Market Segmentation and Product recommendation

Sri Seshadri

10/13/2018

Contents

1. Introduction

The App Happy company is currently in the business of providing B2B analytic apps. It is wanting to diversity its product portfolio and is exploring entering new market in the entertainment space. In order to make an informed decision on new market entry, the company surveyed the consumers using a third party surveyor (Consumer Spy Corp). This report discusses the analysis of the survey to explore the segments of customers in the market. The analysis also explores the products in the market based on the survey and touches on potential opportunities for product differentiation. The analysis also explores possible marketing strategy to reach the various segments of the market. However market sizing and market growth rate is out of scope of this analysis. The report discusses the analytical techniques used to segment the markets and use of typing tools to classify new customers on whom there is no existent data. The report also discusses further research to improve the accuracy of typing tools.

2. Executive Summary

3. Data

The survey conducted by Consumer Spy Corp was extensive, however not all question's responses was shared with us. We had access to responses to questions covering the topics of a) demography such as age, race, education, income, sex and marital status etc. b) Tpe of web-enabled devices and the type of apps the consumers use c) Attitude of customers towards new technology, products and personality. Responses to 16 questions were shared with us. The attitudinal questions solicited response in Likert scale. There were 5.5% and 1.33% missing responses for questions on percentage of your apps that were free and gender respectively. Missing data was imputed for analysis.

4. Feature Engineering

The questions and responses surrounding the attitudes of customers were lumped together using the KJ method or affinity diagram, based on common themes. The engineered features and the formula are shown in the table below. The Survey is previded in the Appendix for reference. It is to be noted that the 6th statement in question 25 has a negative conotation compared to the other statements in the question. Hence the score of statement is given an negative weightage, as seen in the last row of table 1.

Table 1: Feature engineering

Feature	Formula
TechAttitudeMean	mean(q24r1,q24r2,q24r4,q24r5,q24r6)
TechUseSocialMean	mean(q24r9,q24r10,q24r11,q24r12)
TechUseMusiclMean	mean(q24r9,q24r10,q24r11,q24r12)
TechUseVideoMean	mean(q24r8)
ShoppingAttitudeMean	mean(q26r3,q26r4,q26r5,q26r6,q26r18,q26r7,q26r13,q26r14,q26r15,q26r16,q24r3)
AppPrefMean	mean(q24r8,q24r9,q24r10,q24r11,q24r12)
LeaderMean	(q25r1+q25r2+q25r3+q25r4+q25r5+q25r7+q25r8+q25r9+q25r10+q25r11+q25r12-q25r6)/12

5. Feature Selection & Assumptions

Looking ahead, the use of technology transcends age, gender and race. The effect of the three demographics are likely to be diminished in the future. For example, there are gaming app to entertain a 5 year old and a 50 year old. It was evident from the analysis of the data that these demographics weres not a dominant factor in the segmentation of the markets. To reduce the noise in the segmentation, the demographics features were eleminated from the data. The responses to questions on the kind of apps the surveyee uses, the frequency of various website visits, the attitudinal questions and income were used for this analysis.

It is assumed that frequency of websites visits is likely due to consumer not liking the experience of an app on their devices or they spend more time on their computer. With the features selected the following questions are sought to be answered.

- what are the consumers consuming on their devices?
- Where is room for opportunities in terms of a new product?
- What is/are the attitude(s) of the customers towards new content consumption?

The feature selection was further refined by correlation filtering methods. It was found that the AppPrefMean (App preference mean) does not add new information to the data due its correlation with other variables.Hene was removed. Figure 1 shows the correlation matrix.

5.1 Data Preparation

The data was standardized and centered to ensure the features were in the same scale for the eucleadian distances calculations for clustering algorithms. Also based on preliminary analysis, the variation in the response to the question "Do you use any of the following kind of Apps?" (Question 4) overwhelmed the information contribution from other selected features. Hence the data was split into two

