Vincent Pun
MSDS 450 Section 55
Solo 1: The App Happy Company and the Social Market for Entertainment Apps
01/24/2021

Introduction

The purpose of this project is to help The App Happy Company conduct a general attitudinal post hoc segmentation analysis to explore the market for social entertainment apps. App Happy is planning on introducing its first product in the consumer entertainment app category, so it contracted a third party company (Consumer Spy Corporation) to learn about customer behaviors and preferences about technology through focus groups and one-on-one interviews. A survey was created using this information, which will be used as the primary dataset throughout this analysis.

The goal of evaluating the survey data will be to identify groups of customers that exhibit similar characteristics, so App Happy can better understand the market for their upcoming social entertainment app that they are considering developing. Since App Happy does not have any historical information relating to the entertainment app category, the analysis utilizes various clustering techniques to determine how the business markets should be segmented. Seven market profiles will be presented based on the evaluated segmentation scheme, and market targeting guidance will be provided regarding product opportunities in the consumer entertainment app category.

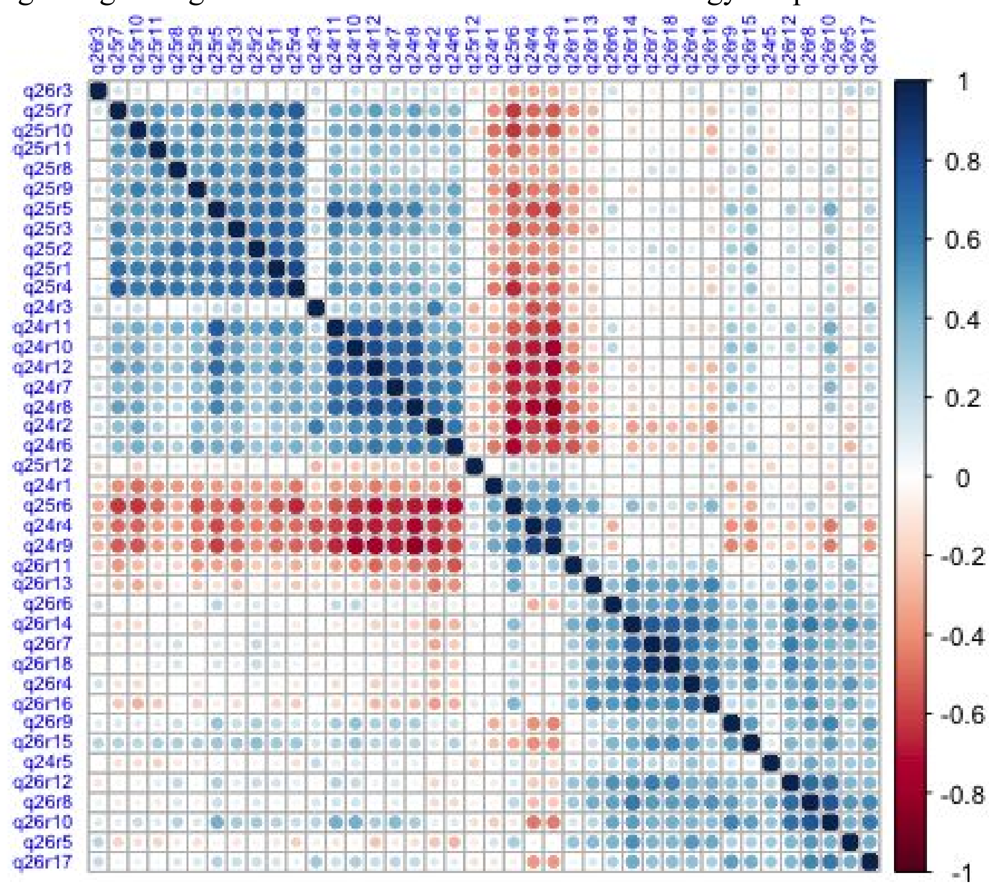
Data Quality/EDA

App Happy's survey research is collected and stored as R data frames; one contains string values that survey participants selected during their interviews, and the other is numerically coded for analytic purposes such as this segmentation scheme. The numerically coded dataset contains a total of 1,800 observed responses, and there are a total of eighty-nine variables that represent the fifty-seven questions in the survey. Questions that allow multiple responses are separated into multiple features (columns). The topics covered in the questionnaire cover areas such as basic demographic information (age, life stages, race), psychographic traits (lifestyle, values, personality), behaviors (needs and decision-making), and attitude towards multiple aspects of technology.

- **Missing Values** - There are only three variables that contain missing values.
    - **Question 5 Part 1** *(29.61% missing)* Upon analysis of the example survey file ("apphappy-quex-sum2014.pdf"), it is seen that question 5 is not included in this version of the survey. Given the large amount of omitted values, further analysis should be done to see whether different versions of the survey were offered to the respondents.

