DATA SCIENCE PROJECT REPORT

TOPIC: CUSTOMER SEGMENTATION

SUBMITTED BY

JAYESH BHALCHIM

AMISHA GOKHALE

UNDER THE GUIDELINES

OF

EXPOSYS DATA LABS

ABSTRACT:

Customer segmentation and pattern extraction is one of the key aspects of business decision support system. In order to grow the business intelligently in competitive market, identification of potential customer should be done timely. This paper proposes an integrated novel approach for determining target customers using predictive model and discover their associative buying patterns using Apriori algorithm. After identification of targeted customers and their associative buying pattern, the business managers take the strategic profitable decisions accordingly.

KEYWORDS:

CRM, Clustering, Customer Segmentation, Associative Mining.

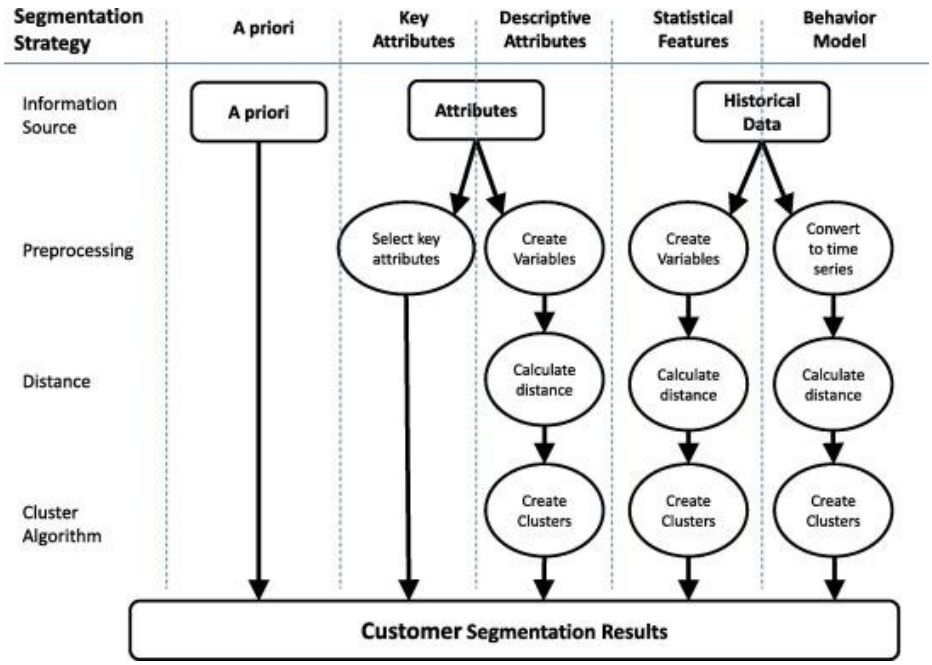
INTRODUCTION

With the evolution of new technologies and increasing growth of e-commerce it is important for every business to adapt new strategies which help them to win the competitive environment. The most asset of any business is customer. In this emerging market it is very difficult to maintain its customer base. To overcome this difficulty every business must focus on customer segmentation. Customer segmentation means categorizing the customers into same group. It acts as a base for Customer Relationship Management (CRM) for businesses to firstly identify their target customers and work on them individually. There are various existing predictive models which provide information on customer segmentation, and which help them to segregate the customers. For a successful business, apart from retaining and adding new customers, making more profit from each customer is key task. Different variety of models exist which help the business managers to implement the strategic policy according to individual customer taste but each one of it are having their own limitations. To excel the business further this paper is proposing a new integrated approach of customer segmentation combined with association mining of different segmented customer which provides more profit to the business.

PROPOSED METHODOLOGY:

Identifying right customer and providing right service at right time and treating different types of customers differently is the key to success in business.

So, a predictive model will be used to segregate customers into different groups based on their transactional data. Once the customers are segregated then their associative buying pattern are identified to enhance the profit for the organization future coming customer.



