Analysis on Impact Scores of Certified B Corporations in the Apparel, Footwear & Accessories Industry

Jitong Li, Ran Zhang

Abstract

Impact Scores of Certified B Corporations were introduced to qualify the sustainable business among corporations. We are concerning about the multi-variables related to the impact scores of Certified B Corporations from United States and out of United States. We use Q-Q plot and chi-square test to test normality and identify outliers, apply the principal component analysis and factor analysis to reduce dimension of the data. Two sample t-test is used to test there are differences between sustainable performance of corporations in United States and out of Unite States. Simultaneously confidence intervals are examined the inequality of sustainable performance in different area of corporations all over the world. Regression tree is used to classify corporations with various sustainable performance to help us handle missing data. Unlike considering the data itself, we also concerned data with outliers and without outliers to understand how outliers affect our analyzing results under same method. This analysis will help us easily assign a new data to a group and the company will know their competitors in the same group.

Key words: Classification, B Corporation, Multi-variable, outliers

1. Introduction

With the increasing attention on sustainability from various sectors, like manufacturers, consumers and governments, performing a sustainable business seems like an imperative and beneficial effort for enterprises to succeed in a competitive environment. Sustainable business refers to a business that strives to meet the triple bottom line (environment, social and economic). Evaluation on the sustainable performance of enterprises is a key issue, and some certifications existing to assist enterprises getting better understand on their sustainable performance and communicating with shareholders. B Corp Certification is one of these certifications used to measure a company's governance, entire social and environmental performance. A certified corporation's overall B impact score consists of five area, governance, workers, community, environment and customers. Since the evaluation on sustainable performance is still a hot and relative new topic, there is limited knowledge sustainable performance, it is needed to get a general insight on various companies' sustainable performance.

2. Methodology

Data was collected originally by the authors from the website of Certified B Corporation. Considering the interests of the authors, only the data of companies in the apparel, footwear & accessories industry was collected. There were 92 related corporations listed in the B Corp Directory and they were divided into two groups, corporations in the United States (Sample 1: 43 observations) and out of the United States (Sample 2: 49 observations). And there were six variables, including overall B Impact Score, Governance Score, Workers Score, Community Score, Environment Score and Customers Score. Data points collected in US and out of US were named as "US data" and "World data" separately for processing the same methods on all analyzing methods except point out in the whole methodology section. These methods will be applied to analyze this dataset as followings:

Data preparation: As all data were collected from two different regions, in US and out of US, some companies may not provide all necessary data for all 6 variables (overall B Impact Score, governance, workers, community, environment and customers). Data points with missing information will be removed to avoid possible mis-leading results.

Normality assessment and outlier detection: Two independent samples (US data, World data) after cleaning up will be applied on the normality assessment separately. All single variable of each sample will be tested by performing Shapiro tests with Q-Q plot; And multivariate tests will be applied with response variable (overall score) and without response variable. Multivariate tests will be performed by Royston's test with a chi-square plot. Outliers will be identified by Mahalanobis distance observed in chi-square test.

Principal component analysis will be used to reduce the dimensionality and to identify the crucial variables for the response for the overall B impact score. As there are total 6 variables include one response variable (overall B impact score), which will not be considered on principal component analysis. All other variables could be written as a linear combination as $Y_i = a_{i1}X_1 + \cdots + a_{ip}X_p = a_i^T X$ and the total will be calculated by $TV = Var(Y_1) + \cdots + Var(Y_p)$. The variance of each variable will be checked and scaled for standardization to achieve a precise result. The standardized data will be applied for principal component analysis by using procomp function in R. The least number of PCs to retain for 70% variance, and each loading will be extracted for specify each PCs. The data with outliers and without outliers will be performed on principal component analysis to compare with each other to see if any influence by outliers on principal component analysis.

