Israel Nolazco

israelnolazco@lewis.edu

Overview of Cluster Analysis

August 29, 2020

Table of Contents

# Background.

# Executive Summary.

# Analysis

* 1. Selecting the number of Clusters
  2. Profiling the Cluster Variables
  3. Assessing Cluster Solution

## Summary

1. Background.

The purpose of this project is to perform various techniques of Cluster Analysis. Cluster Analysis is a unique technique utilized in many industries whose goal is to group objects based on the characteristics they possess. In short Cluster Analysis classifies objects(e.g., respondents, products, or other entities), on a set of user-selected characteristics. If the classification is successful the object within the clusters will be close together when plotted geometrically, and the different clusters will far apart. In this instance, we will use the HBAT database to illustrate the applications of cluster analysis in hopes to understand how customers rate the firm’s performance.

The variable son the HBAT database are presented in the table below:

* X6 Product Quality
* X8 Technical Support
* X12 Sales Force
* X15 New Product Development
* X18 Delivery Speed.

1. Executive Summary

The analysis will look at the number of clusters needed to have a harmonious representation of our data. Then proceeding with hierarchical cluster analysis and finally assessing the cluster solutions and their validation. The next section provides a high-level synopsis of details that are presented in the remainder of this document. Unfortunately, given the technological imparities, we are only going to analyze the data results provided from the book Multivariate Data Analysis than obtaining the data from scratch.This document presents a Cluster Analysis that were obtained from the HBAT dataset. This measures the firm’s performance based on customer feedback.

1. Analysis
   1. Selecting number of clusters

**A close up of text on a white surface

Description automatically generated**

**Figure1.** Percent Changes in Heterogeneity

An analyst must recognize that and have a sort of intuition as far as how many clusters should be enough to represent the data. The dataset HBAT is composed of customers and our interest is in identifying customer types or profiles that may form the basis for different strategies. Therefore, it clear somewhere between two and four clusters should Sufi however, anything higher than six or seven will create too many unique clusters that will not be useful to interpret in an enterprise solution. In this instance, however, the analysis will look at the changes of Heterogeneity between clusters. If there is a large increase of heterogeneity than the analyst selects the prior cluster solution because the new combination is joining quite a different cluster. Additionally, the agglomeration coefficient should be small, since this indicates a homogenous cluster are being merged. Figure 1. Shows the proportionate increase in heterogeneity through the stages of clusters. Keep in mind that these increases are also determined by the coefficients. Therefore, we have come to the conclusion that a four-cluster solution to be the most stable and efficient moving forward with our analysis.

* 1. Profiling the Clustering Variables

A screenshot of a cell phone

Description automatically generated

Figure 2. Profile of four Cluster from hierarchical Cluster analysis.

Let us first look at our F statistic from one-way ANOVA that examines whether that are statistically significant differences between the four clusters on each of the first cluster five cluster variables. As seen in figure 2. The significant F statistics provide initial evidence that each of the four clusters is distinctive. As for the means for the five cluster variables, the results indicate that each of the four-cluster exhibits somewhat distinctive characteristics. Additionally, no cluster contains less than 10% of the observations. Therefore, the clusters are clearly a great representative of our data.

* 1. Assessing Cluster Solution

A screenshot of a social media post

Description automatically generated

Figure 3.Multivariate F Results Assessing Cluster Solution Criterion Validity

It is without a doubt there must be a validity to our cluster analysis, it is then decided to focus on variables that have theoretically based relationship to the clustering variables but were not included in the cluster solution. Given this relationship, we should see significant difference in these variables across the clusters. In this instance we are utilizing measure from the HBAT dataset:

* X19 Satisfaction
* X20 Likelihood to Recommend
* X21 Likelihood to Purchase
* X22 Purchase Level

Figure 3. shows a MANOVA model in a significant F value at 2.23 and P value at 0.01, this provides initial support for the idea that these variables can be predicted by knowing which segment an HBAT customer belongs. The individual univariate F-statistics are also significant further verifying this finding.

Summary

The results of our cluster analysis show it can predict other key outcomes that provide evidence of criterion validity. As mentioned on page 536 of Multivariate Data Analysis, an example of cluster 2 which HBAT believed was receptive to more business based on its cluster profile displays the highest score on each of these key outcome variables. Thus, HBAT will likely find the cluster solution useful in predicting other key outcomes and forming appropriate strategies. In short Cluster Analysis is key in finding a group within a dataset. In this case, the HBAT dataset provided clusters that were useful to understand customer feedback and such results can then be used to fix relationships, push products, or tap into new markets. However, it is also clear, there must be a validation of the cluster and a deep understanding of the data gathered is needed to fully make use of Cluster Analysis.