<u>The whole process is to be carried out in two phases</u>

Phase 1: Customer Segmentation

Phase 2: Extracting associative buying pattern of segmented customer

Phase 1: Customer Segmentation

Step 1: Collecting Customer Data (Transactional data): This step involves the collection of transactional customer data comprises of their static (Eg: Age, Gender etc.) and dynamic data (Eg: Purchase frequency etc.) [1] from shopping vendors.

```
customer_data=read.csv("/home/dataflair/Mall_Customers.csv")
str(customer_data)
```

```
## 'data.frame':    200 obs. of  5 variables:
##  $ CustomerID            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Gender                : Factor w/ 2 levels "Female","Male": 2 2 1 1 1 1 1 1 2 1
...
##  $ Age                   : int  19 21 20 23 31 22 35 23 64 30 ...
##  $ Annual.Income..k..    : int  15 15 16 16 17 17 18 18 19 19 ...
##  $ Spending.Score..1.100.: int  39 81 6 77 40 76 6 94 3 72 ...
```

```
names(customer_data)
```

```
## [1] "CustomerID"             "Gender"
## [3] "Age"                    "Annual.Income..k.."
## [5] "Spending.Score..1.100."
```

Step 2: Pre-processing of Data:

Pre-processing of the data is one of the important steps for the accuracy of predictive model. In this step, the collected data will be cleaned, and relevant features will be extracted. Feature selection is a data reduction technique which is responsible for extracting relevant features required for input vector of predictive model. This acts as pre-processing steps for creating subset of original features by excluding those features which are redundant.

This paper proposes correlation technique for extracting relevant features. Correlation measures the relationship between two features. Basically it simply filters those features which are not redundant to form subset of original features. To measure the association between features correlation coefficient is calculated between two features and based on its value.

Correlation is broadly categorized into three categories as follows:

Positive correlation: If two features are related in such a manner that if one increases other also increases or if one decreases then other also decreases then this is called as positive correlation.

Negative correlation: This correlation occurs when one decreases other increases.

No correlation: It occurs when there is no relationship between two features (i.e. the features are independent.
So consider those features which are independent to form input vector for model.

Step 3: Pass the input vector to the model for training. After training the model will
Divide the customer data into homogeneous segments.

Step 4: Once the model is trained we pass the test data to check its accuracy and efficiency.

Step 5: Now the predictive Model will predict segments of future customer data.

Once the customers are segmented in phase 1 we find hidden associative buying pattern using association mining Apriori technique from the particular segments to excel the profit of organization which is discussed in next section

APPROACH FOR PROPOSED METHODOLOGY

Research Method for customer Segmentation

To identify the target customers Clustering technique can be used for cluster analysis. Clustering is defined as to group data in clusters/segments so that data within segment are similar while data across the segments are dissimilar. Various techniques can be used for clustering like k means, hierarchical, grid-based model-based technique.

In this paper we proposed to use K-means technique for customer segmentation due its following advantages:

This technique suits for the data with numeric features and often terminates at local optimum.

It is highly scalable and efficient for large data sets.
It is fast in modelling and its result is more understandable.

K-MEANS CLUSTERING:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

The k-means clustering algorithm divides the n records into k segments of records called clusters where k ≤ n, so as to minimize the distances between records within a particular cluster.

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.

Determining Optimal Clusters

While working with clusters, you need to specify the number of clusters to use. You would like to utilize the optimal number of clusters. To help you in determining the optimal clusters, there are three popular methods –

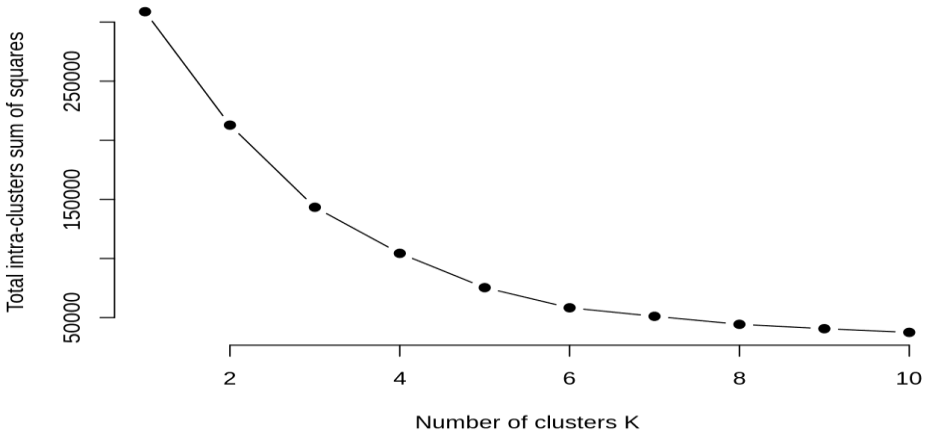
- Elbow method
- Silhouette method
- Gap statistic