Exploratory factor analysis will be applied as another way to reduce the data dimension. Different with principal component analysis, which attempts to capture most of total variance. Exploratory factor analysis tries to maximize variance due to the common factors; Maximum likelihood factor analysis will be used on this project. The model is built with observed variables (X_i), common factor (F), specific errors (u_i) and the factor loadings (λ_i) which is written as $X_p = \lambda_p F + u_p$ and the variance will be calculated as the sum of square of communality and specific variance. Factor analysis with maximum likelihood method will be applied on our data to find the least number of factors needed and Factor rotation will be applied if possible, to identify the different variables mostly contributed to the different factors separately. Besides, comparing the conclusions between the data without removing outliers and with outliers.

Two-sample Hotelling's T^2 test will be used to compare the six scores of two samples and test whether there are differences between sustainable performance of corporations in the United States (US data) and out of United States (World data). The mean vector of 6 variables will be test for equality between these two-independent sample. Normality assessment with all 6 variables will be applied by chi-square test before testing mean vectors and outliers with be identified by Mahalanobis distance. The test will be applied on the sample with outliers and without outliers, the conclusions with two different data (with outliers or not) will be compared to find the influence introduced by outliers.

Simultaneously confidence intervals will be employed to examine whether the scores (2-6) of each sample to the response are equal. Which means that the equality of sustainable performance in different area of corporations in the United States/Out of United States will be examined. The contrast matrix will be built for comparisons among all variables, and the simultaneously confidence intervals will indicate if there is difference between any two variables (include 0) and which variables contributed more to the response variables. Data with outliers and without outliers will be both applied for conclusions comparison.

Regression Tress will be applied to classify corporations with various sustainable performance since it can handle missing data. Initially, the data of sample 1 and 2 will be combined as a same group and then observations will be separated into two groups with the rules of different variables parameters. Two different populations (with outliers and without outliers) will be applied on the regression tree. Two different trees will give us different information about the classification.

3. Results

As the two raw data (US data and world data) were imported and named as US data and World data, results of different methods were applied and shown below.

Data preparation: Overall B impact score, governance, workers, community, environment and customers six variables were imported with NAs. For further possessing data, the formats of variables were all transformed to numeric format. As too many NAs under the customer variable, the whole customer variable was removed. All other data with NAs on any variable were also removed. After that, re-index the data for further processing.

Normality assessment and outlier detection: US data was firstly to test the normality on all variables individually, as shown in figure 1a. As we found in all variables (x2 to x6 correspond to overall B impact score, governance, workers, community and environment) almost normally distributed. After that, the same procedure was applied on world data shown as figure 1b. The same conclusion achieved by the world data which all variables are approximately normally distributed. Then the chi-square test was applied on the US data and world data for multivariate normally test and results are shown in 1c and 1d. Four outliers in US data are “Indigenous Impact Fashion”, “Patagonia”, “The Sox Box” and “Wallaroo Hat Company”. Three outliers in World data are “Carla Fernandes”, “Movin” and “Someone Somewhere”. Both data are normally distributed.

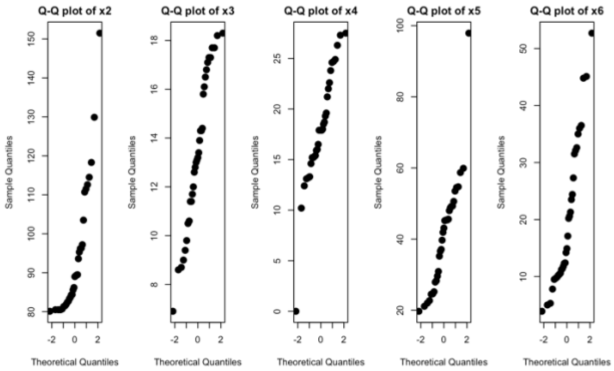


Figure 1a. Q-Q plot of US data.

