Market Analysis of Social Entertainment Apps for App Happy Company

Sunny Wu

Introduction

App Happy Company conducted market research for developing social entertainment app. Survey data were collected from a sample of consumers. The survey questionnaire were based on preliminary qualitative research that included focus groups and one-on-one interviews. This report conducted data exploration towards survey data, and built several clustering models. The K-Means clustering method is selected as the final model and music, social networking, and gaming apps are recommended.

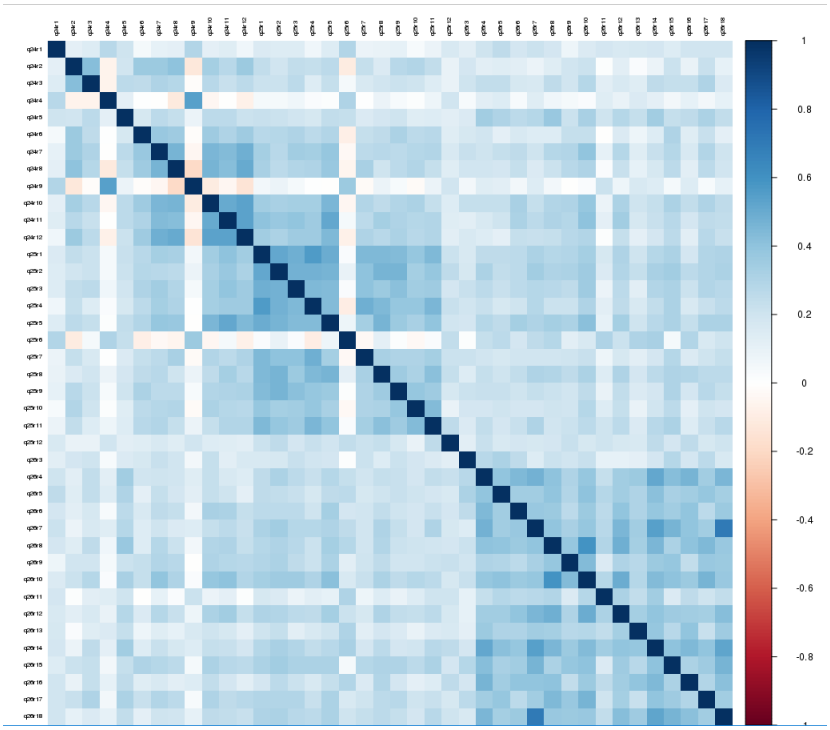
Data Exploration

The survey data is composed of 1800 observations and 80 variables. Forty out of 80 variables are attitudinal data. The attitude data are considered continuous data in this analysis, ranging from 1 (strongly agree) to 6 (strongly disagree). The first observation of this dataset is presented as an example below. The attitudinal data in this dataset for further analysis are presented in questions 24, 25, and 26 (Example: first question in question 24 is marked as q24r1 in the dataset).

```
> head(numdata,1)
     caseID q1 q2r1 q2r2 q2r3 q2r4 q2r5 q2r6 q2r7 q2r8 q2r9 q2r10 q4r1 q4r2 q4r3 q4r4 q4r5 q4r6 q4r7 q4r8 q4r9 q4r10 q4r11 q5r1 q11
1853   1853  2     0     0     1     0     0     0     0     0     0      0     1     0     0     0     0     1     0     0     0      0      0     1    4
     q12 q13r1 q13r2 q13r3 q13r4 q13r5 q13r6 q13r7 q13r8 q13r9 q13r10 q13r11 q13r12 q24r1 q24r2 q24r3 q24r4 q24r5 q24r6 q24r7 q24r8
1853   5     1     1     4     3     4     2     4     4     4      3      4      1     2     3     1     1     6     2     2     2
     q24r9 q24r10 q24r11 q24r12 q25r1 q25r2 q25r3 q25r4 q25r5 q25r6 q25r7 q25r8 q25r9 q25r10 q25r11 q25r12 q26r18 q26r3 q26r4 q26r5
1853     1      2      5      1     1     1     3     1     1     5     1     3     4      3      2      2      3     4     1     1
     q26r6 q26r7 q26r8 q26r9 q26r10 q26r11 q26r12 q26r13 q26r14 q26r15 q26r16 q26r17 q48 q49 q50r1 q50r2 q50r3 q50r4 q50r5 q54 q55
1853     5     4     6     1      5      6      6      1      6      3      6      2   3   2     1     0     0     0     0   6   2
     q56 q57
1853   5   1
```

Data Preparation

All 40 attitudinal data from question 24 (r1-r12), question 25 (r1-r12) and question 26 (r3-r8) are pulled together and a correlation matrix is plotted. The boxes with a warm color indicates a negative correlation. The negative points are q24r4, q24r9, q24r12, q25r6, q25r12, and q26r11.



