Research Inventy: International Journal of Engineering And Science
Vol.10, Issue 9 (September 2020), PP 49-55
Issn (e): 2278-4721, Issn (p):2319-6483, www.researchinventy.com

An application development of Consumer'sData Analysis using K-Mean clustering machine learning algorithm

Tiruveedula GopiKrishna, Mohammed Abdeldaiem
[1]*Department of Computer Science, Adama Science and Technology University, Adama, Ethiopia*
[2] *Department of Information Technology,Tripoli University, Libya*
Corresponding Author:Tiruveedula GopiKrishna

Abstract:*In this end – to – end application development, we will perform one of the most essential applications of machine learning – Consumer Segmentation. In this application project, we will implement consumer segmentation using machine learning algorithm. Whenever you need to find your best customer, consumer segmentation is the ideal methodology. In this machine learning project, discusses the background of consumer segmentation. Then we will explore the data upon which we will be building our segmentation model. Also, we will see the descriptive analysis of our data and then implement several versions of the K-means algorithm.*

Keywords: *Machine Learning, data science,K-means cluster,segmentation,classification.*

Date of Submission: 15-09-2020 Date of Acceptance: 30-09-2020

I. INTRODUCTION

Consumer clustering 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-means clustering which is the essential algorithm for clustering unlabeled dataset. Before ahead in this project, learn what actually customer segmentation is.

Consumer clustering is the process of division of user 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 [1,2,3].

Companies that deploy consumer clustering are under the notion that every user has different requirements and require a specific marketing effort to address them appropriately. Enterprises aim to gain a deeper approach of the consumer 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, enterprises can gain a deeper understanding of consumer 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 [3].

The technique of consumer clustering is dependent on several key differentiators that divide users into groups to be targeted. Data related to demographics, geography, economic status as well as behavioral patterns play a crucial role in determining the enterprise direction towards addressing the various segments [4].

II. IMPLEMENTATION PROCEDURE

1.1 About 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 [5].

1.2 Algorithm steps of 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 [5].

1.3 Implementionphase

In the first step of this implementation of this application, we will perform data exploration. We will import the essential packages required for this role and then read our data. Finally, we will go through the input data to gain necessary insights about it [6].

Program code

```
1. customer_data=read.csv("/home/gkdatabase/Mall_Customers.csv")
2. str(customer_data)
3. names(customer_data)
```

Output:

```
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."
```

We will now display the first six rows of our dataset using the head() function and use the summary() function to output summary of it.

Code:

```
1. head(customer_data)
2. summary(customer_data$Age)
```

Output:

```
head(customer_data)
```

```
##   CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100.
## 1          1   Male  19                 15                     39
## 2          2   Male  21                 15                     81
## 3          3 Female  20                 16                      6
## 4          4 Female  23                 16                     77
## 5          5 Female  31                 17                     40
## 6          6 Female  22                 17                     76
```

```
summary(customer_data$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   28.75   36.00   38.85   49.00   70.00
```

Code:

1. **sd**(customer_data$Age)
2. **summary**(customer_data$Annual.Income..k..)
3. **sd**(customer_data$Annual.Income..k..)
4. **summary**(customer_data$Age)

Output:

```
sd(customer_data$Age)
```

```
## [1] 13.96901
```

```
summary(customer_data$Annual.Income..k..)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   41.50   61.50   60.56   78.00  137.00
```

```
sd(customer_data$Annual.Income..k..)
```

```
## [1] 26.26472
```

```
summary(customer_data$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   28.75   36.00   38.85   49.00   70.00
```

Code:

1. **sd**(customer_data$Spending.Score..1.100.)

Output:

```
sd(customer_data$Spending.Score..1.100.)
```

```
## [1] 25.82352
```

1.4Customer Gender Visualization

In this, we will create a barplot and a piechart to show the gender distribution across our customer_data dataset [6].

