❖ Introduction

The business goal of this segmentation analysis was to support Firm App Happy Company's potential development of a new social entertainment app to break into the consumer entertainment category. In this analysis, we will evaluate few segmentation algorithms, profile based on the segments scheme, and interpret the results in terms of marketing implications and actions.

❖ Data Assessment & Preparation

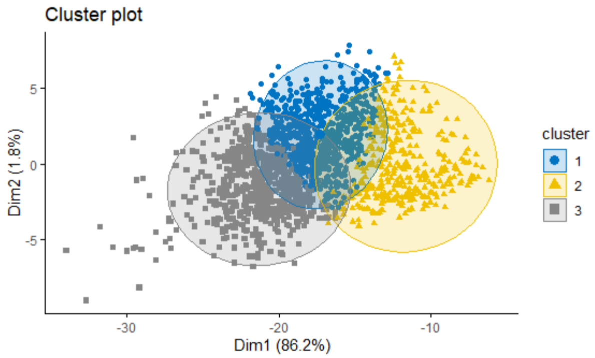
Based on Consumer Spy Corporation (CSC) market survey data, we are working with 2 files, one is numerically coded survey response data, and another is the character strings of the response labels. There were total 57 questions provided to the participants. We do not know the actual questions for all, except the subset of 16 questions which outlines demographics, purchase behaviors, and app download activities.

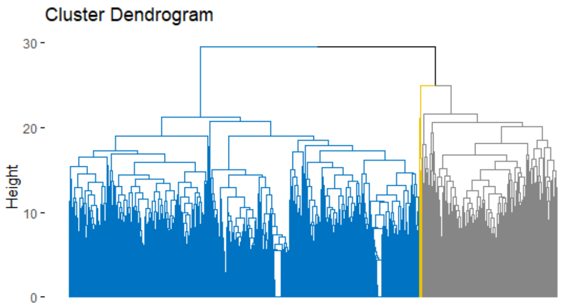
We have total 1800 observations and 89 variables. In which, there were 656 records have NA's. q5r1 has 533 NA's, which is 29.6% of the sample data. But since we do not know the question, and in order to avoid non-response bias. I used a function by adding third level into the column. The random.imp function was used to impute the missing values for q12 and q57.

❖ EDA

Clustering Tendency

Clustering tendency assessment helps evaluating the validity of clustering analysis. This is a quick EDA for the data set and validate for meaningful clusters. I used K=3 as random number. And based on cluster plot and dendrogram, we can see the data does contains meaningful clusters.





By computing the Hopkin statistics and using H=0.50 as the threshold to reject null hypothesis. Results came to 0.5876915, so we can reject null hypothesis, and conclude the data contains meaningful clusters.

Variables Evaluation

Given the App Happy Company is looking for a general attitudinal post hoc segmentation analysis from the data. The task is to discover consumer groups that have marketing meaning and can be used in marketing planning. The segments are generally defined and profiled by consumer attitudes, consumption behavior and demographics. Based on that, we identified q24, q25, and q26 to understand the attitudes toward technology, purchase behavior and app usage and preference.

Upon reviewing all 40 variables related to these three questions, I ran correlation heat map by each question and inspect if there is any highly correlated relationship among the responses and by using 0.50 threshold, I eliminated 8 from the total.

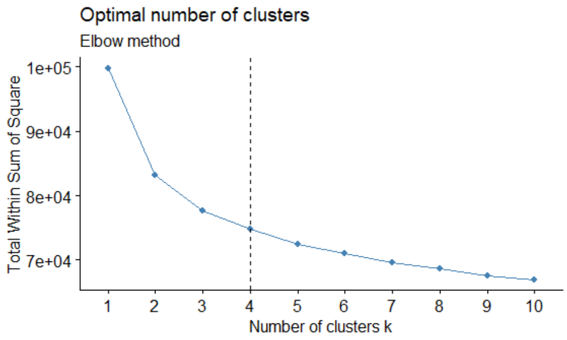
❖ Customer Segmentation

Non-hierarchical

Clustering algorithms can be separated into two, non-hierarchical and hierarchical. For nonhierarchical clustering, the desired number of clusters is specified such as K-means.

Hierarchical does not need to specify that, so I will use hierarchical approach first to understand the cluster estimate from the data.