Detect Outliers

The examinations of outliers shows that some survey respondents have given the same answers (1 or 6) for all questions, and some respondents have answered 1 or 6 for all questions. In this study, all outliers are kept since careful and through study of outliers are not conducted.

Transform Variables – Group Variables

Normalization or log transformation of variables are not conducted in this study since all variables are in the same scale from 1 to 6.

All 40 attitudinal variables are grouped together by computing their means to reduce multicollinearity and simplify the model. The transformations are listed below:

For question 24, 12 variables are reduced to 4 groups.

- 1, 2, 3, 5, 6 -> positive attitude towards technology –> q24a
- 7,8 -> music/TV –> q24b
- 10, 11 -> Internet/Communications -> q24c
- 4, 9, 12 -> negative aspects of technology -> q24d

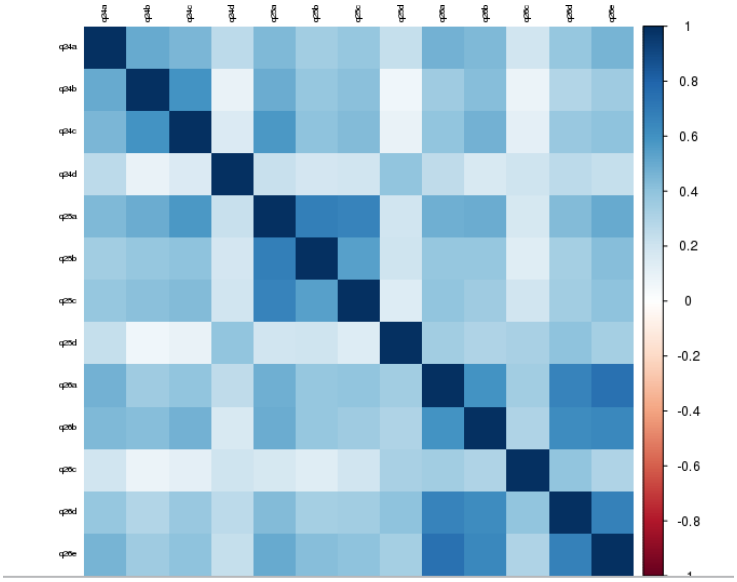
For question 25, 12 variables are reduced to 4 groups.

- 1, 2, 3, 4, 5 -> Leadership -> q25a
- 7,8 -> Control -> q25b
- 9, 10, 11 -> Drive –> q25c
- 6, 12 -> negative -> q25d

For question 26, 16 variables are reduced to 5 groups.

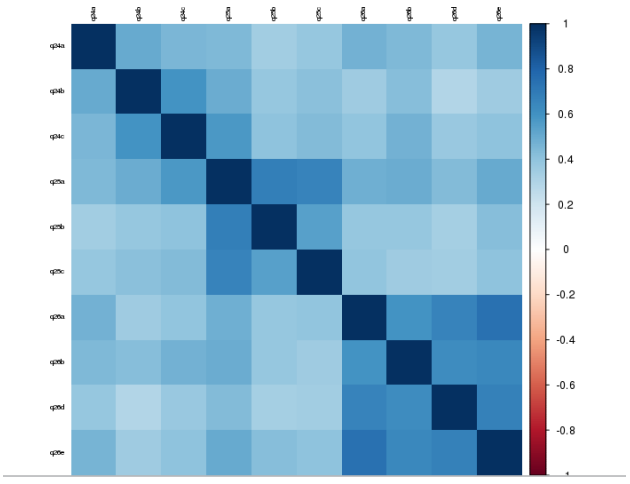
- 3, 4, 5, 6, 7 -> bargain -> q26a
- 8, 9, 10 -> show off -> q26b
- 11 -> children -> q26c
- 12, 13, 14 -> hot -> q26d
- 15, 16, 17, 18 -> brand d -> q26e

The correlation matrix of all grouped variables shows that variables q24d, q25d, and q26c have no correlations with other variables (white box).



