What Are The Commonalities of Successful Social Enterprises?

-- An Empirical Study Based on B Corp Impact Data

Peijin Li

December 2022

Executive Summary

This paper focuses on 5,210 B corporations and aims to answer the question of the common features of successful social enterprises. By studying the common features of successful social enterprises, this paper successfully depicts the picture of B Corps' status quo and their impacts. To capture the features of successful social enterprises, I classify them based on their performance with K-means Clustering and use TF-IDF to find the mutual features of social enterprises in the same cluster. In conclusion, I find significant differences in governance, environmental impact, and community engagement among social enterprises in different industries and countries. For example, social enterprises in the Manufactured Goods, Retail, Wholesale, and Agriculture sectors perform poorly in terms of employee satisfaction assessments; social enterprises in the service sector perform poorly in terms of environmental concerns and lack sustainability concerns. This study provides a reference for policy development for social enterprises and policymakers in different sectors.

1 Introduction

Social enterprises (SEs) have attracted increased attention from researchers and practitioners worldwide. Broadly, SEs are organizations that primarily aim to solve pressing social/environmental problems (e.g., homelessness, youth unemployment, and carbon emissions) while engaging in commercial activities (either partially or fully) to support their operations (Powell, M., 2018). B Corp refers to those social enterprises awarded the B Corp Certification. B Corp Certification is a designation that a business meets high standards of verified performance, accountability, and transparency on factors from employee benefits and charitable giving to supply chain practices and input materials[i].

In the SE context, expanding social impact is considered the main currency or key performance indicator. Moreover, scholars have been studied this topic using different research methods. Briga Hynes used an exploratory study involving four case studies of established social enterprises to learn the key factors influencing the increase of SE impact (Briga Hynes, 2009). Syrus Islam conducted a systematic review of 111 peer-reviewed articles on the topic of SE development. The paper identifies and discusses key insights into organizational growth strategy and ecosystem growth strategy as a means to scale social impact in SEs (Syrus Islam et al., 2021). Michiel van Rijn employed data from the second global and harmonized assessment of social entrepreneurial activity, the Global Entrepreneurship Monitor Adult Population Survey (GEM APS) of 2015, to analyze the measurement of SE impact (Michiel van Rijn et al., 2021).

[i] In order to achieve certification, a company must: 1) Demonstrate high social and environmental performance by achieving a B Impact Assessment score of 80 or above and passing our risk review; 2) Make a legal commitment by changing their corporate governance structure to be accountable to all stakeholders; 3) Exhibit transparency by allowing information about their performance measured against B Lab's standards to be publicly available on their B Corp profile on B Lab's website (B Corp, 2022).

In conclusion, I find a large number of case studies in the study of SE impact, but few use data from open SE evaluation platforms to explore the patterns of social enterprise impact. The disadvantage of case study is the bias and sampling problem. Using data from the open SE evaluation platform avoids this issue by considering all B Corps in the world, which enables more general conclusions. In terms of research methods, a lot of previous studies employ qualitative research methods, and few use unsupervised research methods such as K-means to conduct research in this field; therefore, this study is innovative in the research data and research methods employed.

By conducting this research, I hope to answer the following questions (1) Are there any subgroups of SEs under the B Corp five dimensions standard? (2) What are the common features of SEs in different subgroups?

2 Data

The data employed in this study come from B Corp Impact Data (B Corp, 2022). This study uses a sample of 5,210 B Corporations. The variables included about these B Corporations are its headquarters, certification issue time, industry, B impact overall score, B impact index and a piece of text describing its mission, work, and story. The data I use for this project is the B impact scores, which include five dimensions: governance, workers, community, environment and customer (B Corp, 2022). Governance assesses a company's overall mission, engagement, ethics and transparency around its social/environmental impact. This section also assesses a company's ability to protect its mission and formally consider stakeholders in decision-making through its corporate structure or corporate governance documents. Workers evaluates a company's contribution to the financial security, health and safety, wellness, professional development, engagement and satisfaction of its employees. In addition, this section recognizes business models that are designed to benefit workers. Community assesses a company's involvement and impact on the communities in which it operates, employs, and sources. In addition, this section recognizes business models that are designed to address specific community issues. Environment evaluates a company's overall environmental management practices and their impact on air, climate, water, land and biodiversity. Customer assesses how well a company manages its customers through the quality of its products and services, ethical marketing, data privacy and security, and feedback channels. In addition, the section recognizes products or services that are designed to address specific social issues for or through customers (B Corp, 2022).

From the descriptive statistics result, we can see that the mean of overall score is 93.50, and all of the 5,210 B corporations have an overall score. There are some missing data for the customers and workers variables. However, only a few observations have the issue, it would not cause a problem for the analysis. In the modeling process, I transfer these null values to 0. The mean value of community, customers, environment, governance and workers are 28.02, 11.93, 17.57, 13.75 and 25.26 respectively. The standard deviation of community, customers and environment indicate that SEs are dispersing on these aspects; while the small SD of governance and workers tells the aggregation pattern of SFs on these two indicators (Table 1).

Table 1 Descriptive Statistics of B Corp Impact Index

	Overall Score	Community	Customers	Environment	Governance	Workers
count	5210	5210	4989	5210	5210	4666
mean	93.51	28.02	11.93	17.57	13.75	25.26
std	13.90	13.85	13.94	12.76	4.08	8.23
min	78.2	3.8	0	0	2.6	0
25%	83.2	18.7	2.3	7.6	10.6	20.6
50%	89.05	23.8	4.6	12.7	14.3	24.9
75%	99.4	33.6	19.1	26.1	16.7	29.88
max	183	125.4	79.4	96.9	24.5	72.1

The histograms expose the distribution of the variables of B impact overall score, community, customers, environment, governance and workers. We can learn that governance and workers follow a normal distribution pattern. For B impact overall score, community, customers and customers, there is a tail on the right side of the distribution, indicating a positive skewness value for these four indices (Figure 1&2).

Figure 1 Distribution of B Corp Impact Index: Overall Score, Community & Customers

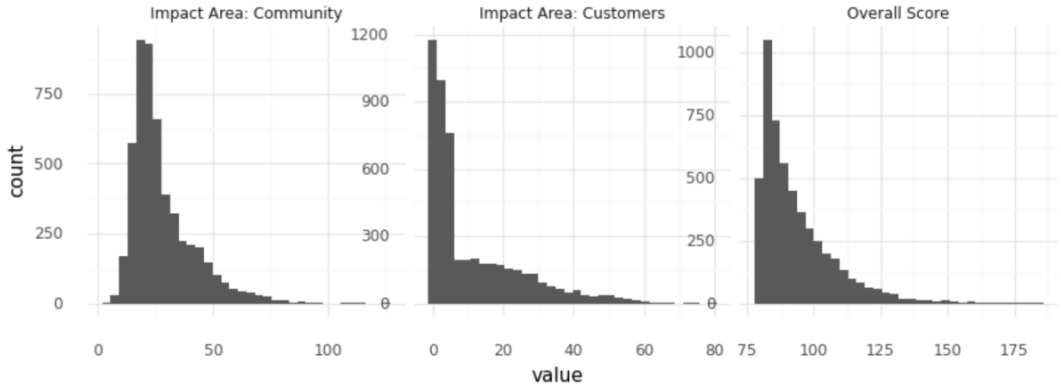
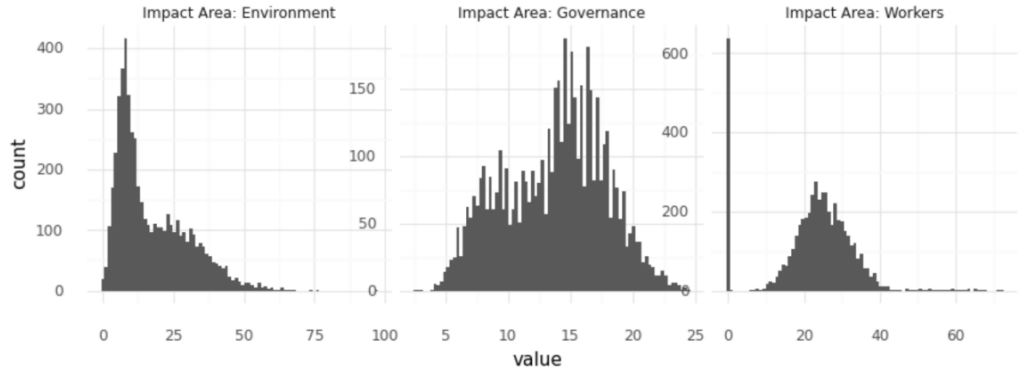


