**Name : Jashwanth Nallaparaju**

**Student Id: 23041020**

**CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING**

**Table of Contents**

[Introduction 3](#_Toc184919399)

[Overview of K-Means Clustering 3](#_Toc184919400)

[Key Steps in K-Means 3](#_Toc184919401)

[Dataset Description 4](#_Toc184919402)

[Implementation Steps 4](#_Toc184919403)

[Step 1: Loading and Exploring the Dataset 4](#_Toc184919404)

[Step 2: Data Preprocessing 5](#_Toc184919405)

[Step 3: Determining the Optimal Number of Clusters 5](#_Toc184919406)

[Step 4: Applying the K-Means Algorithm 6](#_Toc184919407)

[Step 5: Visualizing the Clusters 7](#_Toc184919408)

[Step 6: Saving and Interpreting the Results 8](#_Toc184919409)

[Insights from Clustering 8](#_Toc184919410)

[Challenges and Limitations 9](#_Toc184919411)

[Future Directions 9](#_Toc184919412)

[Conclusion 10](#_Toc184919413)

[Reference 11](#_Toc184919414)

[Appendix 13](#_Toc184919415)

# Introduction

Customer segmentation is a very important concept that forms the core of marketing as well as business intelligence. This involves categorizing customers in relation to their needs and actions, so that various business and marketing strategies can be developed and implemented as well as customers’ satisfaction enhanced while business processes are made more efficient. The most known of all clustering techniques is the K-Means clustering technique and forms the main part of the unsupervised machine learning. In this write-up, the author has used one of the Python implemented customer segmentation algorithms to explain K-Means theory in detail.

The dataset for this project Mall\_Customers.csv includes data about customers and consists of four fields: customer number, gender, age, annual income, and spending score. This is basically aimed at grouping customers based on characteristics that are likely to be of value vis-a-vis marketing promotions and business intelligence.

# Overview of K-Means Clustering

K-Means is an iterative algorithm for clustering of objects in to kkk sets which is defined prior to commencing with this procedure. In KMeans each cluster is represented by a center or centroid and any data point is assimilated to the cluster whose center is closest to it using measures of distances like the Euclidean distance. They are alternatively adjusted until the least within-cluster variance or the inertia, cost function is realized .

## Key Steps in K-Means

1. **Initialize centroids:** Choose randomly kkk data points as initial starting points.
2. **Assign data points:** Classify each data point belonging to the closest centroid of a formed cluster.
3. **Update centroids:** Find new centroids calculating the mean value of all points belonged to each cluster.
4. **Repeat:** Subsequent to the above until there is convergence or the maximum number of iterations have been reached, assign data points to new/old centroid and update centroids accordingly.

K-Means is loved for its simplicity and great performance especially for mid-sized data sets. Yet, it has shortcomings, which include only being applicable to circularly distributed data, high sensitivity to the chosen value of k, and the absence of an automated method of determining the actual value of k.

# Dataset Description

The dataset used in this project contains five columns:

1. **CustomerID:** A customer number which is formed to be unique for each customer.
2. **Gender:** The male or female of the customer.
3. **Age:** The number of years of the age of the customer.
4. **Annual Income (k$):** Total number of thousands of dollars, which customer has received for a year.
5. **Spending Score (1-100):** A numerical value that gets given to customers by the firm evaluating the consumers expenditure spree and enactment.

For clustering, the consideration is given only to the ‘Annual Income’ and ‘Spending Score’ because these features are continuous and are informative of customers’ purchasing capacity and their expenditure tendencies.

# Implementation Steps

## Step 1: Loading and Exploring the Dataset

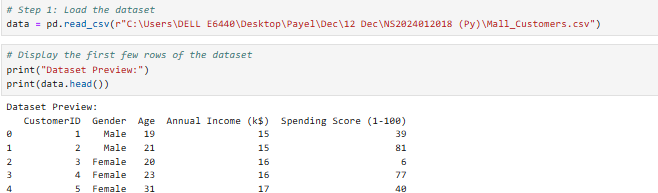
This dataset is read into a Pandas DataFrame for exploratory analysis and preparation of input for the models. Exploratory data analysis begins with?;

Well, it starts with some simple steps such as;

→Data cleaning: Checking for missing values

→Data profiling: Understanding the data type

→Data statistics: Summarizing the given data Histograms and scatter plots are among important known diagrams used for the analysis of distributive features and their mutual correlations if any.

