INTRO

Segmentation analysis involves understanding customers'/individuals' characteristics, in our case, based on the responses given to the App Happy survey conducted by Consumer Spy Corporation (CSC). CSC's raw dataset consists of 1800 observations and 40 variables, which represent 40 questions asked to each individual. Following a brief exploratory data analysis, we aim to conduct multiple clustering models (having k-Means Clustering being our baseline model) and determine customer profiles based on each cluster's patterns and similarities. Finally, we will recommend a typing tool to oversee the 'next step' to the segmentation analysis.

EXPLORATORY DATA ANALYSIS

The Exploratory Data Analysis has been conducted using the demographic data of each survey responder. The results presented in Figure 1 below show that most of the sample group mainly consists of individuals below or around the age of 40, college-educated and white or Caucasian. We see that the individuals' income range is somewhat scattered and unevenly distributed in addition to these demographic characteristics. The sample group population is single or single with a partner, followed by married, and female populations within the sample group are slightly higher than the male population.

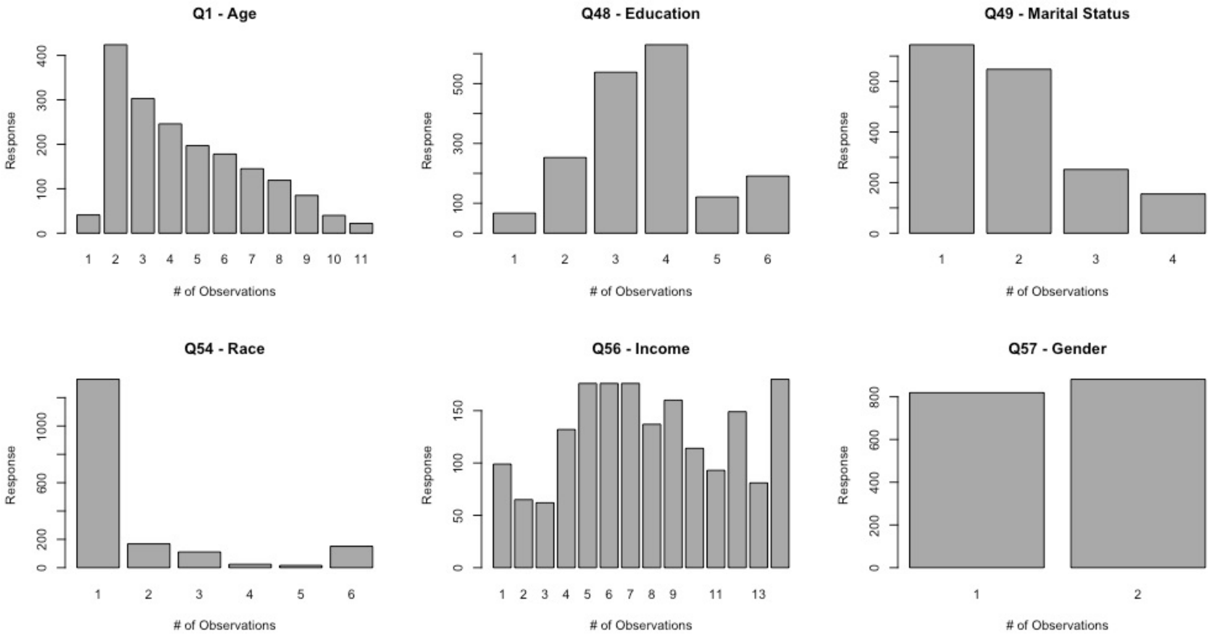

Figure 1: Demographic Data Bar Plots

CLUSTERING

In order to determine the correct number of clusters through the App Happy survey responses dataset, we assessed the Within Cluster Sum of Squares (WSS) value to select the right approach. The Within Sum of Squares is a measure that indicates the compactness of the cluster, and we like it to be as small as possible. The small WSS means that the cluster's data points are close to one another, therefore showing significant similarities.

