DANA 4830 Group 1 Project Report

Can Living Sustainably Bring You Happiness?

Dataset Chosen: Happiness

Full Name	Student ID
Chin Wei Shih (David)	#100334127
Zhixuan Zhang (Eric)	#100338057
Chun Ching Look (Cyrus)	#100347726
Simranjit Singh (Simran)	#100348495
Kailash Sukumaran (Kailash)	#100350193

Aims & Objectives

As developing countries start to enjoy rapid growth from their success in joining the global economy they will inevitably be susceptible to consumerism. Although wanting to consume is a normal human behavior, there must be a need to consume sustainably in order to preserve the state of the planet for future generations. Prolonged efforts in sustainable behavior (SB) and sustainable consumption (SC) could be a pathway to happiness as the act of carrying out environmentally behaviors can provide satisfaction. The aim of this report is to study sustainable behavioral trends in the context of a rapidly developing Vietnam and see how it is associated with the pathways and orientations to happiness as set by Peterson, Park and Seligman (2005). We will (1) review and explore the data set and provide descriptive analysis, (2) use PCA, FA, DA, stepwise regression, and/or MCA to provide inferential analysis, and (3) discuss and provide information based on our inferential analysis and see if it can answer our **Research Question:**

> How do different 'Orientations to Happiness' relate to the adoption of sustainable consumption in Vietnam?

Descriptive Analysis

Our data set consists of three parts. The first part contains 18 questions asking about happiness and uses a 7 point Likert Scale, ranging from 1 (Not like me at all) to 7 (Very much like me). The second part contains 35 questions asking about the participant's sustainable behavior and consumption and also uses a 7 point Likert Scale but adds a NA (not applicable) column. The last part of the data set asks about the participant's demographics as well as some behavioral questions. The behavioral questions uses multiple choice and fill in the blanks questions to ask about the participants frequency in certain transportation methods along with the monthly amount they spend on petrol and utilities. There are a total of 385 rows representing responses from 385 participants.

To develop more accurate statistical analysis, we proceeded to clean the data set. We first checked the missing values of the dataset. There are 1672 missing values in total.

For part 1 of the dataset, there were a total of 37 missing values. A look at the data set found no pattern from the missing values thus we can deduce them to be missing at random. Of the 37 missing values, 25 of them came from rows with only 1 missing value. 3 came from rows with 2 missing values and 2 from rows with 3 missing values. We elected to fill in the missing values of part 1 with the most common answer given, also known as 'mode', for that particular question.

Part 2 of our dataset had a specific column named "NA" for participants to choose. We believe that participants who chose NA for their answer either had the question not be applicable to them or they simply did not understand the question and had no opinion on it. Because of these reasons, we elected to fill all the NA's with the neutral point of the Likert Scale - '4'.

We also found that question #10 in part 2 of the data set had an alarming 90 NA's out of a possible 385 responses. The question asks about whether a participant composts garden or kitchen waste. As the idea of composting, especially garden compost, is more of a western behavior when it comes to sustainability, it might not translate adequately in the Vietnamese context. We have elected to remove this question in our analysis because of its lack of translatability.

In part 3 of the dataset, there were missing values as well. Specifically, we found that question 9 option 1 to option 8, we found out 28 respondents completely skipped ticking these options. This means some respondents did not understand the question fully. Although there were missing values, we did not replace them with new values. We believed it is illogical to handle missing data in demographic variables. In addition, we found that there is a disproportionate amount of female respondents in the dataset. An overwhelming 62.5% of the responses were from females. Further research indicated that this is a unique cultural phenomenon as females in Vietnamese households would be inclined to make more consumption decisions for the family, hence the disproportionate amount of female respondents.

For outlier issues, we checked part 1 and part 2 of the dataset. We remove outliers by measuring Mahalanobis distance. In part 1, there were 22 outliers. After removing part 1 outliers, there were 363 observations left. In part2, there were 25 outliers. After removing part 2 outliers, there were 338 observations left. We didn't remove any outliers in part 3. Out-of-range values were present in the dataset. There are two values in the part 1 which are 4.5 and 6.5, which are not on our Likert Scale. To replace 4.5, we rounded this value to 5; to replace 6.5, we rounded the value to 6.

After a deep data cleaning process, we made a table of summary statistics including mean, minimum, maximum, standard deviation for part 1 and part 2 variables. We later used the clean dataset to do the following inferential analysis.

Table of summary statistics

Variable	Mean	Minimum	Maximum	Standard Deviation
M02	5.27	1	7	1.36
M05	5.33	1	7	1.21
M11	5.04	1	7	1.5
M12	5.6	2	7	1.3
M14	5.03	1	7	1.39
M17	5.17	1	7	1.4
P03	5.03	1	7	1.51
P08	4.99	1	7	1.53
P13	5.49	2	7	1.17
P15	4.44	1	7	1.85
P16	5.4	1	7	1.3
P18	5.4	1	7	1.42
E01	5.44	1	7	1.33
E04	5.03	1	7	1.33
E06	3.97	1	7	1.82
E07	5.32	1	7	1.25
E09	5.46	1	7	1.29
E10	4.61	1	7	1.51
SC1	4.86	1	7	1.66
SC2	4.86	1	7	1.67
SC3	5.02	1	7	1.46
SC4	5.11	1	7	1.54
SC5	5.7	1	7	1.41
SC6	5.82	1	7	1.48
SC7	6	2	7	1.28
SC8	5.54	1	7	1.47

Variable	Mean	Minimum	Maximum	Standard Deviation
SC9	5.16	1	7	1.71
SC11	5.57	1	7	1.38
SC12	4.84	1	7	1.52
SC13	4.46	1	7	1.6
SC14	4.58	1	7	1.58
SC15	4.99	1	7	1.63
SC16	4.44	1	7	1.78
SC17	4.74	1	7	1.42
SC18	4.88	1	7	1.49
SC19	5.13	1	7	1.37
SC20	4.92	1	7	1.44
SC21	4.68	1	7	1.5
SC22	4.24	1	7	1.64
SC23	4.5	1	7	1.51
SC24	4.8	1	7	1.6
SC25	4.9	1	7	1.53
SC26	4.08	1	7	1.71
SC27	5.37	1	7	1.44
SC28	4.39	1	7	1.7
SC29	5.48	1	7	1.35
SC30	4.46	1	7	1.77
SC31	5.28	1	7	1.4
SC32	5.45	2	7	1.38
SC33	5.97	2	7	1.38
SC34	5.58	1	7	1.53
SC35	5.72	1	7	1.48

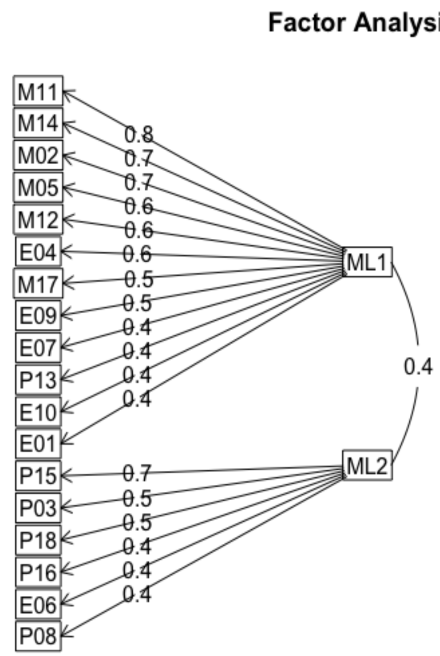
Inferential Analysis & Interpretation

1. Factor Analysis

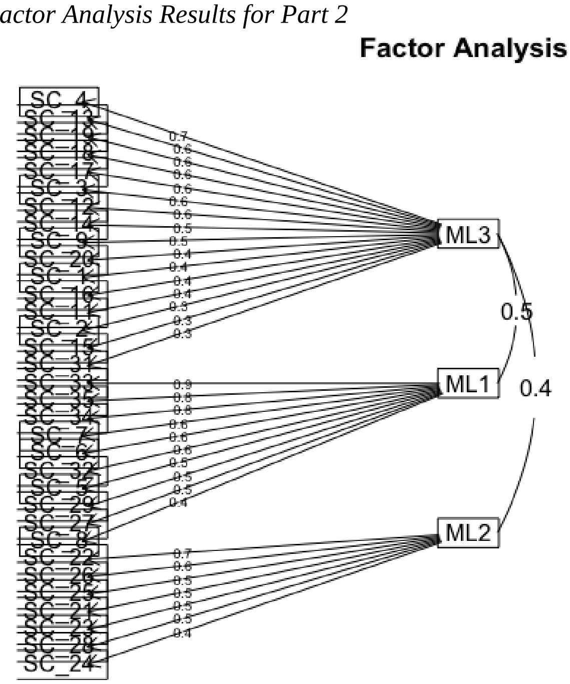
The main purpose of this study is to describe the covariance relationship among the many variables in the terms of a few underlying factors. In order to do so, we run exploratory factor analysis using maximum likelihood as a method of estimation separately for the variables in part 1 and part 2. The reason for doing factor analysis separately is that questions asked in part 1 are related to orientations to happiness while questions asked in part 2 are about adoption of sustainable consumption.