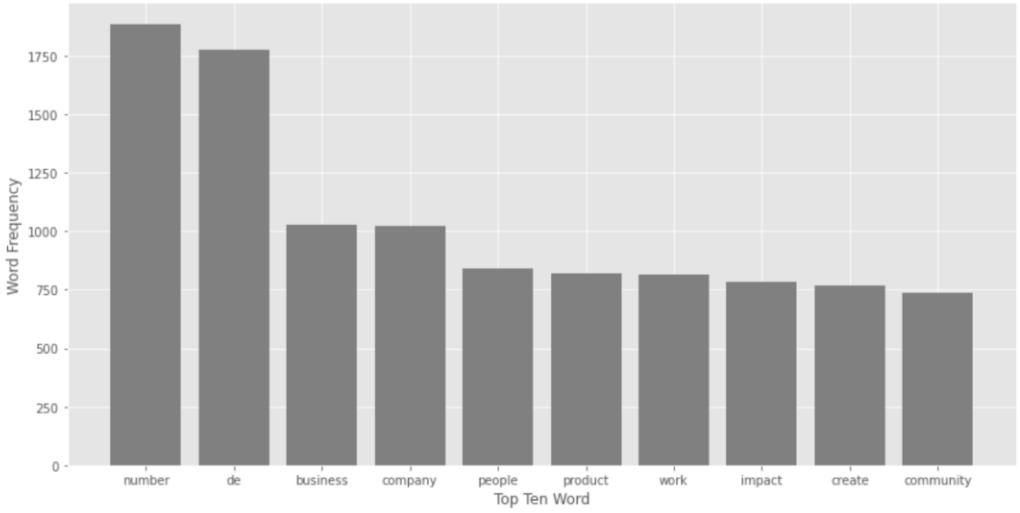
Figure 2 Distribution of B Corp Impact Index: c, Governance & Workers



In order to extract the features of SEs' impact improvement, I employ a simple count matrix and Term Frequency-Inverse Document Frequency (TF-IDF) to represent the description text of each SE. The simple count matrix summarizes the number of times each word appears in the SE description text. This method weights words based on the frequency of a single, which has the limitation of applying to texts of different lengths. TF-IDF makes up for the shortcomings of the count matrix by weighting words based on their relative importance. The TF-IDF value increases proportionally to the number of times the word appears in the document. It is offset by the number of documents containing the word in the corpus, which helps to adjust for the fact that certain words typically appear more frequently (Wikipedia, 2022).

In the memo, I list and visualize the top ten words based on words' frequency (Figure 3). Before conducting the text analysis, I preprocess the text pieces with regularization and remove the stopwords. Figure 3 shows that these top ten frequent words are in the context of social enterprise field, indicating the pre-processing of data has achieved good results.

Figure 3 Most Common Words



3 Methodology

After conducting the descriptive statistics of the numeric variable and preprocessing the text data. Then I will use K-means clustering to aggregate the SEs with similar features. I will employ Principal component analysis (PCA) to realize dimensionality reduction for the five B impact metrics. This step is for the sake of subgroup visualization and classification validation. After that, I use text mining technique to extract the features of sub-grouped successful SEs.

3.1 Analytical Plan

3.1.1 Principal Component Analysis

Principal Component Analysis (PCA) is a popular technique for analyzing large data sets containing many dimensions/features per observation to improve data interpretability. This method helps to retain the maximum amount of information and visualize multidimensional data. By projecting the metrics on a new vector, PCA creates a new vector that maximizes the variance (captures the most features). In the study of global economic policy, FeiDai Peng used PCA to construct a global economic policy

uncertainty index by downscaling economic policy uncertainty indices for twenty primary economies worldwide. This PCA-based index is proven to be positively related to the volatility and correlation of the global financial market, showing its interpretability and consistency (Dai, P.-F et al., 2021). Abdoulkarim Esmaeili and Zainab Shokoohi used PCA to downscale food prices to study the effect of oil prices on world food prices. The study confirms an indirect effect among oil price and food price principal components (Esmaeili, A., & Shokoohi, Z., 2011). In this study, PCA plays a role in Silhouette analysis, it helps to reduce the dimensions for the sake of data visualization.

3.1.2 K-means Clustering

The K-means clustering aims to divide n observations into k clusters, where each observation belongs to the cluster with the closest mean (cluster center or cluster center of mass), as clusters. This step helps to classify SEs by aggregating SEs with similar features. In this project, I would employ Euclidean distances to calculate the similarity of observations. K-means is a computationally expensive method, so the previous downscaling step plays an important role. K-means clustering is relatively simple and easy to implement. It convergent fast with good algorithm interpretability. With the concern for climate issues, K-means clustering has been widely used in recent years for energy analysis. In the study of regional industrial energy management, Gengyuan Liu used K-means clustering to establish a data mining-based regional industrial energy efficiency evaluation model (Liu, G et al., 2018). Because of its easy-to-implement feature and high interpretability, K-means clustering is one of the commonly used methods in exploratory research.

3.1.3 Evaluation and Validation

Before conducting the K-means clustering, I will use Hopkins Statistic to measure the cluster tendency of a data set. Hopkins Statistic ranges from 0 to 1, the closer the value is to 1, the higher tendency that the data is clustered, and a value close to 0 indicates the data is uniformly distributed. 0.5 indicates the data is randomly distributed (Wikipedia, 2022). After that, I will use Silhouette score and elbow chart to determine the value of K. Silhouette analysis can be used to determine the degree of separation between clusters (Wikipedia, 2022). For each sample, I will compute the average distance from all data points in the same cluster(ai) and the average distance from all data points in the closest cluster(bi). And the Silhouette score could be:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

The coefficient takes values in the interval of [-1,1]. A value closer to 0 refers to the sample is very close to the neighboring clusters, a value close to 1 indicates the sample is assigned to the wrong clusters and a value equals to 1 means the sample is far away from the neighboring clusters (Wikipedia, 2022).

4 Findings

4.1 K-Means Clustering

By conducting Hopkins statistic test on the B Corp data, I get a Hopkins statistic of 0.92, which means there is a high cluster tendency in the data.