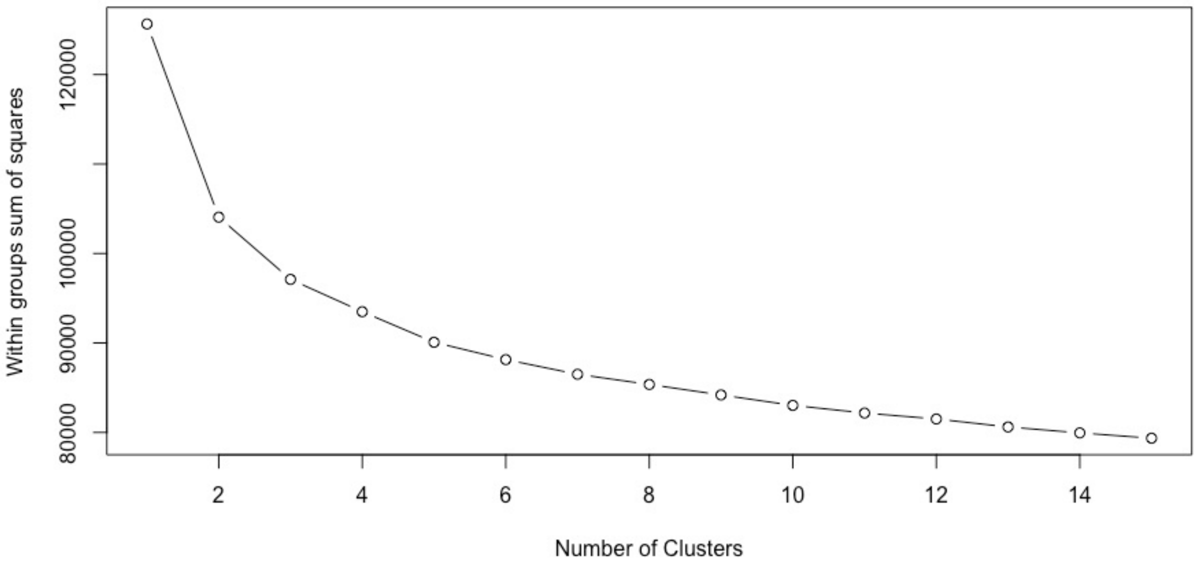

Figure 2: Within Cluster Sum of Squares

The WSS calculated through the survey response data shows WSS measure to be higher than expected. From the graph presented in Figure 2, we see that to achieve 8000 WSS; we need to have about 14 clusters that are less than ideal. For this analysis, we set our threshold to 9000 and move our analysis forward with 5 clusters, with which we get WSS a little less than 9000.