Before conducting the factor analysis, Bartlett's Sphericity test is used to check whether there is enough correlation among the variables and the result of this test is significant which means our data is adequate for the factor analysis.

Factor Analysis Results for Part 1



Factor Analysis Results for Part 2



After applying various kinds of techniques and rotations, we came up with two significant factors for part 1 and 3 factors for part 2 using oblimin rotation.

<u>Part 1 Factors</u>

ML1 - Includes all but 1 question from the meaning and engagement part of the OTH scale. Also includes 1 from the pleasure scale but is negligible. Thus we named this factor: **Meaning and Engagement**.

ML2 - Includes all but 1 pleasure question from the OTH scale. Also contains 1 question from the engagement part of the OTH scale but is negligible as well. Thus we named this factor: **Pleasure**.

Part 2 Factors

ML1 - includes questions that ask about saving water, turning off lights, switching electronics off when not in use, saving fuel, and only using air conditioning when necessary. Thus we named this factor: **Energy Conservation**.

ML2 - includes questions that ask about recycling cans, reusing things, recyclable materials, and refuse plastic bags. Thus we named this factor: **Three Rs** (Reduce, Reuse, and Recycle).

ML3 - includes questions that asks about buying energy efficient equipment, environmental concerns, environmentally friendly, social commitment, and environmental damages. Thus we named this factor: **Environmental Conscious**.

Measurement of the Factors

After naming the factors, our next step is to measure the factors for each observation. As our data is well cleaned we can simply take the average of the participants' responses in questions corresponding to each factor.

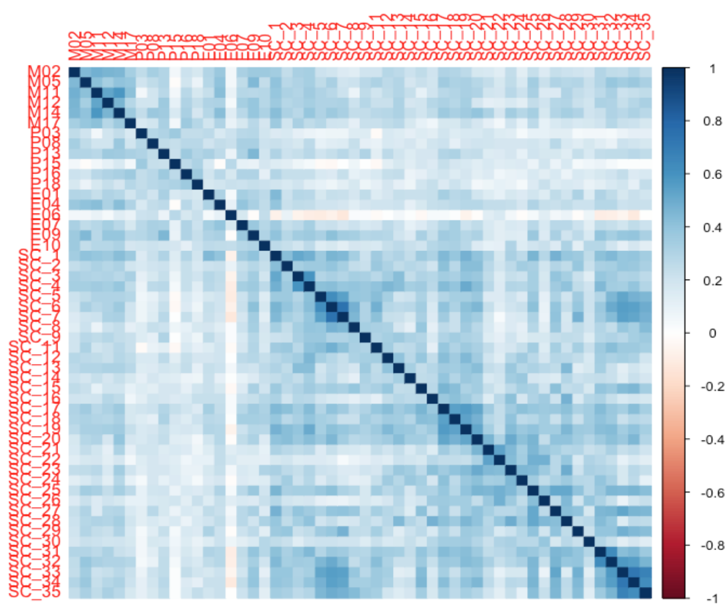
A value of 3.83333 corresponds to the average of all responses provided by a participant when he/she answered the questions on 'meaning' and 'engagement'.

```
> knitr::kable(head(data[,c(88, 89, 90, 93, 92)]), "pipe", align = "ccccc")


| MeaningAndEngagement | Pleasure | EnvironmentalConscious | ThreeRs  | EnergyConservation |
|:--------------------:|:--------:|:----------------------:|:--------:|:------------------:|
|       3.833333       | 5.000000 |         2.7500         | 3.142857 |        5.7         |
|       4.250000       | 5.000000 |         4.0625         | 3.428571 |        5.5         |
|       4.333333       | 2.333333 |         3.2500         | 2.714286 |        5.6         |
|       5.750000       | 4.000000 |         4.7500         | 6.000000 |        6.1         |
|       4.666667       | 5.500000 |         5.3125         | 5.428571 |        6.4         |
|       5.083333       | 5.000000 |         4.9375         | 3.714286 |        6.3         |
```

We can see the first five observations of measured factors. Each numeric value ranges from 1(not very like me) to 7(very like me). Now it is very easy to interpret our data based on the measured value. For example if the value of Environmental Consciousness for a person is close to 1 means the person is not very likely to conduct sustainable behavior that reflects as being environmentally conscious. Advantage of these measurements is that we can directly use this data for further analysis like Regression, LDA etc.

2. Principal Component Analysis (PCA)

It is clear from the plot to the right that correlation among the variables is not significant. PCA will not help us in reducing the number of dimensions significantly or extracting main variables. If we look at the table below, the first component explains 28% of the variance while the variance explained by the component 2 is only 6% and even loadings for these components is not significant which means we would end up with selecting most of the components to explain a significant amount of variance.



Correlation Plot

```
Importance of components:
                          Comp.1     Comp.2     Comp.3    Comp.4     Comp.5     Comp.6     Comp.7    Comp.8
Standard deviation     3.8571604 1.80275889 1.66380894 1.3746336 1.31066655 1.16616059 1.11428964 1.0954065
Proportion of Variance 0.2861093 0.06249884 0.05323577 0.0363388 0.03303552 0.02615251 0.02387772 0.0230753
Cumulative Proportion  0.2861093 0.34860819 0.40184396 0.4381828 0.47121827 0.49737078 0.52124850 0.5443238
                           Comp.9    Comp.10    Comp.11    Comp.12    Comp.13    Comp.14    Comp.15    Comp.16
Standard deviation     1.06833202 1.03922046 1.02979540 1.00397393 0.97396185 0.94046609 0.93003792 0.90233312
Proportion of Variance 0.02194872 0.02076883 0.02039382 0.01938392 0.01824234 0.01700916 0.01663405 0.01565779
Cumulative Proportion  0.56627251 0.58704134 0.60743516 0.62681908 0.64506142 0.66207058 0.67870463 0.69436242
                          Comp.17    Comp.18   Comp.19    Comp.20    Comp.21    Comp.22    Comp.23    Comp.24
Standard deviation     0.89339826 0.86523499 0.8576000 0.83446062 0.82521801 0.81473945 0.80280250 0.78741391
Proportion of Variance 0.01534924 0.01439676 0.0141438 0.01339086 0.01309586 0.01276539 0.01239407 0.01192347
Cumulative Proportion  0.70971166 0.72410842 0.7382522 0.75164308 0.76473894 0.77750433 0.78989841 0.80182188
                          Comp.25    Comp.26    Comp.27    Comp.28     Comp.29     Comp.30     Comp.31     Comp.32
Standard deviation     0.77913513 0.76436816 0.74179807 0.72522857 0.720943422 0.708309688 0.696820964 0.680802635
Proportion of Variance 0.01167407 0.01123574 0.01058201 0.01011455 0.009995373 0.009648127 0.009337682 0.008913312
Cumulative Proportion  0.81349595 0.82473169 0.83531370 0.84542825 0.855423623 0.865071750 0.874409432 0.883322744
                           Comp.33     Comp.34     Comp.35     Comp.36     Comp.37     Comp.38     Comp.39
Standard deviation     0.676360366 0.655125300 0.636484405 0.626277880 0.613453849 0.611963132 0.605960599
Proportion of Variance 0.008797372 0.008253638 0.007790623 0.007542769 0.007237031 0.007201901 0.007061312
Cumulative Proportion  0.892120116 0.900373754 0.908164377 0.915707145 0.922944177 0.930146078 0.937207391
                           Comp.40     Comp.41     Comp.42     Comp.43     Comp.44     Comp.45     Comp.46
Standard deviation     0.589813460 0.570700450 0.567243365 0.552797090 0.541280009 0.516831715 0.498702696
Proportion of Variance 0.006689998 0.006263442 0.006187789 0.005876627 0.005634309 0.005136827 0.004782777
Cumulative Proportion  0.943897389 0.950160831 0.956348621 0.962225248 0.967859557 0.972996384 0.977779160
                           Comp.47     Comp.48     Comp.49     Comp.50     Comp.51     Comp.52
Standard deviation     0.475753890 0.468886540 0.450855659 0.429563545 0.412606343 0.388905707
Proportion of Variance 0.004352726 0.004227973 0.003909054 0.003548555 0.003273923 0.002908609
Cumulative Proportion  0.982131887 0.986359859 0.990268914 0.993817468 0.997091391 1.000000000
```

PCA Components and Proportion Table

Why doing PCA is not recommended for this study

As we know, the main motive of this data set is to explain the interrelationships among the variables and factor analysis is useful for that. On the other hand PCA reduces the dimensions according to the maximum variance in the original variables without explaining how these reduced dimensions can be linked to the original variables (interpretability lost).