Figure 4 Elbow Plot of SSE

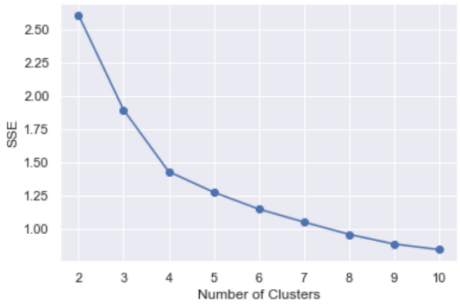
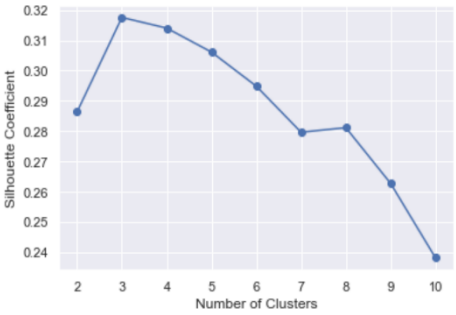


Figure 5 Knee Plot of Silhouette Scores



The SSE decreases rapidly until K is more than 4, which indicates that the distance between each point and the centroid decreases rapidly as K increases before K = 4. When K is greater than 4, the rate of decrease of SSE decreases dramatically, indicating that the distance between each point and the centroid no longer changes significantly as K increases. So, I employ K =4 (Figure 4).

Silhouette score is bounded in the range of -1 to 1. The closest to 1, the separation score is the largest, which indicates that an ideal silhouette coefficient is obtained since the separation score quantifies how dissimilar a sample is to other clusters, and the cohesion score tells how similar it is to the other samples in its own cluster. In this case, the Silhouette score roughly tends to decrease as the value of K increases (there is an increase at K =3), getting closer to 0, which indicates that the phenomenon of cluster overlap is more obvious. When K =3, the Silhouette score increases, which indicates that compared to K =2, dividing the data into 3 clusters can better reflect the similarity of data within clusters and the degree of difference between clusters (Figure 5).

When K = 4, the silhouette score decreases rapidly, indicating that we should update the previous decision and set K = 3 as the parameter of the model. Dividing the data into 3 clusters can better reflect the similarity of data within clusters and the degree of difference between clusters.

In the PCA, the scree plot displays how much variation each principal component captures from the data. From the plot, PC1 captures the most variation (approximately 30% of the variance), PC2 — the second most (approximately 30% of the variance), and so on. Since there are five characteristics in the data, we have five principal components. In the Silhouette score, I employ the first two PCs, which in total capture 60% of the information[ii]. Through the results of Silhouette analysis, I find when K equals to 3, the model reaches the highest Silhouette score[iii].

4.1 Clusters Descriptive Statistics

By classifying B corps into three groups, I get 1902 B corps in Cluster 1, 2377 B corps in Cluster 2, and 931 B corps in Cluster 3. The result of descriptive statistics of B Corp Cluster 1 shows that, in general, the 1st Clustering B Corps perform well in terms of Environment and poorly in terms of Customers. B Corp Cluster 1 has a mean of 30.6 in Environment, one standard deviation above the average of 17.57. And this group of SEs has a mean of 2.43 in Customers, three standard deviations under the average of

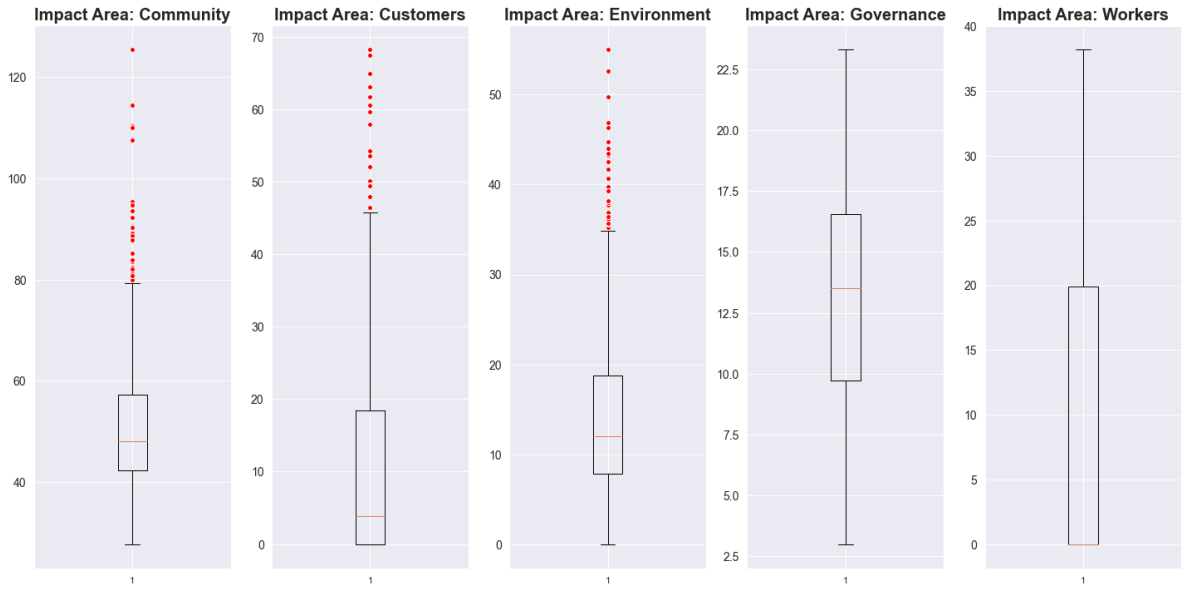
[ii] See Appendix for PCA Scree Plot.
[iii] See Appendix for Silhouette analysis.

11.93 (Table 2&Figure 6). This indicates that this group of SEs are good in overall environmental management practices and their impact on air, climate, water, land and biodiversity, but below average in managing their customers through the quality of their products and services, ethical marketing, data privacy and security.

Table 2 Descriptive Statistics of B Corp Cluster 1 Impact Index

	Community	Customers	Environment	Governance	Workers
Count	1906	1906	1906	1906	1906
Mean	23.02	2.43	30.60	13.01	22.17
Std	7.67	3.00	10.26	3.82	7.02
Average	28.02	11.93	17.57	13.75	25.26