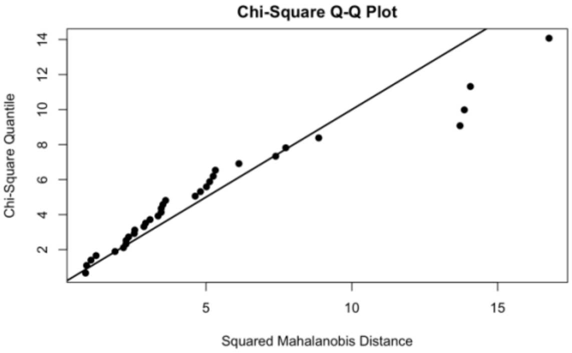


Figure 1c. Chi-Square plot of US data.

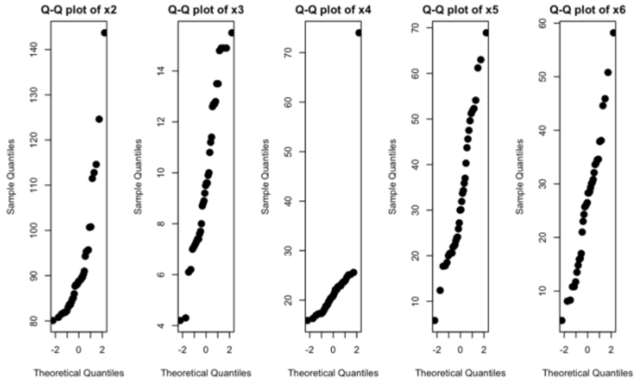


Figure 1b. Q-Q plot of world data.

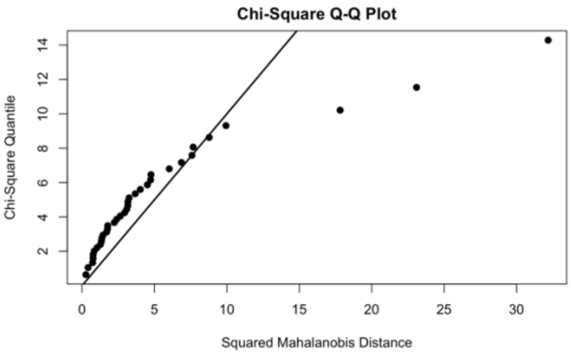
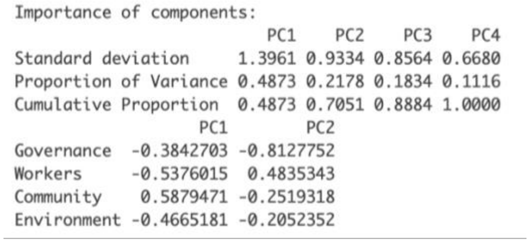


Figure 1d. Chi-Square plot of world data.

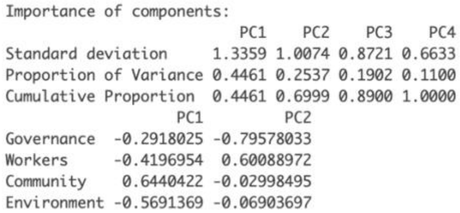
Principal component analysis: As US data and world data both have two sub-samples with outliers, named as US RM data and World RM data. US data and US RM data were applied on the principal component analysis with loadings attached, shown as figure 2a and 2b. For US data, first two PCs will be kept achieving 70.5% variance; For US RM data, first two PCs will be kept to achieving 70.0% variance. Also, World data and World RM data were applied on the principal component analysis, and results were shown in 2c and 2d with loadings attached. For World data, first two PCs will be kept for 69.7% variance; For World RM data, first two PCs will be kept for 70.9%

variance. The data with outliers or without outliers not provide much difference on principal component analysis.