Code:

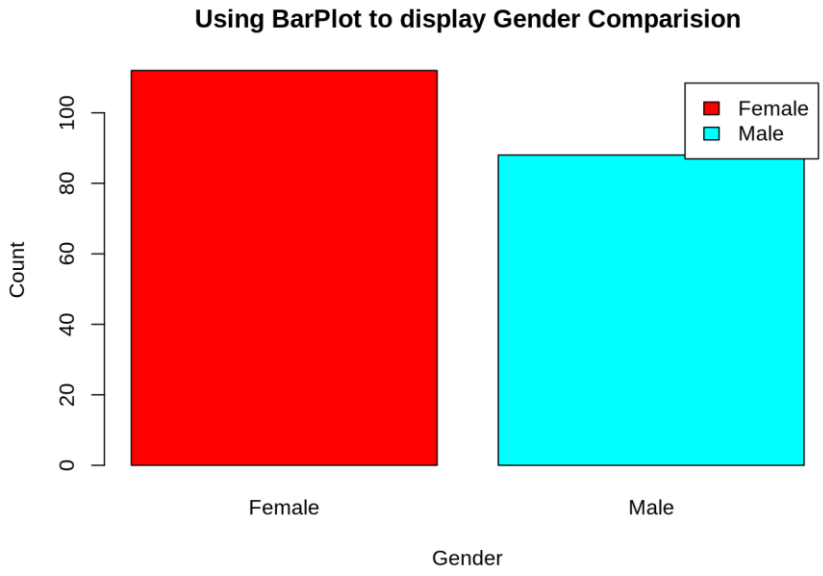
```
a=table(customer_data$Gender)
barplot(a,main="Using BarPlot to display Gender Comparision",
ylab="Count",
xlab="Gender",
col=rainbow(2),
legend=rownames(a))
```

Screenshot:

```
a=table(customer_data$Gender)
barplot(a,main="Using BarPlot to display Gender Comparision",
        ylab="Count",
        xlab="Gender",
        col=rainbow(2),
        legend=rownames(a))
```

Output:

Figure1. Gender comparison



From the above Figure 1 barplot, we observe that the number of females is higher than the males. Now, let us visualize a pie chart to observe the ratio of male and female distribution [6].

Code:

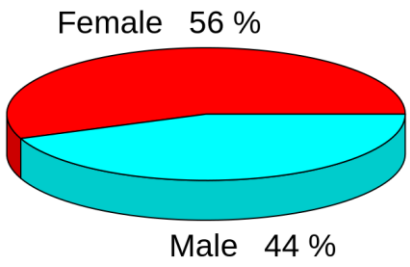
```
pct=round(a/sum(a)*100)
lbs=paste(c("Female","Male")," ",pct,"%",sep=" ")
library(plotrix)
pie3D(a,labels=lbs,
main="Pie Chart Depicting Ratio of Female and Male")
```

output:

```
pct=round(a/sum(a)*100)
lbs=paste(c("Female","Male")," ",pct,"%",sep=" ")
library(plotrix)
pie3D(a,labels=lbs,
     main="Pie Chart Depicting Ratio of Female and Male")
```

Output:

Figure2. depicting ratio of female and male



From the Figure 2 above graph, we conclude that the percentage of females is **56%**, whereas the percentage of male in the customer dataset is **44%**.

1.5Visualization of Age Distribution

Let us plot a histogram to view the distribution to plot the frequency of customer ages. We will first proceed by taking summary of the Age variable [6].

Code:

1. **summary**(customer_data$Age)

Output Screenshot:

```
summary(customer_data$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   28.75   36.00   38.85   49.00   70.00
```

Code:

1. **hist**(customer_data$Age,
2. col="blue",
3. main="Histogram to Show Count of Age Class",
4. xlab="Age Class",

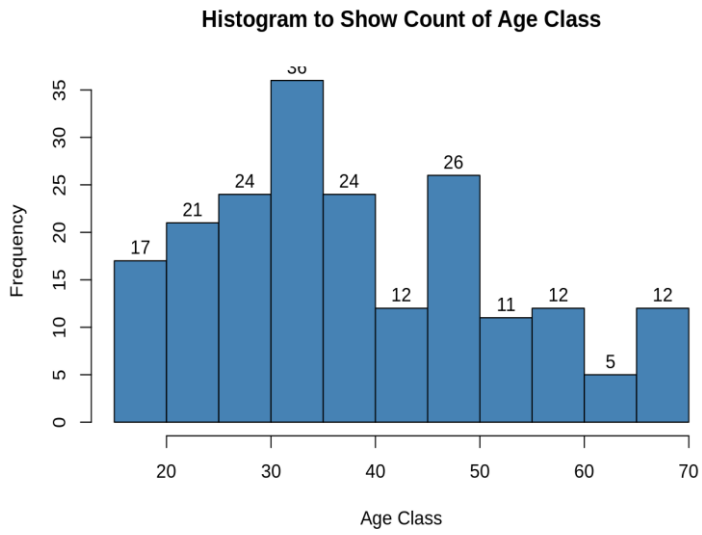
5. ylab="Frequency",
6. labels=TRUE)

