Market Segmentation Project

Veronica Stephens

Objective:

Develop a market segmentation system and supporting materials

Data:

Simmons National Consumer Study (NCS): American family lifestyles, media, and product preferences

Business problem:

Annie's, a socially and environmentally aware company that produces food products is looking to expand brand awareness and adoption for its healthy packaged foods.

Executive Summary:

The variables used in the marketing segmentation system are environmentally conscious, early adopters, volunteer time, read labels, gender, household purchasing power, and watch morning shows (NBC, ABC, CBS). According to the data, consumers whose values most align with an environmentally and socially conscious company that produces healthy food are in cluster 4 (see details below). A marketing strategy from the segmentation system could be advertising geared toward females during ABC and CBS morning shows to sell Annie's food products. These ads should highlight the social, environmental, and health benefits of brand and food choices.

Cluster 4: Environmentally conscious & socially aware
highest environmentally conscious, above average early adopter, highest volunteer time, highest reads labels, lowest number of males, highest number of females, second highest purchasing power (-1.3% below highest), highest watches ABC and CBS morning shows

Programs:

Python (Spyder IDE)
Entire program is in Python, used rpy2 package to run R NbClust package during cluster analysis.

Files:

Program:
exam1_vstephens_181021.py
exam1_vstephens_181021_prgram.pdf

Console output:
exam1_vstephens_181021_output.html
exam1_vstephens_181021_output.pdf

Target population:

U.S. adult population 18 years of age or older (the entire NCS data set)

Variables:

Factor Analysis:

Question set 1 = environmentally conscious (p.132 personal book)
People have a responsibility to use recycled products whenever possible (environRespRecycldProd)
Companies should help consumers become more environmentally responsible (enrironCompHelpConsumEnvironResp)
I am more likely to purchase a product or service from a company that is environmentally friendly (environPurchEnvironFriendlyComp)
I am more likely to buy a product from a company that uses environmentally-friendly methods of advertising (environFriendlyAdPurchase)

Question set 2 = early adopters
I am usually the first of my friends to shop at a new store (firstShopNewStore, p102)
I'm always the first among my friends to have the latest in electronic equipment (firstElecEquip, p.134)
I'm usually the first to try a new health food (firstNewHealthFood, p.135)
I am usually the first among my friends to try new clothing styles (firstNewStyles, p.86)

Cluster Analysis:

Statistical/Cluster Drivers
I am willing to colunteer my time for a good cause (p.134)

I usually read the information on product labels (p.139)

Descriptor/Profile Variables
Sex (p.4 personal book)
RESPONDENT MOST INF IN HH PURCH DECSIONS
CBS THIS MORNING- FREQ OF VIEWING (1 time/week)
GOOD MORNING AMERICA (ABC)-FREQ OF VIEW (1 time/week)
TODAY SHOW (NBC) - FREQUENCY OF VIEWING (1 time/week)

Narrative:

Factor Analysis

Factor analysis was run on eight opinion variables, containing two groups: environmentally conscious and early adopters (see variables, above). The environmentally conscious group contains variables about recycling products, company's environmental responsibility, and likelihood of purchases from environmentally friendly companies. The early adopter group contains variables indicating consumer opinions on being the first to shop at new stores, try new health food and styles, and have new electronic equipment.

All eight variables in the factor analysis have five answer options including: agree a lot, agree a little, neither agree nor disagree, disagree a little, and disagree a lot. Variables were cleaned and combined into a single column (see output files).

Strategy Variable Selection

Step 1
Isolate all opinion questions (agree/disagree) and get the correlations between all variables

Step 2
In order to select groups of highly correlated variables for factor analysis, randomly selected groupings of 5,4,3 variables and measured the Kaiser Meyer Olkin statistic (KMO) for each grouping. Continued until threshold for KMO met (>0.7-0.8).

Step 3
Run principal component analysis on selected variables to see how many factors are within the group of variables. Used eigenvalues and variance explained to interpret which groups were good candidates. After numerous runs, patterns of grouped variables emerged and helped in variable selection.

After two groups of variables were selected, principal component analysis was performed for variable reduction, resulting in orthogonal and uncorrelated factors. The Kaiser Meyer Olkin statistic & Bartlett's Sphericity Test returned results of 0.79 and non-significant respectively, the data is well behaved.

Two factors were extracted with PCA, based on the Kaiser criterion with eigenvalues are greater than or equal to 1. Also, The variance explained by the two factors is greater than 50% at 59%. Extraction communality is also included in the output with all variables have greater than 54% percent of the variance in the variable reproduced by the factors.

The scree plot based on the factor analysis is included in the output files. The plot is interpreted by finding the point at which a rock would stop rolling downhill or at the elbow and subtracting 1. In this case, the scree plot indicates there are two factors.

