# ANA-515 Analytics\_ Project

Customized Marketing by segmenting customer data by (age, gender, and interest) from Mall customer data.

Authors: Ritesh Tikhade

Department of Data Management (Data Analytics), McDaniel College

Professor: Anna Lissitz

Submission Date: May 8, 2022.

1. **Introduction**:

The project is a part of research study done in the coursework of MS in Data Analytics.

What is Customized Marketing: Customized marketing is a very expensive way of marketing and is generally used for niche target groups with high revenue generating potential. Customized marketing is also known as niche marketing as it is for a specific target audience and for customized products/services.

What is Customer Segmentation: Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits. Every customer has different requirements and require a specific marketing effort to address them appropriately. With the help of the customer segmentation companies will target correct audience and generate maximum Profit.

2 Business Idea:

The Primary objective of the project is to analyze Mall customers data from the website <https://www.kaggle.com/datasets/shwetabh123/mall-customers> and segment the customer age, gender and interest which will help to target the correct Audience. Download and import data Using R. I can Generate the segmentation of Customers data.

I am using clustering techniques(K-means), as companies can identify the several segments of customers which allow them to target the potential user. In this project, we will make use of K-means clustering which is the essential algorithm for clustering unlabeled dataset.

Diagram

Description automatically generated

3 Understanding the Dataset:

Mall Customers(6Cust\_ID):

A picture containing graphical user interface

Description automatically generated

Data sets Mall\_Customers.csv (4.29 kB) has 200 rows and 5 columns. Which include Customer id (1-200), age, Genre (Male, Female), Annual income(K$) and spending score (1-100).

1. Challenge:

My Dataset has all the information and has no error, missing values. Most of Raw data from different sources will have different formats focused on presenting specific information. When we brought together all the data, data may have duplication of attributes and blank values where subjects are not present in all systems.

This data preparation step includes eliminate duplicates and errors, remove incorrect or incomplete entries, fill up blank spaces wherever possible, and put it all in a standard format. As per the data protection policies applicable to the business, some data fields will need to be masked and/or removed as well.

5. Exploratory Analysis:

The first part of the project is to see their graphical distribution of Mall customer data and then K means algorithm for clustering with help R language.

We can see the Min, Max, mean and Median of the Mall Customer dataA screenshot of a computer

Description automatically generated with medium confidence

Age distribution of the mall customer data.

Chart, histogram

Description automatically generated

Form Chart and summary data, we can see that age range between 30-35 frequently visited the Mall and age range 60-65 has not visited often. Minimum age is 17 and maximum is 70.

Gender Comparison:

Shape, square

Description automatically generated

From the graph we can conclude that Female more visited the Mall compared to Male.

Annual Income:

Chart, histogram

Description automatically generated

From the above graph and summary data we can see that minimum annual income is 15 and maximum is 137. Average salary of the customer is 60.56.

Histogram of Spending Score.

Chart, histogram

Description automatically generated

From the data and Graph: The minimum spending score is 1, maximum is 99 and the average is 50.20.

we conclude that customers between 40 and 50 have the highest spending score among all the classes.

K mean Algorithm:

We must specify the number of clusters that we need to create. The algorithm selects the k objects at random from the dataset. This object is initial cluster or mean.

The closest centroid obtains the assignment of a new observation. We base this assignment on the Euclidean Distance between object and the centroid.

k clusters in the data point update the centroid through calculation of the new mean values present in all the data points of the cluster. The kth cluster, s centroid has a length of p that contains means of all variables for observations in the kth cluster. We denote the number of variables with p.

Iterative minimization of the total within the sum of squares. Then through the iterative minimization of the total sum of the square, the assignment stops wavering when we achieve maximum iteration. The default value is 10 that the R software uses for the maximum iterations.

There are multiple methods for the determining Optimal Clusters, but we are using Elbow Method. The Main goal is to find the clusters so that the intra-cluster variation stays minimum.

minimize (sum W(Ck)), k=1…k

where:

Ck: kth cluster

W(CK): Intra-cluster Variation

By measuring the total intra-cluster variation, we can evaluate the compactness of the clustering boundary. We can then proceed to define the optimal clusters. First, we calculate the several value of K then we calculate the total intra-cluster sum of square (iss) then we plot ISS based on the number of K which will help us to get number of clusters required. In plot bend or Knee is indication of the optimum number of clusters.

Chart, line chart

Description automatically generated

From the above graph we can see that graph is bend in Elbow from 4 - 6 So, 6 is appropriate number of Clusters.

Segmentation of the Mall customer Using K means:

Chart, scatter chart

Description automatically generated

From the above graph there is distribution of 6 Clusters: The cluster 3 is the customers with the least income and least spending score, similarly, the cluster 4 is the customers with the most income and most spending score.

Conclusion:

form the graphical distribution of age, spending score, Gender. we can say that age between 30 to 35 and customers between 40 and 50 have the highest spending score and women has more visited the mall compared to men. With segmentation, the cluster 4 has most income and spending score. With the help of this data, we can sell the product based on the identified customer group which is nothing but Customized Marketing.