CUSTOMER SEGMENTATION USING K-MEANS ALGORITHM

INTRODUCTION

Customer segmentation is the subdivision of a business customer base into groups called customer segments such that each customer segment consists of customers who share similar market characteristics. This segmentation is based on factors that can directly or indirectly influence market or business such as products preferences or expectations, locations, behaviours and so on. The importance of customer segmentation include, inter alia, the ability of a business to customise market programs that will be suitable for each of its customer segments; business decision support in terms of risky situation such as credit relationship with its customers; identification of products associated with each segments and how to manage the forces of demand and supply; unravelling some latent dependencies and associations amongst customers, amongst products, or between customers and products which the business may not be aware of; ability to predict customer defection, and which customers are most likely to defect; and raising further market research questions as well as providing directions to finding the solutions (Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance kalu ,2015).

Customer segmentation is one of the application of data mining which helps to segment the customers with similar patterns into similar clusters hence, making easier for the business to handle the large customer base. This segmentation can directly or indirectly influence the marketing strategy as it opens many new paths to discover like for which segment the product will be good, customising the marketing plans according to the each segment, providing discounts for a specific segment, and decipher the customer and object relationship which has been previously unknown to the company. Customer segmentation allows companies to visualise what actually the customers are buying which will prompt the companies to better serve their customers resulting in customer satisfaction, it also allows the companies to find who their target customers are and improvise their marketing tactics to generate more revenues from them (Tushar Kansal, Suraj Bahuguna ,Vishal Singh, Tanupriya Choudhury,2018).

Clustering has proven efficient in discovering subtle but tactical patterns or relationships buried within a repository of unlabelled datasets. This form of learning is classified under unsupervised learning. Clustering algorithms include k-Means algorithm, k-Nearest Neighbour algorithm, Self-Organising Map (SOM) and so on. These algorithms, without any knowledge of the dataset beforehand, are capable of identifying clusters therein by repeated comparisons of the input patterns until the stable clusters in the training examples are achieved based on the clustering criterion or criteria. Each cluster contains data points that have very close similarities but differ considerably from data points of other clusters(Chinedu Pascal Ezenkwu, Simeon Ozuomba , Constance kalu,2015).

Clustering process is not one time task but is continuous and an iterative process of knowledge discovery from huge quantities of raw and unorganized data (T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu,2001). For a particular classification problem, an appropriate clustering algorithm and parameters must be selected for obtaining optimum results(MacKay and David, 2003).

The primary objective of this work is to leverage the k-means algorithm to perform customer segmentation. By applying this powerful clustering technique, we aim to divide the customer base into distinct groups based on shared characteristics such as purchasing behaviour, demographics, or other relevant attributes. Through this segmentation, we aim to gain valuable insights into customer preferences and behaviour patterns. This enables to personalize marketing efforts, enhance customer satisfaction, optimize resource allocation, and ultimately drive business growth. By using the k-means algorithm for customer segmentation, we strive to make informed decisions and improve our overall understanding of the customer base.

BACKGROUND

Several studies have explored the use of machine learning algorithms, including K-means clustering, for customer segmentation. Machine learning algorithms can identify patterns and relationships in customer data that traditional statistical methods may not detect, making them a powerful tool for customer segmentation. Customer segmentation is a crucial aspect of marketing and business strategy, as it helps companies understand their customers' behaviour, preferences, and needs. Machine learning techniques, particularly K-means clustering, have emerged as popular approaches to segmenting customers based on their characteristics and behaviour. Jain, Murty, and Flynn's (1999) paper[16], "Data Clustering: A Review," provides a comprehensive survey of clustering algorithms, which are used for grouping data points into clusters or subgroups based on their similarity. The paper reviews over 400 research articles and categorizes clustering algorithms based on their underlying assumptions, types of data, and clustering criteria. The authors begin by defining clustering and explaining its importance in various domains, including pattern recognition, image analysis, data mining, and information retrieval. They then provide a taxonomy of clustering algorithms based on their characteristics and features. The taxonomy includes partitioning algorithms, hierarchical algorithms, density-based algorithms, grid-based algorithms, model-based algorithms, and others. The authors explain each algorithm's working principle, strengths, and weaknesses. The paper also highlights some common clustering criteria used to evaluate the quality of the clustering solutions, including the sum of squared errors, the silhouette coefficient, the purity measure, and the entropy measure. The authors also provide some guidelines for choosing the appropriate clustering algorithm and parameters for a given dataset, such as the size of the dataset, the dimensionality of the data, and the desired number of clusters. The authors then present a detailed discussion of some specific clustering algorithms, such as the k-means algorithm, the singlelink and complete-link hierarchical clustering algorithms, and the density-based spatial clustering of applications with noise (DBSCAN) algorithm. They compare these algorithms in terms of their computational complexity, memory usage, and performance on different types of data. Overall, the paper provides a comprehensive review of the clustering algorithms and criteria and serves as an essential resource for researchers and practitioners in the field of clustering and data mining. The authors' clear and concise presentation of the material and their insightful analysis make this paper an excellent starting point for anyone interested in clustering algorithms. 16 Clustering in Data Mining: A Survey," by Xu and Wunsch (2005)[13] is a comprehensive review paper that provides an overview of various clustering algorithms in data mining. The authors discuss the different types of clustering

algorithms, their advantages and disadvantages, and their applications in realworld scenarios. The paper begins by defining the concept of clustering and its significance in data mining. The authors explain that clustering is the process of grouping similar objects together based on their attributes, with the goal of finding meaningful patterns and relationships within the data. The authors then discuss the different types of clustering algorithms, including hierarchical, partitioning, density-based, and grid-based methods. They provide a detailed description of each algorithm, including its underlying principles and how it operates. They also highlight the strengths and weaknesses of each algorithm, along with their applications in different fields such as biology, medicine, and finance. In addition, the paper discusses the issue of choosing the appropriate distance metric or similarity measure for clustering, as well as the problem of determining the optimal number of clusters in a given dataset. The authors provide a thorough explanation of various validation techniques such as silhouette score, Calinski-Harabasz index, and Davies-Bouldin index that can be used to assess the quality of clustering results. The authors also examine recent advancements in clustering algorithms, such as the use of ensemble clustering, subspace clustering, and clustering with constraints. They highlight the potential benefits and challenges of these emerging techniques and discuss their future implications in data mining. Overall, "Clustering in Data Mining: A Survey" provides an excellent overview of various clustering algorithms and their applications. It is a valuable resource for researchers and practitioners in the field of data mining who want to gain a deeper understanding of clustering techniques and their strengths and limitations. MacQueen's 1967 [14]paper titled "Some Methods for Classification and Analysis of Multivariate Observations" is considered a seminal work in the field of clustering algorithms. In the paper, MacQueen proposed the k-means clustering algorithm, which has since become one of the most widely used clustering methods. The k-means algorithm aims to partition a set of n data points into k clusters, such that each data point belongs to the cluster with the nearest mean. The algorithm starts by randomly selecting k points as initial cluster centres, and then iteratively assigns each data point to the cluster with the nearest centre and updates the centres to the mean of the points in the cluster. The algorithm terminates when the cluster assignments no longer change. MacQueen's paper also introduced the concept of within-cluster variance as a measure of the goodness of the clustering. He suggested using the ratio of the between-cluster variance to the within-cluster variance, known as the F-statistic, to determine the 17 optimal number of clusters. The k-means algorithm has since become a widely used method for clustering data in various applications, such as image segmentation, document clustering, and customer segmentation. One of the strengths of k-means is its simplicity and efficiency, making it suitable for large datasets. Despite its popularity, k-means does have some limitations. One major drawback is its sensitivity to the initial cluster centres, which can result in suboptimal clustering solutions. Various modifications and extensions of the kmeans algorithm have been proposed to address these limitations, such as hierarchical clustering and fuzzy clustering. MacQueen's 1967 paper proposed a simple yet powerful algorithm for clustering data that has since become a fundamental tool in the field of data mining and machine learning. The k-means algorithm continues to be widely used and has inspired numerous extensions and modifications, making it a cornerstone of clustering methods. "Introduction to Data Mining" by Tan, Steinbach, and Kumar[15] is a widely used textbook that provides an in-depth overview of data mining techniques, including clustering methods such as k-means. The authors introduce the concept of clustering as a process of grouping data objects into clusters such that objects in the same cluster are more similar to each other than to those in other clusters. They explain that clustering

is an unsupervised learning technique, meaning that the algorithm attempts to find patterns in the data without being given prior knowledge or labelled examples. The book discusses the k-means algorithm as a popular and simple clustering method. The authors explain that the algorithm partitions a set of data objects into k clusters by iteratively optimizing the sum of squared distances between each data point and its assigned cluster centre. They provide step-by-step instructions for implementing the k-means algorithm and emphasize the importance of selecting the appropriate number of clusters (k) for a given dataset. In addition to describing the k-means algorithm, the book also covers variations of the algorithm, such as k-medoids and fuzzy k-means, as well as techniques for evaluating clustering results. The authors provide examples and case studies throughout the book to illustrate the application of clustering in various fields. Overall, "Introduction to Data Mining" is a comprehensive and accessible resource for learning about clustering and other data mining techniques, including the popular k-means algorithm.

The paper titled "Integration of Machine Learning Techniques to Evaluate Dynamic Customer Segmentation Analysis for Mobile Customers" focuses on customer segmentation in the telecommunications industry. The goal is to develop a business intelligence model that helps mobile providers optimize customer retention and marketing strategies. The paper highlights the use of machine learning algorithms, specifically the C.5 algorithm within naive Bayesian modelling, to segment telecommunication customers based on their billing and socio-demographic aspects. The authors, Cormac and Eleni collected and analysed a large dataset of customer account data to gain insights and improve customer satisfaction, loyalty, and retention. The main contribution of the paper is the improvement in churn prediction by applying machine learning algorithms for customer segmentation. Decision tree rules models were used to identify segmentation rules and predict customers' buying patterns. This allows telecommunication companies to target deals and programs effectively and reduce churn rates. Furthermore, the paper discusses the importance of customer relationship management (CRM) and the application of data mining techniques in telecom churn prediction. The RFM (Recency, Frequency, Monetary) analysis is mentioned as a useful method for behavioural-based customer data mining. The paper also highlights the significance of derived attributes, such as customer demographic profiles, including age groups and location counties, in customer behaviour pattern analysis. Overall, the paper emphasizes the importance of data mining and machine learning techniques in customer segmentation and churn management for mobile providers in the telecommunications industry.

In the research paper titled "Customer Segmentation Using Clustering and Data Mining Techniques," the authors K.R.Kashwan and C.M.Velu proposed a real-time and online system for a specific supermarket to predict sales in various annual seasonal cycles. The system utilizes the k-means clustering technique and the SPSS tool to analyse sales data records and update segmentation statistics on a daily basis. The authors applied the k-means clustering technique to segment customers of a leading supermarket retail household supplier. They collected data from customer transactions over a three-month period and utilized the SPSS tool for analysis. The results showed promising accuracy, with the predicted sales statistics closely matching the actual sales statistics. The convergence of the k-means algorithm can be slow, especially in cases requiring high accuracy. However, by setting a reasonable threshold value, quick results can be obtained without compromising much accuracy. Additionally, the algorithm allows for the modification of the sum of squares of errors (SSE) by splitting or merging clusters, which

is desirable in market segmentation studies. Effective market segmentation leads to better decision-making by accurately predicting customer needs. Clustering analysis can be applied to classify objects such as brands, products, and consumer perceptions. By clustering customers based on various dimensions related to a product, organizations can gain valuable insights into customer behaviour and tailor their strategies accordingly. Overall, the use of the k-means clustering technique in market segmentation research holds significant potential for improving decision-making processes in various industries. Its ability to accurately group customers based on their characteristics and preferences allows organizations to target specific segments effectively and enhance profitability.

The research paper ”Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services”, focuses on customer segmentation in a retail business using the k-Means clustering algorithm. The k-Means clustering algorithm is employed to identify customer segments based on two features: the average amount of goods purchased per month and the average number of customer visits per month. A MATLAB program is developed and trained using a dataset of 100 training patterns from the retail business. The dataset is z-score normalized, and after several iterations, four customer clusters or segments are identified: High-Buyers-Regular-Visitors (HBRV), High-Buyers-Irregular-Visitors (HBIV), Low-Buyers-Regular-Visitors (LBRV), and Low-Buyers-Irregular-Visitors (LBIV).The paper discusses the concept of customer segmentation and its significance in business decision-making, customization of marketing programs, and predicting customer defection. It also explores the role of big data in providing valuable insights and the challenges associated with analysing large volumes of structured and unstructured data. Clustering algorithms, particularly the k-Means algorithm, are highlighted as effective tools for discovering patterns and relationships in unlabelled datasets. The results of the k-Means clustering algorithm show that the clusters converge after 100 iterations, and the algorithm successfully segments the customer data. In conclusion, the paper demonstrates the application of the k-Means clustering algorithm in customer segmentation for a retail business. It emphasizes the benefits of automated approaches using big data and machine learning over traditional market analyses. The findings provide valuable insights into customer behaviour and can guide businesses in developing targeted marketing strategies and improving customer services.

The paper “Customer Segmentation using K-means Clustering “ discusses the application of customer segmentation using clustering algorithms, specifically k-means, agglomerative, and mean shift, for better decision making in marketing strategies. The authors aim to identify customer segments based on their shopping behaviour, using a dataset from a local retail shop with two features: average number of visits and average amount of shopping per year. The k-means clustering algorithm is employed to group customers into clusters based on their similarities. The optimal number of clusters (k) is determined using the elbow method, which analyses the within-cluster sum of squares (WCSS). By plotting the WCSS for different values of k, the authors select the value where the decline in WCSS is most significant, indicating the elbow point and the optimal number of clusters. Furthermore, the silhouette score is utilized to evaluate the quality of the clustering results. It measures how close each sample in one cluster is to samples in neighbouring clusters. A higher silhouette score indicates better-defined and well-separated clusters. The results of the k-means clustering show the formation of five clusters labelled as Careless, Careful, Standard, Target, and Sensible customers. Additionally,

the agglomerative clustering algorithm produces similar cluster patterns. However, the mean shift clustering algorithm reveals two new clusters: High buyers and frequent visitors, and High buyers and occasional visitors. The authors emphasize the importance of customer segmentation for businesses to understand customer behaviour, target specific segments, and customize marketing strategies accordingly. Clustering algorithms, particularly k-means, offer an effective approach to achieve customer segmentation by grouping customers with similar patterns together. The evaluation of clustering results using metrics like the silhouette score provides insights into the quality of the segmentation.

PROPOSED METHODOLOGY

The proposed methodology has been divided into few distinct yet interrelated steps. These are described below :

Data Preparation

Data Pre-processing

Data Pre-processing can be defined as a process of converting raw data into a format that is understandable and usable for further analysis. It is an important step in the Data Preparation stage. It ensures that the outcome of the analysis is accurate, complete, and consistent. It is the first and crucial step while creating a machine learning model.

Prior to clustering, the dataset undergoes a crucial phase of pre-processing and analysis to ensure its suitability for the task. This process involves handling missing values and duplicates, as well as examining the presence of null values and inconsistent datatypes within the data. By addressing these issues, the dataset is refined and prepared for subsequent clustering operations.

To facilitate the clustering and splicing procedures, the column names of the resulting dataframe are explicitly specified. This deliberate specification of column names simplifies the clustering process and enhances the subsequent ability to extract relevant subsets of data. By following these practices, the dataset is optimized for effective clustering analysis, enabling to gain a comprehensive understanding of the provided data.

Column Selection

Feature selection is the process of identifying the most effective subset of the original features to use in clustering. Feature extraction is the use of one or more transformations of the input features to produce new salient features. Either or both of these techniques can be used to obtain an appropriate set of features to use in clustering(Jain, A., Murty, M.N., and Flynn, P.J. 1999).

In the methodology proposed in this research work , feature selection technique is applied to select a subset of columns from a dataframe while preparing the data for machine learning or data analysis tasks, where only certain variables are relevant to the task at hand.

The choice of columns for segmentation varies depending on the specific dataset under analysis. In each case, relevant variables are carefully selected, considering the unique characteristics and objectives of the dataset. For instance, when examining consumer behaviour, columns such as spending score and annual income may be deemed particularly pertinent. These columns are extracted, separated, and organized into a two-dimensional or n-dimensional array, allowing for seamless utilization in subsequent data processing steps. This refined arrangement facilitates further exploration and analysis, enabling researchers to derive valuable insights from the data.