Elbow Method

The main goal behind cluster partitioning methods like k-means is to define the clusters such that the intra-cluster variation stays minimum.

minimize(sum W(Ck)), k=1…k

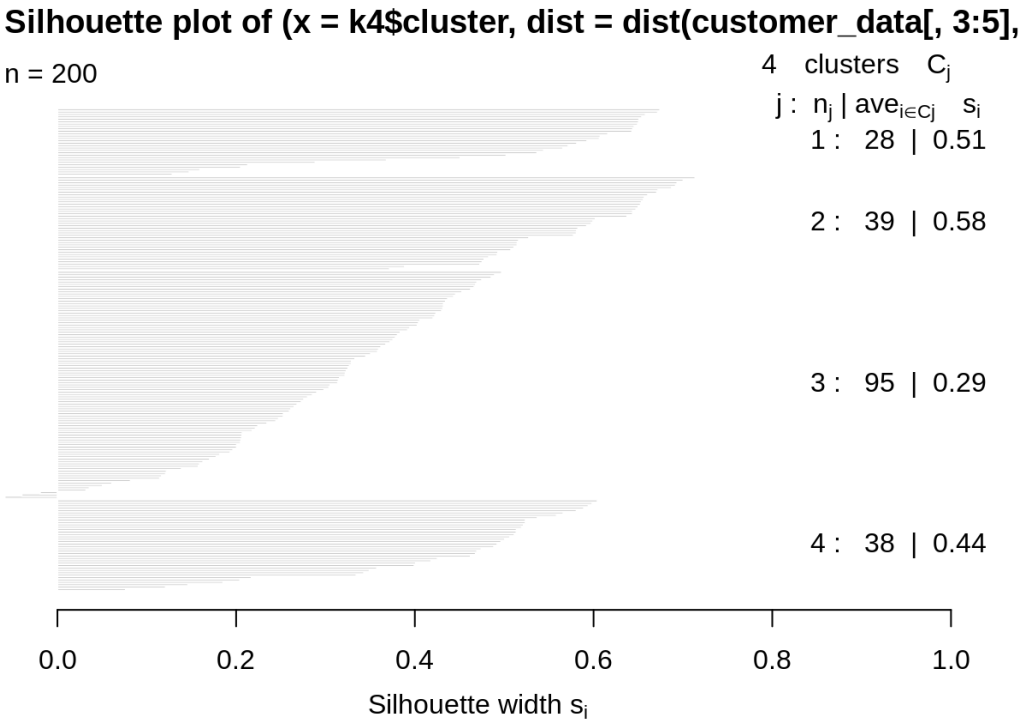
Where Ck represents the kth cluster and W(Ck) denotes the intra-cluster variation. With the measurement of the total intra-cluster variation, one can evaluate the compactness of the clustering boundary.



AVERAGE SILHOUETTE METHOD:

Average Silhouette Method

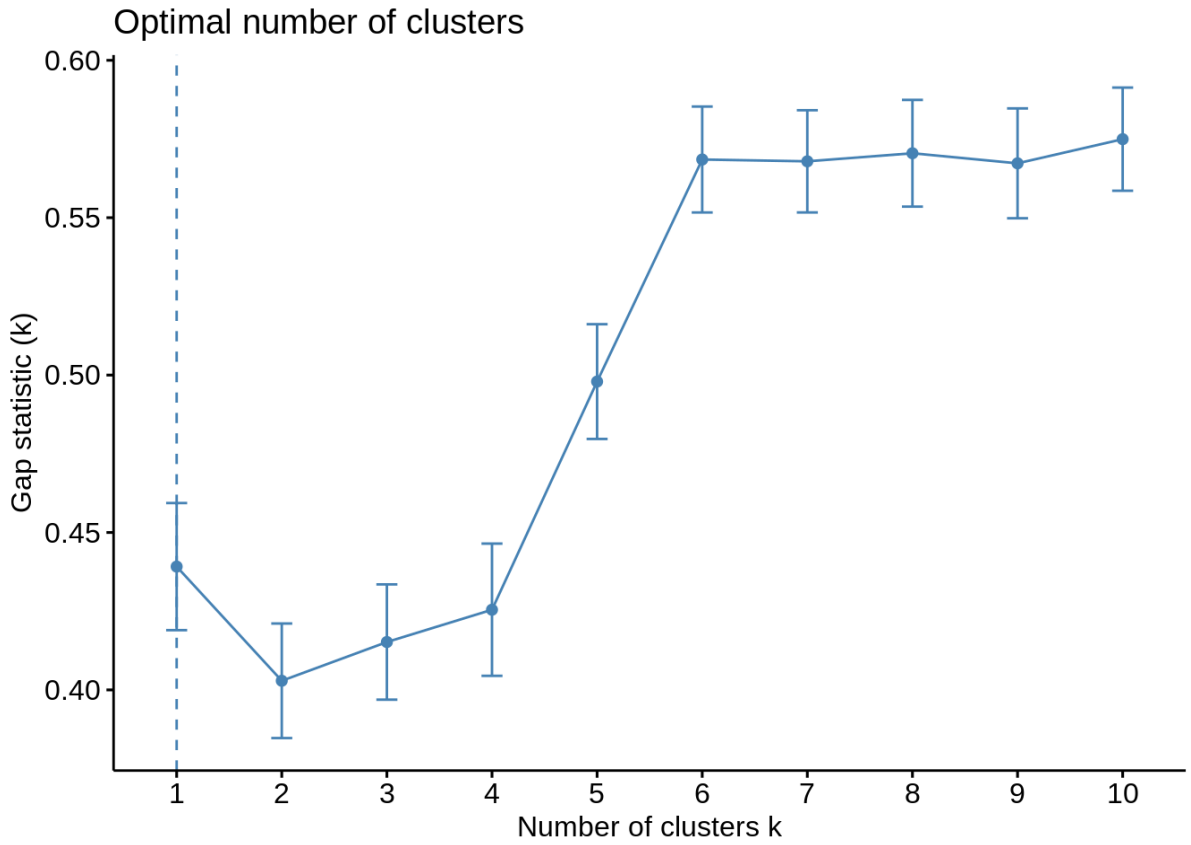
With the help of the average silhouette method, we can measure the quality of our clustering operation. With this, we can determine how well within the cluster is the data object. If we obtain a high average silhouette width, it means that we have good clustering. The average silhouette method calculates the mean of silhouette observations for different k values. With the optimal number of k clusters, one can maximize the average silhouette over significant values for k



clusters.

Gap Statistic Method

In 2001, researchers at Stanford University – **R. Tibshirani, G.Walther and T. Hastie** published the Gap Statistic Method. We can use this method to any of the clustering method like K-means, hierarchical clustering etc. Using the gap statistic, one can compare the total intracluster variation for different values of k along with their expected values under the null reference distribution of data. With the help of **Monte Carlo simulations**, one can produce the sample dataset. For each variable in the dataset, we can calculate the range between min(xi) and max (xj) through which we can produce values uniformly from interval lower bound to upper bound.
For computing the gap statistics method we can utilize the clusGap function for providing gap statistic as well as standard error for a given output.

