**Final Report: Machine Learning I : Capstone Project**

**Executive Summary**

* No outliers and values are evenly spread
* No findings of Correlation through heatmap
* Visually, there is single meaningful cluster
* Clusterwise, three cluster gives highest silhouette score
* PCA component = 2 gives organized cluster visualization
* About half/half is Male/Female
* average age is 54, but there is spread of age groups
* income ranges from 20k to 200k
* spreading score ranges from 1 to 100 and average of 50.59
* It's not a linear relationship silhouette score vs cluster
* There are no children on data, meaning it's left out from its dataset

**Conclusion**

Dataset like customer segmentation on supermarket grocery is simple yet the data is too broadly spread to gain meaningful insights. Its ideal choice is to gain more data on other features like race, opening and closing hours, cost of spending, date and time of shopping, number of children/teen shopping.

Because the data is broad, its ideal choice is to think about bringing in a broad customer base which includes having coupons and food samping on multiple isles as well as placing discounts on school supplies and toys to show children/teen friendly.

The ML approach on using K-means clustering and hierarchical clustering model becomes meaningless with data without correlation, and shows how vulnerable ML is when it comes to handling data that is ambiguous.

Kaggle Dataset: https://www.kaggle.com/datasets/zubairmustafa/shopping-mall-customer-segmentation-data/data

About Dataset

This dataset is designed for learning customer segmentation concepts, such as market basket analysis. It includes basic customer data such as Customer ID, age, gender, annual income, and spending score, which is assigned based on customer behavior and purchasing data. The goal is to help a supermarket mall owner understand their customers better, identify target customers who are likely to converge, and provide insights to the marketing team for strategic planning.

Description of the problem.

The primary challenge faced by the supermarket mall is the lack of a targeted marketing strategy that addresses the diverse needs and preferences of its customer base. Without a clear understanding of the different customer segments, the mall risks deploying generic marketing campaigns that fail to engage potential customers effectively.

Description of the dataset (dimensions, names of variables with their description)

Customer ID: A unique identifier for each customer.

Age: The age of the customer.

Gender: The gender of the customer (Male/Female).

Annual Income (k$): The customer's annual income in thousands of dollars.

Spending Score (1-100): A score assigned to the customer based on their behavior and purchasing data.

Description of the techniques that you will use..

I will apply EDA, Data Preprocessing, K-Means Clustering, Hierarchical Clustering, Principal Analysis Component Analysis (PCA), silhouette score.

Comments and/ or concerns?

As with most data, it would have been better if there were more features, but because it’s simple, it’s better for me to try algorithms with my understanding. Data might be flawed considering the usability score of 9.41.