Determining the Optimum Number of clusters

One crucial aspect of implementing the K-means algorithm is determining the optimal number of clusters. Choosing the correct number of clusters is essential to ensure meaningful and accurate cluster assignments. By employing techniques such as the Elbow Method, Silhouette Analysis, Gap Statistic, information criteria, and considering domain knowledge, practitioners can make informed decisions to identify the optimal number of clusters that best capture the underlying patterns in the data without overfitting or underfitting. The goal is to find the optimal number of clusters that maximizes the intra-cluster similarity while minimizing the inter-cluster dissimilarity.

In the proposed methodology of this research work elbow method and silhouette analysis has been used to determine the optimal k value.

Elbow Method

Elbow method is used for finding optimal value of K for K-means clustering algorithm (Tushar Kansal, Suraj Bahuguna ,Vishal Singh, Tanupriya Choudhury,2018).The elbow graph is a visual tool used to determine the optimal number of clusters to use in a K-means clustering algorithm. In the context of customer segmentation, the elbow graph can help identify the optimal number of customer segments or clusters based on the available customer data.

The elbow graph is plotted with the number of clusters on the x-axis and the Within-Cluster-Sum-of-Squares (WCSS) on the y-axis. The WCSS measures the sum of distances between the data points and their assigned cluster centroid. As the number of clusters increases, the WCSS decreases, since each data point has a closer centroid. However, beyond a certain number of clusters, the improvement in WCSS begins to decrease at a slower rate, resulting in an elbow-shaped curve.

The WCSS formula is defined as:

WCSS = sum(d^2)/n

where d is the distance between an observation and its cluster mean, *n* is the number of observations in the cluster, and *sum(d^2)* is the sum of the squared distances between each observation and its mean.

The "elbow point" on the curve represents the optimal number of clusters to use. In the context of customer segmentation, this optimal number of clusters is the one that produces a reasonable balance between the granularity of the segments and the size of each segment.

For example, suppose we have a dataset of customer purchases and we want to segment them into different groups based on their buying patterns. We can use the elbow method to determine the optimal number of clusters for our dataset. We first apply *k*-means clustering to our data for different values of *k* (number of clusters). We calculate the WCSS for each value of *k* and plot it on a graph. We can see that initially, the WCSS decreases rapidly as we increase the number of clusters. However, after a certain point, the rate of decrease slows down significantly, indicating the optimal number of clusters. Suppose we observe an elbow point at *k*=3. This means that the optimal number of clusters for our dataset is 3, and we can use this information to segment our customers into 3 groups based on their buying patterns. Overall, the elbow graph is an important tool for optimizing customer segmentation using the *k*-means clustering algorithm.

Silhouette Analysis

When determining the optimal number of clusters, if it is unclear where the elbow point lies in the evaluation process, the silhouette plot is employed for validation. The silhouette score is utilized as a quantitative measure to assess the clustering's overall quality. By calculating this score, researchers can gauge how well the data points within each cluster are separated and how cohesive they are within their respective clusters. This approach provides a more refined and objective assessment of the clustering performance, aiding in the selection of the most appropriate value for K, the number of clusters.

Silhouette analysis is a technique used to determine the quality of a clustering algorithm's output. It evaluates the quality of the clustering by measuring the distance between each data point and other data points within the same cluster compared to the distance between that data point and data points in other clusters. The silhouette score ranges from -1 to 1, where a score closer to 1 indicates that the object is well-matched to its own cluster and poorly matched to neighbouring clusters. On the other hand, a score closer to -1 indicates that the object may be better matched to a neighbouring cluster. A score of 0 means that the object is on the boundary between two clusters.

The silhouette score is calculated using two metrics:

- The mean distance between a sample and all other points in the same class.
- The mean distance between a sample and all other points in the next nearest cluster.

The silhouette score is calculated using the following formula:

$S = (b - a) / max(a,b)$

where ‘a’ is the average distance between a data point and other data points in the same cluster(i.e., average intra-cluster distance or cohesion), and ‘b’ is the average distance between a data point and data points in the nearest other cluster(i.e., average inter-cluster distance separation). The ‘S’ value ranges between -1 and 1, where a value of 1 indicates that the data point is very similar to other data points in its cluster, and very dissimilar to data points in other clusters. A value of -1 indicates the opposite, with values around 0 indicating that the data point is equally similar to points in its own cluster and other clusters.

The silhouette score can be plotted for each value of *k* (number of clusters), and the cluster with the highest average silhouette score is considered to be the optimal number of clusters for the data set. The Silhouette graph plots the silhouette score for each cluster against the number of clusters. The graph is used to find the optimal number of clusters, which is the number of clusters with the highest average silhouette score.

In practice, a silhouette score of 0.5 or greater is considered a strong separation between clusters, while a score between 0.25 and 0.5 indicates a reasonable separation, and a score less than 0.25 indicates that the clustering may not be meaningful. The thickness of the silhouette plot representing each cluster is also a deciding point . Plots with moderately similar thickness are better choice as the optimal cluster value .

In the context of customer segmentation, the silhouette analysis can be used to determine the optimal number of customer segments that provide meaningful insights into customer behaviour. By analysing the silhouette scores for different values of *k*, businesses can gain insights into the number of distinct customer segments in their data set, and the quality of the clustering for each segment.

K-Means Clustering

In the context of customer segmentation, k-means can be used to divide a customer base into smaller, more homogeneous groups based on their demographic, psychographic, and behavioural characteristics. To do this, the algorithm requires a set of numerical variables that describe the customers, such as age, income, education, occupation, and buying behaviour.

The algorithm is called *k*-means due to the fact that the letter *k* represents the number of clusters chosen(Kishana R. Kashwan and C.M. Velu,2013). An observation is assigned to a particular cluster for which its distance to the cluster mean is the smallest. The principal function of algorithm involves finding the *k*-means. First, an initial set of means is defined and then subsequent classification is based on their distances to the centres (S. Dasgupta and Y. Freund,2009). Next, the clusters’ mean is computed again and then reclassification is done based on the new set of means. This is repeated until cluster means don’t change much between successive iterations (M. Mahajan, P. Nimbhorkar, and K. Varadarajan,2009). Finally, the means of the clusters once again calculated and then all the cases are assigned to the permanent clusters.

Given a set of observations *(x1, x2, ….., xn)*, where each observation *xi* is a *d*-dimensional real vector. The *k*-means clustering algorithm aims to partition the n observations into *k* groups of observations called clusters where $k \leq n$, so as to minimize the sum of squares of distances between observations within a particular cluster (A. Vattani,2011).

As shown in Table I, the sum of squares of the distance may be given by the equation *arg min* $S = \Sigma i=1 \Sigma sj\ si\ ||xj - \mu j\ ||2$, where μi is the mean of points in Si . Given an initial set, k-means computes initial means *m1* (1) ,…,*mk* (1) and it identifies *k* clusters in given raw data set.

Stated informally, the *k*-means procedure consists of simply starting with *k* groups each of which consists of a single random point, and thereafter adding each new point to the group whose mean the new point is nearest. After a point is added to a group, the mean of that group is adjusted in order to take account of the new point. Thus at each stage the k-means are, in fact, the means of the groups they represent (hence the term *k*-means)(MacQueen, J. B.,1967).

The *k*-means algorithm is very simple and can be easily implemented in solving many practical problems. It can work very well for compact and hyperspherical clusters. The time complexity of *k*-means is *O(NKd)*(Xu, R., and Wunsch, D. ,2005).

Detailed working of K-Means Algorithm used to determine the optimum no of clusters highlighted in this paper:

i. The number of clusters, *k* to be created is chosen.
ii. Initialize the centroids (or means) for each cluster by randomly selecting K observations from the dataset.
iii. Each observation is assigned to the cluster with the nearest mean. The distance between an observation and a mean can be calculated using various distance metrics, but the most commonly used is Euclidean distance. It calculates the straight-line distance between two data points in n-dimensional space. This measure is suitable for continuous data with a Euclidean geometry.
$deuc(x, y) = \sqrt{\sum i=1n(xi-yi)2}$
iv. The within-cluster sum of squares (WCSS)is calculated for each cluster. WCSS is a measure of how spread out the observations are within each cluster, and it is calculated as the sum of squared distances between each observation and its cluster mean.
v. The silhouette coefficient is calculated
vi. The WCSS values are used to create an elbow plot, which shows the relationship between the number of clusters and the WCSS. The elbow plot is used to determine the optimal number of clusters, where the WCSS starts to decrease at a slower rate. This point is called the "elbow" of the plot.
vii. The silhouette coefficients are used to create a silhouette score graph to understand the quality of the cluster.
viii. Steps 3 to 5 are repeated until the maximum number of iterations is reached.

This process is iterated for k values ranging from 2 to 11.

To ensure the selection of the optimal value for k, a silhouette analysis was conducted for each cluster value ranging from 2 to 6. This involved creating a silhouette plot and visualizing the clusters. The silhouette plot shows the silhouette coefficient for each sample, which measures how similar a sample is to its own cluster compared to other clusters. The silhouette coefficient ranges from -1 to 1, with 1 indicating that the sample is well-matched to its own cluster and poorly matched to neighbouring clusters. The plot also shows a vertical line indicating the average silhouette coefficient for all samples, and horizontal bars for each cluster indicating

the range of silhouette coefficients for that cluster. The cluster visualization plot shows the clusters as determined by the k-means algorithm, with each cluster represented by a different colour. The plot also shows the cluster centres as white circles with black edges, and each point is labelled with its cluster number.

Training K-means Algorithm

Training the k-means algorithm with a fixed value of k, determined through silhouette and elbow graph analysis, is a crucial step in unsupervised machine learning. The training process involves optimizing the algorithm to achieve accurate clustering results, focusing on the training aspects.

The first step in training the k-means algorithm with a fixed value of k involves determining the optimal number of clusters. This is typically done using techniques such as silhouette analysis and the elbow method. Silhouette analysis calculates a score for each data point based on its cohesion within the assigned cluster and separation from other clusters. The elbow method examines the sum of squared distances within each cluster as the number of clusters increases and identifies the "elbow" point where the rate of improvement diminishes significantly.

Once the optimal value of k is determined, the training process begins. Randomly selected data points are assigned as the initial centroids for each cluster. The algorithm then iteratively performs the assignment and centroid update steps until convergence is achieved. In the assignment step, each data point is assigned to the nearest centroid based on their proximity in feature space.

After the initial assignment, the centroids are updated by calculating the mean of the data points within each cluster. This iterative process continues until convergence, typically defined by a threshold or a stopping criterion, is reached. The final centroids represent the cluster centres, and the data points are grouped into distinct clusters based on their proximity to these centres.

Training the k-means algorithm with a fixed value of k aims to find the optimal configuration of clusters that minimizes the sum of squared distances between data points and their assigned centroids. The performance of the algorithm heavily relies on the initial centroid selection, the choice of distance metric, and the determination of the optimal number of clusters.

By utilizing silhouette analysis and the elbow method to determine the appropriate value of k, the training process focuses on optimizing the algorithm specifically for that fixed number of clusters. This approach helps to ensure that the clustering results accurately represent the underlying structure of the data.

In conclusion, training the k-means algorithm with a fixed value of k determined through silhouette and elbow graph analysis involves iteratively assigning data points to clusters and updating centroids until convergence is achieved. The process is tailored to optimize the algorithm for a specific number of clusters, resulting in accurate and meaningful clustering results.

The objective function of k-means is to minimize the sum of squared distances between each data point and its assigned centroid:

$J = \sum i=1 \text{ to } n \ ||xi - \mu ci||2$

where xi is the ith data point, μci is the centroid of the cluster to which xi is assigned, and n is the total number of data points.

Numerical Example:

Suppose we have a dataset of 6 two-dimensional points:

(2, 10), (2, 5), (8, 4), (5, 8), (7, 5), (6, 4)

We want to partition this dataset into 3 clusters using the k-means algorithm. We can follow the steps below:

Step 1: Initialize centroids randomly.

Let us randomly initialize the centroids as:

c1 = (2, 10), c2 = (8, 4), c3 = (5, 8)

Step 2: Assign each data point to the closest centroid.

We calculate the Euclidean distance between each data point and each centroid, and assign each data point to the closest centroid:

Cluster 1: (2, 10), (2, 5)

Cluster 2: (8, 4), (7, 5), (6, 4)

Cluster 3: (5, 8)

Step 3: Recalculate the centroid for each cluster.

We calculate the mean of all data points assigned to each cluster, and update the centroid: c1 = (2, 7.5), c2 = (7, 4.33), c3 = (5, 8)

Step 4: Repeat steps 2 and 3 until convergence.

We repeat steps 2 and 3 until no data points are reassigned. After one iteration, the clusters become:

Cluster 1: (2, 10), (2, 5)

Cluster 2: (8, 4), (7, 5), (6, 4), (5, 8)

After another iteration, the clusters become:

Cluster 1: (2, 10), (2, 5), (5, 8)

Cluster 2: (8, 4), (7, 5), (6, 4)

After the third iteration, the clusters do not change anymore, and we obtain the final clusters:

Cluster 1: (2, 10), (2, 5), (5, 8)

Cluster 2: (8, 4), (7, 5), (6, 4)

Cluster 3: (5,8)

The final centroids are:

c1 = (3, 7.67), c2 = (7, 4.33), c3 = (5, 8)

In the methodology proposed in this paper the cluster number for each of the data points were stored as an array which can be used for visualisation and analysis of the given customer database.

Algorithm for training k-means algorithm:

i. Initialize the centroids:

- Randomly select initial centroids for each cluster.

Pseudocode:

```
centroids = initialize_centroids(num_clusters)
```

ii. Assign data points to clusters:

- Assign each data point to the nearest centroid.

Pseudocode:

```
clusters = assign_to_clusters(preprocessed_data, centroids)
```

iii. Recalculate centroids:

- Calculate new centroids for each cluster by taking the mean of all the data points assigned to it.

Pseudocode:

```
new_centroids = calculate_centroids(clusters)
```

iv. Repeat steps 5 and 6:

- Repeat the assignment and centroid recalculation steps until the centroids no longer move or a maximum number of iterations is reached.

Pseudocode:

```
while centroids are changing and iteration_count < max_iterations:
    clusters = assign_to_clusters(preprocessed_data, centroids)
    new_centroids = calculate_centroids(clusters)
    centroids = new_centroids
    iteration_count += 1
```

v. Evaluate the results:

- Analyze the characteristics of each cluster, such as average spending and purchase frequency, to evaluate the clustering results.

Pseudocode:

```
cluster_characteristics = analyze_clusters(clusters)
```

RESULT AND ANALYSIS

Dataset Used

The proposed algorithm was evaluated and analyzed using two datasets. These datasets consisted of comma-separated values (mall customer data) obtained from Kaggle. Specifically, the data corresponding to the 'Annual_Income', 'Spending_Score', 'Annual_Income', 'Age', and 'Customer_Pin' labels were extracted from the dataset and read into a pandas DataFrame. This facilitated convenient manipulation and processing of the data.

Methods and Functions Used

The methods and functions used for the research work are listed in Table 1.

Table 1.Methods and functions used

Job	Method/Function	Library/Package Used
Loading data from CSV file	pd.read_csv()	pandas
Data preprocessing	head()	pandas
	shape	pandas
	info()	pandas
	isnull().sum()	pandas
	iloc[]	pandas
	np.expm1()	numpy
	append()	numpy
	KMeans()	sklearn.cluster
	silhouette_score()	sklearn.metrics
	silhouette_samples()	sklearn.metrics

Clustering	fit_predict()	sklearn.cluster
	fit()	sklearn.cluster
	inertia_	sklearn.cluster
Data visualization	fill_betweenx()	matplotlib.pyplot
	text()	matplotlib.pyplot
	scatter()	matplotlib.pyplot
	axvline()	matplotlib.pyplot
	set_yticks()	matplotlib.pyplot
	set_xticks()	matplotlib.pyplot
	set_title()	matplotlib.pyplot
	set_xlabel()	matplotlib.pyplot
	set_ylabel()	matplotlib.pyplot
	suptitle()	matplotlib.pyplot
	bar()	matplotlib.pyplot
	show()	matplotlib.pyplot
	plot()	matplotlib.pyplot

DATASET 1

The provided dataset consists of 8 columns, namely Customer ID, Gender, Age, Annual Income, Spending Score, Warehouse Pin-code, Customer Pin-code, and Zone, encompassing data for 124 customers. During the coding process, the cluster value is added to the dataset.

