Attribution

Table of Contents

[Code inputs – Readme.txt 3](#_Toc24984808)

[Code Process- 4](#_Toc24984809)

[Data Description 6](#_Toc24984810)

[Data – ‘data\_rated.csv’ 6](#_Toc24984811)

[Data- ‘data\_weighted.csv’ 8](#_Toc24984812)

# Code inputs – Readme.txt

* Folder hierarchy –
* We produce Short-Term Score (**STS**) and Long-Term Score (**LTS**) from the Sustainable Development Goal (**SDG**) score computed from Global Knowledge Graph (**GKG**) database by GDELT
* The code scans “Readme.txt” for variable inputs for conditional code execution. The following are the inputs that can be supplied by the user-

|  |  |  |
| --- | --- | --- |
| Variable | Value | Comment |
| Data\_File\_Name | Data\daily\_new\_v3.1\_morgan.csv | Address of merged daily data downloaded from GDELT in its raw form |
| Pool\_File\_Name | Data\MXUS\_POOL\_2.csv | Address of the file containing portfolio and benchmark weights |
| SASB\_File\_Name | Data\SASB\_Data.xlsx | Address of file containing SICS sector information for relevant companies |
| SASB\_Weights | Data\SASB\_Weights.csv | Address of the file containing weights to use for weighting data with respect to SICS sector classification and relevance |
| Market\_Caps | Bayesian/benckmarkListMktCap2019\_12\_13.csv | Address of the file containing company market capitalizations for analysis |
| Pool\_Year\_Start | 2015 | Start year of analysis date range |
| Pool\_Year\_End | 2018 | End year of analysis date range |
| Output\_Data | 0 | Set as ‘1’ when running the file for the first-time a given data.  Set as ‘0’ for repeat execution. |

# 

# Code Process-

## Operations

* Check if *Output\_Data* flag is set. If yes, it means the data is updated or previously processed data doesn’t exist. Then we do-
  + Apply **Bayesian weights** to data such that noisy data has a lower value. Noisy data refers to the data points when number of news articles are low and the variation between the news is high
  + Time heavy calculations on the raw data for getting Short-Term and Long-Term ratings

**Rating Data** – The process of standardizing data to generate comparable information.  
For rating data we perform the following operations-

1. Perform moving average calculation as described in the section [Data – ‘data\_rated.csv’](#_Data_–_‘data_rated.csv’)
2. Re-Scale data to combat the lowered standard deviations as a result of the above calculations
   * **Weighting data** to reflect SASB materiality (SASB materiality simply specifies what SDG values are relevant for each sector). For more detail, refer section [Data- ‘data\_weighted.csv’](#_Data-_‘data_weighted.csv’)
   * Processing portfolio/benchmark allocation data for the date range
   * And, saving this data for future executions or reference.

* We create the following data sets-
* If flag isn’t set, get previously calculated data.

## Attribution

* Start attribution. We must create the following attributions-
  + **Contemporaneous** – Forward looking, i.e., for Yearn we take the constituent weights (which are based on last year’s ESG scores) and compute a weighted average with Yearn-1 SDG scores. This shows if SDG scores warranted the selection of weights
    - **SASB Weighted**- SDG Scores and Short Term Ratings
    - **Non SASB Weighted** - o
  + **Non-Contemporaneous** - Backward looking, i.e., for Yearn we take the constituent weights (which are based on last year’s ESG scores) and compute a weighted average with Yearn SDG scores. This shows the sustainable performance of the portfolio given the weights
* For columns in SDG, MA-7day, and MA-180day, do-
  + Multiply company scores by the weight of that company in portfolio/benchmark
  + For yearly SDG - Group data by year by adding the values for that year (as weighted by normalized weights, the sum results in weighted average)
  + For GCIS Sector-wise SDG – Group data by year and sector by adding values for each sector in each year
  + Save Year over Year SDG graphs as ‘Rated\_Portfolio\_YoY\_ Portfolio/Benchmark\_’
  + For sector-wise data group 17 SDG scores by averaging to an aggregate SDG score, therefore having 1 SDG score for each sector in each year
  + Save these aggregates in files as reference to graphs as ‘Rated\_Portfolio/Benchmark\_SDG\_Yearly/Sector\_’.csv
  + Save yearly Portfolio vs Benchmark bar graphs for SDG and Sectors as ‘Rated\_SDG\_(Sector)\_2015\_’
  + Export the yearly aggregates for both SDG and Sector along with radar charts to excel for better visualization as ‘Rated\_MA\_7day\_Results.xlsx’
  + Repeat these steps for SASB weighted data