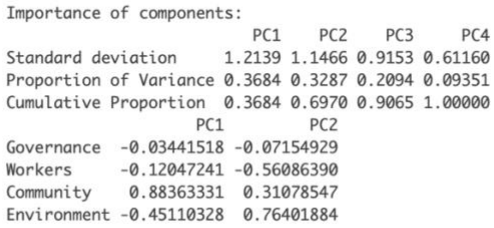
```
Importance of components:
                          PC1    PC2    PC3    PC4
Standard deviation     1.3961 0.9334 0.8564 0.6680
Proportion of Variance 0.4873 0.2178 0.1834 0.1116
Cumulative Proportion  0.4873 0.7051 0.8884 1.0000
                   PC1        PC2
Governance  -0.3842703 -0.8127752
Workers     -0.5376015  0.4835343
Community    0.5879471 -0.2519318
Environment -0.4665181 -0.2052352
```

Figure 2a. US data with outliers.



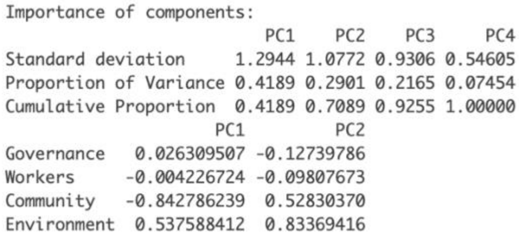
```
Importance of components:
                          PC1    PC2    PC3    PC4
Standard deviation     1.3359 1.0074 0.8721 0.6633
Proportion of Variance 0.4461 0.2537 0.1902 0.1100
Cumulative Proportion  0.4461 0.6999 0.8900 1.0000
                   PC1         PC2
Governance  -0.2918025 -0.79578033
Workers     -0.4196954  0.60088972
Community    0.6440422 -0.02998495
Environment -0.5691369 -0.06903697
```

Figure 2b. US data without outliers.



```
Importance of components:
                          PC1    PC2    PC3     PC4
Standard deviation     1.2139 1.1466 0.9153 0.61160
Proportion of Variance 0.3684 0.3287 0.2094 0.09351
Cumulative Proportion  0.3684 0.6970 0.9065 1.00000
                    PC1         PC2
Governance  -0.03441518 -0.07154929
Workers     -0.12047241 -0.56086390
Community    0.88363331  0.31078547
Environment -0.45110328  0.76401884
```

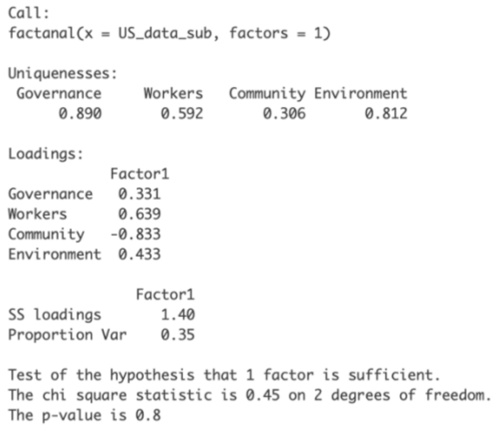
Figure 2c. World data with outliers.



```
Importance of components:
                          PC1    PC2    PC3     PC4
Standard deviation     1.2944 1.0772 0.9306 0.54605
Proportion of Variance 0.4189 0.2901 0.2165 0.07454
Cumulative Proportion  0.4189 0.7089 0.9255 1.00000
                     PC1         PC2
Governance   0.026309507 -0.12739786
Workers     -0.004226724 -0.09807673
Community   -0.842786239  0.52830370
Environment  0.537588412  0.83369416
```

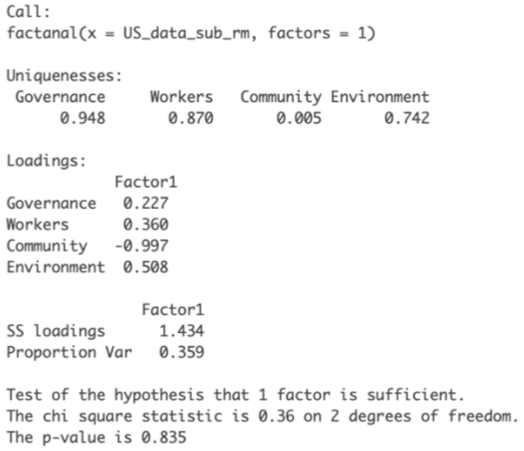
Figure 2d. World data without outliers.