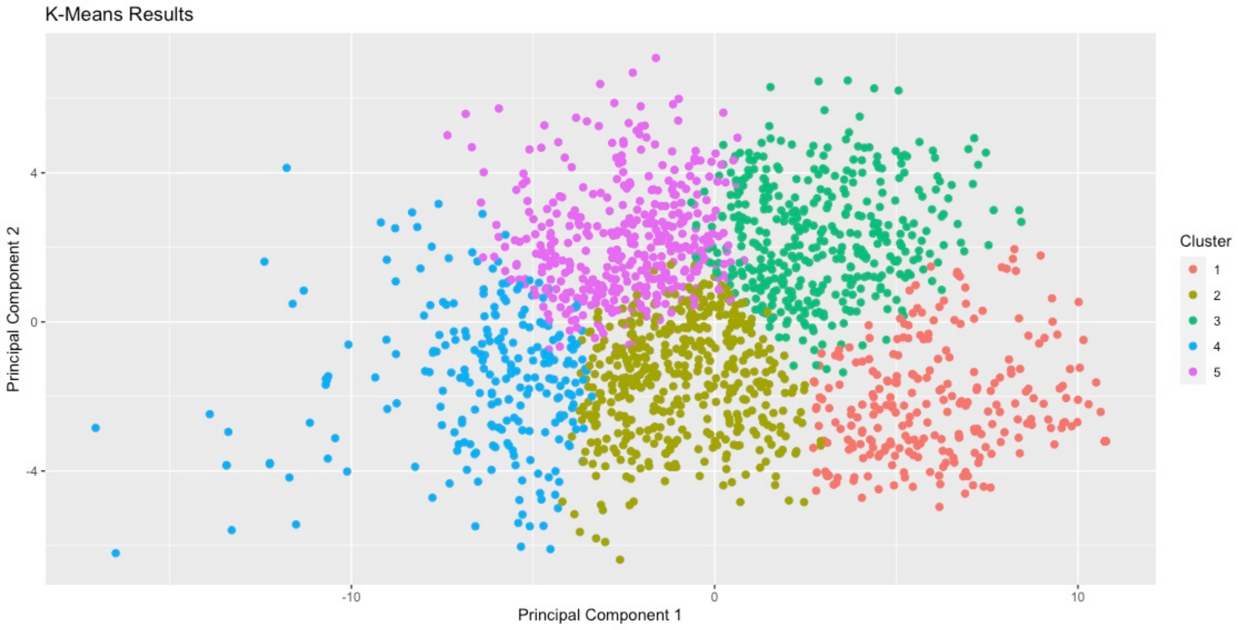

Figure 3: k-Means Clustering

Figure 3 presents the k-Means clustering conducted using all variables under questions 24, 25, and 26 separately. As seen in the plot, there are overlaps between the clusters, some noticeable outliers that indicate individuals who either gave 'strongly agree' or 'strongly disagree' response to all options. Finally, we see that the clusters are not compact and cohesive. In the next section, we will create new features to combine several related responses and attempt to improve the accuracy.

FEATURES

Multiple features were engineered through the responses of questions 24, 25, and 26. Question 24 and 25 were built around obtaining the psychographic information on the individuals related to their lifestyle, relationship with technology, and personality. On the other hand, question 26 was designed to obtain information on the individuals' purchasing behavior, related to their online shopping behavior and whether they are looking for deals and bargains or are willing to spend the extra bucks. These features were engineered using the correlation plot output (see appendix), which presented the variables with high correlations and similarities, therefore indicate possible combinations under the same feature.

The 12 variables of Question 24 were transformed into four. Response options 1-6 were identified as 'positive attitude towards technology', 7-8' interest in music and tv', 10-11 as 'online communication and social media', and 4, 9, and 12 as 'negative attitude towards technology'. The 12 variables under question 25 were also transformed into four features. Response variables 1-5 were identified with 'leadership', 7-8 with 'control', 9-11 with 'drive' and 6 and 12 with negative personality. These featured variables were engineered to have a high-level understanding of the individuals in each cluster and see if individuals with similar personality traits are located in the same cluster. Finally, the question 26 responses group under 5 features as 3-7 being 'bargain shoppers', 8-10 being the ones who like to 'show-off', 11 being people with children, 12-14 being individuals who like to shop for what is hot and trendy, and 15-18 being the group of people who find the brand identity as a crucial aspect dictating their purchasing behavior.

The engineered features have increased the success metrics on the cluster analysis significantly. Through the features, the model is able to cover 80% of the variance using four principal components, which in our initial attempt, pre-feature engineering was achieved through 8 components. The features helped us capture more information in a smaller number of components and better explain the variance in the dataset through being able to determine similarities between observations, which gives us the pattern.

