

*School Of Computer Science and Engineering*

*VIT-AP UNIVERSITY*

*AMARAVATI*

**CUSTOMER SEGMENTATION**

**by**

**K Means Clustering Algorithm**

*Using R programming*

**Guided By:**

Prof: Dr. Scahi Nandan Mohanty

**Submitted by:**

Rampam Greeshma Geethika

20BCD7094

GROUP 36

geethika.20bcd7094@vitap.ac.in

**ABSTRACT:**

Data analytics is a broad term that encompasses many diverse types of data analysis. Any type of information can be subjected to data analytics techniques to get insight that can be used to improve things. Data analytics techniques can reveal trends and metrics that would otherwise be lost in the mass of information. This information can then be used to optimize processes to increase the overall efficiency of a business or system. It is important because it helps businesses optimize their performances. Implementing it into the business model means companies can help reduce costs by identifying more efficient ways of doing business and by storing large amounts of data. A company can also use data analytics to make better business decisions and help analyze customer trends and satisfaction, which can lead to new—and better—products and services.

**KEYWORDS:**

K Means Algorithm, Clustering, Histograms, Boxplots, Polygon graphs

**INTRODUCTION:**

Customer segmentation is one of the most popular data science projects. Customer segmentation is a well-known example of unsupervised learning. Companies use clustering to identify customer groups and target potential user bases. Segment consumers based on common characteristics such as gender, age, hobbies, and purchasing patterns to effectively sell to each group. You can visualize the distribution of gender and age using K-means clustering. Then look at their annual income and spending patterns.

**Literature survey:**

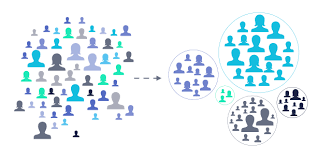
What is Customer Segmentation?

Customer Segmentation is the process of division of the 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.

What is K Means Clustering?

K Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

Customer segmentation techniques rely on several key differentiators to divide customers and Address target groups. Data about demographics, geography, economic conditions, and behavioral patterns play a key role in determining the company's direction to address its various segments.



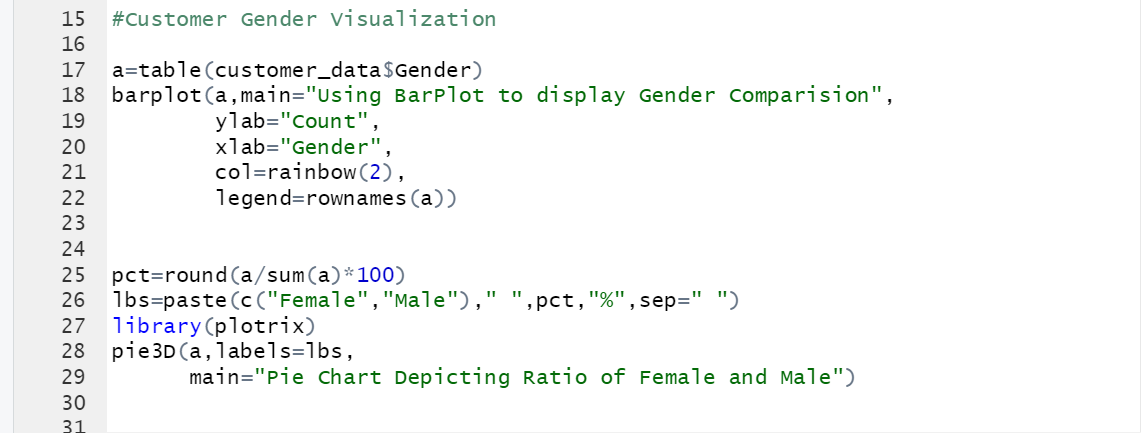
**FIGURE (1)**

**Methodology:**

The dataset is saved as a Mall\_Customers.csv file. This dataset contains 400 records of various types of customers. The events saved in the dataset are unstructured. To perform analysis, reading data is done using the command “read.csv”.

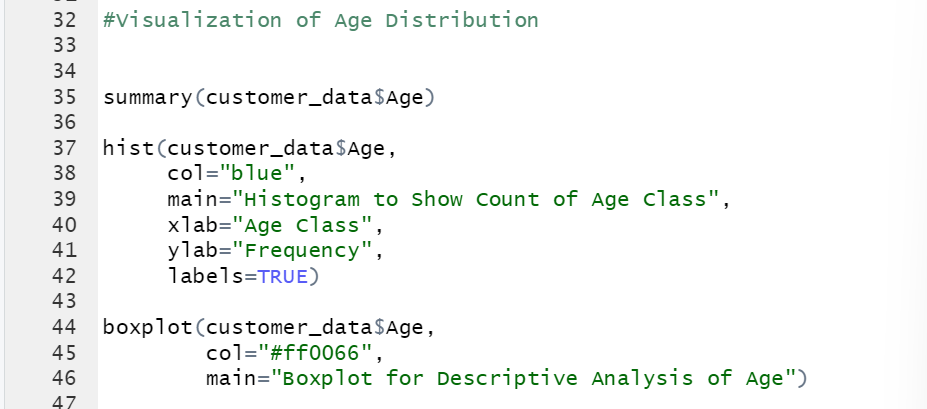
>customer\_data<-read.csv("D:\\R\\Mall\_customers.csv")

