Customer Segmentation using k-Means Clustering

October 22, 2021

A. Introduction

A1. Background

Customer Segmentation reports are considered essential tactical analysis tools and are often used by business owners to optimize products, services, and marketing strategies. To scale efficiently and effectively, companies need to focus their efforts on a specific subset of customers who are most similar to their best current customers. By dividing the customers in a given market into discrete groups based on their buying characteristics, a company can make marketing strategies that could drive an increase in profits and customer satisfaction.

It is critical to develop customer segment hypotheses and variables, and then validate them with a well-developed, scientific research process. In this project, we would take a data-driven approach for the segmentation of customers by utilizing k-Means Clustering based on their annual income and spending score. The aim is to use this algorithm to accurately classify customers and make good business decisions.

A2. Problem

For this project, the question that can be asked: “How to understand customer behavior such that we can segment easily a target audience to determine marketing strategies?”

A3. Target

The intended audience for this project could be private company executives who owns a mall, or other institutions that cater to a mass of people. It can be used to derive business decisions that could be beneficial to making more profit.

B. Data Acquisition and Cleaning

B1. Data Source

For this project, the following data is used:

1. Dataset that contains information about customers taken through membership cards:

Data Source = Kaggle: https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python

description: The dataset contains 200 entries and 5 attributes.

Matrix column entries (attributes):

- CustomerID: number given to customer
- Age: the age of the customer
- Gender: the sex of the customer
- Annual Income: how much the customer earns per year (in thousand dollars)
- Spending Score: the willingness of a customer to spend their money (1-100)

B2. Data Cleaning

The Kaggle Dataset initially have to be downloaded in the data assets of Watson Studio. The file has 20 entries and 5 attributes.

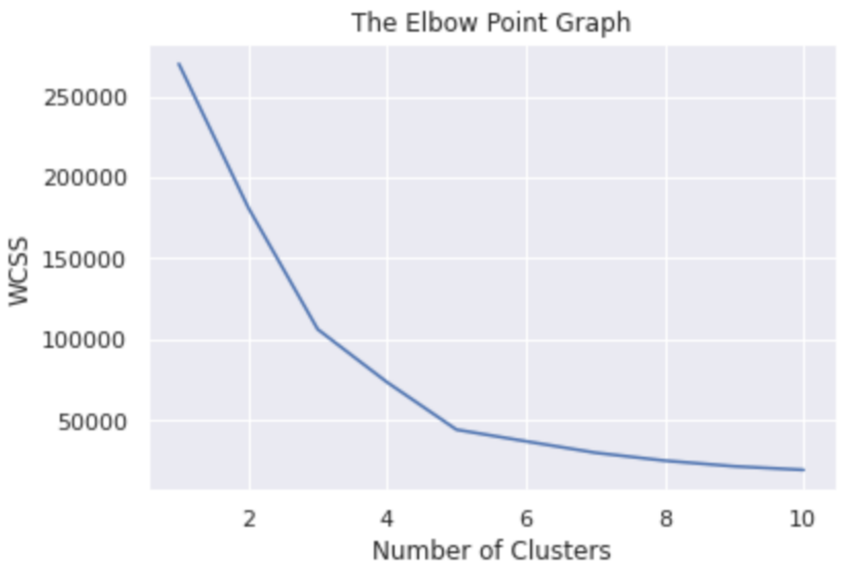
We initially took the information about the dataset. This is to see what type each attribute is and to see if there are attributes that gave missing values. When ensured that there is none, we started the pre-processing stage.

There are attributes in the dataset that are not necessary to converge customers in cluster, namely: CustomerID, age, and gender. For this analysis, we will group the customers based on their annual income and their spending score. Hence, we created an array consisting only of values from these two attributes and labelled this array as X. In the first column of the array consists of the Annual Income values and the second values represent the Spending Score. This array is the data we used for the analysis.