3. Stepwise Regression

Objective

Previously, we mentioned "Orientation of Happiness" (OTH) can have positive and negative relationships with Sustainable Behavior (SB). "Meaning" and "Engagement" are positively related to SB; "Pleasure" is negatively related to SB. To further testify relationships between OTH and different categories of SBs, we apply stepwise regression.

Methodology

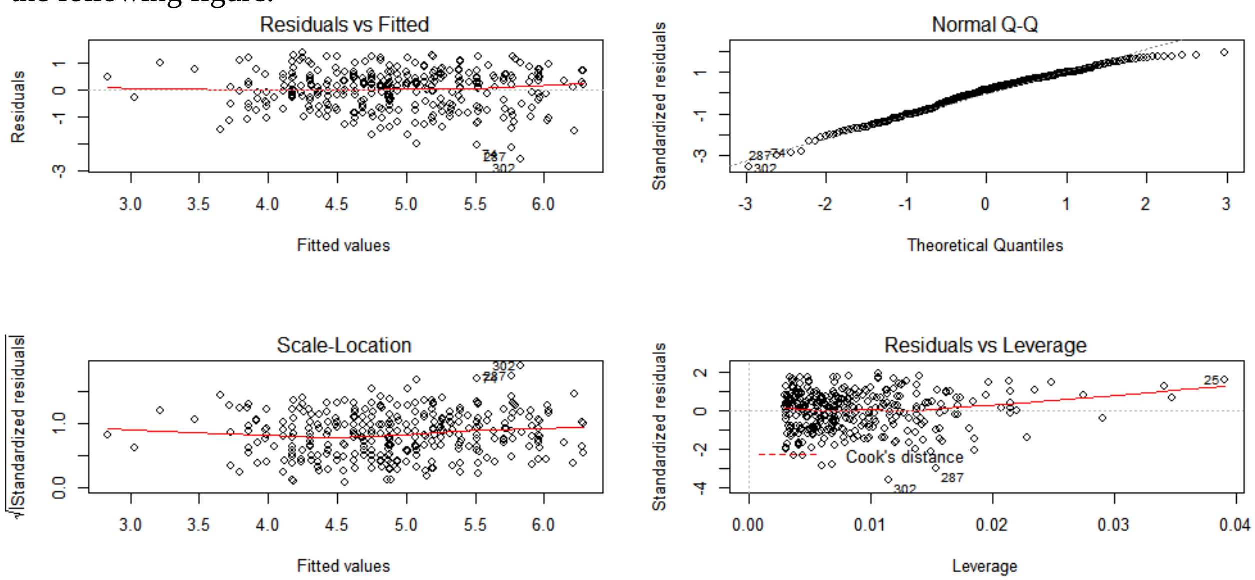
In our analysis, we used the 5 factors derived from our factor analysis to create the models. Part 1 factors belong to OTH while part 2 factors belong to SB. Therefore, part 1 factors are used as predictors while part 2 factors are used as response variables. Part 1 factors are Meaning & Engagement, and Pleasure. Part 2 factors are Environmental Conscious, Three Rs, and Energy Conservation.

There would be 3 stepwise regression models. They all share the same predictors but have different response variables. The first model (sb1) uses Environmental Conscious as the response variable. The second model (sb2) uses Three Rs as the response variable. The third model (sb3) uses energy conservation as the response variable.

We first build three multiple linear regression models to inspect their linear regression assumptions which includes normality, linearity, homogeneity, and homoscedasticity. Then, we removed leverage points and influence points measured by Cook's distance for each of the three models. Using the clean models, we inspect their linear regression assumptions again. Finally, we created three stepwise regression models to discover any important predictors and test our hypothesis.

First Model (sb1) Building

In the inspection of assumptions, we have the following observations. It fails normality as not all the points fall approximately along the reference line in the Normal Q-Q plot. It passes linearity as the red line is close to the dashed line in the residuals vs fitted plot. It fails homogeneity and homoscedasticity as the red line is not horizontal in the residuals vs fitted plot. This shows that the average magnitude of the standardized residual is changing as a function of the fitted values. The spread around the red line varies with the fitted values so the variability of magnitudes varies as a function of the fitted values. It became necessary to remove outliers. Please refer to the following figure.



Inspection of assumptions for sb1

As we removed some outliers measured by the Mahalanobis distance in the beginning, we did not repeat this step. After removing leverage points and influence points, 8 more outliers were removed. We retained 330 observations for model sb1. Redoing the inspection of assumption, there was no significant change as we removed a few outliers only. Still, we wanted to reduce uncorrelated predictors by applying stepwise regression.

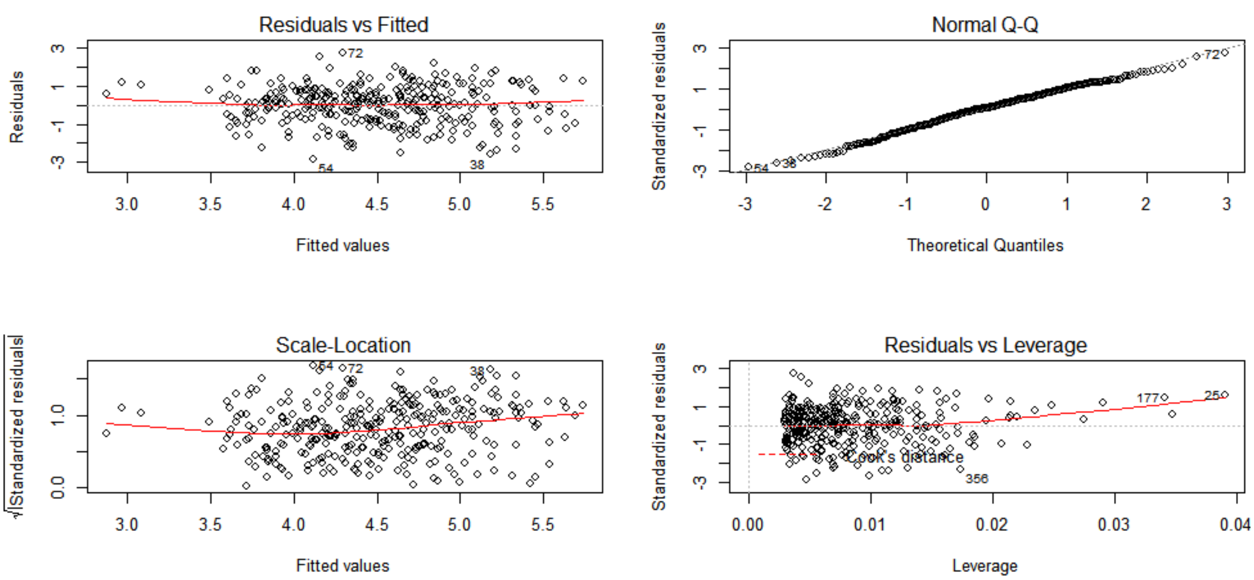
At last, we built a stepwise regression model. After comparing AIC and beta, we confirmed that only 1 Part 1 factor – Meaning and Engagement, is a qualified predictor for the model. Another Part 1 factor – Pleasure should be removed. In the model results, Meaning and Engagement has an extremely small p-value which is approximately 0. It also has a quite large beta (0.6636543) which indicates it is an important predictor. Please refer to the following figure.

```
> lm.step.one.sb1 <- lm(EnvironmentalConscious ~ MeaningAndEngagement, data = inlinersb1); summary(lm.step.one.sb
1)

Call:
lm(formula = EnvironmentalConscious ~ MeaningAndEngagement, data = inlinersb1)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5779 -0.5326  0.1234  0.5047  1.3895 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           0.92320    0.25228   3.659 0.000294 ***
MeaningAndEngagement  0.76437    0.04757  16.068  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7114 on 328 degrees of freedom
Multiple R-squared:  0.4404,	Adjusted R-squared:  0.4387 
F-statistic: 258.2 on 1 and 328 DF,  p-value: < 2.2e-16

> library(QuantPsyc); lm.beta(lm.step.sb1)
MeaningAndEngagement 
           0.6636543 
```

Model Results of sb1

Second Model (sb2) Building

In the inspection of assumptions, we have the following observations. It passes normality as all the points fall approximately along the reference line in the Normal Q-Q plot. It passes linearity as the red line is close to the dashed line in the residuals vs fitted plot. It fails homogeneity and homoscedasticity as the red line is not horizontal in the residuals vs fitted plot. This shows that the average magnitude of the standardized residual is changing as a function of the fitted values. The spread around the red line varies with the fitted values so the variability of magnitudes varies as a function of the fitted values. It became necessary to remove outliers. Please refer to the following figure.



