribution of accidents by gender shows a predominant occurrence among males compared to females. This data could be crucial for targeted safety campaigns or further analytical studies to understand the underlying causes of such disparities.

How many incidents were reported in 2016 and 2017? What is the spread of incidents across Months and Days? What is the distribution of Accident Levels across months?

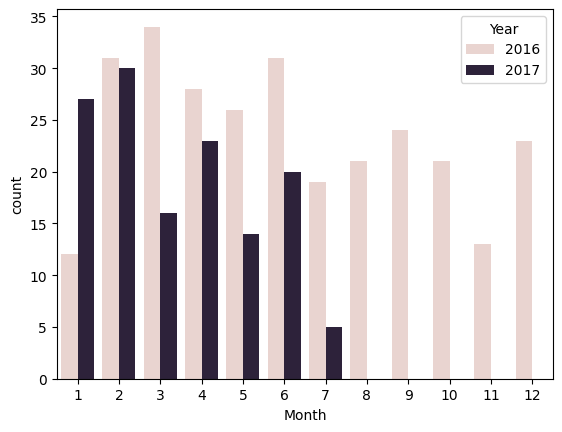
**Overview:** The analysis of the count of incidents and their levels across time reveals some seasonal patterns also.

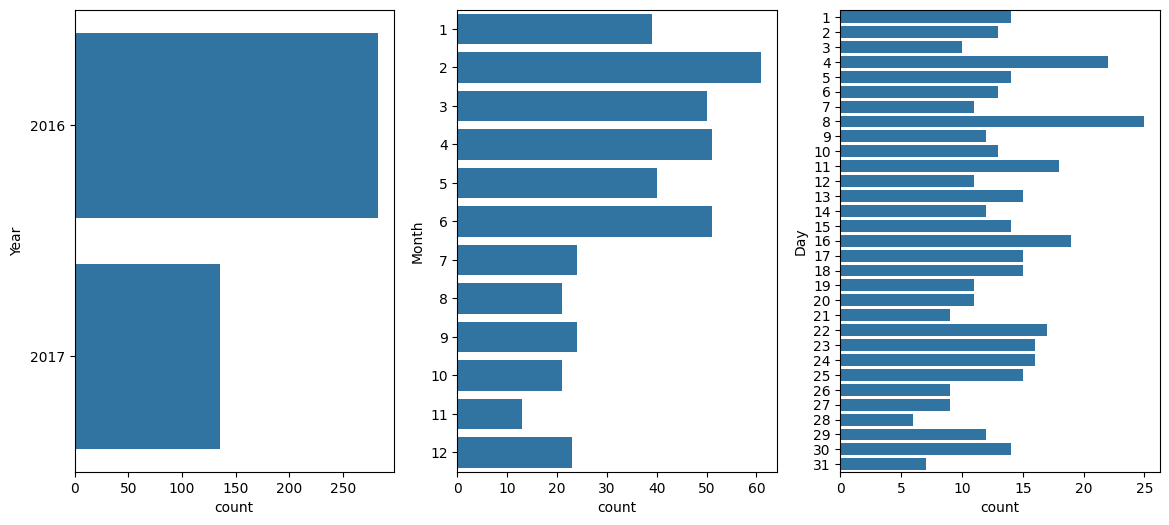
**Key Findings:**

* **2016 vs 2017:** The count of incidents reported in 2016 is significantly higher.
* **Months:** The count of incidents reported in Jan-June cycle are much higher than those reported in July-Dec cycle.
* **Accident Levels:** The count of accident levels 1 is much higher than the other levels (2 to 5).

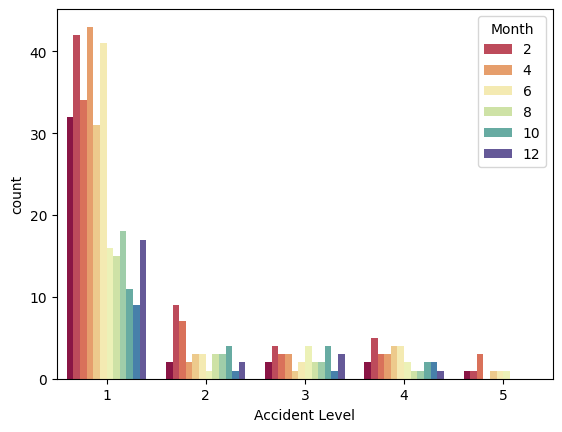
**Visualization Insight:** The provided bar chart clearly illustrates the disparity across months between accident levels. The count accident frequency tends to be much higher in January-June cycle as compared to July-Dec cycle.