Figure 6 Performance of Cluster 1 on the Five Dimensions

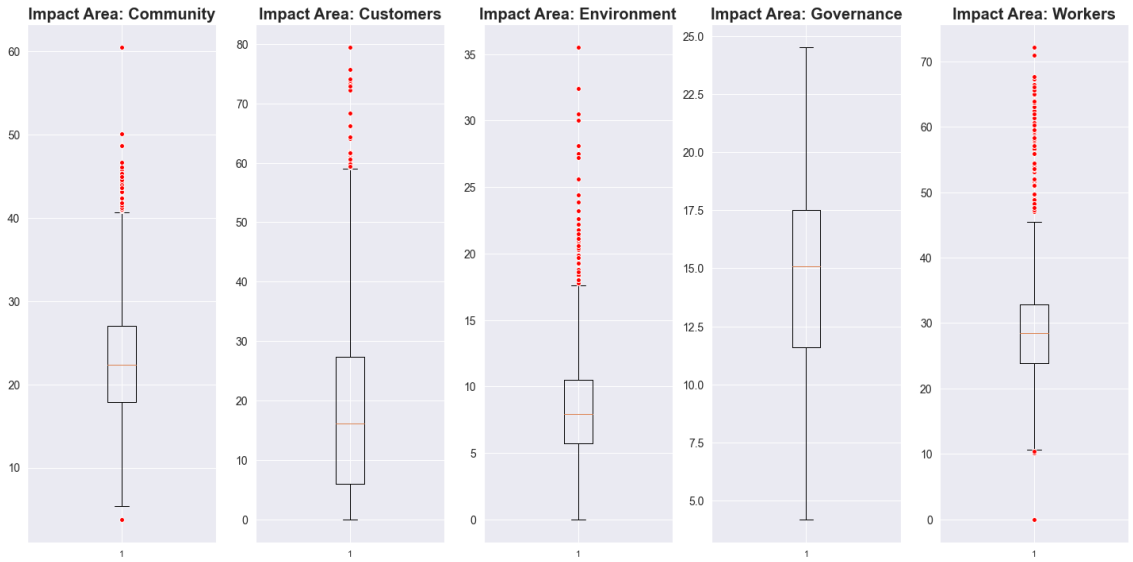


The 2nd Clustering B Corps perform average in many dimensions, poorly in terms of Environment. On dimensions Community, Customers, Governance, and Workers, B Corp Cluster 2's performance is no more than one standard deviation away from the overall mean. This group has a mean of 8.35 in Environment, 2.5 standard deviations lower than the average of 17.57 (Table 3&Figure 7). This indicates that this group of SEs have poor overall environmental management practices and their impact on air, climate, water, land and biodiversity.

Table 3 Descriptive Statistics of B Corp Cluster 2 Impact Index

	Community	Customers	Environment	Governance	Workers
Count	2376	2376	2376	2376	2376
Mean	22.76	18.91	8.35	14.57	28.06
Std	7.04	14.78	3.94	4.04	9.19
Average	28.02	11.93	17.57	13.75	25.26

Figure 7 Performance of Cluster 2 on the Five Dimensions

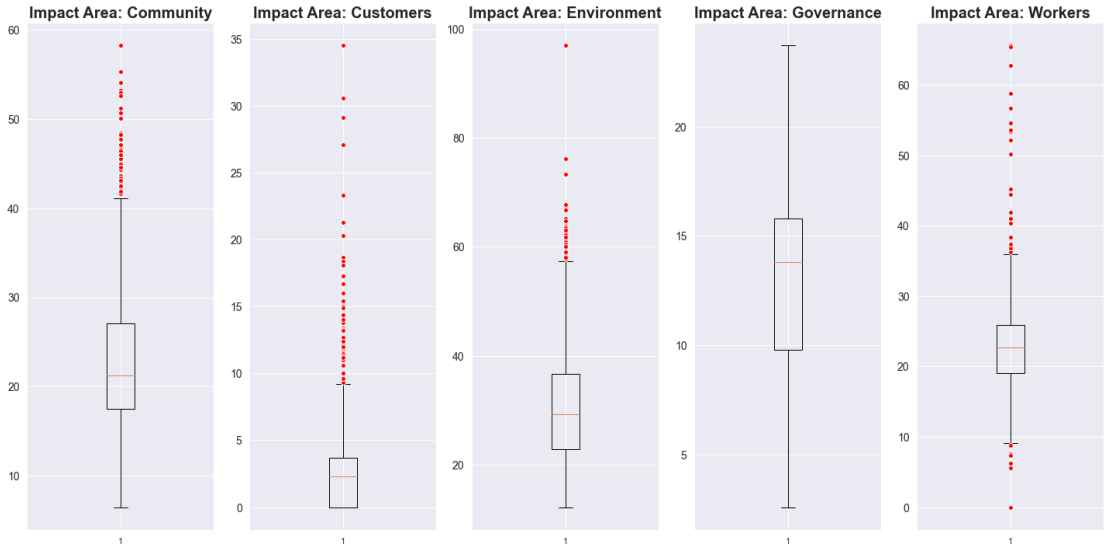


In general, the 3rd Clustering B Corps perform well in terms of Community and poorly in terms of Workers. B Corp Cluster 3 has a mean of 51.27 in Community, two standard deviations above the average of 13.12. And this group of SEs has a mean of 9.62 in Workers, 1.5 standard deviations under the average of 11.93 (Table 4&Figure 8). This indicates that this group of SEs have good engagement and impact on the communities in which they operate, employ and source, and poorer contributions in workers' assessment of a company's contribution to the financial security, health and safety, well-being, professional development, engagement and satisfaction of its employees.

Table 4 Descriptive Statistics of B Corp Cluster 3 Impact Index

	Community	Customers	Environment	Governance	Workers
Count	928	928	928	928	928
Mean	51.27	10.76	14.42	13.20	9.62
Std	13.12	13.77	9.04	4.31	10.72
Average	28.02	11.93	17.57	13.75	25.26

Figure 8 Performance of Cluster 3 on the Five Dimensions



4.2 Text Analysis

4.2.1 Bag of Words

In this part, I focus on the description texts of the three clusters of SEs, and conduct exploratory data analysis on the texts. By counting the word frequency distribution in different clusters respectively, we learn that Cluster 1 B Corps are more focused on their products and sustainability, and work in food industry. Top words include “product”, “company”, “sustainable”, “business”, “food”, etc,. Compared to Cluster 1, Cluster 2 B Corps are less about sustainability. "business", "impact" and "social" are the aspects that Cluster 2 B Corps focuses on. Top words include “work”, “company”, “impact”, “business”, “client”, etc,. Cluster 3 B Corps focus on “community” than the other two Clusters. Top words include “community”, “company”, “social”, “work”, “impact”, etc,[iv].

Figure 9 Cluster Word Cloud

Cluster 1 Cluster 2 Cluster 3

4.2.2 Term Frequency-Inverse Document Frequency

Further, to better capture the meaning in the description, I use TF-IDF to analyze the text. TF-IDF vectorization transforms text data into a numerical vector, taking into account the frequency of each

[iv] See Appendix for word frequency charts of three clusters.

word in the document, the total number of words in the document, the total number of documents, and the number of documents that include each unique word. Thus, unlike term-document matrices that only show the presence, absence, or number of a word in a document, it creates more meaningful text vectors that focus on the weights of words that represent their unique contributions to the document.

Table 5 Word Frequency of Cluster 1

Rank	Word	Rank	word
1	business	11	good
2	company	12	world
3	social	13	de
4	impact	14	sustainable
5	product	15	design
6	community	16	brand
7	people	17	service
8	create	18	coffee
9	help	19	support
10	work	20	change

Most of the top 20 words in Cluster 1 and Cluster 2 are duplicates. And TF-IDF works better in Cluster 3 keyword extraction. The top 10 words in Cluster 1 are "business", "company", " social", "impact", "community", "people", "create", "help" in sequence (Table 5). In the top then words, Cluster 2 also includes "client" and "service", and gives higher weight to "work" (Table 6). This observation echoes the fact that Cluster 2 SEs are doing better than Cluster 1 SEs in terms of Customers. SEs in Cluster 2 are mainly focused on the service sector. The top four industries are Professional & technical services, Information, communication & technology, Financial & insurance activities and Education[v].

