**FINAL PROJECT REPORT**

ON

**CUSTOMER SEGMENTATION**



**Instructor Presented by**

Dr. Chaojiang (CJ) Wu, Ph.D. Pravalika Girneni

Fundamentals of Machine Learning pgirneni@kent.edu

MIS-64060-003

Customer Segmentation

**Problem**

I wanted to focus on how the customers' annual income and annual spending score related. the annual income and annual spending score are the 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 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.

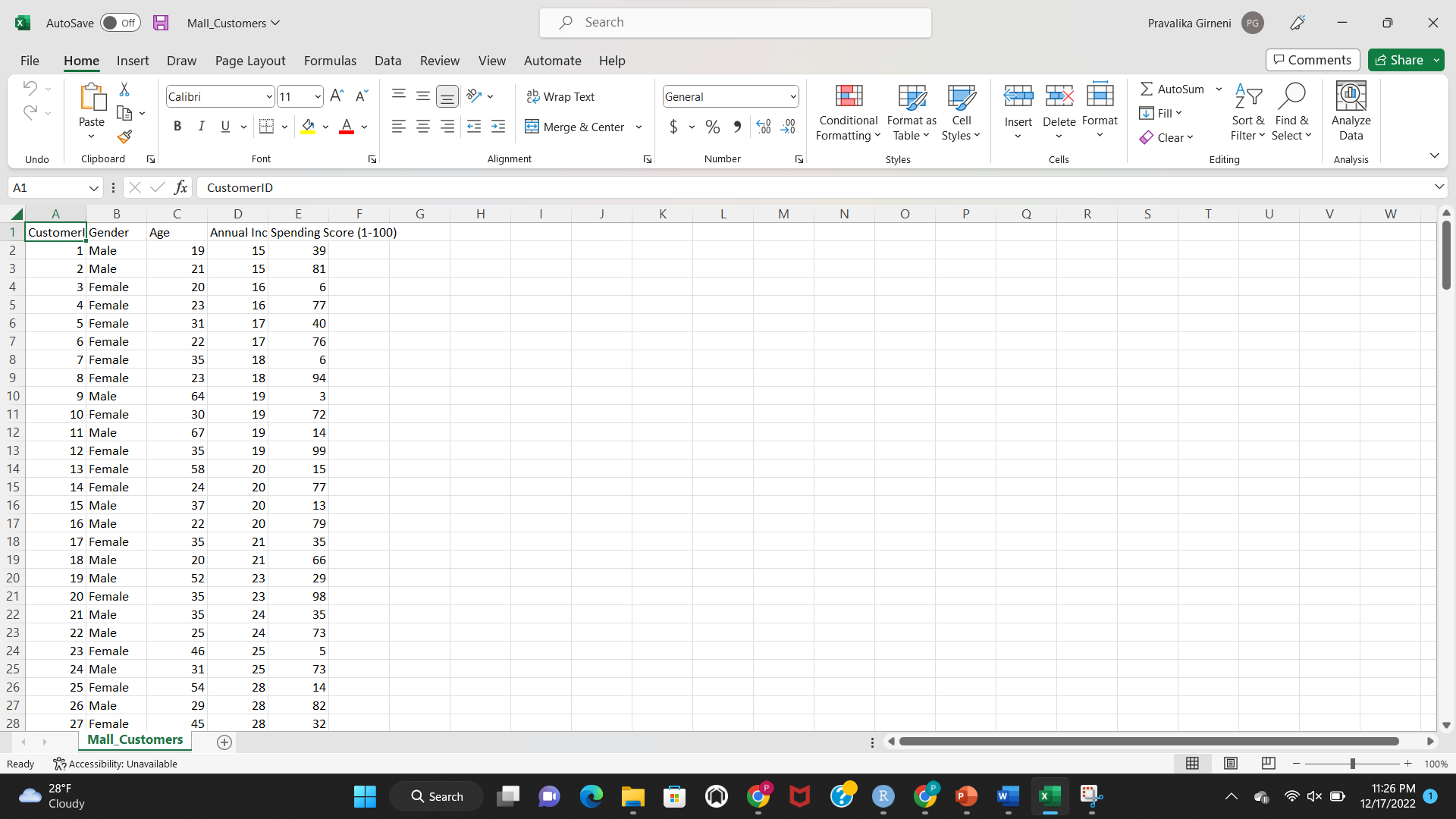
**Mall customer dataset explanation**

I have choose a data from 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. In order to address real-world issues, all of the datasets on Kaggle can incorporate machine learning methods into their datasets. , 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. 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.

