**PROJECT NAME**

**SAFEX.PYBOTS**

Exchanging of Industrial Safety using NLP based Python Bots

**TEAM NAME**

**SAFEX.PYBOTS**

**TEAM MEMBERS:**

**1) APOORV DIXIT**

**2) DIVYASHREE C**

**3) PRINCE**

**4) SRINATH**

**5) SRINAYANI**

**INTERIM REPORT OF CAPSTONE PROJECT ON NLP2 – CHATBOT**

**PROJECT OBJECTIVE:**

Design and build chatbot by employing ML/DL techniques which can help the professionals in determining the potential level and accident level involved in any accident and to highlights the safety risk as per the incident description.

**PROBLEM STATEMENT AND ABSTRACT:**

For enhancing one’s self-esteem, wellbeing, and social mobility, work is very important. However, during work activities if any accidents occurred leads to impairments to worker’s health and which in-turn leads to serious social and economic repercussions.

Globally, it is estimated between the range of 1.8% to 6.0% cost of work-related accidents and ill-health out of gross domestic product. It is also estimated as, around 2.3 million people will die in a year around the world due to work-related activities. Specifically, in Brazil, around 2500 such deaths per year which corresponds to one death for every 3.5hr.

Since human activity is involved, accidents & injuries are common despite all the safety measures and precautions put in place. Such injuries can also prove fatal. Industrial accidents can turn depending on the type of industry. For example, a mere spark in a firecracker factory can burn the whole plant leading to loss of lives and property. Workplace injuries are a big concern for both workers and management. It is imperative to classify industrial incidents into different categories and determine whether the event was merely an accident, due to negligence or by incompetence. This avoids reoccurrences, reduce frequency of occurrence & severity and minimize the effects. To achieve this, we employ exploratory data analysis on a dataset from one of the biggest Brazilian industries and find out the top reasons for industrial accidents, nature of accidents, type of employees being injured and so on. We also aim to develop a chatbot application using natural language processing to classify the accident into various critical risks by looking at the description of the accident.

**DOMAIN:**

Industrial safety. NLP based Chatbot.

**DATA DESCRIPTION:**

This The database is basically records of accidents from 12 different plants in 03 different countries which 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

**LINK FOR THE DATASET**

[**Industrial Safety and Health Analytics Database | Kaggle**](https://www.kaggle.com/datasets/ihmstefanini/industrial-safety-and-health-analytics-database)

**DATA INTERPRETATION**

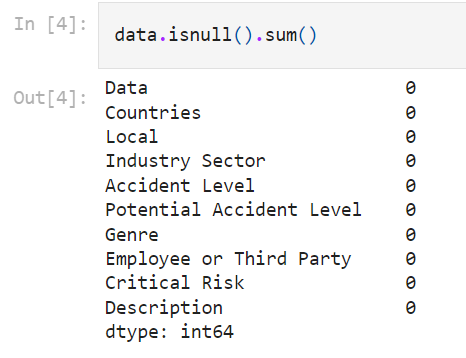
Based on the entire dataset -

1. Shape and Top 5 details



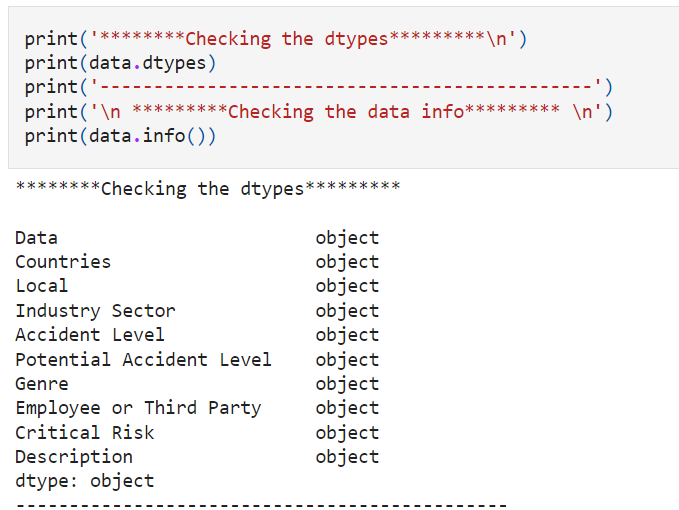
The dataset contains 425 instances and 11 attributes

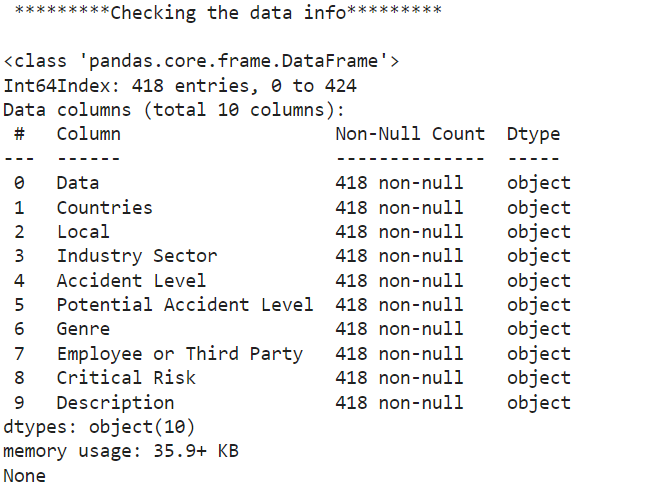
1. Missing Values



There are NO NULL values

1. Checking of Duplicates and Drop if any-
2. Checking of dtypes and data info





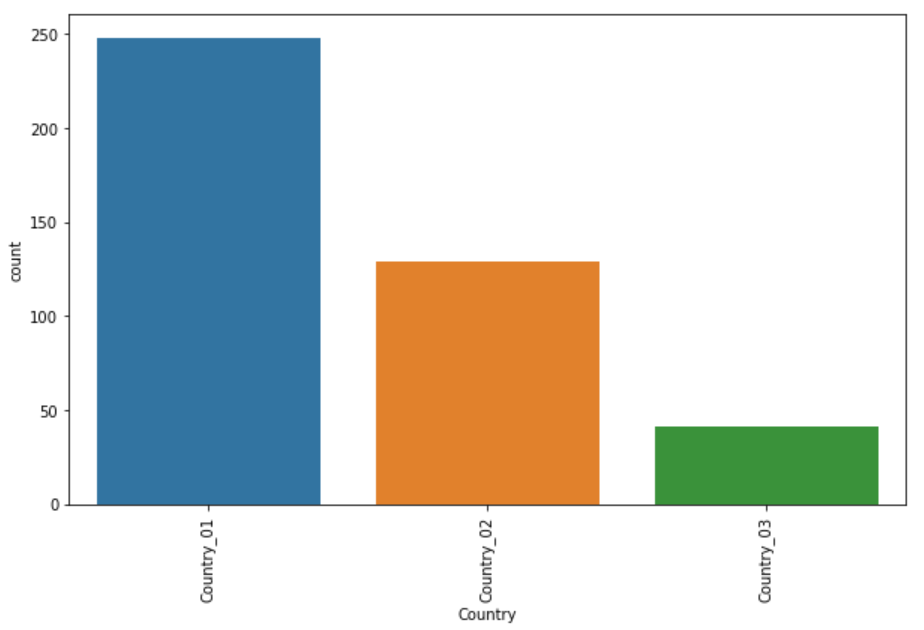
**UNIVARIATE ANALYSIS:**

1. COUNTRY

COUNT PLOT-

It is used to show the counts of observations in each categorical bin using bars. For instance, the count plot () method is used to display the number of accidents happened in country\_01, country\_02 and country\_03 for the country variable.

The following are the observations that are made for the country count plot ()-



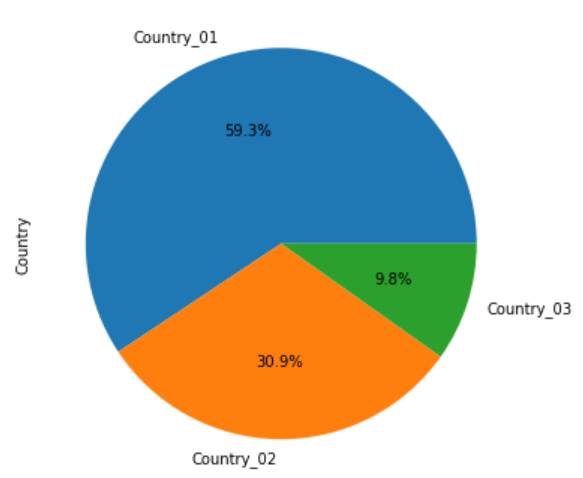
1. About 250 accidents have happened in the country \_01.

2. About 125 accidents have happened in the country\_02 and about 50 accidents happened in country\_03.

3. The maximum number of accidents happened in Country\_01 and least number of accidents happened in Country\_03.

PIE CHART-

A pie chart represents data in a circular graph containing slices of different colours. It is used to study the proportion of numerical data. It shows the proportion of data as a percentage of a whole. For instance, the pie chart for the country variable gives the number and percentage of accidents that have occurred in the country\_01, country\_02 and country\_03.



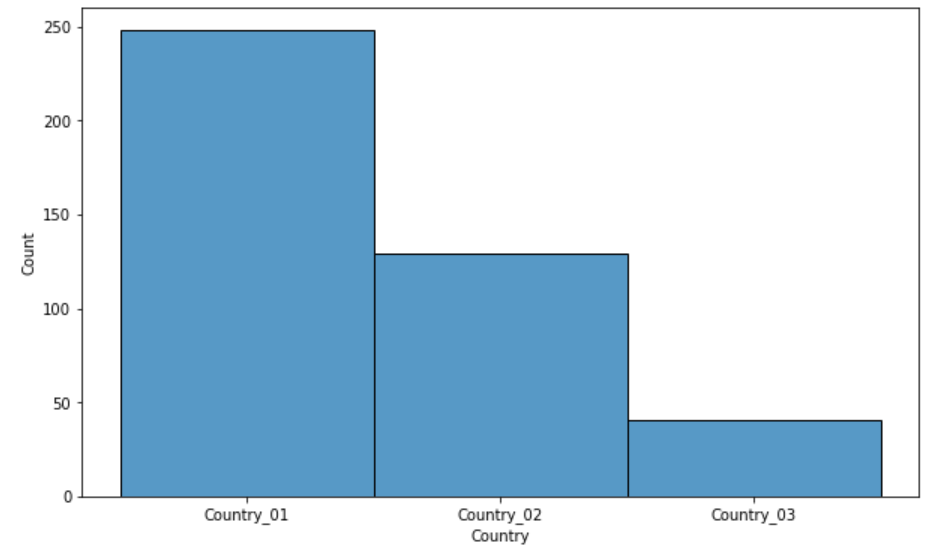
OBSERVATIONS-

1. The number of maximum accidents have taken place in country\_01. i.e.- 59.3%.

2. The least number of accidents took place in country\_03. i.e- 9.8% .

HISTPLOT-

It is used to show the distribution of the datasets. For instance, for the variable country it displays the number of accidents that happened in the country\_01, country\_02 and country\_03.

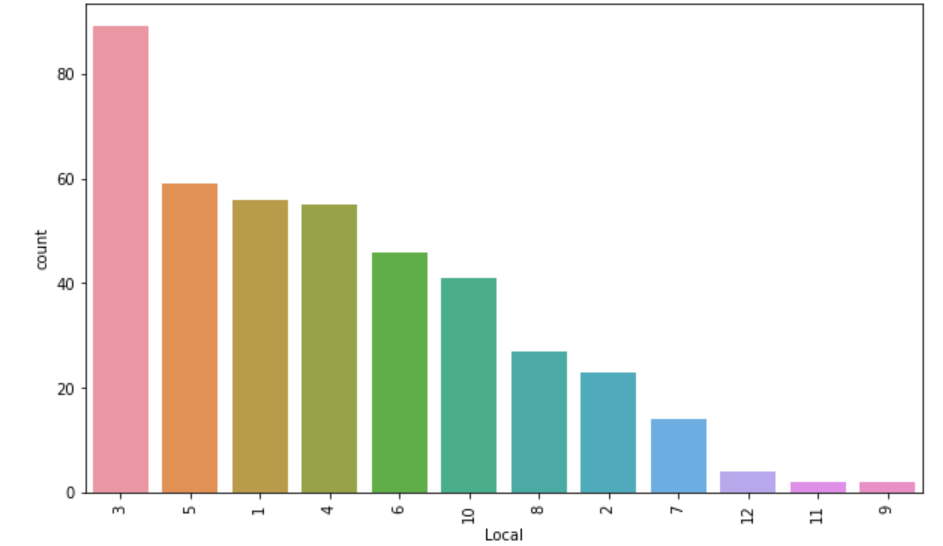


OBSERVATIONS- It can be noted that country\_01 has the maximum number of accidents with a count of 248 and the minimum number of accidents has happened in country\_03 with a count of 41.