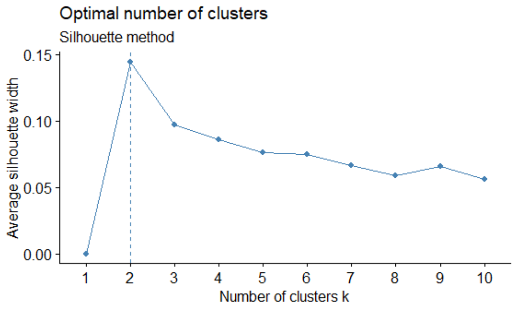
1. Elbow method



The Elbow method looks at the total WSS as a function of the number of clusters. The goal is to measures the compactness of the clustering and minimize WSS. The location of a bend in the plot is generally considered as the appropriate number of clusters.

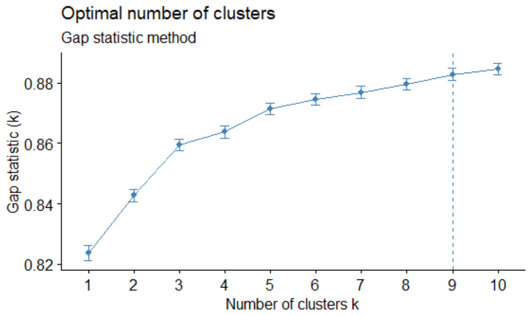
K assigned between 1-10

2. Average silhouette method



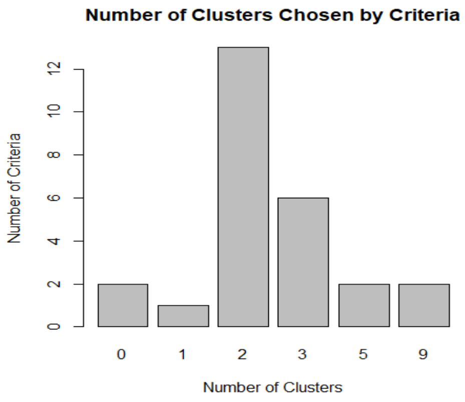
This method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k (Kaufman and Rousseeuw 1990).

3. Gap Statistic Method

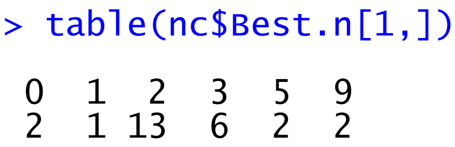


This method estimates of the optimal clusters by valuing to maximize the largest gap statistic. It selects the number of clusters as the smallest value of k as the gap statistic is within one standard deviation of the gap at k+1.

4. NbClust() function- 30 indices

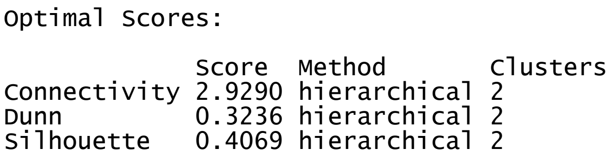


We apply the ‘NbClust’ function to the k-means to identify the most optimized clustering. This function uses 30 indices for choosing the best number of clusters. And the result shows number 2 is the best one, followed by 3 as the second best.



```
> table(nc$Best.n[1,])

 0  1  2  3  5  9
 2  1 13  6  2  2
```



```
Optimal Scores:

            Score  Method       Clusters
Connectivity 2.9290 hierarchical 2
Dunn         0.3236 hierarchical 2
Silhouette   0.4069 hierarchical 2
```

Based on these approaches, we received 3 different optimal clustering recommendations. This suggests the optimal number of clustering is subjective and depends on the method used for measurement and the parameters used for partitioning.

Clustering Analysis

Cluster can be measured by 4 metrics. R has a ClValid package compares clustering algorithms for analysis. I want to focus on the stability measures, which is a special version of internal measures that evaluates the consistency of a clustering result by comparing it with the clusters obtained after each column is removed, one at a time.

```
          Score       Method Clusters
APN 0.002184889 hierarchical        2
AD  8.646506970          pam        8
ADM 0.053719935 hierarchical        2
FOM 1.142168774       kmeans        9
```

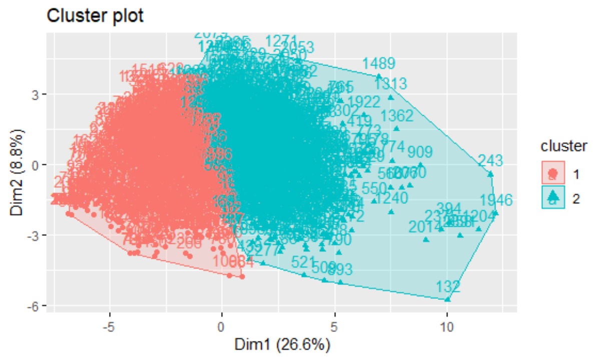
These 4 metrics in gauging the stability. We aim to minimize them as much as we can to 0.

