Dataset Documentation

Dataset Collector Name: Md Abu Sufian
University of Leicester
MSc in Data Analysis for Business Intelligence

Overview of the dataset: In the dataset, there are two key indicator datasets one is the sustainability dataset for the energy sector and the other is the financial indicator.

Sustainable ESG Metrics	**Financial Metrics**
Ticker	Ticker
Company_Name	Expected_return
Exchange_Name	Volatility
Total_GHG_Emissions	cik
Safety_Incidents	calendarYear
Workforce_Safety_Incidents	period
Employees	Face_value
Women%	enterpriseValue
Methane_Emissions	PD
Methane_Intensity_Rate	
Total_Energy_Used	
Energy_generated	
Fresh_Water_Consumed	
Water_Returned	
Women_in_Executive	
Spills	

summary statistics for each Sustainable variable:

1. Ticker: Each company has a unique identifier or "ticker". The number "2" appears twice in the dataset, which might indicate a duplicate row.

2. Company_Name: Each company name is unique, suggesting that each row might represent data from a unique company.

3. Exchange_Name: The dataset seems to consist of companies listed on four different exchanges, with one being the most prominent, having 122 companies.

4. Total_GHG_Emissions: The distribution is heavily skewed towards a certain value (approximately 515 million), but it also shows a significant range.

5. Safety incidents: There's a clear outlier with a significantly higher count of safety incidents (132) than the rest. Other numbers have less frequency.

6. Workforce_Safety_Incidents: This column also has a very high frequency of a specific number (133), possibly suggesting many companies with similar workforce safety incident rates.

7. Employees: The dataset is skewed towards a certain number of employees (approximately 21,104), suggesting that a large portion of the companies have a similar employee count.

8. Women%: There's a clear outlier with a significantly higher frequency (136) than the rest, which could suggest a general trend in the proportion of women employed by these companies.

9. Methane_Emissions: This distribution is heavily skewed towards a certain value (approximately 331,476), but it also shows a significant range.

10. Methane_Intensity_Rate: The dataset is highly skewed towards a particular value (approximately 0.829), suggesting a common Methane Intensity Rate across most of the companies.

11. Total_Energy_Used: The dataset is highly skewed towards a particular value (approximately 236 million), which could be a common total energy used by most of companies.

12. Energy_generated: The dataset is highly skewed towards a particular value (approximately 7.8 million), suggesting that most of the companies generate a similar amount of energy.

13. Fresh_Water_Consumed: The dataset is heavily skewed towards a specific value (approximately 67 million), but it also shows a significant range.

Purpose: This research project aims to explore how machine learning algorithms can be used to integrate sustainability metrics into risk assessment models to improve risk-weighted asset calculation and default probability prediction in companies.

Source: Web scraping
web scraping steps:

1. Start
2. Set up Python environment
3. Import Beautifulsoup and Requests libraries
4. Configure the scraper
5. Send a request to the URL using requests.get()
   - If the request fails, handle the exception, and proceed to the End step
6. Success: The request is successful
7. Parse the HTML using Beautifulsoup
8. Find the desired elements with Beautifulsoup
   - If the elements are not found, log the missing elements, and proceed to the End step
9. Found: The desired elements are found
10. Extract the data
11. Store the data in a Python data structure
12. More pages to scrape?
   - If yes, repeat steps 5 to 11 for the next page
   - If no, proceed to the next step
13. Store the data permanently
14. End of web scraping

This sequence outlines the typical steps involved in a web scraping process, from setting up the environment to storing the scraped data. Remember to handle exceptions and log any errors encountered during the process to ensure a smooth and reliable scraping operation.

Tools: Python, Google Colab, Azure , Different PDF libraries

Not only Python library was main weapon but also A total of 10 PDF libraries were tried in Python to gauge their effectiveness in terms of data extraction from PDF files before libraries were chosen. For each library, the amount of time needed to extract the data was noted. The decision to choose a certain library was not made based on a predefined preference but rather on the details and requirements of the work at hand.

- Pypdf2-3.3 sec
- Pdfrw-.067 sec
- Tabula-42.67 sec
- Pymypdf-.27 sec
- Slate-12.9 sec

- Pikepdf
- PDFMiner is slower than other C/C++-based counterparts such as XPdf, but it took 15 seconds.