Exploratory factor analysis: US data and world data with outliers and without outliers were applied on the factor analysis to gain the least common factor. From the US data without removing outliers, one factor is sufficient with p-value of 0.8. And the factor was explained as the difference between the community and sum of governance, workers and environment (shown as 3a). The result of US data after removing outliers was shown in 3b, which the p-value is 0.835, also indicates one factor is enough and the factor has the same explanation as US data without removing outliers. Outliers here not make much difference on US data. For world data with outliers, the p-value is 0.0338 which smaller than 0.05, indicates one factor is not enough. As we only have 4 variables here, more than one factors cannot be applied (Figure 3c). However, after removing outliers, the p-value is 0.206 which indicates one factor is enough for this model, and the factor represents the difference between the community and sum of governance plus environment (Figure 3d). Therefore, outliers influence the results a lot on world data for factor analysis.



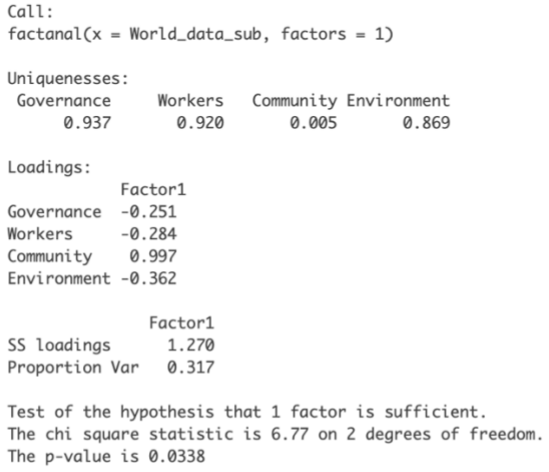
```
Call:
factanal(x = US_data_sub, factors = 1)

Uniquenesses:
 Governance     Workers   Community Environment
      0.890       0.592       0.306       0.812

Loadings:
            Factor1
Governance   0.331
Workers      0.639
Community   -0.833
Environment  0.433

               Factor1
SS loadings       1.40
Proportion Var    0.35

Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 0.45 on 2 degrees of freedom.
The p-value is 0.8
```

Figure 3a. US data with outliers.



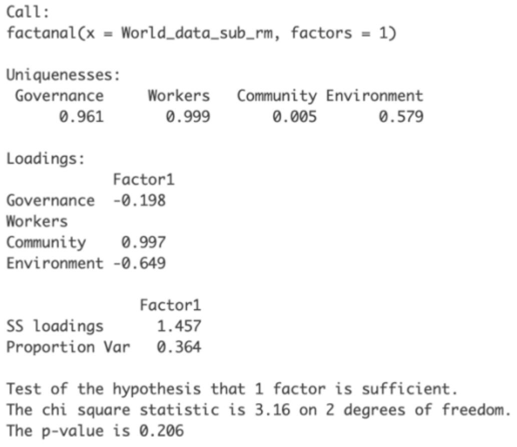
```
Call:
factanal(x = US_data_sub_rm, factors = 1)

Uniquenesses:
 Governance     Workers   Community Environment
      0.948       0.870       0.005       0.742

Loadings:
            Factor1
Governance   0.227
Workers      0.360
Community   -0.997
Environment  0.508

               Factor1
SS loadings      1.434
Proportion Var   0.359

Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 0.36 on 2 degrees of freedom.
The p-value is 0.835
```