Table 6 Word Frequency of Cluster 2

Rank	Word	Rank	word
1	business	11	brand
2	company	12	create
3	impact	13	world
4	social	14	design
5	work	15	investment
6	help	16	provide
7	people	17	change
8	client	18	sustainable
9	community	19	good
10	service	20	build

Cluster3's top 10 ranking keywords for social enterprise are "product", "company", "sustainable", "energy", "food", "business", "people", "create", " design", "world". Unlike Cluster1 and Cluster2 (Table 7), the SEs in Cluster3 ranked "product" first and "sustainable", "energy" and "food " are included in the

[v] See Appendix for industries in which the three B Corp Clusters are located.

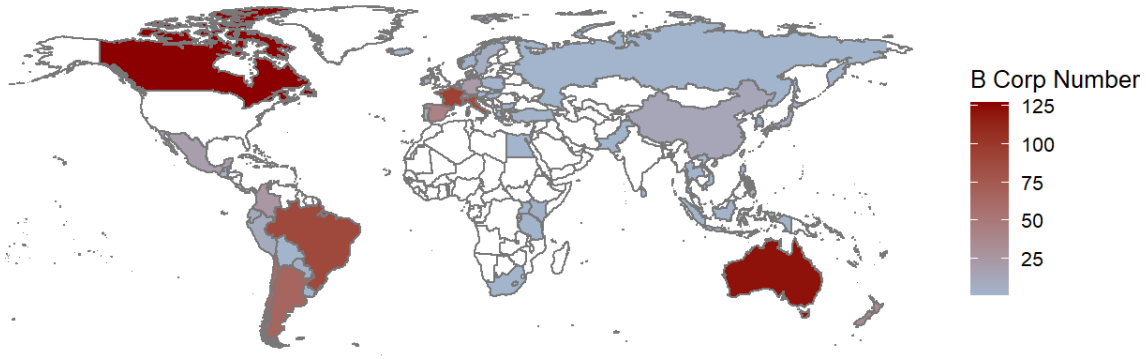
top five words. This is related to the fact that the SEs in Cluster 3 are mainly concentrated in industries Manufactured Goods, Retail, Wholesale and Agriculture, forestry & fishing.

Table 7 Word Frequency of Cluster 3

Rank	Word	Rank	word
1	product	11	organic
2	company	12	community
3	sustainable	13	good
4	energy	14	work
5	food	15	brand
6	business	16	impact
7	people	17	planet
8	create	18	environmental
9	design	19	la
10	world	20	environment

The geographic map shows that B Corp Cluster 3 has the widest distribution in the world, even with the smallest number of SEs. Interestingly, all social enterprises in Russia are classified as Cluster 3 (Figure 10). In addition, in Western Europe and South America, more social enterprises are classified in Cluster 3 than in Cluster 2 and Cluster 1[vi].

Figure 10 Global distribution of B Corp Cluster 3



5 Conclusion

By conducting K-means clustering, I classify the 5,210 B Corporations into three clusters. Each cluster has its notable features. In general, the 1st Clustering B Corps perform well in terms of Environment and poorly in terms of Customers. These B Corps are more focused on their products and sustainability. The 2nd Clustering B Corps perform average in many dimensions, poorly in terms of Environment. and are less focused on sustainability compared to the 1st Cluster. SEs in Cluster 2 are mainly coming from the service sector, such as Professional & technical services, Information, communication & technology,

[vi] See Appendix for geographic map of B Corp Cluster 1&2.

Financial & insurance activities and Education. The 3rd Clustering B Corps perform well in terms of Community and poorly in terms of Workers. This is related to the fact that the SEs in Cluster 3 are mainly concentrated in industries such as Manufactured Goods, Retail, Wholesale and Agriculture, forestry & fishing. B Corp Cluster 3 has the widest distribution in the world, even with the smallest number of SEs. Interestingly, all social enterprises in Russia are classified as Cluster 3. And SEs in Cluster 3 develop the best compared to Cluster 1&2 in Western Europe and South America.

Bibliography

About our standards. bcorporation.net. (n.d.). Retrieved December 11, 2022, from https://www.bcorporation.net/en-us/standards

B Corp certification demonstrates a company's entire social and environmental impact. B Corp Certification demonstrates a company's entire social and environmental impact. (n.d.). Retrieved December 14, 2022, from https://www.bcorporation.net/en-us/certification

B Corp Impact Data - dataset by blab. data.world. (2022, September 6). Retrieved December 11, 2022, from https://data.world/blab/b-corp-impact-data

B impact assessment. bcorporation.net. (n.d.). Retrieved December 11, 2022, from https://www.bcorporation.net/en-us/programs-and-tools/b-impact-assessment

Dai, P.-F., Xiong, X., & Zhou, W.-X. (2021). A global economic policy uncertainty index from Principal Component Analysis. Finance Research Letters, 40, 101686. https://doi.org/10.1016/j.frl.2020.101686

Esmaeili, A., & Shokoohi, Z. (2011). Assessing the effect of oil price on world food prices: Application of principal component analysis. Energy Policy, 39(2), 1022–1025. https://doi.org/10.1016/j.enpol.2010.11.004

Hynes, B. (2009). Growing the social enterprise – issues and challenges. Social Enterprise Journal, 5(2), 114–125. https://doi.org/10.1108/17508610910981707

Islam, S. (2021). Social Impact Scaling Strategies in Social Enterprises: A systematic review and Research Agenda. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3852342

Liu, G., Yang, J., Hao, Y., & Zhang, Y. (2018). Big data-informed energy efficiency assessment of China industry sectors based on K-means clustering. Journal of Cleaner Production, 183, 304–314. https://doi.org/10.1016/j.jclepro.2018.02.129

Powell, M., Gillett, A., & Doherty, B. (2018). Sustainability in social enterprise: Hybrid Organizing in Public Services. Public Management Review, 21(2), 159–186. https://doi.org/10.1080/14719037.2018.1438504

van Rijn, M., Raab, J., Roosma, F., & Achterberg, P. (2021). To prove and improve: An empirical study on why social entrepreneurs measure their Social Impact. Journal of Social Entrepreneurship, 1–23. https://doi.org/10.1080/19420676.2021.1975797

Wikimedia Foundation. (2022, December 9). Silhouette (clustering). Wikipedia. Retrieved December 13, 2022, from https://en.wikipedia.org/wiki/Silhouette_(clustering)#:~:text=The%20silhouette%20value%20is%20a,poorly%20matched%20to%20neighboring%20clusters.

Wikimedia Foundation. (2022, November 20). Hopkins statistic. Wikipedia. Retrieved December 13, 2022, from https://en.wikipedia.org/wiki/Hopkins_statistic

Wikimedia Foundation. (2022, November 21). TF–IDF. Wikipedia. Retrieved December 11, 2022, from https://en.wikipedia.org/wiki/Tf%E2%80%93idf

6 Appendix

Figure 11 PCA Scree Plot

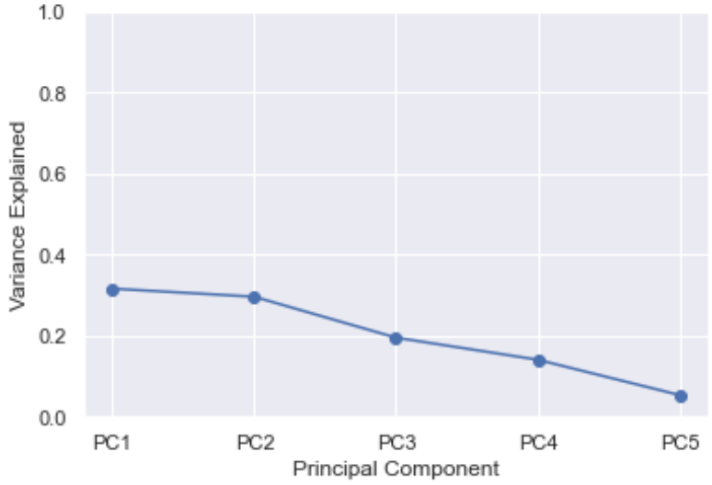


Figure 12 Silhouette Plot (K=2)