Variables/Features:

1. Ticker
2. Company_Name
3. Exchange_Name
4. Total_GHG_Emissions
5. Safety_Incidents
6. Workforce_Safety_Incidents
7. Employees
8. Women%
9. Methane_Emissions
10. Methane_Intensity_Rate
11. Total_Energy_Used
12. Energy_generated
13. Fresh_Water_Consumed
14. Water_Returned
15. Women_in_Executive
16. Spills
17. Expected_return
18. Volatility
19. cik
20. CalendarYear
21. period
22. Face_value
23. Enterprise value
24. PD
25. labels

Data type:

In the project dataset variables are numerical and categorical, not textual data present.

1. Ticker	int64
2. Company_Name	int64
3. Exchange_Name	int64
4. Total_GHG_Emissions	float64
5. Safety_Incidents	float64
6. Workforce_Safety_Incidents	float64
7. Employees	float64
8. Women%	float64
9. Methane_Emissions	float64
10. Methane_Intensity_Rate	float64

11. Total_Energy_Used	float64
12. Energy_generated	float64
13. Fresh_Water_Consumed	float64
14. Water_Returned	float64
15. Women_in_Executive	float64
16. Spills	float64
17. Expected_return	float64
18. Volatility	float64
19. cik	int64
20. CalendarYear	int64
21. period	int64
22. Face_value	float64
23. enterprise value	float64
24. PD	float64
25. labels	int64
26. dtype:	object

Range of each column's values

Variable	Count	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
Ticker	200	99.06	57.451893	0	49.75	99.5	148.25	198
Company_Name	200	99.5	57.879185	0	49.75	99.5	149.25	199
Exchange_Name	200	2.35	0.949742	0	2	3	3	3
Total_GHG_Emissions	200	333494300	243552700	416	2744606	515252800	515252800	515252800
Variable	Count	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
Safety_Incidents	200	27.333238	35.42364	0	26.259322	26.679096	26.679096	410
Workforce_Safety_Incidents	200	2.868642	8.727516	0	1	2.802469	2.802469	122
Employees	200	21441.65	32489.582	0.32	21104.92	21104.92	21104.92	
Variable	Count	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
Women%	200	0.234828	0.054167	0.04	0.234458	0.234458	0.234458	0.519
Methane_Emissions	200	337258.2	478691.3	0.07	331476.2	331476.2	331476.2	6000000
Methane_Intensity_Rate	200	0.839924	0.794554	0.00006	0.829492	0.829492	0.829492	7.38
Spills	200	51.695652	45.642033	0	53.710145	53.710145	53.710145	352
Variable	Count	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
Expected_return	200	0.767424	1.045198	-0.747906	0.122575	0.434972	1.184716	5.773411
Volatility	200	0.617865	0.371588	0.023594	0.383107	0.547757	0.709086	2.434749
cik	200	1104540	584046.9	2178	801652	1191756	1558555	1915657
calendarYear	200	2021	0	2021	2021	2021	2021	2021
period	200	0	0	0	0	0	0	0
Variable	Count	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
Face_value	200	2.68E+11	3.68E+12	0.00E+00	1.32E+07	3.61E+08	1.74E+09	5.21E+13
enterpriseValue	200	9.83E+11	1.35E+13	7.94E+06	4.43E+08	2.18E+09	8.14E+09	1.91E+14
PD	200	3.39E-02	1.10E-01	0.00E+00	4.78E-16	6.51E-07	1.86E-03	6.80E-01
labels	200	1.30E-01	3.37E-01	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.00E+00

Data Pre-processing: There are many steps taken at a pre-processing stage such as cleaning, normalization, imputation of missing values.

The dataset has undergone several pre-processing processes to prepare the data for study. The pre-processing steps and their justifications are as follows:

1. Data Cleaning: - Justification Finding and fixing flaws, inconsistencies, and inaccuracies in the dataset is data cleaning. It guarantees that the data is correct, dependable, and appropriate for analysis.

- Impact: Data cleaning reduces any biases or misleading conclusions that could result from inaccurate data by removing flaws and inconsistencies in the dataset.

2. Normalisation/ Standardisation

- Justification: Normalisation is a technique used to adjust numerical variables to a standard range, which is frequently between 0 and 1. Removing the impact of different scales and units, enables fair comparisons between variables.

Impact: The normalisation process makes sure that variables with various scales contribute evenly to the analysis. It helps improve model performance by preventing variables with greater values from dominating the analysis.