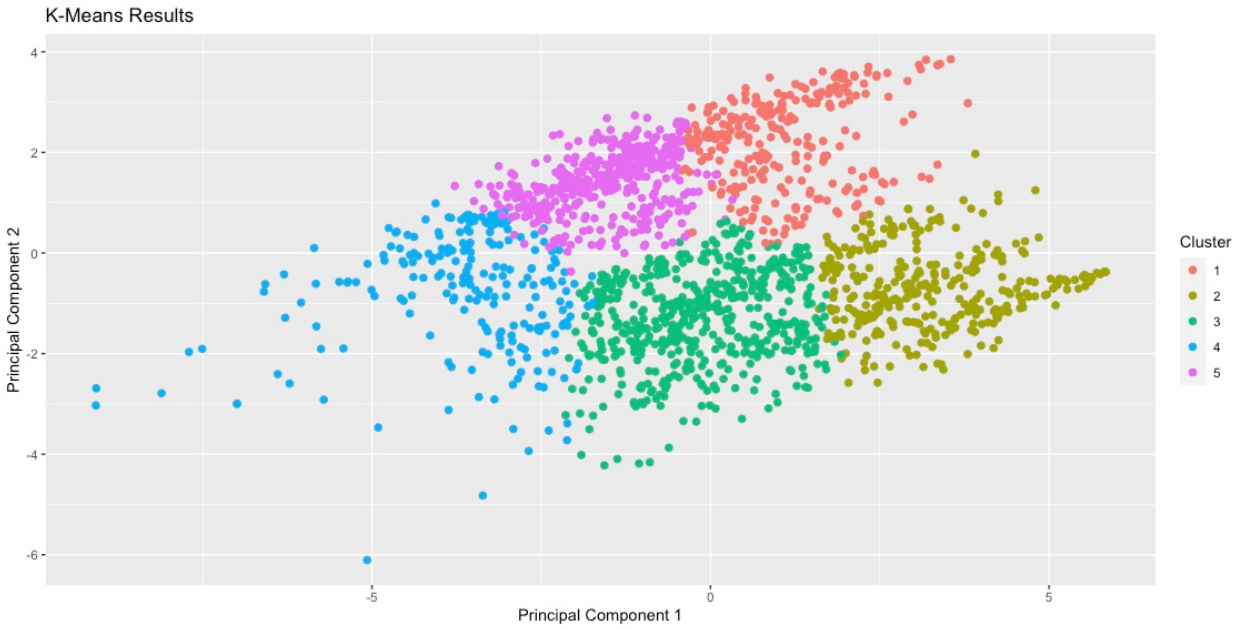

Figure 4: k-Means Clustering using featured variables

As seen in Figure 4, after the feature engineering, the k-means clustering now presents better defined, more cohesive clusters and overlap minimally. The improved clusters and thus cluster silhouettes, which are all above 0.13, are more likely to give us more accurate profiles. We observed the similarities and patterns within each cluster per these psychographic, behavioral features, and demographic variables to recommend four profiles described in the next section.

PROFILES

PROFILE 1 (Middle Age)

Age: above 45
Education: College Educated
Marital Status: Mix of Married, Single and Widower/Divorced
Income Range: 50,000 – 60,000 dollars

Characteristics:

- Don't have a positive attitude towards technology and don't follow the latest advancement and doesn't enjoy purchasing gadgets.
- Don't use social media platforms to communicate or keep in touch with family and friends.
- Don't consider themselves as individuals with leader traits. They are not the decision makers, risk takers and open to trying new things. They also do not consider themselves to be creative, active and optimistic.
- Don't like to shop online, or on the lookout for bargains, deals and promotions.
- They are not likely to spend the extra dollars on extra features or trendy/hot items.
- Their kids do not have any influence on their app downloads.

Summary:

- Not enthusiastic about technology, don't use it every day to save time or to communicate with friends and family.
- Not risk takers. They don't identify as creative and active, are not willing to try new things.
- Not the biggest spenders. They do not like to shop online or spend the extra dollar on what others may find popular.

