450 Solo 1 Report

R Sangole

2018-07-19

Introduction

This paper is organized as follows.

Secion I deliniates the overview of the methodologies used and the challenges faced at a high level. It also explains some technical challenges faced.

Section II explains the data preparation activities.

Section III outlines details of t-SNE approach.

Section IV outlines details of clustering approaches.

Section V talks about the Market Segmentation profiling.

Section VI outlines how to perform classification on the model.

Section I - Overview of Methodologies Used

The approach used for this project follows the CRISP-DM methodolgy, iteratively using exploratory analysis work, modeling work and interpretation of the results. First, the data is cleansed - quality checks are performed, few anomalies are corrected, transformations and summarizations are performed. Thereafter, unsupervized dimension reduction is carried out using t-SNE, described in section III. Many insights are obtained from this analysis. Finally 4 types of models are run for clustering, described in section IV. Segmentation profiling is performed and thereafter some discussion on predictive classificiation models.

Section II - Data Preparation

The original dataset available for analysis comprises of responses from 1800 customers for 16 questions. These questions span objective multi-choice demographic questions (age, gender, education, income) to personality and personal preference related questions on a Likert scale. Since the task at hand is to develop an *attitudinal post hoc segmentation*, it's important to first cleanse these data to responses which are relevant in this analysis. Furthermore, it's important to quality check these data against some rules while also addressing cases of missing values. This section describes the modifications made on the original data.

Data Modifications

RULES There are some inconsistencies in the data which were corrected by simple rules.

- Rule A - If q4 r11 is true, it indicates that the respondant doesn't use any apps. If this is the case, then q11 should be None, and q12 should be blank.
- Rule B - To preserve ordinality of q11, 'none' is set to 0, instead of 6
- Rule C - For responses in q11 where the respondant says 'Dont know how many apps', I've set these to NA, so they can be imputed later.
- Rule D - For q12 (% of free apps), there are values missing (when q11 is None), which are set to 6 (All free apps). This will allow these rows to be used in the clustering.

MISSING VALUES Once the rules are applied, There are 99 missing values in q57 and 53 missing values in q11. Imputation is carried out using the `mice` package, which performs multiple imputation using chained equations, using a random forest method.

RECODING A significant amount of recoding was done on most of the questions. For example:

Almost every original question is modified to more usable and succint groupings. For multilevel variables of Age and Income, individual levels are binned together.

- Q13 - Website visit frequency:
  - Social Visit Freq = Average of Facebook, Twitter, LinkedIn and Myspace
  - Music Visit Freq = Average of Pandora, Vevo, AOL Radio, Last.fm and Yahoo music
  - Video Visit Freq = Average of Vevo, YouTube, and IMDB
- Q24 - Technological Sentiments: The 12 questions are summarized into a few attitudinal basis variables:
  - Positive attitude towards technology
  - Entertainment as a primary use of technology
  - Communication as a primary use of technology
  - Negative view of technology
- Q25 - Personality related questions are grouped into 4 main themes:
  - Leadership view of self
  - Risk taker personality
  - High drive towards life
  - Follower
- Q26 - Shopping trend related questions are grouped into five themes:
  - How important are bargains?
  - How important are brands?
  - Do you believe one earns money to spend on oneself?
  - How much do you love apps?
  - Do children influence your purchases?
- Q2 - Platforms - Apple, Andriod, Windows, or Other

SUBSETTING A total of 27 key variables were taken into the analysis going forward. *Table 1* provides details of which questions fall in which set.

TRANSFORMATION Two types of data transformations are investigated:

1. Min-Max Normalization - This makes every variable vary between 0 and 1. Binary variables do not change their values.
2. Standardization - This makes every variable have a mean of 0 and a standard deviation of 1.

It was found that method two gave better results across the board and is used in the final solution.

Section III - t-SNE

t-SNE, or t-Distributed Stochastic Neighbor Embedding is a non-parametric technique to perform dimensionality reduction suited for high-dimensional large datasets. It maintains the underlying structure (local variation) in higher dimensional data while also capturing the macro-structure of the data. t-SNE has been used for visualization in a wide range of applications, including computer security research, music analysis, cancer research, bioinformatics, and biomedical signal processing. It is often used to visualize high-level representations learned by an artificial neural network. Upon application of t-SNE to the full multivariate dataset, we obtain a 2-dimensional representation as shown to the right.

