**Knowledge Document**

# **Subset Data V10 Details**

* This data file contains the statements released by the companies, approximately there are 8-9k unique companies.
* There are in total 19 columns which were added gradually while doing the research, the versions started from V1, but the latest ones are available on the drive.

# **Details of Columns**

* statement – This column contains the text released by companies either in URLs or PDFs on the website, these were extracted and kept in this column.
* Company – This column is the name of the company which released statements stating what actions they are taking against modern slavery.
* FS\_industry – This column indicates the industry which the company belongs to.
* FS\_hq – This column indicates the headquarters of the company.
* statement\_cleaned – This column is the obtained after preprocessing of the statements released by companies.
* statement\_cleaned\_v2 – Some additionally cleaning was performed in the statement\_cleaned and hence both were kept.
* Company\_Id – This column was created to give the companies a unique id.
* 2022 NAICS Sector – This column indicates that the companies were further categorized into different sectors.
* 2-digit Code – This column indicates the unique number for the NAICS sector
* ID – This column was generated while extracting the financial information of the companies.
* (FY2019 – FY2022) – This column shows the financial performance for that specific year.
* Score\_PerTotalWords – This column shows the score to statements released by companies based on the category words per total words in the statement.
* Weighted\_CategoryScore – This column shows the score based on the categorical words (These words were identified while performing topic modeling) found in the statement.
* Normalized\_Score\_PerTotalWords – This column shows the normalized value of Score\_PerTotalWords
* Normalized\_Weighted\_CategoryScore - This column shows the normalized value of Weighted\_CategoryScore
* 2Word\_Phrase\_Score – This column shows the score of two closed words (Category words) found in the statement.