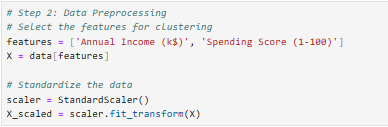


#### Fig 1: Data Upload

(Source: Self-Made)

## Step 2: Data Preprocessing

Data transformation is discussed to be of vital importance in preparation for the application of machine learning. While it comes to implementing cluster centers with the K-Means algorithm, the features have to be standardized with case in order to allow all the features contribute evenly in the clustering exercise. For features Annual Income and Spending Score, StandardScaler from Scikit-learn is applied to make the mean 0 and standard deviation 1.

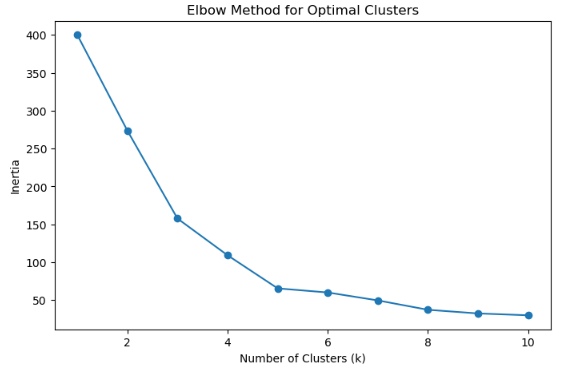


#### Fig 2: Data Preprocessing

(Source: Self-Made)

## Step 3: Determining the Optimal Number of Clusters

The calculated Ackley value of using the Elbow Method, determines the number of clusters k. This method involves applying K-Means with a various number of k‘s and representing thee inertia (sum of distances of the samples to the nearest cluster centre) against thee number of clusters. The zone beyond which the inertia decline rate sharply decelerates is called the “elbow point” to define the best k. For this dataset, as in all the other datasets, the elbow point is usually at k=5.

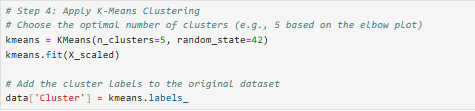


#### Fig 3: Elbow Method for Optional Clusters

(Source: Self-Made)

## Step 4: Applying the K-Means Algorithm

After identifying the right number of segments the K-Means algorithm is used to correctly cluster the customers into kkk clusters. This algorithm is followed by the labelling of each of the customer determined through the clustering process for other uses such as analysis and visualization.

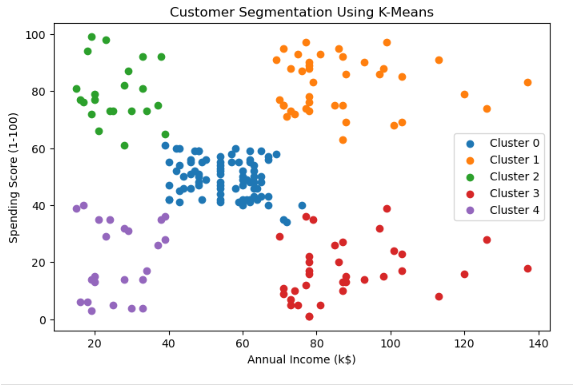


#### Fig 4: K-Means Clustering

(Source: Self-Made)

## Step 5: Visualizing the Clusters

In order to interpret the clustering results a scatter plot matrix is constructed wherein each data point is colored either by its assigned cluster number. The plot has a clear meaning of how consumer are segmented according to their income and other expenditures. Every group refers to a segment for the customer base, including high end spending high-income earners, low end spending low-income earners etc.



#### Fig 5: customer Segmentation Using K-Means

(Source: Self-Made)

## Step 6: Saving and Interpreting the Results

The cluster labels are entered in the same dataset as the new column, and the new dataset with the additions is saved in the form of a CSV file for further work. Several insights can be gained by examining each cluster such as the potential of several customers segments to make easy profits or segmenting and targeting appropriate marketing approaches.

# Insights from Clustering

The K-Means clustering results reveal distinct customer segments:

1. **Low-Income Low-Spenders:** This group of customers has low income and a low willingness to spend money. Some of the key marketing strategies for this group may then be product promotions within these affordable price offerings and any special offers.
2. **High-Income High-Spenders:** Such customers are very valuable and possess above-average levels of consumption. This segment should be targeted by luxury products and specific services.
3. **Moderate-Income Moderate-Spenders:** It is a mainstream group that guarantees constant sales and revenue in the course of business. Promotional techniques applicable for this segment may include bonus card and frequent offers.
4. **High-Income Low-Spenders:** This group of customers is characterized with high total purchasing potential but low Propensity to Treat. Consumers can be motivated to spend through appeals and promises to the consumers as well providing other options.
5. **Low-Income High-Spenders:** In fact, these customers are consequential despite the fact that they do not earn a lot of money. As to this behavior companies can use the installment payment with desired products or affordable luxuries.