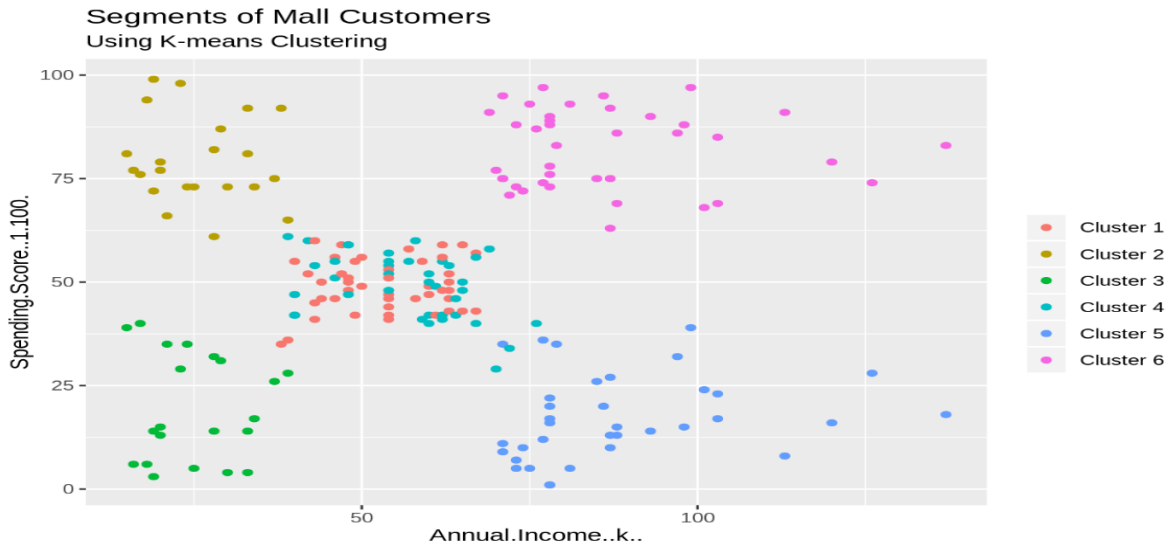


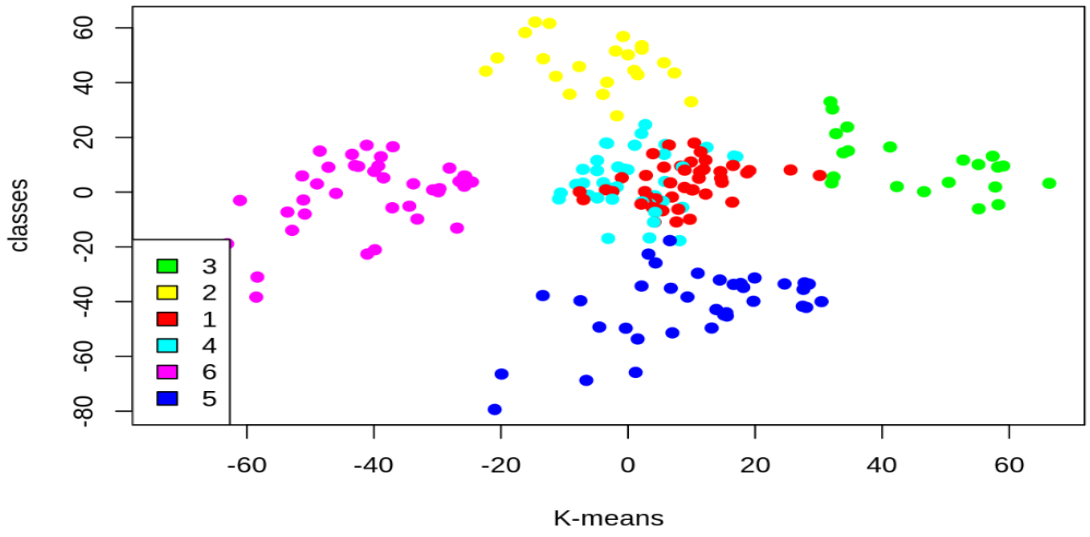
In the output of our kmeans operation, we observe a list with several key information. From this, we conclude the useful information being –