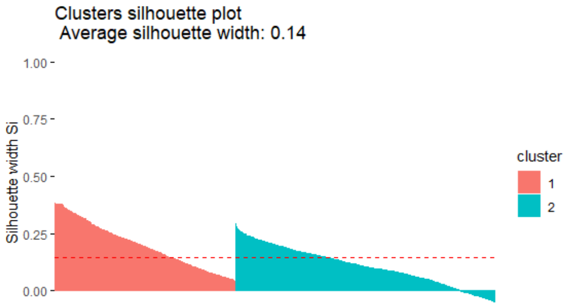
- AD (The average distance) measures the average distance between observations in the same cluster under both cases, full data set and a removal of one column. This metrics can display a value from 0 all the way to infinity.
- APN (The average proportion of non-overlap) measures the average proportion of observations not placed in the same cluster by first cluster on the full data, and then cluster on the data with a single variable removed.
- ADM (The average distance between means) measures the average distance between center of the cluster for both cases, full data set and a removal of one column.
- FOM (The figure of merit) measures the average intra-cluster variance of the deleted variable, where the clustering is based on the remaining variables.

In summary, hierarchical method suggests 2 as the most stable clustering, but 9 if using K-means. A characteristic of statistically robust segmentation is the observations are grouped in similar segments regardless of which method we use. Unfortunately, this is not the case here with the data.

Non-hierarchical Kmeans

Using Non-hierarchical clustering, the desired number of clusters is specified in advance in order to set the number of cluster centers. Each data point is then assigned to its nearest cluster center by minimizing or maximizing a desired criterion, with cluster centroids iteratively recalculated until they remain stable. Our stability results showed, kmeans at 9 as the most optimized. But as a marketer, segmenting and actioning against 9 clusters are not very interpretable, which led me to stay with clustering of 2 which deemed to be stable as well.





Silhouette plot showed we have some Si values in the negative level. Si measures how similar an object is to the other objects in its own cluster versus ones in the neighbor cluster. The values which is close to -1 indicates that the object is poorly clustered, and that assignment to some other cluster would probably improve the overall results. There are 140 of these data points have negative values.

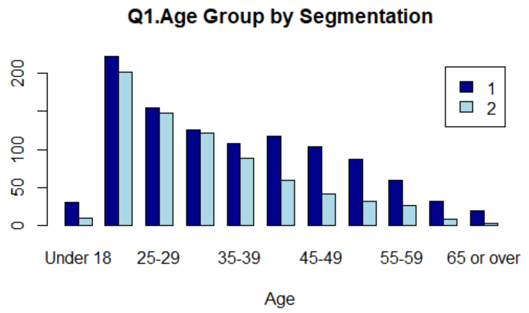
❖ Marketing Implication and Actions

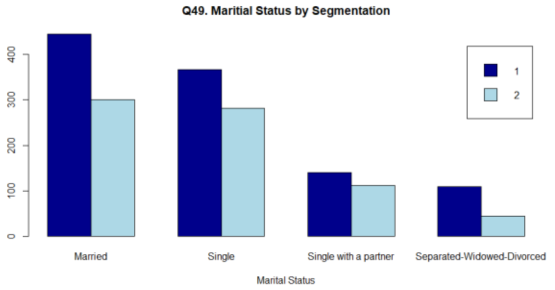
Clustering Profile

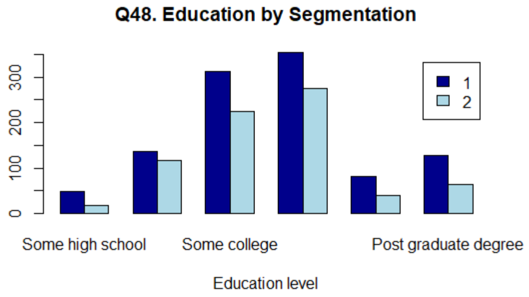
Working with 2 clustering by first reviewing the demographic data. Graphs suggested both segments have respondents clustered in the 18-40 age group, while Segment 1 have 20% more within the “under 18” group. When looking at age group 40 and above, Segment 1 having greater than 50% more respondents than Segment 2. Segment 2 has 75.7% of respondents fall within the age 18-40 group.