CUSTOMER GENDER VISUALIZATION:



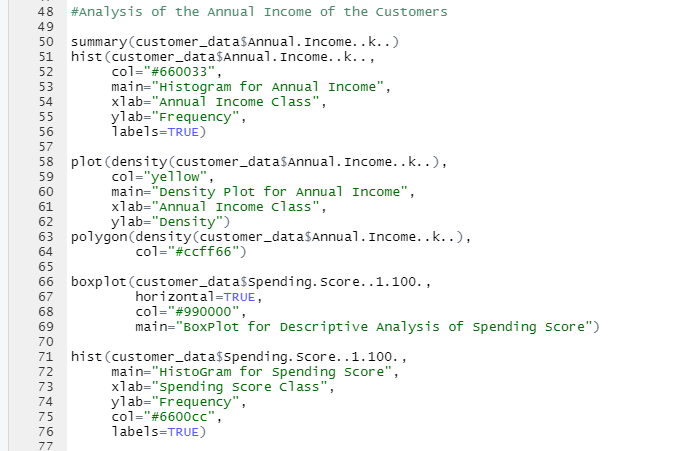
**FIGURE (2)**

VISUALIZATION OF AGE DISTRIBUTION:



**FIGURE (3)**

ANALYSIS OF ANNUAL INCOME OF CUSTOMERS:



**FIGURE (5)**

K MEANS CLUSTERING:

When using the k-means clustering algorithm, the first step is to specify the number of clusters (k) that you want to produce in the final output. The algorithm starts by selecting k objects. A random dataset that serves as the initial centers of the clusters. These selected objects are clusters also called centroids. Then the rest object has the following assignment

Next focus. This centroid is defined by the Euclidean distance that exists between objects.

and cluster means.

#K-means Algorithm

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:

>pcclust=prcomp(customer\_data[,3:5],scale=FALSE) #principal component analysis

>summary(pcclust)

>

>pcclust$rotation[,1:2]

>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")

>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")

>kCols=function(vec){cols=rainbow (length (unique (vec)))