Inspection of assumptions for sb2

As we removed some outliers measured by the Mahalanobis distance in the beginning, we did not repeat this step. After removing leverage points and influence points, 9 more outliers were removed. We retained 329 observations for model sb2. Redoing the inspection of assumption, there was no significant change as we removed few outliers only. Still, we wanted to reduce uncorrelated predictors by applying stepwise regression.

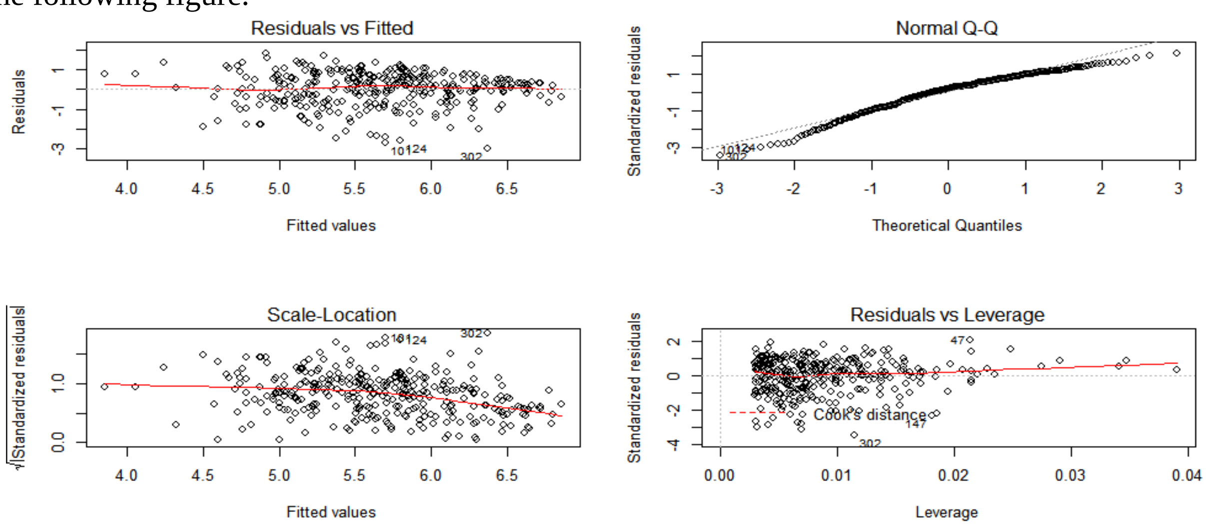
At last, we built a stepwise regression model. After comparing AIC and beta, we confirmed that only 1 Part 1 factor – Meaning and Engagement, is a qualified predictor for the model. Another Part 1 factor – Pleasure should be removed. In the model results, Meaning and Engagement has an extremely small p-value which is approximately 0. It also has a quite large beta (0.6316773) which indicates it is an important predictor. Please refer to the following figure.

```
> lm.step.one.sb2 <- lm(EnvironmentalConscious ~ MeaningAndEngagement, data = inlinersb2); summary(lm.step.one.
sb2)

Call:
lm(formula = EnvironmentalConscious ~ MeaningAndEngagement, data = inlinersb2)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5202 -0.5428  0.1222  0.5235  1.3703 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           1.10226    0.26012   4.238 2.94e-05 ***
MeaningAndEngagement  0.72748    0.04937  14.735  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7206 on 327 degrees of freedom
Multiple R-squared:  0.399,	Adjusted R-squared:  0.3972 
F-statistic: 217.1 on 1 and 327 DF,  p-value: < 2.2e-16

> library(QuantPsyc); lm.beta(lm.step.sb2)
MeaningAndEngagement 
           0.6316773 
```

Model Results of sb2

Third Model (sb3) Building

In the inspection of assumptions, we have the following observations. It fails normality as not all the points fall approximately along the reference line in the Normal Q-Q plot. It passes linearity as the red line is close to the dashed line in the residuals vs fitted plot. It fails homogeneity and homoscedasticity as the red line is not horizontal in the residuals vs fitted plot. This shows that the average magnitude of the standardized residual is changing as a function of the fitted values. The spread around the red line varies with the fitted values so the variability of magnitudes varies as a function of the fitted values. It became necessary to remove outliers. Please refer to the following figure.



Inspection of assumptions for sb3

As we removed some outliers measured by the mahalanobis distance in the beginning, we did not repeat this step. After removing leverage points and influence points, 5 more outliers were removed. We retained 333 observations for model sb3. Redoing the inspection of assumption, there was no significant change as we removed few outliers only. Still, we wanted to reduce uncorrelated predictors by applying stepwise regression.

At last, we built a stepwise regression model. After comparing AIC and beta, we confirmed that only 1 Part 1 factor – Meaning and Engagement, is a qualified predictor for the model. Another Part 1 factor – Pleasure should be removed. In the model results, Meaning and Engagement has an extremely small p-value which is approximately 0. It also has a quite large beta (0.58524790) which indicates it is an important predictor. Please refer to the following figure.

```
> lm.step.one.sb3 <- lm(EnergyConservation ~ MeaningAndEngagement, data = inlinersb3); summary(lm.step.one.sb3)

Call:
lm(formula = EnergyConservation ~ MeaningAndEngagement, data = inlinersb3)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0508 -0.5168  0.1392  0.6160  1.7168 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           2.16954    0.30039   7.222 3.53e-12 ***
MeaningAndEngagement  0.66721    0.05659  11.790  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8543 on 331 degrees of freedom
Multiple R-squared:  0.2958,	Adjusted R-squared:  0.2936 
F-statistic:   139 on 1 and 331 DF,  p-value: < 2.2e-16

> library(QuantPsyc); lm.beta(lm.step.sb3)
MeaningAndEngagement             Pleasure 
          0.58524790          -0.08654328 
```

Model Results of sb3

Conclusion

All 3 stepwise regression models share the same predictor selection results. Based on our analysis, we can show the significance of “Meaning” and “Engagement” but not “Pleasure”. We can conclude that “Meaning” and “Engagement” of OTH are positively related to the adoption of various SB. For “Pleasure” of OTH, we can’t tell the relationship between it and SB.

4. Linear Discriminant Analysis

Objective

Using linear discriminant analysis to develop discriminant functions which uses different combination of independent variables in our dataset to discriminate important categorical variables. This helps us discover the relationship between OTH and SB.

Methodology

In our analysis, we used different combinations of independent variables from partI and part 2 of the questionnaire to predict categorical variables of the participants including Sex, Job type, and education. The dataset would partition 70% of our data into a training set and 30% into a testing set. We used *greedy.wilks()* function from *library klaR* to help select and keep only important **predictors**, then used *lda function()* from *library MASS* in R software to generate linear discriminant function to predict for training and testing set. Finally, contingency tables and accuracy rate for training set and testing set will be compared to test our hypothesis.

Prediction of Sex type

We believe sex type would impact opinions towards happiness and sustainable consumptions. The first attempt to predict sex type used five aggregate factors we concluded from Factor Analysis: Meaning And Engagement, Pleasure, Environmental Conscious, Energy Conservation, and Three Rs.

```
Values calculated in each step of the selection procedure:

  vars Wilks.lambda F.statistics.overall p.value.overall F.statistics.diff p.value.diff
1   v5        0.965                 8.19         0.00462              8.19      0.00462
```

Training Set

```
> confusiontab.one
         Actual
Predicted   0   1
        0   9  13
        1  72 132
> sum(diag(confusiontab.one))/sum(confusiontab.one)
[1] 0.624
```

Testing Set

```
> confusiontab2
         Actual
Predicted  0  1
        0  5  6
        1 33 73
> sum(diag(confusiontab2))/sum(confusiontab2)
[1] 0.667
```

Using Five Factors to Predict Sex

After variable selection, there was only one variable left: V5-Energy Conservation. We conclude from the contingency tables that the accuracy rate of training and test set are 62.4% and 66.7% respectively. We noticed that in our dataset the overall proportion of female participants is 62.5% which indicates that the discriminant function was not good. Then we decided to use more variables with different combinations to predict sex type.

```
  vars Wilks.lambda F.statistics.overall p.value.overall F.statistics.diff p.value.diff
1  E06        0.948                14.26       0.0001975             14.26     0.000197
2  P15        0.926                10.27       0.0000509              6.01     0.014862
3  E07        0.920                 7.48       0.0000804              1.83     0.177052
```

Training Set

```
> confusiontab.one
         Actual
Predicted   0   1
        0  15  23
        1  77 146
> sum(diag(confusiontab.one))/sum(confusiontab.on
[1] 0.617
```