Select base variables

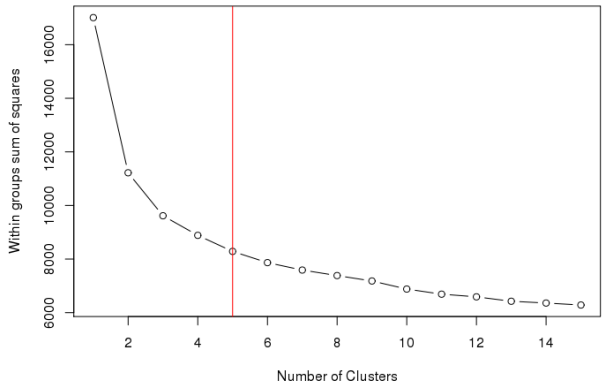
Variables q24a, q24b, q24c, q25a, q25b, q25c, q26a, q26b, q26d, q26e are kept as base variables, but variables q24d, q25d, and q26c are dropped. The correlations among all remaining variables are all positive. Principal component analysis shows that the first 2 principal components could explain 66.18% of total variances of the dataset.



Cluster Analysis

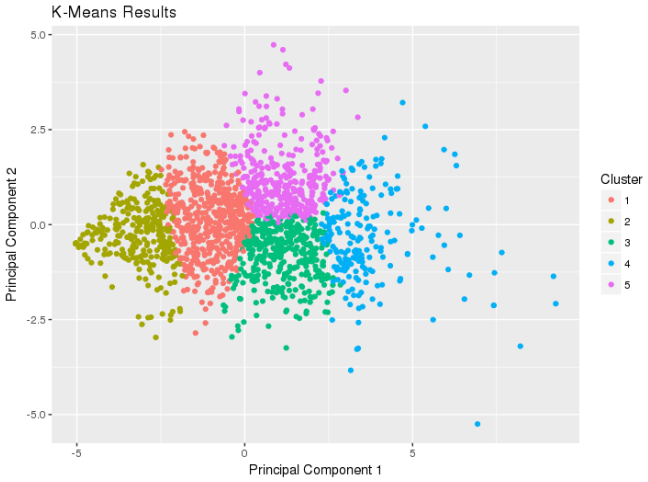
Determine Number of Clusters

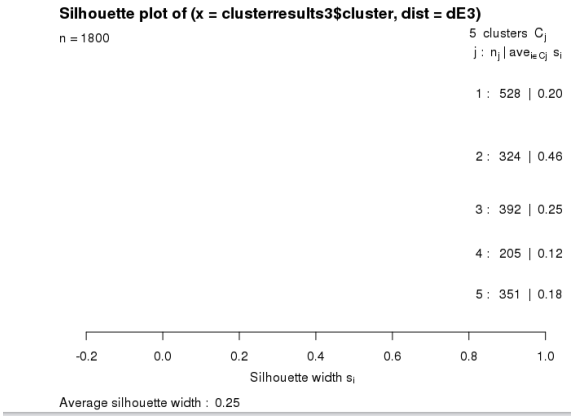
A scree plot is created to determine the number of clusters. The elbow point in this plot is n=5. Therefore, this dataset are divided into 5 clusters.



K-Means Clustering with 5 clusters

The result from K-means with 5 clusters is presented below. The average silhouette width is 0.25. Cluster 2 is the best cluster with silhouette width = 0.46, whereas cluster 4 is the worst cluster with silhouette width = 0.12.