*Note: Rated and SASBWeighted data outputs have different folders. Names of the file may differ from the ones used as example above but will follow the same naming convention. For both Rated and SASBWeighted data, there are three types of outputs- SDG, MA 7-day and MA 180-day.*

# Data Description

## Data – ‘data\_rated.csv’

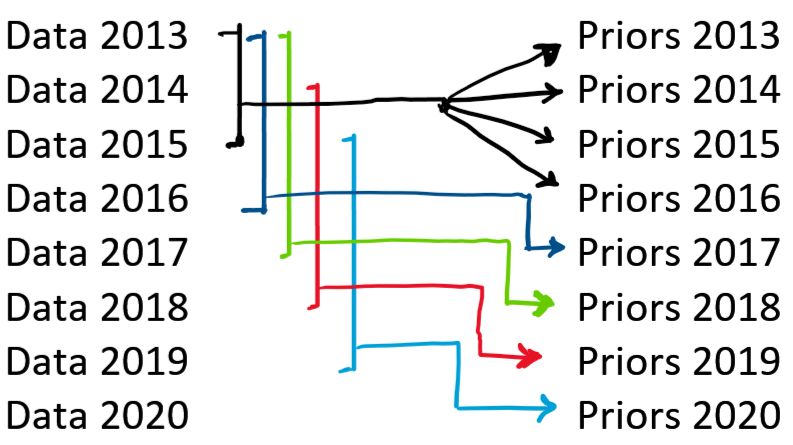
The SDG data in its daily form that reflects Bayesian weights is used to calculate moving averages with window sizes of 7-days and 180-days (MA\_7day, MA\_180day). These averages along with some company information is added back to the original data set as new columns and the file is saved as “data\_rated.csv”. The calculation is performed as shown below-

* First, we use the ‘Count’ and ‘STD’ columns to generate Bayesian weights for each data point. To get Bayesian weighted data we do the following-
  + For each SDG we generate a prior-
  + Then for each data point we can generate a weight by-
  + Here, “Count” is the number of news article used to arrive at a score and “STD” is the standard deviation of the different scores by different articles
  + By applying Bayesian weighting, we effectively reduce the impact of noisy data points (fewer number of articles with very different scores)
  + *Note – For data points with just one news article, we set Standard Deviation to 2 which is a reasonable assumption based on historical data*



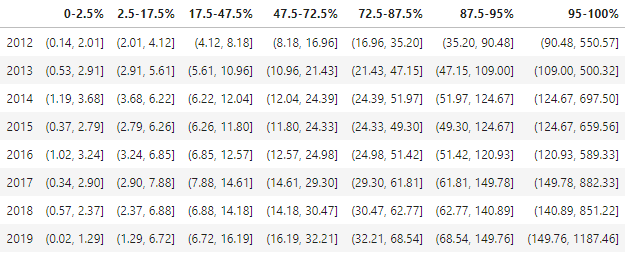
We assigned prior such that for the base case the weight is 0.75

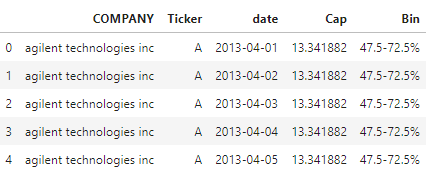
* + As we need to make sure that the ratings once computed do not change, we need to compute priors on a yearly basis before the year starts. We do this as follows-
    - We make use of the GKG-1 data to compute priors in-sample for 3 years and then out of sample
    - We call the number of past years we use to compute the priors for this year the lookback period. This is done to ensure the parameters include past information without getting stale. After much consideration, a lookback period of 5 years was deemed appropriate.
    - We first compute the priors using the first 3 years of data and use it for 1st, 2nd, 3rd, and 4th year.
    - Then 4 years of data for the 5th, then 5 years of data for the 6th, then next 5 years for the 7th (i.e. 2nd to 6th year), and so on.