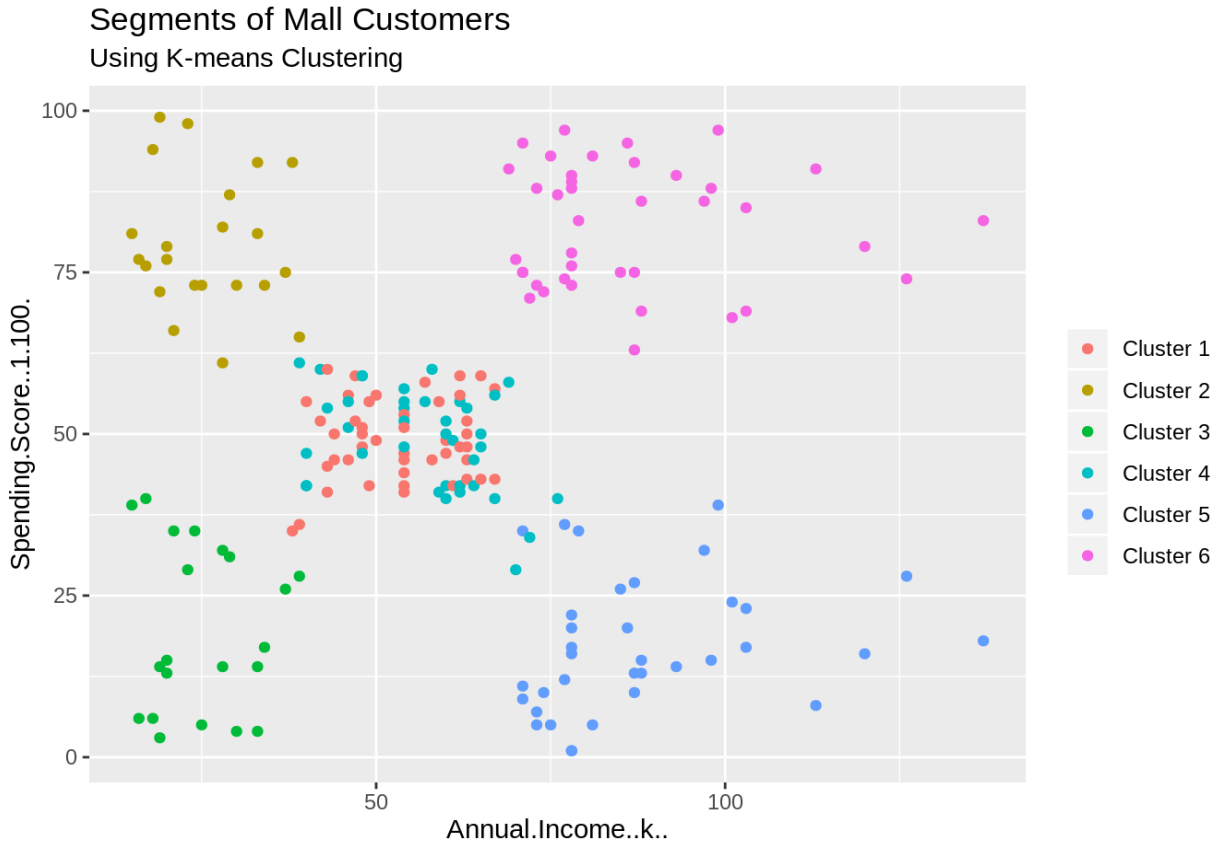
- **cluster –** This is a vector of several integers that denote the cluster which has an allocation of each point.
- **totss –** This represents the total sum of squares.
- **centers –** Matrix comprising of several cluster centers

- **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.

```
ggplot(customer_data, aes(x =Spending.Score..1.100., y =Age)) +
    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")
```

