Customer Segmentation Decision

Ibeh Amaka

2023-10-01

Targeted Customer Campaigns Retail businesses want a targeted marketing campaign to increase sales and customer engagement. They have customer data, including annual income and spending scores. They need to decide how to segment their customer base effectively for the marketing campaign

#knitr::opts\_chunk$set(echo = TRUE) #install all packages used # install.packages(“readr”) # install.packages(“tidyverse”) # install.packages(“tidyr”) # install.packages(“ggplot2”) # install.packages(“dylyer”) # install.packages(“data.table”) # install.packages(“cluster”) # install.packages(“plotly”)

Load the packages

load the dataset

mall <- read\_csv("Data/Mall\_Customers.csv")

## Rows: 200 Columns: 5  
## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (2): CustomerID, Genre  
## dbl (3): Age, Annual Income (k$), Spending Score (1-100)  
##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Performing EDA for this dataset

#EDA  
dim(mall)

## [1] 200 5

##contains 200 observation and 5 variables  
head(mall,10)

## # A tibble: 10 x 5  
## CustomerID Genre Age `Annual Income (k$)` `Spending Score (1-100)`  
## <chr> <chr> <dbl> <dbl> <dbl>  
## 1 0001 Male 19 15 39  
## 2 0002 Male 21 15 81  
## 3 0003 Female 20 16 6  
## 4 0004 Female 23 16 77  
## 5 0005 Female 31 17 40  
## 6 0006 Female 22 17 76  
## 7 0007 Female 35 18 6  
## 8 0008 Female 23 18 94  
## 9 0009 Male 64 19 3  
## 10 0010 Female 30 19 72

tail(mall,10)

## # A tibble: 10 x 5  
## CustomerID Genre Age `Annual Income (k$)` `Spending Score (1-100)`  
## <chr> <chr> <dbl> <dbl> <dbl>  
## 1 0191 Female 34 103 23  
## 2 0192 Female 32 103 69  
## 3 0193 Male 33 113 8  
## 4 0194 Female 38 113 91  
## 5 0195 Female 47 120 16  
## 6 0196 Female 35 120 79  
## 7 0197 Female 45 126 28  
## 8 0198 Male 32 126 74  
## 9 0199 Male 32 137 18  
## 10 0200 Male 30 137 83

sum(is.na(mall))

## [1] 0

str(mall)

## spc\_tbl\_ [200 x 5] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ CustomerID : chr [1:200] "0001" "0002" "0003" "0004" ...  
## $ Genre : chr [1:200] "Male" "Male" "Female" "Female" ...  
## $ Age : num [1:200] 19 21 20 23 31 22 35 23 64 30 ...  
## $ Annual Income (k$) : num [1:200] 15 15 16 16 17 17 18 18 19 19 ...  
## $ Spending Score (1-100): num [1:200] 39 81 6 77 40 76 6 94 3 72 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. CustomerID = col\_character(),  
## .. Genre = col\_character(),  
## .. Age = col\_double(),  
## .. `Annual Income (k$)` = col\_double(),  
## .. `Spending Score (1-100)` = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

#remove white spaces from column names  
colnames(mall) <- gsub(" ", "", colnames(mall))  
##rename columns  
str(mall)

## spc\_tbl\_ [200 x 5] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ CustomerID : chr [1:200] "0001" "0002" "0003" "0004" ...  
## $ Genre : chr [1:200] "Male" "Male" "Female" "Female" ...  
## $ Age : num [1:200] 19 21 20 23 31 22 35 23 64 30 ...  
## $ AnnualIncome(k$) : num [1:200] 15 15 16 16 17 17 18 18 19 19 ...  
## $ SpendingScore(1-100): num [1:200] 39 81 6 77 40 76 6 94 3 72 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. CustomerID = col\_character(),  
## .. Genre = col\_character(),  
## .. Age = col\_double(),  
## .. `Annual Income (k$)` = col\_double(),  
## .. `Spending Score (1-100)` = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

mall <- mall %>%  
 rename(Gender = Genre,  
 AnnualIncome = `AnnualIncome(k$)`,  
 SpendingScore = `SpendingScore(1-100)`)

