Segmenting Supermarket Customers into K-Clusters

I. Introduction

The goal of this assignment is to preform clustering operations using Orange on a suitable data set. The data set submitted for this assignment is a set analyzing customers of an anonymous supermarket. The metrics provided within the data set will be used to determine whether or not the customers of this supermarket can be well sorted into clusters which a marketing team could use to better understand the customers of the supermarket. Through the use of clustering, we can segment the total set of customers which will help the marketing team better target their advertisements or promotions. This is a common application of clusters that this assignment will explore on a much smaller scale.

The future sections of this report describe the dataset, the methodology, results along with a discussion, and a conclusion. Section II contains a description of the dataset used for this analysis as well as a detailed table of all the attributes alongside a histogram of the three major attributes. The methodology for analysis is presented in section III. In section IV contains a report and discuss of the results. Finally, section V provides conclusions.

II. Data Description

There are five columns in total that our data set provides are depicted in Table I. The first will be a unique identifier that records the individuals. The next two are common attributes used to characterize the individuals which are Gender and Age. The fourth attribute is an important numeric attribute and that is Annual Income, which will play an important role in helping us create clusters. The data is represented by only counting in the thousands as explained in the table below. The last and arguably most important attribute is the Spending Score which is determined for us as by the supermarket as an interval between 1 and 100. Using distinctive attributes such as Age, Annual Income, and Spending Score, we will create meaningful clusters later on in sections II and III.

TABLE I.

Attribute	Type	Example Value	Description
Customer ID	Nominal (primary key)	4	Record identifier
Gender	Numeric (binary)	1	Age of customer Male = 0 Female = 1
Age	Numeric (integer)	23	Reported age
Annual Income ($k)	Numeric (integer)	16	Annual Income in thousands. Ex: 16 = $16,000
Spending Score (1 – 100)	Numeric (Interval)	77	Score assigned to customer based on behavior and purchasing data

The attributes recorded in Table I can be better understood when their frequencies are analyzed. Using histograms, the distribution of data collected for Age, Annual Income, and Spending Score is easily visualized. Figures I, II, and III of page two display such data.