3. Missing Values imputation:

- Rationale: Missing values can occur in the dataset due to various reasons such as data collection from web scraping that's why incomplete data and lots of missing values. Imputing missing values is necessary to handle these gaps and ensure the availability of complete data for analysis.

- Impact: Imputing missing values helps retain valuable information that would otherwise be lost. It allows for a more comprehensive analysis by utilising the available data, reducing the potential bias introduced by omitting incomplete cases.

Data Quality Assessment:

Consistency of data set:

1. Ticker: Each company has a unique identifier or "ticker". The number "2" appears twice in the dataset, which might indicate a duplicate row.

2. Company_Name: Each company name is unique, suggesting that each row might represent data from a unique company.

3. Exchange_Name: The dataset seems to consist of companies listed on four different exchanges, with one being the most prominent, having 122 companies.

4. Total_GHG_Emissions: The distribution is heavily skewed towards a certain value (approximately 515 million) but also shows a significant range.

5. Safety_incidents: There's a clear outlier with a significantly higher count of safety incidents (132) than the rest. Other numbers have less frequency.

6. Workforce_Safety_Incidents: This column also has a very high frequency of a specific number (133), possibly suggesting many companies with similar workforce safety incident rates.

7. Employees: The dataset is skewed towards a certain number of employees (approximately 21,104), suggesting that many companies have a similar employee count.

8. Women%: There's a clear outlier with a significantly higher frequency (136) than the rest, which could suggest a general trend in the proportion of women employed by these companies.

9. Methane_Emissions: This distribution is heavily skewed towards a certain value (approximately 331,476) but also shows a significant range.

10. Methane_Intensity_Rate: The dataset is highly skewed towards a particular value (approximately 0.829), suggesting a common Methane Intensity Rate across most of the companies.

11. Total_Energy_Used: The dataset is highly skewed towards a particular value (approximately 236 million), which could be a common total energy used by most of companies.

12. Energy_generated: The dataset is highly skewed towards a particular value (approximately 7.8 million), suggesting that most of the companies generate a similar amount of energy.

13. Fresh_Water_Consumed: The dataset is heavily skewed towards a specific value (approximately 67 million), but it also shows a significant range.

Data Validation Summary:

Accuracy Check:
No inaccuracies were found in the "merged_data" DataFrame. The values in all columns meet the specified criteria.

Completeness Check:
The "merged_data" DataFrame is complete, with no missing values in any of the columns. All columns have non-null values.

Consistency Check:
Duplicate rows were identified in the "merged_data" DataFrame. These duplicate rows may indicate data entry errors or inconsistencies. Further investigation and appropriate actions are recommended to address these duplicates.

The data validation process for the "merged_data" DataFrame indicates that the data is accurate, complete, and consistent, except for the identified duplicate rows. Taking appropriate steps to resolve the duplicate entries will help improve the overall data quality.

Describe any outlier detection or treatment methods have been used:

For this dataset normalisation or standardisation technique has been applied. Creating a scatter plot shows not normal distribution which has a significant impact.

The code calculates the Z-scores for each value in the columns and identifies the outliers based on a threshold value

The Z-score formula is:

Z-score = (x - mean) / standard deviation

By substituting the values of each outlier into this formula and calculating the Z-score,the calculation has been done in Python, then the outcome determined how many standard

deviations away from the mean each outlier value which allows us to assess the magnitude of the outliers

Here is a summary of the outliers found in the specified columns of the 'merged_data' DataFrame:

1. Total_GHG_Emissions: No outliers were found in this column.
2. Safety_Incidents: There are 3 outliers with the following characteristics:
    - Ticker: 190, Company_Name: 195, Exchange_Name: 1
    - Total_GHG_Emissions: 1613.0
    - Safety_Incidents: 144.0
    - Workforce_Safety_Incidents: 2.802469
    - Employees: 0.32
    - Women%: 0.234458
    - Methane_Emissions: 65.0
    - Methane_Intensity_Rate: 0.829492
    - Spills: 53.710145
    - Other column values: Varying
3. Workforce_Safety_Incidents: There is 1 outlier with the following characteristics:
    - Ticker: 0, Company_Name: 0, Exchange_Name: 3
    - Total_GHG_Emissions: 561.0
    - Safety_Incidents: 410.0
    - Workforce_Safety_Incidents: 122.0
    - Employees: 105600.0
    - Women%: 0.26
    - Methane_Emissions: 331476.0
    - Methane_Intensity_Rate: 0.007
    - Spills: 5.0
    - Other column values: Varying
4. Employees: There is 1 outlier with the following characteristics:
    - Ticker: 139, Company_Name: 136, Exchange_Name: 3
    - Total_GHG_Emissions: 167000000.0
    - Safety_Incidents: 75.0
    - Workforce_Safety_Incidents: 2.0
    - Employees: 432000.0
    - Women%: 0.0721
    - Methane_Emissions: 428100.0
    - Methane_Intensity_Rate: 0.45
    - Spills: 53.710145
    - Other column values: Varying
5. Women%: There are 7 outliers with the following characteristics:
    - Ticker, Company_Name, and Exchange_Name vary for each outlier