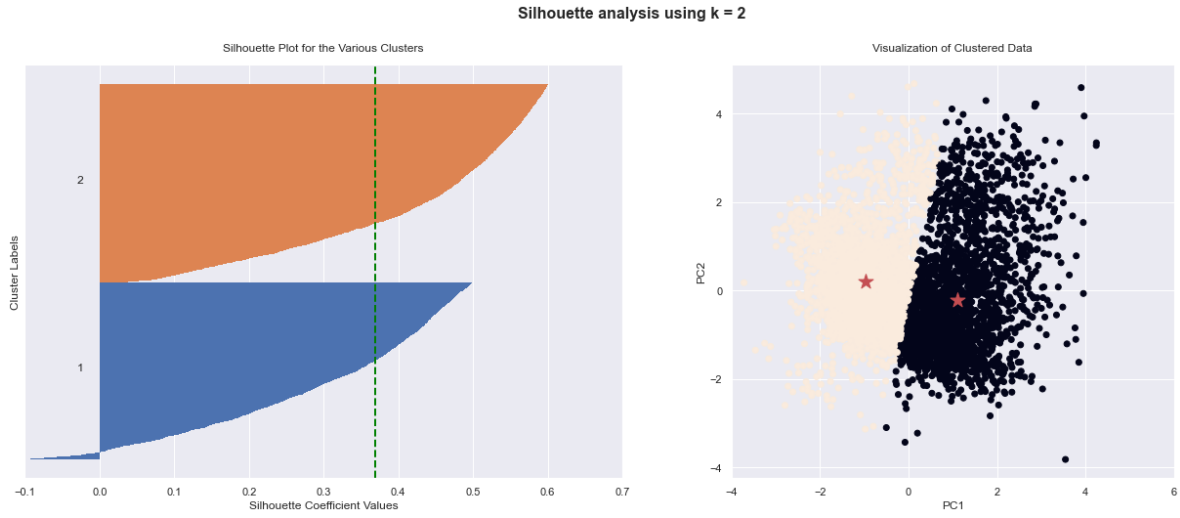


Figure 13 Silhouette Plot (K=3)

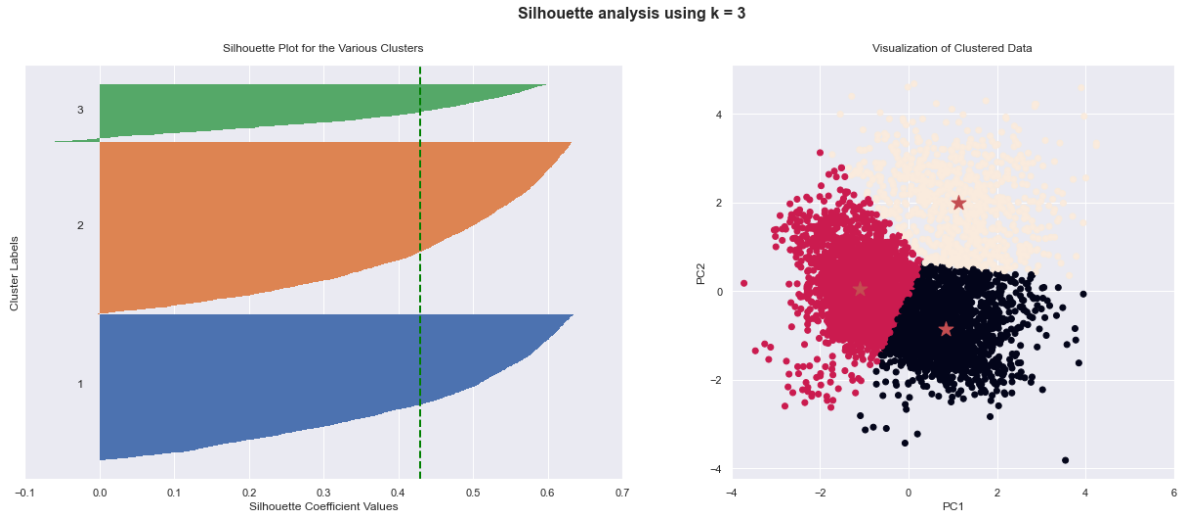


Figure 14 Silhouette Plot (K=4)

