NEOMA Business School

TRAN Francis
WILLSON Joseph

Alexandre COUREAUD

Sustainable Finance & Investment

Prediction of a workplace wellbeing index

Msc Finance & Big Data 2022-2023

Step 1 : Critical Analysis of data and the chosen model

As mentioned on the subject, the goal of this first step is to analyze the quality of the dataset based on three different points which are its relevance, completeness and accuracy.

Relevance

The goal of the project is to create a well-being index at work. This notion can be defined as a measure to evaluate well-being and the satisfaction of the employee in a company based on multiple factors about the quality and balance of the work life. It can help to identify the strengths and weaknesses of the working environment and how a company can improve the well-being of their employees. To do so, in theory we will need some data about the social environment, employee health, work atmosphere quality and so on. This is the case on the dataset as mentioned on the data description, we have columns about human rights, economic development, reputation risk. That's why we considered that the given dataset is relevant for our study. Now let's focus on the second point of the analysis.

Completeness

There are more than 33 000 rows and 276 columns in our dataset. First of all we checked for the number of nan values of every column. As you can see, there are lots of columns for which the number of nan values is important. That's why we decided to only keep the column for which there were at least 33% of the data because we considered that when the number of nan values of a column was above the ⅔ of the total of our rows, it wouldn't be complete enough for our model implementation. However, we don't exclude the possibility that among the removed columns some of them might have been useful for our prediction. After this data cleaning we dropped approximately 100 columns and for the numerical columns left, we replaced the nan values by the mean of their columns. After this data cleaning, we now have a dataset ready for the implementation of our machine learning model.

HRts2.5 score	33282
HRT2.5 Criterion L score	33282
HRT2.5 Criterion I score	33282
HRT2.5 Criterion R score	33282
SEDOL 2 code	30876
HR3.1 score	29663
HRS3.1 Criterion L score	29663
HRS3.1 Criterion I score	29663
HRS3.1 Criterion R score	29663
ENV2.6 score	29230
ENV2.6 Criterion L score	29230
ENV2.6 Criterion I score	29230
ENV2.6 Criterion R score	29230
HR3.3 score	26161
HRS3.3 Criterion L score	26161
HRS3.3 Criterion I score	26161

Columns with the most NAN values

Research Questions

How often ESG data is updated and when it is updated ?

The frequency of the data updates can impact the prediction of our well-being index because we know that ESG factors may have significant impact on the employee well-being. Let's consider a company for which ESG data are only updated annually then it might not reflect recent changes or risk incidents which can impact employee well-being. Then our machine learning model might not be accurate based on outdated ESG data. On the other hand, if the ESG data of a company are updated quarterly, machine learning might not reflect enough changes in employee attitudes during the intervening months between every update. So results can be inaccurate and can lead to wrong decision-making.

How to deal with different historical periods ?

By having a dataset of different periods, we will compute the same calculations of our index and compare the results for the different years. We expect to observe some variations of our index which mean that the time period can have a significant impact on our prediction. The most important is to take the same column on every dataset for the calculation of our workplace wellbeing index.

Treatment of securities with financial data and without ESG data

As the goal of this project is to define an workplace index based on the ESG data, we've decided not to exclude any ESG data for the analysis as they mainly composed our dataset. We thought that by excluding them, the results of our prediction won't be accurate enough to be interpreted. So we've kept treating the securities with the ESG data.

