Chris Pelkey
PREDICT 450-55

Introduction

Set forth in the proceeding is an analysis of data received from Consumer Spy Corporation by the consumer company App Happy. App Happy is a company that provides B2B analytic apps and is trying to break into the consumer entertainment app industry. Outlined in the following pages is an exploratory data analysis, a market segmentation approach, recommendations based on that segmentation and a description of typing tools to help segment future customers as they begin to interact with App Happy.

Exploratory Data Analysis

An initial dive into the data reveals that there are 1,800 observations in the dataset provided by Consumer Spy Corporation with 89 variables. These variables include identifying information on the participants and include:

- Ages
- Phone type
- Type of apps that they use
- How many apps they use
- What percentage of apps are free
- What websites they visit most

In addition to this demographic information, attitudinal data is taken on the consumers in order to help segment them by their opinions. This data is measured on a 1 to 6 Likert scale ranging from strongly agree to strongly disagree. Questions from this set include whether or not the respondents think of themselves as a opinion leader, whether they keep up with technological developments, whether they think it is worth spending a few extra dollars to get app features, etc.

In the initial dive it can be noted that there are some anomalous data that have chosen to be ignored for the sake of this research. Some users answered only in extremes while taking the survey, others did not appear to have closely have read the questions and have conflicting answers, while others appeared to only have wanted to strongly agree with the questions as examples. Because we have no real indication as to why this occurred in the data, the inputs are taken at face value and incorporated into the mix as is. With most cluster research, there may be some reason for this in the data and we do not want to eliminate potential patterns from the segmentation.

Market Segmentation

As mentioned previously, there are both attitudinal and demographic questions present. Throughout the data are questions that measure many aspects of how the respondent feels about themselves and their shopping behavior. In order to best find clusters that will be beneficial for App Happy, the process of this report focuses in on those customers who believe that they are influential, on the cutting edge, and who are willing to be lavish with their expenses. Therefore, the clustering revolved around the 11 questions presented in **Table 1**.

Number	Question
q24r1	I try to keep up with technological developments
q24r2	People often ask me advice when they are looking to buy technology or electronic products
q24r3	I enjoy purchasing new gadgets and appliances
q25r1	I consider myself an opinion leader
q25r3	I like to offer advice to others
q25r5	I'm the first of my friends and family to try new things
q26r8	I can't get enough Apps
q26r10	I love showing off me new Apps to others
q26r12	It's usually worth spending a few extra dollars to get extra App features
q26r13	There's no point in earning money if I'm not going to spend it
q26r16	I tend to make impulse purchases

Table 1 – Questions used from the survey

By winnowing to these questions we are able to identify power users and influencers, those people who are willing to spend money and who view themselves as influential to their friends and family.

After the questions were selected the data was run through a rigor of tests in order to determine hoe cohesive the data is. **Figure 1** and **Figure 2** represent how the data are correlated with each other. As can be seen, there are some instances where there are strong correlations, some medium correlations and a few weak correlations. There are no instances in the representations where there is an inverse correlation present.

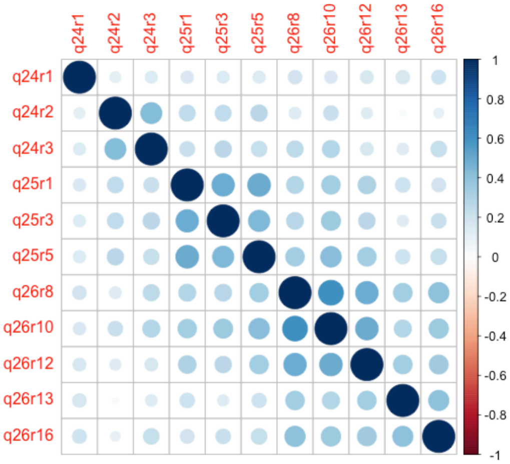
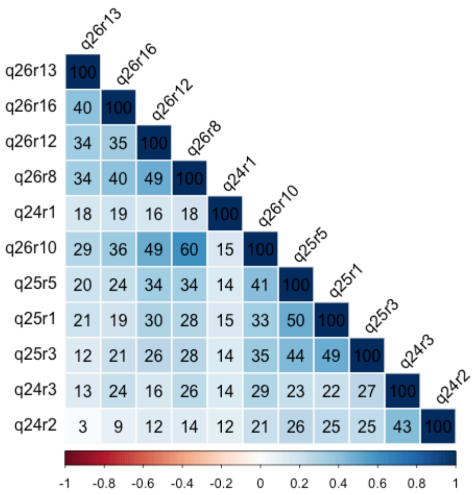


Figure 1 – Correlational plot of data



	q26r13	q26r16	q26r12	q26r8	q24r1	q26r10	q25r5	q25r1	q25r3	q24r3	q24r2
q26r13	100										
q26r16	40	100									
q26r12	34	35	100								
q26r8	34	40	49	100							
q24r1	18	19	16	18	100						
q26r10	29	36	49	60	15	100					
q25r5	20	24	34	34	14	41	100				
q25r1	21	19	30	28	15	33	50	100			
q25r3	12	21	26	28	14	35	44	49	100		
q24r3	13	24	16	26	14	29	23	22	27	100	
q24r2	3	9	12	14	12	21	26	25	25	43	100

Figure 2 – Correlational plot of data

K means clustering

A test was run on the data to determine the appropriate number of clusters that would come out of this data. Using the within group sum of squares, we are able to determine that either 5 or 8 clusters (**Figure 3**) are most appropriate to use in the clustering approach for the data. An initial dive through silhouette plots suggested that 5 is the more appropriate number.