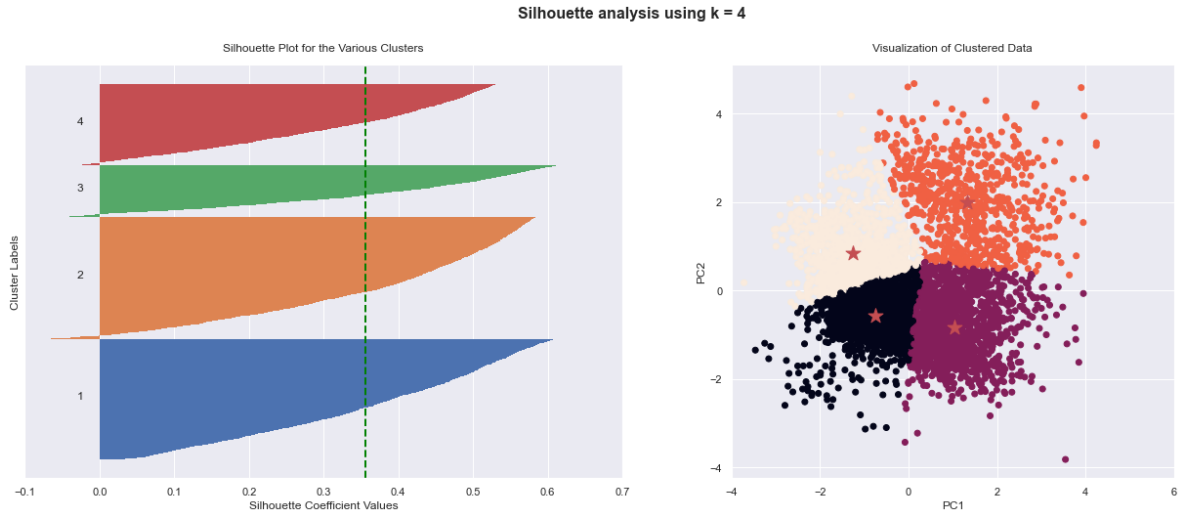


Figure 15 Word Frequency of Cluster 1

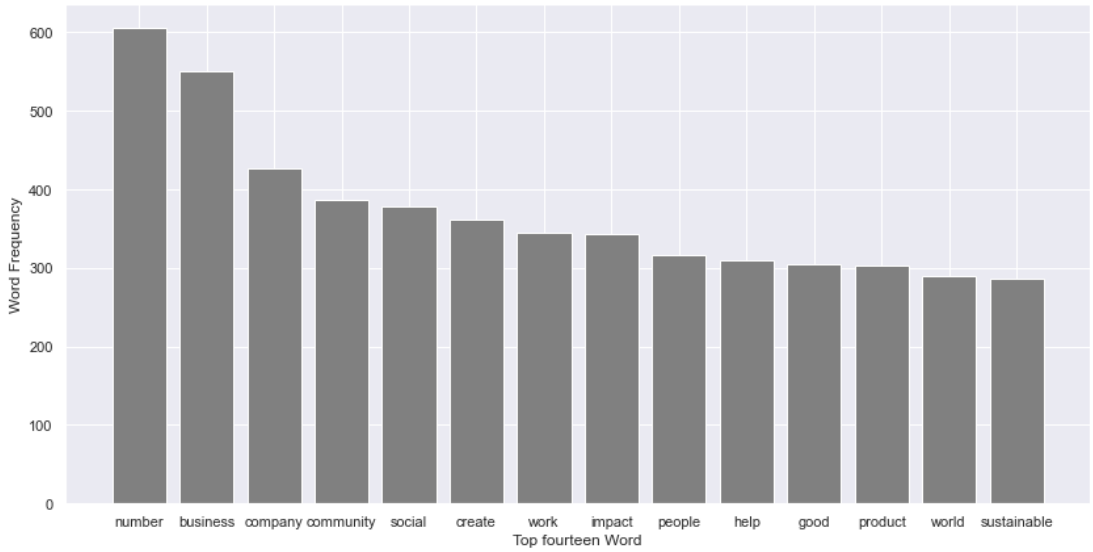


Figure 16 Word Frequency of Cluster 2

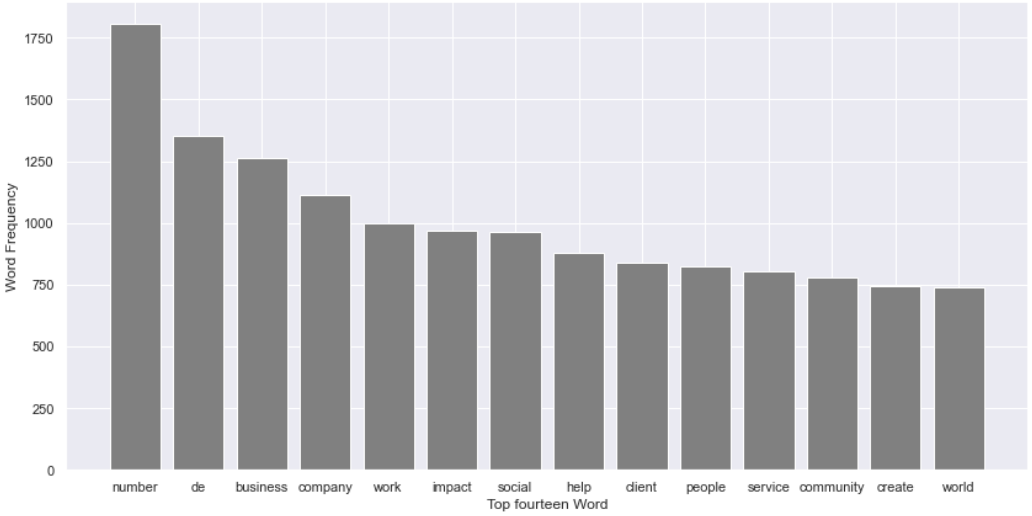


Figure 17 Word Frequency of Cluster 3

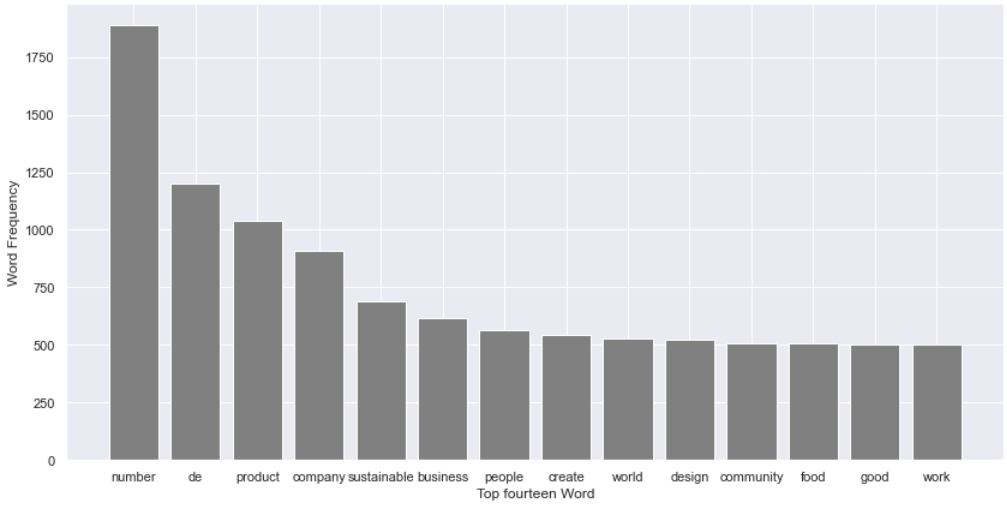


Figure 18 Cluster 1 Industry Distribution