Testing Set

```
> confusiontab2
         Actual
Predicted  0  1
        0  5  6
        1 22 49
> sum(diag(confusiontab2))/sum(confusiontab2)
[1] 0.659
```

Above: Using All Variables from Part1 to Predict Sex

	vars	Wilks.lambda	F.statistics.overall	p.value.overall	F.statistics.diff	p.value.diff
1	SC_19	0.959	10.14	0.001641135	10.14	0.00164
2	E06	0.932	8.59	0.000249896	6.79	0.00973
3	SC_15	0.911	7.73	0.000060213	5.67	0.01806
4	SC_23	0.897	6.77	0.000036004	3.61	0.05849
5	SC_28	0.873	6.81	0.000006016	6.35	0.01239
6	SC_12	0.858	6.41	0.000002903	3.98	0.04717
7	E01	0.842	6.22	0.000001081	4.53	0.03426
8	SC_14	0.833	5.80	0.000000945	2.56	0.11111
9	M12	0.826	5.39	0.000001050	1.94	0.16525
10	SC_7	0.817	5.12	0.000000955	2.35	0.12693
11	SC_32	0.806	4.99	0.000000593	3.22	0.07402
12	SC_11	0.799	4.76	0.000000616	1.96	0.16244

Training Set

```
> confusiontab.one
         Actual
Predicted   0   1
        0  38  24
        1  46 125
> sum(diag(confusiontab.one))/sum(con
[1] 0.7
```

Testing Set

```
> confusiontab2
         Actual
Predicted  0  1
        0  9 17
        1 26 58
> sum(diag(confusiontab2))/sum(confusiontab2)
[1] 0.609
```

Above: Using All Variables from Part1&2 to Predict Sex

As we can see from above that after adding more variables, the discriminant function still didn’t perform well in separating sex type. Even with all the variables from part 1 and 2, there was only little improvement for the training set and a slight drop in accuracy in the testing set. At this point, we can conclude that sex type is not separable using variables in our data.

Prediction of Job type

We performed similar approach predicting for job type as predicting for sex. Unlike sex which has only two levels, the job type variable has seven levels which could make it harder to separate. First, using five factors from Factor Analysis to predict job type and then using all variables from Part 1 and 2.

	vars	Wilks.lambda	F.statistics.overall	p.value.overall	F.statistics.diff	p.value.diff
1	v5	0.896	4.16	0.000567	4.16	0.000567
2	v1	0.841	3.22	0.000191	2.33	0.033904
3	v2	0.795	2.81	0.000100	2.00	0.066641

Training Set

```
> confusiontab.one
         Actual
Predicted  1  2  3  4  5  6  7
        1 84 22 33 15 34  5  5
        2  1  2  1  0  0  0  0
        3  6  1  7  2  0  0  0
        4  0  0  0  0  0  0  0
        5  0  1  0  0  2  0  0
        6  0  0  0  0  0  0  0
        7  0  0  0  0  0  0  0
> sum(diag(confusiontab.one))/sum(confusiontab.one)
[1] 0.43
```

Testing Set

```
> confusiontab2
         Actual
Predicted  1  2  3  4  5  6  7
        1 45  7 24  6 17  0  0
        2  1  0  0  0  0  0  0
        3  4  1  2  1  1  0  0
        4  0  0  0  0  0  0  0
        5  0  0  0  0  1  0  0
        6  0  0  0  0  0  0  0
        7  0  1  0  0  0  0  0
> sum(diag(confusiontab2))/sum(confusiontab2)
[1] 0.432
```

Using Five Factors to Predict Job

Coefficients of linear discriminants:

	LD1	LD2	LD3	LD4	LD5	LD6
E06	-0.0702	-0.3214	-0.30933	-0.0922	0.1436	-0.1438
SC_4	0.2811	0.1699	-0.11473	-0.1005	0.0973	0.1752
SC_30	0.3335	-0.1812	0.08059	0.2116	0.1419	0.0211
E04	-0.5843	-0.0106	0.10393	-0.0514	0.2380	0.1985
SC_33	0.1353	0.0173	-0.11488	-0.4301	-0.0271	-0.4389
SC_28	-0.2546	0.2194	0.03614	0.1440	0.1561	-0.1066
E07	0.3069	0.2272	0.26597	0.4108	-0.2519	-0.0490
SC_11	0.1431	-0.0948	0.30987	-0.1014	-0.2288	-0.3086
SC_22	-0.0131	-0.2176	0.06081	0.0749	-0.3903	0.0172
SC_18	-0.2361	0.1697	-0.45051	-0.3265	-0.3151	0.0476
SC_29	-0.0604	-0.2310	-0.36641	0.1292	-0.1481	0.0578
SC_15	0.1387	-0.0236	0.37926	0.0124	0.0287	0.1334
E01	0.2103	0.2342	-0.14735	-0.0256	0.2578	-0.2704
P13	0.3081	-0.2639	0.18382	-0.1896	0.3791	0.2534
SC_25	0.2792	0.1079	0.03477	-0.3292	0.1927	0.2723
SC_1	-0.2299	-0.1815	-0.00806	0.0962	-0.1704	-0.2246
P03	-0.1666	0.1343	-0.20862	-0.3035	0.1556	-0.0656
SC_31	-0.2656	0.2168	0.20973	0.0578	-0.0212	-0.1010
SC_17	0.1805	0.0178	-0.10564	0.3077	0.2595	0.1594
SC_24	0.0386	0.0329	-0.33538	-0.1535	-0.1224	0.2398
SC_23	-0.1745	-0.1410	0.16221	-0.0933	0.2393	-0.3619
M14	0.1664	0.1018	-0.08900	0.2988	-0.2859	-0.1702
P15	-0.0462	0.0507	0.17246	0.1998	-0.1730	0.0251

Training Set

```
> confusiontab.one
         Actual
Predicted  1  2  3  4  5  6  7
        1 78 10 18  3 10  0  2
        2  4 11  2  0  1  0  0
        3  8  1 27  1  4  1  0
        4  1  1  0  9  2  0  0
        5  2  3  2  1 14  0  0
        6  1  0  0  1  0  4  0
        7  0  0  0  0  0  0  2
> sum(diag(confusiontab.one))/sum(confusiont
[1] 0.647
```

Testing Set

```
> confusiontab2
         Actual
Predicted  1  2  3  4  5  6  7
        1 30  3 11  1 16  0  1
        2  6  0  1  0  2  0  0
        3  7  2  5  4  2  0  0
        4  2  1  1  4  1  0  0
        5  2  2  0  0  3  0  0
        6  0  1  0  0  0  0  0
        7  0  0  0  0  0  0  0
> sum(diag(confusiontab2))/sum(confusic
[1] 0.389
```

Using All Variables from Part1&2 to Predict Sex

We can see from the outputs above, the accuracy rates using only five factors for both training and testing set are about 43% which is very low. But after we added more predictors, the training set accuracy rate for training set increased to 64.7%, but accuracy rate for testing set decreased to 38.9%. One possible reason is that we might have over fitted our model by using too many variables. We can say that job type is not separable using variables in our data.

Prediction of Education

	vars	Wilks.lambda	F.statistics.overall	p.value.overall	F.statistics.diff	p.value.diff
1	v3	0.914	4.31	0.000908	4.31	0.000908
2	v2	0.884	2.91	0.001529	1.56	0.171156

Training Set

```
> confusiontab.one
         Actual
Predicted   1   2   3   4   5   6   7
        1   0   0   0   0   0   0   0
        2   0   0   0   0   0   0   0
        3   0   0   0   0   0   0   0
        4   0   2  13 147   3  43  26
        5   0   0   0   0   0   0   0
        6   0   0   0   0   0   0   0
        7   0   0   0   1   0   0   0
> sum(diag(confusiontab.one))/sum(confusiontab.one)
[1] 0.626
```

Testing Set

```
> confusiontab2
         Actual
Predicted  1  2  3  4  5  6  7
        1  0  0  0  0  0  0  0
        2  0  0  0  0  0  0  0
        3  0  0  0  0  0  0  0
        4  1  2  7 53  3 26 12
        5  0  0  0  0  0  0  0
        6  0  0  0  0  0  0  0
        7  0  0  0  0  0  0  0
> sum(diag(confusiontab2))/sum(confusiontab2)
[1] 0.51
```