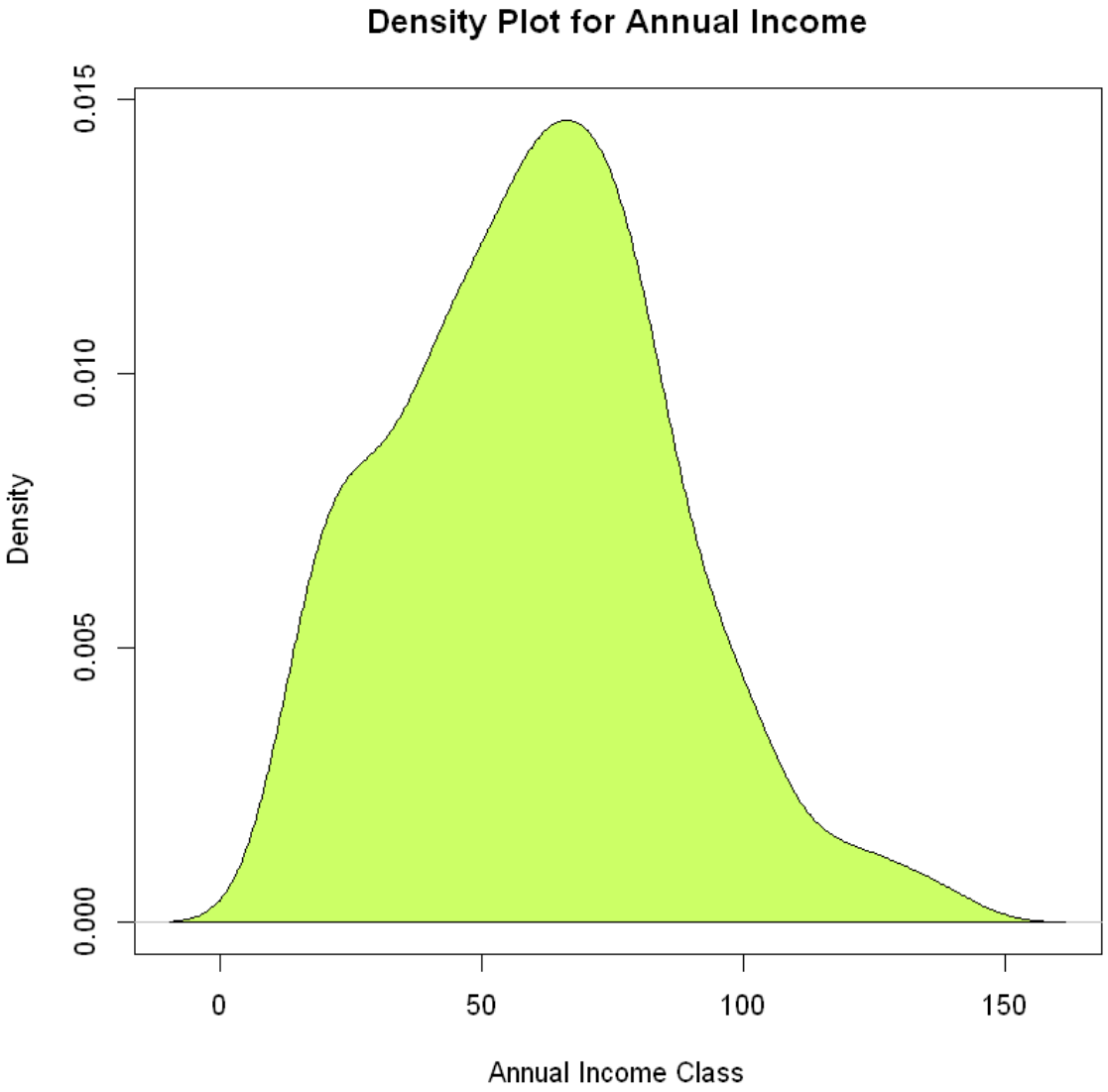



classes
60
40
20
0
-20
-40
-60
-80
3
2
1
4
6
5
-60
-40
-20
0
20
40
60
K-means


Segments of Mall Customers
Using K-means Clustering
Spending.Score..1.100.
100
75
50
25
0
50
100
Annual.Income..k..
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6

Summary

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.



CONCLUSION:

In competitive market of e-commerce, the problem of identifying potential customer is gaining more and more attention. To address this problem timely, this paper proposes a study on integrated novel approach based on clustering using K-means and associative mining using Apriori technique. After identification of targeted customers and their associative buying pattern, the business managers take the strategic profitable decisions accordingly. This integrated model could be directly brought into implementation for providing better profitable margins from sales.

REFERENCES 1. XIANG-BIN YAN, YI-JUN LI, Customer Segmentation based on Neural Network with Clustering Technique, 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006 (pp265-268).

2. Ling Luo, Bin Li et. al. “Tracking the Evolution of Customer Purchase Behaviour Segmentation via a Fragmentation-Coagulation Process”, Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17).

3. Mark K.Y.Mak,George T.S.Ho,S.L.Ting,“A Financial Data Mining Model for extracting Customer Behaviour”, INTECH open access publisher, 23 July 2011.

4. Juni Nurma San, Lukito Nugroho, Ridi Ferdiana, P.InsapSantosa, “A Review on Customer Segmentation Technique on E-Commerce”, Advance Science Letters, Vol.4, 400-407,2011.

5. Kishana R. Kashwan, Member, IACSIT, C.M.Velu, “Customer Segmentation using clustering and Data Mining Techniques”, International Journal of Computer Theory and Engineering, Vol.5, No.6, December 2013.

6. LuoYe, CaiQiuru, XiHaixu,LiuYijun and Zhu Ghuangping , “Customer Segmentation for Telecom with the k-means Clustering Method”, Information Technology Journal 12(3):409-413,2013 .