App Happy Market Segmentation Analysis

Yining Feng

January 23, 2021

Executive Summary

The App Happy Company wants to better understand what they believe is the market for a new social entertainment app they are thinking of developing. A general attitudinal post hoc segmentation analysis is thus devised to develop and evaluate a segmentation scheme for App Happy based on customer survey data. The data set consists of 1,800 customers' Apps usage motives, demographic variables, and varying attitudes of customers towards technology.

1. Technology Adoption Trend

Smartphone ownership across all ages

Smartphone ownership among survey respondents varies substantially by age: 63% of 55- to 59-year-olds own iPhone, but that share falls to 34% among teenagers under 18-year-olds. 44% of 18- to 24-year-olds own Android, but that share falls to 14% among seniors who are over 65-year-olds. Smartphone adoption drops off considerably among adults in their mid-60s and beyond. In addition, the utilization rate of Entertainment Apps plummets down to 27% among seniors over 65-year-olds in contrast to 54% among teenagers under 18-year-olds. Nevertheless, Social Networking Apps are consistently popular among all age groups with the highest utilization rate among the majority of customers, with 87% of 40- to 44-year-olds and 59% of those ages 65 and older who are active users.

Smartphone ownership is also highly correlated with household income and educational attainment. Nearly 56% of customers whose annual household income is $50,000 or more say they own iPhones, compared with 38% of those living in households earning less than $30,000 a year. Additionally, more than a half of those with bachelor's or advanced degrees report owning iPhones (57%), compared with 45% of those who have some college experience and 37% of those who have high school diplomas or less.

Demographics of Social Media Users

Social media is increasingly becoming an important platform where users find news and information, share their experiences, and connect with friends and family. Social media use is

also relatively common among those who are at least high school graduate (79%) and those whose annual household income is $50,000 or more (76%).

Out of all social media sites surveyed, social networking site Facebook is particular popular across all ages. Based on survey responses, 64% of customers ages 65 and up say they use social networking site Facebook very often. On the contrary, only 5% of older adults reported using social media Twitter very often. Still, a majority of seniors do not use social media very often, and the share that do is considerably smaller than that of overall customer population.

As with other forms of social media, teenagers under 18-year-olds are more likely than middle-aged customers (between ages 45 and 60) to use social media site YouTube. More than six-in-ten (66%) teenagers under the age of 18 say they use YouTube very often, compared with 36% of middle-aged customers. YouTube only remains appealing to the majority of customers under the age of 40 (62%), whereas just 18% of seniors say that they use online streaming site Netflix very often. But 77% of seniors ages 65 and older say they are always checking on friends and family through Facebook or other networking websites.

Demographic Factors in Apps Adoption

Social networking apps are the most popular type of apps among those ages 65 and up – albeit the share of seniors who subscribe to social networking apps (59%) is still lower than the share of subscribers in any other age group. Besides social networking apps, general news apps are equally popular among seniors. Another type of popular apps among seniors is music and sound identification apps, with more than half of seniors (55%) as subscribers.

Apps adoption among other users varies substantially across a number of demographic factors – most notably age, household income, and educational attainment. While music and sound identification apps are the most commonly used type of apps among teenagers under 18-year-olds, specific publication news apps are the type of apps they are least interested in. Nevertheless, social networking apps are the most used type of apps among middle-aged customers (75%) and the overall surveyed population (81%), even though the majority of the overall survey population (71%) say that they have more than 11 apps on their smartphones

/iPods Touch/Tablets. Accordingly, social networking apps are unanimously the most favored type of apps across all ages.

Apps adoption rates also differ considerably by household income and educational attainment. More than half of customers whose annual household income is $50,000 or more say they use general news apps (58%) or shopping apps (51%). Those shares drop to 38% for general news apps adoption and 37% for shopping apps adoption, respectively, among customers living in households earning less than $30,000 a year. Customers with post graduate degree are far more likely than those with high school educations or less to say they use specific publication news apps (53% vs. 20%) or shopping apps (61% vs. 33%).

2. General Attitudinal Analysis

Most customers surveyed have a positive outlook about technology and the benefits it can provide. These customers think that technology products have improved and benefited them in many ways; 66% of those customers say that they enjoy using technology to give them more control over their lives. They are particularly interested in technologies that can enable them to enjoy music in everyday life (89%), keep up with TV shows (92%), connect with friends and family (90%), and look for web tools and Apps that help them save time (93%).