Using Five Factors to Predict Education

In the last part of linear discriminant analysis, we tried to predict for education. Same as Job type, Education also has seven levels. After variable selection, there are only two variables left. Although the accuracy rates are 62.6% and 51% in our tables, we can see that discriminant function predicted only level 4.

```
Values calculated in each step of the selection procedure:

    vars Wilks.lambda F.statistics.overall p.value.overall F.statistics.diff p.value.diff
1  SC_15        0.899                 4.16  0.000549774869              4.16      0.00055
2  SC_30        0.833                 3.53  0.000051411763              2.92      0.00923
3  SC_25        0.780                 3.18  0.000011183946              2.48      0.02427
4    P08        0.733                 2.97  0.000002856734              2.37      0.03088
5    E04        0.689                 2.84  0.000000831958              2.28      0.03712
6    M05        0.648                 2.75  0.000000236421              2.28      0.03713
7  SC_22        0.618                 2.62  0.000000172040              1.77      0.10689
8   SC_9        0.588                 2.52  0.000000107125              1.83      0.09504
9  SC_11        0.551                 2.52  0.000000023228              2.41      0.02850
10  SC_7        0.520                 2.49  0.000000008311              2.12      0.05173
11 SC_20        0.491                 2.45  0.000000003476              2.03      0.06259
12   M17        0.467                 2.41  0.000000001890              1.88      0.08587
13 SC_16        0.442                 2.38  0.000000000970              1.91      0.08132
14  SC_4        0.423                 2.33  0.000000000825              1.60      0.14764
15 SC_28        0.403                 2.29  0.000000000564              1.72      0.11755
16  SC_1        0.386                 2.25  0.000000000542              1.51      0.17658
```

```
> confusiontab.one
         Actual
Predicted  1  2  3  4  5  6  7
        1 78 10 18  3 10  0  2
        2  4 11  2  0  1  0  0
        3  8  1 27  1  4  1  0
        4  1  1  0  9  2  0  0
        5  2  3  2  1 14  0  0
        6  1  0  0  1  0  4  0
        7  0  0  0  0  0  0  2
> sum(diag(confusiontab.one))/sum(confusiontab.
[1] 0.647
```

```
> confusiontab2
         Actual
Predicted  1  2  3  4  5  6  7
        1 30  3 11  1 16  0  1
        2  6  0  1  0  2  0  0
        3  7  2  5  4  2  0  0
        4  2  1  1  4  1  0  0
        5  2  2  0  0  3  0  0
        6  0  1  0  0  0  0  0
        7  0  0  0  0  0  0  0
> sum(diag(confusiontab2))/sum(confu
[1] 0.389
```

Using All Variables from Part1&2 to Predict Education

We tried to add more predictors, it caused the same problems as predicting for Job type that too many predictors have over fitted our model. We can conclude that Education is not separable using five factors variables in our data.

Conclusion

In conclusion, the independent variables in our dataset are not supportive to use discriminant function to separate sex, job, and education. In other words, discriminant analysis provided insignificant results. We do not have enough evidence to say sex, job, and education have impacts on orientations towards happiness and sustainable behaviors.

5. Correspondent Analysis (CA)

Correspondence analysis (CA) is an extension of principal component analysis suited to explore relationships among qualitative variables (or categorical data). Like principal component analysis, it provides a solution for summarizing and visualizing data sets in two-dimension plots. Since the happiness dataset is more of categorical data, it makes sense to conduct a Correspondence analysis and extract most of the insights out of it.

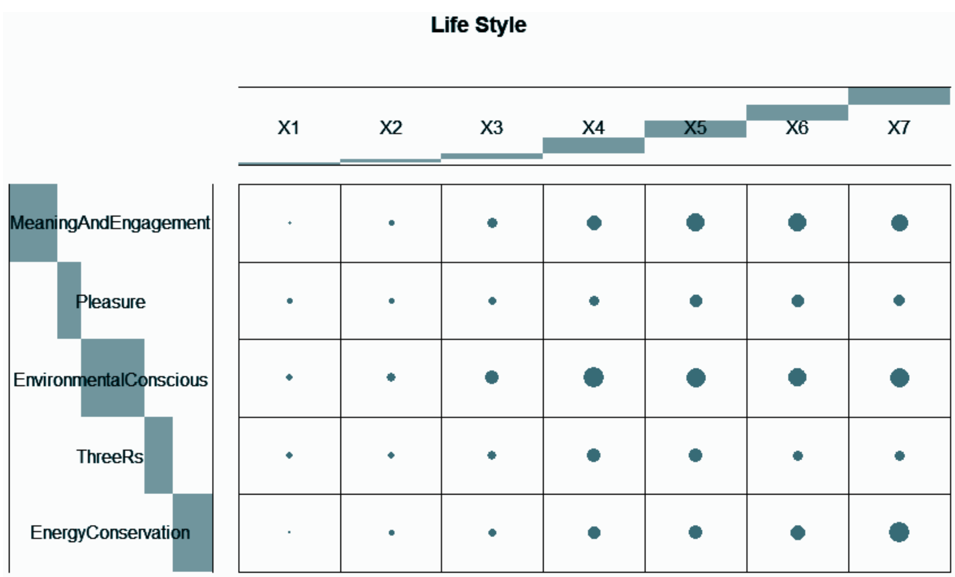
In order to start with Correspondence analysis, we needed to first derive the contingency table. We can derive the contingency table using the five factors we extracted using Factor Analysis, namely **Meaning and Engagement, Pleasure, Environmental Conscious, Three Rs** and **Energy Conservation**. With the factors mentioned earlier, we can group the questions which are in line with the factors and using the aggregate method we are able to derive the contingency table below.

	X1	X2	X3	X4	X5	X6	X7
MeaningAndEngagement	31	110	293	724	1038	1019	841
Pleasure	106	107	183	335	480	449	368
EnvironmentalConscious	144	243	536	1258	1102	1010	1115
ThreeRs	131	149	282	598	539	352	315
EnergyConservation	20	85	179	496	531	713	1356

Where the rows are the factors derived from the factor analysis and the columns are Likert Scales (1-7).

It provides factor scores (coordinates) for both row and column points of the contingency table. These coordinates are used to visualize graphically the association between row and column elements in the contingency table.

We can use balloon plot to graphically derive the insights from contingency table:



From the balloon plot it's evident that Vietnamese people were more environmentally and energy conscious, though less importance were given to recycle, reuse and reduce methodology maybe because these complicated methods are less promoted by the government to the people of Vietnam.

Statistical significance

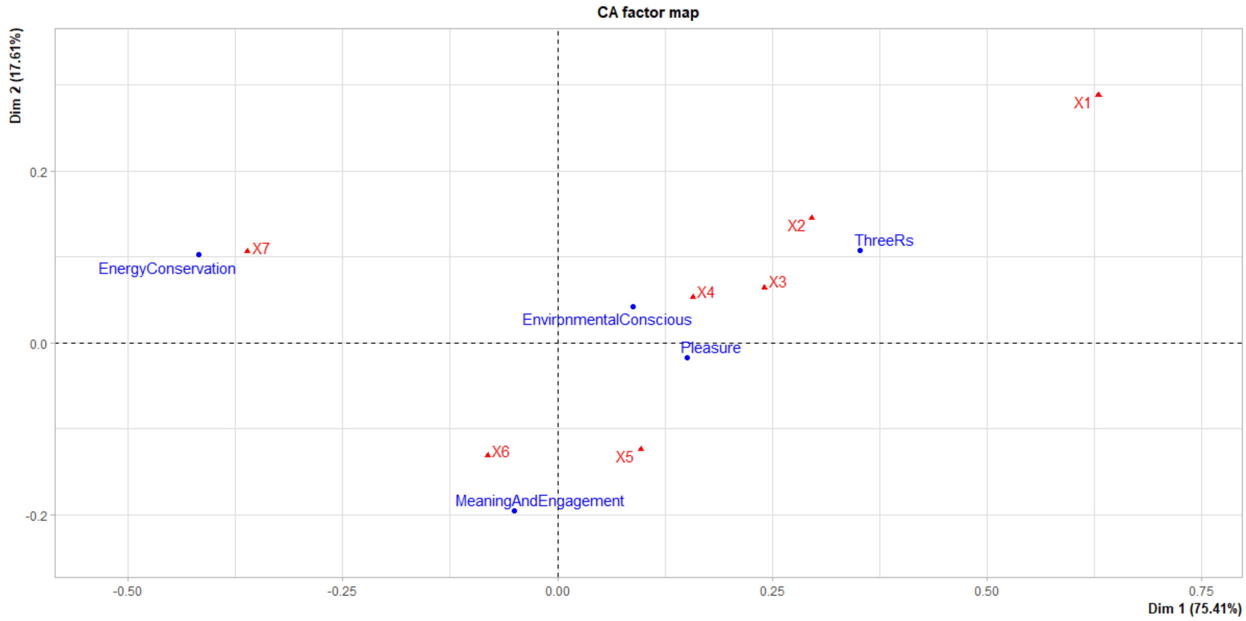
To interpret correspondence analysis, the first step is to evaluate whether there is a significant dependency between the rows and columns. In order to verify the contingency theoretically we can perform a chi-square test, if the p-value is more than 0.05, then we can conclude that there is no relation between the columns and rows. A high chi-square statistic means a strong link between row and column variables. In OTH survey data's contingency table p-value is as less as **2.2e^-16**, making it evident that there is significance between the rows and columns.

```
        Pearson's Chi-squared test

data:  mca
X-squared = 1299.9, df = 24, p-value < 2.2e-16
```