Figure 3b. US data without outliers.



```
Call:
factanal(x = World_data_sub, factors = 1)

Uniquenesses:
 Governance     Workers   Community Environment
      0.937       0.920       0.005       0.869

Loadings:
            Factor1
Governance  -0.251
Workers     -0.284
Community    0.997
Environment -0.362

               Factor1
SS loadings      1.270
Proportion Var   0.317

Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 6.77 on 2 degrees of freedom.
The p-value is 0.0338
```

Figure 3c. World data with outliers.



```
Call:
factanal(x = World_data_sub_rm, factors = 1)

Uniquenesses:
 Governance     Workers   Community Environment
      0.961       0.999       0.005       0.579

Loadings:
            Factor1
Governance  -0.198
Workers
Community    0.997
Environment -0.649

               Factor1
SS loadings      1.457
Proportion Var   0.364

Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 3.16 on 2 degrees of freedom.
The p-value is 0.206
```

Figure 3d. World data without outliers.

Two-sample Hotelling's T^2 test: Comparisons of two independent samples was applied by two-sample Hotelling's T^2 test. For data with outliers, the p-value is 1.55e-5 which smaller than the

0.05, so we reject H_0. The mean vector of US data and world data are different. For data without outliers, we concluded with the same conclusion but with a bit of higher p-value with smaller degrees of freedom, which is 6.745e-5. Therefore, there is significance different between sustainable performance of corporations in United States and out of United States. The outliers here not affected the result on testing the mean vectors from US data and world data.

Simultaneously confidence intervals: The distribution of different variables contributed to the performance score were tested by simultaneously confidence intervals. For US data, the p-value with 5.68e-12 indicates there are big difference between four variables. And the confidence interval indicates the community contributed most to the performance score, and the governance contributed least. For data out of United States, the p-value is 5.1e-16 for the hypothesis test of all equally contributed to the performance. Therefore, there is significant different among the contributions from four variables. From the confidence intervals, we can conclude that the community contributed most, and the governance contributed least, which is the same pattern as US data.

Regression Tress: For classifying the corporations, data from United States and out of United States were combined together, one with outliers (Figure 4a) and the other one without outliers (Figure 4b) to form two trees. In figure 4a, data with outliers, only two variables were used for classification. However, after removing outliers, more classifications were introduced. The community, which contributes most to the performance was examined twice for classification.

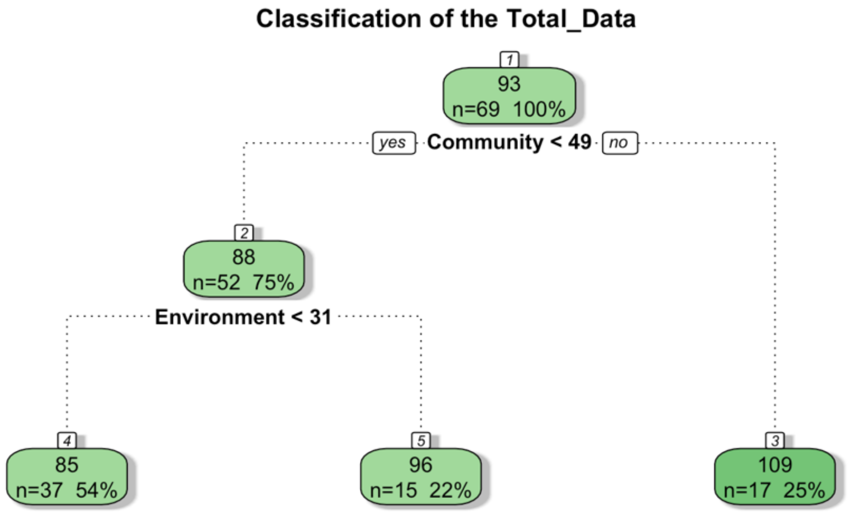


Figure 4a. Regression tree of total data with outliers.

