Customer Segmentation Final Report

**(a). Problem Statement:**

For my final project for this course, I chose to work with the topic of customer segmentation. I wanted to specifically focus on the annual income of the customers in correlation to their annual spending score. These two segments were chosen as the focal point for my final project because I wanted to address the problem of annual incomes accurately reflecting an individual’s spending score. This was the problem statement that I decided to focus on because I intended to show that the annual incomes of an individual are not necessarily a predictor of annual spending scores. Thus, just because I customer has a high or low annual income doesn’t necessarily mean that their annual spending score will directly correlate. There is the possibility that it could be the same, but also the possibility that it could be the opposite of what one would predict it to be.

**(b). Explanation of Real-World Data:**

Since the final exam project was geared towards working with real-world data, I decided to work with data from Kaggle. I chose to work with Kaggle because it is a reputable platform where real-world data sets can be found so that further analysis can be conducted that is related to data-science. Additionally, all the datasets on Kaggle can integrate machine learning algorithms into their datasets to solve real-world problems. As mentioned above, I chose to work with the topic of customer segmentation. I did this using real-world data in the form of a .csv file below. The .csv file included the following segments: Customer ID, Gender, Age, Annual Income, and Spending Score. My focus was clustering customers by their annual income in correlation to their annual spending score. The real-world data can be referenced below with an image of the .csv file as well as a reference to Kaggle.

**Reference to the real-world data set in .csv format and the link to Kaggle:**

**Graphical user interface, application, table, Excel

Description automatically generated**

<https://www.kaggle.com/search?q=cutomer+segmentation>

**(c). Approach:**

I approached the project by first loading the necessary packages required to complete my coding which was the library(factoextra). This package was installed so that the extraction and visualization process of the exploratory multivariate data would be easier. Then, I read the .csv file and selected the numerical variables for scaling which focused on the annual income and the annual spending score. After I scaled the numerical values of the real-world dataset, I found the value of K-means through unsupervised learning. I chose to use unsupervised learning to find the value of K-means because it can be: (1) easily integrated into datasets and when reading results, (2) adapts well to changes when clustering, (3) works well with hyper spherical-clusters, (4) increases time execution, (5) produces tighter clusters, (6), has a good computational cost, and improves the clustering accuracy. When implementing the unsupervised learning, the value of K ended by being 5. This meant that there were going to be clusters in my analysis.

First, I chose to integrate the Elbow Method to find the value of K-means. This method was implemented as it is a heuristic used in determining the number of clusters in a data set. The method plots the explained variation as a function of the number of clusters, and then, picking the elbow of the curve as the number of clusters to use. As this Elbow Method is not always optimal in determining the value of K-means (especially when the data is not very clustered), I decided to also attempt to find the optimal value of K-means with using the Silhouette Method.

Unlike the Elbow Method, the Silhouette Method looks at the separation distance that exists between the different clusters. When the code is used, the silhouette plot shows a measure of the proximal distance of how close points are in one cluster in relation to neighboring clusters. Therefore, allowing the distinct parameters of each cluster to be clearly visualized. References to my approach are in the following images below.

After finding the value of K-means is equal to 5, I then clustered the five segments with the optimal value of K-means that was found using the Elbow Method and Reaffirmed using the Silhouette Method.

**Step 1: Installing Libraries and Finding the Value of K-means**

#First, I will read the .csv file.   
library(factoextra)

## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

cust<- read.csv("Customers.csv")  
head(cust)

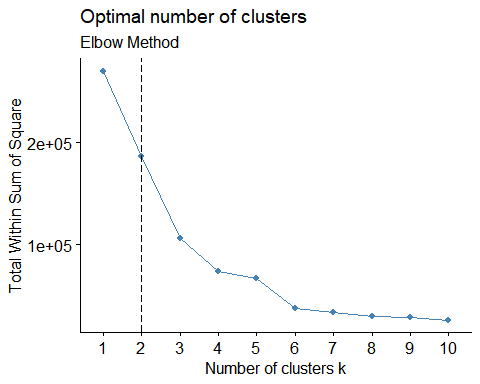
## CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100.  
## 1 1 Male 19 15 39  
## 2 2 Male 21 15 81  
## 3 3 Female 20 16 6  
## 4 4 Female 23 16 77  
## 5 5 Female 31 17 40  
## 6 6 Female 22 17 76