	CustomerID	Gender	Age	Annual_Income	Spending_Score	Ware_Pin	Customer_Pin	Zone	clusters
0	1	Male	19	15	39	121003	507101	d	4
1	2	Male	21	15	81	121003	486886	d	4
2	3	Female	20	16	6	121003	532484	d	4
3	4	Female	23	16	77	121003	143001	b	0
4	5	Female	31	17	40	121003	515591	d	4

Table.2 : First five rows of the dataset 1

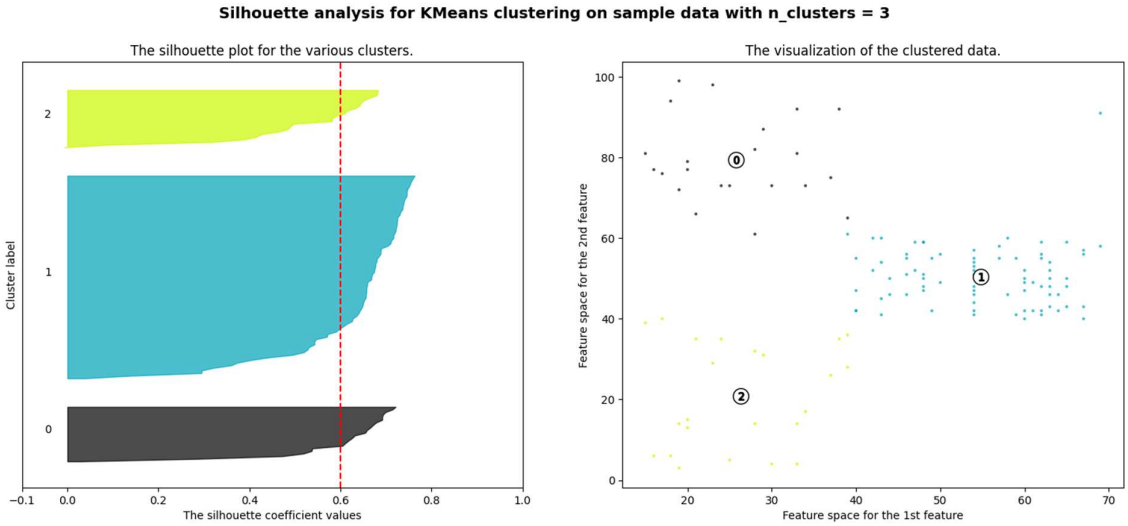
Analyzing the Results :

CASE 1 : In the first case we applied K-Means Clustering on the parameters –

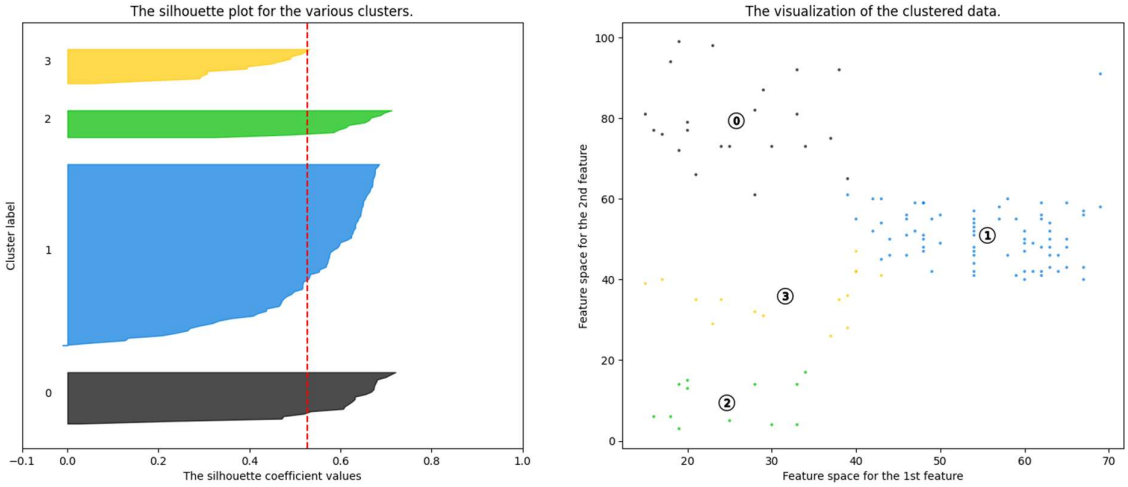
Annual Income vs Spending Score and the results looked like:

Figure 1 :proves how 3 is the optimal number of clusters.

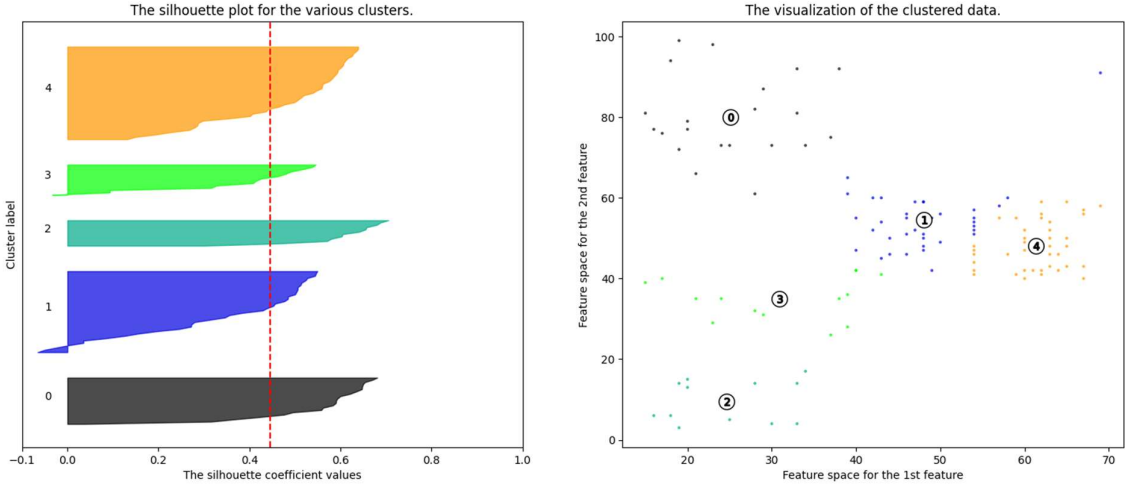
```
For n_clusters = 2 The average silhouette_score is : 0.7051708416017349
For n_clusters = 3 The average silhouette_score is : 0.6595557835052082
For n_clusters = 4 The average silhouette_score is : 0.6918601183884051
For n_clusters = 5 The average silhouette_score is : 0.6936287837961752
For n_clusters = 6 The average silhouette_score is : 0.6922766282170691
```



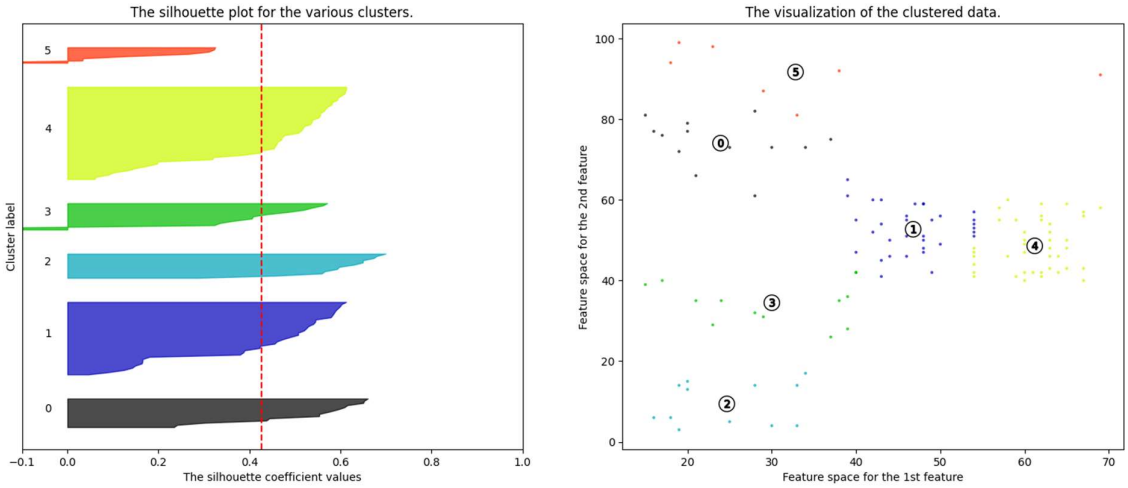
Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



Silhouette analysis for KMeans clustering on sample data with n_clusters = 5



Silhouette analysis for KMeans clustering on sample data with n_clusters = 6



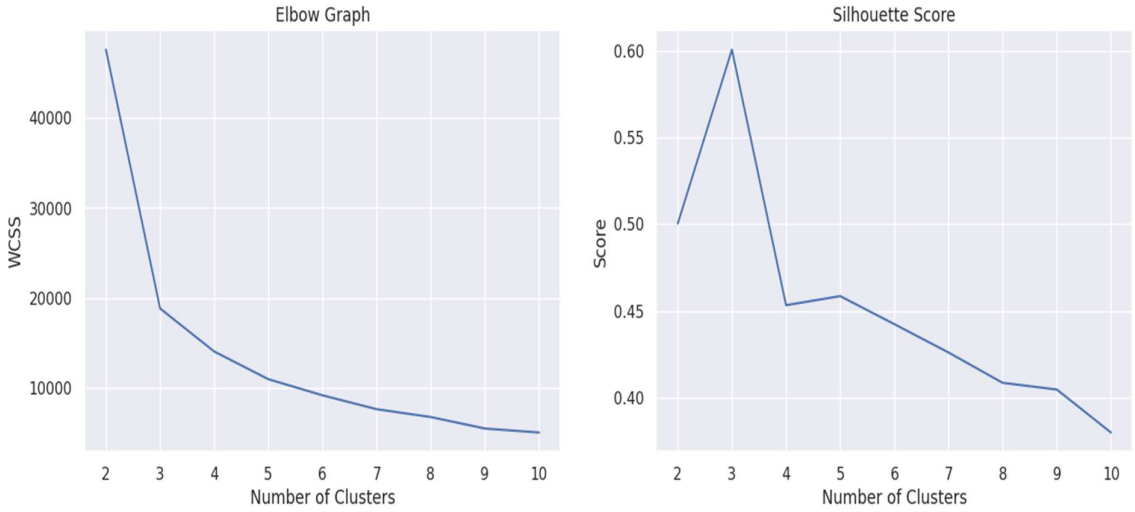


Figure 2: Elbow Method and Silhouette Score for optimal number of Clusters

From this elbow graph and silhouette score we can say that the optimum number of clusters are 3.

The Final clusters of the scatter plot look like –

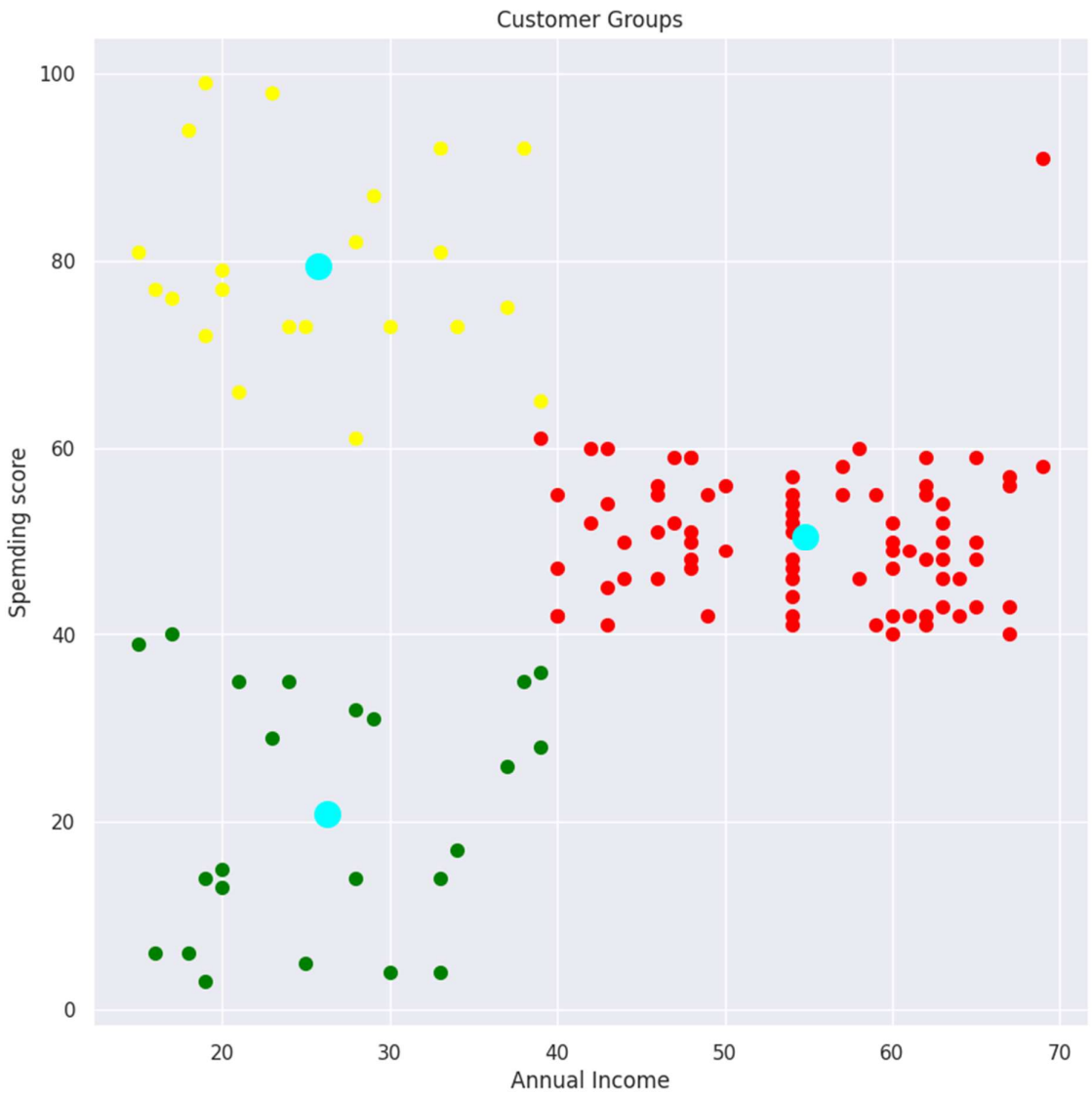


Figure 3: Final scatter plot

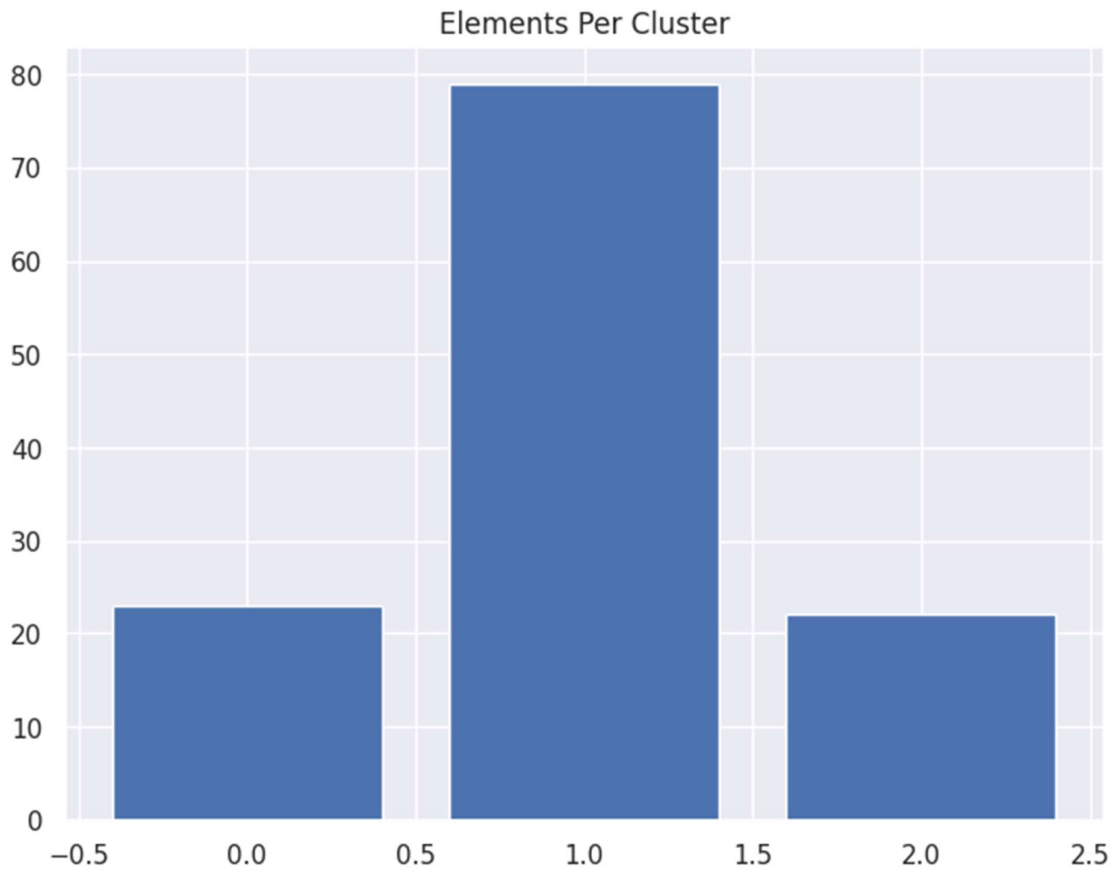


Figure 4: Elements per cluster

Based on the generated output, we can make the following observations:

1. Green Cluster (0.0): This cluster represents customers with an annual income of <=400k and a spending score of <=40. It appears as a scattered cluster, indicating that there are relatively fewer customers interested in the products offered by this particular business. To attract customers from this group, the business could consider implementing various strategies such as offering discounts or loyalty benefits.