It is by getting adequate information about these segments that business can be in a position invest adequately, enhance customer satisfaction and thus increase its profitability.

# Challenges and Limitations

While the K-Means algorithm provides valuable insights, it has some limitations:

1. **Predefined kkk:** The practical disadvantage of this approach lies in the fact that the number of clusters k is defined before the classification and may be problematic in cases of vague division of the material.
2. **Sensitivity to Outliers:** Outliers may dominate the computation of centroids thus distorting the value of the clustering.
3. **Assumption of Spherical Clusters:** K-Means also has the problem that assumes the clusters are spherical and of equal sizes, which is not possible in most of our real data sets.
4. **Scalability:** While useful in clustering moderate size databases, K-Means can be much less effective for very large databases owing to its computational costs.

In order to cope with these issues, there is the possibility to use another clustering algorithms, for example DBSCAN or Gaussian Mixture Models depending of the characteristics of the data and the aim of the project.

# Future Directions

This project can be extended in several ways:

1. **Feature Engineering:** If more features are used for clusters, for example, customer histories or geographical location data, then these clusters are going to be more accurate and useful.
2. **Dynamic Clustering:** Dynamic ways of computing kkk like silhouette score should be employed in order to enhance the reliability of the cluster results.
3. I**ntegration with Predictive Analytics:** When clustering is used in coordination with other predictive models, it becomes possible for businesses to make future assessments of customers and make better decisions.
4. **Visualization Enhancements:** Power BI/ Table you could be developing interactive dashboards to make the results more understandable to stakeholders.

# Conclusion

The tutorial on Python and the K-Means clustering technique for the clustering of customers prove that unsupervised machine learning plays a pivotal role in business analytics. This is because customers are often segmented by their income and their spending patterns and thus information acquired under this segment is useful for the business in its marketing activities.

Finally, by making clear preprocessing process, the appropriate choice of kkk, and presenting the results it is possible to propose the general framework for the analysis of clusters within the framework of this project. As much as one can argue that K-Means has its shortcomings it is the most basic yet efficient approach to undertake customer segmentation. Subsequent studies and live experiences with other methods and characteristics might increase the benefits and usefulness of clustering.

This project also provides a useful guideline for students or employees who plan to learn how to use machine learning to solve some business problems, stating the current key message of data-driven business decisions.

# Reference

Tabianan, K., Velu, S. and Ravi, V., 2022. K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. Sustainability, 14(12), p.7243.

Pradana, M.G. and Ha, H.T., 2021. Maximizing strategy improvement in mall customer segmentation using k-means clustering. Journal of Applied Data Sciences, 2(1), pp.19-25.

Abdulhafedh, A., 2021. Incorporating k-means, hierarchical clustering and pca in customer segmentation. Journal of City and Development, 3(1), pp.12-30.

Wu, J., Shi, L., Lin, W.P., Tsai, S.B., Li, Y., Yang, L. and Xu, G., 2020. [Retracted] An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K‐Means Algorithm. Mathematical Problems in Engineering, 2020(1), p.8884227.

Aryuni, M., Madyatmadja, E.D. and Miranda, E., 2018, September. Customer segmentation in XYZ bank using K-means and K-medoids clustering. In 2018 International conference on information management and technology (ICIMTech) (pp. 412-416). IEEE.

Shirole, R., Salokhe, L. and Jadhav, S., 2021. Customer segmentation using rfm model and k-means clustering. Int. J. Sci. Res. Sci. Technol, 8, pp.591-597.

Zhao, H.H., Luo, X.C., Ma, R. and Lu, X., 2021. An extended regularized K-means clustering approach for high-dimensional customer segmentation with correlated variables. Ieee Access, 9, pp.48405-48412.

Syakur, M.A., Khotimah, B.K., Rochman, E.M.S. and Satoto, B.D., 2018, April. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In IOP conference series: materials science and engineering (Vol. 336, p. 012017). IOP Publishing.