1. Data with responses to question 4 removed - Used for attitudinal segmentation and prooduct offering opportunities
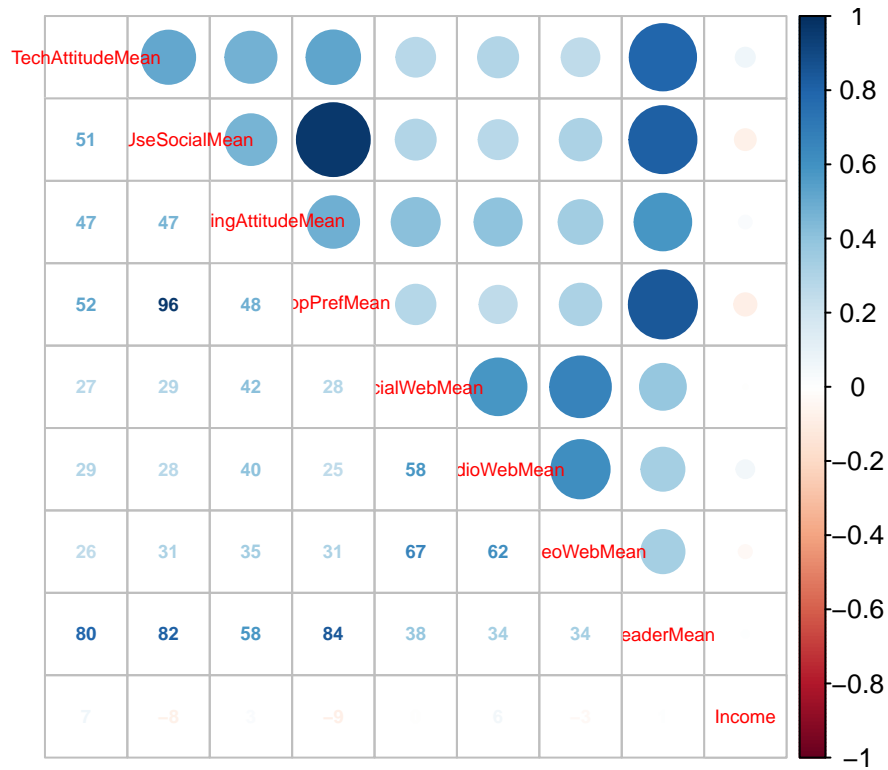2. Data with all selected features - Used for segmenting customers based on use of apps in the market.

Figure 1: Correlation plot

6. Market Segmentation Analysis

In this section the analytical techniques used for market segmentation is discussed and the results are interpreted and recommendations to AppHappy are made regarding product offereing and required further study.

6.1 Attitudinal Segmentation

6.1.1 Hierarchical clustering

Hierarchical clustering is used to determine the inherent structure in the data and used to get a sense on number of segments in the market. 25 bootstrap samples are drawn and dendrograms are cut at height 10 for each of the samples. The height to cut the dendrogram was based on sampled dendrogram. Also the The number of clusters were evaluated based on within sum of squares (WSS) by number of cluster.

As shown in figure 2, a three cluster solution is a good starting point for segmentation.