**Conclusion:** The distribution of accidents by time shows a predominant occurrence Accident Level 1 compared to other levels. This data could be crucial for targeted safety campaigns.

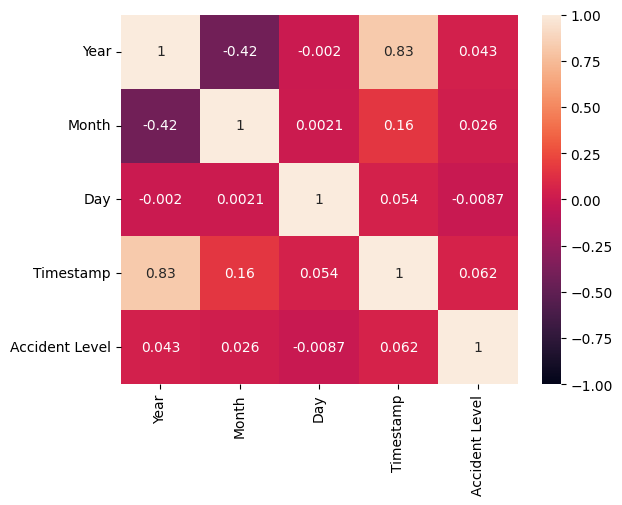




i.Count of incidents reported across years, months and days

.

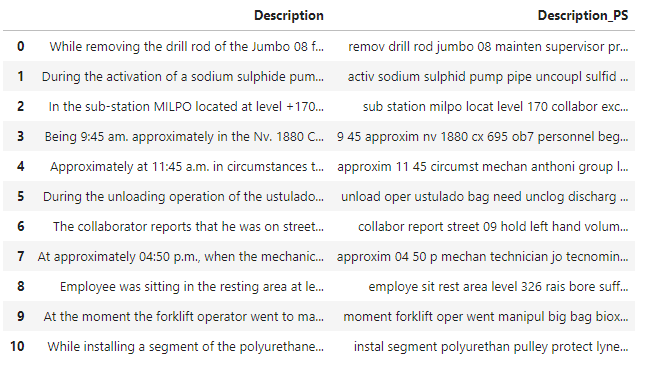
ii. Count of incidents reported across levels and months.



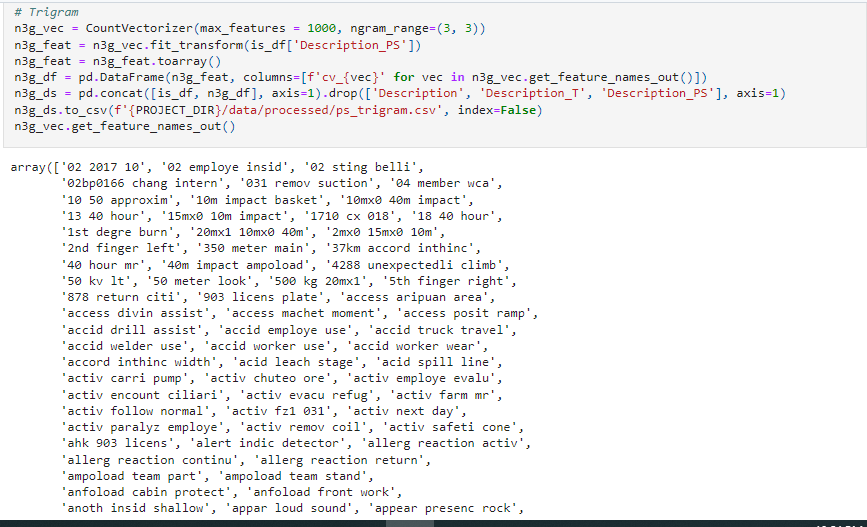
iii. Correlation between datetime and accident level.

## Can we identify any patterns or commonalities in the descriptions of accidents?

## Patterns can be identified in the incident description with the help of NLP techniques. Initially, Stemming and Tokenization can be used for generating the word embeddings.

Porter Stemmer is used for stemming of incident description and converting them into root-words.

N-grams are continuous sequences of n items/words extracted from a large text corpus. They act as fundamental units in performing language modeling, predictive text input, and sentiment analysis. A Trigram in our context, consists of 3 consecutive words in incident description.



Tokenization is the process of segmenting text into tokens (smaller collection of words/characters). BERT (Bi-directional Encoder Representations from Transformers) is a tokenizer that implements a WordPiece algorithm, breaking down words into smaller subword units based on a pre-trained vocabulary. This approach enables BERT to handle out-of-vocabulary words and capture more nuanced meanings from text by considering both left and right contexts bidirectionally