C. Exploratory Data Analysis

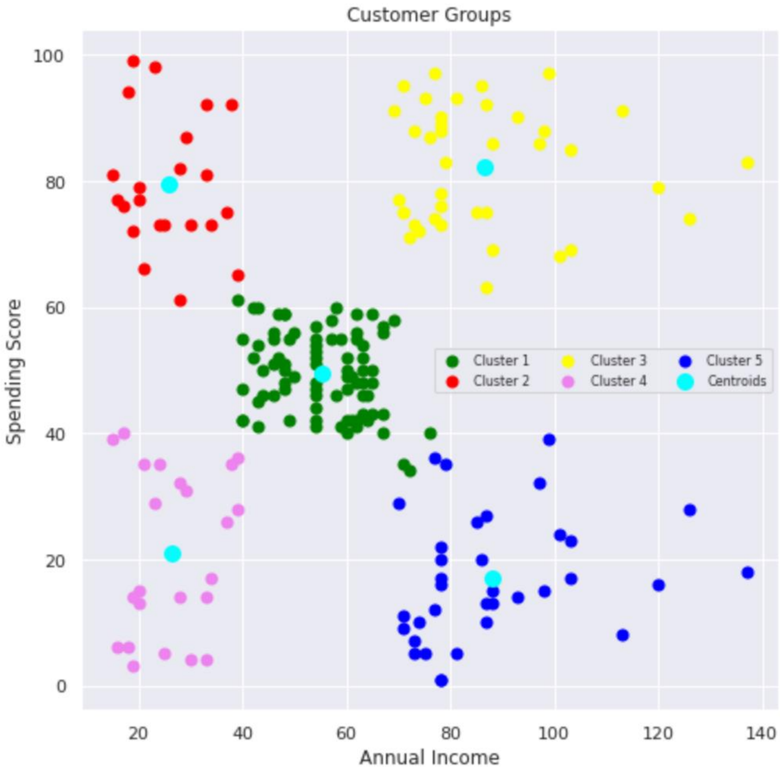
Before we actually go through the k-Means algorithm, we must choose the appropriate number of clusters for our dataset. To do this, we applied the Elbow Method. In this method, there are varying number of clusters (k) from 1-10. For each value of k, the Within-Cluster Sum of Square (WCSS) is calculated. This is the sum of squared distance between each point and the centroid in a cluster. As the number of clusters increases, the WCSS value will start to decrease. When the WCSS is plotted with the k value, the graph will rapidly change at a point and would create an elbow shape. The k value corresponding to this point is the optimal k value or an optimal number of clusters.

Looking at the graph below, we can deduce that the optimal number of clusters for this dataset is 5.



Now that the number of clusters is established, we can train the k-means clustering model. To do this, we used the KMeans function from sklearn.cluster library. We specified the number of clusters to form as 5 and the method for initialization as "k-means++". Then we computed cluster centers and predicted cluster index for each data in the array X. This returned an array of labels for each data point based on their cluster.

After training the model, we plotted all the clusters and their centroids. Below shows how the 200 customers were clustered.



In the graph, we can discern that Cluster 2 has low annual income yet has a high spending score; and Cluster 5 has a high annual income yet a low spending score. Cluster 4 have a less annual income and a low spending score. This makes sense as people with low income are less likely to spend money in supermarkets or malls. Cluster 3 consists of people with more annual income and more affinity to spend. Lastly, Cluster 1 has a moderate annual income and a moderate spending score.

If we were to use the output in the graph to make marketing strategies based on a target group, we can conclude that mall owners can give customers classified in Clusters 4 and 5 more discounts or promotions and offers to engage them to buy more. Engaging these customers with said strategies could help procure more profits for business owners and better customer satisfaction.

D. Conclusion

Classifying customers based on two factors – annual income and spending score – is enough for business owners to make great marketing strategies that could drive profits. To determine which customers would benefit more if given discounts and/or offers, companies can give these customers motives to spend more. From this clustering algorithm, we can segment customers and use the output to determine companies can decide what other services they can provide that would incline customers to make allocate more of their income to these services.