The survey data also imply that these “Tech Fans” are also very likely to consider themselves as opinion leaders, or they are likely the first of their friends and family to try new things, or they like to offer advice to others. The correlation plot shown in Figure 1 further suggests that there is a very strong positive correlation (0.63) between those customers’ positive attitudes towards technology and their extraverted personality (decisive, adventurous, solicitous). In addition, these customers also tend to spend money on a variety of things such as new Apps and extra App features, as evidenced by the strong positive correlation (0.56) between their extraverted personality and their highly motivated purchase behavior (impulse purchase or trend-chasing behavior). These customers are characterized by their shared enjoyment derived from shopping. They are substantially more likely (73%) to say that they are influenced by what is hot and what is not: 55% of those shoppers say that they prefer to buy designer brands, and 53% of those shoppers say that they are often attracted to luxury brands.

Overall, more than seven-in-ten (72%) of all customers surveyed indicate that they try to keep up with technological developments and they enjoy purchasing new gadgets and appliances. Notwithstanding, 52% of customers say that they are overwhelmed by technology and they think there is too much information out there today from the Internet and sites like Facebook.

3. Market Segmentation Analysis

The main objective of market segmentation is to divide customers into homogeneous groups which have similar characteristics such as purchasing behavior, attitudes towards technology, apps preferences etc. A scree plot (Figure 2) is used to determine the number of factors to retain in an exploratory factor analysis (FA) or principal components to keep in a principal component analysis (PCA). Based on the scree plot, I decided to specify five clusters for each clustering method. The results of five clustering methods applied are given in Table 1. It is evident that of the agglomerative hierarchical clustering methods, average-linkage (AL) puts almost all the observations into a single cluster, whereas single-linkage (SL) and complete-linkage (CL) are somewhat better at distributing the observations among the five clusters. (Figure 3) On the other hand, the K-means clustering, which is illustrated by the scatterplot shown in Figure 4 and biplot shown in Figure 5, demonstrates better results, as evidenced by the value of $\frac{between\_SS}{total\_SS}$ as 78.6%. Nevertheless, pam is closest to the true configuration of the data. The pam silhouette plot for five clusters is given in Figure 6 and the average silhouette width is 0.25.

Cluster	Single Linkage	Average Linkage	Complete Linkage	K-Means	pam
1	769	1,204	1,213	502	471
2	547	539	223	403	465
3	301	53	180	377	376
4	158	2	150	315	256

5	25	2	34	203	232

Table 1. *Comparison of results of different clustering algorithms applied to the Landsat image data. The data consist of six groups of 1,800 observations measured on 91 variables. Prior to clustering, all variables were standardized. The five derived clusters are designated 1–5. The agglomerative hierarchical clustering methods are single-linkage (SL), average-linkage (AL), and complete-linkage (CL), and the nonhierarchical methods are K-Means and partitioning around medoids (pam). Each column in this table gives the cluster sizes distributed among the five clusters, ordered from largest cluster-1 to smallest cluster-5.*

For the given market segmentation problem, 5 clusters are analyzed for various consideration. Cluster-1 represents 35- to 39-year-olds single male customers with incomes between $50,000 and $60,000. They are typically Caucasian or African American college graduates who sometimes visit social media websites such as Facebook, YouTube, and Netflix. These customers also show some interests in gaming, social networking, general news, music and sound identification apps, but 76% —99% of existing apps on their iPhones are free and they are not enthusiastic shoppers.

Cluster-2 represents 30- to 34-year-olds single male customers with incomes between $60,000 and $70,000. They are typically Caucasian or African American with some college experience who visit social media websites such as Facebook and YouTube very often. Sometimes, they also visit other websites including Twitter, MySpace, Pandora Radio, Vevo, IMDB, LinkedIn, Netflix, Yahoo Entertainment and Music. These customers are active Internet users and they also show some interests in entertainment, TV Show, gaming, social networking, general news, shopping, music and sound identification apps. Although 51% —75% of existing apps on their iPhones are free, they can be enthusiastic shoppers.

Cluster-3 represents 30- to 34-year-olds single female customers with incomes between $60,000 and $70,000. They are typically Caucasian or African American college graduates who visit social media websites such as Facebook and YouTube very often. Sometimes, they also visit other websites including Twitter, Pandora Radio, IMDB, and Netflix. These customers also show some interests in entertainment, gaming, social networking, general news, shopping, music

and sound identification apps. Although 51% —75% of existing apps on their iPhones are free, they can be enthusiastic shoppers.