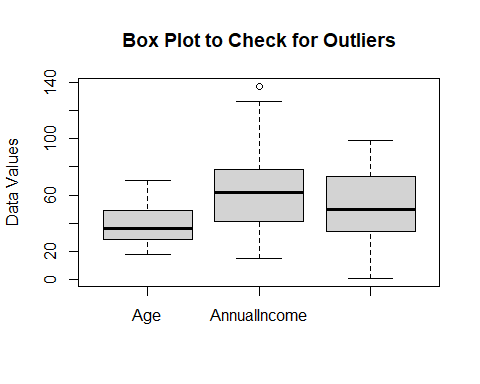
#Get the summary of the data however we do not need customerid  
#i will need this at the end  
mall\_new = mall[1:3]  
mall = mall[-1]  
summary(mall)

## Gender Age AnnualIncome SpendingScore   
## Length:200 Min. :18.00 Min. : 15.00 Min. : 1.00   
## Class :character 1st Qu.:28.75 1st Qu.: 41.50 1st Qu.:34.75   
## Mode :character Median :36.00 Median : 61.50 Median :50.00   
## Mean :38.85 Mean : 60.56 Mean :50.20   
## 3rd Qu.:49.00 3rd Qu.: 78.00 3rd Qu.:73.00   
## Max. :70.00 Max. :137.00 Max. :99.00

##check for categorical varaible  
#cat <- sapply(mall, is.character)  
#cat\_name <- names(cat[cat])  
#checking for outliers

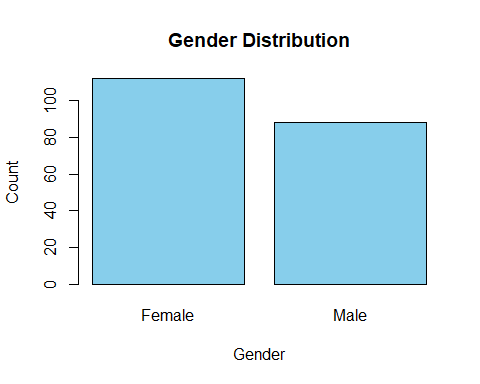
Lets check for outliers using box plot

numeric\_data <- mall[,sapply(mall, is.numeric)]  
boxplot(numeric\_data,   
 main="Box Plot to Check for Outliers",  
 ylab="Data Values")



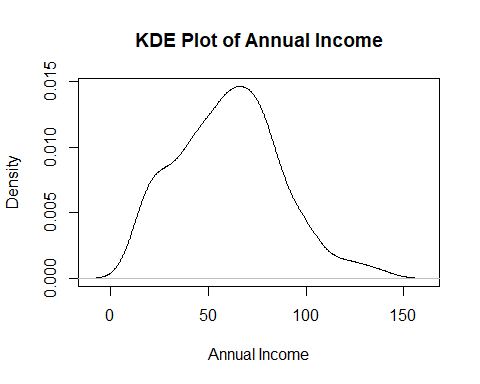
From the box plot we can see that there is an outlier in annual income since we have one outlier in annual income,we neglect it as it may be the person with higher annual income.

gender\_counts <- table(mall$Gender)  
barplot(gender\_counts,   
 main = "Gender Distribution",  
 xlab = "Gender",  
 ylab = "Count",  
 col = "skyblue", # Bar color  
 names.arg = names(gender\_counts))

 from the results we see that females are more than males

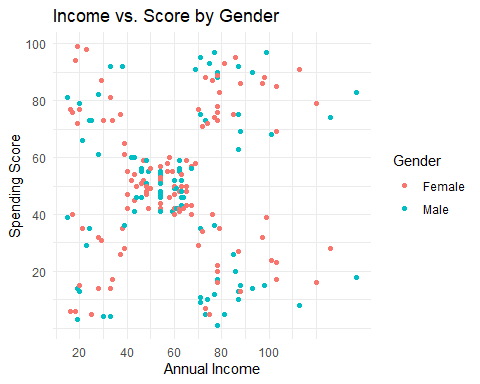
Calculate the kernel density estimate

density\_est <- density(mall$AnnualIncome)  
  
# Create a KDE plot using the plot() function  
plot(density\_est,   
 main = "KDE Plot of Annual Income",  
 xlab = "Annual Income",  
 ylab = "Density")

 The data is most concentrated at 60 which is the peak, the tail is from 120 to 150 where data are less concentrated.