- ○ **Question 12** *(1.33% missing)*. This is not a large amount and is likely to be caused by user error. No action necessary.
- ○ **Question 57** *(5.5% missing)*. Survey participants are asked to identify their gender. To decrease nonresponses, more options may be given to participants regarding self-identification.

- **Variable Relationships** - There are multiple questions that are grouped within questions 24, 25, and 26, and they account for a total of forty-one out of the eighty-nine variables in the dataset. Participants are asked attitudinal questions about their preferences on technology use, values, and general spending habits. Numerical responses that range from 1 to 6 (strongly agree to strongly disagree) are recorded. It is seen below that many of these variables are positively correlated with each other, and there are a few questions that display a strong negative correlation as well. Clustering techniques in this analysis focus specifically on these variable relationships, as they provide the most valuable insights regarding both general consumer behavior and technology adoption.



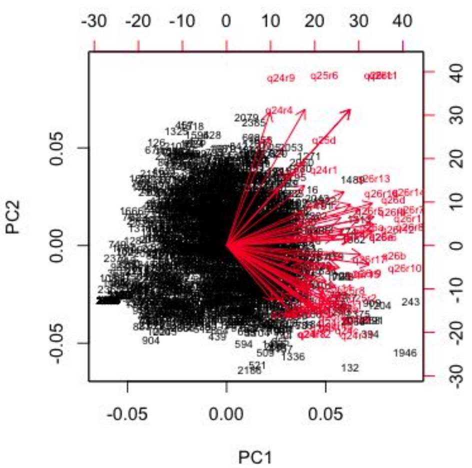
Principal Component Analysis

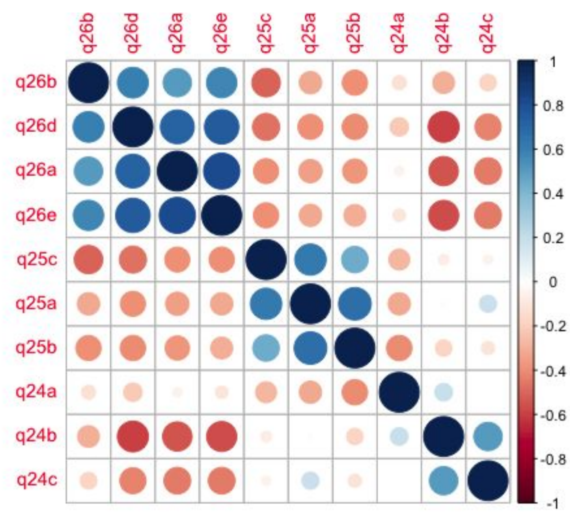
```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     4.9616303 2.9003277 2.16438298 2.09642816 1.81353387
Proportion of Variance 0.2911674 0.0994920 0.05540669 0.05198212 0.03889962
Cumulative Proportion  0.2911674 0.3906594 0.44606614 0.49804826 0.53694788
```

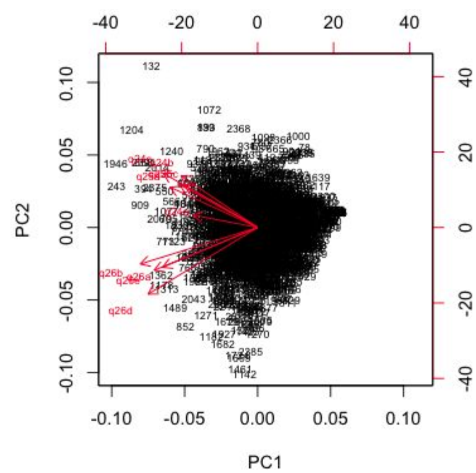
Original Data

```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     2.2412919 1.1094891 0.87543650 0.70192856 0.67170710
Proportion of Variance 0.5315736 0.1302605 0.08109906 0.05213776 0.04774484
Cumulative Proportion  0.5315736 0.6618340 0.74293310 0.79507086 0.84281570
```