#Taking the quantitative variables in order to scale.   
cust1<-cust[,4:5]  
head(cust1)

## Annual.Income..k.. Spending.Score..1.100.  
## 1 15 39  
## 2 15 81  
## 3 16 6  
## 4 16 77  
## 5 17 40  
## 6 17 76

#Finding the value of K-means using unsupervised learning. Wanted to use the simplest, but most accurate method possible.   
fviz\_nbclust(cust1,kmeans,method="wss")+geom\_vline(xintercept = 2,linetype= 5)+labs(subtitle = "Elbow Method")

**Step 2: Integration of the Elbow Method for Optimal Value of K-means**



fviz\_nbclust(cust1,kmeans,method ="silhouette") + labs (subtitle = "Silhouette Method")

**Step 3: Integration of the Silhouette Method for Optimal Value of K-means**

Chart, line chart

Description automatically generated

**Step 4: Clustering with the Value of K-means**

#Here, I will set the seed for kmeans.   
set.seed(1)  
k5<-kmeans(cust1, centers = 2, nstart = 25)  
k5$centers

## Annual.Income..k.. Spending.Score..1.100.  
## 1 79.60000 50.12727  
## 2 37.28889 50.28889

#Thus, K= 5, meaning that there will be 5 clusters.

#Clustering the data from .csv file.   
custclus<-kmeans(cust1,5)  
custclus

## K-means clustering with 5 clusters of sizes 81, 35, 22, 39, 23  
##   
## Cluster means:  
## Annual.Income..k.. Spending.Score..1.100.  
## 1 55.29630 49.51852  
## 2 88.20000 17.11429  
## 3 25.72727 79.36364  
## 4 86.53846 82.12821  
## 5 26.30435 20.91304  
##   
## Clustering vector:  
## [1] 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 5  
## [38] 3 5 3 5 3 5 1 5 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [75] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [112] 1 1 1 1 1 1 1 1 1 1 1 1 4 2 4 1 4 2 4 2 4 1 4 2 4 2 4 2 4 2 4 1 4 2 4 2 4  
## [149] 2 4 2 4 2 4 2 4 2 4 2 4 2 4 2 4 2 4 2 4 2 4 2 4 2 4 2 4 2 4 2 4 2 4 2 4 2  
## [186] 4 2 4 2 4 2 4 2 4 2 4 2 4 2 4  
##   
## Within cluster sum of squares by cluster:  
## [1] 9875.111 12511.143 3519.455 13444.051 5098.696  
## (between\_SS / total\_SS = 83.5 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

#Now, I will visualize the clusters.   
ggplot(cust1, aes(x = Annual.Income..k..,y = Spending.Score..1.100.)) +geom\_point(stat = "identity", aes(color=as.factor(custclus$cluster)))+ scale\_color\_discrete(name="k",breaks=c("1", "2", "3", "4", "5"),labels=c("Cluster1", "Cluster2","Cluster3", "Cluster4", "Cluster5"))+ ggtitle("Customer Segmentation")

**(d). Results:**

After running both the Elbow Method to find the optimal number of clusters and Silhouette Method, I was able to segment each of the clusters by their annual income and their annual spending score. In the first cluster, the customers who earn a medium annual income, have a medium annual spending score. This shows an exact correlation between the customers annual income and their annual spending score. The second cluster reflects customers who earn a high annual income, have low spending score. As a result, we can see that just because a customer earns a high annual income does not mean that they will have a high annual spending score too. The third cluster in this dataset is the opposite of my second cluster as the third cluster shows costumers that earn a low annual income, have a high annual spending score. Thus, annual income earned does not result in what one would predict their annual spending score to be. The fourth cluster shows that customers who earn a high annual income, have a high annual spending score. The fifth cluster shows that customers who earn a low annual income, have a low annual spending score. It can be concluded that clusters one, four, and five have the predicted correlations that one would expect. The customer segmentation plot produced by the Silhouette Method can be referenced below.

**Customer Segmentation Plot Produced by the Silhouette Method:**

Chart, scatter chart

Description automatically generated

**(e). Conclusions/Summary:**

My final conclusions are that there are times when a customer annual income, will not accurately correlate to the predicted annual spending score. As we have seen in the results, there are two clusters where the customers annual income, does not correlate with the predicate annual spending score. The assumption that I made going into this project was that all the clusters would have direct correlations which was not the case.