PROFILE 2 (Young Parents)

Age: 25 - 40
Education: College Educated
Marital Status: Most likely to be married with children
Income Range: 60,000 – 70,000 dollars

Characteristics:

- Have a positive attitude towards technology, follow the advancements, enjoy using technology and buying new gadgets.
- Music and TV shows are important for them and they follow the latest on these using their technological equipment
- Uses social media platforms to communicate with and check in on friends and family.
- They consider themselves as opinion leaders and decision makers.
- They are creative, optimist and active. They are willing to try new things and offer their advice on friends and family.
- They like to shop online and always on the lookout for new deals, bargains, sales and packages.

- They like to spend the money they make on extra features, luxury items, and on what they find trendy/hot.
- Their kids have influence on which apps they download.

Summary:

- Technology enthusiast, likes to have the latest gadgets, follows the advancements. Uses social media to communicate
- Are creative, active and optimistic. They are willing to try new things and advice family and friends.
- Likes to shop online. On the lookout for deals and bargain sales but also likely to spend extra bucks on luxury items and whatever is trendy and luxurious.
- Their kids influence which apps they download.

PROFILE 3 (Single and Child-free Adults)

Age: 35 - 50
Education: College Education, also likely to have post graduate degree
Marital Status: Mostly single or single with partner
Income Range: 60,000 – 70,000 dollars

Characteristics:

- Has somewhat a positive attitude towards technology. Are not enthusiastic about advancements but still are aware of what's happening in tech.
- Uses social media to communicate with friends and family
- Stays neutral in regard to their personality traits. They don't identify to be neither optimistic nor pessimistic. Neutral towards trying new things and being the opinion leaders.
- Do not like to shop online nor they are on the lookout for deals and promotions. They are not likely to buy luxury items or whatever is trendy.
- Their kids do not influence which apps they download. They are likely to be child-free.

Summary:

- Stays connected, likes technology and uses social media.
- Shows neutral attitude regards to personality traits. Do not indicate that they are leaders or are open to new things
- Not online or bargain shoppers. Also, not spenders when it comes to extra features and luxury items.
- Most likely don't have kids.

PROFILE 4 (Young Adults)

Age: 18 - 35
Education: College Education
Marital Status: Mostly single
Income Range: 70,000 dollars + (highest earners)

Characteristics:

- They are technology enthusiasts. Follow the latest advancements, likes to purchase gadgets and overall enjoys using technology and apps.
- Music and TV are important to them.
- Social media users. They use social media platforms to communicate with friends and family.

- They are outgoing. Identify as active, creative and optimistic.
- They are risk takers and like to try new things.
- Online shoppers. They enjoy shopping and on the lookout for deals and promotions, but also are willing to spend on luxury items and whatever is trendy.
- Brand identity is important to them and they are likely to buy from brands they think is close to their style.
- Their kids do not affect their app downloads. Likely to not have kids.

Summary:

- Loves technology, enjoys using tech and follows advancements.
- Uses social media, enjoy music and follows tv shows.
- Open to new experiences, are optimistic creative and active. Are risk takers.
- Enjoy shopping, likely to follow deals but also buy luxury items. Brand identity is important part of their purchase decision making.
- Do not have kids.

Per five clusters obtained through the model, the above four profiles represent the market segments which are Middle Aged Individuals (Profile 1), Young Parents (Profile 2), Single Adults (Profile 3) and Young Adults (Profile 4). The results of the segmentation analysis suggest 4 different profiles thus 4 different marketing approaches for each type of market.

TYPING TOOL RECOMMENDATIONS

Supervised learning methods following the effort on unsupervised clustering are appropriate for continuing the segmentation on an ongoing basis. The new data classification helps the new information be integrated into the existing segmentation model and improve the depth and width of the analysis. k-Nearest Neighbor (kNN) model. The model should use demographic data such as Age, Gender, Income, Marital Status, Education as the variables the kNN model with which the new data will be placed in the relevant clusters in the existing clustering analysis. kNN using demographic features is the recommended supervised learning method as the typing tool for its easy integration to the existing clustering model as well as its simplicity, speed (depending on k), and applicability.