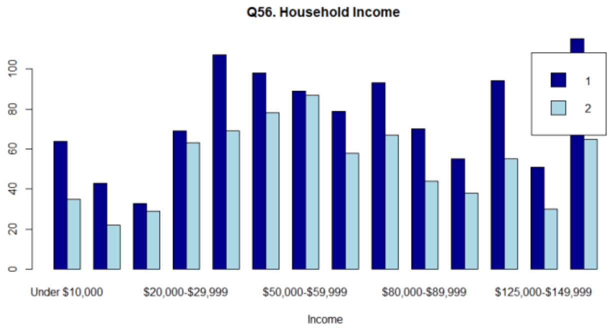
Segment 1 also has a slightly higher income than segment 2 >$80,000, also reflective of the age disparity, income >$20,000 is higher than segment 2. On the education side, Segment 1 slightly skewed higher educational level than Segment 2.

76.6 % of respondents in Segment 1 is white and Caucasian, while 69.9% in Segment 2.









Here is a table by reviewing some key characteristics on preferences among 2 customer profiles.

		1060 Respondents	740 Respondents
General Attitudinal	**App Volume**	66% have more than 10 apps Don't need as much apps, and less bothered by what's the latest	77% have more than 10 apps Largely feel they can't get enough of them, prefer cool apps
	App Cost	Really enoy free apps	Have about half of the apps are free but do not mind paying for some
	App Category	60% do not use Linkedin 43% using youtube 9.1% uses yahoo entertainment and music	88% frequent user in social media app 69% frequent user in youtube 30% enjoy yahoo entertainment and music
	Technology Sentiment	Feel too much information on he internet today	Feel stronly that the mobile as a source of entertainment, and music is very important to them
	Purchase Behavior	Prefer more planning rather than packaged deals No heavy preference for designer brands	Tendency for impulse buy and enjoy shopping all the time. Believe the brands they buy reflect their style, and enjoy luxury items
	Others	Responsibility and effort in earning money is important	Children has an impact somewhat to what they download

Initial Marketing Recommendation

Some initial marketing actions that App Happy company can do is tap into Segment 2 customers since they are considered as heavy users in the social media space. They are big in following the latest and greatest apps out there, and cool effect of the features are important to this group. Another interesting element in Segment 2, 96.7% agreed to the statement "Music is important part of my life". If a social entertainment app can be developed with focus on music and social interaction, that can be very attractive to this cluster. Using 4 P's as the guiding principles:

Product: Social interaction with music entertainment as the key feature.

Place: Via social platform as Facebook, YouTube and streaming services such as Netflix.

Price: If there are cool features and represent the latest and greatest, pricing an app as not free would still be attractive to the customers.

Promotion: No promotion initially since this group believe what they purchase reflect their style and image.

❖ Classification

App Happy would also like to use typing tools to classify future data into the segments that we have defined here. Since this analysis is a post hoc segmentation scheme specific, existing customers are segmented based upon their behaviors and attitudes, attributes that Kotler and Keller (2012) refer to as behavioral considerations. If that information is not available for new customers, then we will need to rely on demographic information to conduct our classification. We can use logistic regression since there are only 2 clustering, and we can also use Random Forests and K-Nearest Neighbors as classification methods. One challenge I noticed based on this survey data, there is not a very distinct differentiation on demographic data between 2 clusters, but app download frequency, and the category of apps perhaps can be supplemental.

