Derek Hughes
PREDICT 450 – Sec 55
Solo #1
Summer 2015

INTRO

Transitioning into a new market category can be a challenge for any business. App Happy, a company that traditionally designs for B2B applications, is interested in developing social entertainment apps for the general public. To improve their understanding of the new market, they created a survey questionnaire that included focus groups and individual interviews. We will use this qualitative research data to create a general attitudinal post hoc segmentation analysis so the company may design an effective marketing campaign.

A general attitudinal post hoc segmentation analysis requires a post-hoc descriptive approach. Here segments are identified by creating groups of consumers that are homogeneous along a set of measured characteristics (Kamakura & Webel, 2000). Furthermore, we will need to use descriptive instead of predictive statistical methods, as we want to find patterns in the data as a whole instead of based on a defined response variable. Finally, post hoc means we will determine the marketing segments after we have the results from the survey.

Classification methods for post hoc segmentation analysis include clustering methods that are non-overlapping, overlapping, Fuzzy techniques, ANN, and mixture models. We will use non-overlapping clustering methods for this segmentation analysis so each customer will be defined into a single segment. Non-overlapping clustering methods can be hierarchical and nonhierarchical. The difference is that hierarchical methods begin with a single cluster and link clusters in successive stages, while nonhierarchical methods begin with a predetermined number of clusters and iteratively assign and reassign members into clusters until the method is complete. Simply put, members assigned in a cluster in a hierarchical method will stay in that cluster until the method is complete, but members in a nonhierarchical method can be reassigned to clusters until the method is complete. Our analysis will use at least one method from each type of non-overlapping clustering method.

DATA EXPLORATION

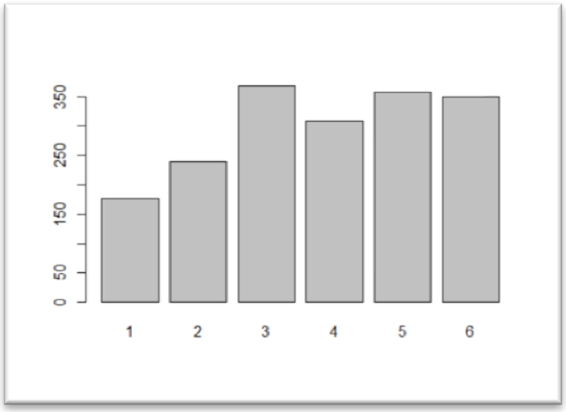
The survey data, as a whole, has a few different types of response data. One version is in the form of character strings and another is a numeric version which uses numbers to represent specific responses. Furthermore, in the numeric form, the data is ordered, non-ordered, and binary for different responses. In addition, the data has 1800 observations and comprises 89 different variables.

The analysis calls for attitudinal variables. By researching the survey questions and responses we identified three questions, questions 24, 25, and 26 that comprise the attitudinal variables of the consumers taking the survey. Each questions has approximately 12 “sub-questions” that require a numeric response that represents attitudinal beliefs ranging from strongly agree to strongly disagree. Since these are ordered responses, the results can be averaged and be used to extract relevant information.

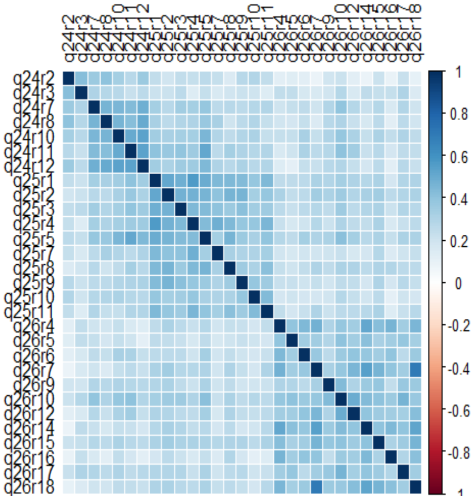
VARIABLE REDUCTION / SELECTION

Our goal is to use relevant attitudinal variables to create market segment clusters to use for deriving effective marketing schemes. We can do this by checking for natural clusters within each variable, checking correlations between variables, and using our subject expertise to further group relevant variables together.

The second step to selecting the right number and type of variables is to look for natural clustering of the variable based on the consumer responses for that variable. Specifically, we use the barplot() function to show the frequencies of the responses for each variable. If the frequency plot shows highly skewed distribution, that indicates a natural clustering for that variable and it will be useful for our segmentation analysis. An equally distributed frequency of responses is not very valuable for our purposes, as it doesn’t identify any significant differences, and therefore doesn’t show any distinct opportunities in which to take advantage of any attitudinal beliefs. After reviewing each attitudinal variable, we removed seven variables that did not indicate any natural clustering.