2. Red Cluster (1.0): This cluster comprises customers with an annual income in the range of 400k to 700k and a spending score in the range of 40-60. It forms a concentrated cluster, suggesting that a larger number of customers from this group are interested in the products offered by the business. To capitalize on this customer segment, the business could introduce memberships or provide specific benefits to keep them engaged and invested.

3. Yellow Cluster (2.0): This cluster consists of customers with an annual income of <=400k and a spending score in the range of 60-100. Similar to the green cluster, it appears as a scattered cluster, indicating a lower level of interest among customers in this group for the business's products. To attract customers from this segment, the

business could consider implementing strategies such as improved advertising, offering flashy offers, and allocating loyalty points.

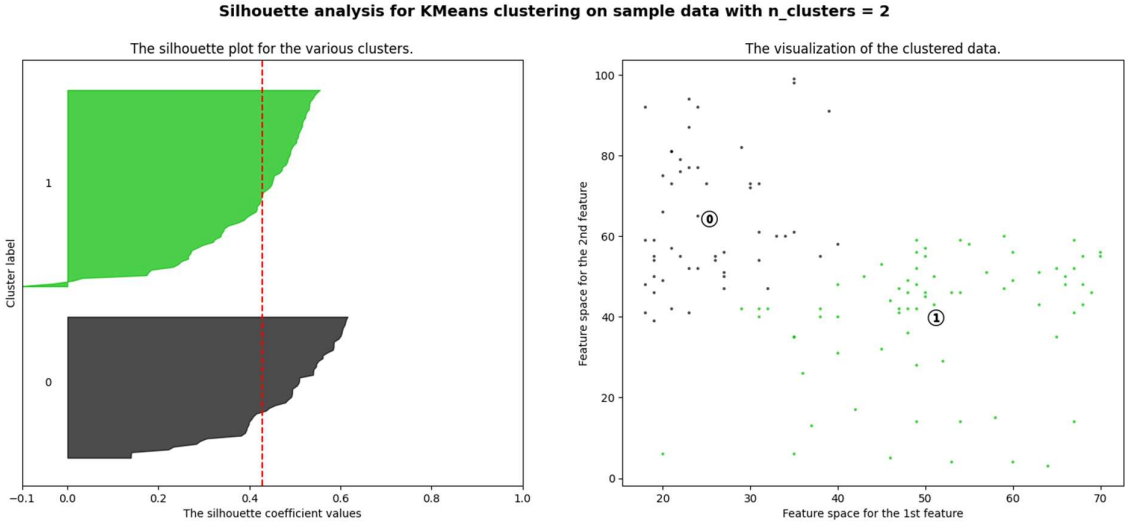
By carefully analyzing these clusters and implementing targeted strategies, the business can aim to enhance its customer base and optimize its offerings to cater to different customer preferences.

CASE 2 : In the second case we applied K-Means Clustering on the parameters –

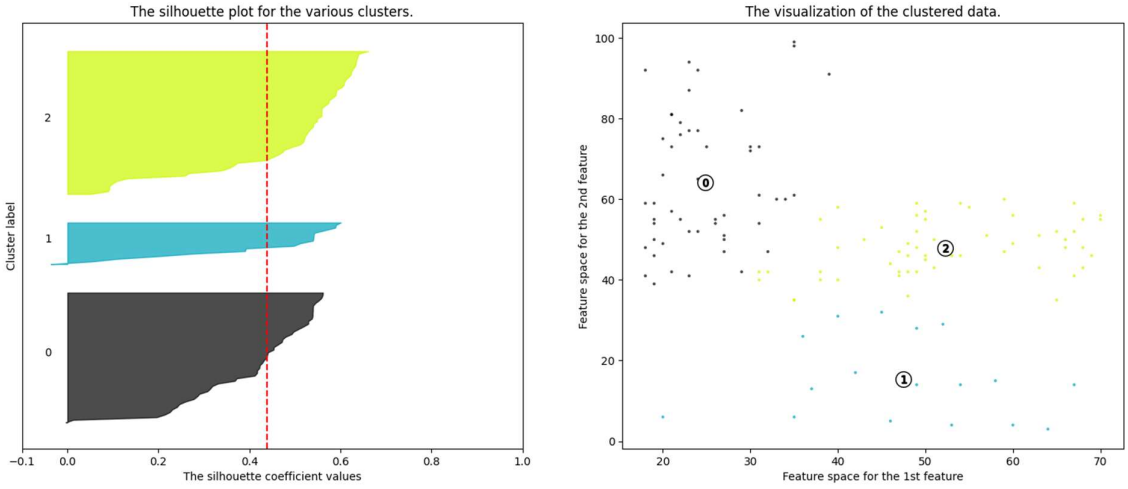
Spending Score vs Age and the results looked like :-

Figure 5 :proves how 4 is the optimal number of clusters.

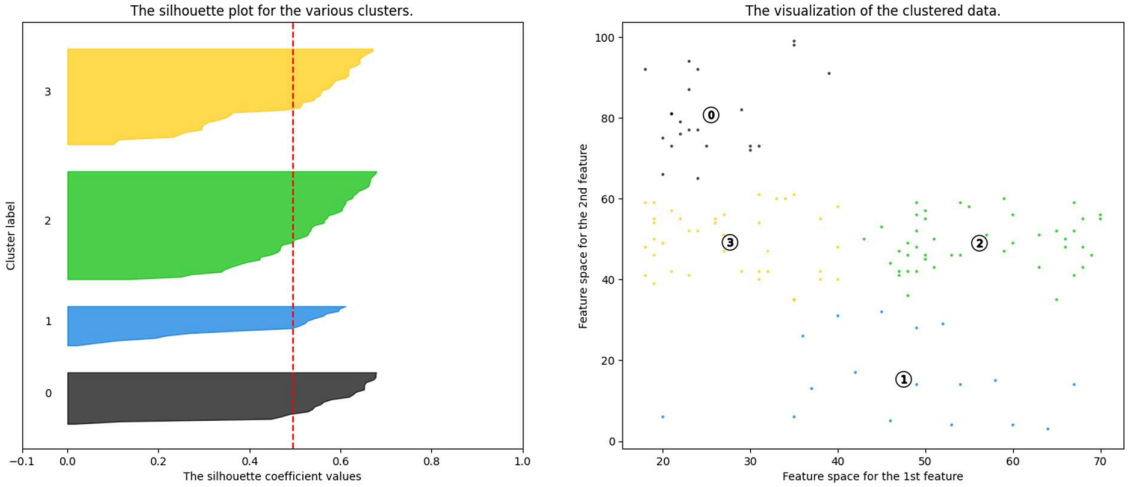
```
For n_clusters = 2 The average silhouette_score is: 0.4273768903935447
For n_clusters = 3 The average silhouette_score is: 0.4392875019111278
For n_clusters = 4 The average silhouette_score is: 0.4953949368236104
For n_clusters = 5 The average silhouette_score is : 0.45172627429691875
For n_clusters = 6 The average silhouette_score is : 0.4500112897884164
```



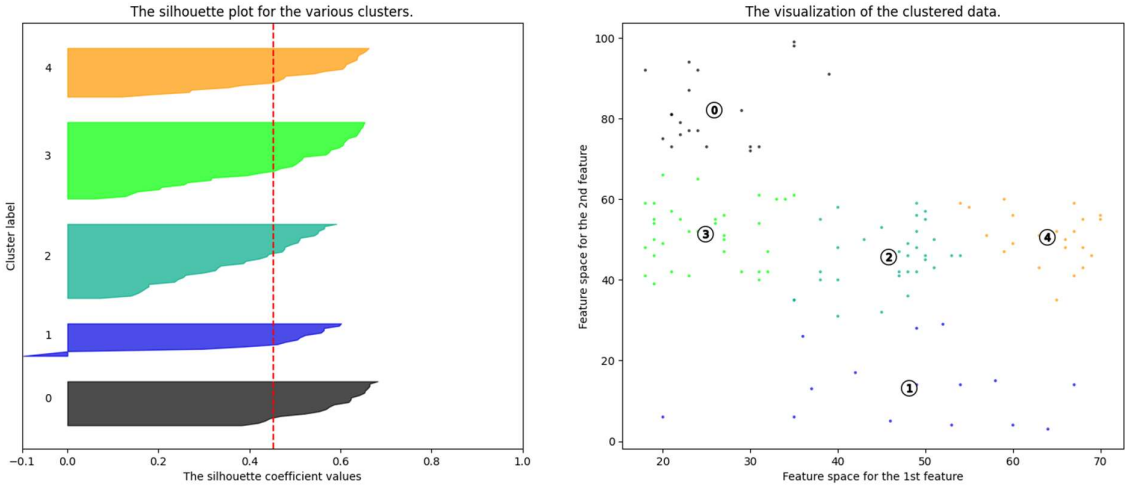
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3

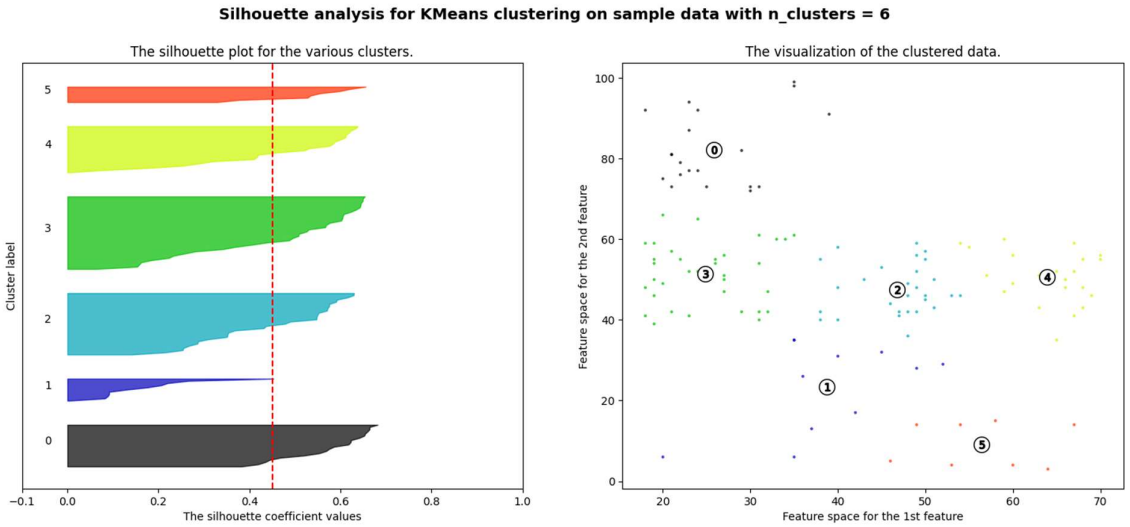


Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



Silhouette analysis for KMeans clustering on sample data with n_clusters = 5





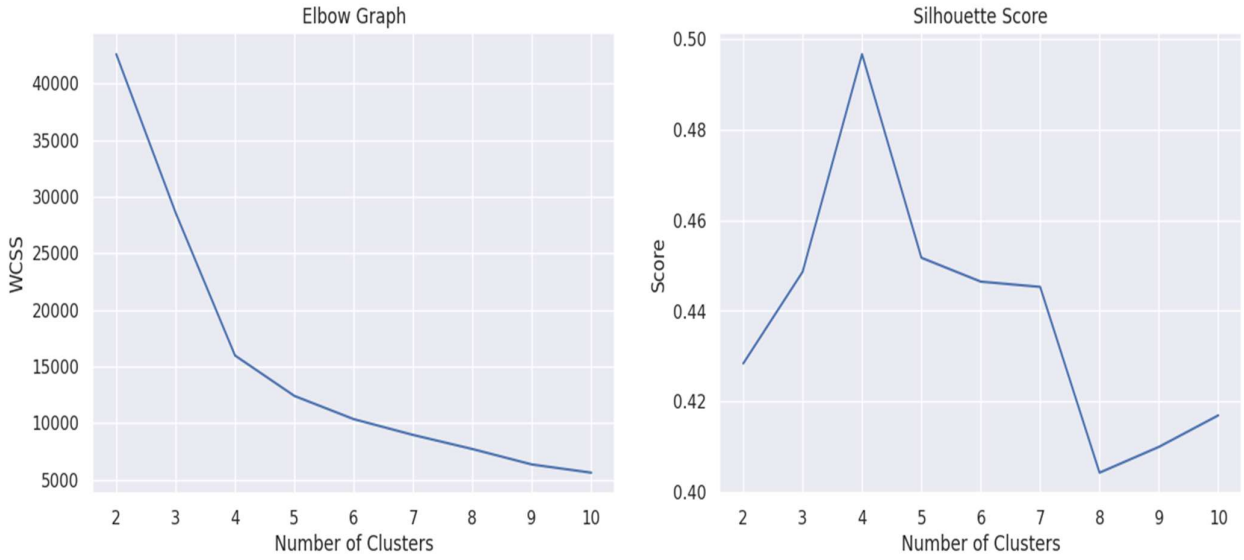


Figure 6: Elbow Method and Silhouette Score for optimal number of Clusters

From the elbow graph and silhouette score we can say that the optimum number of clusters here will be 4.