2. LOCAL-

COUNT PLOT-

The local variable determines the region of the accident. The countplot() method helps to calculate the number of accidents happened in region wise.



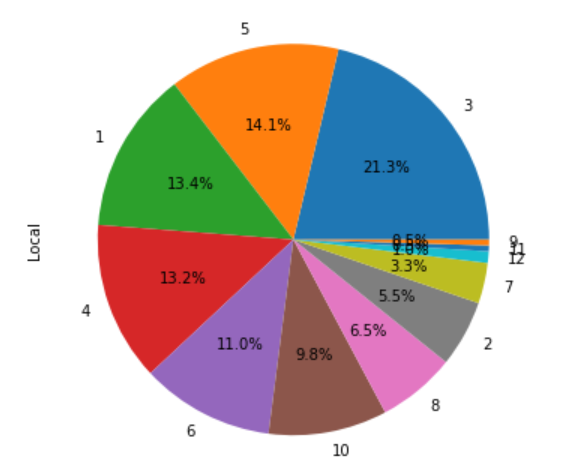
OBSERVATION-

1. It is observed that in the local\_03 has the maximum number of accidents with a count of about 90.

2. The least number of accidents happened in the local\_09 region.

PIECHART-

The pie chart for the local variable gives the number and percentage of accidents that have occurred in the local\_01, local\_02 and so on upto local\_11



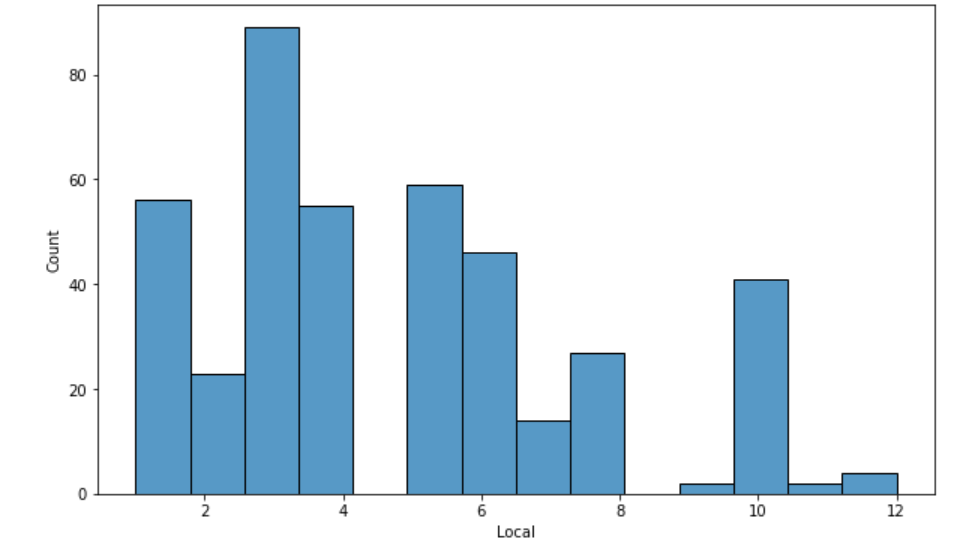
OBSERVATIONS-

1.From the pie chart it can be observed that at the local \_03 maximum number of accidents have taken place with about 21.3%.

2. The least number of accidents have happened in local\_09. i.e-0.5%

HISTPLOT-

The histplot() of the local variable is useful to determine the distribution.



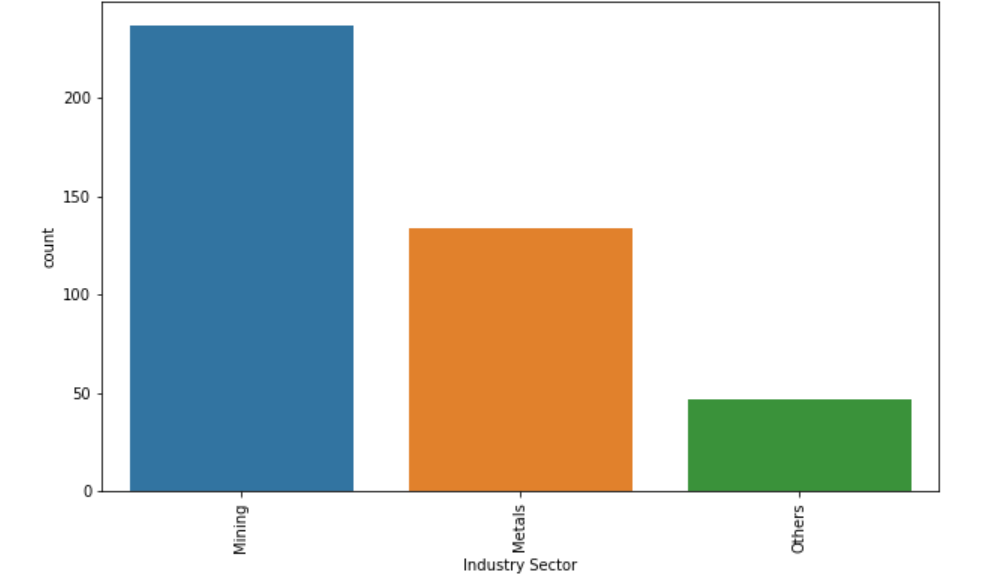
OBSERVATIONS-

1. It is observed that the number of accidents at local\_03 is about 90.

3.INDUSTRY SECTOR-

COUNTPLOT-

A countplot() method in industrial sector is used to determine the number of accidents that had happened due to different industrial sectors such as mining, metals and others.

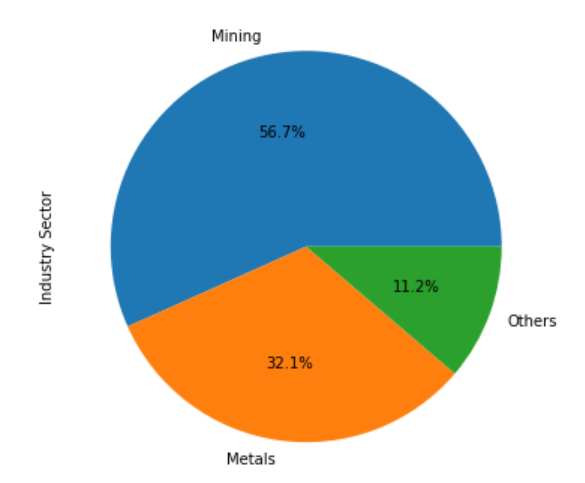


OBSERVATIONS-

It can be determined that maximum number of accidents happened due to mining sector (237) than metals and others.

PIECHART-

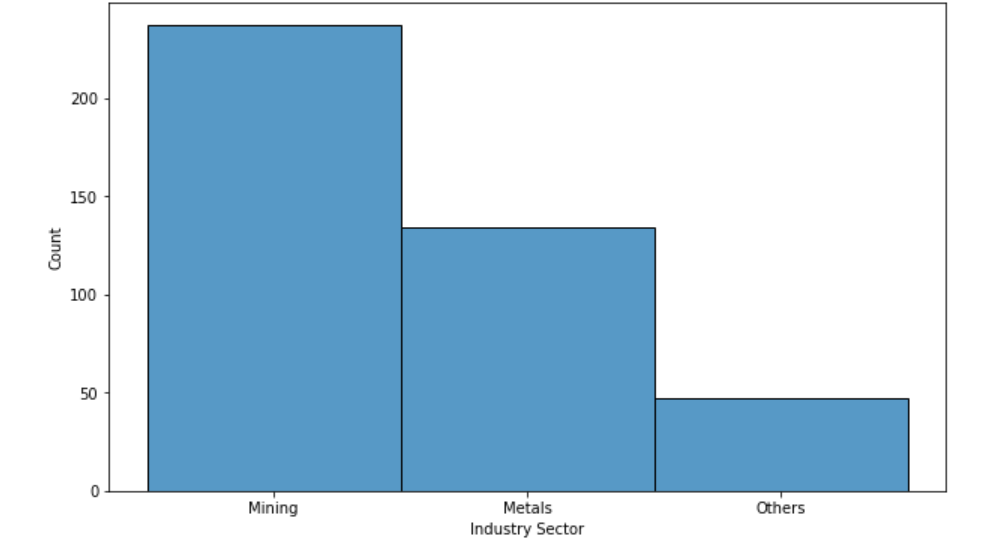
A piechart is used to determine the percentage of accidents that were caused due to mining, metals and other industrial sector.



OBSERVATIONS- It can be concluded that the maximum number of accidents happened due to mining i.e- 56.7% of the total accidents.

HISTOGRAM-

A histplot is used to determine the distribution of different industrial sectors.

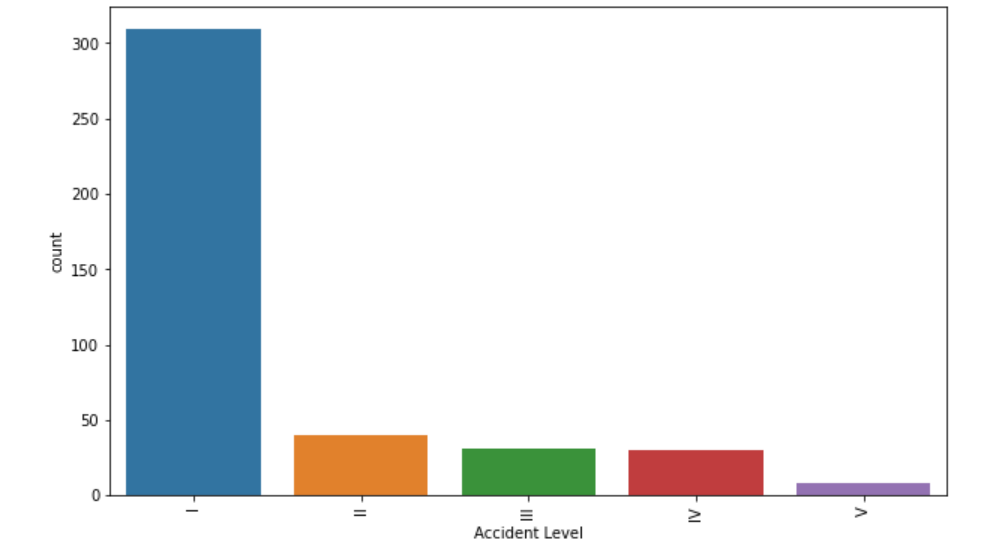


OBSERVATIONS- It can be noted that maximum number of accidents happened due to mining sector.

4.ACCIDENT LEVEL-

The accident level provided the information about the severity of the accident. Where, I represents less severity and V represents high severity.

COUNTPLOT- The countplot() of the accident level determines the number of accidents that has happened at level I, II, III, IV, and V.



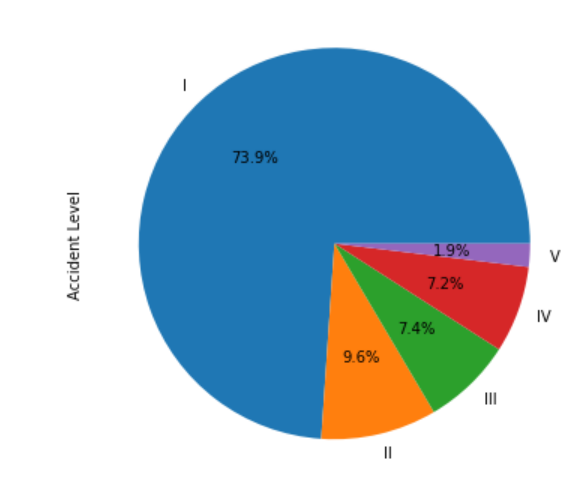
OBSERVATIONS-

1. It can be observed that the number of accidents happened at level I (300)are higher than level V(less than 50).

2. Hence, it can be concluded that maximum number of less severe accidents took place.

PIE CHART-

The pie chart helps to find out the percentage of accident level I, II, III, IV and V accidents out of all the accidents.