Next, we used a Pearson correlation plot to analyze the correlation between the remaining variables. We used a colored correlation plot to make it easier to identify correlations. Quantitatively, we used an arbitrarily selected cutoff value of .40 to determine if a variable was significant and relative to our analysis. Thus, if a variable did not show a correlation equal to or greater than .40 with any other variable, it was removed. After removing all non-relevant



variables, we ran the correlation iteration again to check and remove any variables that were not significant with the new grouping of variables.

Our final variable selection/reduction process uses our subject expertise to try to find ways to group the remaining variables. We looked at each question and compared it to every other question for similarities. If we found a potential grouping of questions/variables, we checked the correlations between each variable to be grouped together. Again using our cutoff value, if all variables being considered for a group are correlated at or above .40 then we combined them into a group. This approach combines a subjective approach (subject expertise) with an objective statistic to confirm our assumptions.

We created six groups with this method: internet use for friends, leaders, risk takers, luxury shoppers, app lovers, and shopping lovers. To create the groups, we simply added the response values from each variable together and took the average. Thus, the new group response values are an average of the values from every variable included in the group.

CLUSTERING METHODS

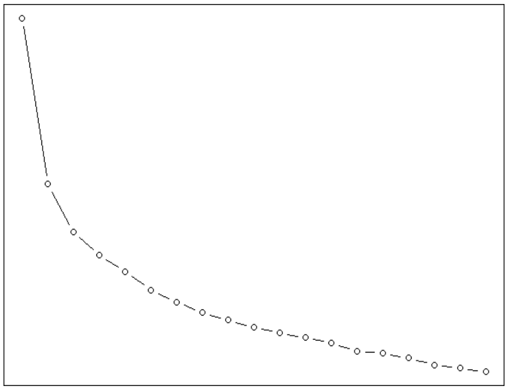
Now that we have our variables selected, we will use three clustering methods to discover marketing segments to be exploited. One will be hierarchical and two will be nonhierarchical (K-means, PAM). Nonhierarchical methods are often considered to be superior to hierarchical methods because they can handle outliers and irrelevant variables better. However, nonhierarchical methods generally require that the number of clusters be determined beforehand and this can cause issues with performance if not selected carefully.

K-means is one of the most commonly used nonhierarchical clustering methods used for marketing segmentation. It is a non-overlapping clustering method which means each customer will be placed into only one cluster. It is an extremely efficient algorithm that uses a predetermined number of clusters, computes their centroids, and aims to reduce the sum of squares within the clusters. The process reassigns observations at each iteration and concludes when no more observations are reassigned. Some assumptions of K-means are the number of clusters, that SSE is the correct method to minimize, all clusters have the same SSE, and that all variables have the same importance. The advantages of K-means are that it is computationally faster and may produce tighter clusters than hierarchical methods. That disadvantages are that it suffers with non-spherical clusters, difficult to predict what number of clusters should be, difficult to compare quality of clusters produced because different initial centroids can produce different final clusters.

PAM is similar to K-means only that it uses the median instead of the means for calculations (mediods instead of centroids). It also calculates using a dissimilarity-based distance instead of a squared-Euclidean distance. The advantage of the PAM method is that it’s more robust to outliers, missing values, and other data anomalies. Its disadvantage is that it works well on small data sets but is not really efficient enough for large data sets.

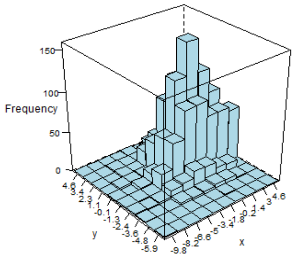
Hierarchical clustering builds a cluster hierarchy that looks similar to a tree called a dendrogram. It allows us to evaluate data on different levels of granularity and is made up of agglomerative and divisive approaches. Agglomerative being a bottom-up approach that starts with one point and recursively merges with other points over time, and divisive being a top-down approach that starts with one cluster and divides it over time. The advantages are the ability to see detailed levels of granularity, easily handles many types of similarity or distance, and its applicability to any attribute type. Disadvantages are that its termination criteria can be vague and most hierarchical algorithms do not revisit intermediate clusters with the intent to improve them.