The history of the securities making up the index not available

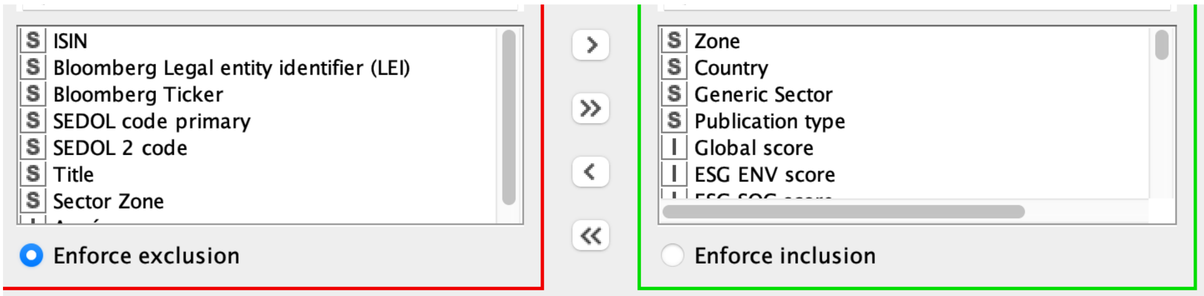
If the predictions are based on the history of security then not having access to this historical data can be problematic. First of all, the index prediction may not be accurate to represent the current state of the market and can lead to mistakes when we need to make decisions about investments. Secondly, it could be difficult to identify trends for the data which could limit us to create predictions about the market. Finally the non-access to the historical data can create difficulties to access to the global performance of our index or to make comparisons with benchmarks.

KNIME

We've checked the software to analyze the existing model which is a random forest. The input variables are the ESG data of 2018 and 2019. Then all the data manipulation is explained on the software (conversion string to number, deal with missing values…) and the output are excel files to store the predictions.

Input file of the KNIME model

With KNIME, to evaluate the performance of the model, we can check the case "Scorer" which gives the confusion matrix and accuracy stats. Depending on the results we can know if the model is appropriate or not. Also we can check the variables selected for the implementation of the model, for example there are the country, zone and the Bloomberg Ticker were excluded.



Selected columns for the implementation of the KNIME model

Interpretation of the results

Now, let's interpret the output return by the existing model on the Accuracy statistics excel sheets.

RowID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
++	14	37	165	37	0,274509804	0,274509804	0,274509804	0,816831683	0,274509804		
+	11	39	164	39	0,22	0,22	0,22	0,807881773	0,22		
ø	23	28	137	65	0,261363636	0,450980392	0,261363636	0,83030303	0,330935252		
-	2	48	192	11	0,153846154	0,04	0,153846154	0,8	0,063492063		
--	17	34	168	34	0,333333333	0,333333333	0,333333333	0,831683168	0,333333333		
Overall										0,264822134	0,080341613

Results tables of the KNIME model

The results of our classification were the different labels (++,+,...) were assigned to different instances with the criteria. We have the TruePositives, FalsePositives, TrueNegatives and FalseNegative which represent the number of values which were correctly predicted as positive, wrongly predicted as positive, correctly predicted as negative and wrongly predicted as negative. Then the recall, precision, sensitivity, specificity, F-measure, accuracy and Cohen's kappa are the KPI to evaluate the performance of our model. The F-measure represents the mean of the precision and recall, the accuracy is the proportion of correct prediction to the total number of instances. It is the overall performance of our model. Finally, Cohen's kappa is a KPI that evaluates the agreement between the observed value and expected prediction by taking into consideration the possible random chance.

When we look at the results, we can see that the model performs poorly overall because the accuracy is only equal to 0.08. The others KPI show diversified results for the different labels but no one showed a consistent good performance. The Cohen's kappa is very low which represents a poor agreement between the real and expected predictions.

We can conclude that the model must be improved by using different input or another machine learning model. Also it's possible that with a better dataset, the measures may have better values.

Now we will see the results of our confusion matrix to proceed to a second performance evaluation of our model.