Clusters in the 25 bootstrap samples – cut height 10
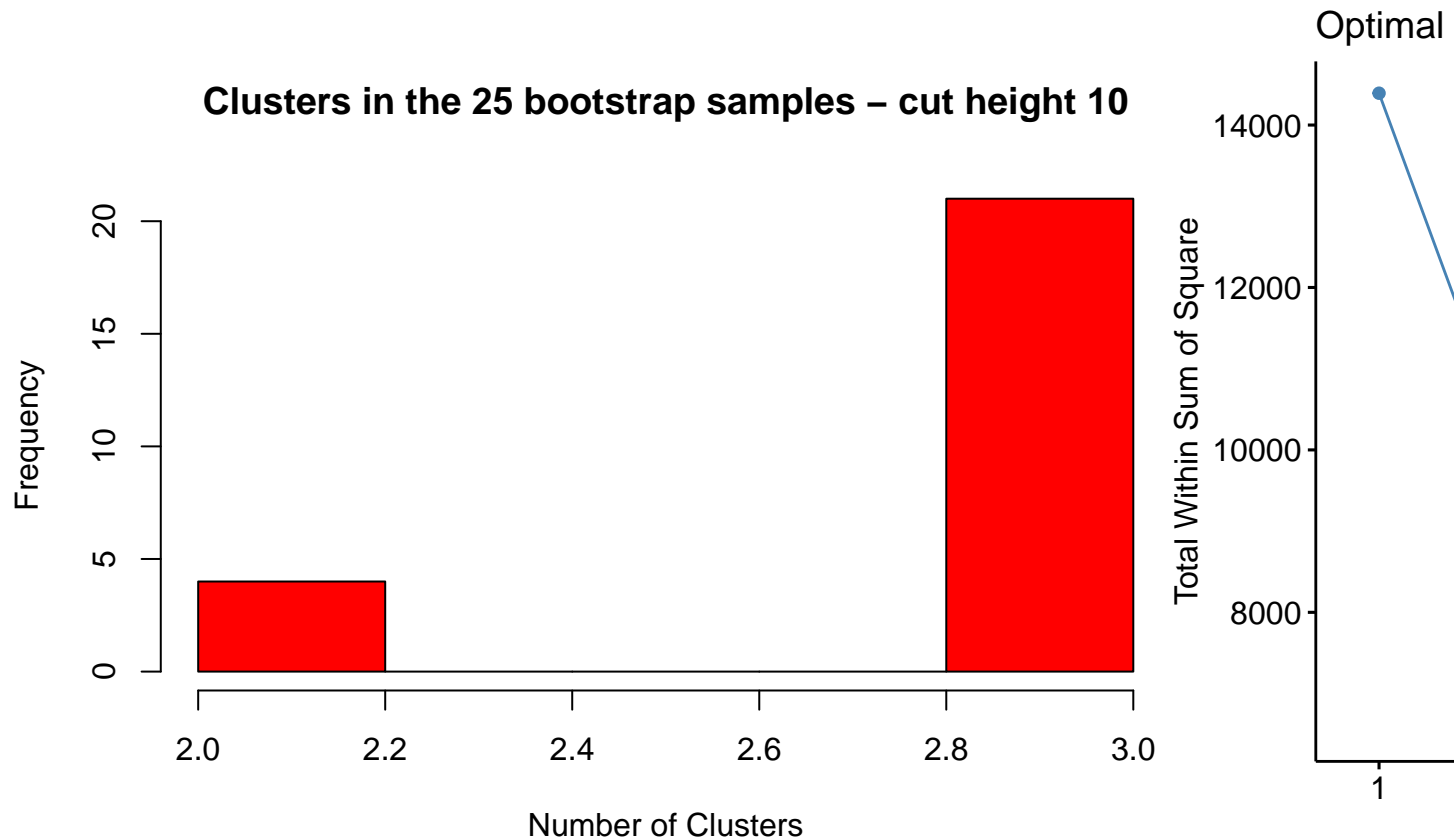
The Hierarchical clustering technique however lumps 92% of the consumers into cluster 1. This is evident in the Sihouette plot shown in figure 3.

6.1.2 K-means

We attempt the K mean clusering with gap statistics to determine the number o clusters to see if it agrees with the observation in the hierarchical clustering techniques.As shown in figure 4, the clusters seem to be a single blob with various densities.

6.1.3 Partition About Mediods (PAM)

We try if there are inherent clusters using the PAM method. Likewise there is a blob of clusters with varying density. We employ the Within Sum Squares metric to assess number of clusters to improve speed of computation for PAM.

6.1.4 DBSCAN

The varying densities lead us to attempt Density based clustering technique to see if there are inherent patterns in the data. Here too we are left with a single blob.

T-distributed Stochastic neighborhood embedding (TSNE)

We attempt to use T-SNE to bring out inherrent clusters in the data and the PAM method is used to cluster the SNE scores. The PCA technique gave a similar result as below.