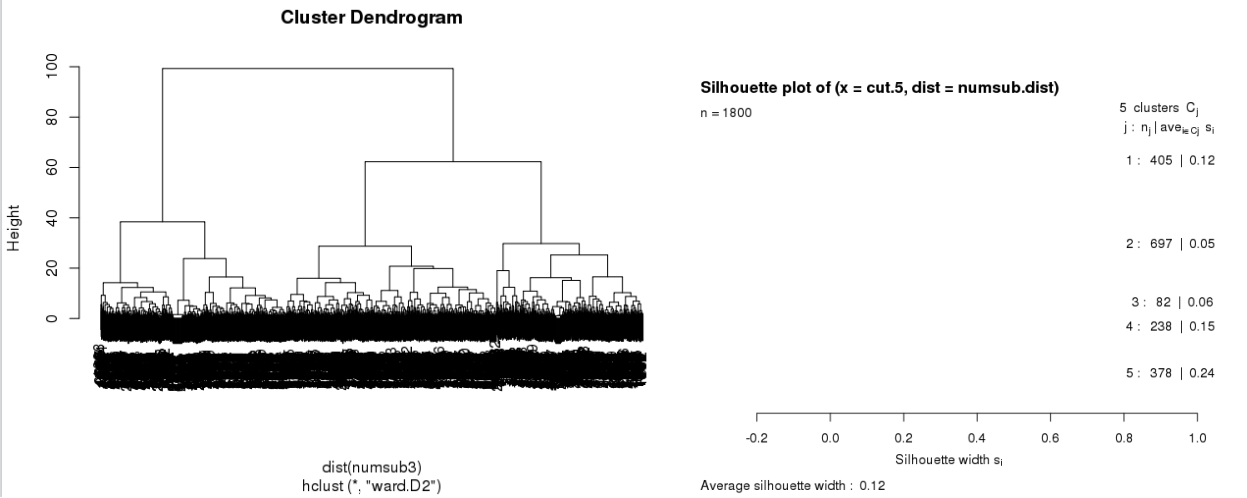




Hierarchical Clustering

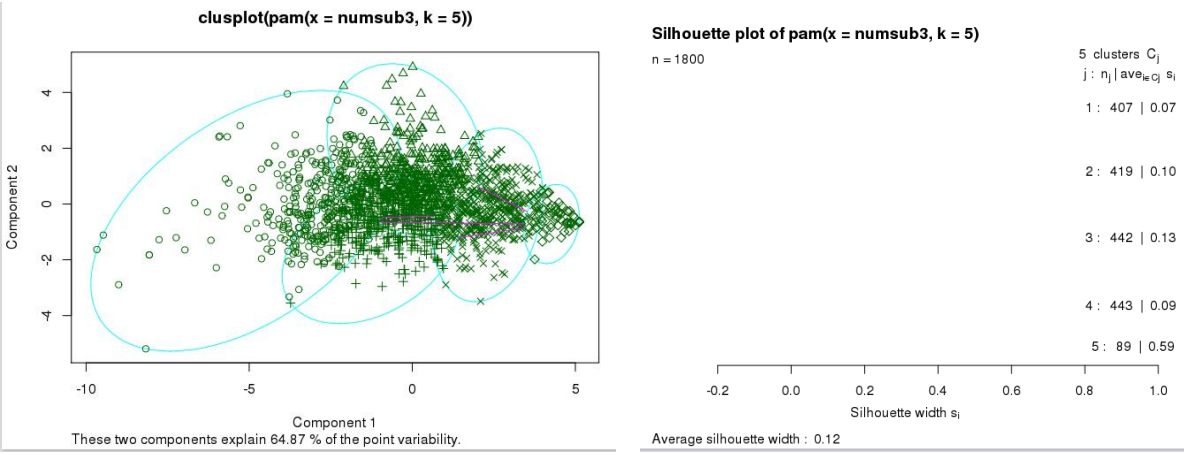
Hierarchical clustering has lower R^2 value, adjusted R^2 value and lower average Silhouette width than K-means clustering.

	R^2	adj-R^2	Average Silhouette Width
K-Means	0.5133	0.2594	0.25
hierarchical	0.4734	0.2197	0.12



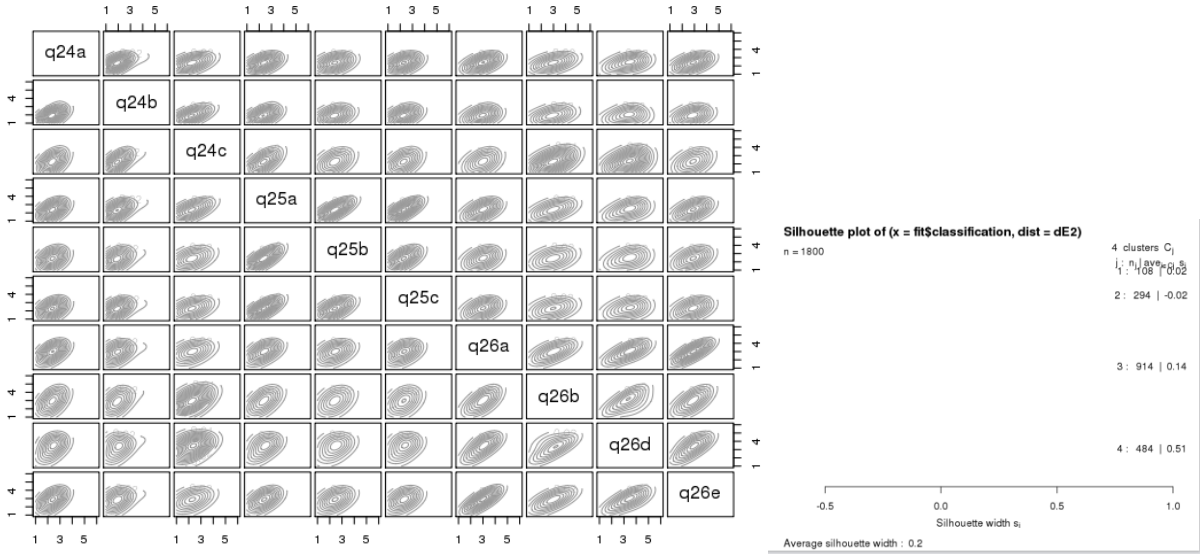
Partitioning Around Methods (PAM) Clustering