RowID	++	+	Ø	-	--
++	14	13	10	7	7
+	19	11	9	7	4
Ø	12	21	23	13	19
-	1	2	4	2	4
--	5	3	5	21	17

Confusion matrix of the KNIME model

Let's define the following term such as :

1. VP are the correct predicted value of a positive class (++)
2. FP are the wrong predicted value of a positive class whereas it is in fact another class
3. VN are the correct prediction of a negative class (--)
4. FN are the wrong prediction of the negative class whereas it is in fact another class

We can see that the model predicted correctly 14 values of the VP but failed to predict 19 values of class + whereas it was FP. Then for example the model correctly predicted 17 values of class – as VP but failed to predict 5 values of – as FP. We can see that the confusion matrix proved that our machine learning model managed to make some accurate predictions for the ++ and – labels but didn't make some accurate predictions for other classes. The number of false positives for the Ø and - labels is important which can help to identify the weaknesses of our model and to define some improvement.

Creation of our workplace wellbeing index

With the 186 columns left, we will define the formula of our workplace wellbeing index. We have some columns about the Human resource domain, environmental strategy, Business behavior domain for example. For each of this column category data, we have a score from 0 to 100 about the leadership, implementation and results of these categories. We've considered that for the creation of our index, we will define the same weight importance for every category. We've selected 7 columns which are : "Global score", "HRS Domain score", "ENV score", "C&S score", "CIN score", "CG score" and "HRts score". To define a relevant index, it was necessary to take into consideration multiple different data. The workplace wellbeing is not only about the ESG data or the physical health and safety of the employees but also about how they are treated in the company's life.

Human rights is an important concept which protects people from exploitation, discrimination, racial abuse… It includes freedom of speech, the possibility to participate and elaborate decision-making for the company… If employees' human rights are well respected then the promotion of workplace wellbeing will be easier. For Corporate governance, it is linked to the practices and processes implemented to control and guide the company. With good corporate governance, the firm can be managed in a transparent and ethical way so that a culture of trust and respect can be well developed. For community involvement it refers to how a company engages with different communities for which it operates. (partnerships, volunteering…). It is important for the implementation of our index because it can build a positive relationship for the company with other communities. Then for the employees, it can create a trust and loyalty atmosphere thanks to their contribution. In addition, it can reflect social and environmental issues which can lead to a positive impact for the health and wellbeing of them.

We created a new column called 'Workplace index' and we computed the mean of our 7 columns.

Accuracy

We considered that the accuracy was about the result of the implemented machine learning model. We've created a new column called 'global index' which is the value we want to predict and we implemented a random forest by taking for the X variable all the numerical columns. We obtained a R-squared of 0.99, a MAE of 0.30 and a RMSE of 0.39. The R-squared represents the proportion of variance in our Y variable explained by the X variable in our machine learning model. As the value is close to 1, it means that all of the variance of Y is explained by X which is a good point. The Mean Absolute Error measures the average of absolute difference between our prediction and the actual values of the Y variable. Its calculation is based on the mean of the absolute differences between the prediction and actual value. The Root Mean Squared Error measures the standard deviation of the difference between our prediction and actual values of our Y variable. Its calculation is based on the square root of the mean of the squared differences between the prediction and actual value. In our case, the MAE and RMSE are low which indicate a better performance of our model because they represent small errors between our prediction and actual values.

Step 2 : ESG rating criteria classification

For the second step of this project, we will go into more detail about the implementation of our model and more precisely we will focus on the variables which have the more impact on our index prediction. To do so, we will first compute a feature importance technique. This is a measure about the importance of each feature in our Y variable prediction. It will help us to identify which columns are the most relevant in our case. The calculation is based on the reduction in impurity obtained from splitting a specific feature at each node of the decision tree. Features which result in large reductions of impurity will be considered as the most important. The feature importance values are finally normalized to obtain a value between -1 and 1 which represents the proportion of prediction power of our model attributed for every feature. A high feature importance includes that the feature is very important in the prediction of our target variable.

```
Global score: 0.9807890481646225
CIN score: 0.009598647388882522
HRts score: 0.0015021305288057188
ESG GOV score: 0.0004115758179601284
HRS Domain score: 0.00036470226387107593
CG score: 0.0003612068972878565
CG I-score: 0.00030312167310139244
HRts L-score: 0.00030097189682016764
ENV2 Weight: 0.0002572922673121176
HR L-score: 0.0002503818921292626
```

Top 10 most important features

We can see that some of the columns defined in our index formula have the most significant impact on our prediction (Global Score, CIN score…). On the other hand, it is interesting to focus the column about the Leadership and implementation (CG I-score, HRts L-score and HR L-score). The implementation score for the corporate governance domain is an extent to which the company has or not implemented good governance practices. This can impact the workplace index. The leadership score for the human rights domain is the company's capacity to respect human rights for their employees and also contribute to the prediction of our index.