Cluster-4 represents 25- to 29-year-olds single female customers with incomes between $60,000 and $70,000. They are typically Caucasian or African American with some college experience who visit social media websites such as Facebook and YouTube very often. Sometimes, they also visit other websites including Twitter, Pandora Radio, IMDB, Netflix, Yahoo Entertainment and Music. These customers also show interests in various types of apps including entertainment, TV Show, gaming, social networking, general news, specific publication news, shopping, music and sound identification apps. Although 51% —75% of existing apps on their iPhones are free, they are "Tech Fans" and therefore they can be enthusiastic shoppers.

Cluster-5 represents 35- to 39-year-olds single female customers with incomes between $50,000 and $60,000. They are typically Caucasian or African American college graduates who sometimes visit social media websites such as Facebook and YouTube. These customers also show some interests in gaming, social networking, music and sound identification apps, but 76% —99% of existing apps on their iPhones are free. They do not show much interest in technology products and they are not enthusiastic shoppers.

Since almost all customers show high interests in social networking apps, I would recommend App Happy developing a new social entertainment app like Facebook that can satisfy customers' needs. Nevertheless, this recommendation can be subject to under-coverage bias since 67% of survey respondents are under 40 years old. Members of the middle-aged (between ages 45 and 60) or senior population (ages 65 and up) can be inadequately represented in the sample. Moreover, the cluster analysis based on survey data might not be representative of those unable to participate in the survey. Therefore, the cluster analysis results can be subject to nonresponse bias and overrepresent respondents' opinions and attitudes.

4. Classification Model

If the basis variable data is not available for App Happy's new customers, I would suggest researching what data was available for new customers (named "available new customer data")

— this dataset could include variables such as age, gender, income, and geographic location like State or Zip Code. For the customers that were in the segmentation study (for which the basis variables were available), I would try creating a series of logistic regression models, one for each segment. The advantage of the logistic regression model is that it cannot only predict the probability of each customer segment but also estimate the marginal effect of each explanatory variable. For example, if App Happy estimate that there could be five target customer segments in the potential market (Segments A, B, C, D, and E), I would suggest creating a logistic regression model for each one of those segments.

Since this study focuses on new customer segmentation analysis, the outcome of a prediction is assumed to be dichotomous: target segment and non-target segment. A binary logistic regression model is used for each cluster of the regression and classification tree to estimate the likelihood of each outcome. The target variable can be set as: y=1 for target segment, while y=0 for non-target segment. The coefficient of the explanatory variable and Odds Ratio (OR) can be used to explain the relationship of explanatory variables (e.g., gender, annual household income, and apps preference) and target variable (target customer segment), which are estimated by the maximum likelihood estimation method. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure. For example, the explanatory variable "Gender (1=Male, and 2=Female)" has a coefficient of 1.41 and OR of 4.1, which suggests that the probability of a male customer resulting in the target segment 3.1 times higher than a female customer.

The first logistic regression model would be employed to predict the likelihood that a consumer was in segment A (using only available new customer data as predictors), the second model would be deployed to predict the likelihood that the consumer was in segment B (using only available new customer data as predictors), and so on. In this way, each consumer would be run through the five logistic regression models, and the predicted segment for each consumer would be based on whichever model one scored the highest when given that consumer's inputs. Before launching this typing tool, I would recommend ensuring that this typing tool had an acceptable rate of accuracy by calculating how often it correctly predicted the segment for the customers who were in the original segmentation analysis.