Table 1: Question Sets

AggregatedQuestions	Set1	Set2	Set3	Set4
q1_age	x	x	x	x
q11_appnum			x	
q12_freeapppc		x	x	
q48_edu	x	x	x	
q49_marital	x	x	x	x
q50r1_nochild	x	x	x	x
q56_income	x	x	x	x
q57_mf			x	x
q13_visitfreq_social		x		x
q13_visitfreq_music		x		x
q13_visitfreq_video		x		x
q24_tech_posatt	x	x		x
q24_tech_enter	x	x		x
q24_tech_comm	x	x		x
q24_tech_negatv	x	x		x
q25_prsnlty_leader		x		x
q25_prsnlty_risk		x		x
q25_prsnlty_drive		x		x
q25_prsnlty_follower		x		x
q26_shopsavvy_bargain	x	x	x	x
q26_shopsavvy_brands	x	x	x	x
q26_shopsavvy_earn2spend	x	x	x	x
q26_shopsavvy_applover	x	x	x	x
q26_shopsavvy_children	x	x	x	x
q2_apple	x		x	
q2_andriod	x		x	
q2_windows	x		x	
q2_tablet			x	
q2_other			x	
q4_use_music_apps			x	
q4_use_tv_apps			x	
q4_use_game_apps			x	
q4_use_social_apps			x	
q4_use_news_apps			x	
q4_use_shop_apps			x	
q4_use_none_apps			x	
q54_white	x	x		
q54_black	x	x		
q54_asian	x	x		
q54_hawai	x	x		
q54_native	x	x		
q54_other				
q55_latino	x	x		

This plot can be overlayed with, or colored with the any of the explanatory variables in our dataset to gain insights into the structure of these data. For example, if the plot is overlayed with the variable for race, we can see some remarkably clear distinctions in the data:

- A, B, C = White
- D = African American
- E, G = Latino
- F = Asian
- H = Hawaiian

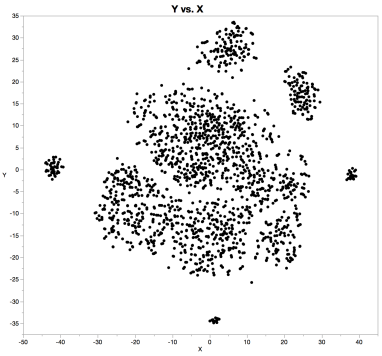
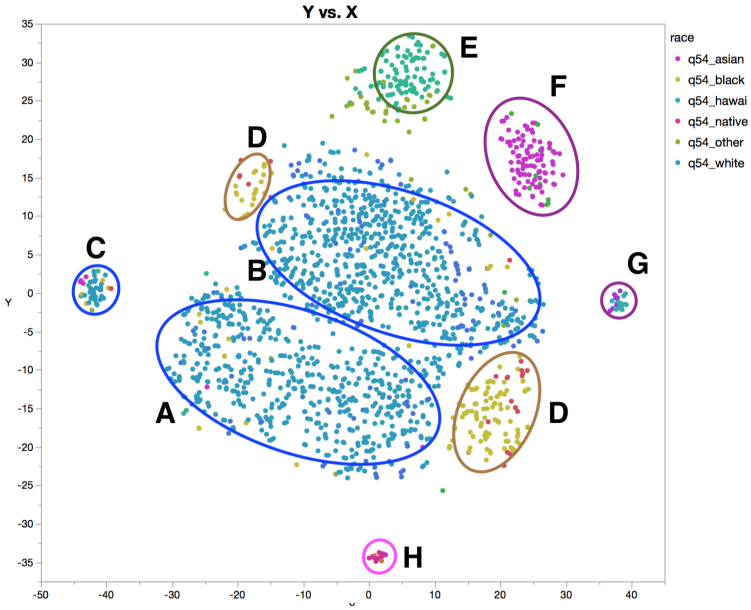