**Overview of data**

* Customer ID
* Gender
* Age
* Annual Income
* Spending Score

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

****

**Dataset link**

[**https://www.kaggle.com/code/naren3256/kmeans-clustering-and-cluster-visualization-in-3d/data**](https://www.kaggle.com/code/naren3256/kmeans-clustering-and-cluster-visualization-in-3d/data)

**Approach**

I started the project by loading the packages I would need to finish the coding, which was the library (factoextra). This package was installed to make it simpler to extract and visualize the exploratory multivariate data. I next examined the.csv file and chose the numerical scaling variables, concentrating on the annual income and spending score. I discovered the utility of K-means after scaling the real-world dataset's numerical values. through unsupervised learning. I chose to use K-means is a good choice for unsupervised learning because it can be easily integrated into datasets and when reading results, adapts well to changes in clustering, works well with hyper spherical clusters, increases execution time, produces tighter clusters, has a low computational cost, and increases clustering accuracy. The final value of K after implementing unsupervised learning was 5. This implied that my analysis will contain clusters.

I first choose to combine the Elbow Method in order to determine the value of K-means. This method was used since it is a heuristic for determining the number of clusters in a set of data. The method plots the explained variance as a function of the number of clusters, selecting the elbow of the curve as the appropriate number of clusters. Because the Elbow Method is not always the ideal for doing so, I decided to additionally try to determine the optimal value of K-means using the Silhouette Method (particularly when the data are not strongly clustered).

The Silhouette Method takes into account the separation distance between the distinct clusters, in contrast to the Elbow Method. The silhouette changes once the code is applied a measurement of the proximal distance, or how close points in one cluster are to those in surrounding clusters, is shown on the silhouette plot. As a result, it is easy to see each cluster's unique characteristics in great detail. The following graphics serve as references for my methodology.

After determining that K-means is equal to 5, I grouped the five segments using the Elbow Method's optimal K-means value and the Silhouette Method's reaffirmation.

##calling the required library

**library**(factoextra)

Reading the csv file

Mall\_Data<- read.csv("C:/Users/girne/Downloads/Mall\_Customers.csv")

##printing the top portion data file

head(Mall\_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

##Taking the quantitative variables in order to scale.

Mall\_Data1<-Mall\_Data[,4:5]

head(Mall\_Data1)