Combined Features



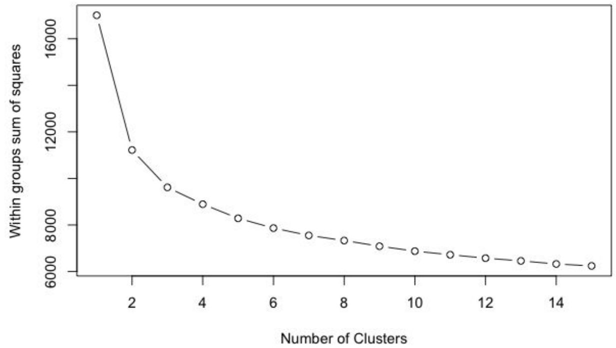


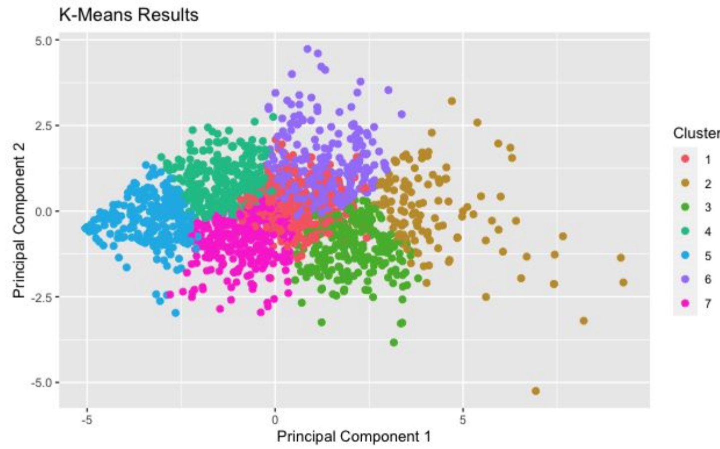


Performing principal component analysis was important in reducing the dimensionality of the data. The initial PCA computation shows that the first two components only capture roughly 39% of the variance from all the variance in the data. The initial biplot to visualize the results of PCA shows a dispersed relationship between the variables with mixed associations (both positive and negative) to the principal components.

After grouping attitudinal variables that were highly correlated with each other, PCA was performed again, and a significant increase in explained variance was observed in the first two principal components (66.2% variance explained), as shown above (see Appendix for brief descriptions of each combined variable). Furthermore, the subsequent biplot contains arrows (variables) that show clearer relationships to each of the principal components, and the stronger correlations between variables are also evident in the correlation plot above. Also, questions that were not strongly correlated in any direction were removed for further dimensionality reduction.

Cluster Analysis





A scree plot of the second principal component solution (top left) is used to visualize the total variance accounted for in each component when conducting a k-Means clustering analysis. A steep drop-off is seen between the first and second clusters, which reiterates the influence and explained variance of the first two principal components in the PCA calculation. The downward curve levels off between component 5 through component 8. After several iterations, it was found that producing seven clusters creates distinct groups that have minimal overlap.

While the results from the k-means approach produced clear groups, it was compared with two other clustering methods to ensure that the best method was used regarding App Happy's survey research data. PAM was another partitioning method that was used with a specified number of clusters set at seven, and hierarchical clustering was performed with a set height of seven.

R-squared, a goodness-of-fit measure, and the average silhouette width of each method are compared to determine the optimal solution. K-means performed the best for both evaluation techniques, so its results are used to profile the seven different groups in this segmentation scheme.

Clustering Method	**R-Square**	**Avg. Silhouette Width**
K-Means	0.5545 (best)	0.22 (best)
Hierarchical Clustering	.5163	0.12
PAM	.4251	0.17

Segment Profiles

To create thorough market segments and targets, the k-means results are recombined with some of the features found in the original survey dataset. Visualizations and interpretations of features related to age, income, education, race, and family information are compared between each cluster (can be seen in R Markdown file). Also, averages are calculated for the responses to the questions that were used in the clustering analysis for each of the seven groups, which is used to draw comparisons between the cluster profiles. Each profile is explained below:

1. **Late Majority** (16.4% of the market). Second-largest homogenous group. Relatively high-income. This group is the least racially diversified group. They rank about average compared to the other groups concerning technology use, influencing others, and consumption of new products.
2. **Laggards** (6.9% of the market). This group consists of adults (40-44) with relatively lower income. Technology is not very important in their lives, they tend to not make leadership decisions, and shopping is not something that they consider enjoyable. They are the lowest income earning group of the seven.
3. **Followers** (15.3% of the market). Young (18-24) and has the least education relative to other profiles. Second least tech-savvy group. This group makes less leadership and controlling decisions in their communities, but they are willing spenders and open to shopping for new trends.
4. **Early Adopters** (14.8% of the market). Group 4 consists of high-income earning young (30-34) families with some college education. Music, television, and social communication are important to them. While they are not particularly influenced by trends, they are driven, optimistic, and willing to make purchases that align with their personal brand/values.
5. **Innovators** (15.8% of the market). Group 5 is young (18-24), educated, diverse, and most consistently in agreement with integrating new technology with their lifestyles. Customers in this profile may have young children ranging from the ages of 0 to 12, see themselves as leaders in their communities, and are active consumers of apps and popular trends. Majority female.
6. **Adult Majority** (14.2% of the market). Highest levels of education, more mature families (children over 18), comfortable with using technology for entertainment and family reasons. Willing leaders, but are unlikely to make many purchases, especially related to tech.
7. **Avid Consumers** (16.6% of the market). Young (18-24), well-educated, diversified, majority male. Noticeably more single and without children. They are very open to making purchases, exploring trends, and using their phones for entertainment.

