Customer Segmentation

Damaris Wanyoike

5/7/2022

Business Understanding

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 use this approach to gain a deeper understanding of the customer preferences as well as requirements for discovering segments that would reap maximum profits.Some of the objectives that we hope to answer are:Identifying the high spending customers, identify the age group of customers who usually go to the mall.

Data Origin/Description

The Mall customer data set is drawn from the data flair website. It contains information about people visiting the mall. The data set consists of 200 observations and 5 variables namely:gender,customer id, age, spending score and annual income. It collects insights from the data and group customers based on their behaviors.

# this provides the summary statistics and standard deviation of the customer's age.  
summary(customer\_data$Age)

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

sd(customer\_data$Age)

## [1] 13.96901

#this provides the summary statistics and standard deviation of the customers' annual income.  
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

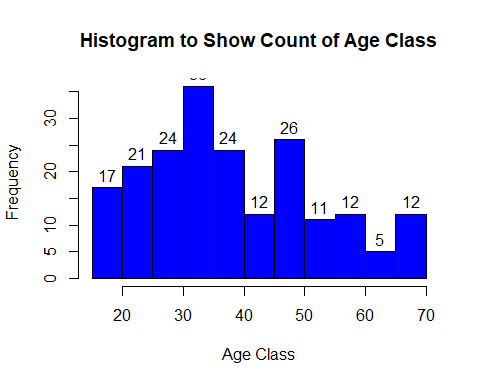
Data Preparation

In this particular example, there was no missing values. The only error was the gender column name had been misspelt.Some of the common issues that arise from using raw data include; having duplicate data, different data formats, poorly organized data(unstructured), incomplete data and having difficulty accessing the data.

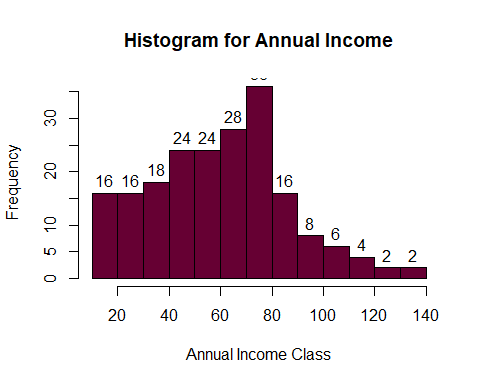
Modelling

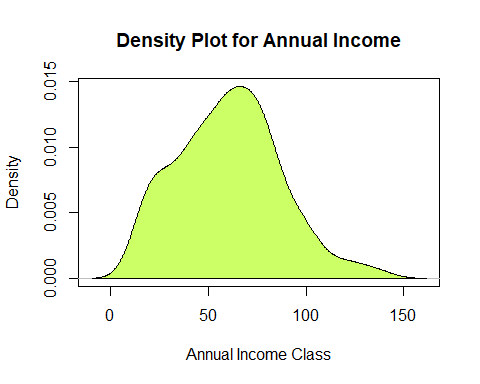
To answer the objectives, we will employ graph visualizations such as bar charts, histograms and pie charts to check the distribution of the age, annual income and annual spending of the customer data set.We will also use the K-Means Algorithm to implement the customer segmentation.

Customer Gender Visualization



The histogram above shows that most customers are in the age range between 30 to 35. The minimum age is 18 whereas the maximum age is 70.

Analysis of the annual income of Customers. 



From the two visualizations we can conclude that, people earning an average income of 70,have the highest frequency count in the histogram distribution and from the density plot, we observe that annual income has a normal distribution.

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. Then, the remaining objects have an assignment of the closest centroid.

#computes gap statistic  
k6<-kmeans(customer\_data[,3:5],6,iter.max=100,nstart=50,algorithm="Lloyd")  
k6

## K-means clustering with 6 clusters of sizes 38, 21, 45, 39, 35, 22  
##   
## Cluster means:  
## Age Annual.Income..k.. Spending.Score..1.100.  
## 1 27.00000 56.65789 49.13158  
## 2 44.14286 25.14286 19.52381  
## 3 56.15556 53.37778 49.08889  
## 4 32.69231 86.53846 82.12821  
## 5 41.68571 88.22857 17.28571  
## 6 25.27273 25.72727 79.36364  
##   
## Clustering vector:  
## [1] 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2  
## [38] 6 2 6 3 6 3 1 2 6 3 1 1 1 3 1 1 3 3 3 3 3 1 3 3 1 3 3 3 1 3 3 1 1 3 3 3 3  
## [75] 3 1 3 1 1 3 3 1 3 3 1 3 3 1 1 3 3 1 3 1 1 1 3 1 3 1 1 3 3 1 3 1 3 3 3 3 3  
## [112] 1 1 1 1 1 3 3 3 3 1 1 1 4 1 4 5 4 5 4 5 4 1 4 5 4 5 4 5 4 5 4 1 4 5 4 5 4  
## [149] 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5  
## [186] 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4  
##   
## Within cluster sum of squares by cluster:  
## [1] 7742.895 7732.381 8062.133 13972.359 16690.857 4099.818  
## (between\_SS / total\_SS = 81.1 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

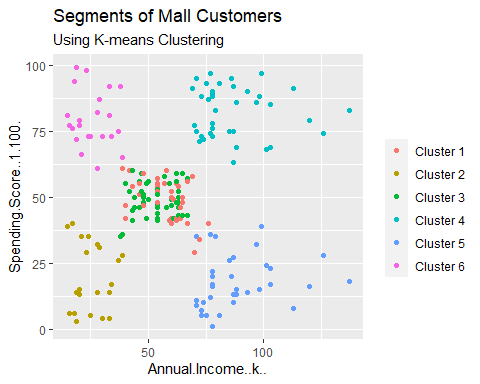
pcclust=prcomp(customer\_data[,3:5],scale=FALSE) #principal component analysis  
summary(pcclust)

## Importance of components:  
## PC1 PC2 PC3  
## Standard deviation 26.4625 26.1597 12.9317  
## Proportion of Variance 0.4512 0.4410 0.1078  
## Cumulative Proportion 0.4512 0.8922 1.0000

pcclust$rotation[,1:2]

## PC1 PC2  
## Age 0.1889742 -0.1309652  
## Annual.Income..k.. -0.5886410 -0.8083757  
## Spending.Score..1.100. -0.7859965 0.5739136

#Visualizes the clusters.  
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")



From the above visualization, we observe that there is a distribution of 6 clusters as follows –

Cluster 1 – This cluster represents the customer\_data having a high annual income as well as a high annual spend.

Cluster 2 – This cluster denotes a high annual income and low yearly spend.

Cluster 3 – This cluster denotes the customer\_data with low annual income as well as low yearly spend of income.

Cluster 5 – This cluster represents a low annual income but its high yearly expenditure.

Cluster 4 and 6 – These clusters represent the customer\_data with the medium income salary as well as the medium annual spend of salary.