On the other hand the less significant variables are for example the HRts 2.5 Weight, CG3.1 Weight. These variables were not considered on the formula of our index and as we considered an equal weight for every term of our formula the weight column won't have a huge feature importance.

In addition, we've implemented permutation importance. This technique assesses the importance of every feature and the calculation is based on a random permutation of the values for every feature on the dataset and by calculating the decrease of the model's performance. The measure is then based on the decrease caused in the model's performance by the permutation of that feature. In more detail, we've trained our model on the training dataset, we evaluated the model's performance. Then for every feature, we randomly permute its values on the dataset and we calculate the model's performance on this new permuted dataset. Finally we subtract the permuted measurement with the original performance to have the decreased performance and we rank the features. We can see that the results obtained are similar to the previous method.

ENV score	0.000207733258072329
C&S score	0.00019719649505639536
HRts I-score	0.00018813316431870453
ESG SOC score	0.00017374555564275695
HR I-score	0.00016519764190764397
ESG ENV score	0.00012752185691022877
C&S L-score	0.00011709138970178445
HRts R-score	0.00011600976759357229
ENV I-score	0.00011133701624893577
CIN I-score	0.00010206115991545639

Top 10 most important features (permutation importance)

To finish with, let's compute the correlation matrix of our index to see if the more correlated features are the one with the highest feature importances. This is the case here.

Global score	0.9908351929123588
ESG SOC score	0.9119314434892979
HRS Domain score	0.8655253766840337
ENV score	0.8593018063751793
HRts score	0.858771183333268
ESG ENV score	0.8457112822218996
C&S score	0.8380503601185245
HR I-score	0.837600421608724
ENV I-score	0.8361470800159095
C&S I-score	0.8355733168222191

Score correlation of our index

Analyze of the characteristics

On this part, we will analyze the importance of some characteristics in the prediction of our index. First of all let's see how the country can impact the value of our index. We've computed the mean of the index predicted for every country. Depending on it, the governance policy, environment implications of the company can be different. That's why we obtained such diversified value (15 to 62). This is the top 10 country with the highest value of our index.

Country	Workplace index
Ivory Coast	62.785714285714285
France	46.45479049252634
Netherlands	43.22674047264211
Portugal	42.15824175824176
Finland	41.60269214068606
Italy	40.48427476138319
United Kingdom	40.20212102308172
Spain	39.90699782451052
Germany	39.76601904761905
Norway	39.0301724137931

Top 10 country with the highest index

Now let's see the importance of the activity sector. We can see that the Gas & Water Utilities and Development Banks have the highest score index. These sectors have a good index because of their job stability (gas sector is often considered as a stable sector with a long-term employment same as the development banks). Also there is competitive compensation (high salaries and benefits) and a good work-life balance proposed to the employee. These reasons can highly contribute to a good score for the index.

Sector Zone	Workplace index
Gas & Water Utilities	58.06493506493507
Development Banks	53.789610389610395
Specific Purpose Banks & Agencies	51.79140328697851
Diversified Banks	51.64273946811014
Air Transport	51.22857142857143
Diversified Banks Asia Pacific	49.68377253814147
Waste & Water Utilities	49.5527950310559
Building Materials Latin America - Added	48.285714285714285
Mining & Metals Latin America	47.714285714285715
Forest Products & Paper	47.15714285714286

Top 10 sector with the highest index

Now let's see the impact of the size of capitalisation of the company for our index prediction. To do this, we will compute the mean of our index for every ISIN code.