* + - These priors and then used to compute weights for each data point. They remain the same as more data is added to the database.
* Second, we need to split the data sets according to companies as news from one company is not relevant for the other and thus the moving window average is applied separately to each company’s data.
* Then, for each SDG in each company we need weighted MA\_7day and MA\_180day. Following is an example of calculating MA\_7day for SDG\_3 (labelled as MA\_7day\_3):



* These ratings have been down weighted due to Bayesian scheme and moving average. To interpret these results, it is imperative that we scale these ratings to not only allow interpretation but also cross comparison. Our goal is to scale the scores such that the standard deviation is close to 1.
  + Just like priors, we need a re-scaling criterion that does not change the ratings as more data is added to the data set.
  + Also, it was observed that the variability in small sized companies was higher than the big ones. If we used a common scaling factor, the comparability between these classes would not improve and inadequate.
  + Thus, we use the market capitalization of the companies as an indicator of the news volume. High cap companies will have more news coverage and thus the confidence in the rating will be high. On the other hand, low cap companies will have less coverage and thus low confidence.
  + We start by getting the market caps of the companies each year for the entire period. This will help us update the company classification each year and is necessary as we will need to account for growth of a company.
    - We use WRDS – Compustat – Capital IQ to get the monthly price and monthly shares outstanding for the month of November from 2012 to 2019
    - Due to many reasons there might be some companies that do not have this data for some of the years. (Merger, privatized, unlisted, etc) We need to make reasonable assumptions to fill in the gaps.
      * If we do not have a market cap for a year after having it for any year before, we can reasonably assume it to be the last available value.
      * If we are missing a market cap at the beginning of the period for a company, a reasonable assumption is that the company is small. Thus, when assigning companies to groups (bins) for classification, we just assign these companies to the group representing the smallest sized companies.
    - Then we assign the companies to different bins based on their market caps. These bins are-  
      0-2.5%, 2.5-17.5%, 17.5-47.5%, 47.5-72.5%, 72.5-87.5%, 87.5-95%, 95-100%
      * The bins are defined as quantiles rather than static values to avoid overstuffing of one set of groups and depriving the other. This ensures appropriate scaling.
      * These quantiles are used to define bin edges in terms of market cap values so that companies can be sort into these groups using market caps.  
        
      * Then after forward filling all the missing values, we assign the rest to bin 0-2.5%
    - We then append the bin information to the data. Here we must keep in mind that the bin that was decided based on 2012 cap will be assigned to 2013 data points. This allows us to compute static ratings in 2013 daily without changing the previously calculated ratings.



* + - With yearly updated binning information, we can proceed to re-scale the data. We follow the same principle we did when computing priors. For each year, with a lookback period of 5 years and 3 years of data in-sample we get the standard deviation of each bin.
    - For instance, for 2019’s scaling factors for SDG 1-17 and SDG\_Mean, we take the SDG data from 2014-2018, divide them into bins and compute standard deviation. Then get the scaling factor by computing the inverse and multiplying the 2019 data by these scalars with the respective bins.



* + - For 2019 data, we have the company market caps from 2018 and thus their bins. Based on what group the company belongs to, we pick the scaling factors from the above table and multiply it to the data.
* Lastly, we combine MA\_7day columns, MA\_18day columns and some additional information columns to the data set and save it.