Lets see a scattered plot Scatter plot for the Annual Income & SpendingScore by gender

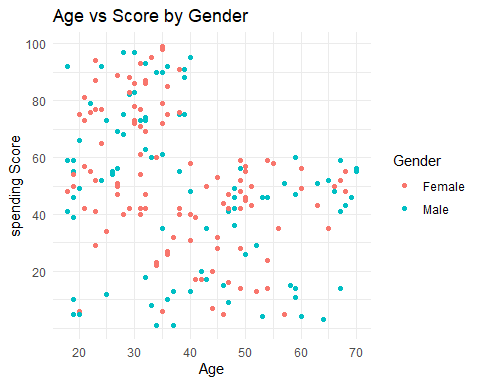
ggplot(mall, aes(x = AnnualIncome, y = SpendingScore, color = Gender)) +  
 geom\_point() +  
 labs(title = "Income vs. Score by Gender", x = "Annual Income", y = "Spending Score") +  
 theme\_minimal() +  
 scale\_x\_continuous(breaks = seq(20, 100, by = 20)) +  
 scale\_y\_continuous(breaks = seq(20, 100, by = 20))



We can see that most of the customer data points lies at annual income(40-70) and spending score (40-60). This shows that their income is corresponding to their spending score

#scatter plot for age and spending score

ggplot(mall, aes(x= Age, y=SpendingScore,colour=Gender))+  
 geom\_point() +  
 labs(title= "Age vs Score by Gender", x = "Age", y = "spending Score") +  
 theme\_minimal () +  
 scale\_x\_continuous(breaks = seq(10, 100, by = 10)) +  
 scale\_y\_continuous(breaks = seq(20, 100, by = 20))



We interpret that the age (40-60) of having spending score around (20-60), the age (20-40) of having higher spending score around (40-100) and the age (60-70) has balanced spending score around (40-60) respectively .

DATA PREPROCESSING

changing gender to 1 and 0 male = 1 female is 0

mall$Gender <- ifelse(mall$Gender =="Male", 1, 0)

Variables often have different units and scales. For example, “AnnualIncome” might be in thousands of dollars, while “SpendingScore” could be on a scale from 0 to 100. Scaling ensures that variables are on a similar scale, making comparing and analysing them easier.

scale\_col <- c("AnnualIncome", "SpendingScore")  
mall[scale\_col] <- scale(mall[scale\_col])  
head(mall,5)

## # A tibble: 5 x 4  
## Gender Age AnnualIncome SpendingScore  
## <dbl> <dbl> <dbl> <dbl>  
## 1 1 19 -1.73 -0.434  
## 2 1 21 -1.73 1.19   
## 3 0 20 -1.70 -1.71   
## 4 0 23 -1.70 1.04   
## 5 0 31 -1.66 -0.395

Decision Steps: Data Analysis: Analyze the customer data, including annual income and spending score, to understand the distribution and characteristics of the customer base. Segmentation Method: Decide on a segmentation method. In this case, the business will use K-means clustering to group customers based on their annual income and spending score

we will use our target features which is annual income and spending score

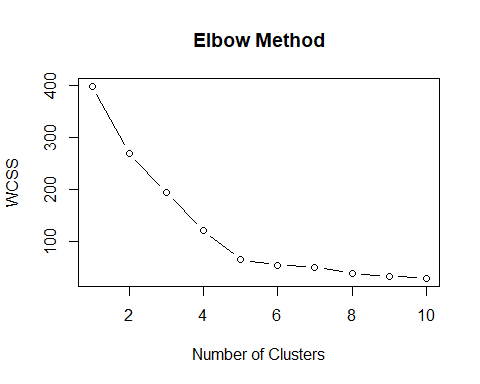
new\_mall = mall[3:4]

Number of Clusters: Decide on the number of clusters (segments) to create. For this, i will run the Elbow Method to determine the optimal number of clusters.

wcss <- numeric(10) # Initialize a vector to store WCSS values  
for (i in 1:10) {  
 kmeans\_model <- kmeans(new\_mall, centers = i)  
 wcss[i] <- kmeans\_model$tot.withinss  
}  
wcss

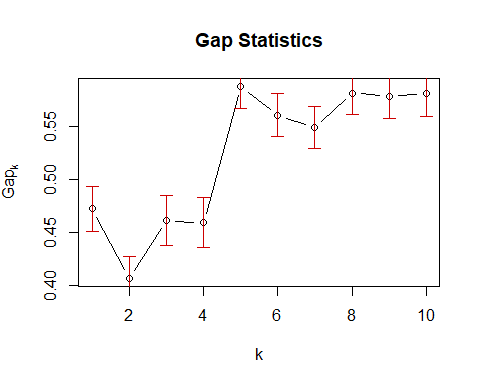
## [1] 398.00000 269.53790 194.27040 122.21498 65.24057 55.09894 50.90937  
## [8] 39.26966 33.75952 29.56752

plot(1:10, wcss, type = "b", main = "Elbow Method", xlab = "Number of Clusters", ylab = "WCSS")