FR0004125920	69.6
AU0000009771	68.61904761904762
XS0993154748	68.08163265306122
FR0013451333	68.0
FR0000125924	66.85714285714286
FR0010725549	65.95714285714286
XS0213876146	65.74285714285715
GB00B05KYV34	65.71428571428571
FR0000476087	65.60714285714286
IT0001415246	65.34285714285714

Top 10 ISIN code with the highest index

Respectively these ISIN code corresponds to the following companies : AMUNDI, UNIBAIL, FMO, FDJ, AGF, Caisse des dépôts, EIB, PLC, LAPOSTE, FINANTIERI.
We can see that there are some French companies on this list and most of the companies are in the financial sector which confirms the previous results obtained of the decomposition of the sector by index. Finally we will compute the same analyzes with historical data based on different years (2020-2021-2022). We followed the same method except that the HRts score column is not available on these dataframe so we only took the 6 others columns and calculated their mean for the creation of our index.

Pays	First Data set	déc-22	déc-21	déc-20
Ivory Coast	62.79	68.33	68.33	68.33
France	46.55	47.4	46.27	43.18
Netherlands	43.23	47.4	44.87	42.17
Portugal	42.16	51.12	50.37	47.48
Finland	41.60	46.85	45.93	42.98
Italy	40.48	47.46	45.54	40.64
Inited Kingdor	40.20	39.74	38.81	36.49
Spain	39.91	48.47	46.9	43.33
Germany	39.77	40.72	40.18	37.41
Norway	39.03	41.62	41.49	36.38

Variation of our index for the top 10 country

We can see that from one year to another, there is some variation on our global index which means that time trends can influence the prediction of our workplace wellbeing index.

Can ESG data be used to forecast Workplace Wellbeing ?

We know that ESG data is highly used to evaluate the performance of a company and its impact on the society and environment. In addition, the workplace is a key element of the social part of ESG and it is important for the company to be interested in the forecasting of workplace wellbeing. By identifying the key factors, it can help to implement some good predictions. There is a huge relationship between ESG data and workplace wellbeing because the ESG data give important overview about the social impact on their employees. It includes employee health, safety for example. They can be used to measure in which way companies are addressing issues and how they are or are not trying to create a positive workplace for their employees. If we consider a company with a high turnover rate, then there might be a negative impact on the workplace. Research showed that companies prioritizing employee wellbeing were more likely to tend to better financial performance over the time. The study by Harvard Business Review demonstrated that when the employee engagement was high then the total shareholder return was 22% higher compared to other companies.

In addition, another study proved that with a strong commitment for diversity represents a 2.3 times higher cash flow for every employee than companies without strong commitment. Based on this relationship between the data, we can say that ESG data can forecast a workplace wellbeing index. This is the main idea of our work. We selected an ESG dataset from different companies and implemented our own formula based on different scores such as the environment, social, human rights. After the computation of our Random Forest model, we've seen that performance metrics were giving very accurate results meaning that the predictions are consistent. We also declined our analysis based on different characteristics such as the country, sector and showed that the global index can have huge variation based on these variables. However ESG data can often be self)reported by companies so we can have in our dataset lots of inconsistencies and inaccuracies. Also ESG data are not standardized so every company can use their own metrics to measure similar factors which is then difficult to make some relevant comparisons. Finally the implementation of workplace wellbeing is complex as the number of factors influencing this score is high.

We can say that yes ESG data can help to forecast workplace wellbeing as the relation between ESG data and workplace wellbeing is strong, however, there are some limitations which need to be overcome to obtain some accurate prediction. It is important to check the quality of the data used and to implement the appropriate machine learning model.