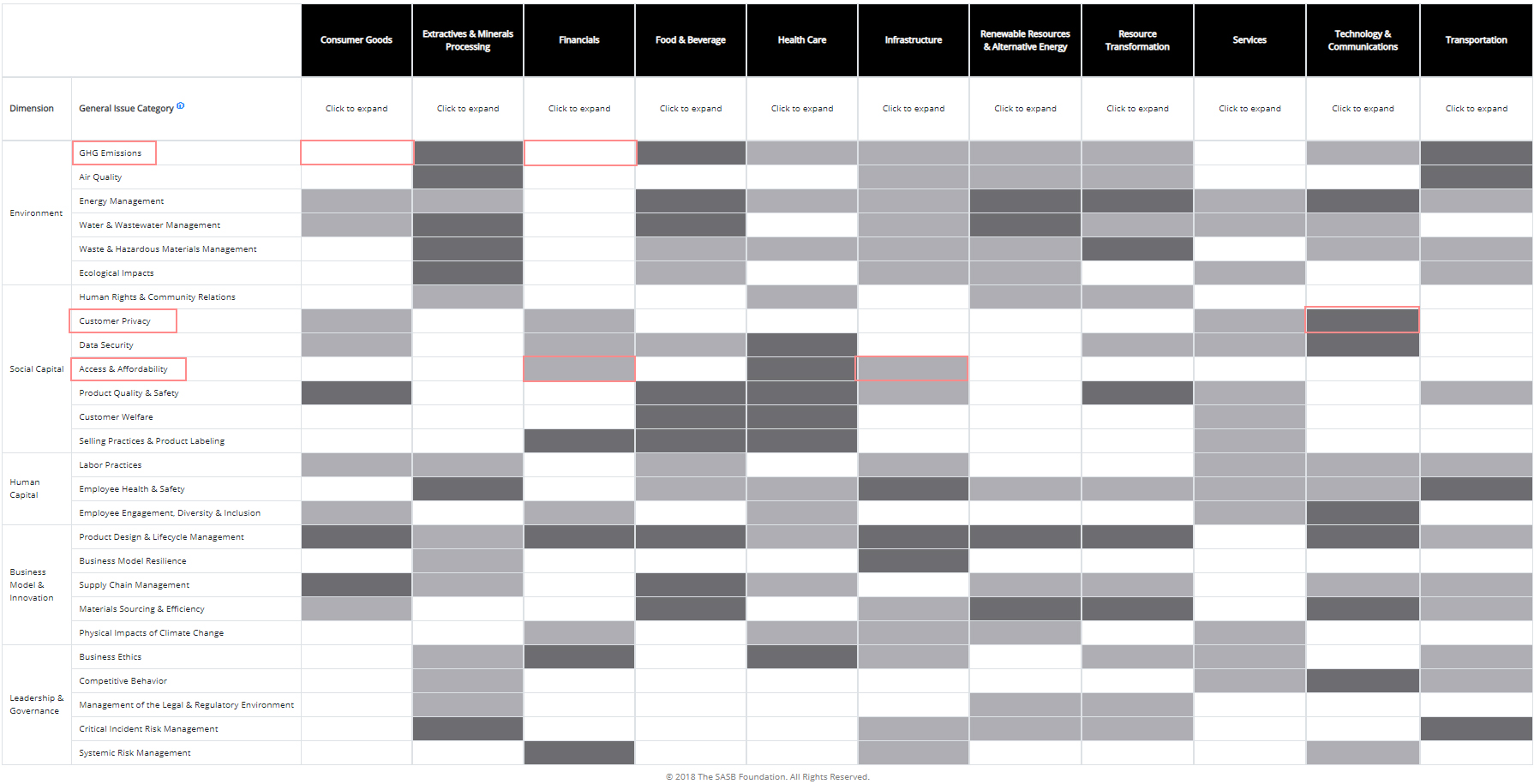
Please find below a short summary of all columns in data\_rated.csv for your reference-