Performed varimax rotation to discriminate between the factors better. Chose varimax rotation because the factors are more likely to result in well separated clusters. The original and rotated factor loadings are shown in the output. The rotated loadings show clear separation from 0.74 to 0.87. The factor analysis resulted in two extracted factors: environmentally conscious and early adopters. The factor scores were saved prior proceeding to cluster analysis.

Cluster Analysis

In addition to the two factors created in factor analysis, two statistical or cluster driver variables were added (see variables above). The driver variables capture a consumer's wiliness to volunteer for a good cause and likelihood of reading information on labels.

Strategy Variable Selection

Step 1
Isolate all opinion questions (agree/disagree) and get the correlations between all variables and the two factors.

Step 2
In order to get distinct clusters, ranked opinion variables by correlation to two factors and created a list of possible drivers and corresponding correlation measure.

Step 3
Added the drivers to and ran an initial cluster analysis examining the cluster means, CCC, Pseudo T squared, and Gap statistics to assess possible variables.

The two driver variables in the cluster analysis have five answer options including: agree a lot, agree a little, neither agree nor disagree, disagree a little, and disagree a lot. Variables were cleaned and combined into a single column (see output files). No standardization was used.

Performed K-means clustering with two factors and driver variables, setting the range of clusters from 2 to 10. The cubic clustering criteria (CCC), Pseudo T squared and gap statistics are plotted in the output. The optimal number of clusters for each statistic is 4, 2,4 respectively. For CCC and gap the first local maximum is the ideal number of clusters, for Pseudo T squared it is one cluster less than the first local maximum.

The cluster analysis was run with 4 clusters and the corresponding cluster assignments were added to the data frame containing the two factors and driver variables. The output shows the table of means for the factors and driver variables. While there is some separate among cluster means, ideally these differences would be greater.

Profile Variables

In addition to the two factors and driver variables, five profile variables were added (see variables above). The profile variables capture a consumer's gender, their purchasing power within their home, and if they have watched the CBC, ABC, or NBC morning show once per week. The five driver variables are binary. Variables were cleaned and combined into a single column (see output files). No standardization was used.

The cluster analysis was run with 4 clusters and the additional profile variables. The output shows the table of means for the factors, driver, and profile variables. The means for the factors (environmentally conscious, early adopter) are z scores and are converted into probabilities below. The means are a proportion of responses for the binary profile variables.

Cluster Summaries

Cluster 1: lowest environmentally conscious at 0.02, 0.49 early adopter, slightly above neutral to volunteer time, lowest score for read labels below neutral toward disagree, highest number of male consumers at 72%, second lowest purchasing power at 57%, lowest rate watches NBC morning show at 1.8%, lowest watches ABC morning show at 1.9%, lowest watches CBS morning show at 1%

Cluster 2: mid environmentally conscious at 0.49, 0.92 early adopter, lowest score slightly above neutral to volunteer time, slightly below neutral to read labels, male consumers at 66.7%, lowest purchasing power at 55%, highest rate watches NBC morning show at 2.3%

Cluster 3: mid environmentally conscious at 0.54, lowest early adopter at 0.06, between neutral and agree to volunteer time, slightly above neutral to reading labels toward agree, male consumers at 67%, highest purchasing power at 62%

Cluster 4: highest environmentally conscious 0.99, 0.65 early adopter, most likely to volunteer time, most likely to read labels, lowest number of male consumers at 65%, second highest purchasing power at 60.7%, highest rate watches ABC morning show at 3.7%, highest watches CBS morning show at 2%

Cluster Names

Cluster 1: Mid Early Adopting Males
lowest environmentally conscious, mid early adopter, lowest read labels, highest number of males, lowest number of females, lowest NBC, ABC, CBS morning shows

Cluster 2: Early Adopters & mid environmentally conscious
highest early adopter, mid environmentally conscious, lowest volunteer time, lowest purchasing power, highest watches NBC morning show

Cluster 3: Mid environmentally conscious with high purchasing power
lowest early adopter, mid environmentally conscious, highest purchasing power

Cluster 4: Environmentally conscious & socially aware
highest environmentally conscious, above average early adopter, highest volunteer time, lowest number of males, highest number of females, second highest purchasing power (-1.3% below highest), highest watches ABC and CBS morning shows

Additional Commentary

Had significant difficulty adding drivers that produced clear difference in cluster means or best cluster number with CCC, Pseudo T squared, and gap analysis.

For CBS, NBC, ABC morning shows could combine the watched once per week variable with additional weekly frequency variables to obtain consumers who watched a show at least once per week.

Could gather and order binary profile variable by response rate in order to choose better profile variables.