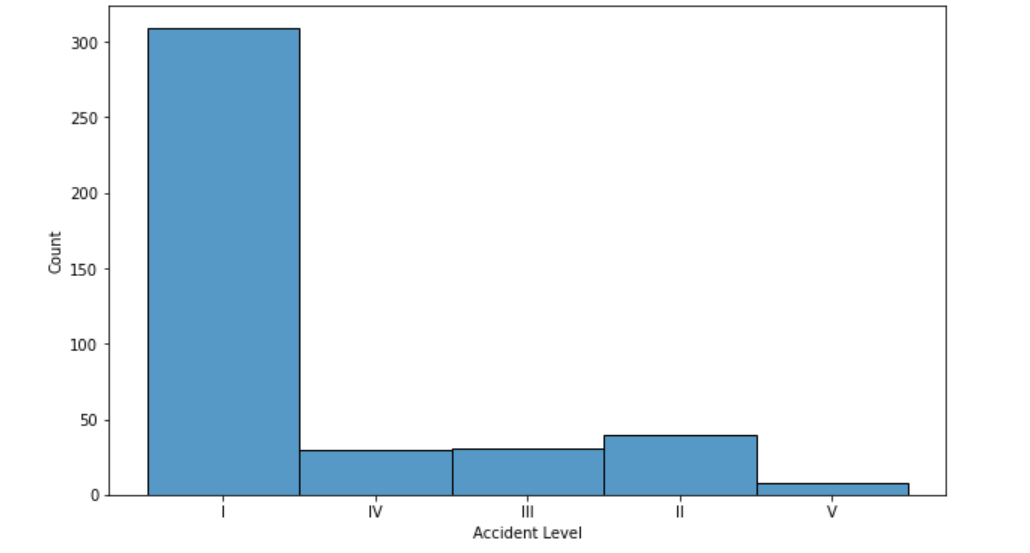


OBSERVATIONS-

1. It can be noted that maximum percentage of less severe accidents took place. i.e accident level I with 73.9%.

HISTOGRAM-

The histogram plot could be used to plot the distribution of accident level I, II, III, IV and V.

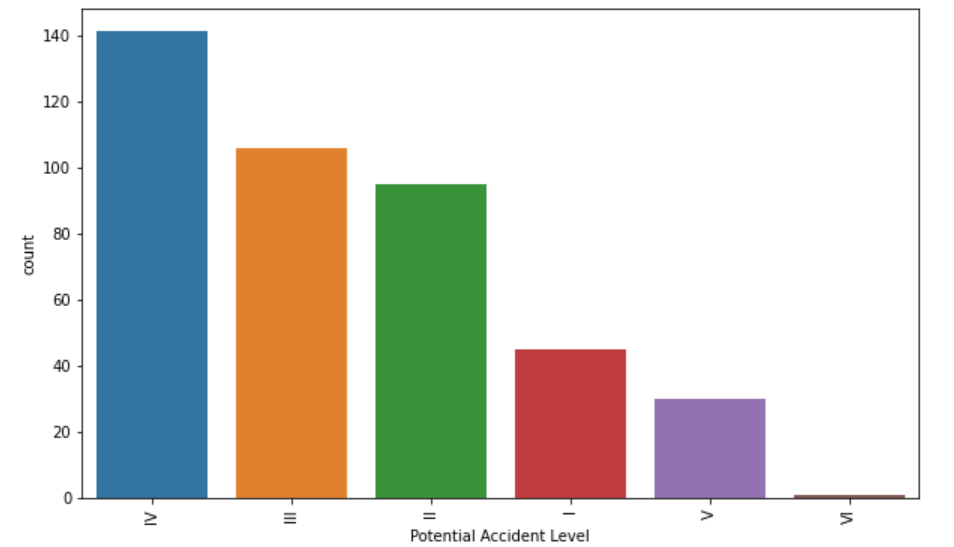


OBSERVATIONS- The number of accidents with less severity(309) are higher than more severity(8) with a huge difference.

POTENTIAL ACCIDENT LEVEL-

The potential accident level is based on the accident level with other factors. Here, Potential accident level I is less severe and VI is more severe.

COUNT PLOT- It determines the number of potential accident level caused by all levels.

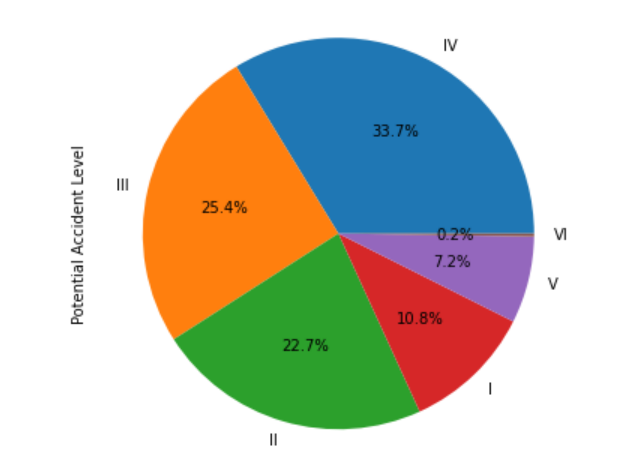


OBSERVATIONS-

The number of accidents due to potential accident level IV are maximum with a count of 141.

PIECHART-

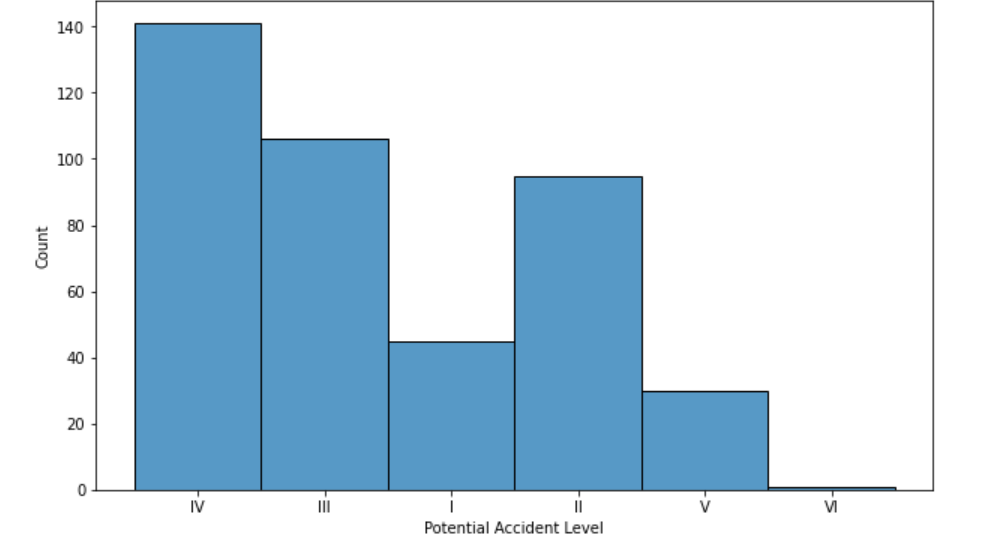
The pie chart determines the percentage of potential accidents caused by level I to VI.



OBSERVATIONS- It can be determined that the percentage and number of accidents caused due to potential accident level IV are maximum.

HISTOGRAM-

It gives the distribution of the potential accident level.

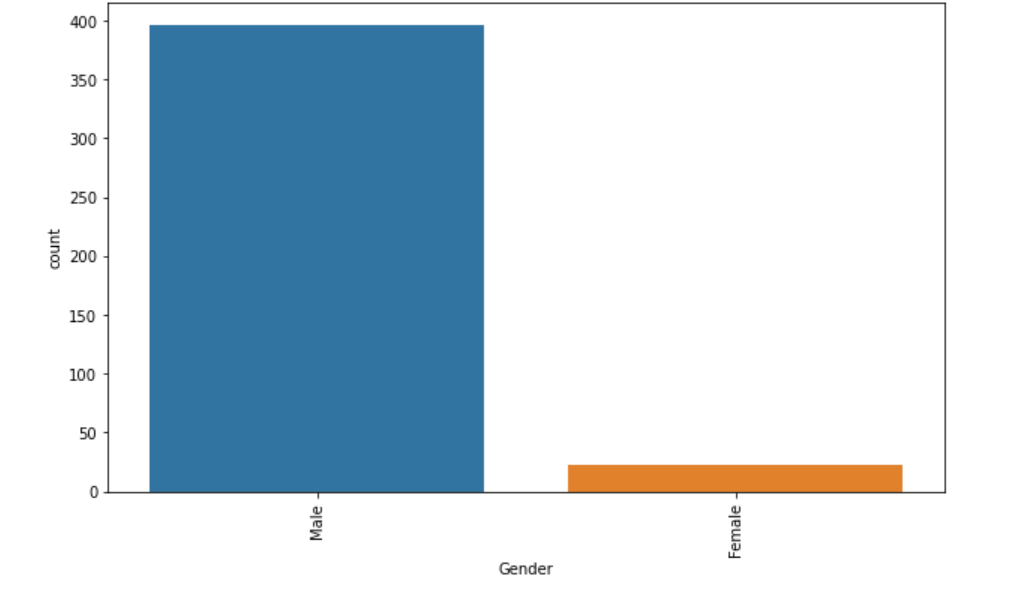


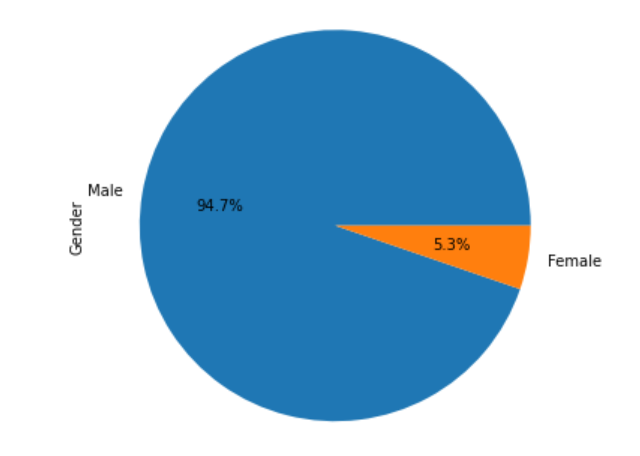
OBSERVATIONS-

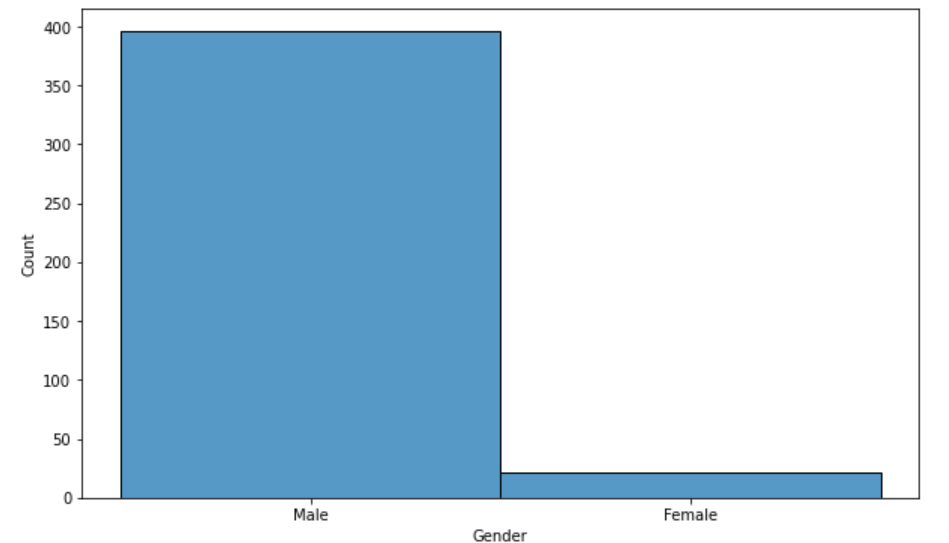
It can be determined that the number of accidents caused by potential accident level IV are maximum.

5. GENDER-

The gender is used to determine the number of accidents caused to both female and male.



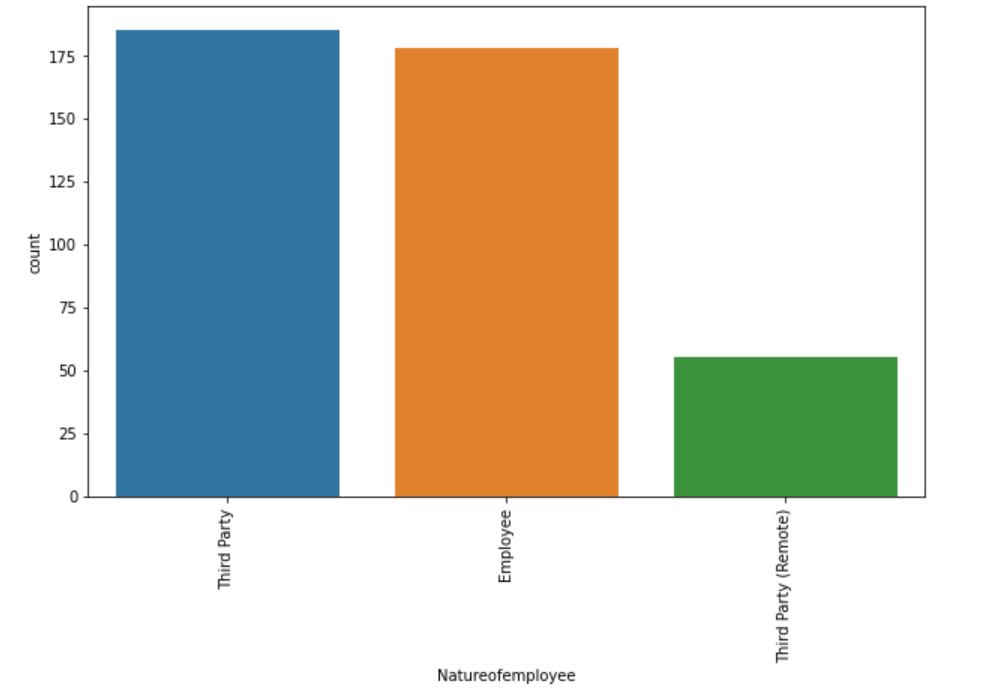




6. NATURE OF EMPLOYEE-

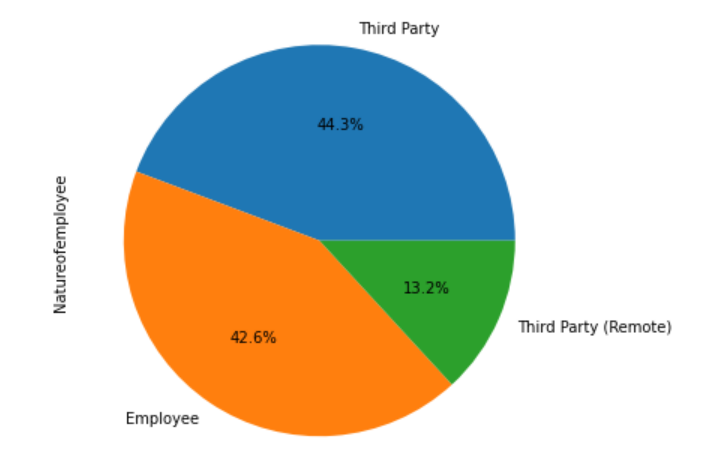
The nature of employee is divided into 3 types that is either third party, engineering and third party (remote).