There is strong significance between the rows and columns in OTH survey contingency table

Evaluating Quality of Fit

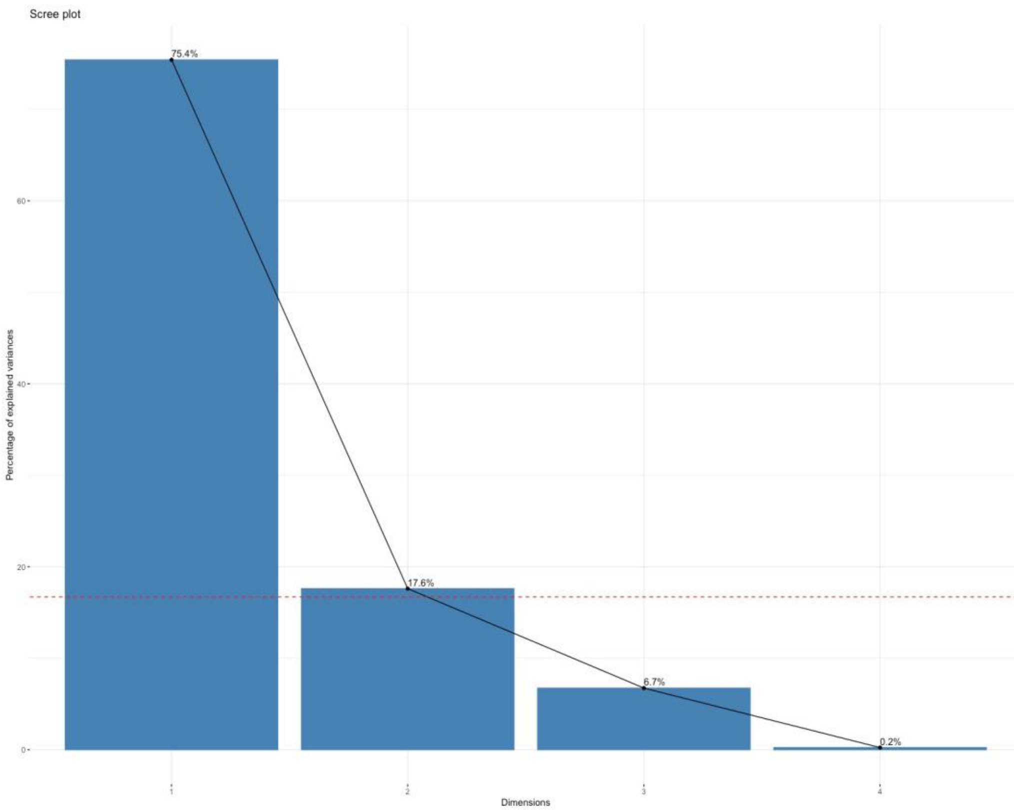
The distance between any row points or column points gives a measurement of their similarity (or dissimilarity). Row points with similar profiles are close in proximity on the factor map. The same holds true for column points. Factor map plot can be used to evaluate the quality of CA fit of OTH survey data, the plot is given below.



X-axis has an inertia of **75.41%** and y-axis has an inertia of **17.61%**. We can use distance metric in order to group the similarity and attributes of the group. It's evident that energy conservation, meaning and engagement scores highly among our audience, being placed in the same side of the plot also makes it evident that both are the same group of audience. On the other hand, recycling, reuse, and reduce and environmental consciousness did not score as high or takes a neutral point among the audience. Pleasure in this case remains near the origin, making it tough to determine which group it belongs to. The next step for the interpretation is to determine which row and column variables contribute the most in the definition of the different dimensions retained in the model.

Eigenvalues / Variances

We examine the eigenvalues to determine the number of axes to be considered. *Eigenvalues* correspond to the amount of information retained by each axis. Dimensions are ordered decreasingly and listed according to the amount of variance explained in the solution. Dimension 1 explains the most variance in the solution, followed by dimension 2 and so on. The cumulative percentage explained is obtained by adding the successive proportions of variation explained to obtain the running total.



Scree Plot

Dimensions 1 and 2 explain approximately 75.4% and 17.6% of the total inertia respectively. This corresponds to a cumulative total of **93%** of total inertia retained by the 2 dimensions. The higher the retention, the more subtlety in the original data is retained in the low-dimensional solution. It makes sense to retain only first-two dimensions.

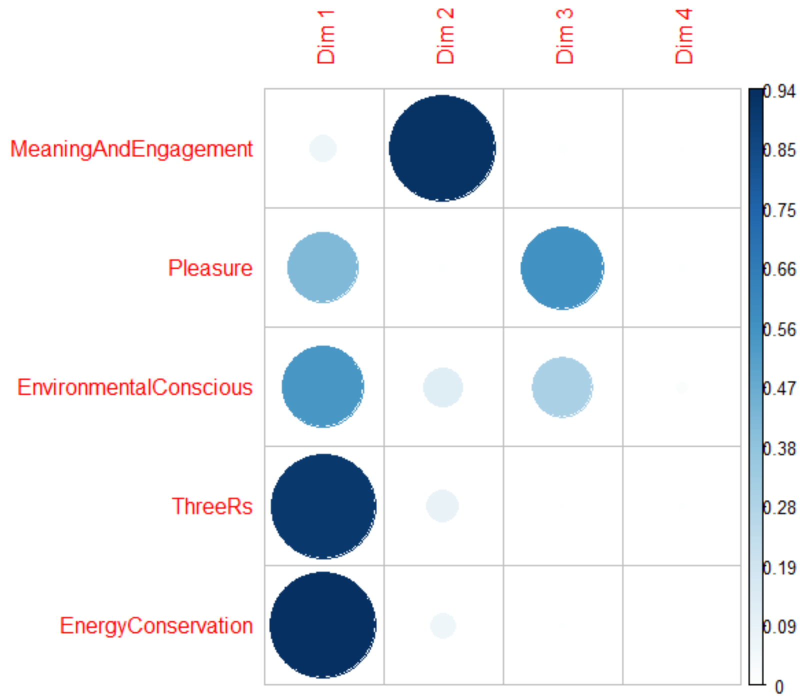
Quality of representation of rows

The result of the analysis shows that the contingency table has been successfully represented in low dimension space using correspondence analysis. Dimensions 1 and 2 are sufficient to retain 93% of the total inertia (variation) contained in the data.

	Dim 1	Dim 2	Dim 3	Dim 4
MeaningAndEngagement	0.06055141	0.931059743	7.280797e-03	0.0011080529
Pleasure	0.42231331	0.005482881	5.685890e-01	0.0036147609
EnvironmentalConscious	0.54970606	0.128089711	3.084594e-01	0.0137447983
ThreeRs	0.91070940	0.085026635	3.396287e-05	0.0042300052
EnergyConservation	0.94111931	0.056655858	2.056597e-03	0.0001682307

Summary of Dimensions

The function row$cos2 can be used to obtain the inertia of rows being distributed on each dimension, we can see that the first two dimensions take up most of the variance and dimension three contributes only to Pleasure. This table can be graphically represented below.



Row$cos2 Plot

From the balloon plot, it's evident that dimension-1 takes up the variance from recycle, reduce, reuse, environmental consciousness and energy conservation. Dimension-2 takes up the variance only from the meaning and engagement, dimension-3 takes up the variance from pleasure. Dimension-4 takes up negligible variance from any of the sub categories of OTH.

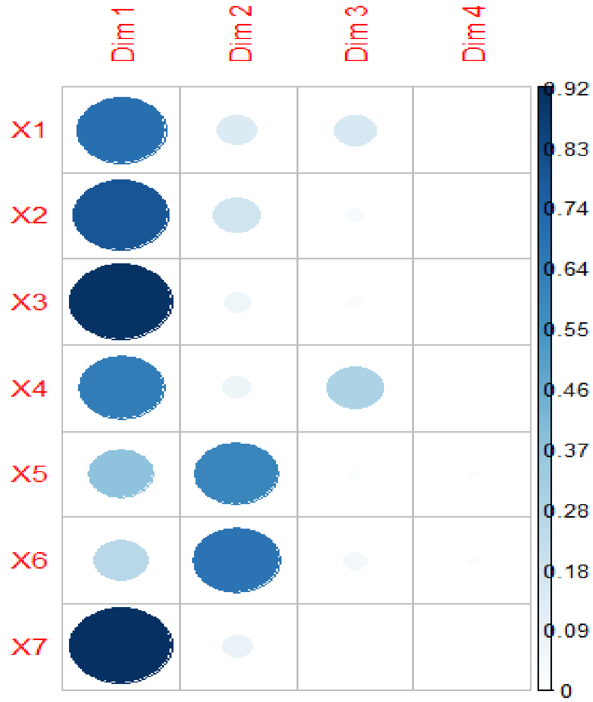
Quality of representation of columns

As stated, two dimensions 1 and 2 are sufficient to retain 93% of the total inertia (variation) contained in the data. Now we can see the distribution of inertia of the Likert Scale on the dimensions.