Given the seven profiles found from this segmentation scheme, it is clear that certain groups such as the Innovators and Early Adopters find the integration of technology with their day-to-day activities more enjoyable. Additionally, it is observed that those same groups that enjoy using technology also believe that they often share, lead, and influence others regarding through their values, experiences, and acts of self-expression. Purchasing approaches was not always correlated with a group's affinity for technology and applications; this is evident in the Adult Majority group.

Effectiveness in market segmentation can be achieved if target markets are "conceptually distinguishable and respond different to marketing-mix elements and programs (Kotler & Keller, 2012, p.263). For example, the Innovators should react much differently than the Laggards if they were to be presented with the same promotional campaign incentivizing users to test App Happy's social entertainment app. Other criteria that should be considered include measurability of segment characteristics, segment size relative to the mass market, and accessibility through available channels and platforms.

Given that The App Happy Company is still in an early stage of discovering its potential consumer base for their entertainment app, it may be beneficial to target multiple segments to better understand the market. These segments include the Innovators, Avid Consumers, Early Adopters, and Adult Majority; in total, they make up 61% of the total markets. Demographic characteristics that these groups have in common is that they are all generally well-educated, may have young children, and have higher incomes. Since the current recommendation is to target over half of the entire market for discovery purposes, it may be helpful for App Happy to increase levels of customization in their marketing strategy after a clearer understanding of the social entertainment app's objectives are available.

Additionally, it is important to understand that the survey research represents findings at a given point in time, and market behavior should be reassessed iteratively given the pace of digitization throughout many aspects of society. Finally, this segmentation scheme is heavily dependent on whether the survey was conducted on a sample that is representative of the population that App Happy is striving to generalize. How are participants chosen to participate for this survey? Were there any limitations (i.e. geographic) when Consumer Spy Corporation hosted its focus groups and one-on-one interviews?

Classification

After applying segmentation schemes to discover new target markets, App Happy will need the ability to classify new consumers to the defined segment profiles without relying on survey research data. For this to happen, classification techniques are used to observe labelled data to make predictions on new observations.

While new prospective consumers may not answer questions regarding their affinity for technology and entertainment apps, App Happy may still be able to collect data regarding basic demographic information for these individuals. Through supervised learning methods such as the Naive Bayes classifier, probabilities of cluster profiles can be calculated as a function of predictor variables such as age, income, education, and family size. Cluster profiles will be assigned to the unlabelled data according to the combined probabilities based on these features.

Appendix: Cluster Observations

Questions 24, 25, and 26 have been ranked from 1 to 7 based on their levels of agreement (strongest to weakest).

Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Total Count	279	117	261	252	269	241	282
% of Market	16.4%	6.9%	15.3%	14.8%	15.8%	14.2%	16.6%
Age (1)	30-34	40-44	18-24	30-34	18-24	35-39	18-24
Gender (57)	Slightly more F	More M	Slightly more F	More F	More F	Slightly more F	More M
Income (56)	60-69	30-39	50-59	70-79	50-59	60-69	60-69
Education (48)	Some college	Some college	Some college (Least Educated)	Some college	Some college	Some college (Most Educated)	Some college
Race (54)	Least diversified	Neutral	Diversified	Neutral	Diversified	Neutral	Diversified
Latino (55)					More		More
Marital Status (49)	Married	Married	Neutral	Married	Neutral	Married	Single
Children (50)	No Children	Mixed	Mixed	0-6,6-12	0-6,6-12	Children, 18+	No Children
24a Tech Positivity	4	7	6	2	1	5	3
24b Music TV	5	7	6	2	1	3	4
24c Internet/Co	5	7	6	2	1	3	4
25a Lead	5	7	6	2	1	3	4
25b Risk/Contr	5	7	6	2	1	3	4
25c Driven	5	7	6	2	1	3	4
26a Bargain/Deals	4	7	5	3	1	6	2
26b New Apps	4	7	5	3	1	6	2
26c childimpact	5	7	3	4	1	6	2
26d spend for trends	5	7	4	3	1	6	2
26 willing to spend	4	7	5	3	1	6	2

References

1. Kotler, P., & Keller, K. L. (2016). Marketing management. Boston, Massachusetts: Pearson.
2. Chapman, C. & Feit, E. (2015). R for Marketing Research and Analytics. New York: Springer.