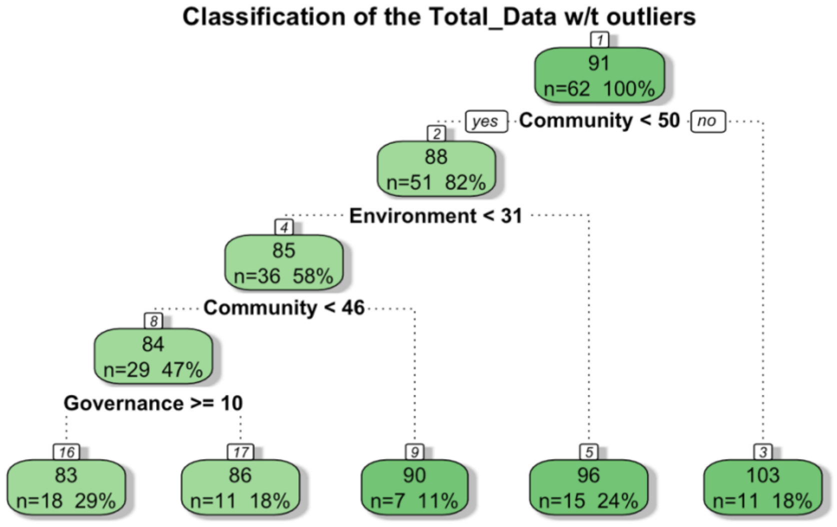


Figure 4b. Regression tree of total data without outliers.

4. Discussion

In our real life, we cannot obtain the perfect data like we dealt with in class. So, it is necessary to do the data preparation and cleaning before data analysis. For variable contains many NAs, that variables should be omitted directly. For some data points with NAs, we could also delete it if we still have 30 or more data points left (CLT theorem applied). Otherwise, no matter how to make up the NAs, some mis-leading conclusions will be obtained. Here, we used two methods for reducing dimensions: Principal component analysis and factor analysis. Principal component analysis provides more information about variance captured, without assuming the normality of the data. However, the principal component analysis, we didn't find any difference on the data with outliers and without outliers. For factor analysis, we applied the maximum likelihood method for common factors, so we have to assess the normality of the single variables and multi-variables. And for looking for the least factor here, on the world data, the results are totally different between the world data with outliers and without outliers. As we only have four variables here, factor analysis cannot be improved a lot. As more variables we have, factor analysis could perform better. For testing difference from different regions, if possible, we should collect more data with more specific regions. Here, we didn't observe big difference for In that way, we could use compare mean vectors from multiple independent populations and find the difference by paired comparisons. Simultaneously confidence intervals which indicates the difference within the sample, we could use Bonferroni confidence interval to compensate the errors caused by the multi-variables next time. Regression tree is a good way for visualizing the classification, we noticed that these two trees with outliers and without outliers give us totally different classifications. Like with the outliers, we cannot specify each region clearly. For example, here there are 3 different groups, some of the samples will be classified into different groups due to the influence by the outliers. After removing outliers, we have much more specified classifications. With the tree after removing outliers, we can classify data more precisely. This classification will help a company get a better understand on their sustainable process and competitive analysis. For instance, in the future, when a company get its scores, it can be easily assigned to a group and the company will know their competitors in the same group. Also, if we have data in the future, we can find more closely group of the data. In that way, we can save time and money to find the correct corporation to collaborate or compete. In conclusion, data preparation is important before data analysis. Normality assessment is necessary as outliers may cause large influence on the analysis results.

5. References

1. Certified B Corporation, retrieved from https://bcorporation.net/directory.

2. Multivariate Statistics with R by Paul J. Hewson.

3. Applied Multivariate Statistics with R by Daniel Zelterman. New York: Springer.

4. Modeling Longitudinal Data by Robert E. Weiss. New York: Springer.

5. All lecture notes from ST 537 “Applied Multivariate and Longitudinal Data Analysis” by Dr. Arnab Maity.