The final scatter plot clusters look like -

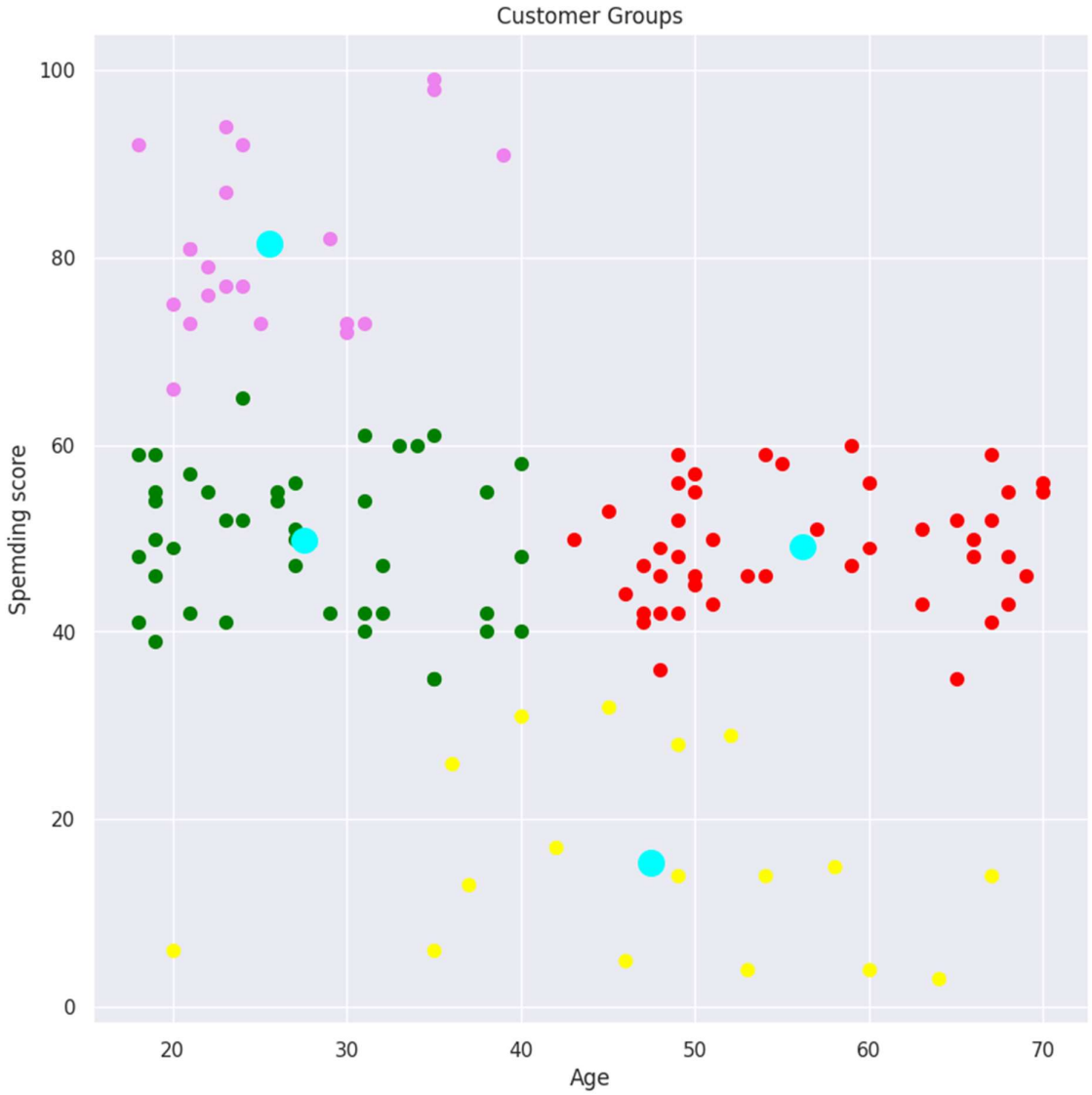


Figure 7: Final scatter plot

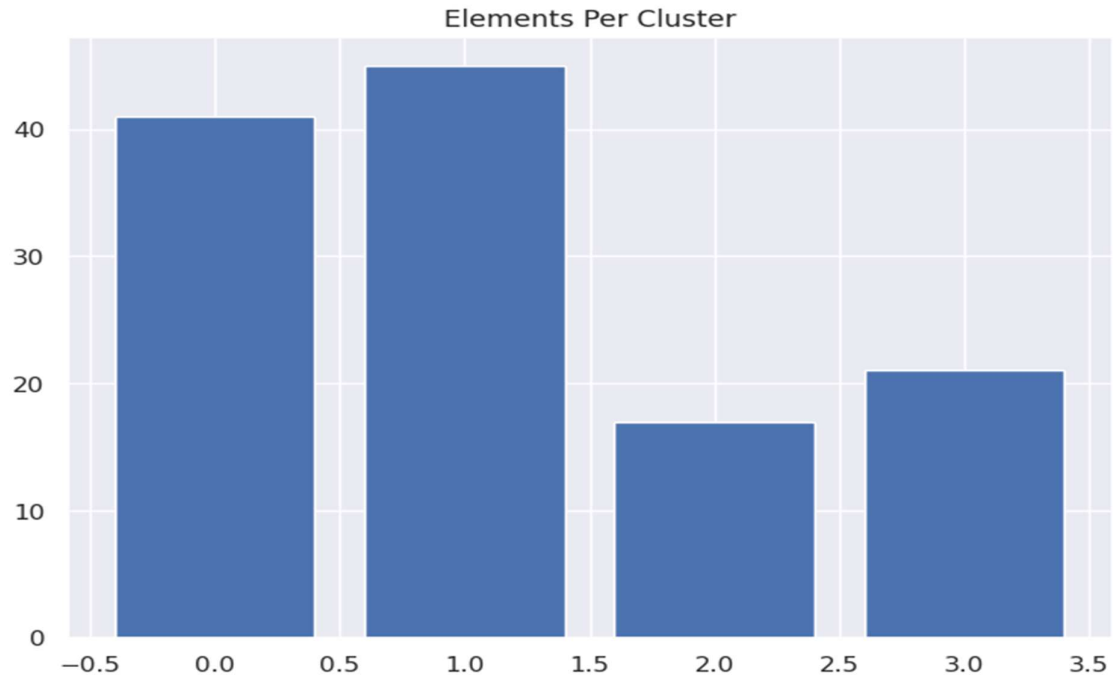


Figure 8: Elements per cluster

Based on the generated output, we can make the following observations:

1. Green Cluster (0.0): This cluster represents a concentrated group of customers with a medium spending score and a younger age. These individuals appear to be quite loyal to the business. To maintain their investment and engagement, the business may consider implementing flashy offers or promotions.

2. Red Cluster (1.0): This cluster comprises a concentrated group of customers with a medium spending score and a higher age. Similar to the green cluster, these customers exhibit loyalty to the business. To retain their loyalty and engagement, the business could consider implementing membership benefits.

From analyzing Cluster 2 and 3, we can conclude that customers with a medium spending score show a higher level of investment in the business compared to those with low or high spending scores.

3. Yellow Cluster (2.0): This cluster consists of customers with a lower spending score (<40), and they form a scattered cluster. Notably, there is a higher proportion of older customers showing interest in the business within this group. To attract younger customers, the business could offer flashy offers and discounts. For the older customer segment, incorporating membership benefits and loyalty points may prove effective.

4. Violet Cluster (3.0): This cluster is formed by customers with a high spending score and a lower age. To maintain their investment and attract more customers from this group, the business could consider incorporating membership benefits and flashy offers.

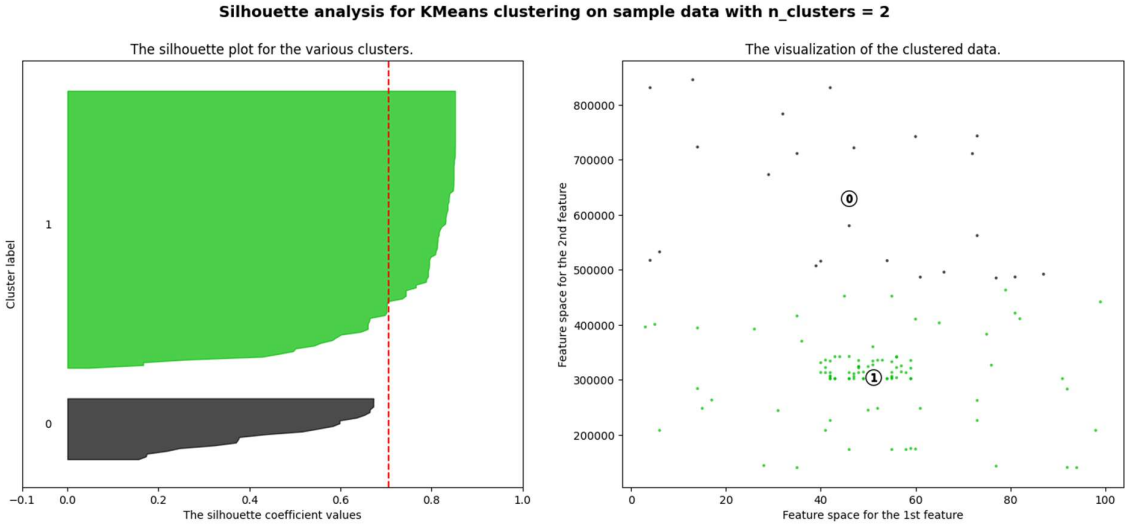
By understanding and targeting these clusters, the business can optimize its strategies to cater to different customer segments, foster loyalty, and attract a wider customer base.

CASE 3 : In the third program we applied K-Means Clustering on the parameters –

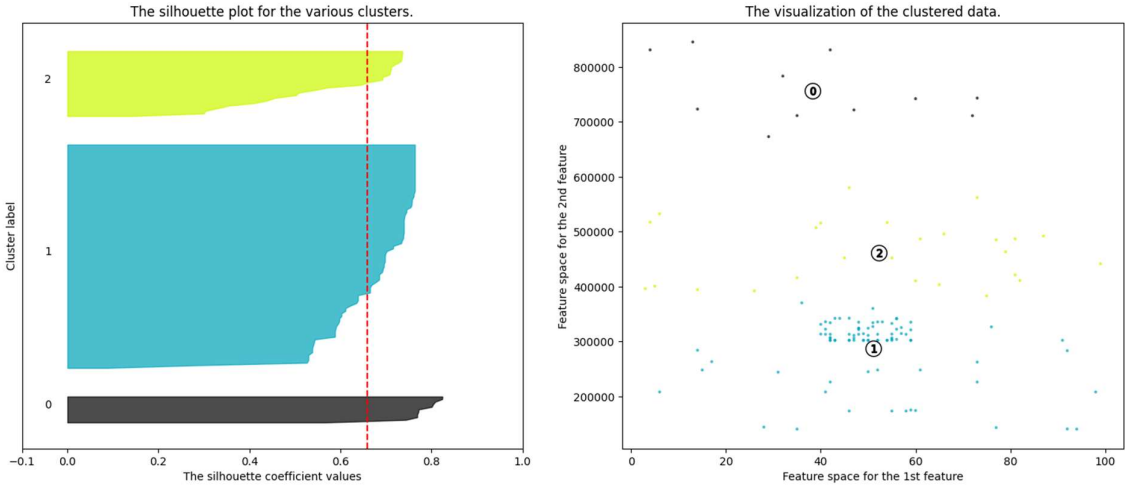
Pin-Code vs Spending score and the results looked like :-

Figure 9 :proves how 6 is the optimal number of clusters.

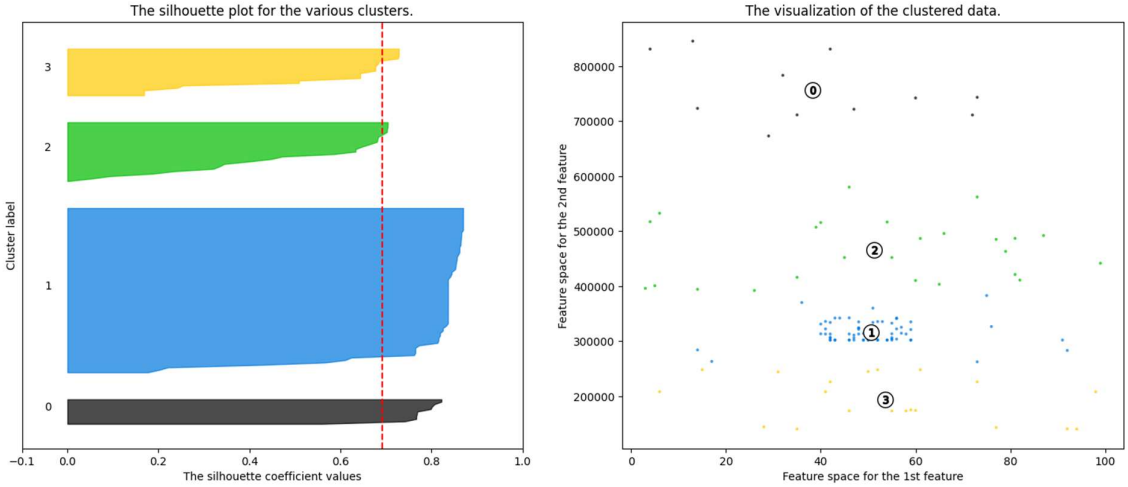
```
For n_clusters = 2 The average silhouette_score is: 0.7051708416017349
For n_clusters = 3 The average silhouette_score is: 0.6595557835052082
For n_clusters = 4 The average silhouette_score is: 0.6918601183884051
For n_clusters = 5 The average silhouette_score is : 0.6936287837961752
For n_clusters = 6 The average silhouette_score is : 0.6922766282170691
```



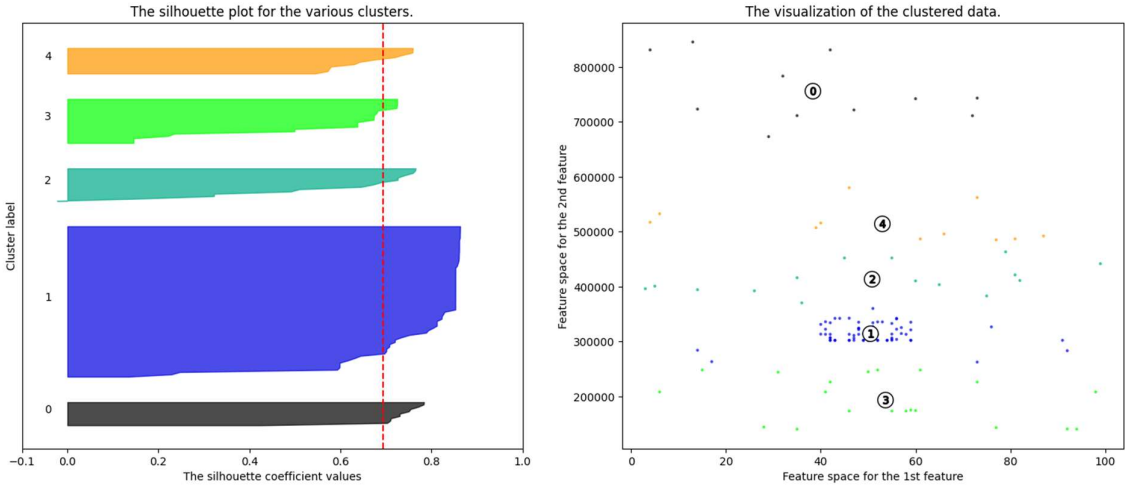
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3

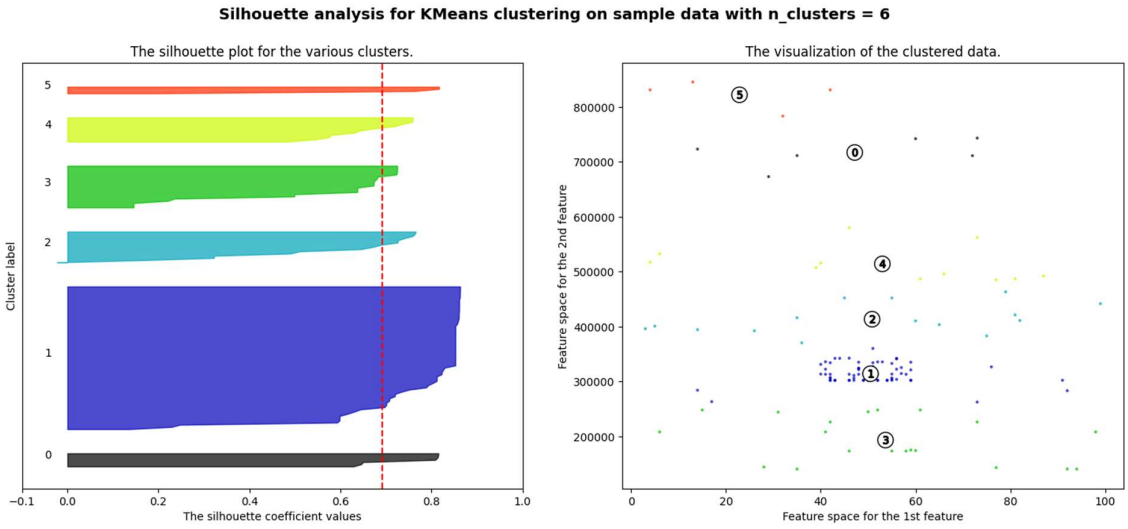


Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



Silhouette analysis for KMeans clustering on sample data with n_clusters = 5





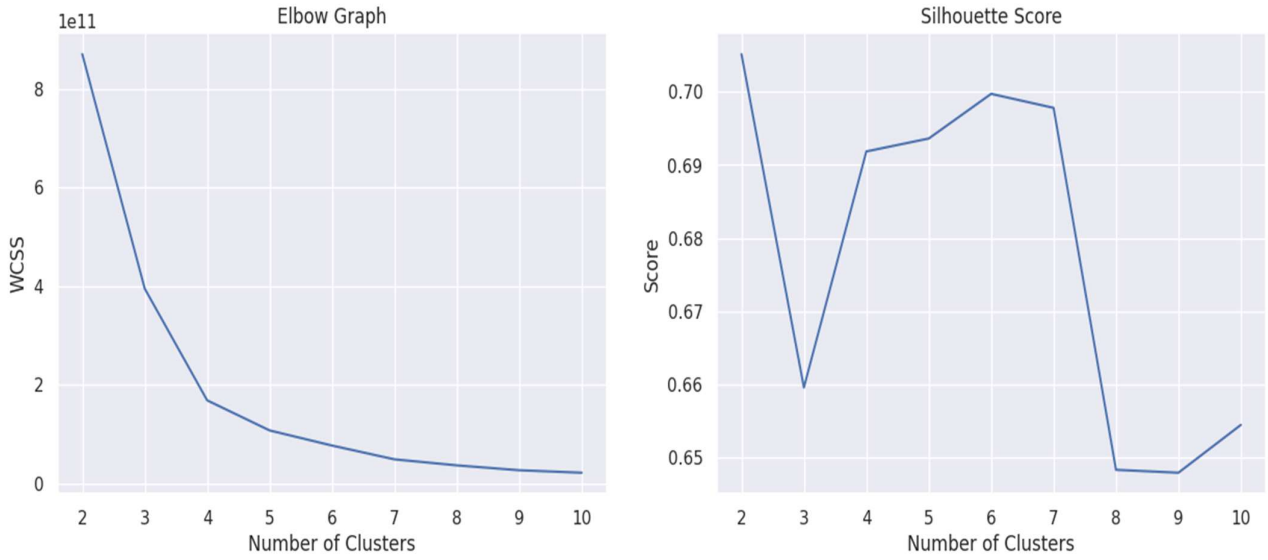


Figure 10: Elbow Method and Silhouette Score for optimal number of Clusters

From the elbow graph and silhouette method we can say that the optimum number of clusters are 6.