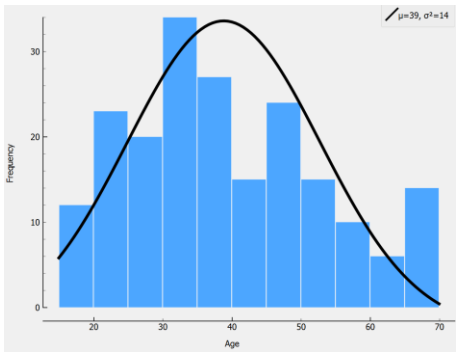

Fig. 1. Frequency distribution of Age

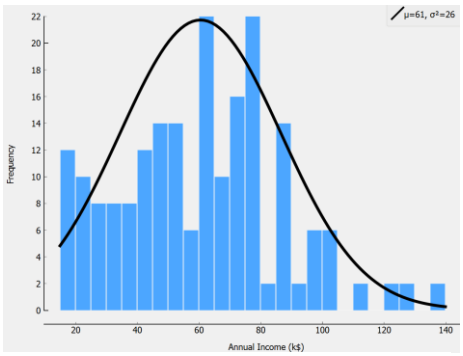

Fig. 2. Frequency distribution of Annual Income

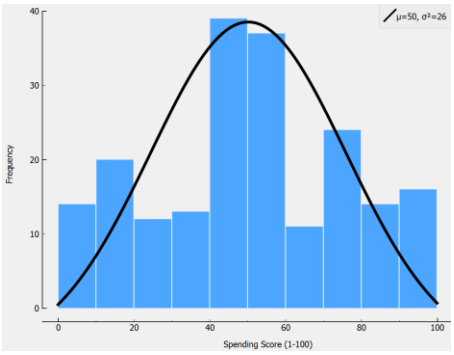

Fig. 3. Frequency distribution of Spending Score

III. Methodology

The data collected by this anonymous supermarket was analyzed in the following ways. First, the possible number of clusters was determined by their silhouette scores and the cut off for the possible number of clusters was set once the difference in scores became less distinguishable. From this method, similar to an elbow method, the number of possible meaningful clusters was determined to be between three and six clusters. Second, scatter plots were designed for each of the four possible number of clusters. These scatter plots used Annual Income and Spending Score as their x and y axis. Cluster sizes three and four, as shown in Figures 4 & 5, displayed clusters which provided distant groups that were too large and the scatter plot of six clusters, in Figure 7, displayed indistinct groups. As such, the K-Means cluster of size five proved to be the most optimal number as shown in Figure 6. Finally, after the cluster size was determined, box plots were used to determine the centroid or average individual profile of each cluster based on Age, Annual Income, and Spending Score. The results of which will be discussed in section IV.