Appendix

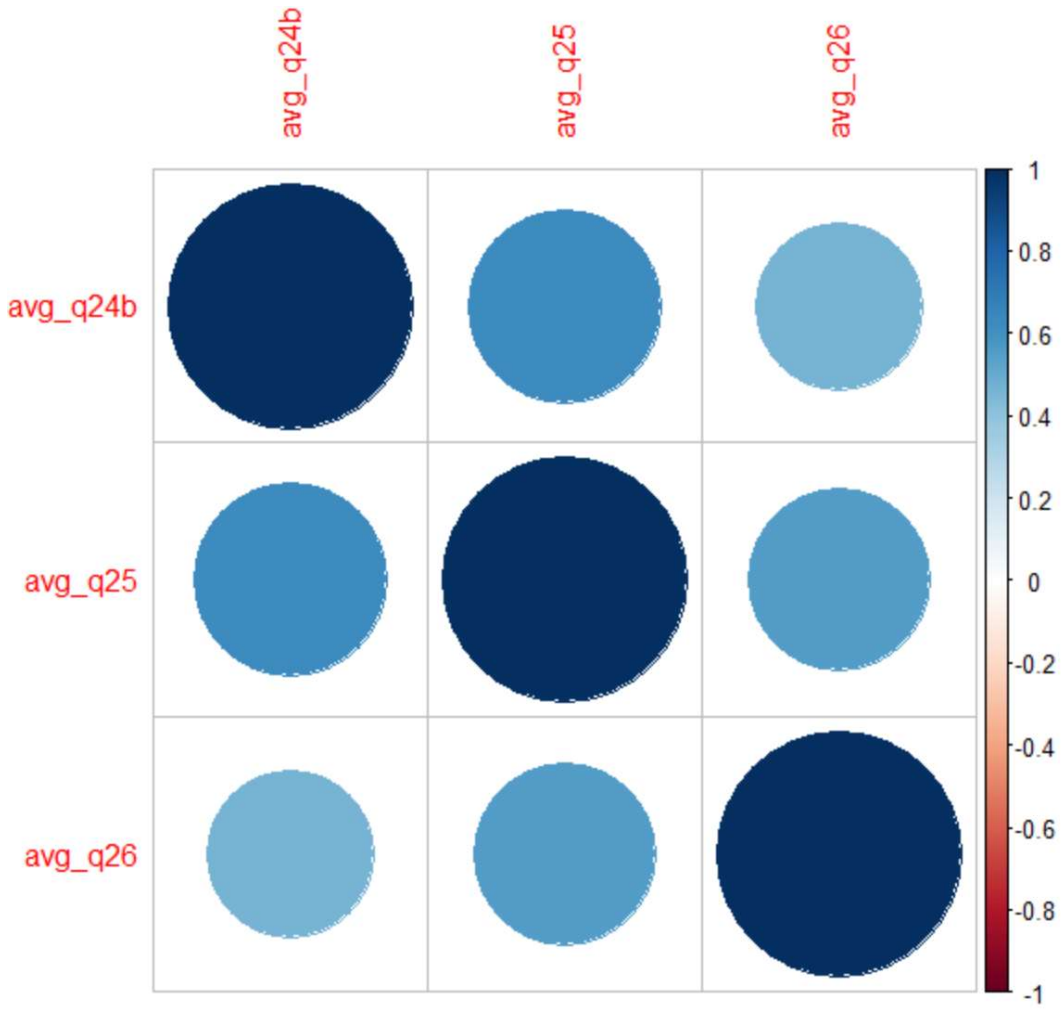


Figure 1. *Correlation plot of customers' responses to questions 24, 25, 26. The variable "avg_q24b" refers to customers' average Likert Scale score (six point rating scale) derived from their responses to statements 2, 6, 7, 8, 10, 11, 12; the variable "avg_q25" refers to customers' average Likert Scale score derived from their responses to all statements except for statements 6 and 12; the variable "avg_q26" refers to customers' average Likert Scale score derived from their responses to all statements except for statements 3 and 11.*