❖ Conclusion

For this segmentation exercise, we have conducted both hierarchical and non-hierarchical clustering methods. A characteristic of statistically robust segmentation is that observations are grouped in similar segments regardless of which method we use. Based on various methods, we received 3 different clustering estimates. This suggests further evaluation of the survey and even sampling method should be revisited to ensure segmentation stability and avoiding sampling biases. First and foremost, a clear definition of the social entertainment application is critical before any primary marketing research. Data challenge we faced was not having a full list of the 57 questions, which limited the basis variable selections as well as the amount of profiles we can create for the segmentation. In addition, relevancy of a survey is also important. For example, q13 listed some obsolete app such as “AOL Radio”, which has been discontinued years ago. A better way in constructing that question would be creating categories of the apps, such as social media, music entertainment, video app, and streaming app, etc. This will allow the survey to remain relevant for years to come.

About 95 percent of new products fail. While The success of companies weighs in the balance of how much they understand and sympathize with the real-world setting of their most valued customers, and their ability to turn this understanding into actionable product that makes their customers’ lives easier. Marketing segmentation is truly a blend of art and science.

Reference:

Brock, Guy, Vasyl Pihur, Susmita Datta, and Somnath Datta. 2008. “ClValid: An R Package for Cluster Validation.” *Journal of Statistical Software* 25 (4): 1–22. https://www.jstatsoft.org/v025/i04.

Nobel, Carmen. Clay Christensen’s Milkshake Marketing. 2011. *Harvard Business School.* https://hbswk.hbs.edu/item/clay-christensens-milkshake-marketing

Han, Jiawei, Micheline Kamber, and Jian Pei. 2012. *Data Mining: Concepts and Techniques*. 3rd ed. Boston: Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00016-2.

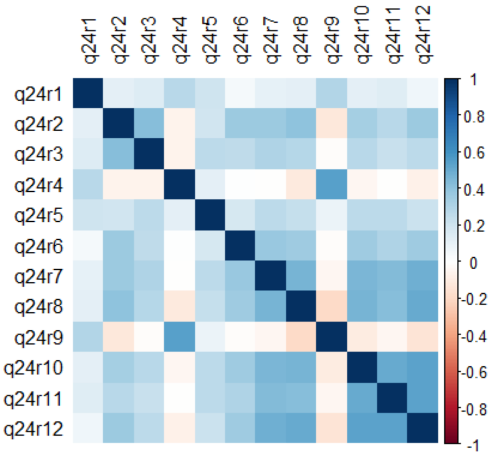
Charrad, Malika, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. 2014. “NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set.” *Journal of Statistical Software* 61: 1–36. http://www.jstatsoft.org/v61/i06/paper.

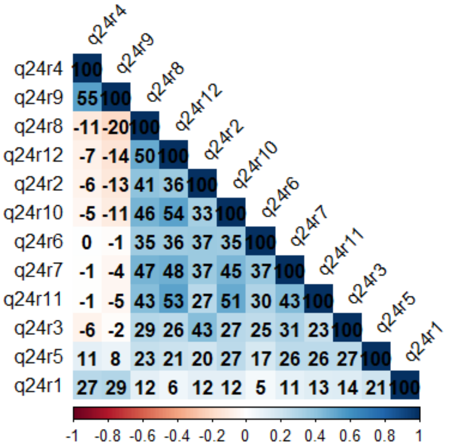
Kaufman, Leonard, and Peter Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*.

Kotler, P., & Keller, K. (2012). *Marketing management* (15th ed.). Boston, MA: Prentice Hall