PAM clustering generates a 5 cluster model with average silhouette width of 0.12, and the first 2 prinicpal components explains 64.87% of total variances in the dataset.



Model Based Clustering

Model based clustering method divides the data into 4 clusters and the average silhouette width is 0.2.



Model Selection

The best model is the model with the largest average silhouette width. K-Means method is selected as a winner.

Methods	K-Means	hierarchical	PAM	Model based
Average Silhouette Width	0.25	0.12	0.12	0.2

The cluster results are also compared to understand the cluster similarities. The results are presented below. K-means cluster is most similar to PAM method and least similar to model based method.

K-means vs PAM cluster	K-means vs hierarchical	K-means vs Model based
0.41	0.26	0.19

Study the Profile of the Clusters

The result from K-means clustering method is used to create profiles. No correlation between the 5 clusters and age groups are found.

Correlations between clusters and people's favorite types of apps are found. For music and sound identification apps (q4r1), groups 1, 2 and 5 are highly interested >70%, while groups 3, and 4 have moderate interest >50%. For TV apps(q4r2), although all groups have low interest, group 2 has relatively high interest of 39.8%.

We recommend App Happy to develop social networking apps, which attract all clusters with > 60% interest, followed by gaming and music apps. The apps to stay away from are publication news apps and TV check-in apps.

This study result is valid because of close match between social trend and our study result.

	cluster	numdata$q4r1	numdata$q4r2	numdata$q4r3	numdata$q4r4	numdata$q4r5	numdata$q4r6
1	1	0.7689394	0.19128788	0.5189394	0.2708333	0.8219697	0.8901515
2	2	0.7469136	0.39814815	0.6635802	0.4876543	0.8518519	0.8456790
3	3	0.6326531	0.14030612	0.3954082	0.2244898	0.7142857	0.7372449
4	4	0.5463415	0.05853659	0.2292683	0.1121951	0.5512195	0.6146341
5	5	0.7378917	0.09401709	0.3390313	0.1082621	0.7264957	0.8518519

	numdata$q4r7	numdata$q4r8	numdata$q4r9	numdata$q4r10	numdata$q4r11
1	0.5795455	0.5303030	0.4109848	0.06818182	0.003787879
2	0.6450617	0.6574074	0.4444444	0.04320988	0.003086420
3	0.4260204	0.4030612	0.2474490	0.09183673	0.025510204
4	0.3170732	0.1951220	0.2000000	0.06341463	0.117073171
5	0.5612536	0.4188034	0.3789174	0.11680912	0.022792023

Conclusion and Recommendation

App Happy has conducted a successful survey to do cluster analysis. Outliers are detected in the dataset and kept in data analysis process. Outliers might be removed by incorporating expert opinion.

This study has grouped the attitudinal data and deleted unrelated variables to improve model accuracy. The variances explained by the first 2 principal components has improved from 36.14% (raw data) to 66.18% (transformed data). Data analysis recommends dividing the dataset into 5 clusters. The K-means clustering method is the best method with largest average silhouette width. The R^2 value of K-means method is 0.51 and adjusted R^2 value is 0.2594.

The cluster analysis shows that App Happy should prioritize developing social networking, gaming and music apps, and also avoid newspaper and TV apps. Moreover, the clusters in this study shows no correlation between cluster groups and age groups.

I recommend trying multiple variable transformation methods such as grouping variables and assigning them different weights, or convert 6-point scale variable to other scales. Moreover, more demographic data should be incorporated into data exploration, such as genders and races.

Last but not least, the experts opinions from AppHappy Company's marketing, business technical, and engineers professionals should be discussed and incorporated in the final selection of clustering.