>return (cols[as.numeric(as.factor(vec))])}

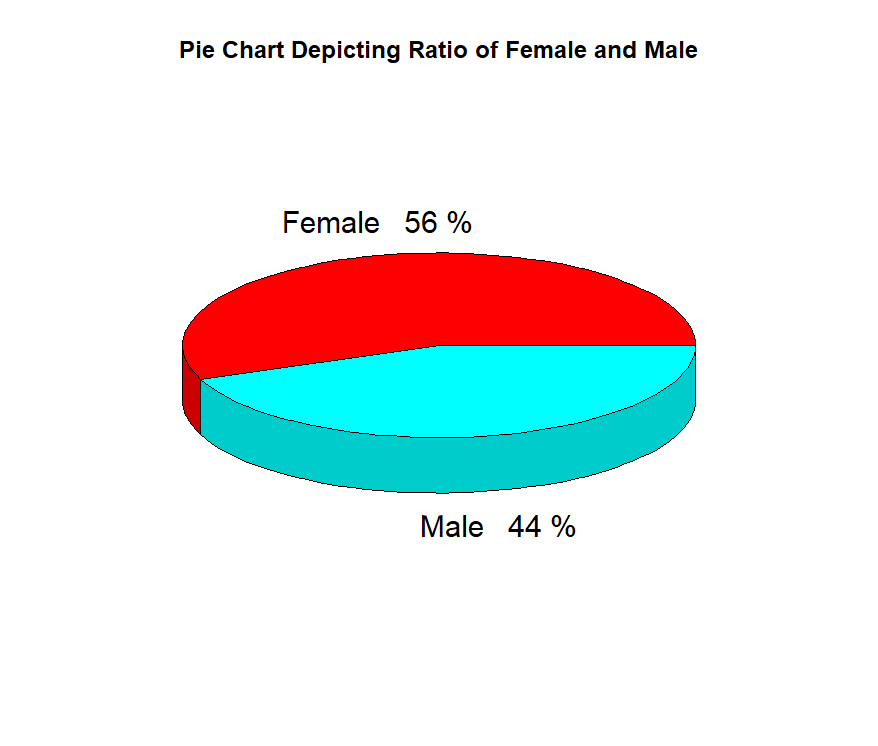
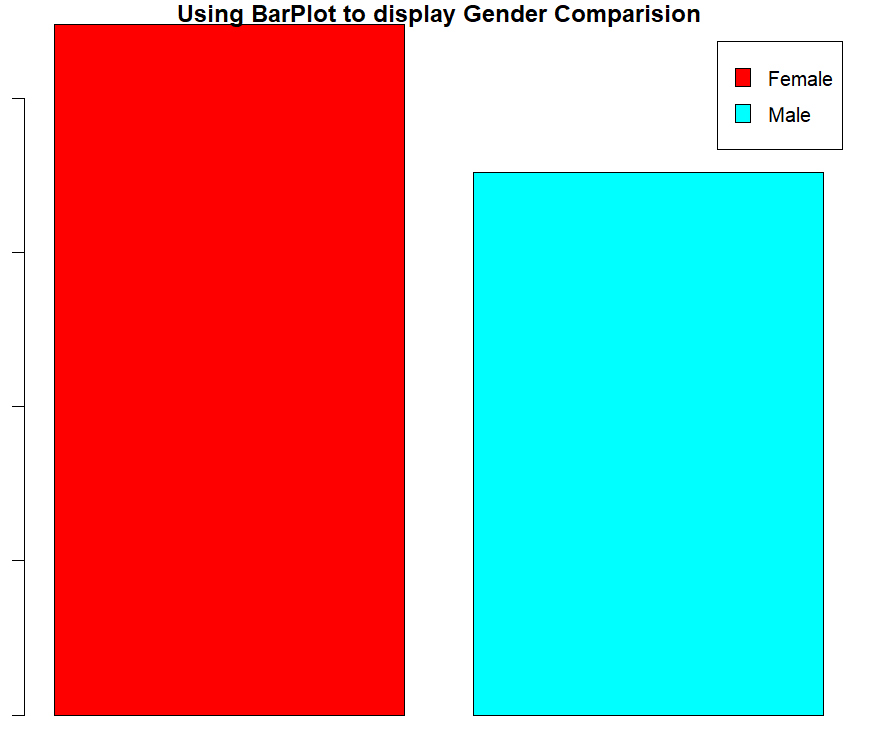
>digCluster<-k6$cluster; dignm<-as.character(digCluster); # K-means clusters

>plot(pcclust$x[,1:2], col =kCols(digCluster),pch =19,xlab ="K-means",ylab="classes")

>legend("bottomleft",unique(dignm),fill=unique(kCols(digCluster)))