COUNT PLOT()- It is used to determine the number of accidents happened according to the nature of the employee such as third party, engineering and third party (remote).



OBSERVATIONS-

It can be noted that the third party type of employee has faced many accidents and third party(remote) has faced the least accidents.

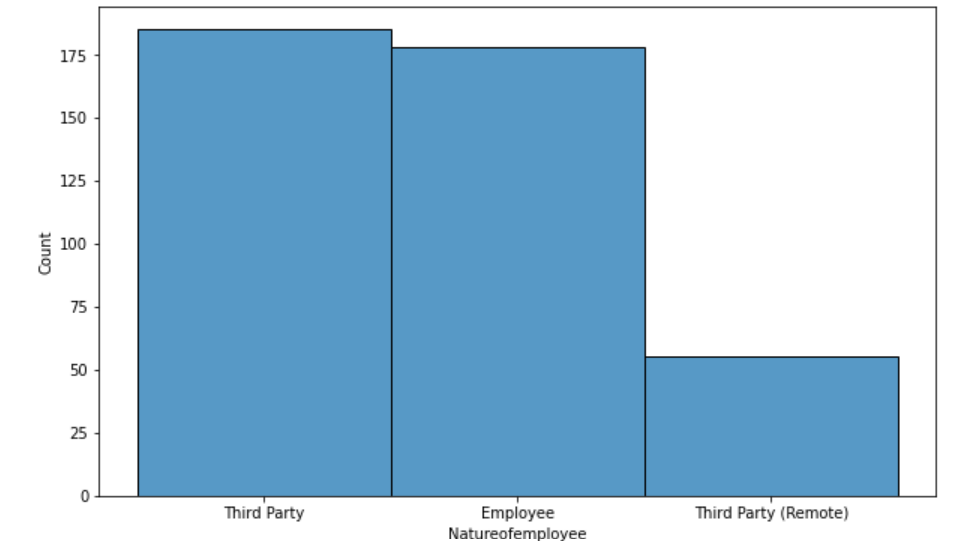
PIE CHART-

OBSERVATIONS-

It can be noted that the third party employee has happened to face most of the percentage of accidents. i.e- 44.3%.

HISTOGRAM-

It is used to show the distribution of the nature of employee such as third party, employee and third party(remote).

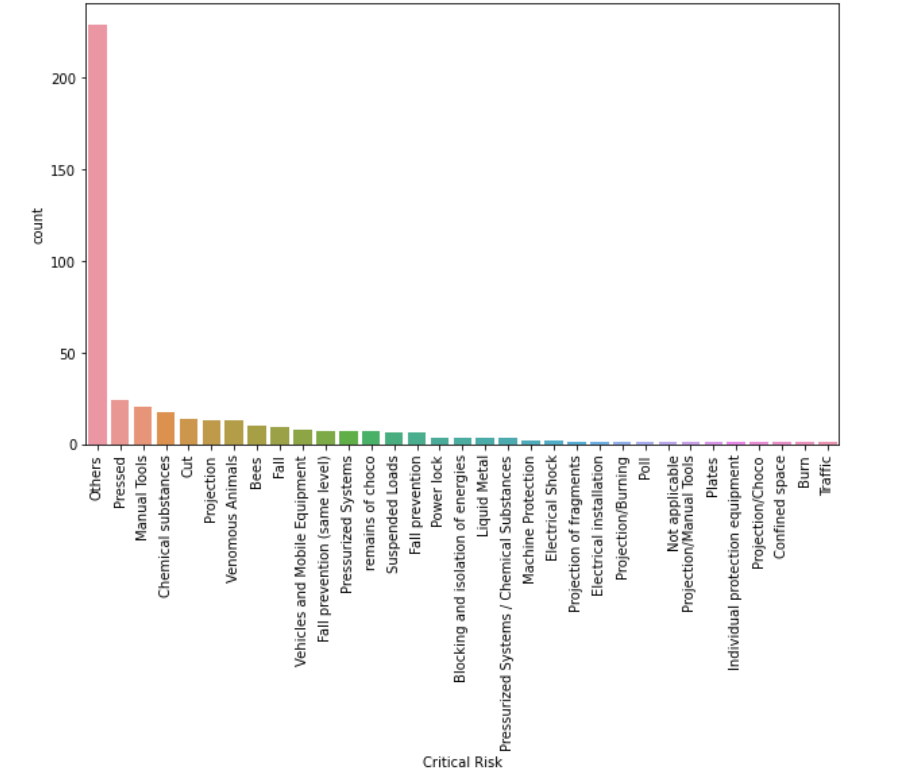


7. CRITICAL RISK-

It is used to determine the number of accidents caused by different critical risks such as pressed, pressurized system, electric shock and others.

COUNT PLOT-

The countplot() method is used to give the count of the number of accidents caused by different critical risks

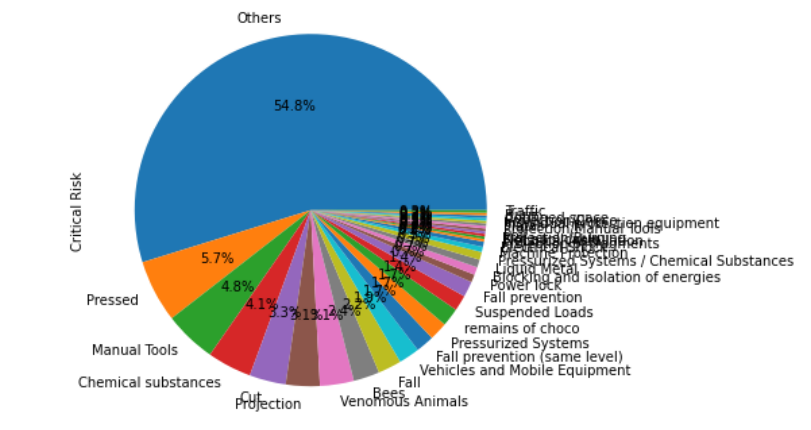


OBSERVATONS-

It can be noted that the maximum number of accidents have taken place with the critical risk “others” and minimum number of accidents took place with risk poll or traffic.

PIE CHART-

It is used to show the distribution of the different critical risks.

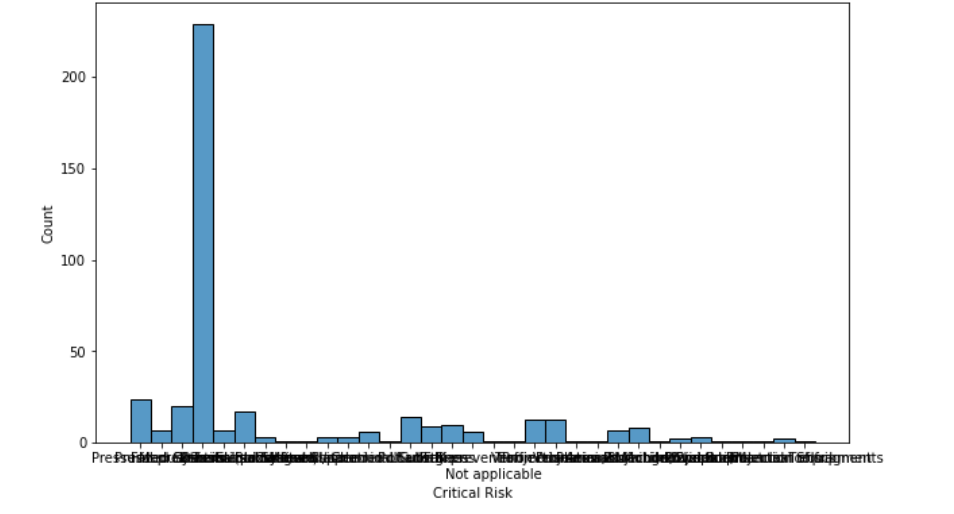


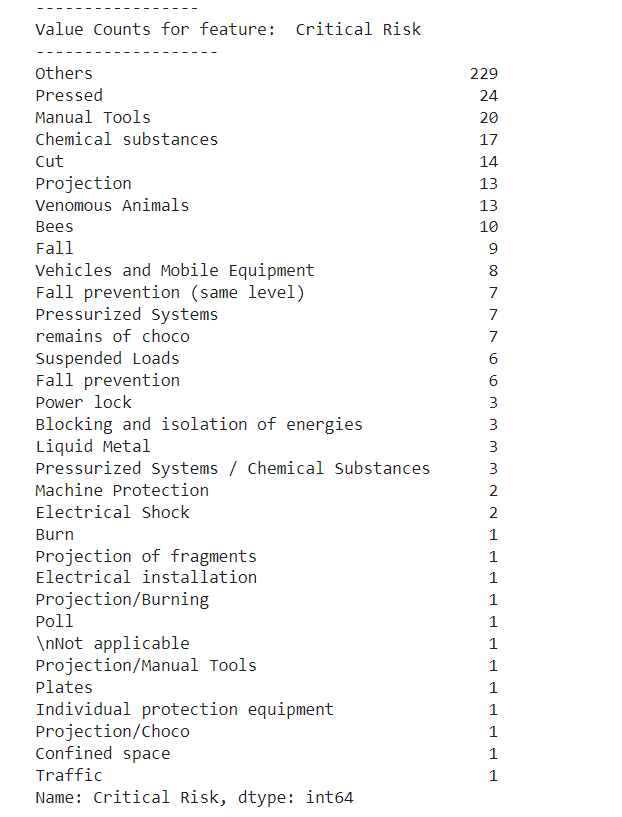
OBSERVATIONS –

It can be noted that most of the percentage of accidents have taken place with the critical risk “others”.

HISTPLOT-

It is used to show the distribution of the critical risks.

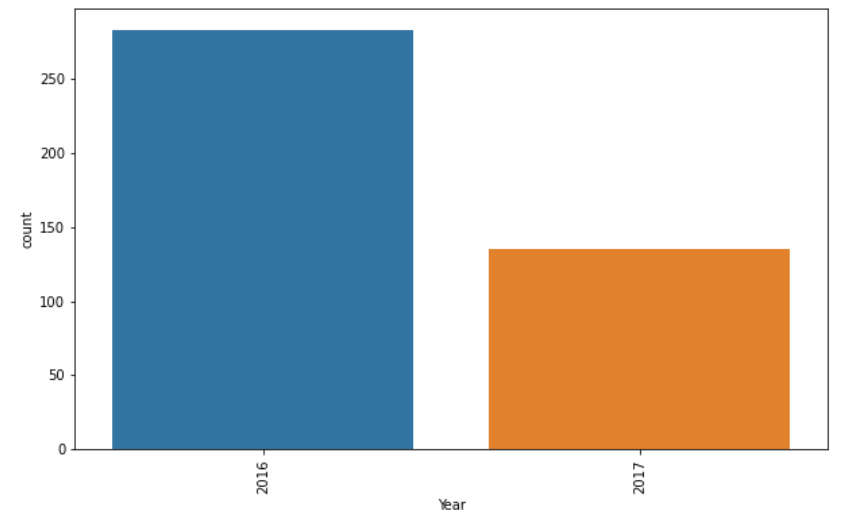




8. YEAR-

The year attribute is used to determine the number of accidents that had happened according to the years .i.e- 2016 and 2017.