Figure 1: t-SNE representation



There are many more insights which can be quickly sought through analyzing multivariate data through t-SNE. While it's tough to represent all of them visually in a report, this hand representation attempts to explain the clusters observed:

Some of the key takeaways:

- ~ 100% of Asian respondants use Apple devices
- ~ 0% of Black respondants use Windows phones, and ~0% of Black respondants use tablets
- Cluster E - 100% latino cluster is largely an Apple device user, with ~0% windows usage
- Cluster C - 100% of this cluster do not use any apps
- Cluster C is also mostly White, 60 years +, and richer
- A majority of Android users are White
- Everybody plays games regularly, regardless of gender, race, or age

- Younger crowds are more brand aware than older crowds
- A large majority of Asian users are TV related App users, and music related app users
- Irrespective of gender or marital status - brand awareness, app lovers and belief in earning money to spend on oneself go hand in hand
- Most folk think of themselves are thought leaders
- As age increases, folks tend to be more risk averse

Section IV - Clustering Approaches

There are two main categories of clustering approaches tried - non-hierarchical and hierarchical. For the non-hierarchical clustering approach, two types of models are attempted - `k-means` and `pam`. For the hierarchical clustering approach, the `hclust` approach is used with a variety of agglomeration techniques. All three appoaches do require the user to decide the number of clusters. For `kmeans`, the elbow method, the average silhouette width and the r-square values were used to decide the number of clusters. For `pam`, the average silhouette width is used. The the non-hierarchical methods, this number needs to be decided before execution. For the hierarchical clustering approach, the number of clusters can be gauged by visual inspection of the dendogram. For the `hclust` approach, the average silhouette width is used as a deciding factor. Along with the number of clusters, which question-set (between the four sets of questions described in Section II) gives the best possible clustering option needs to be decided. Finally, model based clustering is attempted using the `mclust` method in R.

The final decision was made as a balance between:

- the average silhouette width of the technique,
- the individual silhouette widths of each cluster,
- the balance of number of members in each cluster,
- the amount of variation explained in the first two principal components,
- the explainability of the resulting cluster sets,
- the visual separation of the clusters when plotted on the first two principal components

DISTANCE METRIC For hierarchical methods, a distance matrix needs to be calculated. While for continous variables, the commonly used distance metric is either Euclidean distance or Manhattan distance, when there is a mix of continous and categorical variables, the Gower distance is used, using the `daisy` function in the `cluster` package. The Gower distance was applied to both the continuous data as well as the categorical data (coded as ordinal variables where appropriate, and recoded as dummy variables where ordinality does not make sense).

Euclidean and Manhattan distances were also investigated as options since even the categorical variables are converted to numerics, but Gower did work better.

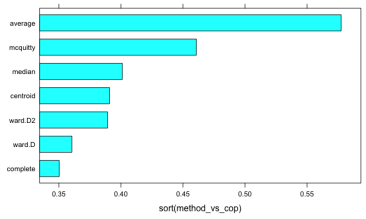
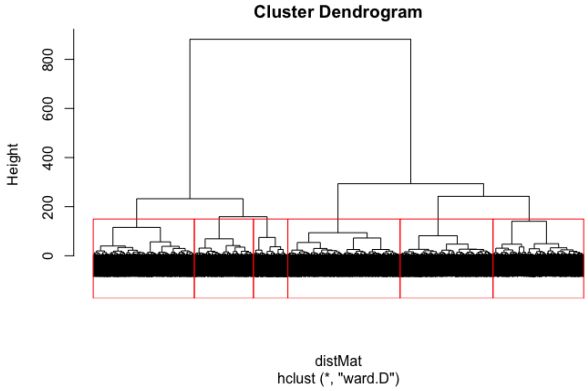


Figure 2: Correlation coeff between the cophenetic and the distance matrix always showed us that average is the best linkage, however it performed the worse when trying to interpret the hclust results

Hierarchical Agglomerative Clustering

HAC is performed using `hclust` on the distance matrix calculated using `daisy`. To decide between the different linkage methods (single, complete, average, ward.D and ward.D2), the dendrograms obtained using each method is plotted and visually inspected. Depending on the question set used, ward.D and ward.D2 outperform all others when judged by the number of balanced clusters seen. Single, complete and average linkaging result in either very long chains, or many clusters with only handful of members. The plot below shows the dendrogram for question set 4 with the `ward.D` linkage method.