**Results and Discussion:**

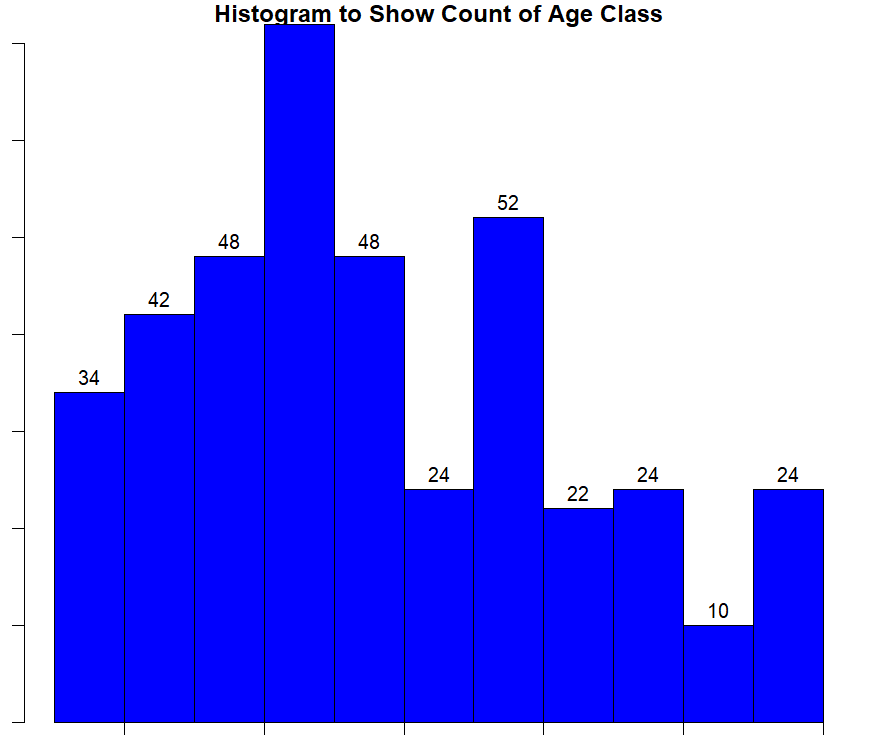
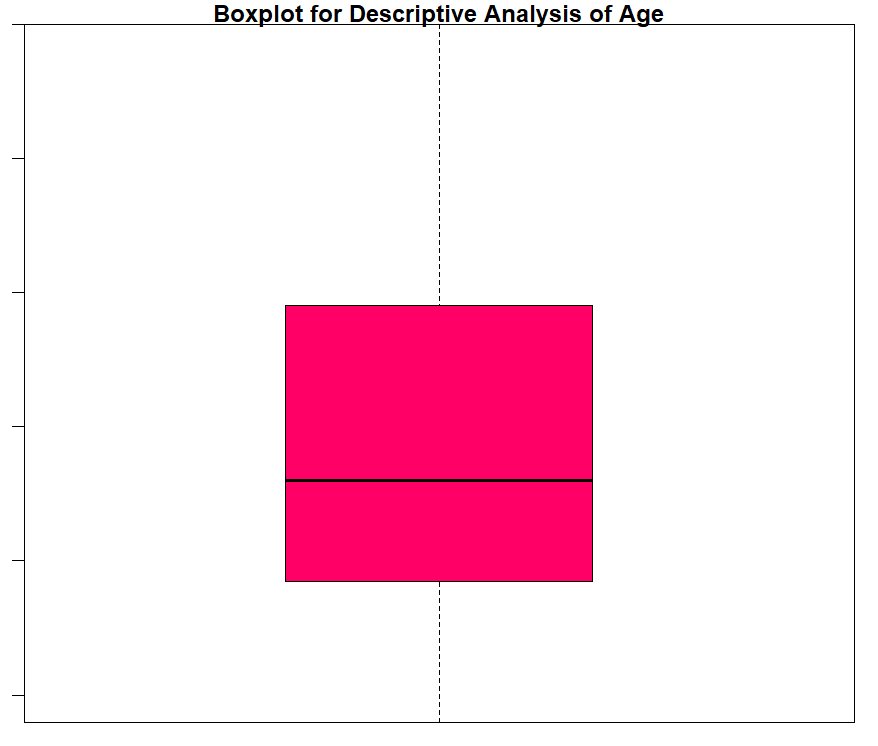
**CUSTOMER GENDER VISUALIZATION:**

****

**FIGURE (6) Gender Comparison FIGURE (7) Gender Ratio**

From the above barplot and pie charts, we can collude that the number of females is more than the number of males and the percentage of females is 56%, whereas the percentage of males is 44%.