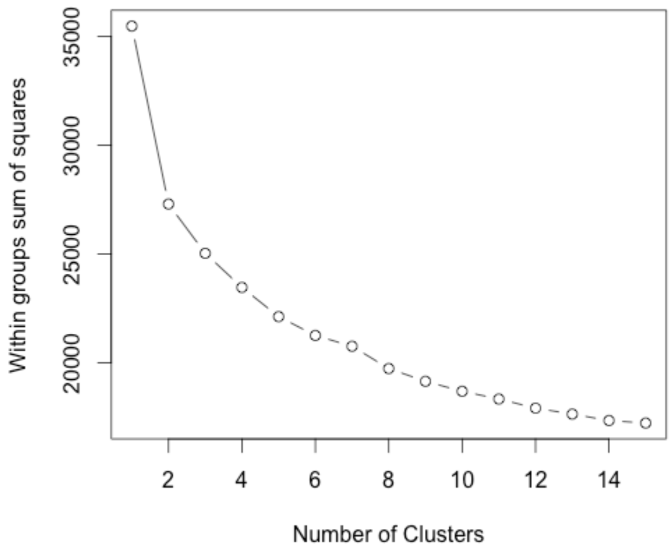


Figure 3 – Scree plot for clustering

The initial approach to clustering was to approach with a K means clustering technique using 5 clusters as the K. The R-Square value for the model is 0.3765. As you can see from **Figure 4** the average silhouette width is 0.21, with some very good scores mixed in.

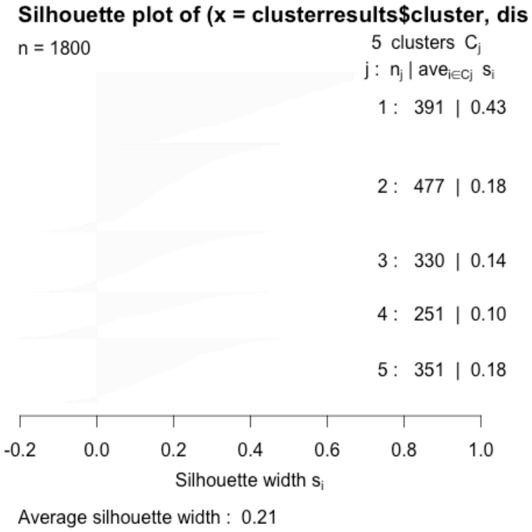


Figure 4 – Silhouette plot for K-Means clustering

This seems to be a very good clustering technique for this data, however we will run a couple more techniques to see if perhaps there is a better approach.

Hierarchical clustering

This method produced tepid silhouette width plotting, as can be seen in **Figure 5**. It appears that the average width is 0.02 and some of the clusters like clusters 3 and 4 are even inappropriately labeled. This is a terrible result and therefor will not be used when creating the market segmentation for the App Happy data. The R-Square for the data is 0.2222.

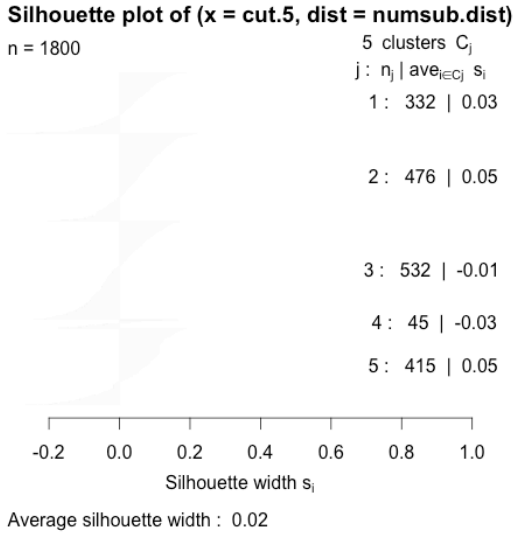


Figure 5 – Silhouette plot for Hierarchical clustering

PAM clustering

Again, this type of clustering did not perform very well, especially in comparison to the K means method. It did do better that the hierarchical clustering approach however, with an average silhouette length at 0.08, see **Figure 6**. Cluster 5 actually pulled in a really great width of 0.47 and none of the widths were negative like in the hierarchical method.