## Annual.Income..k.. Spending.Score..1.100.

## 1 15 39

## 2 15 81

## 3 16 6

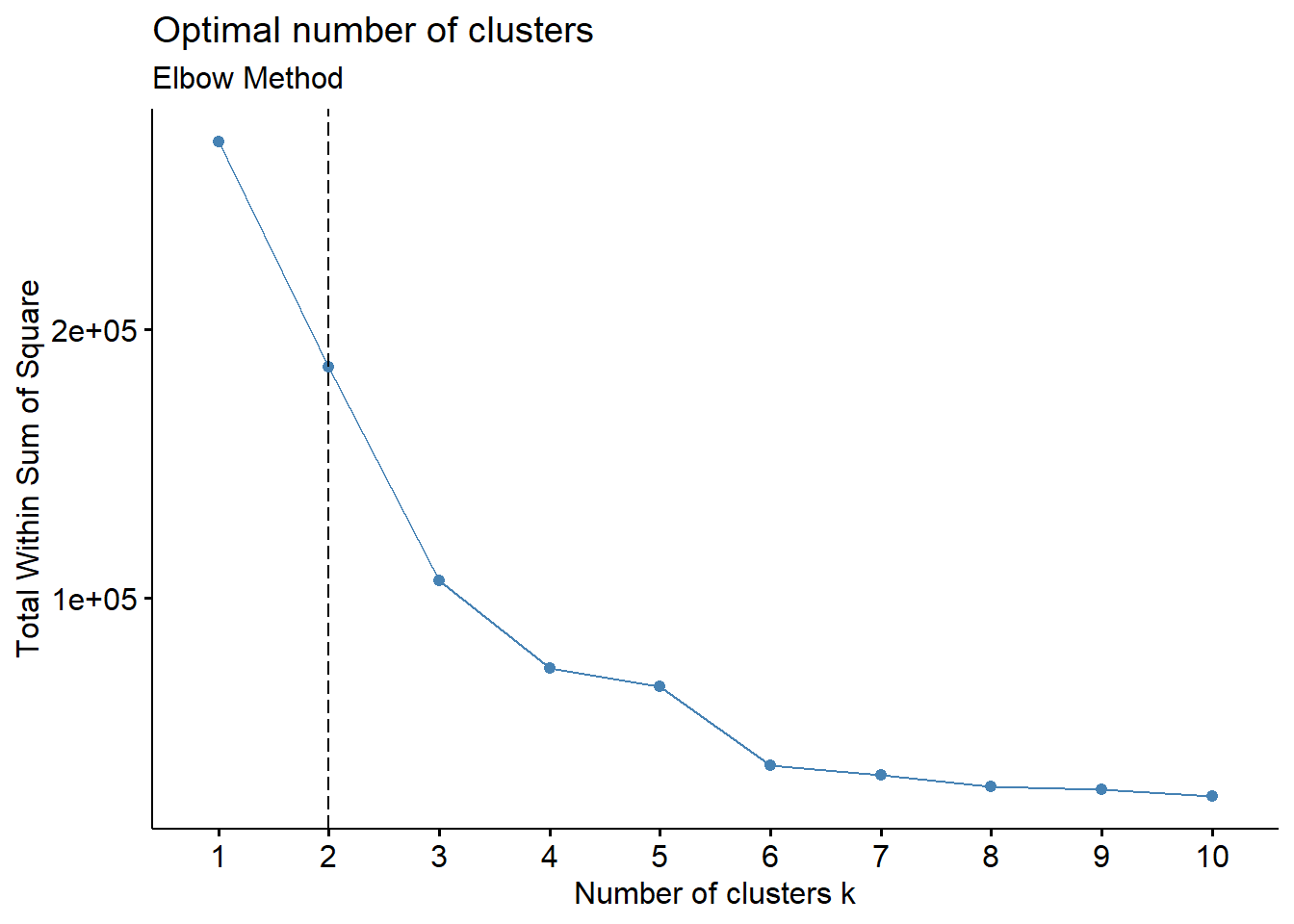
## 4 16 77

## 5 17 40

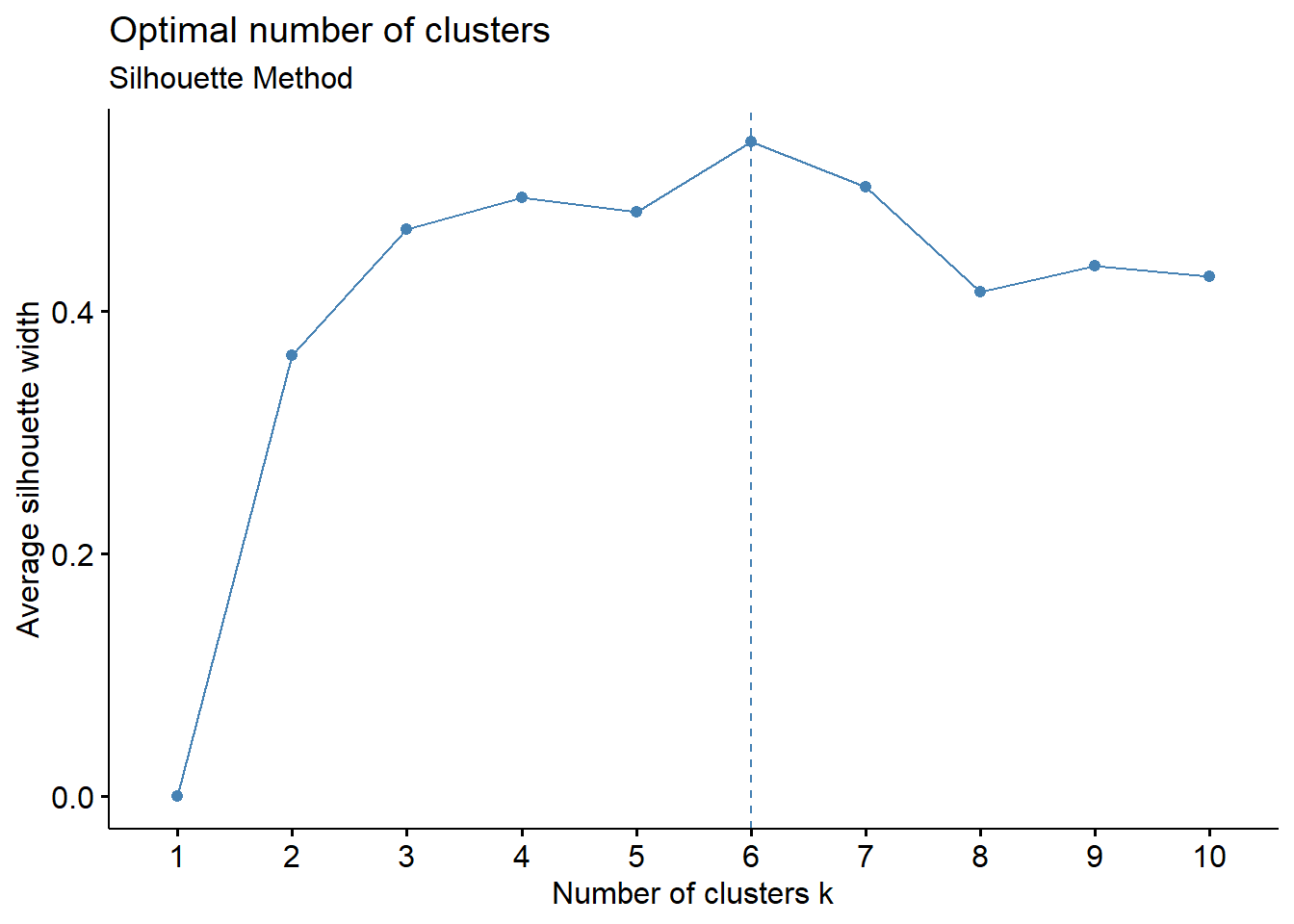
## 6 17 76

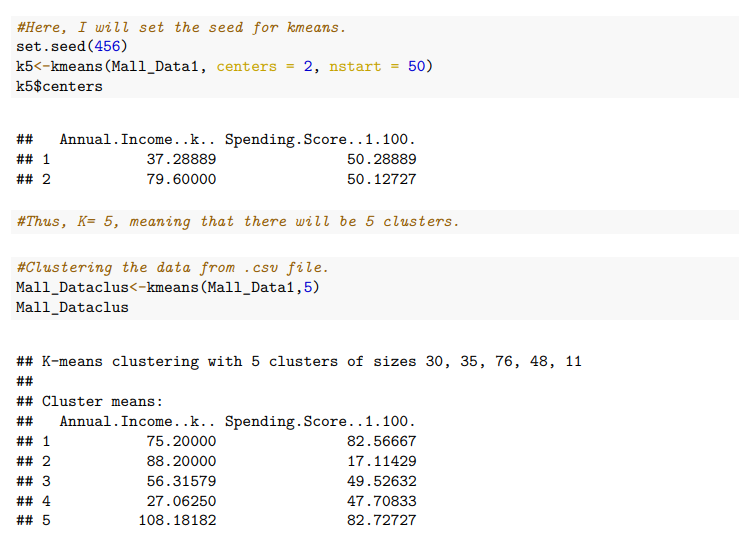
#Finding the value of K-means using unsupervised learning. Wanted to use