Critical opinion of our work

The first step of our research was to do some data cleaning on the dataset. By removing the column with the most important NAN values, we've considered only keeping the most complete column and we replaced the NAN values by the mean of their column. On the other hand among the columns deleted such as the HRT2.5 score, C&S 2.2 score which are sub-criteria of ESG data, maybe some of them would have been more useful in the prediction of our index if they would have been more complete. Then in the definition of our index formula we simply computed a mean based on the different columns used but we didn't take into consideration the different weights available on the dataset such as the HR Weight or ENV Weight… We considered that the Environment, Social and Governance Part of the ESG had the same balance on the formula. Again if maybe we focus more on the social part by a more important pondering on the formula, then the value of our index prediction would have been different.

Moreover, for the choice of the Machine Learning model, we focused our implementation on a Random forest because it can handle huge datasets with lots of features and observations, it can deal with missing data without the need of imputation or can handle non-linear relationships. Maybe with different models, the prediction would have been different. Finally, by adding more external data sources, it would have been easier to determine if our predictions are accurate or not, especially for the country analysis. The cultural, economic and social factors which can influence employee wellbeing can be different from one country to another. For example, some countries are more focused on the cultural emphasis on work-life balance and on the other hand, some of them are focused on productivity and achievement. That's why by including these differences in our work prediction (GDP per capita, financial data), there could have been some adjustment about our random forest prediction.

To conclude, we are still satisfied with the work we've implemented and presented in this report as we were able to deal with the dataset and showed some accurate predictions. We are totally aware about all the possible ways to improve our index prediction as they were mentioned in this report.

ANNEX

Data Cleaning and calculation of our index

```
nan_counts = df.isna().sum()
nan_counts
#nan_counts.to_csv("nan_counts.csv")
```

```
Vigeo Key                                   0
ISIN                                        0
Bloomberg Legal entity identifier (LEI)  4673
Bloomberg Ticker                         6780
SEDOL code primary                       4068
                                         ...
ENV Domain rating                           0
C&S Domain rating                           0
CIN Domain rating                           0
CGV Domain rating                           0
HRT Domain rating                           0
Length: 276, dtype: int64
```

```
mask = nan_counts < 10000
# create a new dataframe with only the selected columns
new_df = df.loc[:, mask]
new_df
```

	Vigeo Key	ISIN	Bloomberg Legal entity identifier (LEI)	Bloomberg Ticker	SEDOL code primary	
0	DE0005545503	DE0005545503	5299003VKVDCUPSS5X23	1U1	5734672	1&
1	DE0005545503	DE0005545503	5299003VKVDCUPSS5X23	1U1	5734672	1&1 Dr
2	DE0005545503	DE0005545503	5299003VKVDCUPSS5X23	1U1	5734672	1&1 Dr

```
#creation of our index
new_df['Workplace index'] = ((new_df['Global score'] + new_df['HRS Domain score'] +
                              new_df['ENV score'] + new_df['C&S score'] +
                              new_df['CIN score'] +new_df['CG score'] +
                              new_df['HRts score']
                                 )/ 7)
```