The figure to the right shows the clusters for a 6 cluster solution. However, if the average silhouette widths are considered to determine the number of clusters, we get a two class solution.

k-Means Clustering

k-Means clustering was run with two methods -

1. The number of clusters is determined by the elbow method, and the cluster start points were randomly selected, or
2. The number of clusters is determined by the elbow method, and the cluster start points are determined through `hclust`

Depending on the question set used and the transformation used, either method 1 or method 2 gave better separation in the PCA plot. With the final selected model, method 1 worked better.

Partitioning Around Medoids

The performance of the `pam` method varyied greatly depending on the selected question set and transformation. If the average sil width was used as a determining factor, the number of clusters varied from 10 for set 1 to 2 for set 2. Additionally, the PCA plot did not show adequate separation of the clusters.

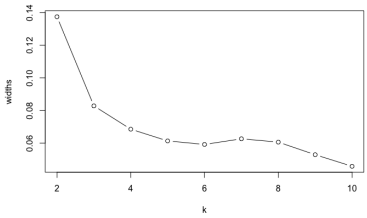


Figure 3: hclust - set 4

From here on, all the graphs are for question set 4 unless otherwise noted

Figure 4: hclust - set 4

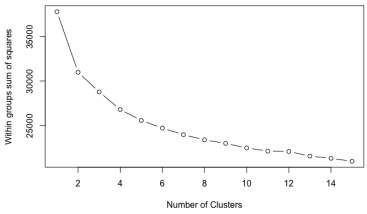


Figure 5: kmeans - set 4 - showing possibly 5 or 6 clusters

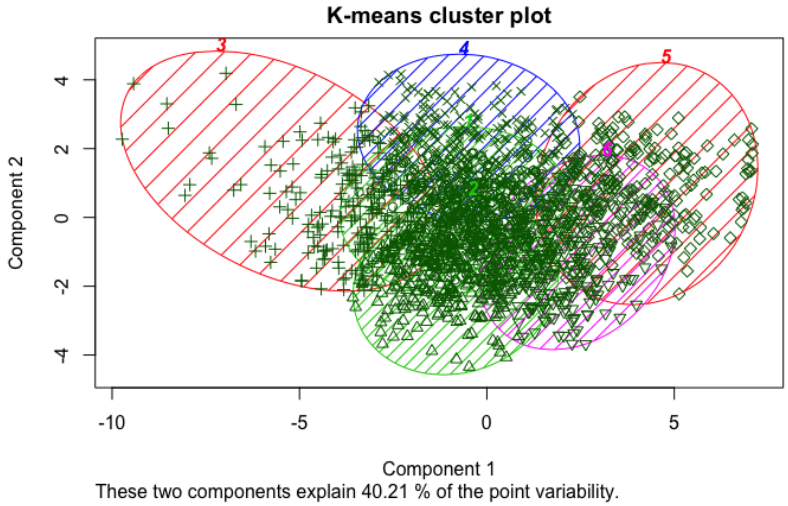


Figure 6: kmeans - set 4 - showing 6 clusters - method 1

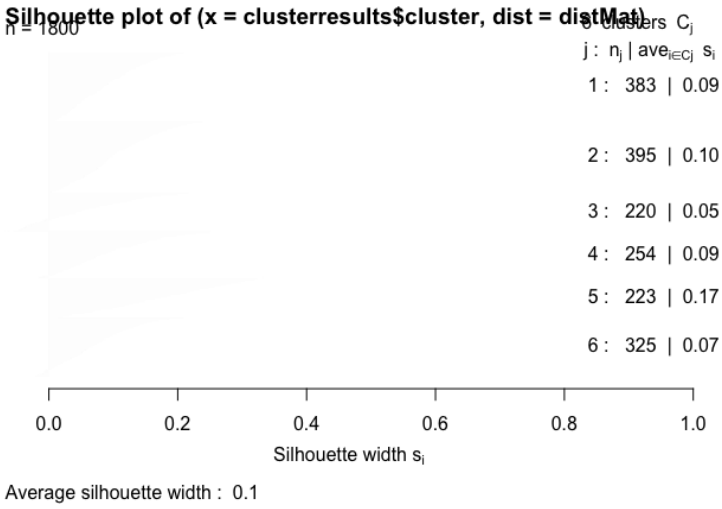


Figure 7: kmeans - set 4 - showing the average sil width for each cluster

Model Based Clustering