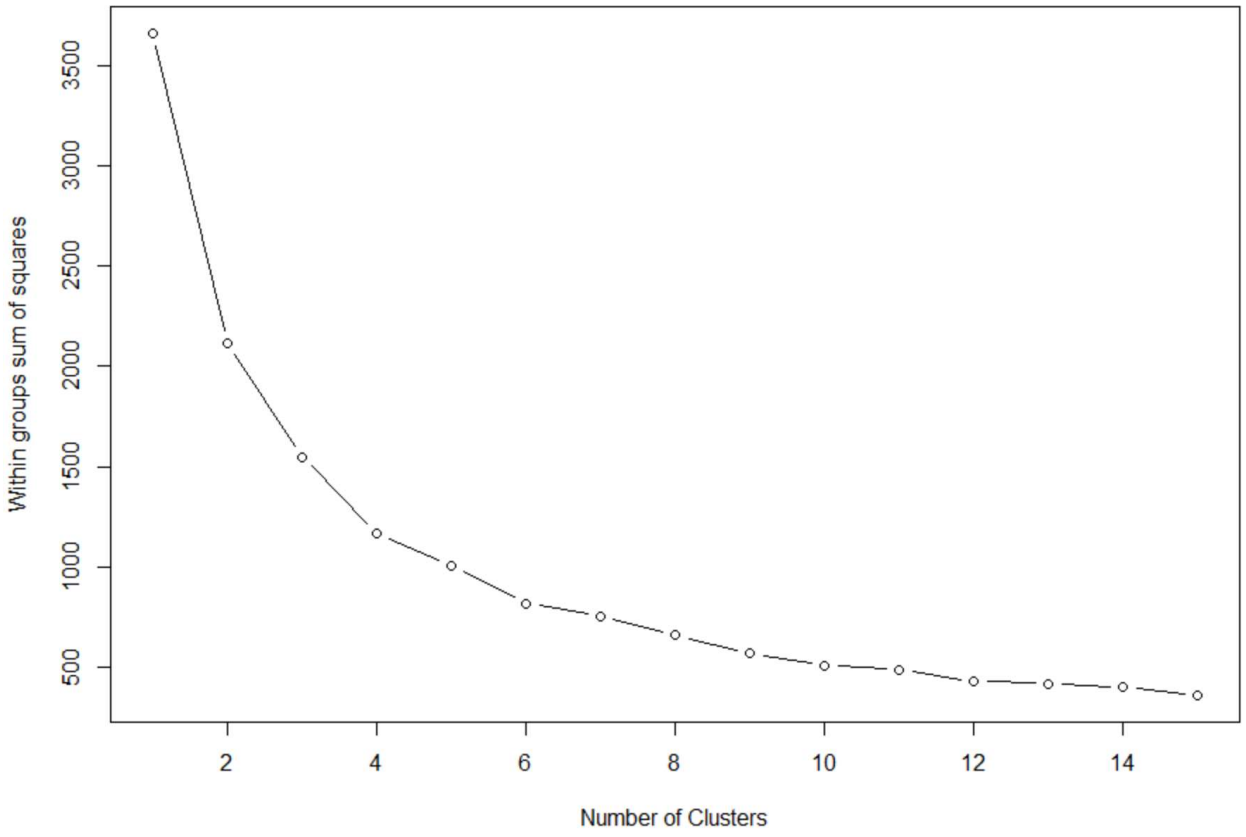
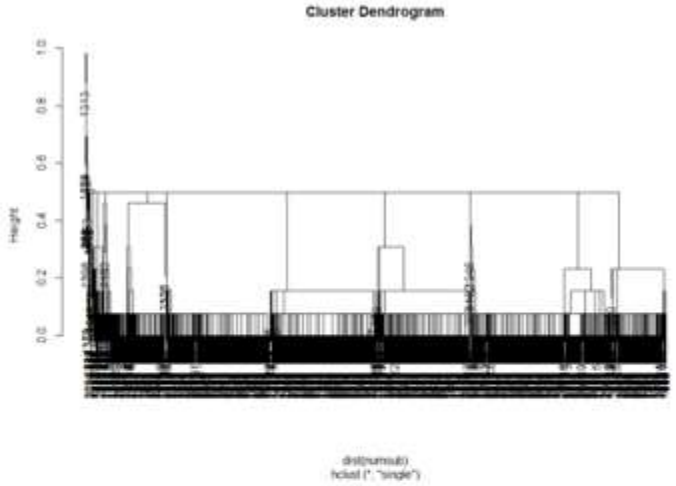
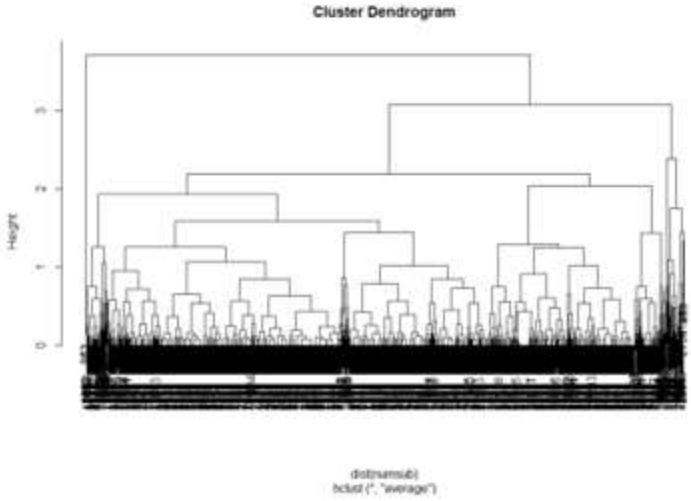


Figure 2. *Scree plot visualizing the variance explained (within groups sum of squares), proportion of variation, by each Principal component (cluster) from principal component analysis (PCA).*


Cluster Dendrogram
Height
dist(numsub)
hclust (*, "single")


Cluster Dendrogram
Height
dist(numsub)
hclust (*, "average")

