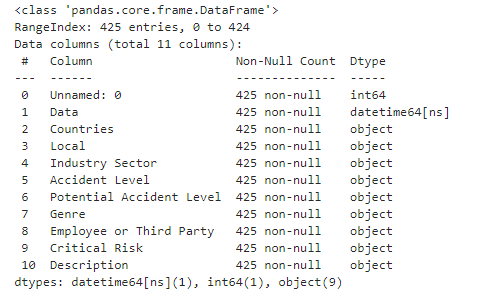
**Summary of the dataset -**

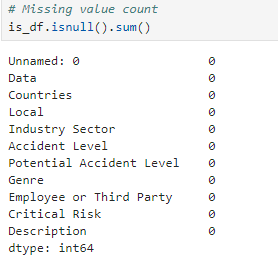
* All the columns of data are of “object” data type except the first column



* There are 425 rows and 11 columns in the dataset



* No missing values in the dataset. Hence, no missing value treatment is required



* 7 duplicates rows have been removed and removed 'Unnamed: 0' column from the dataset and left with 418 rows and 10 columns
* Summary of each column in the data –

1. 3 country types in the “countries” column
2. 12 Local cities where plant is located
3. 5 accident levels types and 6 potential accident level
4. 3 employee types
5. 3 industry sector types

**EDA –**

**Local** – Highest Manufacturing plant – Local\_03

Lowest Manufacturing plant – Local\_09

**Country –**

Country\_01 -59%

Country\_02 – 31%

Country\_03 – 10%

**Industry sector –**

Mining sector 57%

Metal sector 32%

Other sector 11%

**Gender –**

More men working in this industry as compared to women

**Data pre-processing –**

To make data ready for the model development, all the variable should be converted into the numerical form. Following methods are used to convert the categorical variables in numerical from –

* Label encoding (Gender/Accident Level/Potential Accident Level)
* Dummy variable creation (Country/Local/Industry Sector/Employee or Third Party/Critical Risk)
* TF-IDF method (Description)
* NLP preprocessing of the “Description” column. Below are the steps -

1. Converted to lower case
2. Removed white-spaces and non-alphanumeric characters
3. Removed stop words
4. Used lemmatization to convert the word tokens into the root word
5. TF-IDF algo is used to convert the description column into the vector form after using the above cleaning steps