Silhoutte plot –Hierarchical clusters

n = 1800
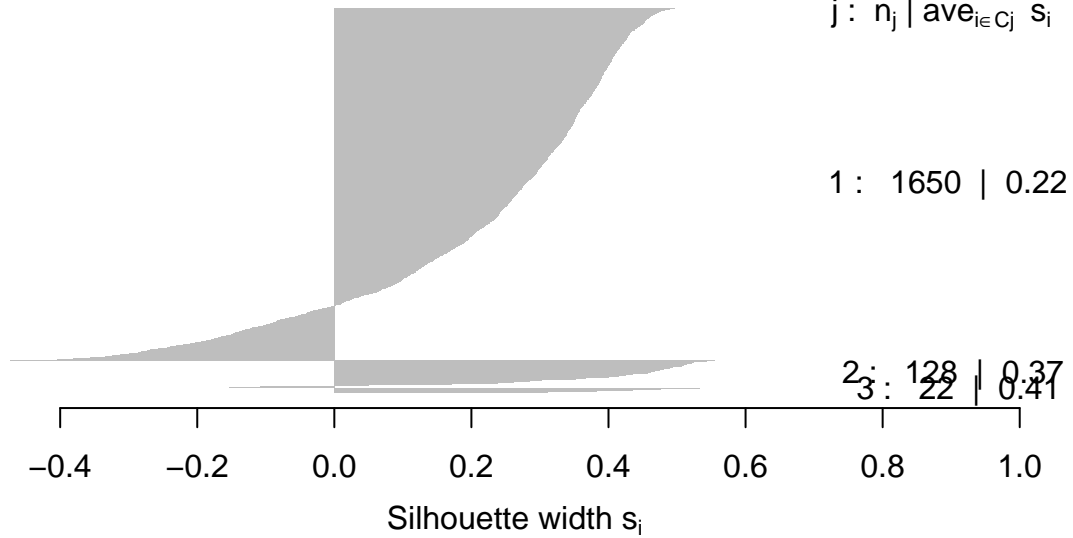
3 clusters C_j

$j : n_j \mid \text{ave}_{i \in C_j} \ s_i$

1 :   1650 | 0.22

2 : 128 | 0.37
3 :   22 | 0.41

Silhouette width s_i

Average silhouette width : 0.23

Figure 2: Cluster purity and samples - Hierarchical clustering

Optimal numbe

Gap statistic (k)

Number of clusters

K – means with 3

Dim2 (17.7%)