CONCLUSION

In conclusion, we could generate four profiles within the market that App Happy team should use as their target audience through our EDA, feature engineering, and clustering analysis. Each profile has similarities with others and unique selling points, which aims to shape the marketing approach and make the market/audience targeting more precise and accurate. In addition, we recommend the k-Nearest Neighbor methods as the typing tool, which will use the common demographic-related question responses as features to conduct the classification of the new data on an ongoing basis and simplify the integration of the new data into existing clusters.

APPENDIX:

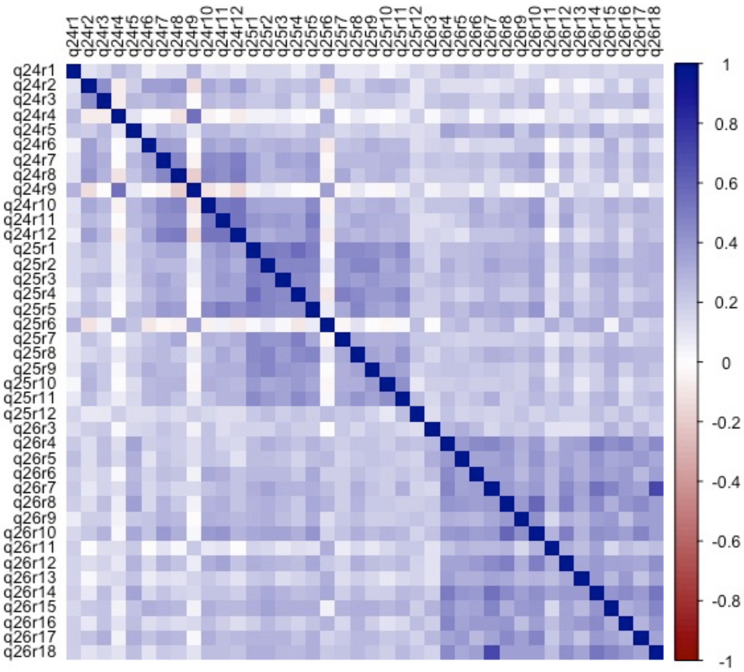

Figure A: Correlation Plot pre feature engineering

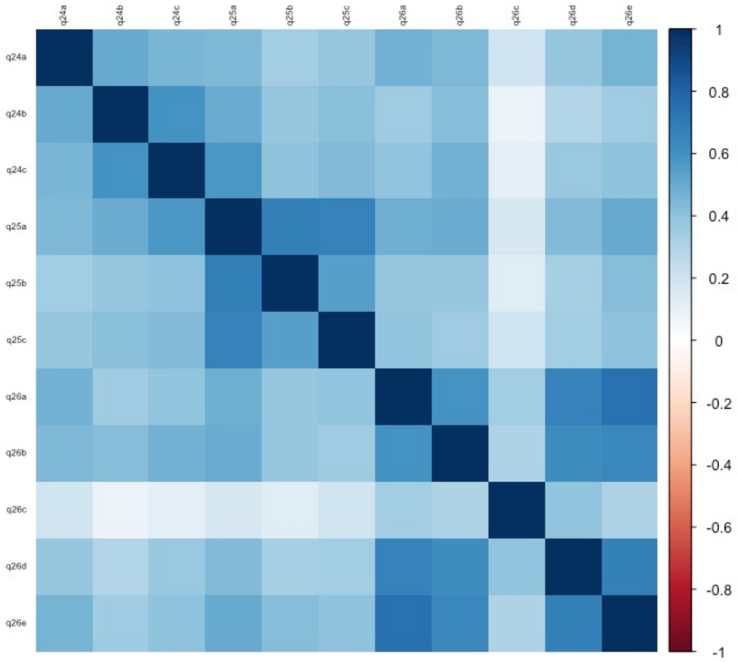

Figure B: Correlation Plot post feature engineering