The `mclust` method selects the optimal number of clusters during the analysis. This method was the most stable returning an expected number of clusters equal to three no matter which questions set or transformation. Though, this method has the lowest average sil score amongst all other methods. Furthermore, this method resulted in very poor separation on the clusters as seen in this plot.

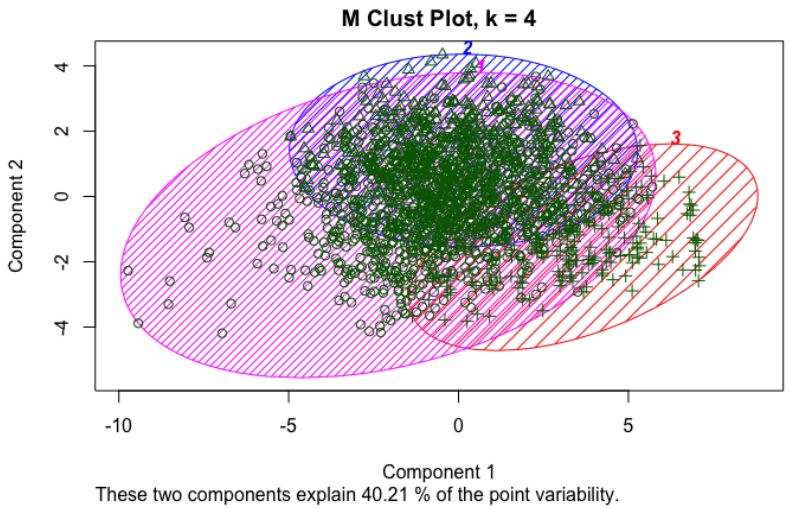


Figure 8: kmeans - set 4 - showing possibly 5 or 6 clusters - method 1

Model Comparison

A total of 16 models are run - 4 types of models across 4 sets of questions.
The critical performance parameters for all models are shown below.

SetId	Linkage	VariationExplained2Pc	HclustAvgSilWidth	AvgSilWidthK
1	ward.D	28.30	0.093	5
2	ward.D2	31.87	0.120	6
3	ward.D2	25.22	0.060	2
4	ward.D	40.20	0.140	2

SetId	Linkage	KmeansElbow	KmeansRsq	KmeansAvgSilWidth	KmeansK
1	ward.D	6	0.270	0.114	6
2	ward.D2	8	0.130	0.133	2
3	ward.D2	6	0.287	0.088	7
4	ward.D	6	0.346	0.156	2

SetId	Linkage	PamAvgSilWidth	PamK	MclustAvgSilWidth	MclustK
1	ward.D	0.100	10	0.09	3
2	ward.D2	0.088	9	0.02	3
3	ward.D2	0.067	5	0.02	3
4	ward.D	0.136	2	0.09	5

Few key takeaways:

1. The percentage of variation explained by the first two components are highest for set 4
2. The average sil width is also the highest for all models for set 4
3. The k-means r squared value is also the highest for set 4 models

Based on these results, the final selected model for market segmentation profiling is the k-means model, using number of clusters selected by the elbow method, using cluster start locations identified by hierarchical clustering using the ward.D agglomeration method. The separation between the clusters is also the quite high as seen in figure 6.