the simplest, but most accurate method possible.

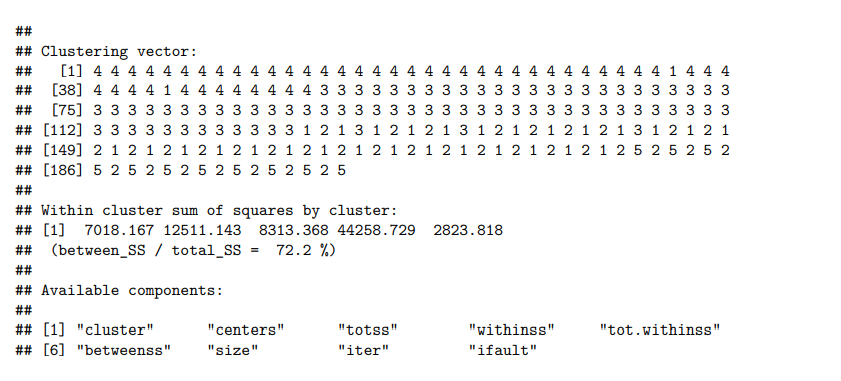
fviz\_nbclust(Mall\_Data1,kmeans,method="wss")+geom\_vline(xintercept = 2,linetype= 5)+labs(subtitle = "Elbow Method")

****

fviz\_nbclust(Mall\_Data1,kmeans,method ="silhouette") + labs (subtitle = "Silhouette Method")