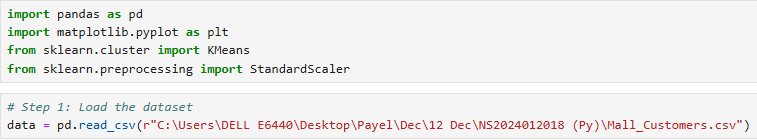
Brahmana, R.S., Mohammed, F.A. and Chairuang, K., 2020. Customer segmentation based on RFM model using K-means, K-medoids, and DBSCAN methods. Lontar Komput. J. Ilm. Teknol. Inf, 11(1), p.32.

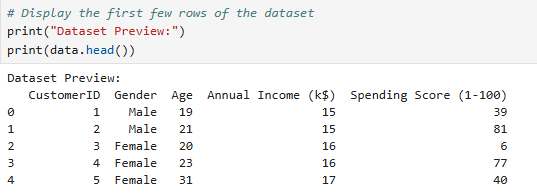
Kalusivalingam, A.K., Sharma, A., Patel, N. and Singh, V., 2020. Enhancing Customer Segmentation through AI: Leveraging K-Means Clustering and Neural Network Classifiers. International Journal of AI and ML, 1(3).

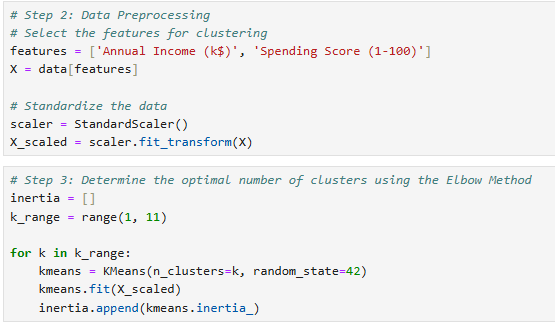
Mihova, V. and Pavlov, V., 2018, October. A customer segmentation approach in commercial banks. In AIP conference proceedings (Vol. 2025, No. 1). AIP Publishing.

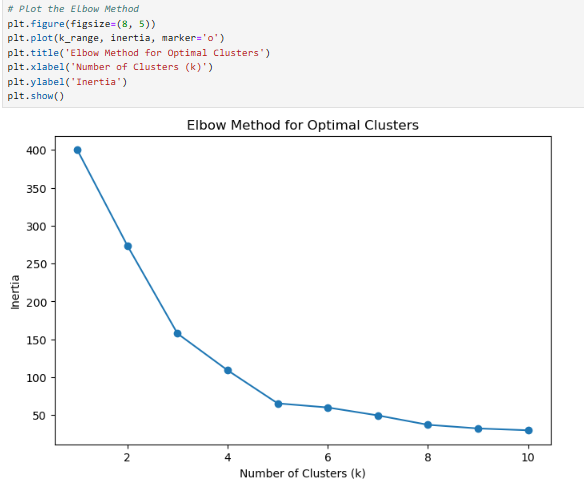
Anitha, P. and Patil, M.M., 2022. RFM model for customer purchase behavior using K-Means algorithm. Journal of King Saud University-Computer and Information Sciences, 34(5), pp.1785-1792.

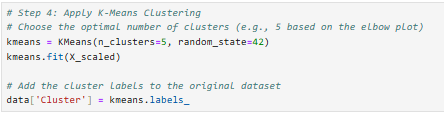
# Appendix

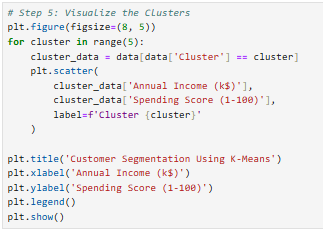


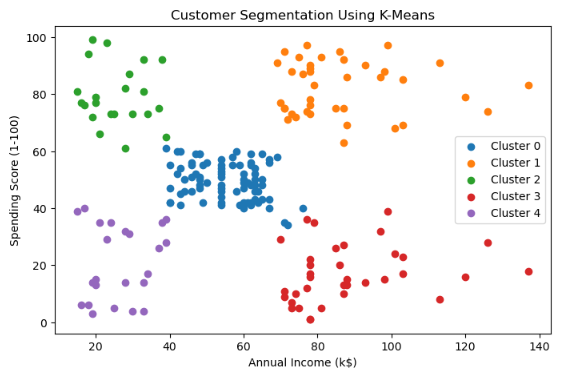












GitHub