**Customer Segmentation using K-means Clustering**

Qussay Al-Qirim

The aim of this report is to use k-means clustering on a dataset composed of the mall’s customers. As described by Hartigan & Wong, the purpose of k-means is to minimize the sum of squares within a cluster (1979); thereby reducing the variation within each cluster.

The dataset at hand provides information about two hundred customers that shop at the mall. Missing values were omitted and the Gender variable was converted to binary integers, with zero representing male and one representing female. CustomerID was dropped as it was not relevant to the cluster analysis. The average customer is 38 years old with an annual income of $60,000 and a spending score of 50.

**Cluster 200 customers to 2 group based on their demographic and purchasing record.**

Through k-means clustering where k = 2, the customers were grouped into two clusters. Described by figure 6, the age group in cluster one was younger than in cluster two. The annual income and spending score in cluster one was much higher than in cluster two. However, there was a similarity that both clusters were female majority.

**Choosing two variables and visualizing the clustering.**

Annual Income and spending score were chosen as displayed in figure 7, the spending score for cluster 1 is much higher than cluster 2 regardless of annual income level. Indicating that younger aged people are less sensitive to price with respect to their level of income in their purchase decision making.

**Describe a buyer persona for each cluster**

Cluster 1 Persona Name: Young & Motivated

Background (Figure 6): Female majority comprised of youthful individuals with an average age of 29 years. They have an above average annual income at $62,000 followed by a high spending pattern with an average score of 73.62.

Story: These customers are young, motivated females in the prime of their career. Successful, wealthy and enjoy lavishing themselves.

Goals: To continue their career success and increase their wealth.

Challenges: Increasing workload pressure and quality expectations.

How to connect: Through social media.

Opportunity: Provide female focused products and services to ease the work pressure such as Gyms, restaurants, and cafes.

Cluster 2 Persona Name: Family

Background (Figure 6): Female majority of middle age averaging at 46 years. Their annual income is below average at $59,000 with a low spending pattern with an average score of 32.

Story: Customers in this bracket are middle aged family-oriented women. They are most likely married and are with children.

Goals: To raise the family and focus on wellbeing of her children.

Challenges: Very little time for self-care

How to connect: Word of mouth, TV and social media.

Opportunity: Provide more products that are focused on children’s wellbeing (i.e. clothes).

**How about 3, 4 or 5 clusters? Visualize the difference.**

As displayed in figure 8, there are overlaps between the groups when clustering 3,4 and 5 groups. As the number of clusters increase, the overlap between each cluster equally increases. Four clusters represent a tipping point where different customer segments are properly represented. After which, differentiating the customer segments becomes more difficult as observed with five clusters.

**What is the optimal number of clusters?**

Through the Elbow method as observed in figure 9, four clusters would be the optimal number of clusters. Furthermore, the descriptive statistics in Figure 10 show clear differences in the variables’ averages between four clusters. Whereas the opposite with five clusters. As such, the cluster visuals coupled with their descriptive statistics and further supported by the Elbow method show that four clusters would be the optimal number.

**Conclusion:**

In conclusion, k-means was used to identify cluster groups and their respective averages were analysed to better understand the buyer persona. Furthermore, the elbow method was used to further support the optimal number of clusters.

**References:**

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1), 100. https://doi.org/10.2307/2346830

‌

**Appendix:**

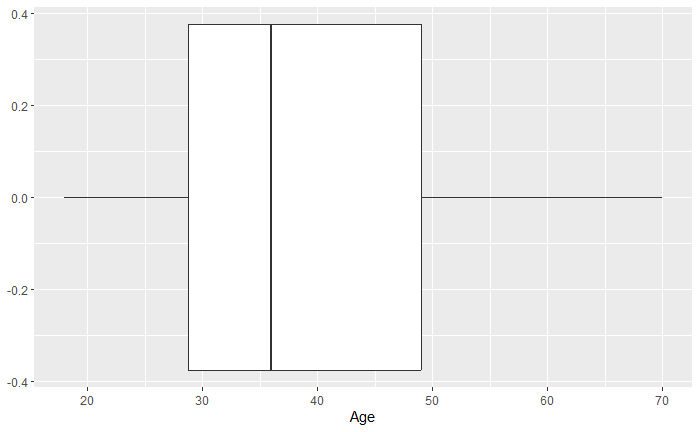


Figure : Boxplot Chart of Customers Age

Chart, histogram

Description automatically generated

Figure :Boxplot Chart of Customers Annual Income

Chart, histogram

Description automatically generated

Figure : Boxplot Chart of Customers Spending Score

A screenshot of a computer

Description automatically generated with low confidence

Figure : Male to Female count comparison (0 representing Male and 1 Representing Female)

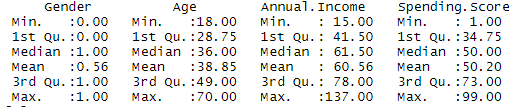


Figure : Descriptive Statistics of Customer Dataset

Text

Description automatically generated

Figure : Average value for variables of the 2 Clusters

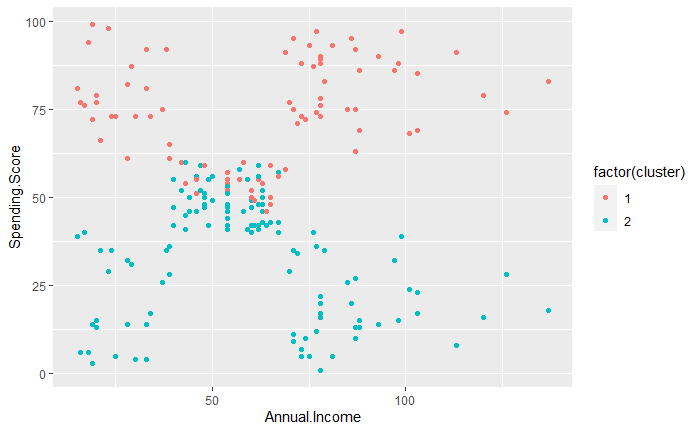


Figure : Visualizing the two clusters on the dataset using Spending Score and Annual Income Variables

A picture containing chart

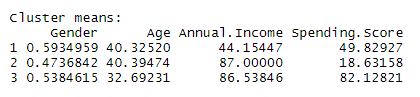
Description automatically generated

Figure : Visualization of 2,3,4,5 clusters in a grid

Chart, line chart

Description automatically generated

Figure : Optimal Number of Clusters using the Elbow Method



Text, letter

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Figure : Descriptive Statistics of Clusters 3,4,5