Section V - Market Segmentation Profiling

Now that the clusters are identified, each of the basis variables can be investigated based on these clusters. Visually, these basis variables are investigated using boxplots and histograms to derive some interpretation behind the meaning behind each cluster. Each of the basis variables can be summarized by a mean, for example, as shown in a sample of variables in the table below:

Group1	Q1Age	Q49Marital	Q50R1Nochild	Q56Income
1	1.8	0.1	0.9	2.4
2	2.9	0.8	0.1	3.2
3	2.5	0.4	0.4	2.5
4	1.5	0.2	0.7	1.9
5	1.8	0.4	0.5	2.5
6	2.2	0.5	0.4	2.6

As an example, the top few basis variables which show these patterns:
Examining these basis variables, we can come to a few conclusions:

- Cluster 1 respondents have a child, Cluster 2 don't
- Cluster 5 respondents are the most brand savvy, while those in cluster 3 are the least. Cluster 5 respondents are self aware, shop for the latest brands, and believe that they earn money to spend on themselves.
- On similar lines, cluster 5 are phone app lovers, while those in 3 arn't
- For the most part, all clusters have respondents who view themselves with strong personalities, though cluster 3 and 4 show lower agreement than the other clusters
- Cluster 1 is unmarried while cluster 2 is almost all married respondents

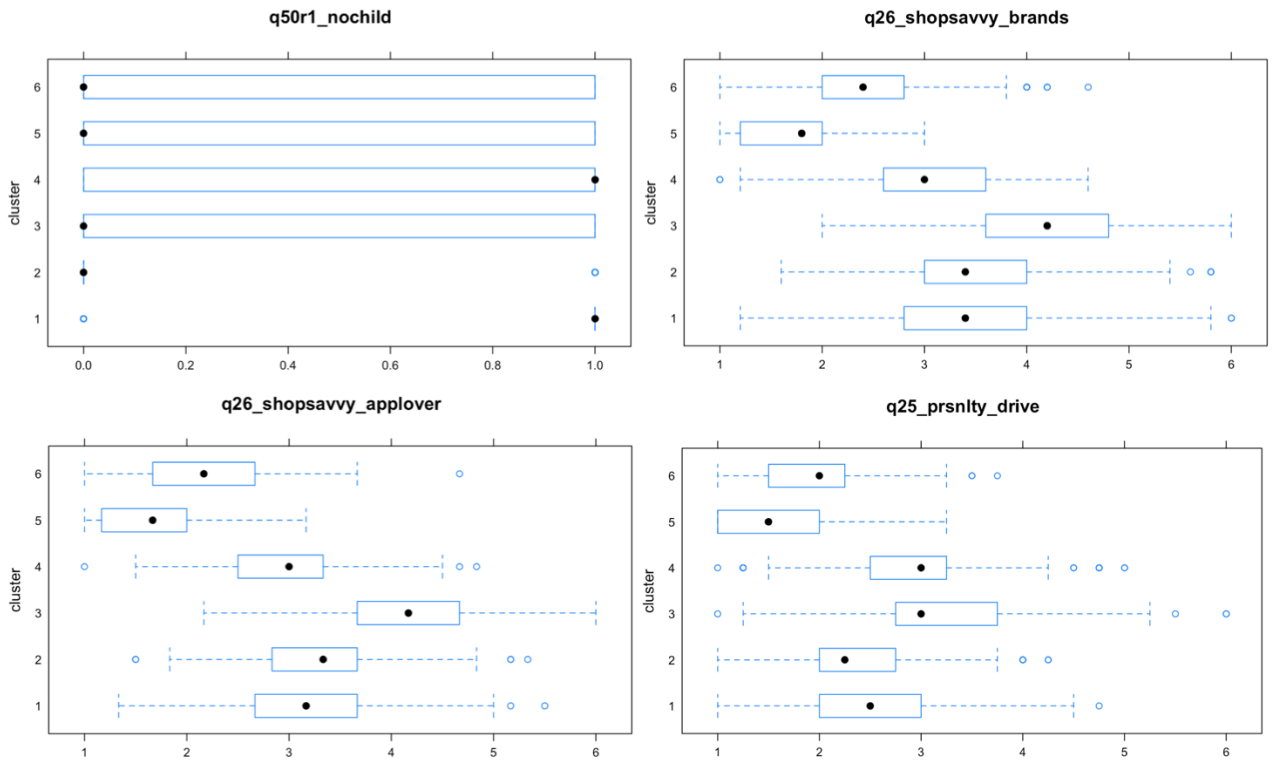


Figure 9: Basis variables boxplots

- Cluster 5 is very strongly a frequent user of phone applications - social, video, music, and for communication. In contrast, cluster 4 has a stronger negative view of technology
- Clusters 4 and 5 are the youngest, predominantly respondents who are 18-24 years old. Cluster 2 is older respondents: 40 - 55 years old.