	Dim 1	Dim 2	Dim 3	Dim 4
X1	0.6970364	0.14625346	0.156710145	4.424190e-08
X2	0.7816627	0.18916899	0.028835035	3.332750e-04
X3	0.9094433	0.06366625	0.025021872	1.868583e-03
X4	0.6415046	0.07201690	0.286441118	3.742081e-05
X5	0.3692117	0.60073441	0.014371502	1.568240e-02
X6	0.2557450	0.67853813	0.051181895	1.453493e-02
X7	0.9193661	0.07931386	0.001160154	1.598479e-04

Likert Scale Dimensions

The function col$cos2 can be used to get the inertia of rows being distributed on each dimension, we can see that the first-two dimension takes up most of the variance and dimension-three contributes only to X4, this table can be graphically represented below.

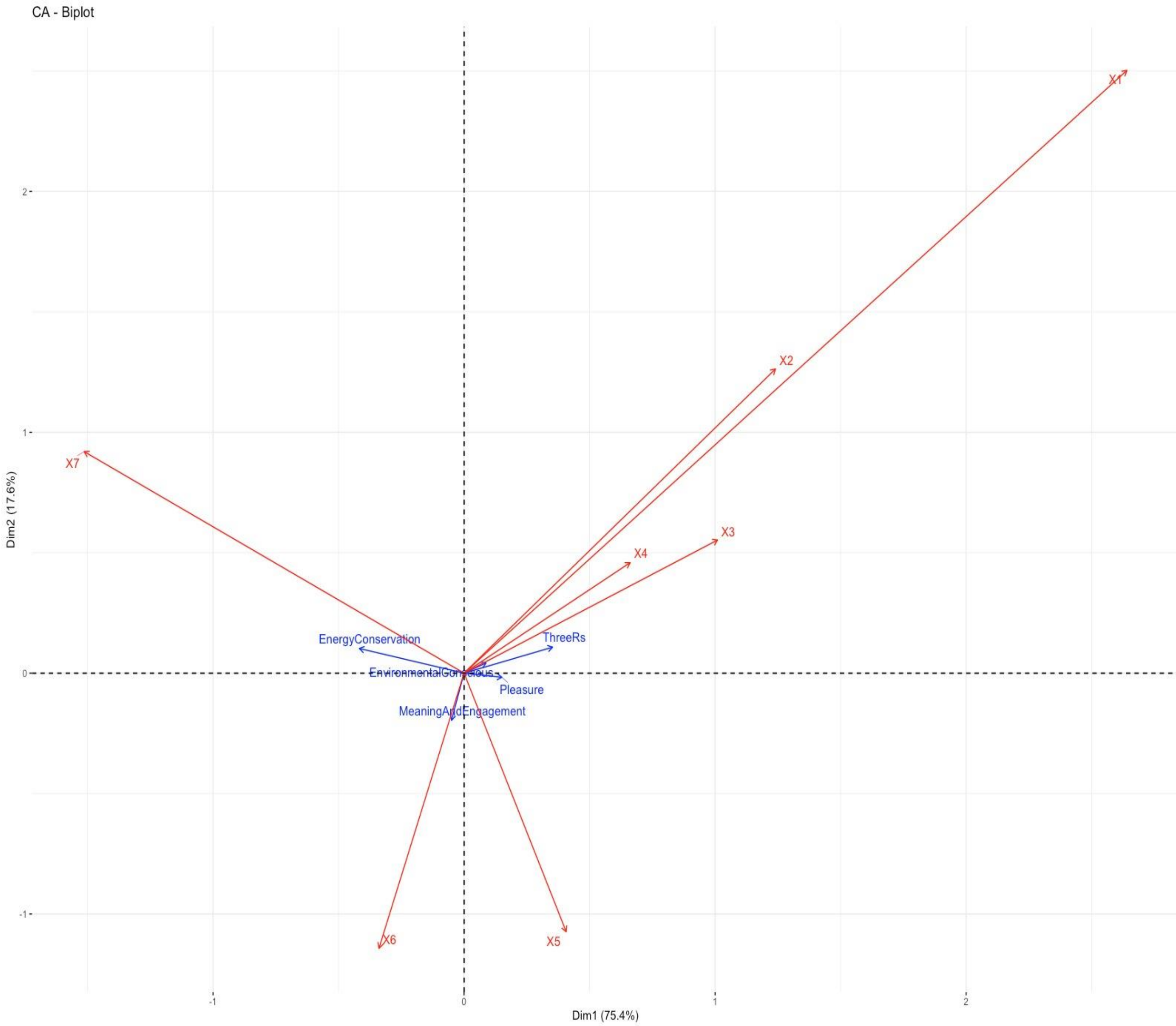


Col$cos2 Plot

From the balloon plot, it's evident that dimension one and two takes up the variance from all the Likert Scale variables, dimension three and four takes up very less inertia.

Biplot

Biplot is a graphical display of rows and columns in 2 dimensions. We have already described how to create CA biplots. To make an asymmetric biplot, rows (or columns) points are plotted from the standard coordinates and the profiles of the columns are plotted from the principal coordinates. When the angle between two arrows is acute, then there is a strong association between the corresponding row and column.



Asymmetric biplot

Biplot helps us to infer the relationship and similarity between the groups and the attributes. Energy conservation, environmental consciousness, and meaning and engagement are very like our audience making them fall into one category of people. Whereas reduce, reuse, and recycle are not very like our audience. Pleasure is in the abnormal direction of Likert scale vectors. This might be because when the questions were translated from Vietnamese to English and back, this set of OTH questions which belong to subset pleasure would have been misinterpreted and the original intent might have changed.

Discussion & Conclusion

The results from our analysis did not produce and replicate the three factor structure as originally intended by the study. When administered in the context of Vietnam, our factor analysis results could differentiate between life of pleasure from life of meaning and life of engagement, but the analysis could not differentiate between life of meaning and life of engagement itself. Although it was not able to extract a distinct structure with three factors, it did however, differentiate between the positive associations and negative associations in the context of sustainable behavior.

A few explanations as to why the three factor structure could not be replicated include 1) the three paths to orientations of happiness might be highly related within itself, 2) there might not necessarily be a single pathway to happiness as suggested by the OTH scale, and lastly 3) the context and original intention of the OTH scale might have lost its effectiveness after being translated, thus, leading to a suboptimal dimensional structure of the scale.

A recommendation we would make is to reduce the Likert scale's range of 1 to 7 to 1 to 5. We believe this would create a more precise picture of the respondent's attitudes towards OTH and SC. In the descriptive analysis part of our data set, we found a lot of the questions' mean were above the neutral point (4) and the proportion of respondents who chose answers above 4 outweighs those that chose below the neutral point. Since this study tries to replicate the same methods used in western society to see where the parallels are, this might not be feasible. Another suggestion we could make is to combine the responses from 1 to 3, thus making the new scale's range from 1 to 5 as well. Where 1, not very like me at all, is a combination of the old scale's 1, 2, and 3. Where 2 is the old neutral point (4), 3 is equal to the old 5, 4 is equal to the old 6, and 5 is equivalent to the old 7. This would provide a more balanced answer and hopefully we could find different insights on Vietnamese's OTH and SC.

Our factor analysis of part 2 of the questionnaire returned only three factors in the context of sustainable consumption. They are 1) environmental consciousness, energy conservation, and the Three Rs (reduce, reuse, and recycle). These factors reflect the multitude of ways in which people in Vietnam get involved with sustainable consumption. They do so by purchasing environmentally friendly products, reducing waste and recycling, conserving water and fuel, and saving energy. The commonality between these factors suggests that the Vietnamese are more inclined towards sustainable consumption in monetary cost saving areas such as saving energy, water and fuel. Purchase of sustainable goods and waste management appear to be slightly less popular among the Vietnamese.

Our findings from regression as well as CA supported the hypotheses about a positive relationship and association between SC specifically with regards to energy conservation and orientations towards *a meaningful and engaged life*. But we were unable to find significance between the negative association between pleasure and SC.