**# R-Programming-Customer-Segmentation**

Customer-Segmentation Project

**# Introduction**

Customer Segmentation is one the most important applications of unsupervised learning. Using clustering techniques, companies can identify the several segments of customers allowing them to target the potential user base. In this machine learning project, we will make use of k-mean Clustering which is the essential algorithm for clustering unlabelled dataset.

**# SCOPE**

  Whenever you need to find your best customer, customer segmentation is the ideal methodology. We will perform one of the most essential applications of machine learning – Customer Segmentation. In this project, we will implement customer segmentation in R.

**# PROJECT SPECIFICATION**

  - R Studio version 1.2.5033

**# HARDWARE SPECIFICATIONS**

  - Microsoft® Windows® 7/8/10 (32- or 64-bit)

  - 3 GB RAM minimum, 8 GB RAM recommended;

  - 2 GB of available disk space minimum

  - core processor of i3 minimum or above.

**# DATASET**

  - Mall\_Customers.csv

**# PACKAGES REQURIED:**

  - plotrix

  - purr

  - cluster

  - gridExtra

  - grid

  - nbClust

  - factoextra

  - ggplot2

  - dplyr

**# 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.

Companies that deploy customer segmentation are under the notion that every customer has different requirements and require a specific marketing effort to address them appropriately.

Companies aim to gain a deeper approach of the customer they are targeting. Therefore, their aim has to be specific and should be tailored to address the requirements of each and every individual customer. Furthermore, through the data collected, companies can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. This way, they can strategize their marketing techniques more efficiently and minimize the possibility of risk to their investment. The technique of customer segmentation is dependent on several key differentiators that divide customers into groups to be targeted.

Data related to demographics, geography, economic status as well as behavioral patterns play a crucial role in determining the company direction towards addressing the various segments.

**# Customer Gender Visualization:**

In this, we will create a barplot and a piechart to show the gender distribution across our customer\_data dataset. A bar chart represents data in rectangular bars with length of the bar proportional to the value of the variable. R uses the function barplot() to create bar charts. R can draw both vertical and Horizontal bars in the bar chart. In bar chart each of the bars can be given different colors

**# K-means Algorithm**

 While using the k-means clustering algorithm, the first step is to indicate the number of clusters (k) that we wish to produce in the final output. The algorithm starts by selecting k objects from dataset randomly that will serve as the initial centers for our clusters. These selected objects are the cluster means, also known as centroids. Then, the remaining objects have an assignment of the closest centroid. This centroid is defined by the Euclidean Distance present between the object and the cluster mean. We refer to this step as “cluster assignment”.

  When the assignment is complete, the algorithm proceeds to calculate new mean value of each cluster present in the data. After the recalculation of the centers, the observations are checked if they are closer to a different cluster. Using the updated cluster mean, the objects undergo reassignment. This goes on repeatedly through several iterations until the cluster assignments stop altering. The clusters that are present in the current iteration are the same as the ones obtained in the previous iteration. Summing up the K-means clustering –

  - We specify the number of clusters that we need to create.

  - The algorithm selects k objects at random from the dataset. This object is the 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 points 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 k-th 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 stop wavering when we achieve maximum iteration. The default value is 10 that the R software uses for the maximum iterations.

  - we calculate the clustering algorithm for several values of k. This can be done by creating a variation within k from 1 to 10 clusters. We then calculate the total intra-cluster sum of square (iss). Then, we proceed to 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. Let us implement this in R as follows –

**#### Code: library(purrr) set.seed(123)**

function to calculate total intra-cluster sum of square iss <- function(k) { kmeans(customer\_data[,3:5],k,iter.max=100,nstart=100,algorithm="Lloyd" )

$tot.withinss }

k.values <- 1:10 iss\_values <- map\_dbl(k.values, iss)

plot(k.values, iss\_values, type="b", pch = 19, frame = FALSE, xlab="Number of clusters K", ylab="Total intra-clusters sum of squares")

# Visualizing the Clustering Results using the First Two Principle Components

A line chart or line plot or line graph or curve chart is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments. It is a basic type of chart common in many fields. Used across many fields, this type of graph can be quite helpful in depicting the changes in values over time. We are going to use ggplot for depicting the line plot.

Code: set.seed(1) ggplot(customer\_data, aes(x =Annual.Income..k.., y = Spending.Score..1.100.)) + geom\_point(stat = "identity", aes(color = as.factor(k6$cluster))) + scale\_color\_discrete(name=" ", breaks=c("1", "2", "3", "4", "5","6"), labels=c("Cluster 1", "Cluster 2", "Cluster 3", "Cluster 4", "Cluster 5","Cluster 6")) + ggtitle("Segments of Mall Customers", subtitle = "Using K-means Clustering")

**# CONCLUSION**

In this data science project, we went through the customer segmentation model. We developed this using a class of machine learning known as unsupervised learning. Specifically, we made use of a clustering algorithm called K-means clustering. We analysed and visualized the data and then proceeded to implement our algorithm. Hope you enjoyed this customer segmentation project of machine learning using R