****

****

****

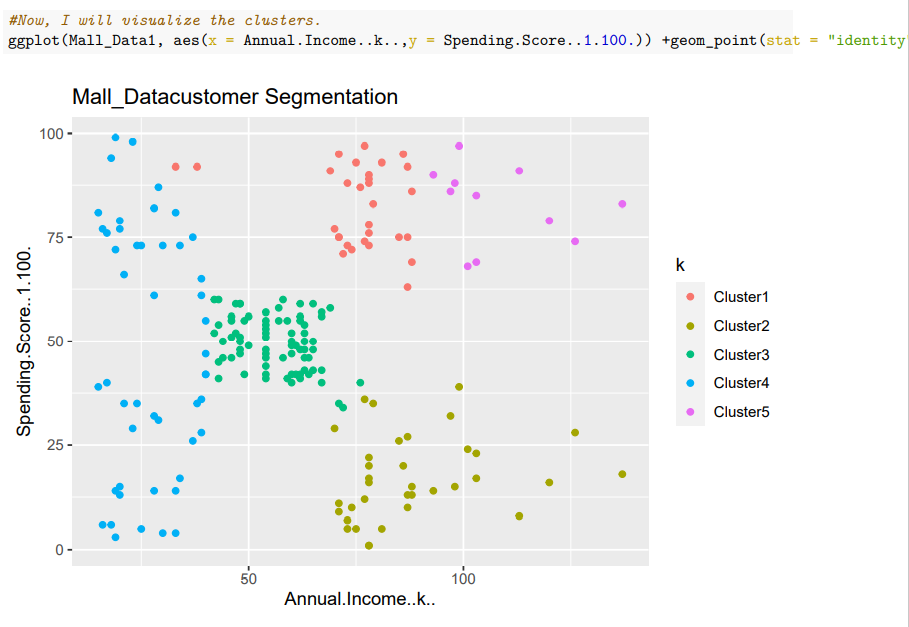
**Result**

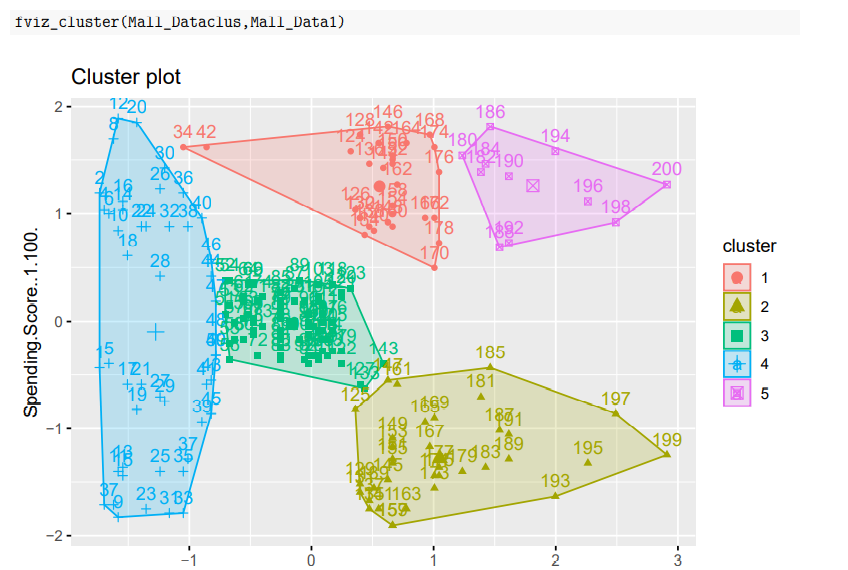
I was able to divide each of the clusters by their yearly income and annual spending score after using the Silhouette Method and the Elbow Method to determine the ideal number of clusters. Customers in the first cluster with medium yearly incomes also have medium annual expenditure scores.

This demonstrates a perfect link between the customer's annual income and the amount they spend each year. Customers in the second cluster are those with high annual incomes and low expenditure scores

As a result, a customer's high annual income does not necessarily translate into a high annual expenditure score. This dataset's third cluster is the opposite of my second cluster in that it displays customers with low annual incomes but high annual spending scores. As a result, annual revenue produced does not produce the annual spending score that one might anticipate.

The fourth cluster demonstrates that clients with high yearly incomes also have high annual spending scores. Customers with low annual incomes have low annual expenditure scores, according to the fifth cluster. Therefore, it may be said that clusters 1, 4, and 5 have the expected correlations. Below is a link to the consumer segmentation plot that the Silhouette Method generated.





**Conclusion**

My final findings are that there are instances where a customer's predicted year expenditure score and yearly income do not connect well. The data shows that there are two clusters where the customers' yearly income does not match the predicate annual expenditure score. I had assumed that all of the clusters in this experiment would be connected directly, but that was not the case.