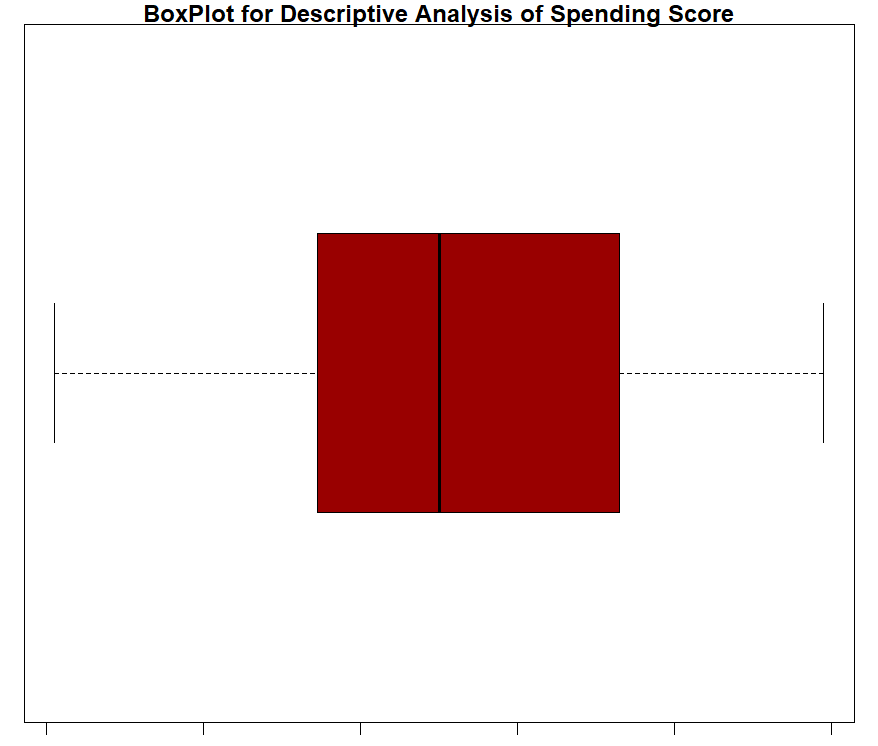
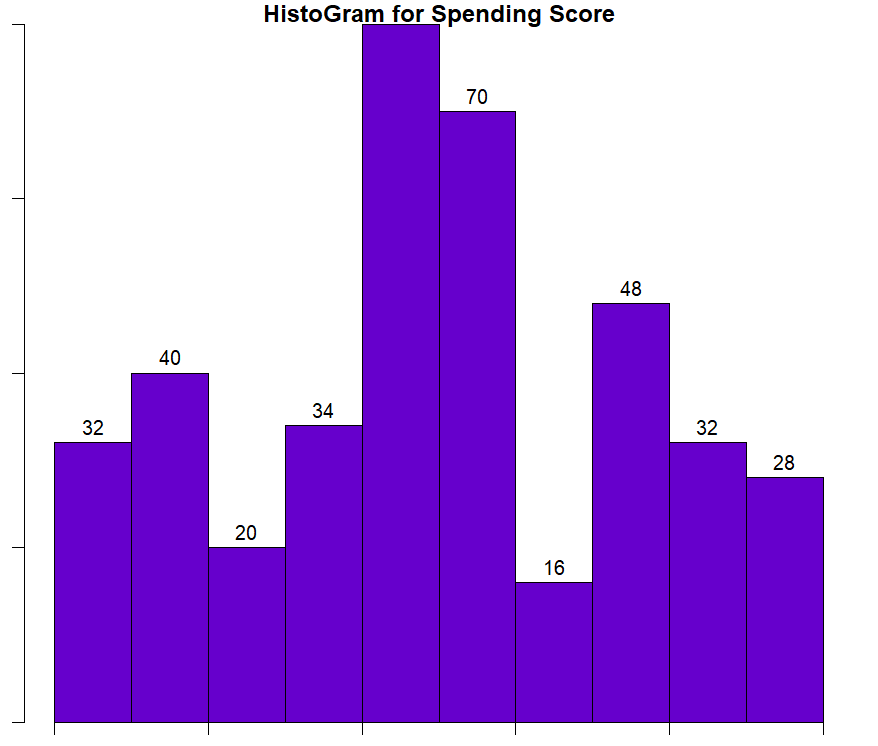
VISUALIZATION OF AGE DISTRIBUTION:

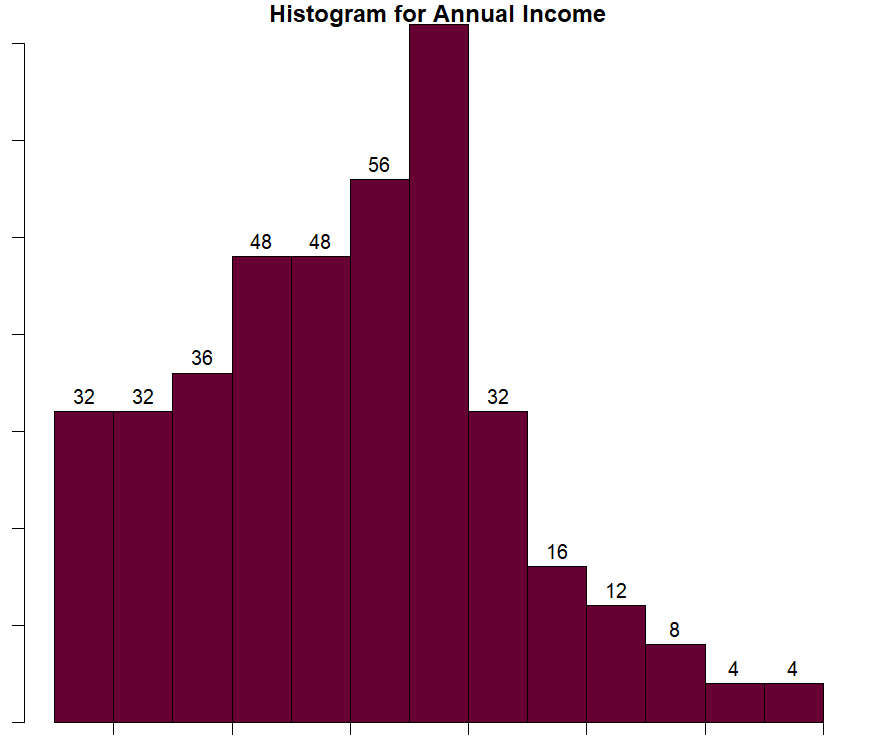
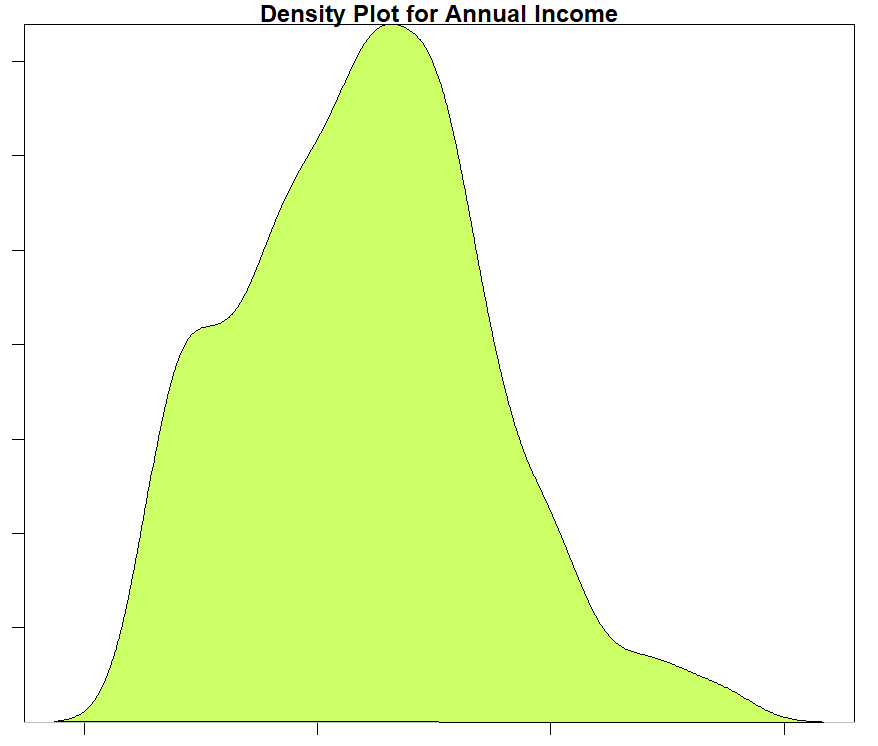
**FIGURE (8) FIGURE (9)**