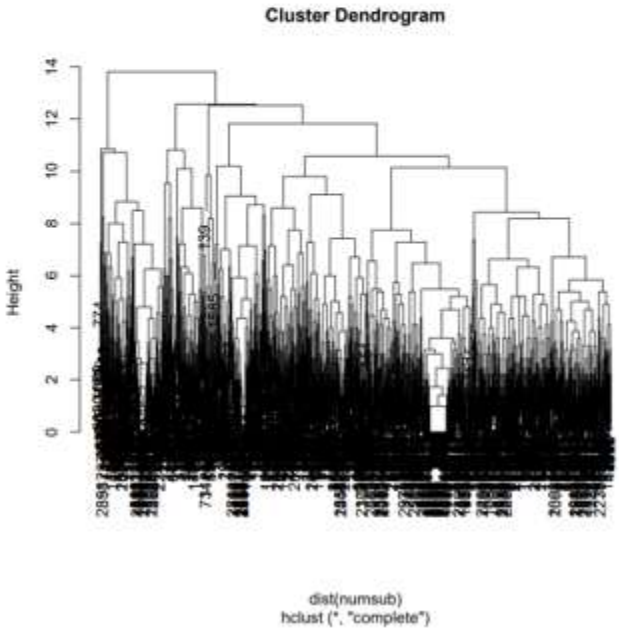

Cluster Dendrogram
Height
dist(numsub)
hclust (*, "complete")

Figure 3. *Dendrograms from hierarchical clustering of customers' responses to survey questions 24, 25, 26. Top panel: single linkage. Center panel: average linkage. Bottom panel: complete linkage.*