APPENDIX

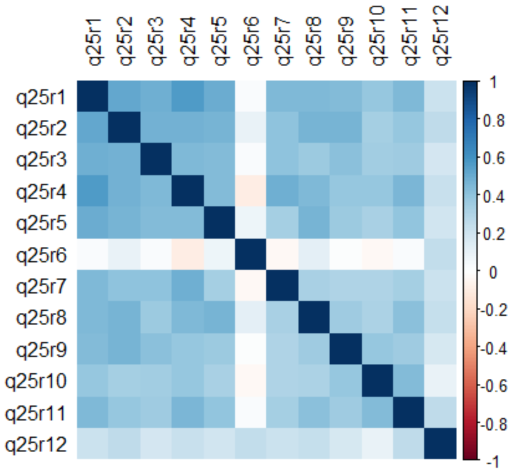
Q24. HEATMAP

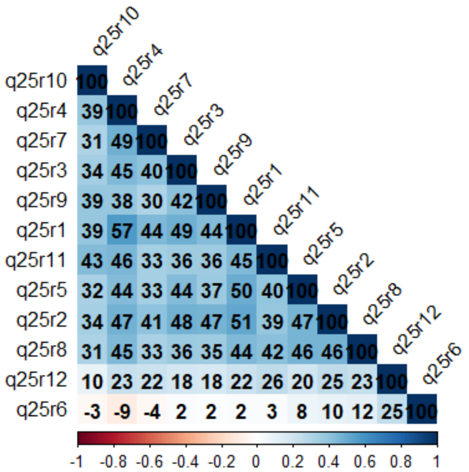




	q24r4	q24r9	q24r8	q24r12	q24r2	q24r10	q24r6	q24r7	q24r11	q24r3	q24r5	q24r1
q24r4	100											
q24r9	55	100										
q24r8	-11	-20	100									
q24r12	-7	-14	50	100								
q24r2	-6	-13	41	36	100							
q24r10	-5	-11	46	54	33	100						
q24r6	0	-1	35	36	37	35	100					
q24r7	-1	-4	47	48	37	45	37	100				
q24r11	-1	-5	43	53	27	51	30	43	100			
q24r3	-6	-2	29	26	43	27	25	31	23	100		
q24r5	11	8	23	21	20	27	17	26	26	27	100	
q24r1	27	29	12	6	12	12	5	11	13	14	21	100

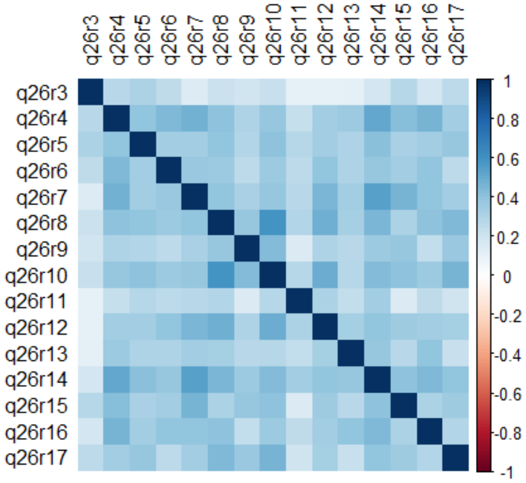
Q25. HEATMAP

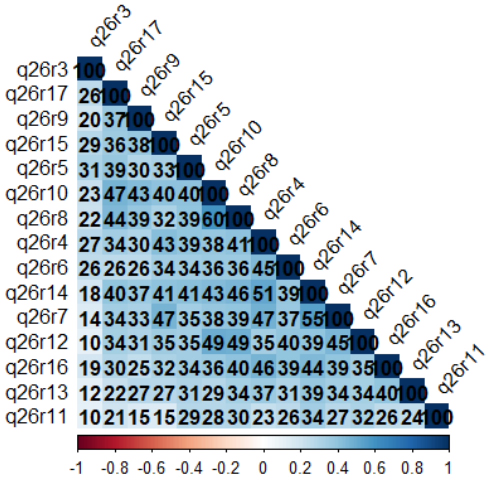




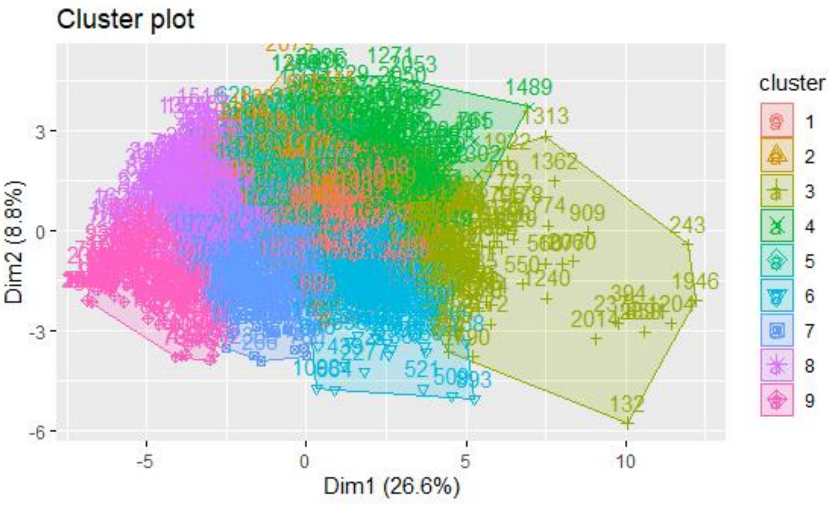
	q25r10	q25r4	q25r7	q25r3	q25r9	q25r1	q25r11	q25r5	q25r2	q25r8	q25r12	q25r6
q25r10	100											
q25r4	39	100										
q25r7	31	49	100									
q25r3	34	45	40	100								
q25r9	39	38	30	42	100							
q25r1	39	57	44	49	44	100						
q25r11	43	46	33	36	36	45	100					
q25r5	32	44	33	44	37	50	40	100				
q25r2	34	47	41	48	47	51	39	47	100			
q25r8	31	45	33	36	35	44	42	46	46	100		
q25r12	10	23	22	18	18	22	26	20	25	23	100	
q25r6	-3	-9	-4	2	2	2	3	8	10	12	25	100