The final scatter plot clusters look like -

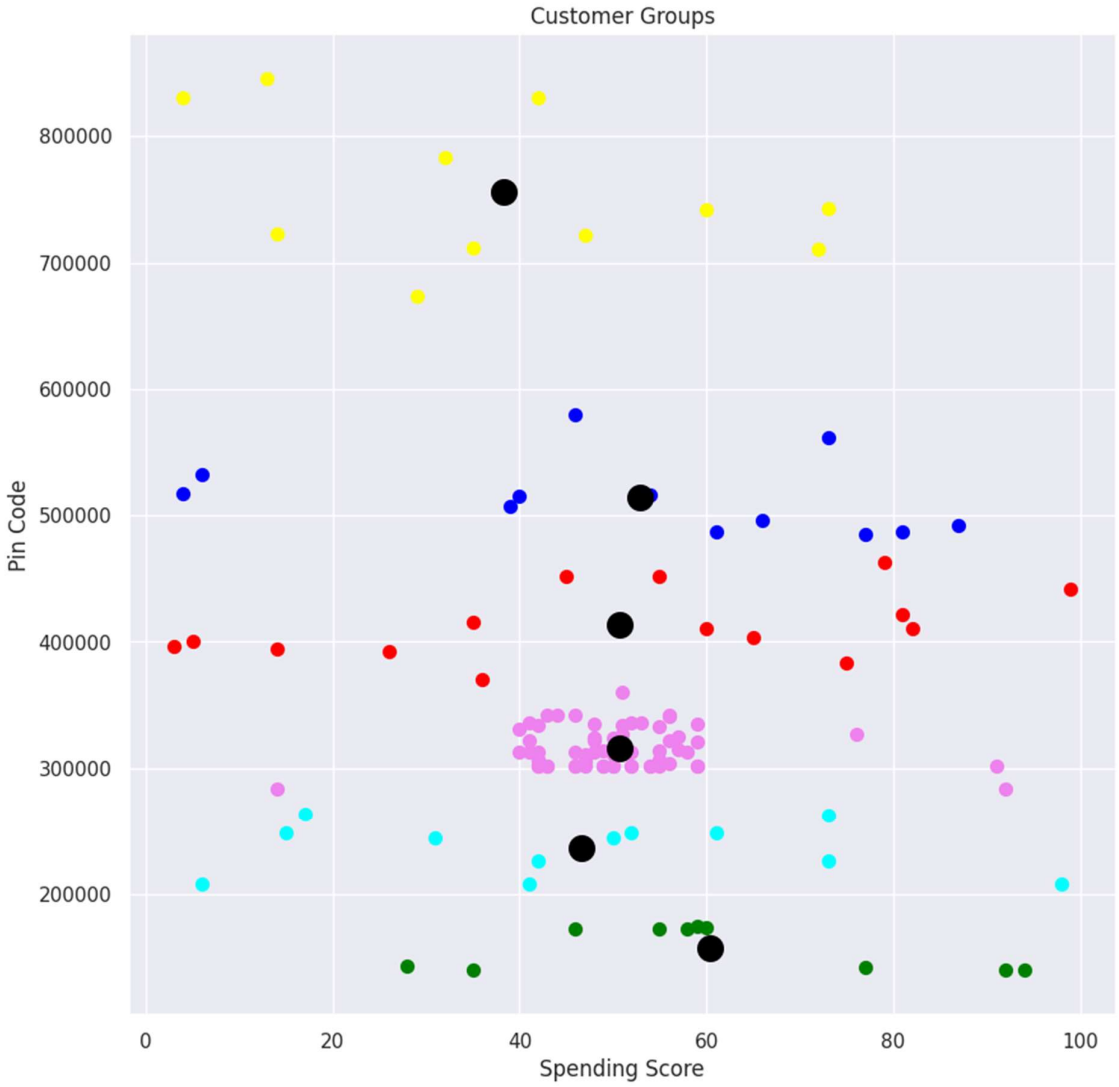


Figure 11: Final scatter plot

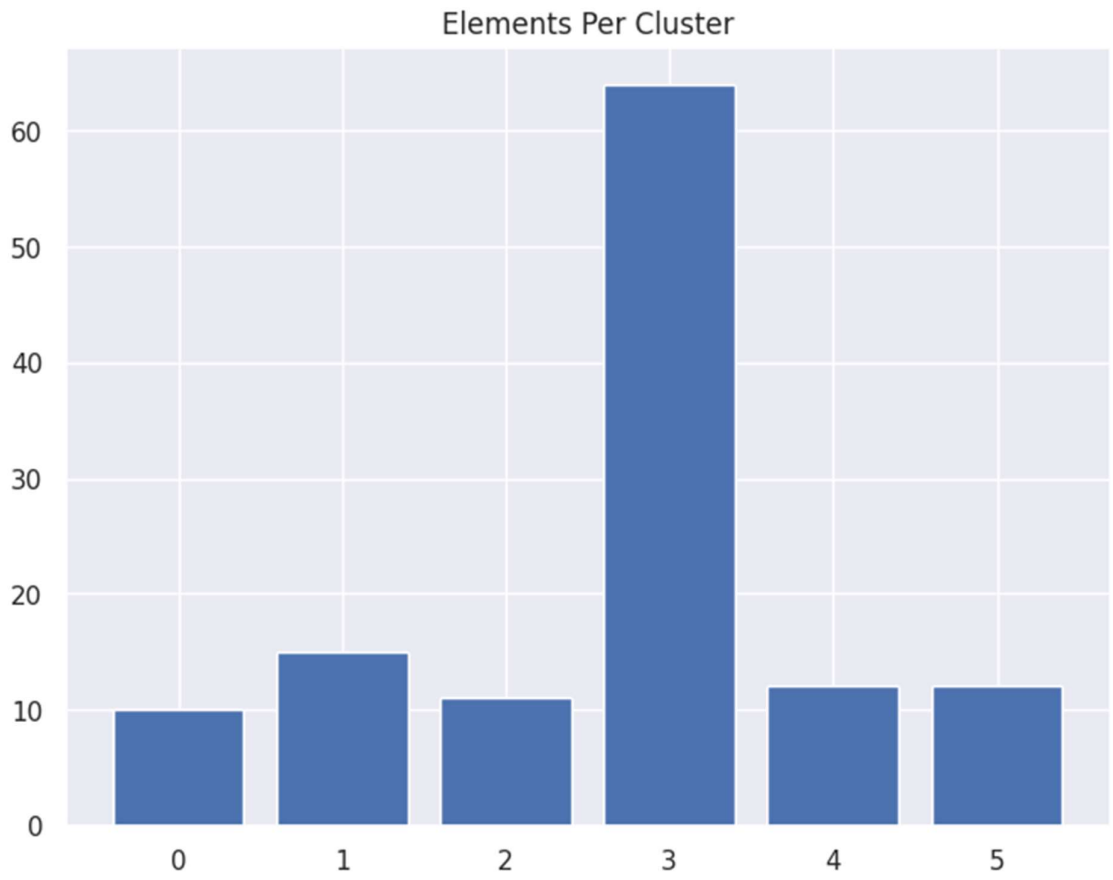


Figure 12: Elements per cluster

• Green Cluster (0.0): The customers residing in this region form a scattered cluster, indicating their relatively low interest in purchasing products from this business. To increase customer engagement, strategies such as providing free delivery services and offering discounted deals can be implemented.

• Red Cluster (1.0): The customers residing in this region also form a scattered cluster, suggesting their diminished interest in buying products from this business. To enhance customer investment, offering free delivery services and discounted deals can be effective strategies.

• Yellow Cluster (2.0): The customers residing in this region form a highly scattered cluster, indicating their low level of investment in the products offered by this business. It is possible that these customers are located in areas where the business does not have outlets or delivery options are limited. To attract these customers, the business may consider opening an outlet in this area or expanding their delivery range.

• Violet Cluster (3.0): This cluster is the most concentrated, indicating that customers residing in this region with a moderate spending score (40-60) are highly invested in purchasing products from this business. It can be inferred that the business outlets are conveniently located in this region, making the products easily accessible. To further attract customers with lower spending scores, the business can implement discount offers, while for customers with higher spending scores, implementing membership benefits and loyalty points can be effective strategies.

• Blue Cluster (4.0): The customers residing in this region form a scattered cluster, suggesting their relatively low interest in buying products from this business. To increase their investment in the products, the business can offer free delivery services and discounted deals.

• Cyan Cluster (5.0): Similar to the blue cluster, the customers residing in this region also form a scattered cluster, indicating their decreased interest in purchasing products from this business. To enhance their engagement, the business can provide free delivery services and discounted deals.

By analyzing the characteristics and preferences of customers in each cluster, the business can tailor its marketing and operational strategies to effectively engage customers, expand its customer base, and boost overall sales.

Dataset 2:

The dataset provided comprises 8 columns: Customer ID, Gender, Age, Annual Income, Spending Score, Warehouse Pin-code, Customer Pin-code, and Zone. It contains data for a total of 462 customers. During the coding process, the cluster value was appended to the dataset.

Due to the substantial size of the dataset, consisting of 462 customer data entries, we utilized a bar graph to facilitate the analysis of the customer distribution across different clusters.

	CustomerID	Gender	Age	Annual_Income	Spending_Score	Ware_Pin	Customer_Pin	Zone	clusters
0	1	Male	55	55	72	515631	434011	Kurnool Region	1
1	2	Male	43	57	44	515631	613712	Kurnool Region	3
2	3	Female	30	66	31	515631	403813	Kurnool Region	4
3	4	Female	30	38	92	515581	687163	Kurnool Region	2
4	5	Female	34	73	23	515581	353527	Kurnool Region	4

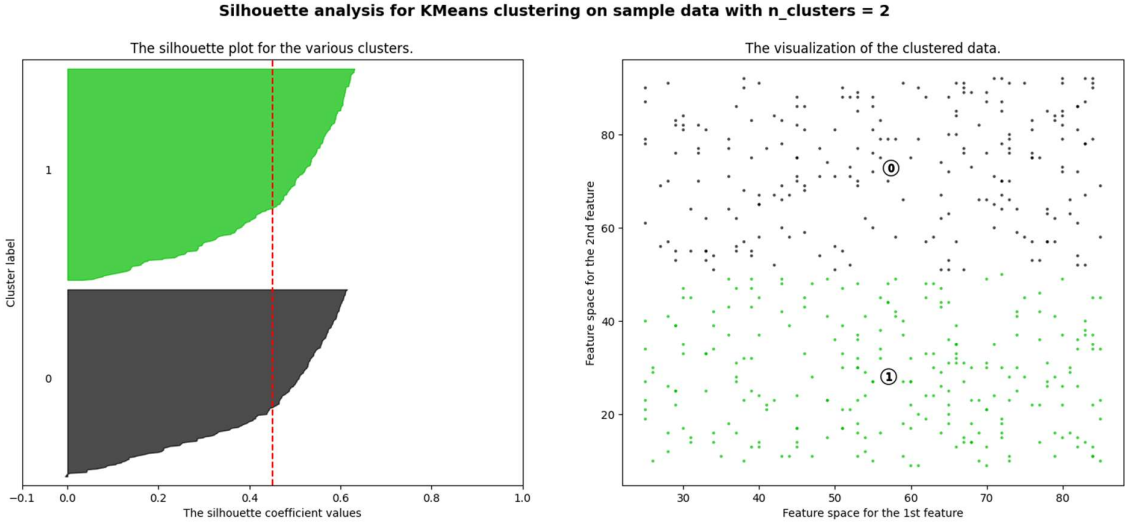
Table 3 : First five rows of the dataset 2

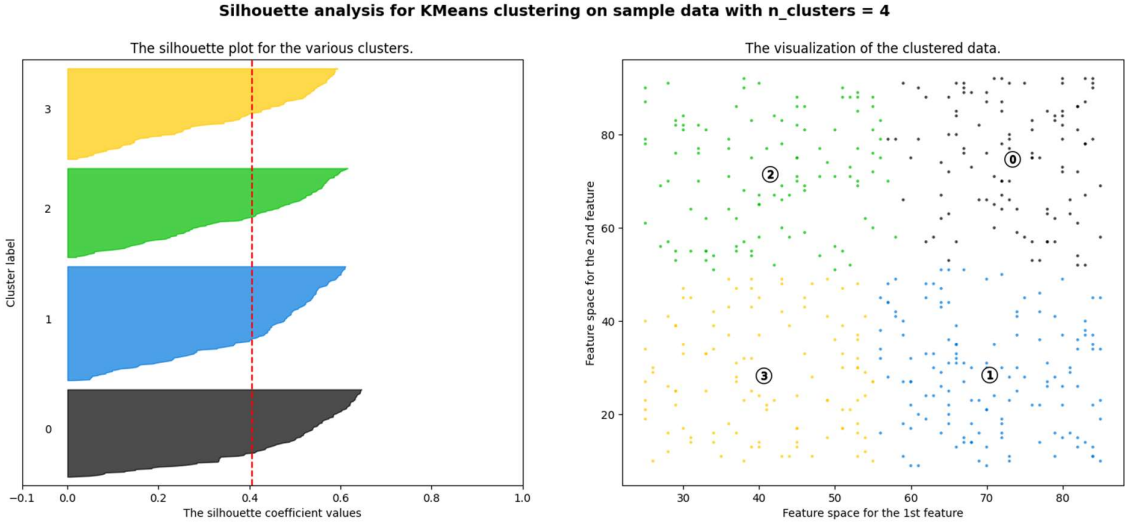
Analyzing the Results :

CASE 1: In the first program we applied K-Means Clustering on the parameters – Annual Income vs Spending Score and the results looked like:

Figure 13 :proves how 5 is the optimal number of clusters.

```
For n_clusters = 2 The average silhouette_score is : 0.4501030908768755
For n_clusters = 3 The average silhouette_score is : 0.4204397683865185
For n_clusters = 4 The average silhouette_score is : 0.4049933674155039
For n_clusters = 5 The average silhouette_score is : 0.4116081952655662
For n_clusters = 6 The average silhouette_score is : 0.402872885781395
```