COUNT PLOT- The countplot() method is used to predict the number of accidents taken place in 2016 and 2017

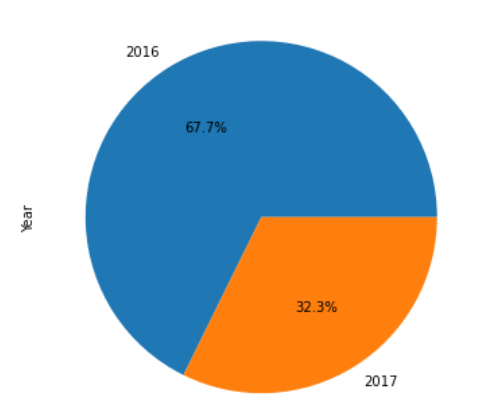


OBSERVATIONS –

It can be noted that the number of accidents that happened in 2016 (above 250) are higher than 2017(150).

PIE CHART-

It is used to show the percentage of accidents that had taken place according to the years.

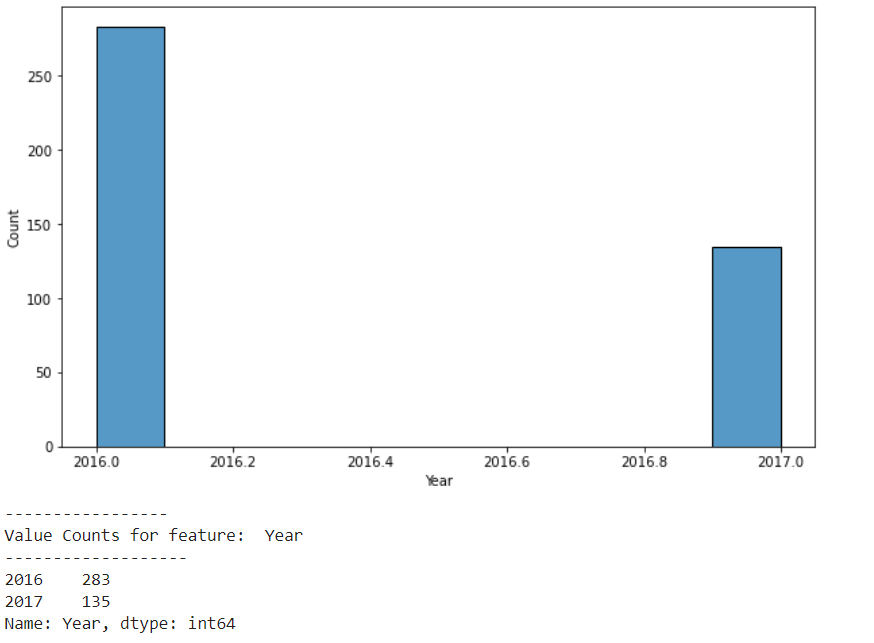


OBSERVATIONS-

It can be noted that the percentage of accidents occurred in 2016 are maximum

HISTPLOT-

It shows the distribution of the accidents that had taken place according to the years

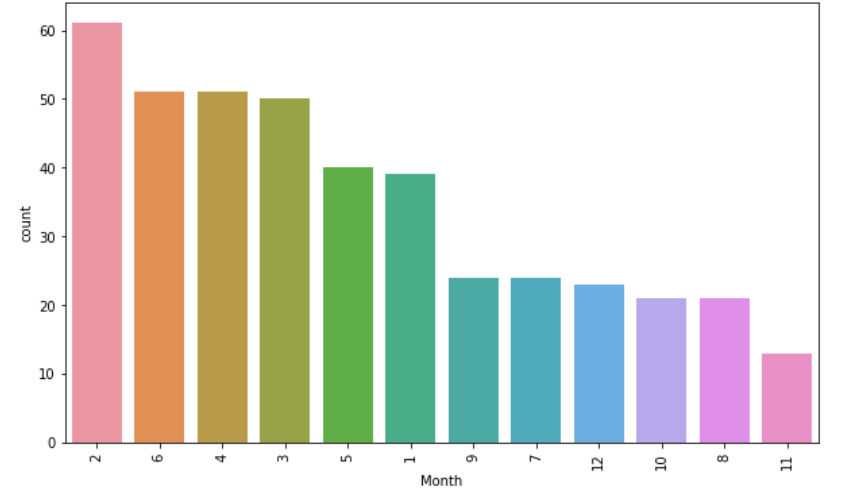


10. MONTH-

It is used to determine the number of accidents that had occurred in all the months of the years.

COUNTPLOT-

It is used to determine the number of accidents that had occurred according to their months.

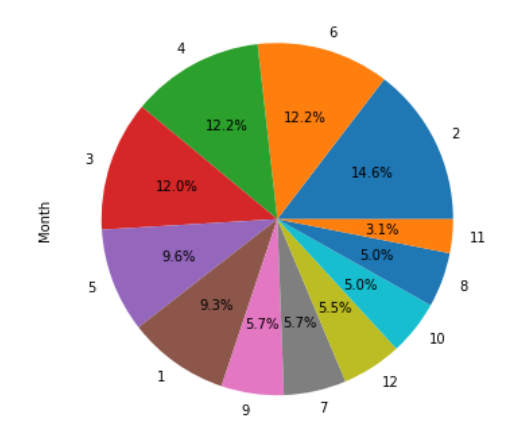


OBSERVATIONS –

It can be noted that the maximum number of accidents happened in Feb and minimum number of accidents occurred in November.

PIECHART-

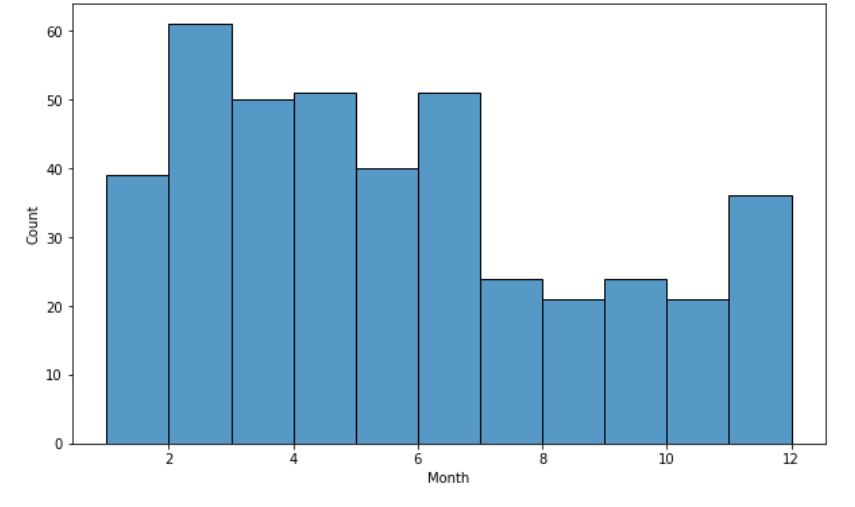
It is used to show the number of accidents that had happened in all the months in terms of percentage.



OBSERVATION- It can be noted from the above piechart that the maximum percentage of accidents happened in the month February.

HISTOGRAM-

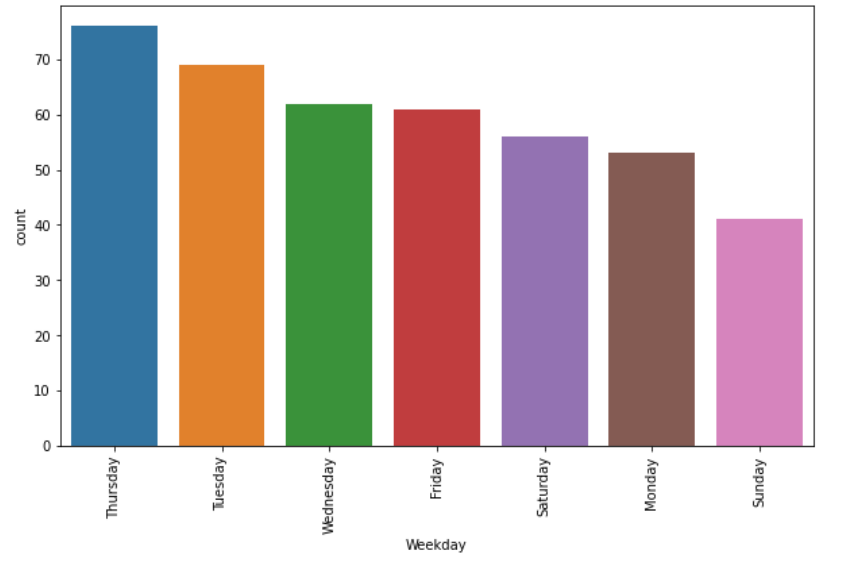
The histplot is used to determine the distribution of accidents according to the months.



11. Weekday- It is used to determine the number of accidents on the different days of a week.

COUNTPLOT-

The countplot() method is used to determine the number of accidents that had occurred in the days of the week.

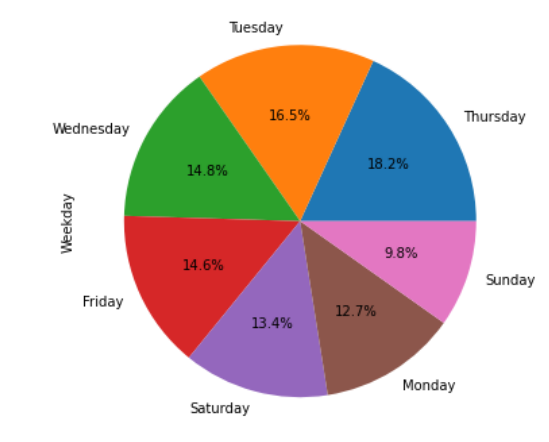


OBSERVATIONS –

It can be determined that maximum number of accidents have taken place on Thursday with a count of about 76 and least number of accidents took place on Sunday.

PIECHART-

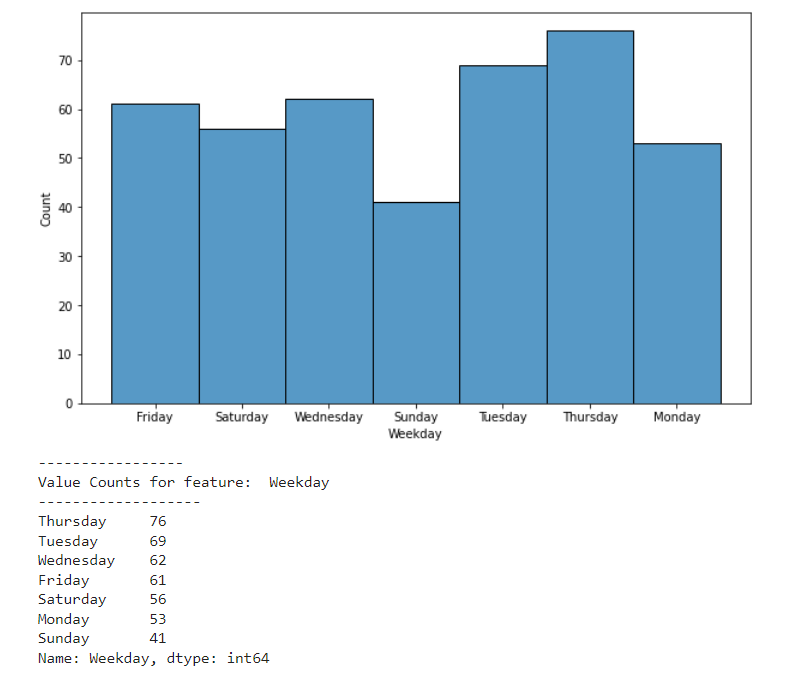
It is used to determine the percentage of accidents that took place in days of a week.



OBSERVATION – It can be observed that maximum percentage of accidents happened on Thursday with 16.5% and least number of accidents happened on Sunday.

HISTPLOT-

It is used to determine the distribution of accidents according to the days.



OBSERVATION –

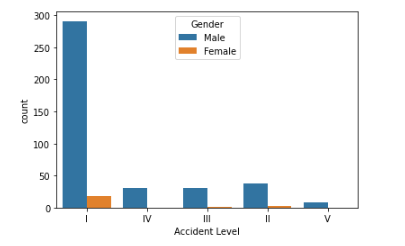
It can be noted that maximum number of accidents have taken place on Thursday

**BIVARIATE ANALYSIS:**

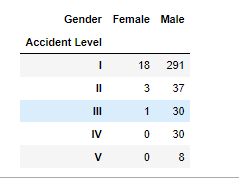
**1 Gender vs RestAll**

* 1. **Gender vs Accident level**

Count Plot:



Cross table Analysis:

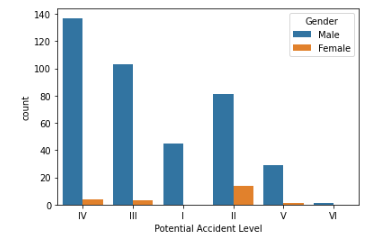


Observation –

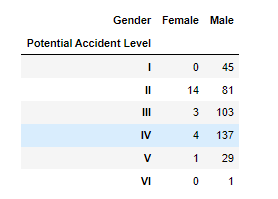
From the above plots, it can be determined that the most of the accidents happened at level I with gender male.