Going through all the basis variables, we can come up with a few market segmentation profiles in order to develop a customized marketing plan. At a very high level of abstraction, the cluster results can be summarized as follows:

Cluster	Age	Marital Status	Children	Income	App Usage	Attitude towards technology	Personality	Leadership	Shopping Trends
1	18-24	Not married	Yes	Mixed	Mixed	Adopters of new technology & apps, very low negative attitude towards new technology	Moderate views	Trend leaders	Moderate views
2	40-55	Married	No	70k-100k	Mixed	Mixed	Moderate views	Trend leaders	Moderate views
3	35-45	Mixed	Mixed	Mixed	Mixed	Lowest positive attitude towards technology	Moderate views	Trend leaders	Non brand savvy
4	18-24	Not married	Yes	30k-70k	Mixed	Mixed	Moderate views	Mixed	Moderate views
5	18-24	Mixed	Mixed	Mixed	High Social & Video App Usage	Adopters of new technology for entertainment & communication	Highest risk takers, strong personalities & leadership view	Trend followers	Very strong brand savviness & App lovers
6	25-40	Mixed	Mixed	Mixed	Mixed	Highest positive attitude towards apps for entertainment & communication	Highest risk takers, strong personalities & leadership view	Trend leaders	Strong brand preferences, App lovers

Figure 10: Summary Table

RECOMMENDATIONS AppHappy should use these clusters to generate some customized marketing plans. A few key insights are that there is a mixed view towards technology, phone app utilization and adoption. This means that AppHappy's marketing strategy should be a combination of technology-driven and conventional approaches. Furthermore, as shown in the t-SNE analyses, there are very distinct groupings of age, race and phone platform. Thus, depending on the services/products being marketed, the apps developed to market to cluster 5 and 6 can be further customized. Both these clusters are young, driven individuals who are single or married. Cluster 2 is older married individuals who have much higher income than the rest. This group also has moderate views across the board. Group 3 has the lowest positive attitude towards technology adoption and usage. This group needs to be targeting using conventional marketing methods. Combining the results from the clustering approach with the t-SNE investigation, AppHappy can focus on Asian and Latino respondants by developing marketing campaigns which focus on Apple devices, can reach Black respondants by focusing on non-Windows phones.

Section VI - Classification & Predictive Modeling

To develop a classification scheme which AppHappy can use to make predictions on customers for which it doesn't have data for, AppHappy can do some of the following activities.If we can generate any demographic/geographic information about the new customer base, we can analyze the existing known customer base using similar variables, assume that the new customer base follow a similar distribution. These data can be combined with the labels created by the k-means clustering algorithm. A predictive model using various techniques can be developed, viz. statistical methods like multinomial logistic regression models, or simpler machine learning routines like boosted classification trees, especially using `xgboost` which is a powerful computationally light package to develop boosted classification trees. `xgboost` is particularly powerful at handling multiclass problems with class imbalances, redundent variables and small or large data. Where tools like `xgboost` fall short is interpretablity of the model. Techniques like classification trees using `rpart` or simpler logistic regression routines will enable the company to interpret which predictor variables affect the customer classes.

Alternatively, we can survey a portion of the new customers (randomly, or stratified on some key variables like age or income). We hypothesize that the smaller group of surveyed customers represent the entirety of the new customer base. Of course, if the cost of false positive is low, we can always mass market to the full base, while working on generating data for the new customer base.

Section VII - Wrap Up

We applied a number combinations of clustering techniques and distance metrics on the data provided. Market segmentation profiles were developed from a select few questions out of the larger group of original questions. Unsupervised clustering method of k-means along with dimension reduction techniques of t-SNE provide many insights into the groupings of customers upon which AppHappy can make actionable marketing plans.

References

1. https://lvdmaaten.github.io/tsne/
2. Gower, J. C. (1971) A general coefficient of similarity and some of its properties, Biometrics 27, 857–874