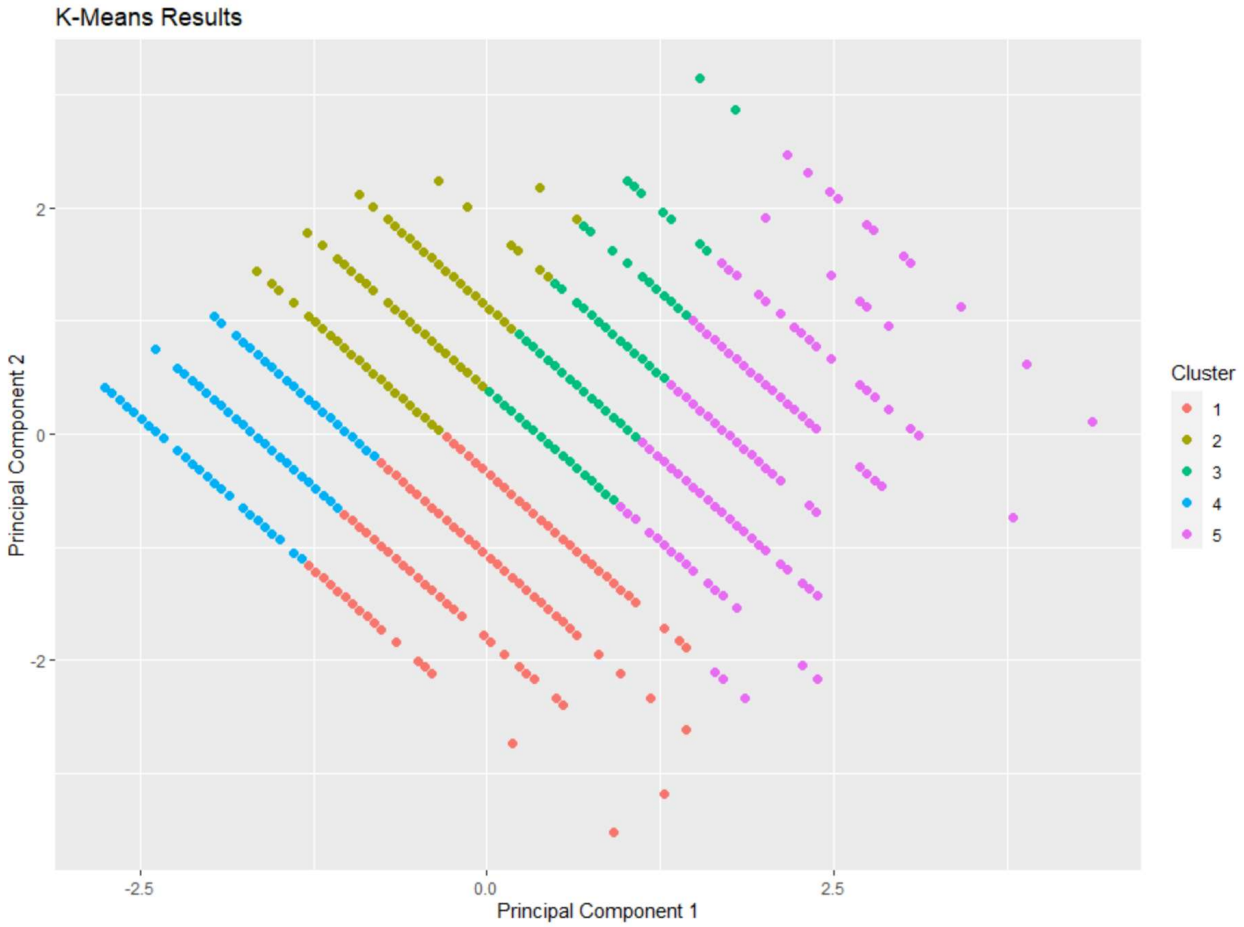


Figure 4. *Scatterplot of first two principal components of customers' responses to survey questions 24, 25, 26, with points colored to identify the clusters found in the data. The five derived clusters are A. coral pink; B. olive green; C. light green; D. light blue; E. purple*.

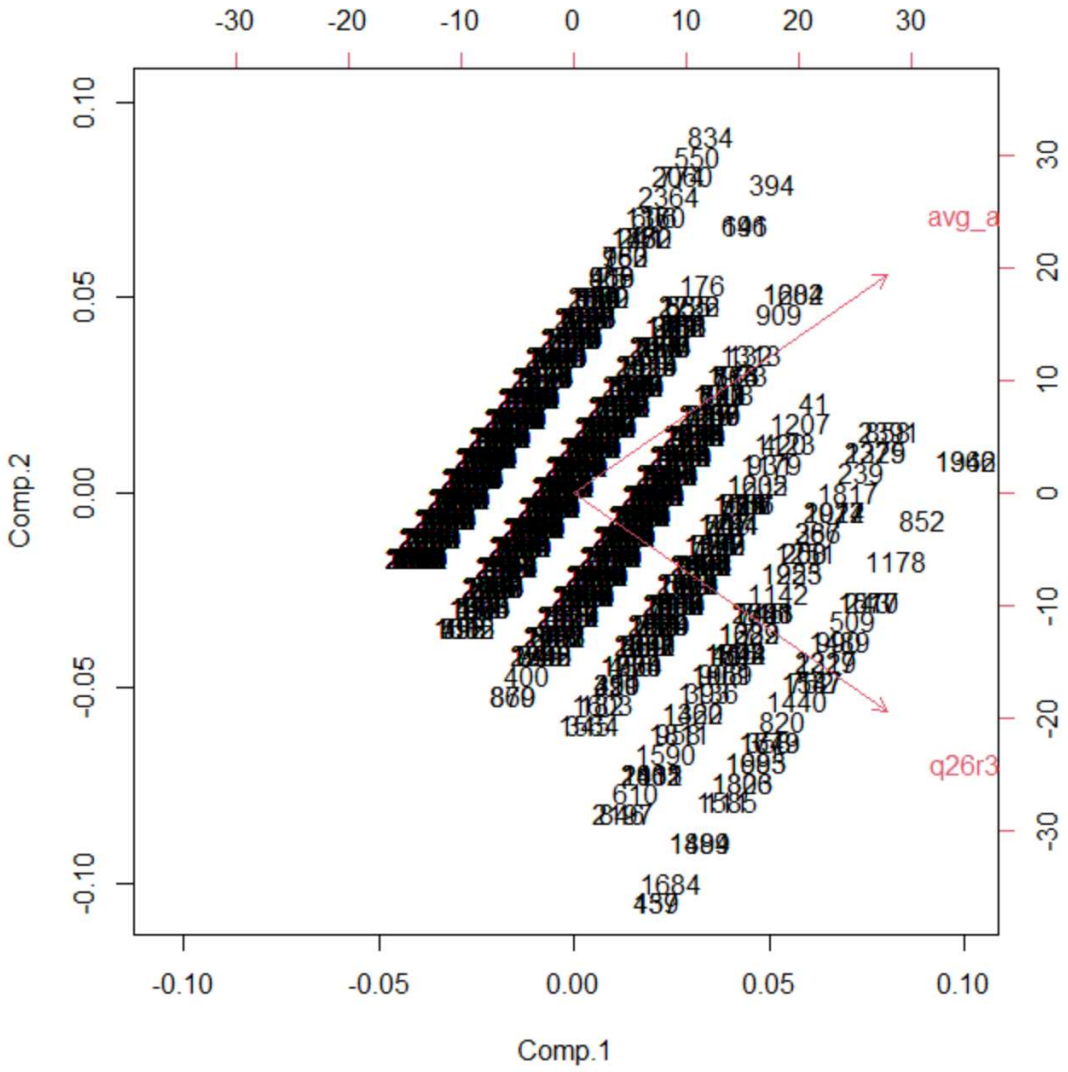


Figure 5. *Biplot of first two principal components of customers' responses to survey questions 24, 25, 26, with the variable "avg_a" and the variable "q26r3" both contribute to principal component 1 (PC1). The variable "avg_a" represents customers' average Likert Scale score (six point rating scale) derived from their responses to statements 2 & 7 of question 24 and statements 14, 15, 17 of question 26, whereas the variable "q26r3" refers to customers' average Likert Scale score (six point rating scale) derived from their responses to statement 3 of question 26.*