* **‘date’** – Daily date record. Useful in aggregating data, re-indexing after operations and reference
* **‘Ticker’** – Ticker symbols of the company. Helps filter data as more efficient than filtering with company names
* **‘ISIN’** – Security identification number, for users’ reference
* **‘COMPANY’** – Column containing company names used in grouping data along with ‘Ticker’ and ‘date’. Also, for users’ reference
* **‘GICS Sector’** – Each company can be classified into 1 of 11 GICS sectors and this column contains that information. We use this to generate sector-wise attributions
* **SDG\_1…SDG\_17** – Daily scores of companies based on how they impact each sustainable development goal. Is set to ‘NaN’ for days with no news.
* **‘SDG\_Mean’** – To get a combined SDG score of a company for any given day, we calculate the mean of SDG 1-17 and store in this column
* **BW\_SDG\_1…BW\_SDG\_17 –** Bayesian Weighted Daily SDG scores weighted according to the reliability of the score based on the number of news articles that day and the variability of the news
* **‘BW\_SDG\_Mean’ –** Bayesian weighted scores combined for a single day. It’s the mean of the BW\_SDG\_1 to BW\_SDG\_17 columns
* **MA\_7day\_1…MA\_7day\_17** – Contains 7-day weighted moving average of standardized SDG scores as explained above
* **‘MA\_7day\_Mean’** – Like SDG\_Mean, this column contains mean of MA\_7day 1-17 for each day
* **MA\_180day\_1…MA\_180day\_17** – Contains 180-day weighted moving average of standardized SDG scores as explained above
* **‘MA\_180day\_Mean’** – Mean of MA\_180day 1-17 for each day
* **SDG\_std 1-17** – This field records the daily standard deviation of the SDG score for a company. For instance, SDG\_1\_std for ‘Apple Inc’ measures the daily variation in SDG\_1 score for ‘Apple Inc’ in term of standard deviation
* **SDG\_count 1-17** – Measures the number of news articles that were mapped to an SDG for a Company. For instance, SGD\_1\_count will be 3 for ‘Apple Inc’ if for that day there were 3 instances where an article resulted in a SDG\_1 score for ‘Apple Inc’

## 

## Data- ‘data\_weighted.csv’

We calculate these weights for each sector by using this table from this table-

*­­*

This table depicts how material an issue is for each industry sector. **Blanks** – This issue isn’t material for any of the companies in this sector. For example, “GHG Emissions” is not an important driver for companies in Consumer goods, Financials or Services sectors. Therefore, we set the weight for these cells to ‘**0**’

**Light Gray** – This issue is likely to be material for less than 50% of the companies in this sector. For example, “Access and Affordability” is material for less than 50% companies in the Financials and Infrastructure sectors. If we consider a uniform distribution of companies as none is mentioned, we can reasonably assume that 25% of companies in the distribution are affected by this issue. Thus, we assign a score of ‘**1**’ to these cells

**Dark Gray** – This issue is likely to be material for more than 50% of the companies in this sector. For example, “Consumer Privacy” issue materially affects more than 50% of the companies in ‘Technology & Communications” sector. Again, we can reasonably assume 75% companies in the sector are affected. We note that 75% companies are 3x the 25% companies of light gray issues. Keeping relative weights in mind, we set the score of dark grays to ‘**3**’

*Using the following values (blank=0, light gray=1, dark gray=3)-*



The SDG column contains the SDGs that are related to their corresponding issues. So, to get the final score for one SDG, we take the mean of all the issue scores under one sector. For example-

***SDG\_1 score for Consumer goods*** *= Average (highlighted cells)*

Once this is done, we get a table of scores for each sector with each SDG-



This table contains the scores that are on a scale from **0 to 3**. SASB sector weights assign varying level of confidence to each SDG. For example, for the companies in “Financials” sector, SDG (4, 9, 16) are relatively more important than others. For each day, we match the sector information in the data to the above weights. So, if the data row contains “Consumer Goods” as sector, we get that row from the SASB Weights table-



We then proceed to multiply SDG columns with those weights elementwise based on what sector the company is part of. Then we re-calculate the short-term and long-term scores based on these materiality-adjusted scores. The rest of the structure of ‘data\_weighted.csv’ is same as that of ‘data\_rated.csv’.