From the above two visualizations, we conclude that the maximum customer ages are between 30 and 35. The minimum age of customers is 18, whereas, the maximum age is 70.

ANALYSIS OF THE ANNUAL INCOME OF THE CUSTOMERS:



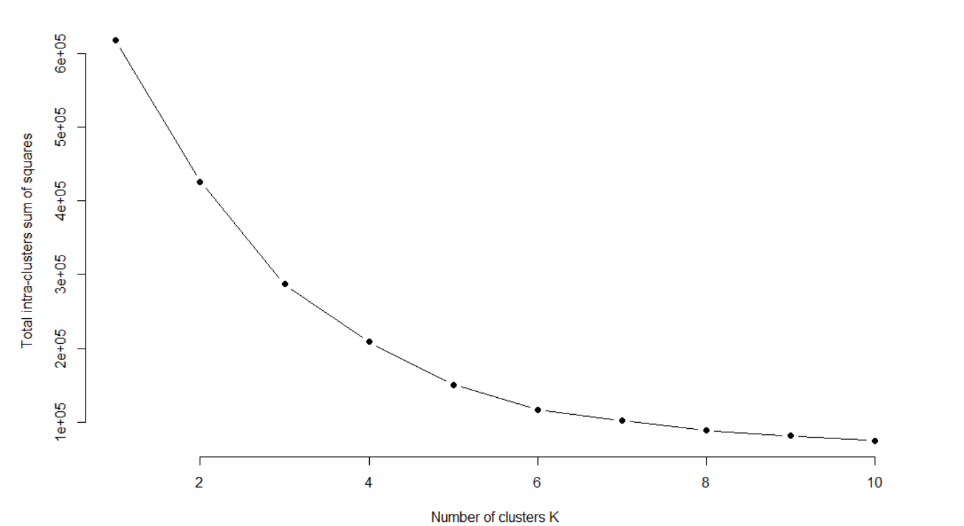
**FIGURE (10) FIGURE (11)**



**FIGURE (12) FIGURE (13)**

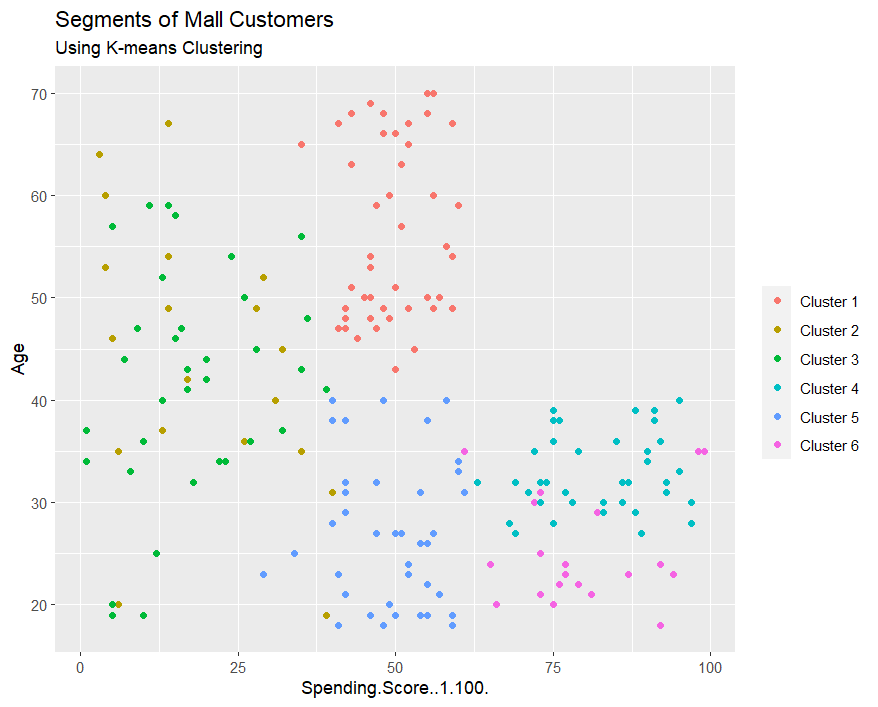
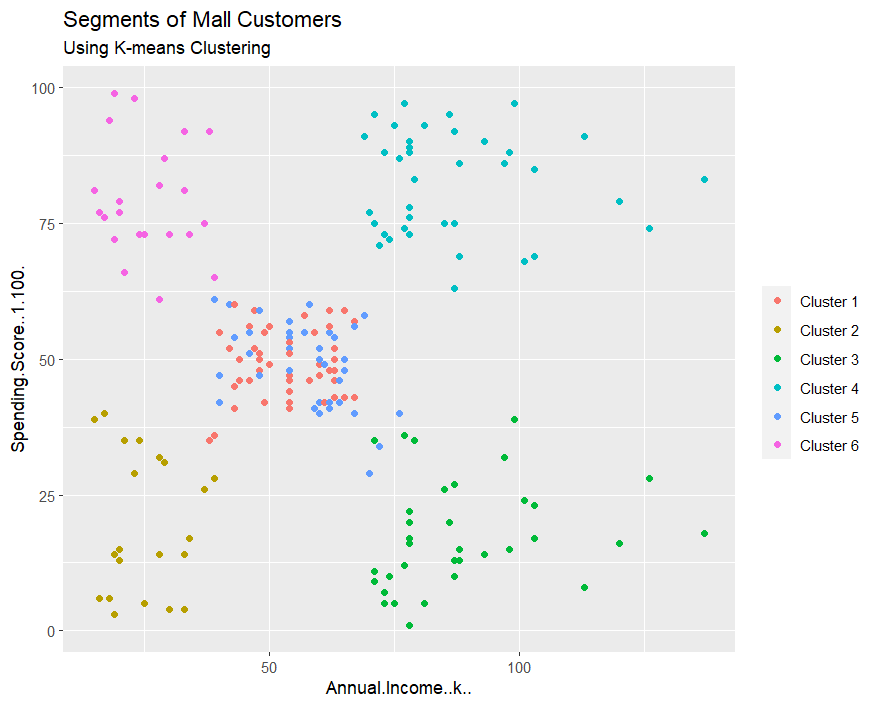
From the above descriptive analysis, we conclude that the minimum annual income of the population is 15 customers, the maximum annual income is 137, and the average annual income of 70 has the highest frequency in the histogram distribution. everyone's average annual income customers are 60.56. In the kernel density plot above,

**Income is normally distributed.**

**K MEANS CLUSTERING:**

**FIGURE (14): Number of clusters K**

VISUALIZING THE CLUSTERING RESULTS USING THE FIRST TWO PRINCIPLE COMPONENTS:



**FIGURE (15) Annual Income Clustering FIGURE (16) Spending Amount clustering**

From the visualization above, we can see that there are distributions of 6 clusters as follows:

Clusters 6 and 4 – These clusters represent average salary income customer data

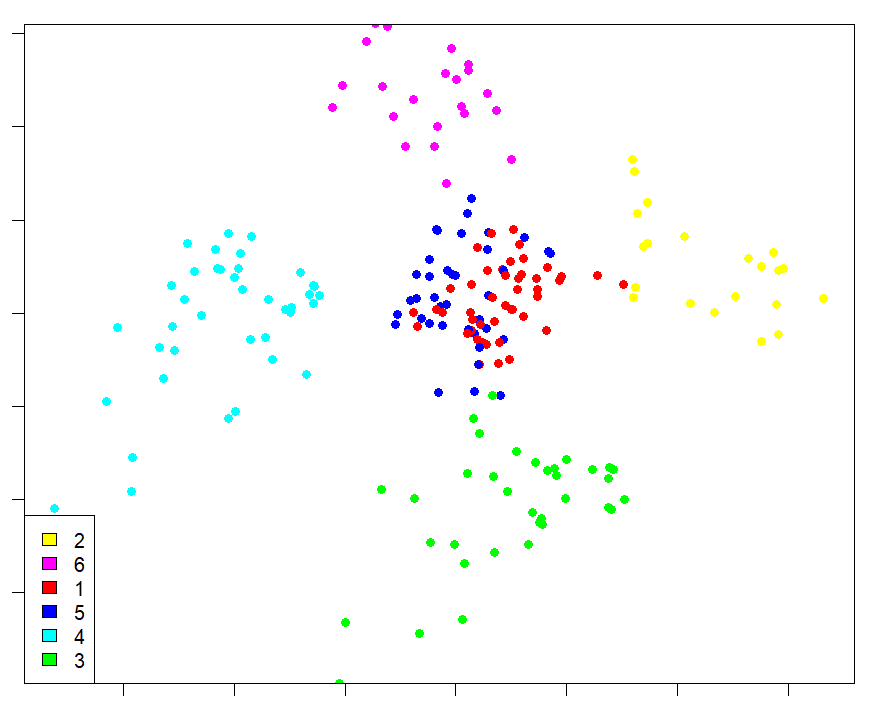
and average annual payroll expenditure.

Cluster 1 - This cluster represents the customer’s data who have high annual income and high annual expenditures.

Cluster 3 - This cluster represents the customer’s data who have low annual income and low annual expenditures.

Cluster 2 - This cluster represents the customer’s data who have high annual income and low annual expenditures.

Cluster 5 - This cluster represents the customer’s data who have low annual income and high annual expenditures.



**FIGURE (17) K Mean Visualization**

With the help of clustering, variables can be better understood by careful decision. Identifying customers allows companies to release products and income, age, spending patterns, etc. In addition, more complex patterns such as product ratings are considered for better segmentation.

**Future Scope:**

* More effective marketing strategy
* Optimizing the customer journey
* Predict customer behavior
* Personalizing the customer experience
* Improves customer loyalty and retention
* Supports Product Development
* Improves conversion Metrics

**Conclusion:**

Overall, the goals and applications defined in this study were achieved by machine learning models covering most aspects of average research work in the field of artificial intelligence for environmental science tasks. This work also demonstrates the importance of consulting a data scientist before starting monitoring, as datasets that are not suitable for the task requested are a common problem.

In general, regression models were able to show consistent trends and overall correlations

However, it provides a poor model for some measurements and quality. Random Forest (RF)

performs best and is recommended by scientists and scientists engineers working with environmental data.

A K-Nearest Neighbour (KNN) model has been used for data imputation and has been used successfully. It has been proposed to other researchers for this task. However, it is worth noting that many of the neighborhood data used for this study are not universal and may be found in different sets that are suitable for different datasets.

The classification model works well and can make very accurate predictions

A model for determining the season of the sample and the land use of the area where the sample was collected.

**References in APA style**

1. Applied Soft Computing

Volume 113, Part B, December 2021, 107924

Customer segmentation using K-means clustering and the adaptive particle swarm

optimization algorithm

YueLia XiaoquanChua DongTiana JianyingFenga WeisongMu

1. Expert Systems with Application

Volume 181, 1 November 2021, 115122

High utility itemset mining using binary differential evolution: An application to customer segmentation

Gutha JayaKrishna

1. Journal of Retailing and Customer Services

Volume 61, July 2021, 102588

Customer segmentation by web content mining

JinfengZhouac, JinliangWeib BugaoXu

1. Research on customer segmentation model by clustering

[ICEC '05: Proceedings of the 7th international conference on Electronic commerce](https://dl.acm.org/doi/proceedings/10.1145/1089551) August 2005 Pages 316–318

Authors: Jing Wu, Zheng Lin

1. Customer Segmentation using K-Means Clustering

Publisher IEEE

[Tushar Kansal](https://ieeexplore.ieee.org/author/37086921754); [Suraj Bahuguna](https://ieeexplore.ieee.org/author/37086919108); [Vishal Singh](https://ieeexplore.ieee.org/author/37086917437); [Tanupriya Choudhury](https://ieeexplore.ieee.org/author/37085794967)

[All Authors](javascript:void())­­­­