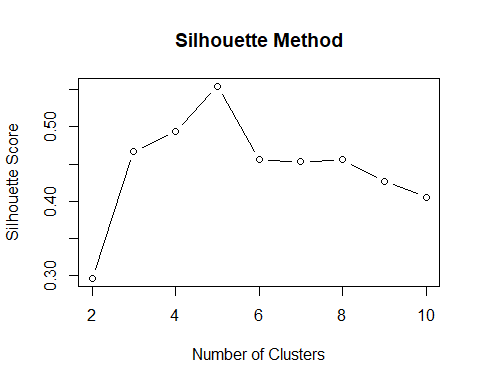
Another method to check the accuracy of the k-means

gap\_stat <- clusGap(new\_mall, FUN = kmeans, nstart = 25, K.max = 10)  
plot(gap\_stat, main = "Gap Statistics")



Want to be sure of our cluster

silhouette\_scores <- numeric(10) # Initialize a vector to store silhouette scores  
for (i in 2:10) {  
 kmeans\_model <- kmeans(new\_mall, centers = i)  
 silhouette\_scores[i] <- cluster.stats(dist(new\_mall), kmeans\_model$cluster)$avg.silwidth  
}  
plot(2:10, silhouette\_scores[2:10], type = "b", main = "Silhouette Method", xlab = "Number of Clusters", ylab = "Silhouette Score")



From the Plot above we see that the k-means is 5 Which is the best fit for the clusters

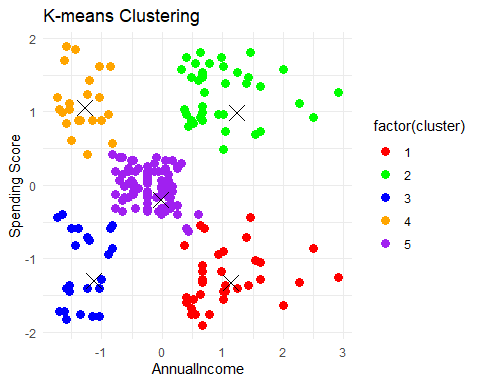
optimal\_clusters = 5  
kmeans\_model <- kmeans(new\_mall, centers = optimal\_clusters)  
print(kmeans\_model)

## K-means clustering with 5 clusters of sizes 22, 23, 35, 81, 39  
##   
## Cluster means:  
## AnnualIncome SpendingScore  
## 1 -1.3262173 1.12934389  
## 2 -1.3042458 -1.13411939  
## 3 1.0523622 -1.28122394  
## 4 -0.2004097 -0.02638995  
## 5 0.9891010 1.23640011  
##   
## Clustering vector:  
## [1] 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2  
## [38] 1 2 1 2 1 2 4 2 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4  
## [75] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4  
## [112] 4 4 4 4 4 4 4 4 4 4 4 4 5 3 5 4 5 3 5 3 5 4 5 3 5 3 5 3 5 3 5 4 5 3 5 3 5  
## [149] 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3  
## [186] 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5  
##   
## Within cluster sum of squares by cluster:  
## [1] 5.217630 7.577407 18.304646 14.485632 19.655252  
## (between\_SS / total\_SS = 83.6 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

# Run K-means clustering with the chosen number of clusters  
final\_kmeans\_model <- kmeans(new\_mall, centers = optimal\_clusters)  
# Get cluster assignments for each data point  
cluster\_assignments <- final\_kmeans\_model$cluster  
##add the predicted values to the features  
new\_mall$cluster <- cluster\_assignments

Visualize the model

centroids <- aggregate(new\_mall[c("SpendingScore", "AnnualIncome")], by = list(cluster = cluster\_assignments), FUN = mean)  
clr <- c("red", "green", "blue", "orange", "purple")  
# Create a scatter plot of clusters with centroids  
ggplot(new\_mall, aes(x = AnnualIncome, y = SpendingScore, color = factor(cluster))) +  
 geom\_point(size = 3) +  
 geom\_point(data = centroids, aes(x = SpendingScore, y = AnnualIncome), color = "black", size = 5, shape = 4) +  
 labs(title = "K-means Clustering", x = "AnnualIncome", y = "Spending Score") +  
 scale\_color\_manual(values = clr) + # Use the same cluster colors as needed  
 theme\_minimal()