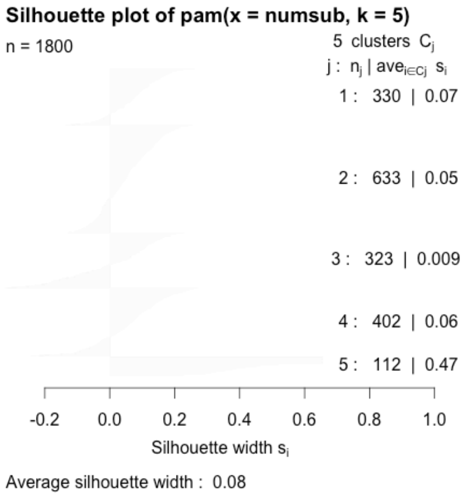


Figure 6 – Silhouette plot for PAM clustering

Segmentation, profile and recommendations

Reviewing the previous sections, it is pretty clear that by measure of silhouette width, we should be using the K means method to determine our clusters. The results of the clusters can be seen below in **Table 2**.

	q24r1	q24r2	q24r3	q25r1	q25r3	q25r5	q26r8	q26r10	q26r12	q26r13	q26r16
1	2.3760	1.3504	1.5192	1.5038	1.5141	1.4425	1.8363	1.5550	1.7596	1.7801	1.8184
2	2.7966	1.8197	2.0335	2.7065	2.6122	2.6520	3.2935	2.8637	3.1468	2.7254	2.8407
3	3.8121	1.3333	1.8091	1.9394	1.6879	1.8939	2.9606	2.0667	2.9000	4.6455	3.5515
4	3.8765	2.6693	4.7450	3.0916	3.0996	3.2032	4.3108	3.9004	3.8167	3.7849	3.9243
5	3.4701	1.5812	1.8348	2.8319	2.4872	2.9430	4.7094	4.2051	4.4473	4.4416	4.1652

Table 2 – Cluster results and attitudinal averages

The main purpose of the clustering for App Happy was to determine power users, who would take their want of spending money, influencing friends and family and belief that they are cutting edge and segment it out from the rest of the users. Looking at the data in order to pull out these users we would like to find a cluster that has relatively low average scores, closer to the 1 of strongly agree on the Likert scale for these questions. In the K means set, you can see from **Table 2** that this is specifically Cluster 1. Looking closely at the demographic data we can see what these users have a profile of in **Table 3**.

Question	Score	Translation
q1	3.913043478	In the mid to late 30's age range
q11	3.163682864	Has between 11-30 apps on their phone
q12	3.808184143	Around 40% of those apps were free
q48	3.524296675	Have some college/college graduate
q49	1.846547315	Married
q50r1	0.511508951	Half of them do not have children
q50r2	0.240409207	24% have one child under 6
q50r3	0.199488491	20% have children 6-12
q50r4	0.138107417	14% have children 12-17
q50r5	0.092071611	9% have children over 18
q54	1.808184143	This statistic is skewed, but most respondents are white
q56	7.751918159	Make around $55,000-$65,000

Table 3 – Breakdown of target segmentation

The profile suggests that we should be targeting married, Caucasians in their mid-30's. More than likely they will not have children and will be in the mid $50k income bracket. They will also most likely have some college education or will have graduated. These people are the power users that App Happy should target to get people who are willing to spend money in apps, and spread the word to family and friends if the apps are successful. It should also be noted that in the K means clustering exercise that this cluster had the highest silhouette width of all of the clusters at 0.43.

Taking a closer look at the relevant data, we can then make a recommendation as to what types of app might be most useful to these consumers so App Happy has a place to start. Looking at **Table 4**, we can see which types of apps are very popular within this group of people. It would be our recommendation for App Happy, in targeting the power user to begin focusing in on apps in the following areas: Music and Sound Identification, Gaming Apps, Social Networking Apps, Entertainment Apps and General News Apps. By targeting these apps specifically, App Happy will be able to draw in the users that it is trying to get.

App type	Average Using (%)
Music and Sound Identification Apps	0.74168798
TV Check-In Apps	0.350383632
Entertainment Apps	0.649616368
TV Show Apps	0.457800512
Gaming Apps	0.831202046
Social Networking Apps	0.833759591
General News Apps	0.639386189
Shopping Apps	0.593350384
Specific News Publication Apps	0.450127877
Other	0.058823529
None	0.00511509

Table 4 – Breakdown of app type

Future segmentation

For future reference App Happy should begin to ask for demographic data from its costumers so that it can begin to segment these new people into the appropriate clusters. We would recommend using a classification and regression tree (CART) model to help place people in their appropriate segment moving forward. The best part of using a CART model is that it can handle outliers very well when noisy data can sometimes be misclassified.

This method takes the incoming data and follows a classification tree to determine which users would fall into the target cluster. It simply looks at an incoming variable and pushes your decision making scheme one way or another down a tree until you reach an end node. These methods are very simple. Following the structure of the tree is as simple as following a workflow structure and making the appropriate decisions as you proceed.

Conclusion

The above documentation takes a in-depth look at data as provided by the Consumer Spy Company to B2B company App Happy. As App Happy tries to break into consumer apps, as opposed to analytic apps, they will be well poised to target power users who are willing to spend money on apps and influence friends and family. The use of CART modeling will help App Happy to segment future incoming customers into this segment and will allow App Happy to target these new customers accordingly.