Code screenshot:

```
hist(customer_data$Age,
     col="blue",
     main="Histogram to Show Count of Age Class",
     xlab="Age Class",
     ylab="Frequency",
     labels=TRUE)
```

Output:

Figure3. histogram to show count of age class



Code:

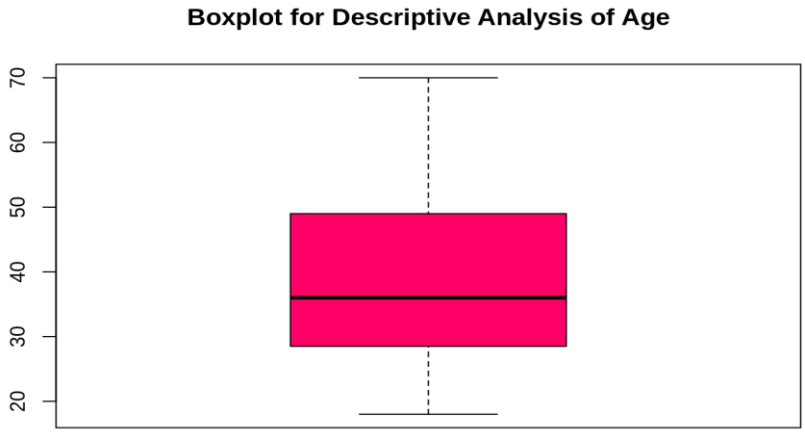
1. **boxplot**(customer_data$Age,
2. col="ff0066",
3. main="Boxplot for Descriptive Analysis of Age")

Output:

```
boxplot(customer_data$Age,
        col="#ff0066",
        main="Boxplot for Descriptive Analysis of Age")
```

Output:

Figure4. Descriptive analysis of age



III. RESULTS AND DISCUSSIONS

The above two visualizations from Figure 3 and Figure 4, we conclude that the maximum consumer ages are between 30 and 35. The minimum age of consumers is 18, whereas, the maximum age is 70.

IV. CONCLUSION

With the help of clustering, we can understand the variables much better, prompting us to take careful decisions. With the identification of consumers, enterprises can release products and services that target intended users based on several parameters like income, age, spending patterns, etc. Furthermore, more complex patterns like product reviews are taken into consideration for better segmentation in future .In this implementation 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 analyzed and visualized the data and then proceeded to implement our algorithm.

.

ACKNOWLEDGEMENT

The authors are grateful to the "Dataflair onlineproject Development for entended support.

REFERENCES

[1]. Abdulla, I.Q., 2014. Synthesis and antimicrobial activity of Ibuprofen derivatives. Natural Science 6,Shraddha, S.et al. 2014, "A Review ON K-means DATA Clustering APPROACH" International Journal of Information and Computation Technology. 47–53.

[2]. Y,S,Patail , M.B. Vaidya 2012, "A Technical survey on Clustering Analysis in Data mining" International Journal of Emerging Technology and Advanced Engineering..

[3]. Himanshu Gupta, Dr.Rajeev Srivastav 2014, "K-means Based Document Clustering with Automatic 'K' Selection and Cluster Refinement" International Journal of Computer Science and Mobile Applications..

[4]. Aloise D., Deshpande A., Hansen P., Popat P.: NP-hardness of Euclidean sum-of-squares clustering. Machine Learning, 75, 245−249 (2009).

[5]. www.https://www.tutorialandexample.com/k-means-clustering-algorithm.

[6]. An intelligent market segmentation system using k-means and particle swarm optimizationC. Chiu, Y. Chen, I. Kuo, H. Ku Computer Science 2009

Tiruveedula GopiKrishna, et. al. "An application development of Consumer'sData Analysis using K-Mean clustering machine learning algorithm." *International Journal of Engineering and Science*, vol. 10, no. 09, 2020, pp. 49-55.