DETERMINING AMOUNT AND QUALITY OF CLUSTER SEGMENTS

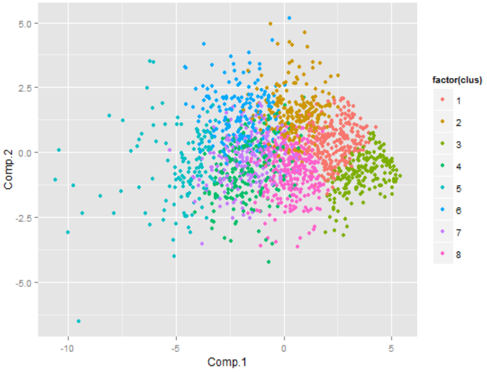


Because K-means requires a pre-selected number of clusters to begin the method, we used a Scree plot to determine the approximate amount of clusters to create in K-means.

We used all nineteen variables remaining after our variable reduction and selection methods to begin the Scree plot. Generally, using between 10-20 variables is considered appropriate, as too few and we can't trust the results as much and too many makes it too difficult to explain the results. Based on the Scree plot, we decided to use eight clusters. This is because eight clusters is where the rate of reduction flattens- less reduction of the within groups sum of squares is obtained with each additional cluster after that point.



It isn't enough to simply select the number of clusters, as we need to determine if the data supports that amount of clusters because we can often simply partition the data (instead of finding real clusters) when natural clusters do not really exist. We use a PCA method to check for at least eight clusters in the data. The 3D graph shows that there are at least



eight clusters in the data so our K-means and PAM methods should certainly find them. We can also determine the cluster quality with a silhouette plot of our cluster results. It used to determine the frequency and average distance from the cluster centroid to determine the cluster quality. Values closer to 1 indicate good clusters and value closer to 0 indicate bad clusters. For our K-means results, one cluster was at a .41 level while the remaining were .15 or less. This indicates that one cluster was of average quality while the remaining were not very good clusters. With more time for trial and error with the variable selection, removal, and grouping process would improve these values.

OUTPUT / RESULTS

After determining the cluster amount to use for K-means and PAM (non-hierarchical methods), we determine the goodness of fit evaluation criteria by determining the sum of squares between each cluster. By dividing the between clusters sum of squares by the total sum of squares we can achieve a ratio that estimates how well our clusters fit the data. Ideally we'd like to have values of .60 or higher for any model. Fortunately with this ratio we can compare how well each method fits the data, but that isn't everything when deciding which method to use. We must consider the clusters created by each method and how well they fit our situation based on our subject expertise.

Between Sum of Squares/Total Sum of Squares = goodness of fit value
K-means = .4038   PAM = .3208   Hierarchical = .2832

We selected the K-means as our "best" cluster because the goodness of fit value was better and the clusters were as relevant as the others. It should be noted that none of these goodness of fit values were particularly good. With more time for trial and error with variable selection, removal, and grouping these values would improve.

PROFILING / RECOMMENDATIONS

We use the cluster centroids to create profiles and now include the demographics and behavior variables to create a cluster "profile." From these clusters we can create a marketing strategy for the type of apps that might be appealing to each cluster. A cluster received a profile description for a variable if their average response agreed or strongly agreed with the variable.

Cluster 1- tech savvy, enjoy new apps, music important, learns more about TV shows, creative, optimistic, uses Internet for family/friends interaction, leadership traits, age 30-34, college degree, single, most with kids, kids mainly 6 yrs. old or older, non-latino, 60-70K income, majority women.

Recommended: A new app that allows them to entertain their young children with creative or educational games may be appealing. These are leaders that are tech savvy and enjoy new apps, but are not particularly prone to risky behavior so the app shouldn't be too much out of the mainstream in terms of why, where, and how apps are used.

Cluster 2- strongly tech savvy, enjoy new apps, music important, learns more about TV shows, creative, optimistic, active, likes cool apps, considers phone entertainment, uses Internet for family/friends interaction, leadership traits, risk taker, 30-34 age, college degree, single, with kids mostly under 6, 60-70K income, majority women

Recommend: An app that allows these educated, active, risk taking, younger family members to enjoy entertainment while tending to their young kids. Maybe something related to facilitating an active lifestyle as a young parent.

Cluster 3 – This cluster showed average attitudinal response variables for all variables. It include 30-34 year olds with some college, mainly single, most without kids but most under 12 if with kids, 50-60K income level, and majority women

Recommend: More research is needed for this cluster because attitudinal responses were average for every variable making it difficult to identify a niche or clear opportunity. Maybe an app that is simple to use and is relatively neutral in terms of creativity and free. Frankly we'd probably focus on other clusters first because expendable income and education is low and motivational, risk taking, and active behaviors are average at best.