cluster
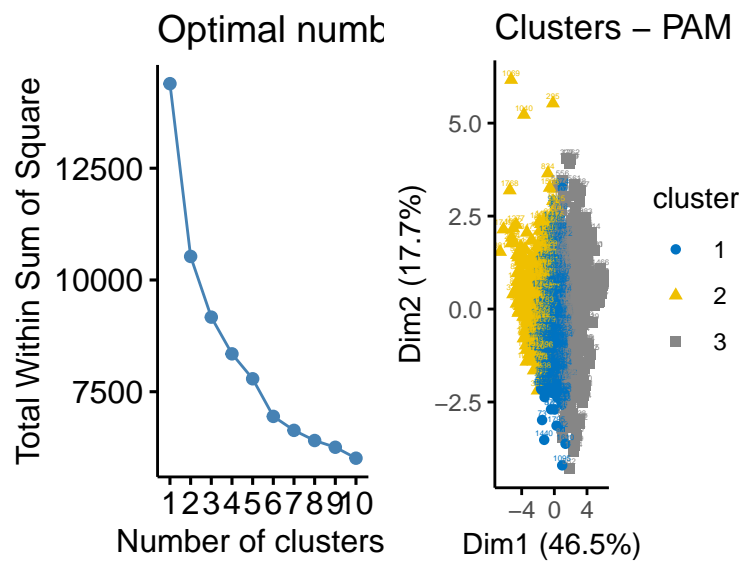
● 1
▲ 2
■ 3

Dim1 (46.5%)
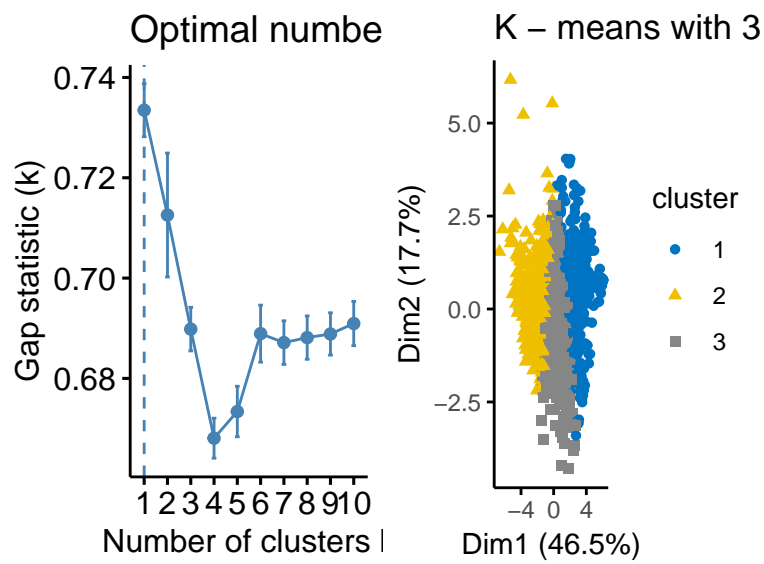
Figure 3: K means clusters
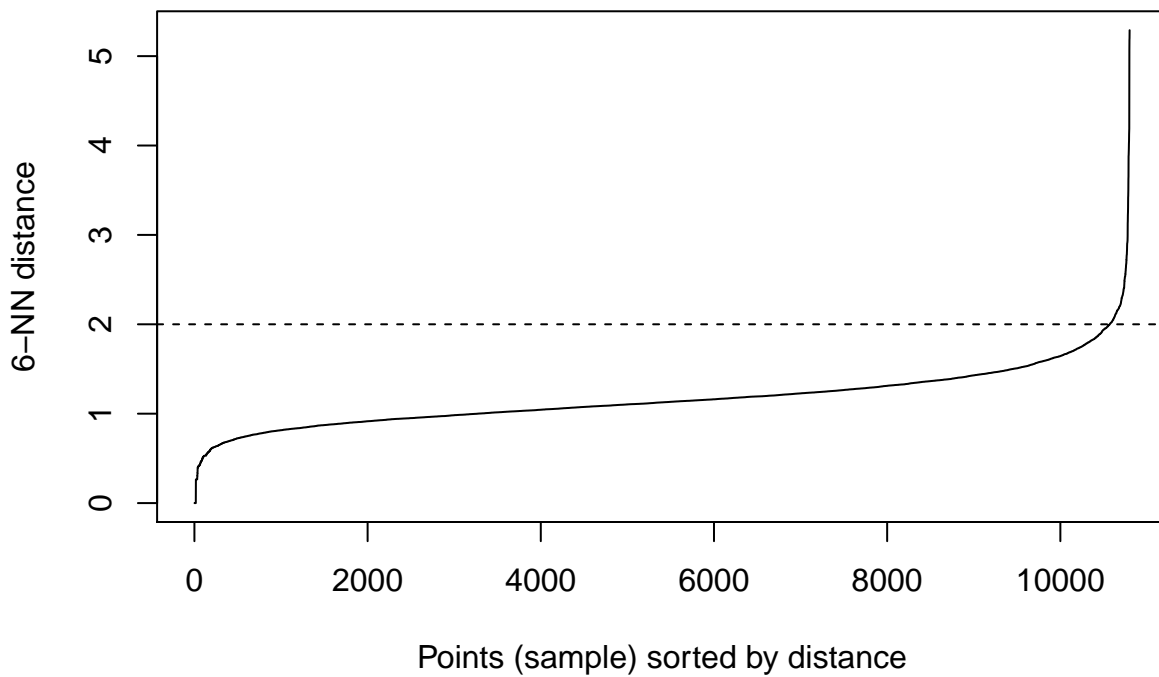
Figure 4: PAM clusters
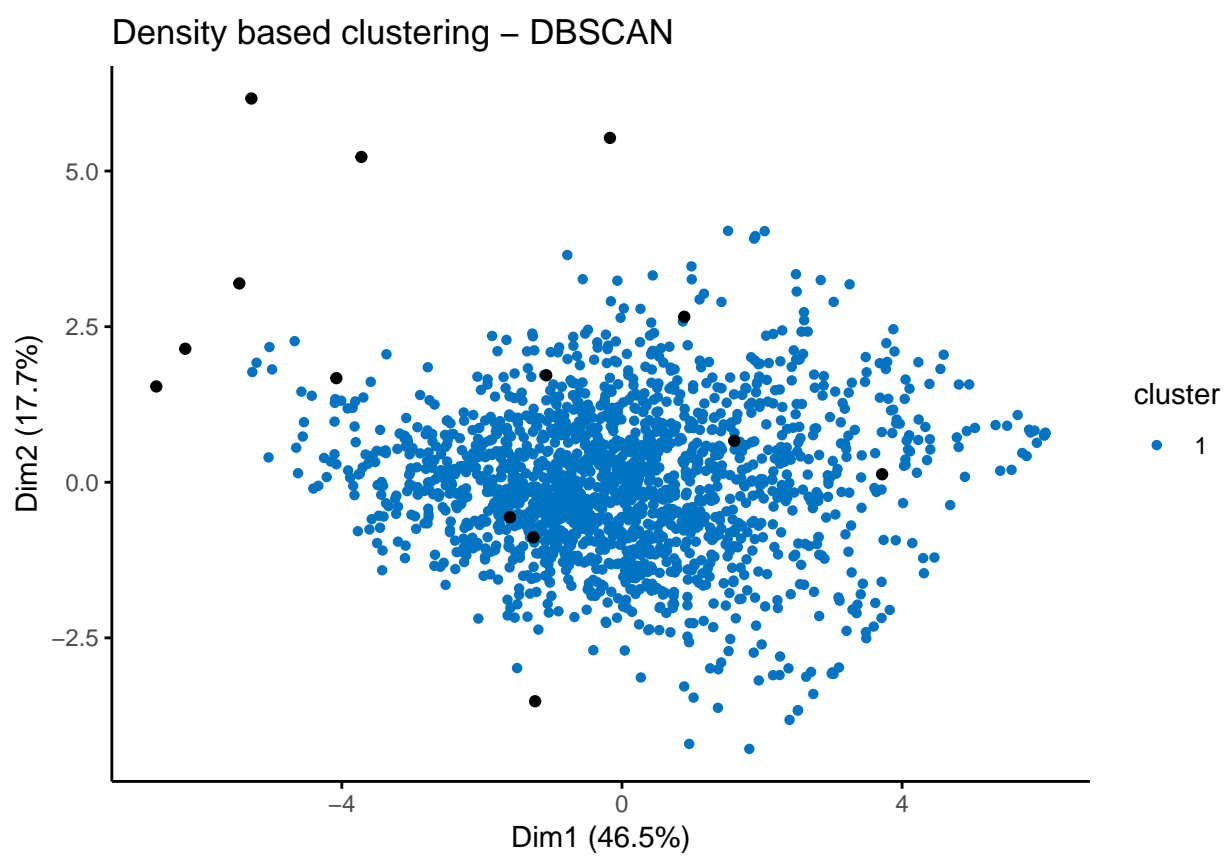