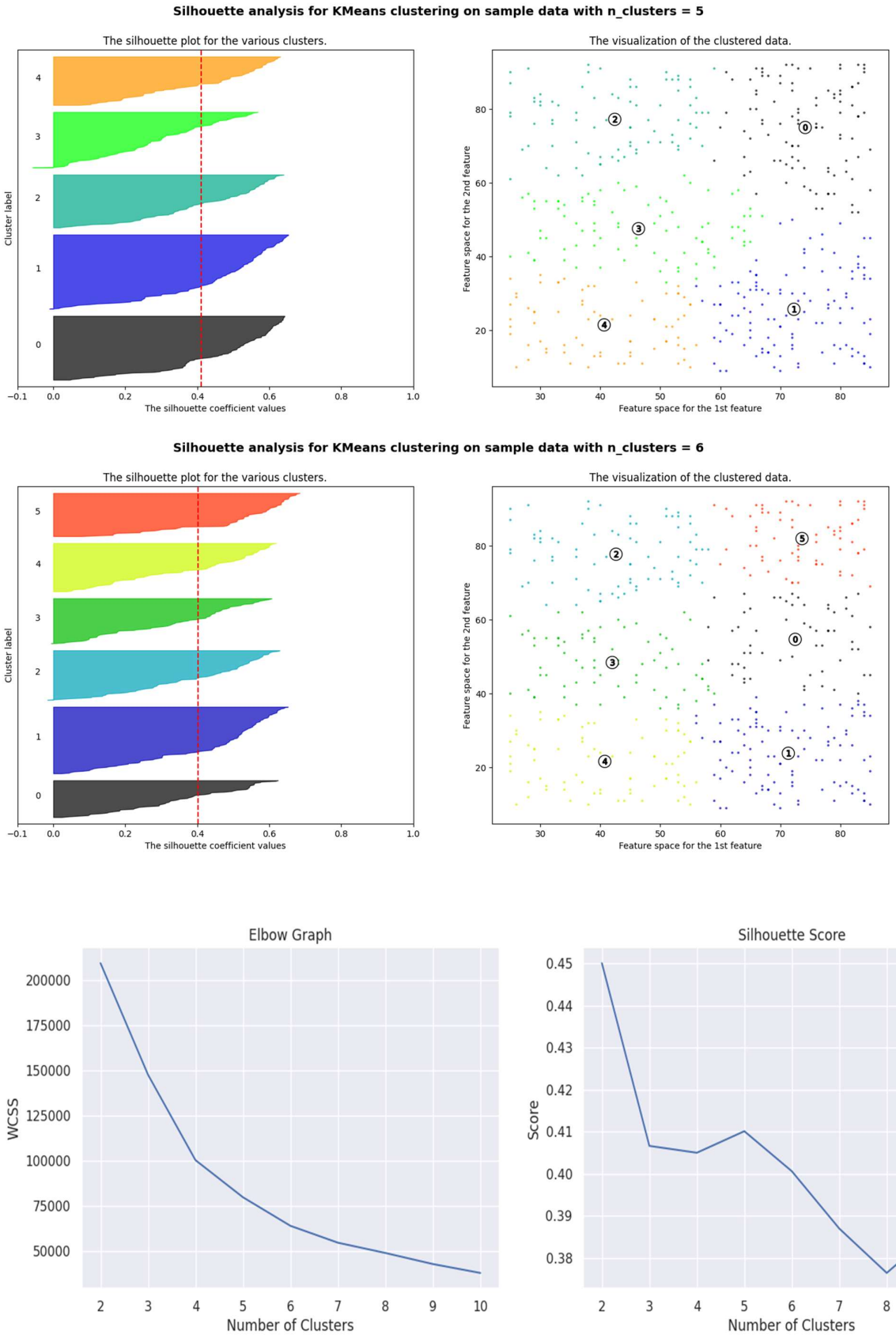


Figure 14: Elbow Method and Silhouette Score for optimal number of Clusters

From this elbow graph and silhouette score we can say that the optimum number of clusters in this case will be 5.

The final scatter plot clusters look like –

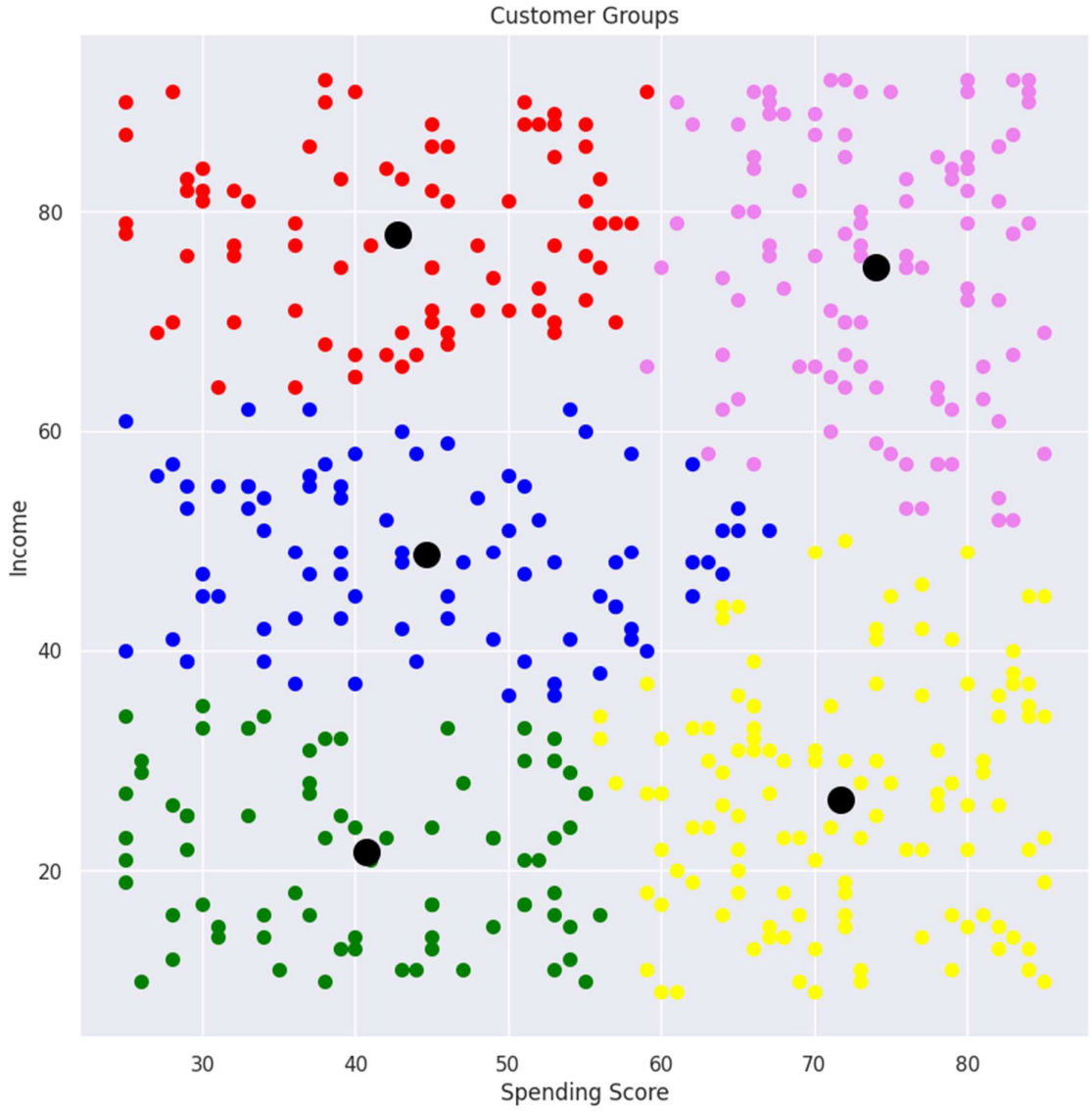


Figure 15: Final scatter plot

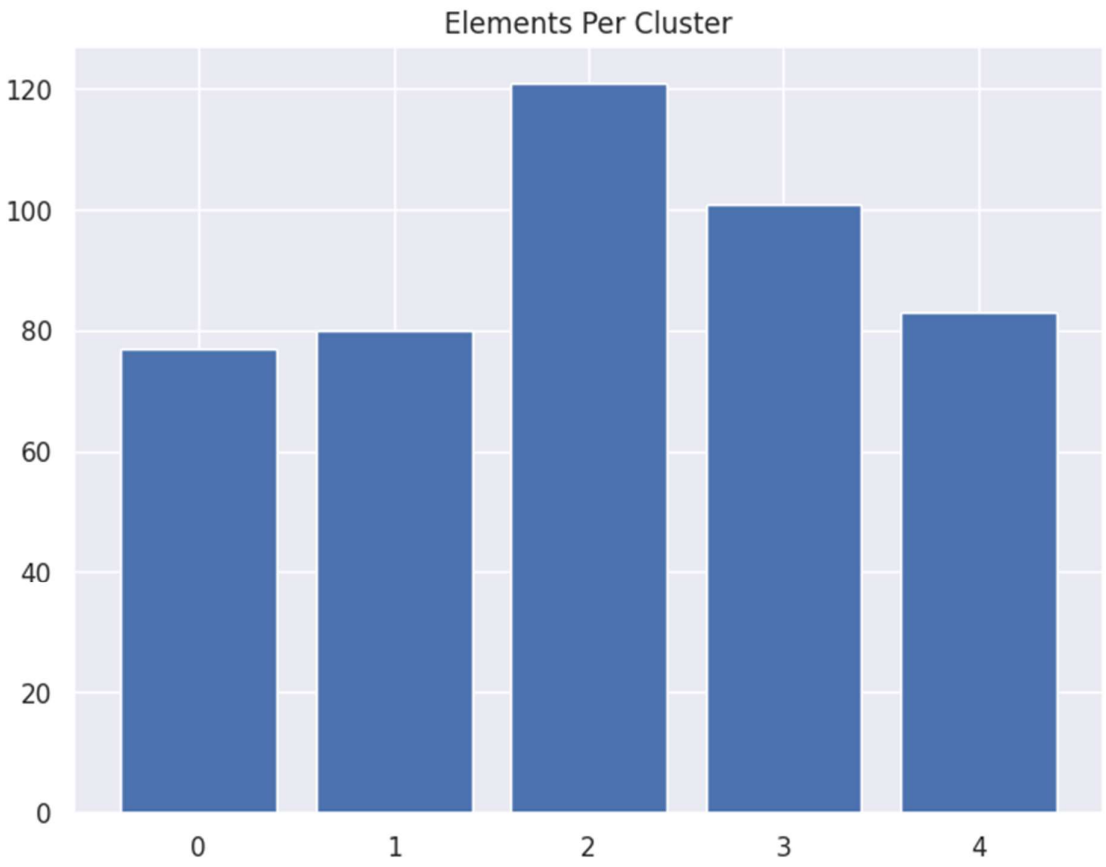


FIGURE 16: Elements per cluster

In the dataset, the clusters can be described as follows:

0.0 - Green Cluster: Customers in this cluster exhibit a low spending score and low income. They appear to be less invested in purchasing products from this business compared to other customer groups.

1.0 - Red Cluster: Customers in this cluster have a high income but a low spending score. They show a slightly higher level of investment in buying products from this business compared to the previous cluster.

2.0 - Yellow Cluster: This cluster stands out as the most concentrated one. It consists of customers with a high spending score and low income. These customers exhibit a strong interest and investment in purchasing products from this business.

3.0 - Violet Cluster: This cluster comprises customers with both a high spending score and high income. They are highly invested in buying from this business and form the second most concentrated cluster.

4.0 - Blue Cluster: This is the third most concentrated cluster, consisting of customers with a moderate income and a low to medium spending score.

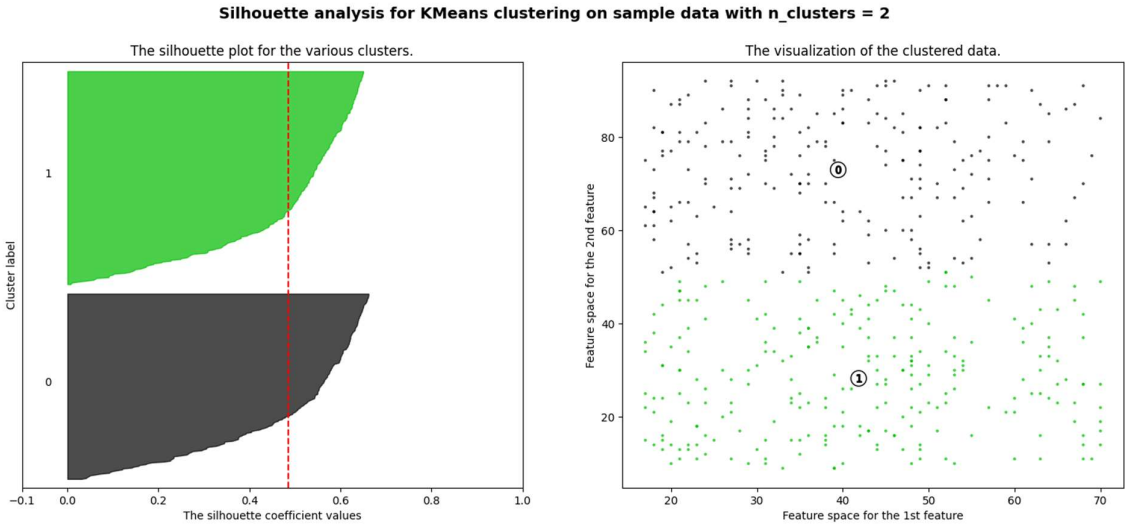
By understanding the characteristics and behaviors of customers in each cluster, the business can tailor its marketing and sales strategies to effectively engage and cater to the specific needs of different customer segments.

CASE 2 : In the second case we applied K-Means Clustering on the parameters –

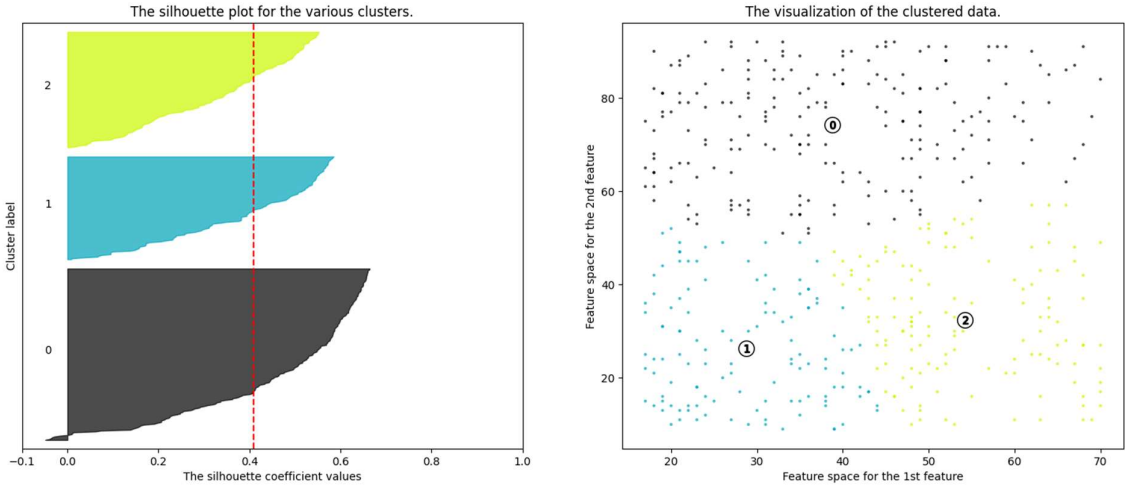
Spending Score vs Age and the results looked like :-

Figure 17 :proves how 5 is the optimal number of clusters.