* 1. **Gender vs Potential Accident Level**

Count Plot:



Cross table Analysis:

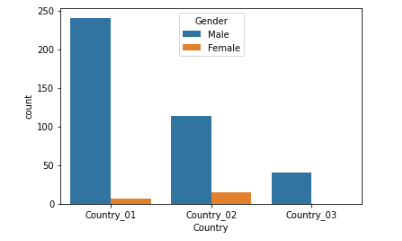


Observation –

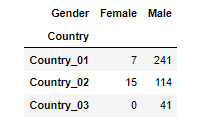
From the above,it can be determined that most of the potential level accidents happened to male compared to female, of which Potential Accident Level of IV is dominant.

* 1. **Gender vs Country**

Count Plot:



Cross table Analysis:

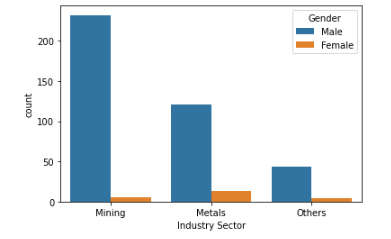


Observation –

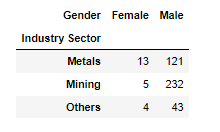
From the above countplot, it can be determined that the maximum number of accidents took place in country\_01 to males and they are about 241.

* 1. **Gender vs Industry Sector**

Count Plot:



Cross table Analysis:

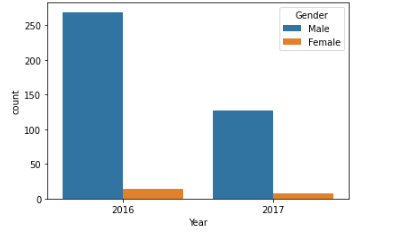


Observation –

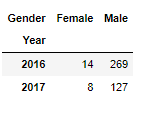
From the above plots, it is evident that most of the accidents happened to Male in the mining sector, around 232.

* 1. **Gender vs Year**

Count Plot:



Cross table Analysis:

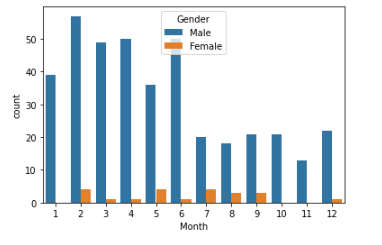


Observation –

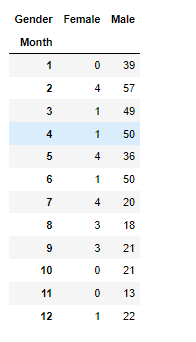
From the above countplot, it is clearly evident that maximum accidents took place in 2016 to the male when compared to female with a count of 269.

* 1. **Gender vs month**

Count Plot:



Cross table Analysis:

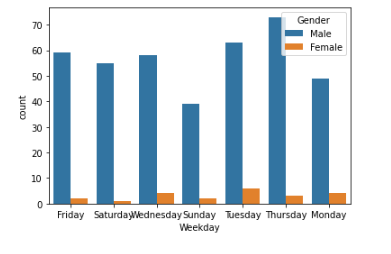


Observation –

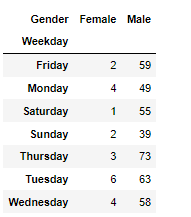
From the above count plot, it is determined that maximum number of accidents happened to male in the month feb with a count 57.

* 1. **Gender vs weekday**

Count Plot:



Cross table Analysis:

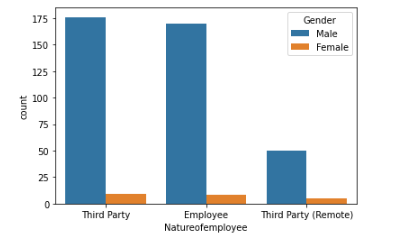


Observation –

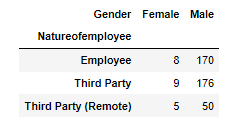
Max accidents happened to male on thursday with a count of more than 73

* 1. **Gender vs Nature of Employee**

Count Plot:



Cross table Analysis:

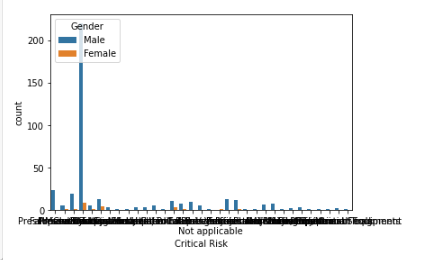


Observation –

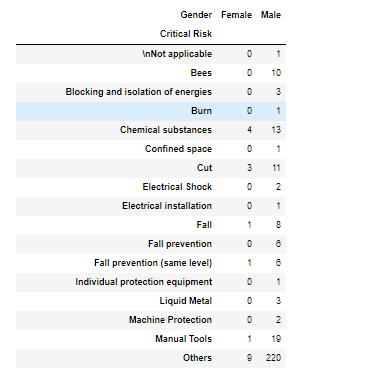
From the above output, it is clearly evident that maximum accidents happened to third party male employees. i.e- 176.

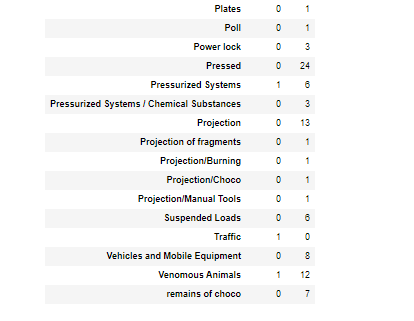
* 1. **Gender vs Critical Risk**

Count Plot:



Cross table Analysis:





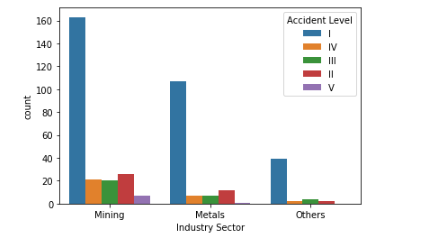
Observation –

Critical Risk of type "Others" is dominant across both Male and Female Genders.

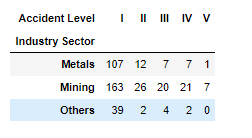
**2 Industry sector vs RestAll**

* 1. **Industry Sector Vs Accident Level**

Count Plot:



Cross table Analysis:

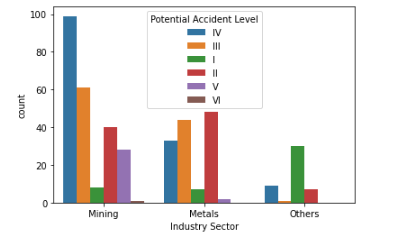


Observation –

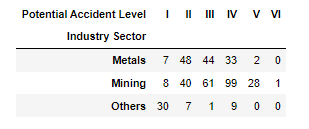
Maximum number of accidents happened in the mining sector with accident Level I. i.e- 163.

* 1. **Industry sector vs potential accident level**

Count Plot:



Cross table Analysis:

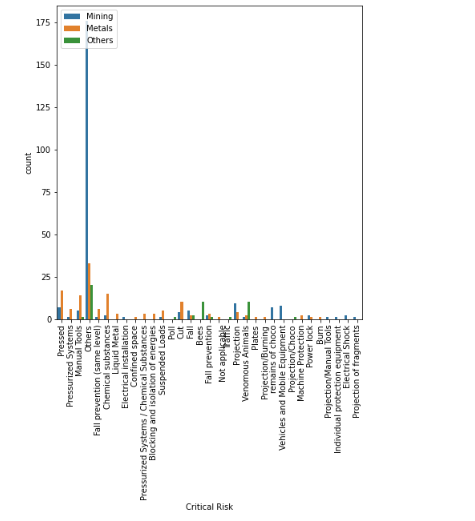


Observation –

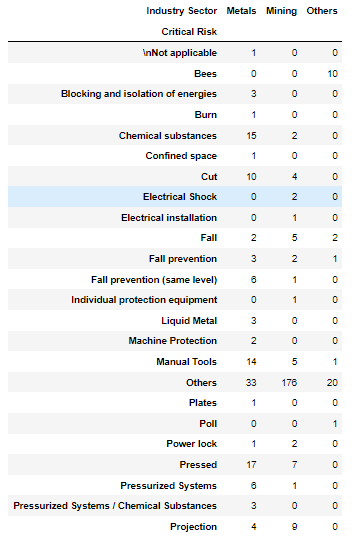
Maximum number of accidents happened in the potential accident level 4 and mining sector with a count 99. Minimum number of accidents took place in the mining sector at a potential accident level 6..

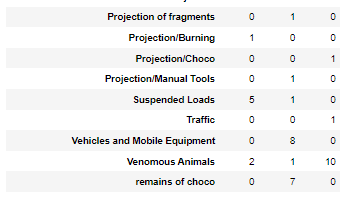
* 1. **Industry Sector vs Critical Risk¶**

Count Plot:



Cross table Analysis:



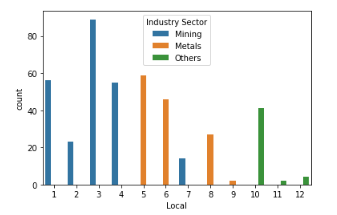


Observation –

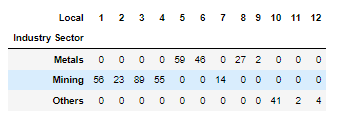
From the above count plot, it is evident that maximum number of accidents happened in mining with a critical risk of others. i.e- about 175.

* 1. **Industry sector vs Local**

Count Plot:



Cross table Analysis:

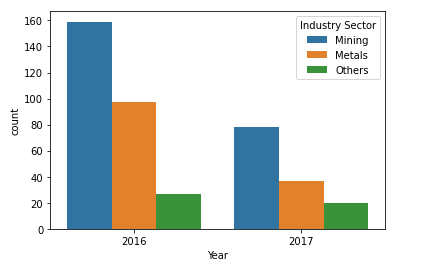


Observation –

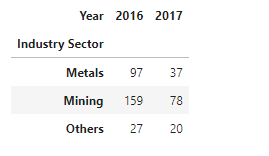
Many accidents happened with a local 3 and industrial sector mining. i.e- more than 80. Least accidents took place with local 11 and industrial sector others.

* 1. **Industry Sector Vs Year**

Count Plot:



Cross table Analysis:



Observation –

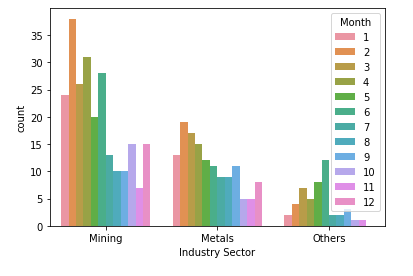
From the above plot, the following could be determined

* The number of accidents taken place in year 2016 for mining sector is 160.
* The number of accidents taken place in year 2016 wrt metals sector is about 100.
* The number of accidents taken place in the year 2016 wrt others sector is about 30. Hence, it can be determined that maximum accidents took place in mining sector in the year 2016.
* The number of accidents taken place in the year 2017 wrt mining sector is 80.
* The number of accidents taken place in the year 2017 wrt metals sector is about 40.
* The number of accidents taken place in the year 2017 wrt others sector is 20.

Hence, it can be determined that max accidents took place in mining sector in the year 2017.

* 1. **Industry Sector Vs Month**

Count Plot:



Cross table Analysis:

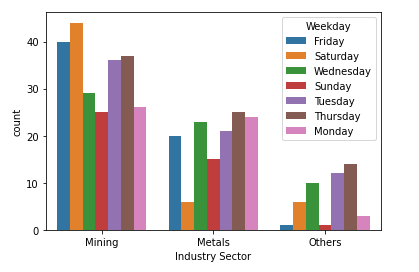


Observation –

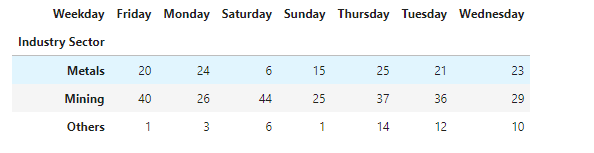
Maximum number of accidents happened in the month feb and mining sector. The least number of accidents took place in the others sector and month december..

* 1. **Industry sector vs weekday**

Count Plot:



Cross table Analysis:

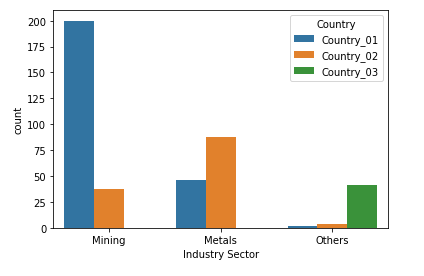


Observation –

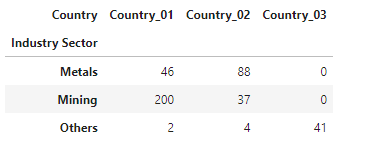
Maximum number of accidents hapenned on the day saturday in the mining sector. i.e- more than 40. The least number of accidents happened on the day sunday in the others sector.

