Customer Grouping Project using R

Introduction

Customer Grouping helps in identifying the potential customer base for selling a product. It helps in segregation of customer base into several groups of individuals with a similarity in different ways that are relevant to marketing such as gender, age, Income and miscellaneous spending habits.

Dataset : customer_data
It is consist of columns

Column Names	Data Type
Customer ID	Int
Gender	Factor
Age	Int
Annual.Income..K..	Int
Spending.Score..1.100.	Int

Key Steps :
First we have find out the proportion of gender, age, annual income & spending using histogram & Pie.
Then we used k-means algorithm and then plotted the ggplot for k-means clustering to provide the better visualization to help to target the appropriate customer base for product.

Analysis

Using clustering techniques, we can identify the several segments of customers allowing to target the potential customer base. In this project I have used ***K-means clustering*** which is the essential algorithm for clustering unlabelled dataset.

I have used standard deviation and summary function to understand the dataset. Histogram & Pie chart are being used to display the distribution of gender, age, Annual income & Spending habits across customer_data dataset.

Then I have calculated the number of clusters that would be required for K-means algorithm. I have used elbow method to find the optimal number of clusters. The main goal behind cluster partitioning method k-means is to define the clusters such that the intra-cluster variation stays minimum.

minimize(sum W(Ck)), k=1…k

Where Ck represents the kth cluster and W(Ck) denotes the intra-cluster variation. With the measurement of the total intra-cluster variation, one can evaluate the compactness of the clustering boundary.

First, I had calculated the clustering algorithm for several values of k. This was achieved by creating a variation within k from 1 to 10 clusters.

Then I calculate the total intra-cluster sum of square (variable : iss).

Then, I plot iss based on the number of k clusters. This plot denotes the appropriate number of clusters required in our model. In the plot, the location of a bend or a knee is the indication of the optimum number of clusters.

Then using the optimal cluster i.e. 6 I have calculated k-means operation. This had provided the following information :

- ✓ **cluster –** This is a vector of several integers that denote the cluster which has an allocation of each point.
- ✓ **totss –** This represents the total sum of squares.
- ✓ **centers –** Matrix comprising of several cluster centers
- ✓ **withinss –** This is a vector representing the intra-cluster sum of squares having one component per cluster.
- ✓ **tot.withinss –** This denotes the total intra-cluster sum of squares.
- ✓ **betweenss –** This is the sum of between-cluster squares.
- ✓ **size –** The total number of points that each cluster holds.

Finally using the first 2 principal components, I had visualized clustering using k-means clustering.

Result

This technique divides the customers into various groups to be targeted. Data related to geography, economic status as well as age group play a crucial role in determining the company direction towards addressing the various segments.

Conclusion

By using the clustering we could understand the variables much better and in turn this helps in taking the careful decision. By identification of target customers, companies could decide the relevant services or products to be launched.