Figure 5: Choice of eps for DBScan
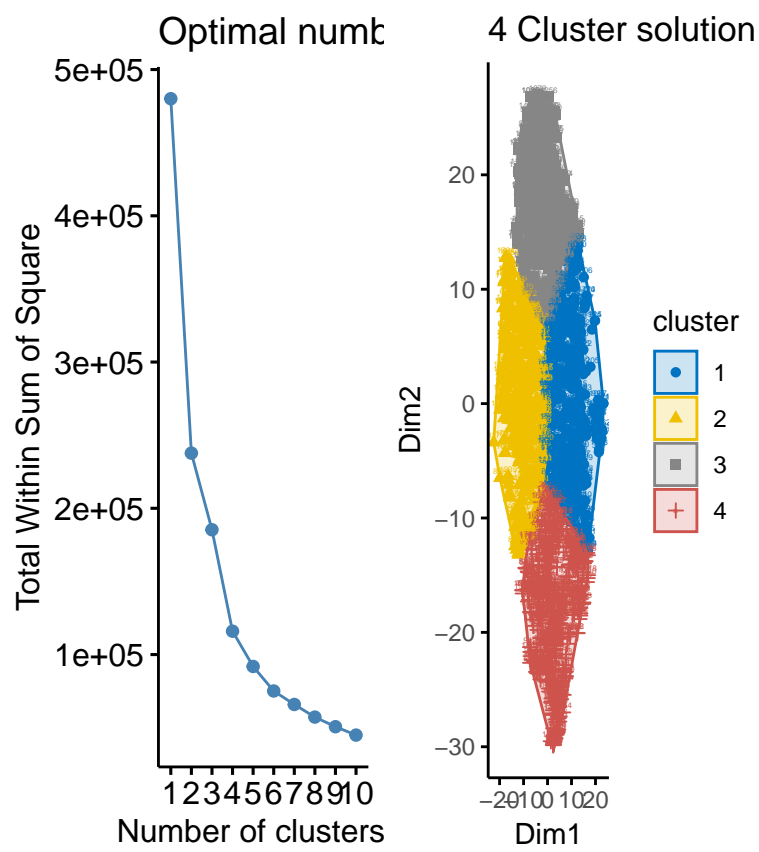
Figure 6: DBSCAN Clustering

Figure 7: tsne scores clustering- PAM

6.2 Attitudinal Segmentation interpretation

The clusters are profiled by using decision trees to augment proper interpretation of the customer segments.
The decision tree is shown in the below figure. The interpretation of the segments is summarized in the
below.

Segment	Percent_Surveyee	MarketingStrategy	Clu
Busy,Upper middle class shopper	29.5%	Internet Radio	
Middle income,late adopter,social media user	23.9%	Ad spaces on social media video page	
Above average income late adopter, web video consumer	24.2%	Ad spaces on social media video page	
Non social follower	22.3%	Unknown	