* 1. **Industry sector vs Nature of Employee**

Count Plot:



Cross table Analysis:

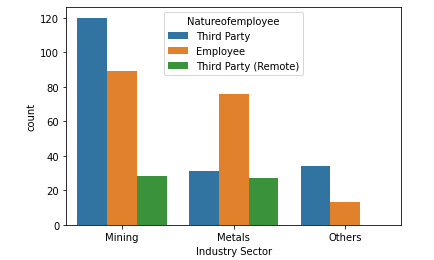


Observation –

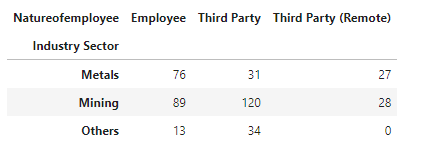
From the above count plot, it is evident that the maximum number of accidents took place in country\_01 and mining sector.i.e- 200. The least number of accidents took place in country \_01 and others sector.

* 1. **Industry Sector Vs nature of employee**

Count Plot:



Cross table Analysis:



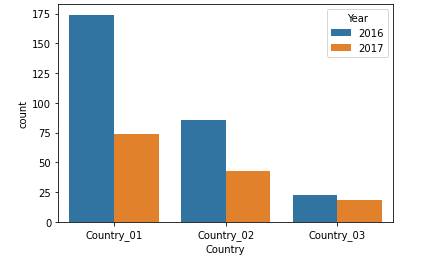
Observation –

From the above count plot, it is clearly evident that the maximum accidents took place in the mining sector with the third party employee type. i.e- about 120. The least number of accidents took place in the others sectors with the nature of employee as employee.

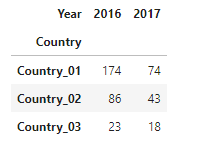
**3 Country vs RestAll**

* 1. **Country vs Year**

Count Plot:



Cross table Analysis:

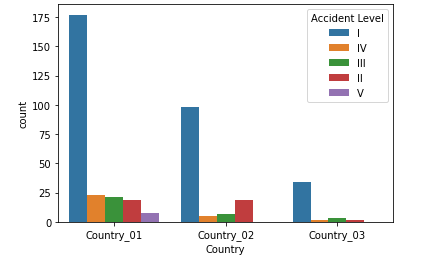


Observation –

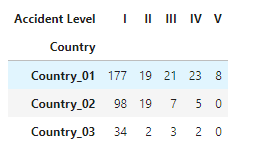
From the above output, the following can be determined-

* The number of accidents taken place in country\_01 and year 2016 is 174.
* The number of accidents taken place in country\_01 and year 2017 is about 74.
* The number of accidents taken place in country\_02 and year 2016 is more than 86.
* The number of accidents taken place in country\_02 and year 2017 is about 43.
* The number of accidents taken place in country\_03 and year 2016 is about 23.
* The number of accidents taken place in country\_03 and year 2017 is about 18.
  1. **Country Vs accident level**

Count Plot:



Cross table Analysis:

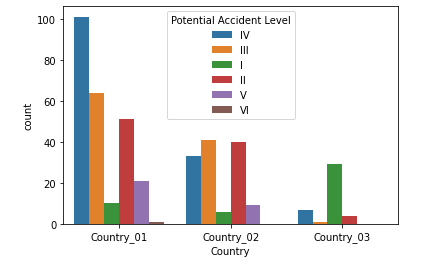


Observation –

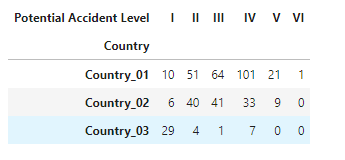
From the above count plot, it is clearly evident that the maximum number of accidents took place in accident level 1 and country\_01.

* 1. **Country Vs Potential Accident Level**

Count Plot:



Cross table Analysis:

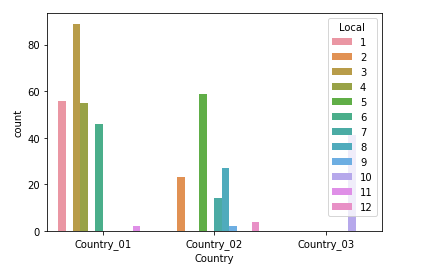


Observation –

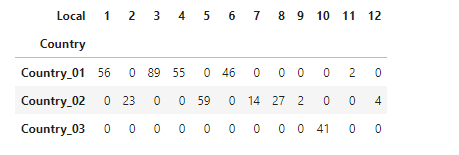
From the above plot, it is evident that the maximum accidents occurred in country\_01 and potential accident level 3

* 1. **Country Vs Local**

Count Plot:



Cross table Analysis:

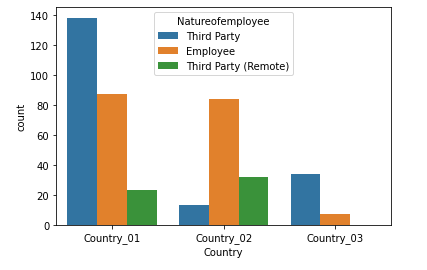


Observation –

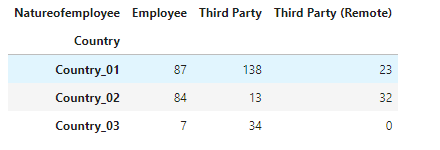
Country 1 is more dominant in local 3 region and least dominant in Local 12.

* 1. **Country Vs Nature Of Employee**

Count Plot:



Cross table Analysis:

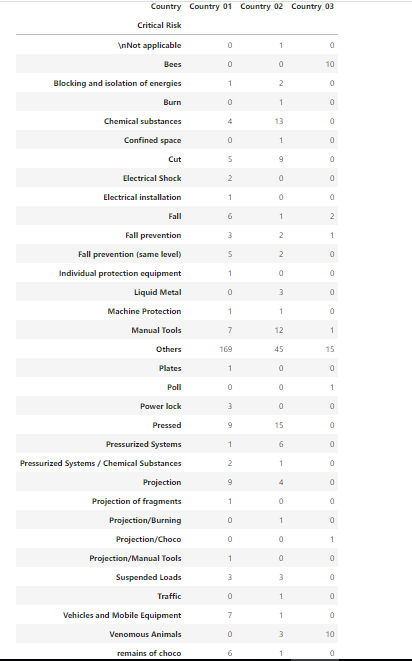


Observation –

Accidents in Country 01 is more dominant in Third Party type of employee, country 03 is least dominant in Third Party (Remote).

* 1. **Country Vs Critical Risk**

Cross table Analysis:



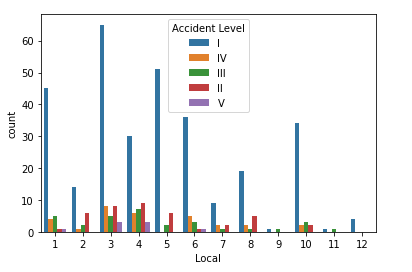
Observation –

Country 01 is more dominant in Others Critical Risk and Critical Risk is least dominant in Country 03

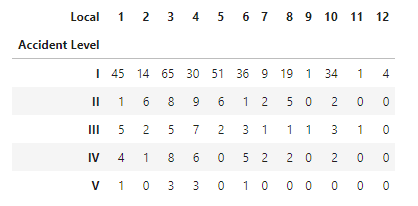
**4 Local Vs Rest All**

* 1. **Local Vs Accident Level**

Count Plot:



Cross table Analysis:

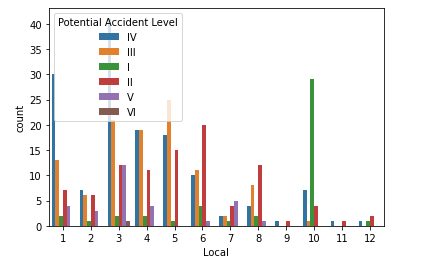


Observation –

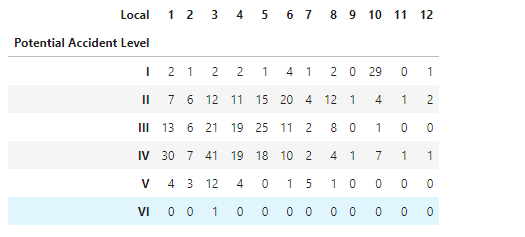
Accident level 1 is more dominant in Local 2 region with 65 accidents, while Accident Level V is least across all Locals

* 1. **Local Vs Potential Accident Level**

Count Plot:



Cross table Analysis:

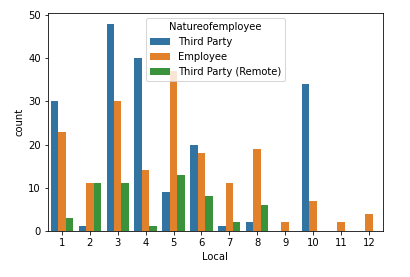


Observation –

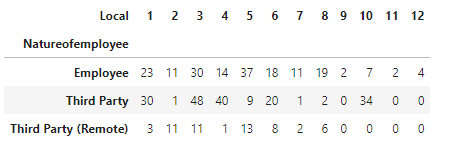
Overall Local 3 is more prone to Multiple potential accidents, while local 12 is the least.

* 1. **Local Vs Natureofemployee**

Count Plot:



Cross table Analysis:

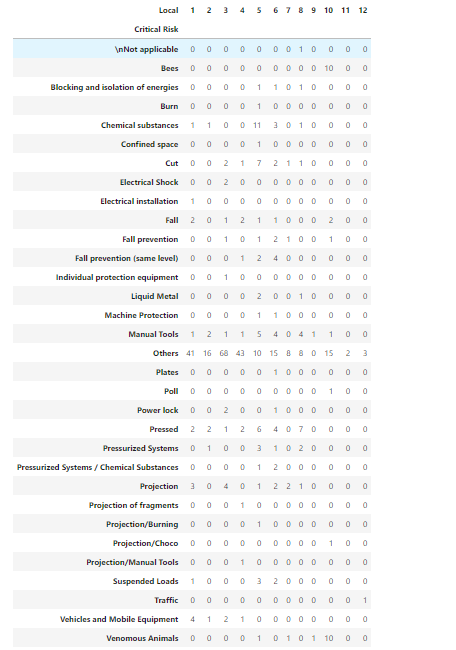


Observation –

Type Employee is more dominant across all Locals, while Type Third Party(Remote) is least dominant across all Locals.

* 1. **Local Vs Critical Risk**

Cross table Analysis:

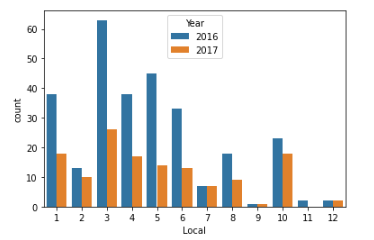


Observation –

Critical Risk of type "Others" is dominant across all Locals.

* 1. **Local Vs Year**

Count Plot:



Cross table Analysis:



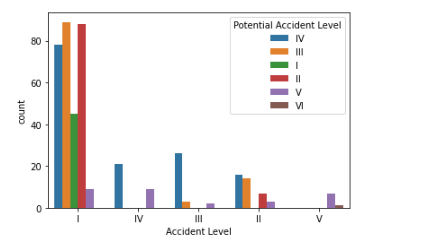
Observation –

Year 2016 has more accidents across all Local regions compared to 2017.

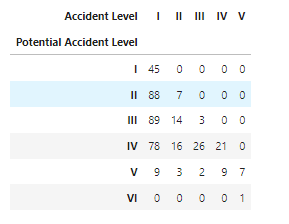
**5 Accident Level Vs Rest All**

* 1. **Accident Level Vs Potential Accident Level**

Count Plot:



Cross table Analysis:

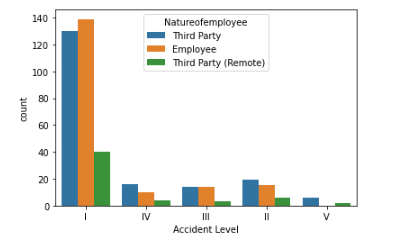


Observation –

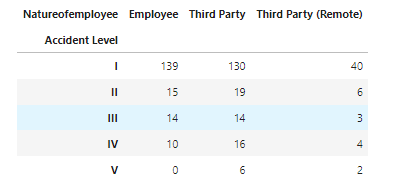
Accident Level I is more related to Potential Accident levels of I, II, III, IV, V, VI.