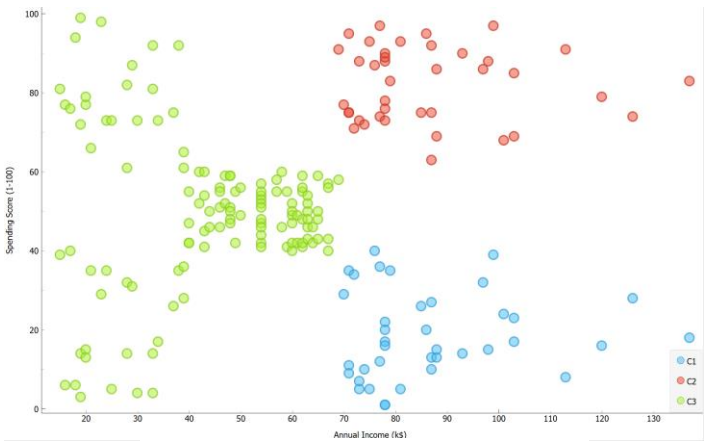

Fig. 4. Scatter Plot
Spending Score vs. Annual Income
3 Clusters

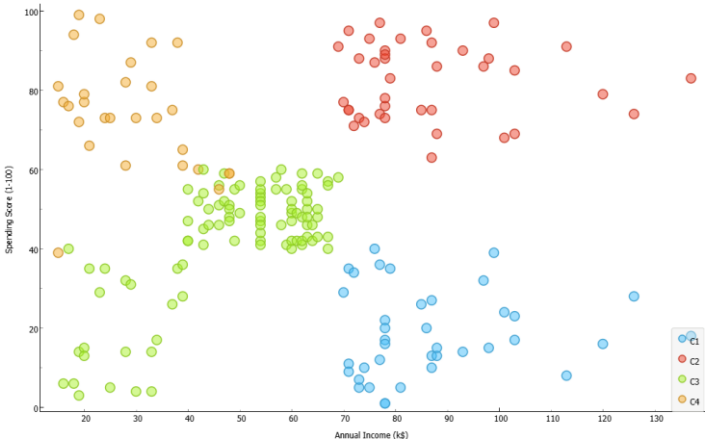

Fig. 5. Scatter Plot
Spending Score vs. Annual Income
4 Clusters

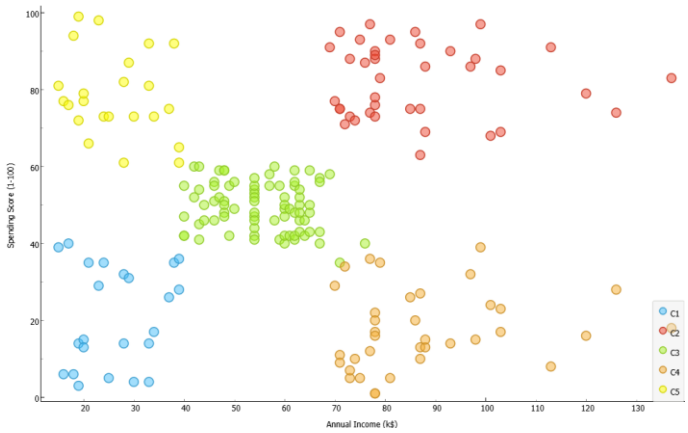

Fig. 6. Scatter Plot
Spending Score vs. Annual Income
5 Clusters

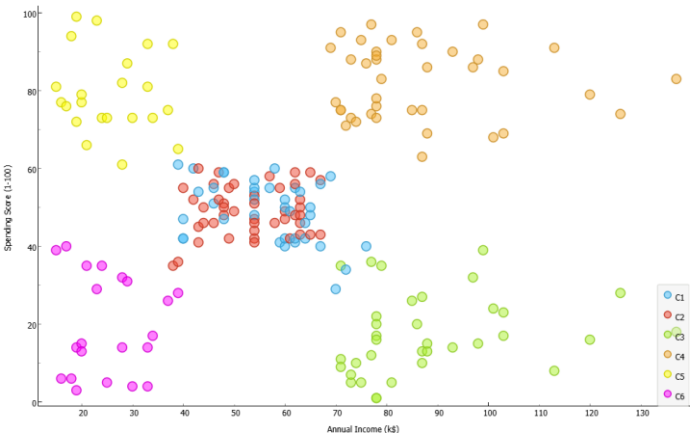

Fig. 7. Scatter Plot
Spending Score vs. Annual Income
6 Clusters

The scatter plot in Figure 6 was chosen as the most optimal numbering of clusters. From this scatter plot, additional information was extrapolated by applying a color map to visually determine the boundaries of the five clusters and a color scheme was applied based on Age, replacing the cluster coloring. These plots are illustrated in Figures 7 & 8 respectively and their significance will also be discussed in Section IV. Scatter plots were purposefully chosen as the chief plotting tool because of their ease in visualizing groups of data points and box plots were chosen as they were a good match for determining the numerical significance of each cluster.

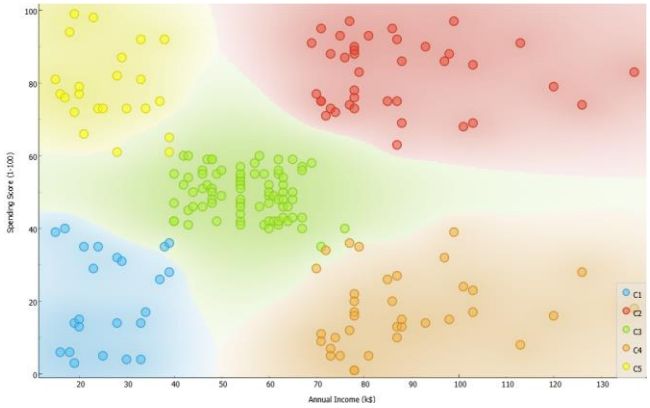

Fig. 8. Scatter Plot
Spending Score vs. Annual Income
5 Clusters Color Map

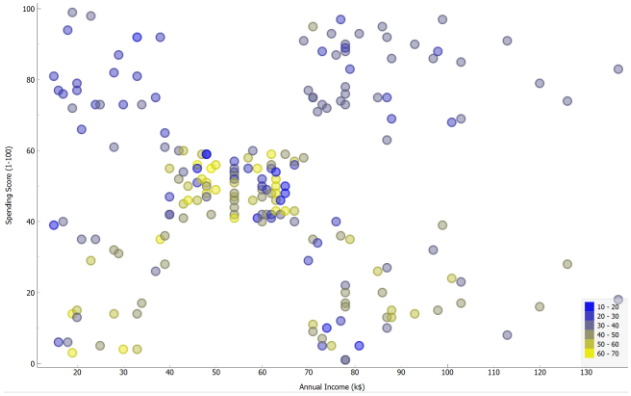

Fig. 9. Scatter Plot
Spending Score vs. Annual Income
Colors Representing Age

IV. Results and Discussion

Figure 8 shows that the regions of these clusters are well defined with little overlap between the clusters. This gives a good degree of confidence to the membership of a data point belonging to a certain cluster. Also, Figure 9 shows us that few of the customers who have Spending Scores over sixty are above the age of fifty. Even though the number of customers who are above the age of fifty is less than the number of customers who are below in the age of fifty, as shown in Figure 1, the near complete exclusion of customers above age fifty who can boast a high Spending Score may imply that the older a customer is the less likely they are to possess a high Spending Score.