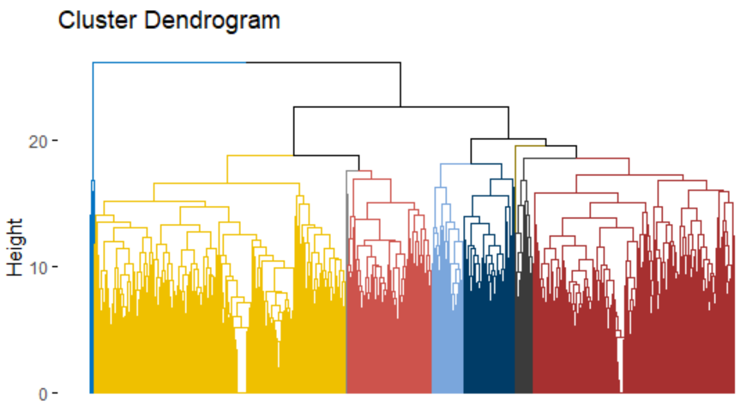
Q26. HEATMAP

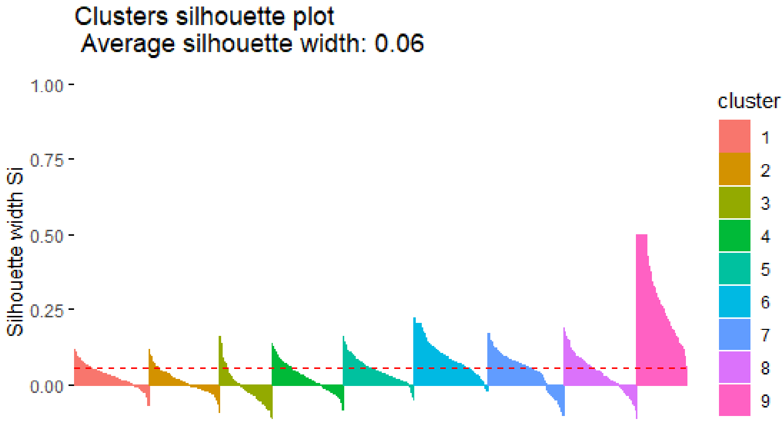




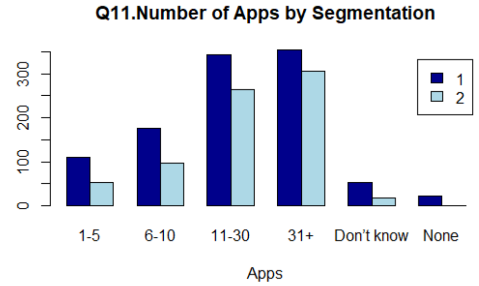
K-means clustering k=9 visualizations

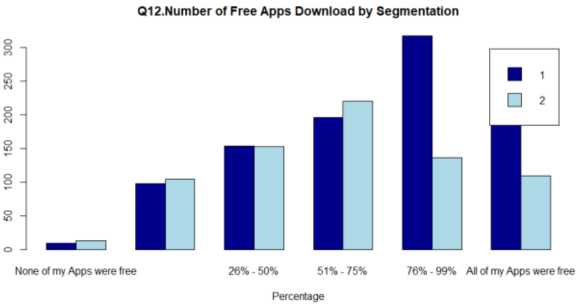






Cluster profile barplot


Q11.Number of Apps by Segmentation
Don't know
None
Apps


Q12.Number of Free Apps Download by Segmentation
None of my Apps were free
All of my Apps were free
Percentage