* 1. **Accident Level Vs Natureofemployee**

Count Plot:



Cross table Analysis:

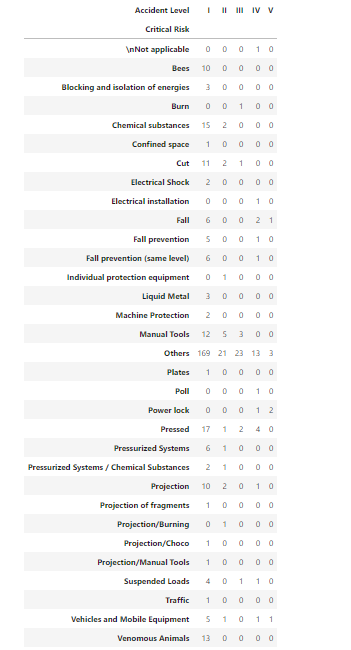


Observation –

Accident Level I is more dominant across all Employee types, where Level V is least across all types.

* 1. **Accident Level Vs Critical Risk**

Cross table Analysis:

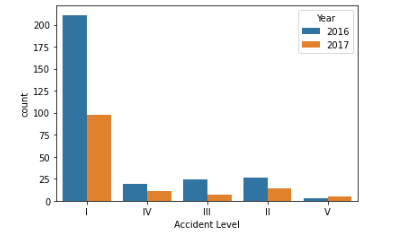


Observation –

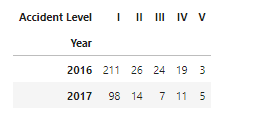
Accident Level I is more dominant with Other critical Risk type

* 1. **Accident Level Vs Year**

Count Plot:



Cross table Analysis:

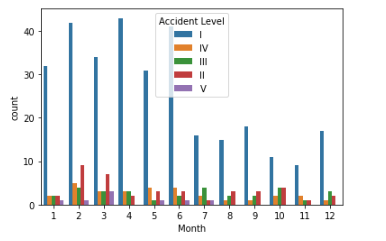


Observation –

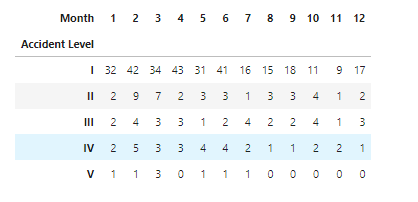
Accident Level I is more dominant in across 2016 and 2017 years, and Level V is minimum.

* 1. **Accident Level Vs Month**

Count Plot:



Cross table Analysis:

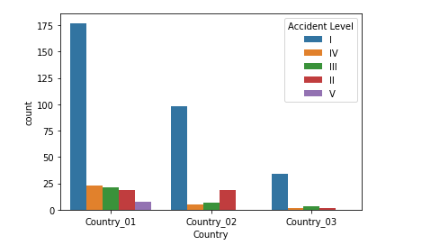


Observation –

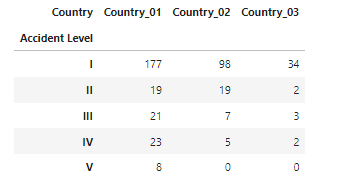
Accident Level 1 dominates across all Months while Level V is minimum.

* 1. **Accident Level Vs Country**

Count Plot:



Cross table Analysis:



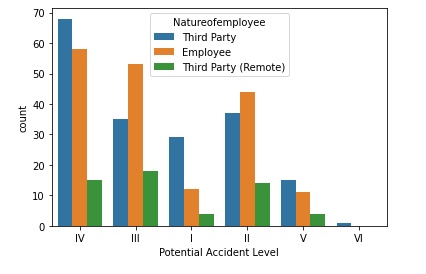
Observation –

Accident Level I is more dominant across all Countries, while Accident Level V is least dominant across all countries.

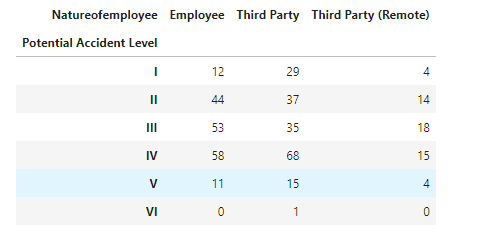
**6 Potential Accident Level Vs Rest All**

* 1. **Potential Accident Level Vs Natureofemployee**

Count Plot:



Cross table Analysis:

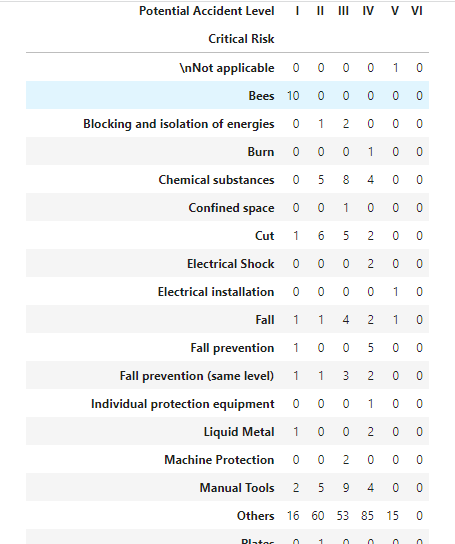


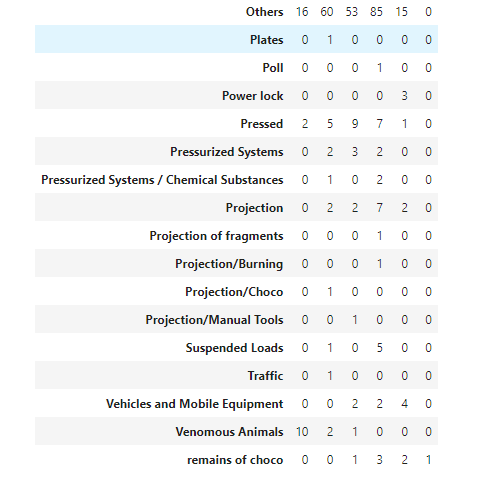
Observation –

Potential Accident level IV dominents in ThirdParty, while VI is least dominant in Third Party(Remote) across all

* 1. **Potential Accident Level Vs Critical Risk**

Cross table Analysis:



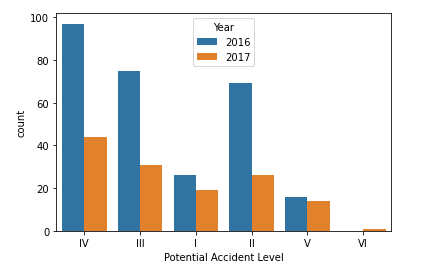


Observation –

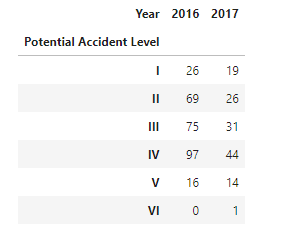
Among all Critical Risk with Type as "Others" is dominant across all Potential Accident Levels

* 1. **Potential Accident Level Vs Year**

Count Plot:



Cross table Analysis:

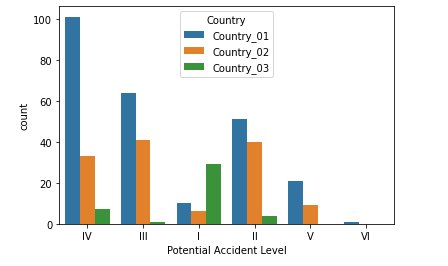


Observation –

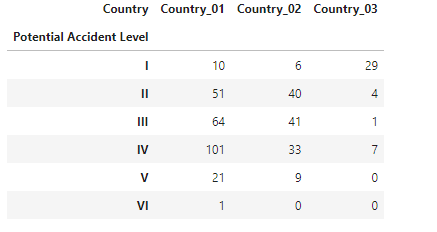
There is Decrease in Number of accidents across all Potential Accident level from 2016 to 2017. Potential Accident level IV is dominant in both 2016 and 2017

* 1. **Potential Accident Level Vs Country**

Count Plot:



Cross table Analysis:



Observation –

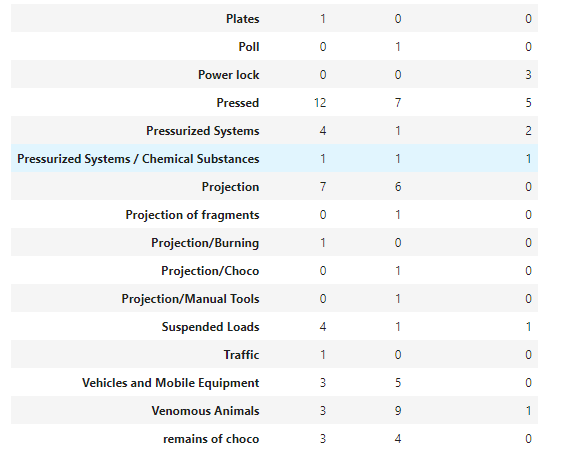
Potential Accident Level IV is dominant across all countries, while with VI least number of accidents happened.

**7 Natureofemployee Vs RestAll**

* 1. **Natureofemployee Vs Critical Risk**

Cross table Analysis:





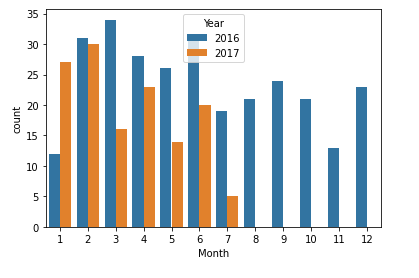
Observation –

Critical Risk of type "Others" is dominant across all Types of Employees

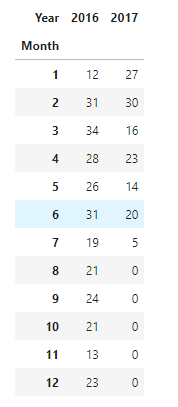
**8 Year Vs RestAll**

* 1. **Year vs month**

Count Plot:



Cross table Analysis:

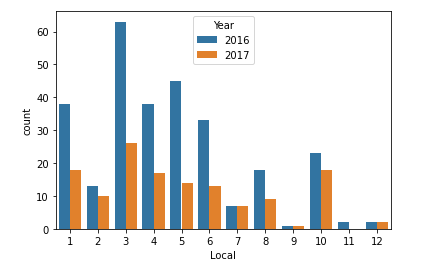


Observation –

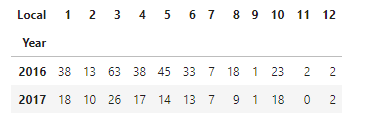
From the above plot, it is evident that the max accidents happened in the year 2016 and march.

* 1. **Year vs Local**

Count Plot:



Cross table Analysis:

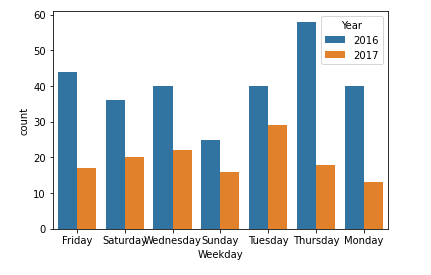


Observation –

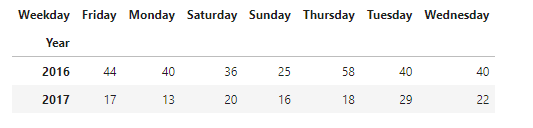
From the above plot, it can be determined that the maximum accidents took place in the local 3 and year 2016.

* 1. **Year vs Weekday**

Count Plot:



Cross table Analysis:

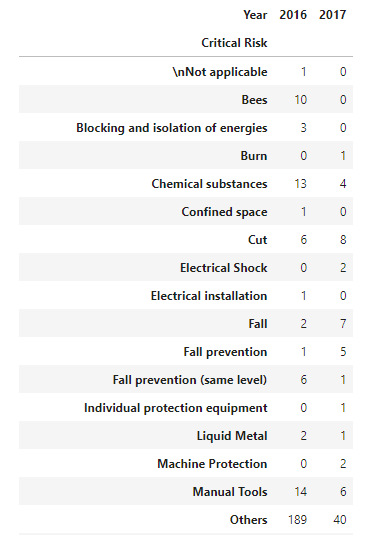


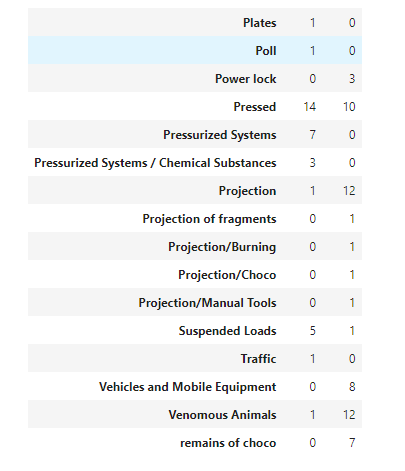
Observation –

From the above plot, it is clearly evident that maximum number of accidents took place on Thursday and year 2016.

* 1. **Year vs Critical Risk**

Cross table Analysis:





Observation –

From the above plot, it is clearly evident that maximum number of accidents took place with "Others" and year 2016.