Cluster 4 – strongly tech savvy, enjoy new apps, music important, enjoy learning more about TV programs, likes cool apps, uses internet for friends/family, 40-44 year old, college degree, single, most without kids but if have kids most either below 6 or above 18, 60-70K income, majority women

Recommend: this app might be fairly complex, trendy, and possibly a bit of a novelty app for music lovers for generation Xers.

Cluster 5 – tech savvy, does NOT enjoy new apps, music is important, enjoys learning more about TV programs, creative, optimistic, active, likes cool apps, uses internet for friends/family, leadership traits, 45-49 age, college degree, most single without kids, if with kids most over 18, 80-90K income, most women

Recommend: Only cluster that does not enjoy new apps so recommendation would be to create an app that improves upon an already existing app that is popular with this age group. This app could require a fee as well because this is the highest income cluster, most don't have kids, or their kids are already over 18 so expendable income is more available..

Cluster 6 – This cluster is almost identical to cluster 3 except age is higher (40-44) and those with kids are over 18 instead of under 12.

Recommend: Again, motivation, education, income, and active behaviors are average at best here, so we recommend focus on other clusters first or to do more research on general interests of 40-44 year olds, as there are not many unique attitudes or behaviors here.

Cluster 7 – This cluster was unique because it strongly agreed with every variables except only just agreed with liking packaged deals, liking online shopping, shopping in general, cool apps, luxury shopping, and showing/sharing new apps. Age range is 30-34, college educated, mostly single but if kids most below 6, and income is between 60 and 70K.

Recommend: This should be the first cluster to focus on because it agrees or strongly agrees with every variable. This allows for a creative, new app by the designers without many restrictions. Furthermore, this group enjoys shopping, shopping for high end items, wants to stand out, active, risk taking, educated, and loves new apps, showing them to others, and think it's worth purchasing them. A fee based app that improves their ability to shop for luxury or designer items or to interact with family and friends may be worth considering. Additionally, the developers could take more risks to create a unique app here because this cluster likes to be leaders, take risks, loves showing others their cool apps, and think it's worth paying for apps.

Cluster 8 – This cluster is almost identical to cluster 7 except they only agree (not strongly agree) with every variable. Another profile difference is that they only had an average response for the active behavior variable. Uniquely, this was the only cluster that was majority men (although most were close to evenly distributed, this was the only one with more men than women).

Recommend: Take a similar approach to that of cluster 7 except tailor it more towards men and temper down the amount of uniqueness of the app compared to cluster 7. It may even be worth reworking the analysis again to find a way to combine cluster 7 and 8 into one cluster because they are so similar.

It should be noted that these results are limited by the survey questions asked, the manner in which the responses were provided to the respondents, the integrity with which the respondents responded to the answers. It may not represent the entire market, as only those available for the focus group (assuming focus group was selected appropriately) and to volunteer for interviews were represented. This may indicate only a sample of enthusiastic consumers and not the market as a whole. Finally, a survey is only a snapshot in time and may or may not be relevant in the future.

TYPING TOOL

Now that we have our clusters, we need to define a way in which to classify new customers into a segment. It isn't feasible to provide the survey to each new consumer; however, since the demographics are unique for each cluster, we can use basic demographic variables to estimate which cluster a new consumer would most likely occupy. There are two relatively easy to implement methods to use for classifying new consumers with a typing tool: discriminant analysis and CART.

For discriminant analysis, we could find the Fisher coefficients for each variable because this is a standard output option for most statistical software. We would then create a series of linear equations, one for each segment. By multiplying the respondents results for each demographic by the coefficients for each segment, we could classify the customer into the segment which produces the highest score. If we have all the demographic information for the new customer, this could be easy to implement.

The second consideration is a CART method using the classification trees. It takes a bit longer to develop and is more computationally intensive, but classification trees can be very useful as typing tools. Here we would ask a simple “yes” or “no” question for each demographic variable of the new customer. Based on the responses, the consumer would travel through the branches until he/she reaches an ending node that indicates a particular segment/cluster. CART works in this analysis because each cluster is composed of a unique set of demographic variables; eventually the consumer will be fitted into a segment. This approach is very intuitive for the business user and, as such, provides another easily implementable solution as a typing tool.

REFERENCES

Kamakura, W.A. & Wedel, M. (2000). *Market segmentation: Conceptual and methodological foundations.* Boston, MA: Kluwer Academic Publishers.