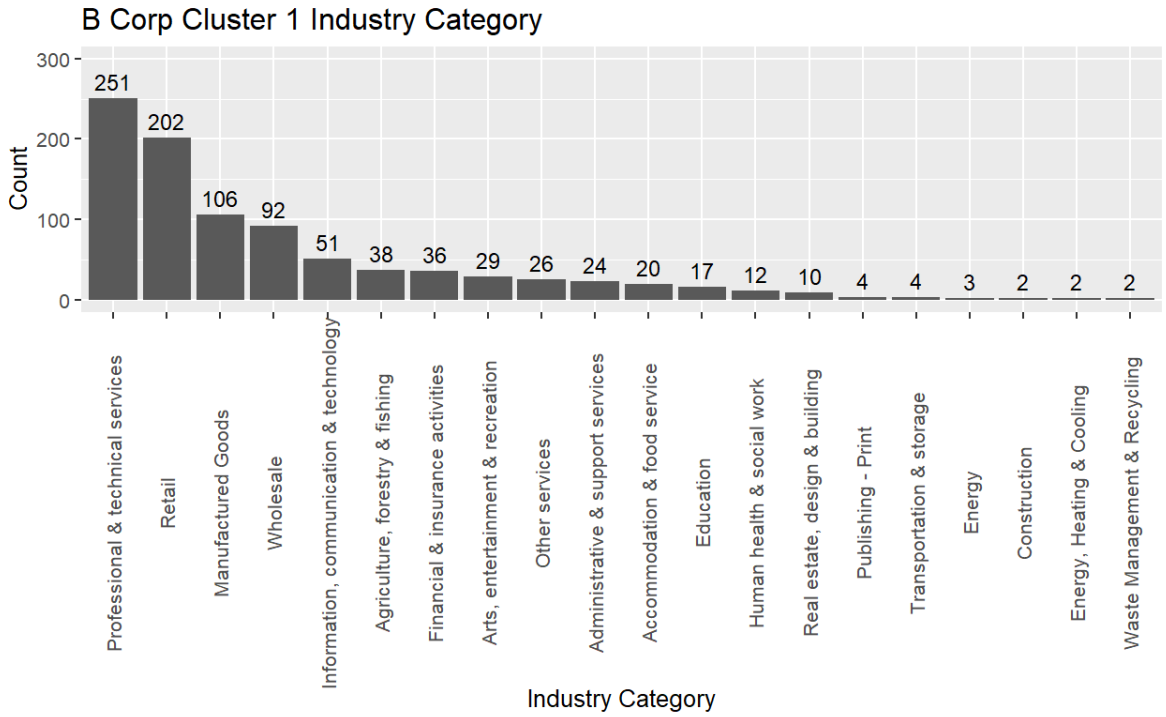


Figure 19 Cluster 2 Industry Distribution

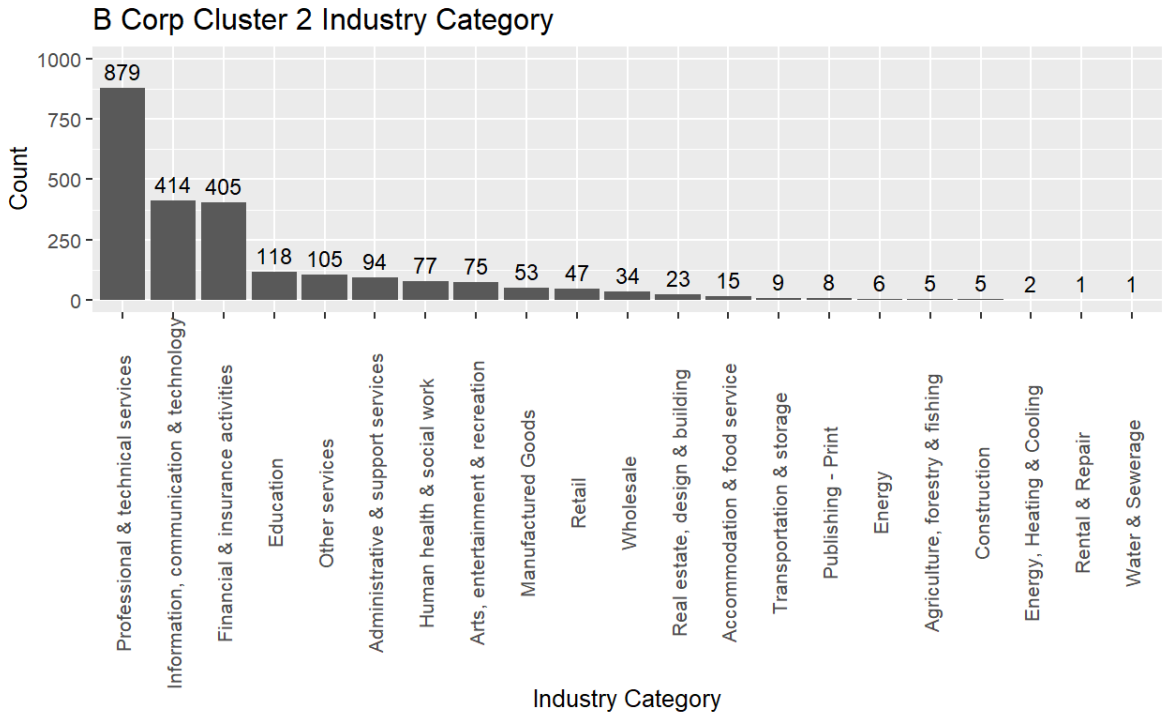


Figure 20 Cluster 3 Industry Distribution

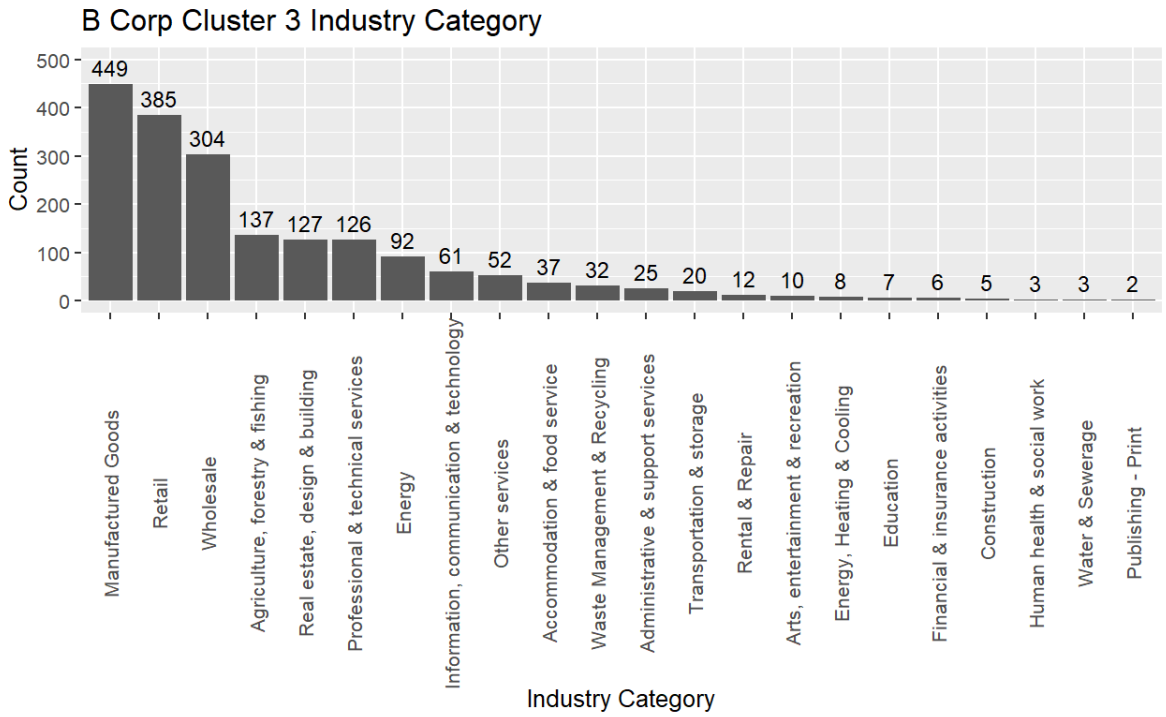


Figure 21 Global distribution of B Corp Cluster 1

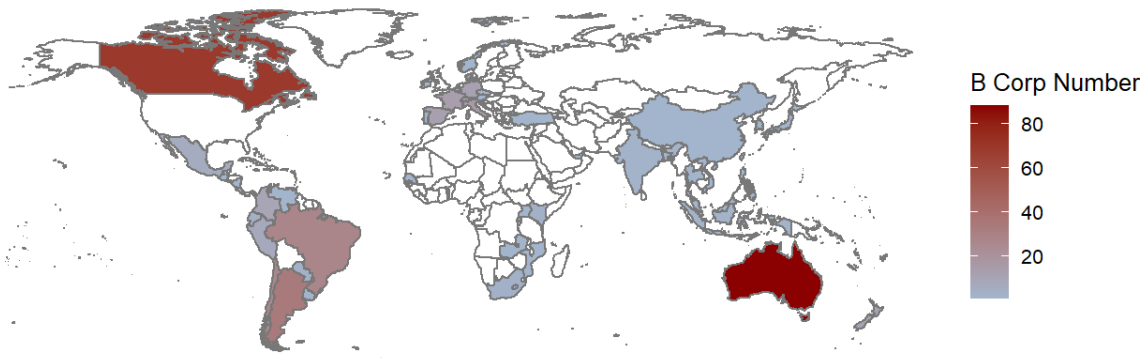


Figure 22 Global distribution of B Corp Cluster 2