Implementation of our Random forest

```
import pandas as pd
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error

# Separate the X and Y variables
X = new_df.select_dtypes(include='number').drop('Workplace index', axis=1)
Y = new_df['Workplace index']

# Fill missing values with the mean of their respective columns
X = X.fillna(X.mean())

# Create a random forest regression model and fit it to the data
model = RandomForestRegressor(n_estimators=100, random_state=42).fit(X, Y)

# Predict Y using the trained model
Y_pred = model.predict(X)

# Compute and print the evaluation metrics
r2 = r2_score(Y, Y_pred)
mae = mean_absolute_error(Y, Y_pred)
rmse = mean_squared_error(Y, Y_pred, squared=False)

print('R-squared:', r2)
print('MAE:', mae)
print('RMSE:', rmse)
```

```
R-squared: 0.9994372968976696
MAE: 0.18729845968344577
RMSE: 0.2981887638276105
```

```
#definition of the feature importances
importances = model.feature_importances_
importance_dict = dict(zip(X.columns, importances))
sorted_importances = sorted(importance_dict.items(), key=lambda x: x[1], reverse=True
for feature, importance in sorted_importances:
    print(f'{feature}: {importance}')
```

```
Global score: 0.9807890481646225
CIN score: 0.009598647388882522
HRts score: 0.0015021305288057188
ESG GOV score: 0.0004115758179601284
HRS Domain score: 0.00036470226387107593
CG score: 0.0003612068972878565
CG I-score: 0.00030312167310139244
```

```
#permutation importance
from sklearn.inspection import permutation_importance

result = permutation_importance(model, X, Y, n_repeats=2, random_state=42, n_jobs=-1)

# Create a dataframe to display the permutation importances
importances = pd.DataFrame({'Feature': X.columns, 'Importance': result.importances_mean

# Sort the dataframe by importance in descending order
importances = importances.sort_values('Importance', ascending=False)

# Print the permutation importances
print(importances)
```

```
              Feature    Importance
1        Global score  1.369167e+00
21          CIN score  3.290827e-02
29         HRts score  3.983211e-03
4       ESG GOV score  7.479954e-04
5    HRS Domain score  7.470952e-04
```

Calculation of the country and sector index

```
import pandas as pd

# assume 'new_df' is the name of your DataFrame
grouped = new_df.groupby('Country')['Workplace index'].mean()

# print the result
print(grouped)
grouped.to_csv('country.csv')
```

```
Country
Australia                   32.538006
Austria                     33.892711
Belgium                     37.122231
Bermuda                     30.623377
Brazil                      32.960357
                              ...
Turkey                      31.288866
United Arab Emirates        20.175510
United Kingdom              40.202121
United States of America    31.718290
Venezuela                   29.964286
Name: Workplace index, Length: 61, dtype: float64
```

```
import pandas as pd

# assume 'new_df' is the name of your DataFrame
grouped = new_df.groupby('Sector Zone')['Workplace index'].mean()

# print the result
print(grouped)
grouped.to_csv('sector.csv')
```

```
Sector Zone
Aerospace                                   39.053655
Aerospace Asia Pacific                      23.441558
Aerospace Emerging Market                   21.969925
Aerospace Middle-East Africa                27.857143
Aerospace North America                     29.170732
                                              ...
Travel & Tourism North America              30.000000
Waste & Water Utilities                     49.552795
Waste & Water Utilities Asia Pacific        30.321429
Waste & Water Utilities Emerging Market     33.311224
Waste & Water Utilities North America       35.817460
```

Using dataset of different time period

```
# Implementation for different years
df2020 = pd.read_csv("202012.csv",delimiter = ';', encoding='latin-1', error_bad_lines=
df2021 = pd.read_csv("202112.csv",delimiter = ';', encoding='latin-1', error_bad_lines=
df2022 = pd.read_csv("202212.csv",delimiter = ';', encoding='latin-1', error_bad_lines=
```

...

```
df2020['Workplace index'] = ((df2020['Global score'] + df2020['HRS Domain score'] +
                              df2020['ENV score'] + df2020['C&S score'] +
                              df2020['CIN score'] +df2020['CG score']

                                 )/ 6)
```

```
df2021['Workplace index'] = ((df2021['Global score'] + df2021['HRS Domain score'] +
                              df2021['ENV score'] + df2021['C&S score'] +
                              df2021['CIN score'] +df2021['CG score']

                                 )/ 6)
```

```
df2022['Workplace index'] = ((df2022['Global score'] + df2022['HRS Domain score'] +
                              df2022['ENV score'] + df2022['C&S score'] +
                              df2022['CIN score'] +df2022['CG score']

                                 )/ 6)
```

```
import pandas as pd

# assume 'df2022' is the name of your DataFrame
grouped = df2020.groupby('Country')['Workplace index'].apply(lambda x: round(x.mean(),

# print the result
print(grouped)

# export the result to a CSV file
grouped.to_csv('country2020.csv')
```