Cluster Interpretation: cluster 1 (red) The data points in this cluster represents the customers with an low annual income who tends to have low spending (-2 to -0.5). This customers are balanced customers.Thy keep to budject based on their income leavel. Target for discounts

custer 2 (green). The data points in this cluster represents the customers with high income that has low spending (-2 to -1). They have high income but spend less, how do we market our products to make them spend more since they still have high disposable income. Target for promotions

cluster 3 (blue). The data points in this cluster represents the customers that have average income and avaerage spending. This are the balanced ones (-0.5 to 0.5). Target for promotions

cluster 4 (orange). The data points in this cluster represents the customers with high income and high spending. (1 to 2). This are the target for premium products

#cluster 5 (purple) The data points in this cluster represents the customers with low income and high spending (1 to 2). (Target for promotions) Marketing Strategy: Based on the cluster interpretation, I decide on a marketing strategy for each segment. High-income customers should receive exclusive offers on premium products. Moderate-income customers should receive promotions and loyalty rewards. Low-income customers should receive discounts and incentives.

Now is to select the customers based on customer ID for targeted marketing

# Now, you can use rbind without errors  
combined\_data <- cbind(mall\_new,new\_mall)

now filter customers in cluster 1

cluster\_1\_customers <- combined\_data %>%  
 filter(cluster == 1)  
head(cluster\_1\_customers,5)

## CustomerID Gender Age AnnualIncome SpendingScore cluster  
## 1 0125 Female 23 0.3594175 -0.820957 1  
## 2 0129 Male 59 0.3974914 -1.517996 1  
## 3 0131 Male 47 0.3974914 -1.595445 1  
## 4 0135 Male 20 0.4736391 -1.750342 1  
## 5 0137 Female 44 0.4736391 -1.672893 1

#cluster 1 is target for discount  
cluster\_2\_customers <- combined\_data %>%  
 filter(cluster == 2)  
head(cluster\_2\_customers,5)

## CustomerID Gender Age AnnualIncome SpendingScore cluster  
## 1 0124 Male 39 0.3213436 1.5799549 2  
## 2 0126 Female 31 0.3594175 1.0378135 2  
## 3 0128 Male 40 0.3974914 1.7348525 2  
## 4 0130 Male 38 0.3974914 0.9603648 2  
## 5 0132 Male 39 0.3974914 0.9603648 2

#cluster 2 is target for promotions  
cluster\_3\_customers <- combined\_data %>%  
 filter(cluster == 3)  
head(cluster\_3\_customers,5)

## CustomerID Gender Age AnnualIncome SpendingScore cluster  
## 1 0001 Male 19 -1.734646 -0.4337131 3  
## 2 0003 Female 20 -1.696572 -1.7116178 3  
## 3 0005 Female 31 -1.658498 -0.3949887 3  
## 4 0007 Female 35 -1.620425 -1.7116178 3  
## 5 0009 Male 64 -1.582351 -1.8277910 3

#cluster 3 is target for promotions  
cluster\_4\_customers <- combined\_data %>%  
 filter(cluster == 4)   
head(cluster\_4\_customers,5)

## CustomerID Gender Age AnnualIncome SpendingScore cluster  
## 1 0002 Male 21 -1.734646 1.1927111 4  
## 2 0004 Female 23 -1.696572 1.0378135 4  
## 3 0006 Female 22 -1.658498 0.9990891 4  
## 4 0008 Female 23 -1.620425 1.6961281 4  
## 5 0010 Female 30 -1.582351 0.8441916 4

#cluster 4 is target for premium products  
cluster\_5\_customers <- combined\_data %>%  
 filter(cluster == 5)  
head(cluster\_5\_customers,5)

## CustomerID Gender Age AnnualIncome SpendingScore cluster  
## 1 0044 Female 31 -0.8208730 0.4182234 5  
## 2 0047 Female 50 -0.7827991 0.1858770 5  
## 3 0048 Female 27 -0.7827991 -0.1239180 5  
## 4 0049 Female 29 -0.7827991 -0.3175400 5  
## 5 0050 Female 31 -0.7827991 -0.3175400 5

#cluster 5 is target for promotions