    - Total_GHG_Emissions, Safety_Incidents, Workforce_Safety_Incidents, Employees, Methane_Emissions, Methane_Intensity_Rate, and Spills vary for each outlier
    - Women% ranges from 0.04 to 0.519
    - Other column values: Varying
6. Methane_Emissions: There are 2 outliers with the following characteristics:
    - Ticker: 75, Company_Name: 78, Exchange_Name: 3
    - Total_GHG_Emissions: 205969.0
    - Safety_Incidents: 12.0
    - Workforce_Safety_Incidents: 0.0
    - Employees: 1536.0
    - Women%: 0.089
    - Methane_Emissions: 3015926.0
    - Methane_Intensity_Rate: 0.829492
    - Spills: 1.0
    - Other column values: Varying
7. Methane_Intensity_Rate: There are 3 outliers with the following characteristics:
    - Ticker, Company_Name, and Exchange_Name vary for each outlier
    - Total_GHG_Emissions, Safety_In

Data Transformation:

In the Document a transformation technique has been applied to the dataset, such as scaling(maximin scaling), and label encoding as categorical variables.

The application of data transformation techniques, such as scaling (specifically Min-Max scaling) and label encoding, serves specific purposes and can have various impacts on the dataset. Here are the reasons behind these transformations and their impacts:

1. Scaling (Min-Max Scaling):
 - Reason: Min-Max scaling is applied to transform numerical variables into a standardized range, typically between 0 and 1. This scaling technique is useful when variables have different scales or units, and it aims to ensure that all variables contribute equally to the analysis.
 - Impact: Min-Max scaling allows for a fair comparison and interpretation of variables by eliminating the influence of their original scales. It helps prevent variables with larger values from dominating the analysis and ensures that all variables have a similar impact.

2. Label Encoding:
- Reason: Label encoding is used to convert categorical variables into numeric representations. It assigns a unique numerical label to each category within a variable, allowing algorithms to process categorical data as numerical input.

- *Impact:* Label encoding enables the inclusion of categorical variables in machine learning models that require numerical input. By encoding categories as numbers, algorithms can perform calculations and comparisons on the transformed data. However, it's important to note that label encoding introduces an arbitrary numerical order, which may imply an unintended ordinal relationship among categories.

These transformation techniques enhance the usability and compatibility of the dataset for various analyses. Scaling ensures comparability and equal weightage of variables, while label encoding enables the inclusion of categorical variables in numerical-based models. However, it is crucial to consider the specific characteristics and limitations of each transformation technique to ensure the appropriate interpretation and analysis of the transformed data.

Data Integration: There was nothing about surveys, or experiments involved. Data used in this study was derived from multiple sources, including financial information from SEC filings, ESG values scraped from sustainability reports, and Yahoo Finance (Fig. 3) [Source: Morningstar]. The data collection process involved web scraping, using APIs, and manual data collection from different companies. This data was then cleaned and pre-processed before being utilized for analysis using machine learning algorithms.

Data Privacy and Ethics considerations: Web scraping time privacy and ethical considerations related to the dataset, including anonymization, data protection measures, or compliance with regulations like GDPR have ensured that personal or sensitive information is handled appropriately.

Dataset Usage and Limitations: The dataset was analysed using financial and sustainability indicators with the end goal of incorporating sustainability metrics into machine learning-based risk assessment models. The resulting model would consequently aid in enhancing the calculation of risk-weighted assets and predicting default probability in companies. Nevertheless, the dataset was not devoid of limitations. The main issue was its size, which was relatively small, leading to certain constraints. These included potential biases, limited sample size, and restrictions related to the timeframe of data collected.