With these points in mind we turn our attention towards Figure 6 and the inferences that may be made from the scatter plot of clusters. Clusters C1 and C4 are clearly split apart from one another by their levels of Annual Income and yet they have very similar low Spending Scores. The flip side is true for clusters C2 and C5 who are also clearly distinguishable from one another by Annual Income levels and yet both groups have high Spending Scores. The members belonging to the remaining cluster C3 have both an average Annual Income level and an average Spending Score.

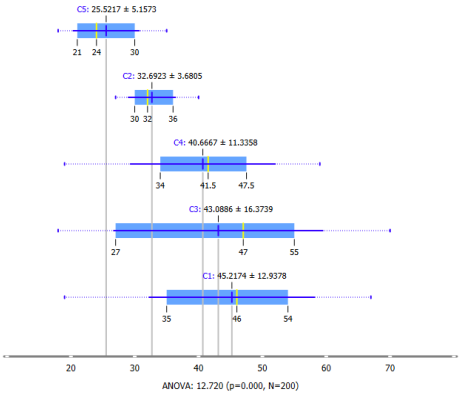

Fig. 10. Box Plot
Age Distribution of Clusters

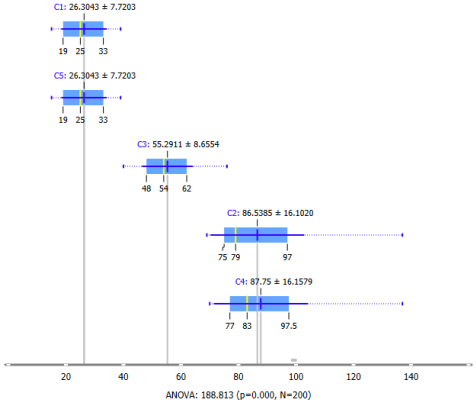

Fig. 11. Box Plot
Annual Income Distribution of Clusters

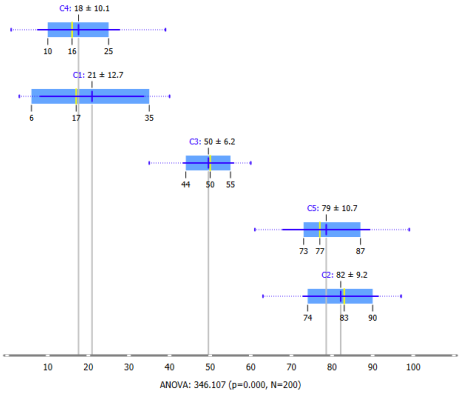

Fig. 12. Box Plot
Spending Score Distribution of Clusters

Using the box plots in Figures 10, 11, and 12 we are able to construct a profile of what the average member of each cluster looks like. From Figures 10 & 12 an interesting correlation can be made for clusters C2 & C5 as well as C1 & C4. In Figure 10 we see that clusters C2 & C5 are very close to each other in terms of age and when compared to Figure 12 we see that that those same two groups share very similar Spending Scores which in this case high. For clusters C1 & C4 the opposite is true where these two clusters represent a distinctly older population of customers, but have much lower Spending Scores. Across all three of the Figures, cluster C3 is almost always the average or middle of the box plot distribution in the three major features of the customer population.

V. Conclusions

The results found in Section IV can be leveraged by a marketing in order to better market sales and promotions to the cluster of customers that will respond the best. If the Spending Score is to be believed as an accurate measure of an ideal customer than it would stand to reason that the best cluster to promote to is clusters C2 and C5. If the products or services being marketed require customers to have high Annual Incomes than cluster C2 would be the best group of customers to target. Otherwise, if the product or service is meant to be marketed towards lower income customer then clusters C3 and C5 would the optimal group. In all other instances, clusters C1 and C4 are the groups of customers that the marketing team may wish to avoid. In all instances, Gender did not yield any meaningful results.