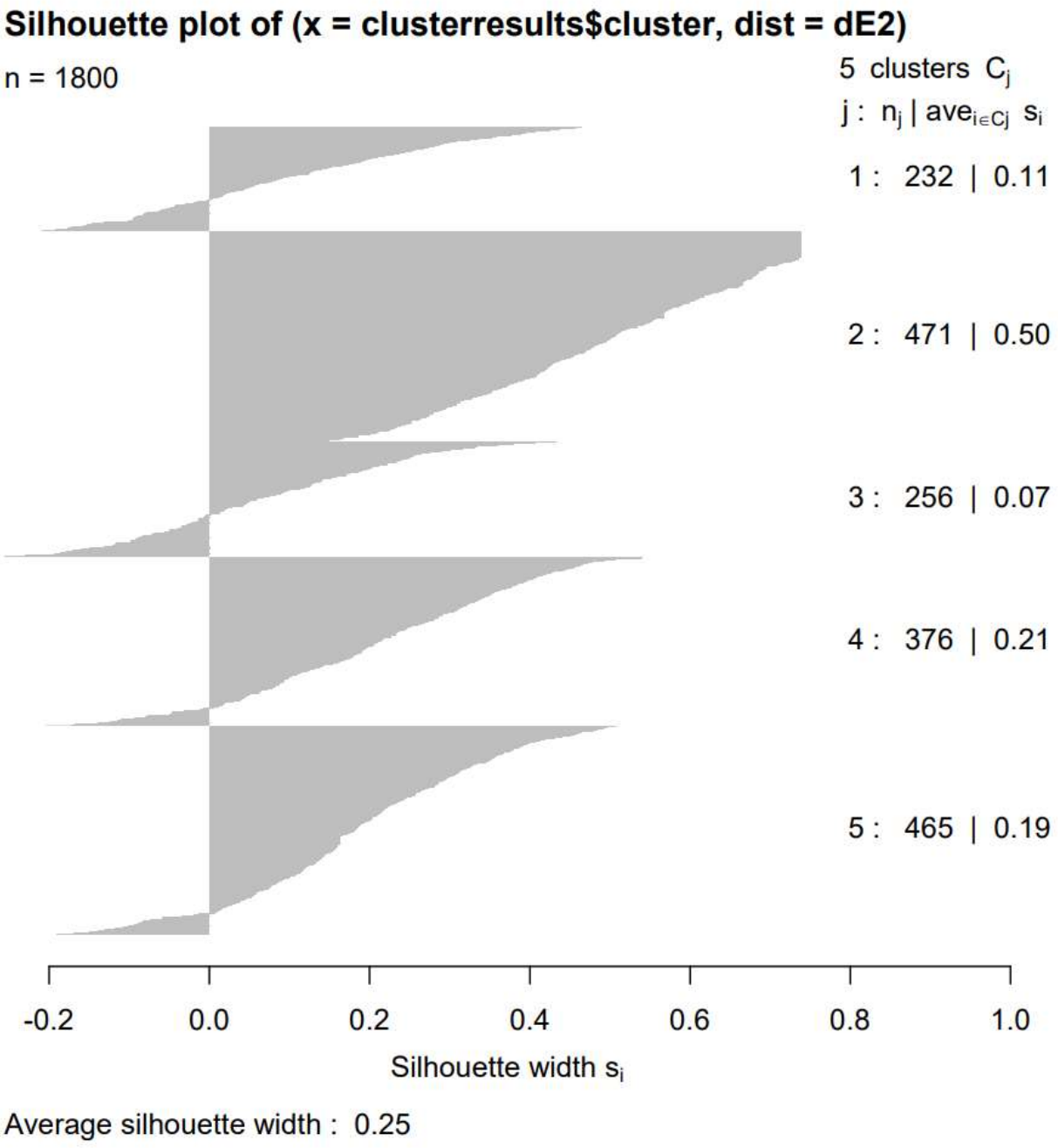


Figure 6. *Silhouette plot for grouping customers' responses to survey questions 24, 25, 26 using the partitioning around medoids (pam) clustering method with K=5 clusters*.