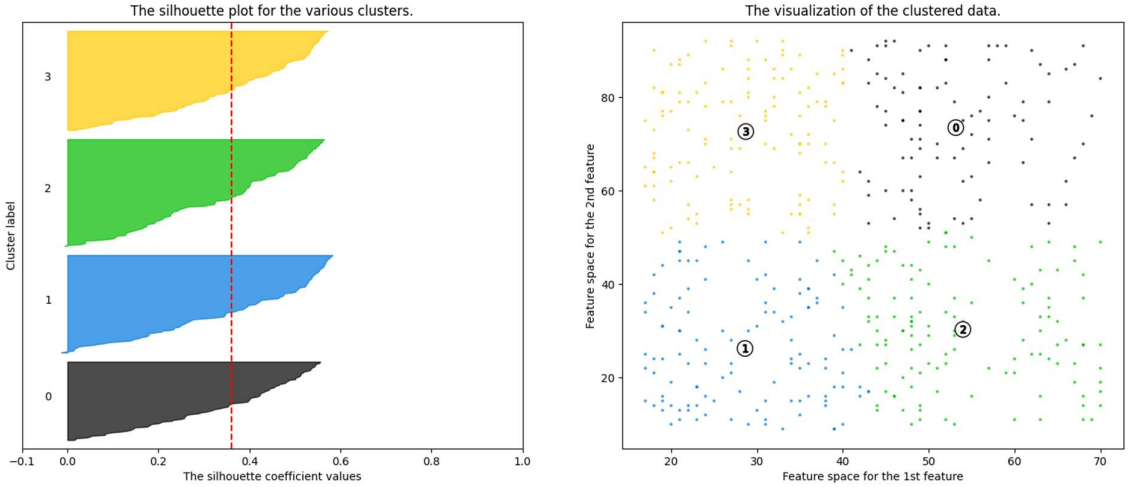
```
For n_clusters = 2 The average silhouette_score is : 0.4858793595690374
For n_clusters = 3 The average silhouette_score is : 0.409303131229153
For n_clusters = 4 The average silhouette_score is : 0.3597251354382038
For n_clusters = 5 The average silhouette_score is : 0.37918852030118116
For n_clusters = 6 The average silhouette_score is : 0.3846591558553363
```



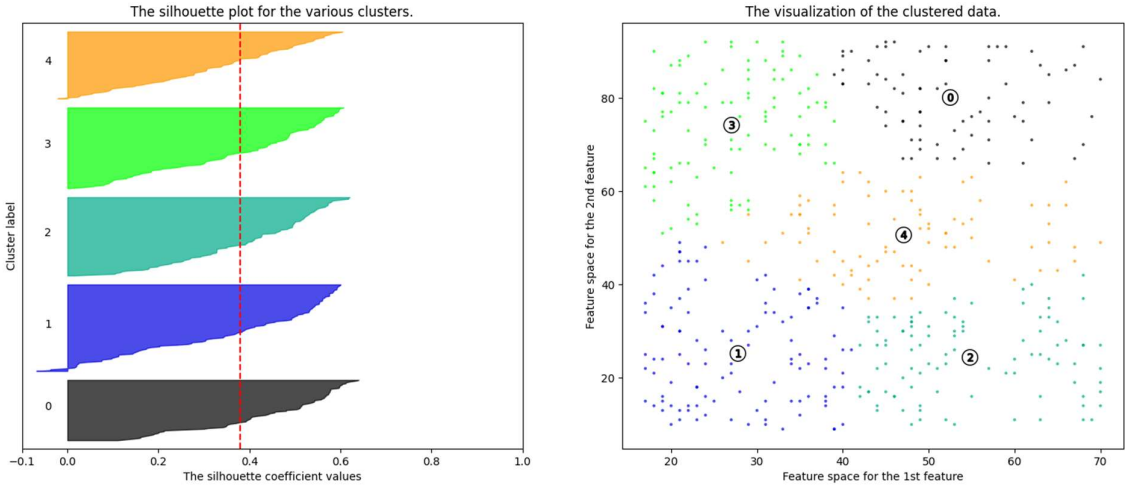
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3

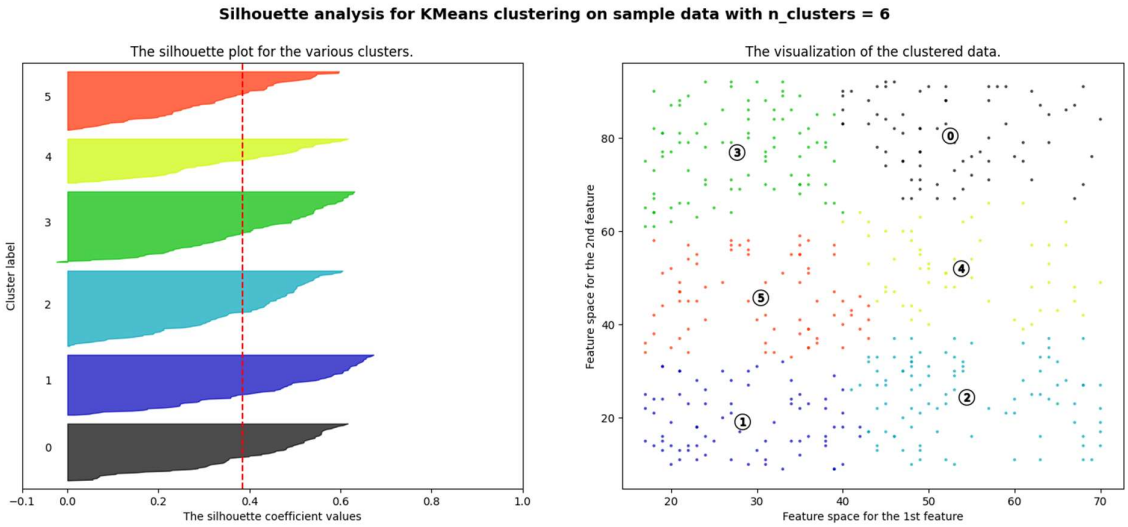


Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



Silhouette analysis for KMeans clustering on sample data with n_clusters = 5





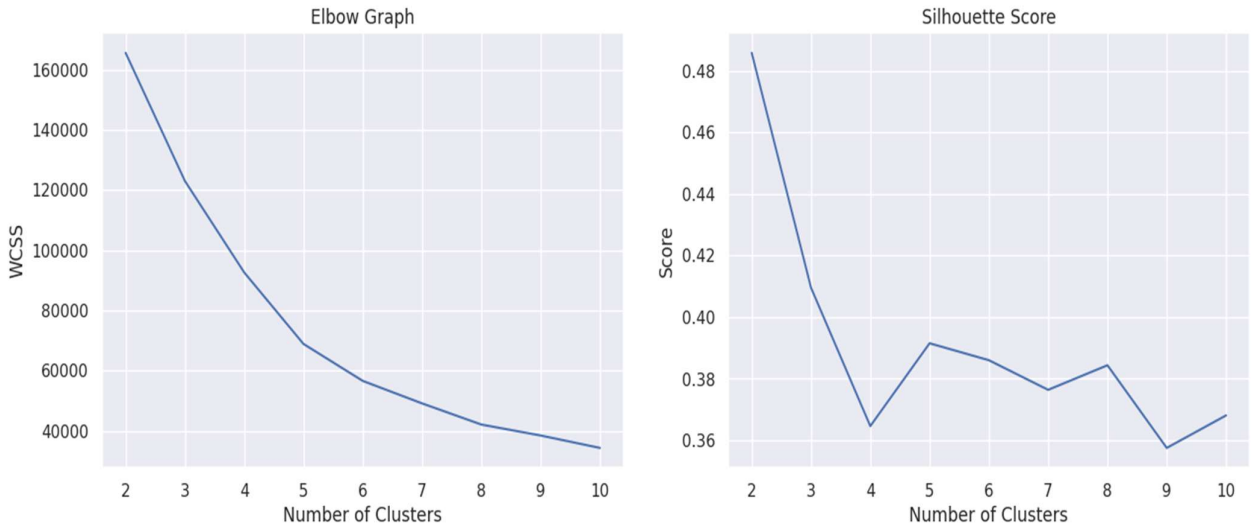


Figure 18: Elbow Method and Silhouette Score for optimal number of Clusters

From this elbow graph and silhouette graphs we can say that the optimum number of clusters is 5.

The final scatter plot clusters look like –

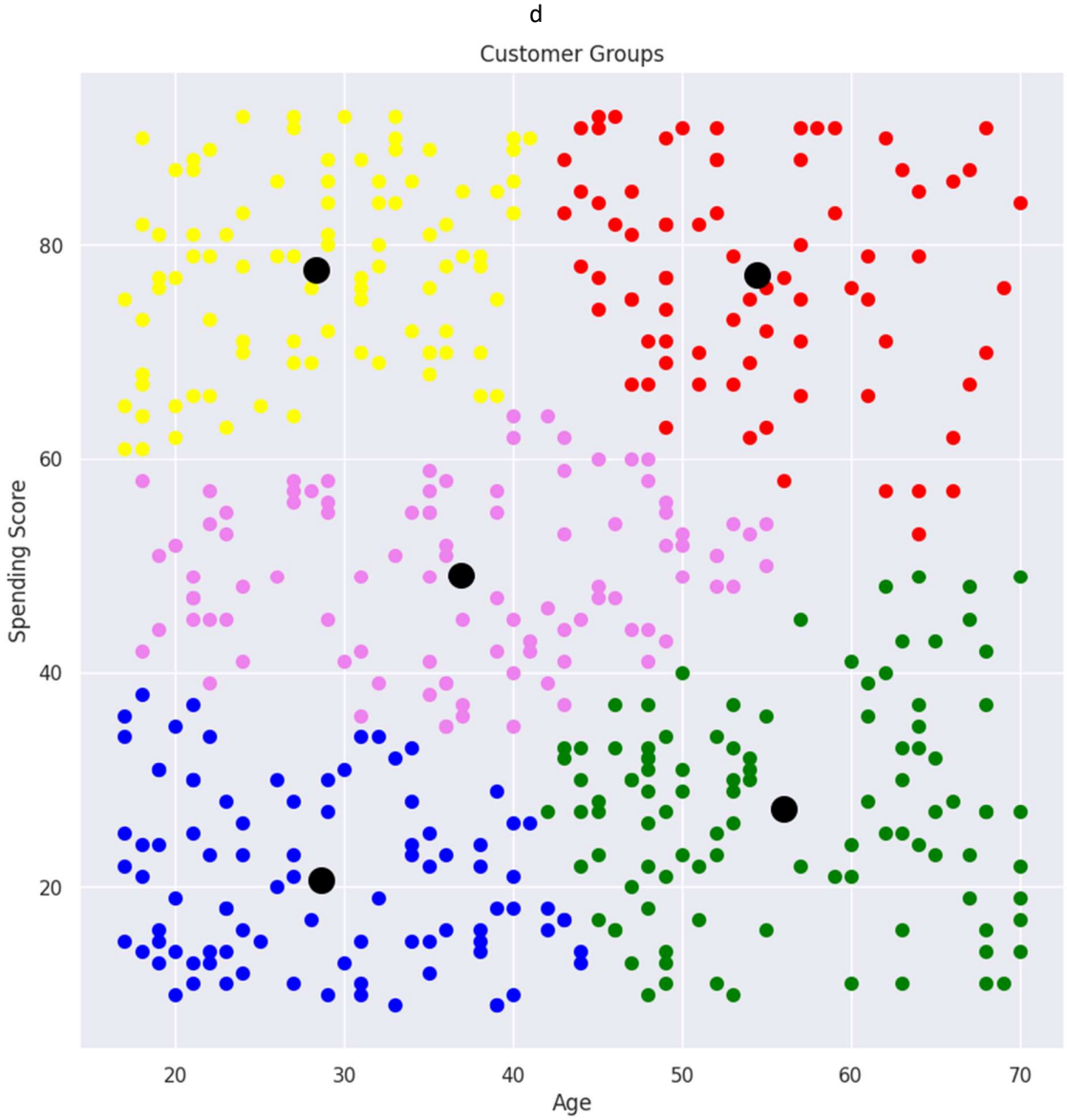


Figure 19: Final scatter plot

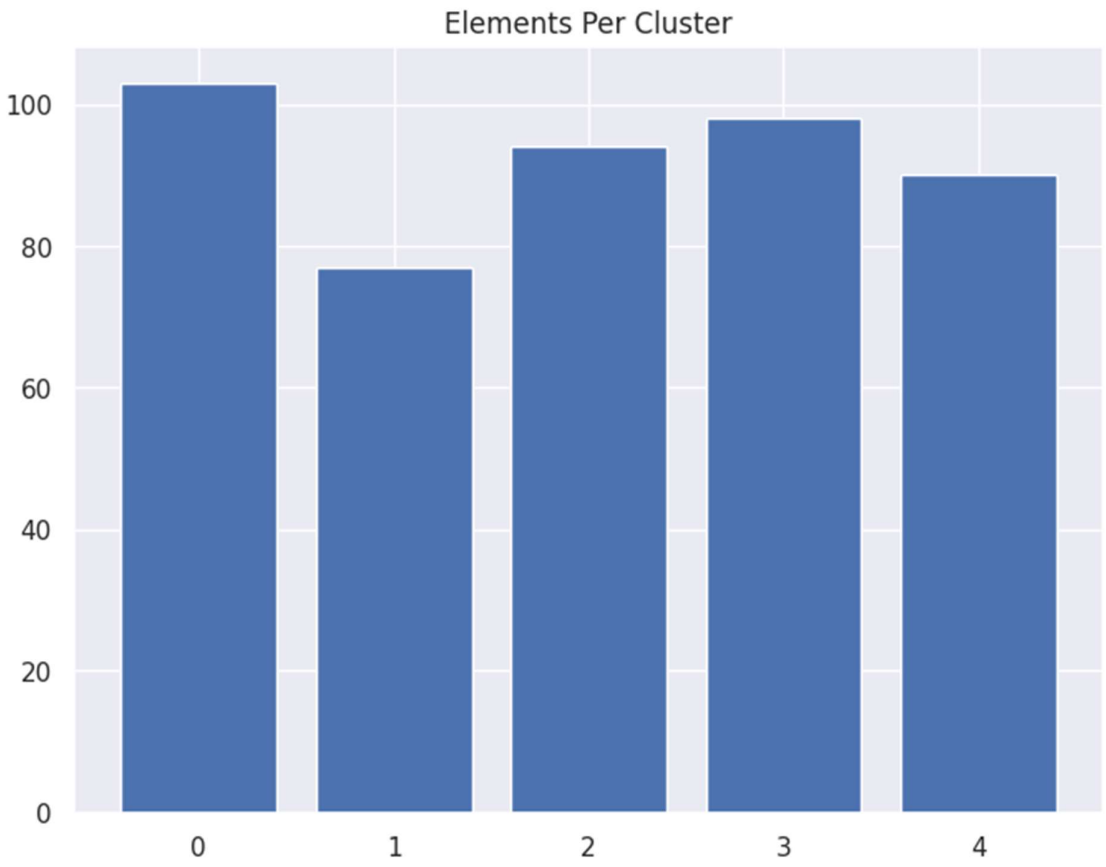


FIGURE 20: Elements per cluster

In the given dataset, the clusters can be described as follows:

0.0 - Green Cluster: This cluster contains the largest number of customers, indicating that older individuals with a low spending score are more invested in this business. It suggests that the products offered by this business are budget-friendly and attract customers who prioritize affordability.

1.0 - Red Cluster: This cluster is the least concentrated, indicating that customers with older age and a higher spending score are less invested in buying products from this business compared to other customer groups. It suggests that this particular customer segment may have different preferences or priorities when it comes to purchasing decisions.

2.0 - Yellow Cluster: This cluster is the third most concentrated, consisting of customers with a lower age and a high spending score. It suggests that younger customers with a higher disposable income are particularly interested and invested in purchasing products from this business.

3.0 - Violet Cluster: This cluster is the second most concentrated, encompassing customers of almost all age groups with a moderate spending score. It suggests that customers across various age ranges show a consistent level of investment in this business.

4.0 - Blue Cluster: This cluster primarily includes younger individuals with a low spending score and is the second most scattered cluster after the red cluster. It implies that customers in this group may have lower purchasing power or different preferences compared to other clusters.

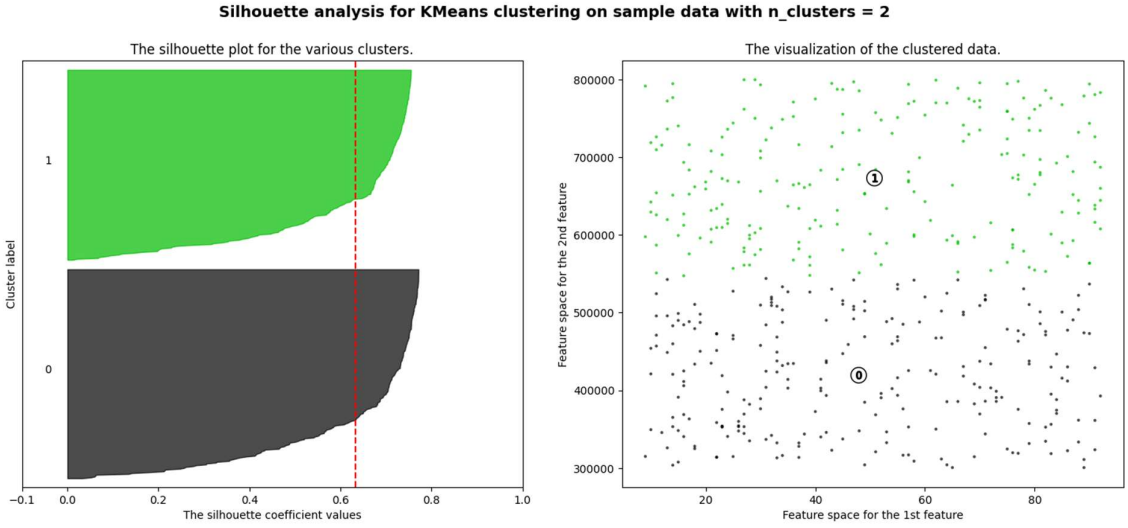
By understanding the characteristics and preferences of customers in each cluster, the business can tailor its strategies to effectively target and engage specific customer segments, thus maximizing customer satisfaction and overall business growth.

CASE 3 : In the third program we applied K-Means Clustering on the parameters –

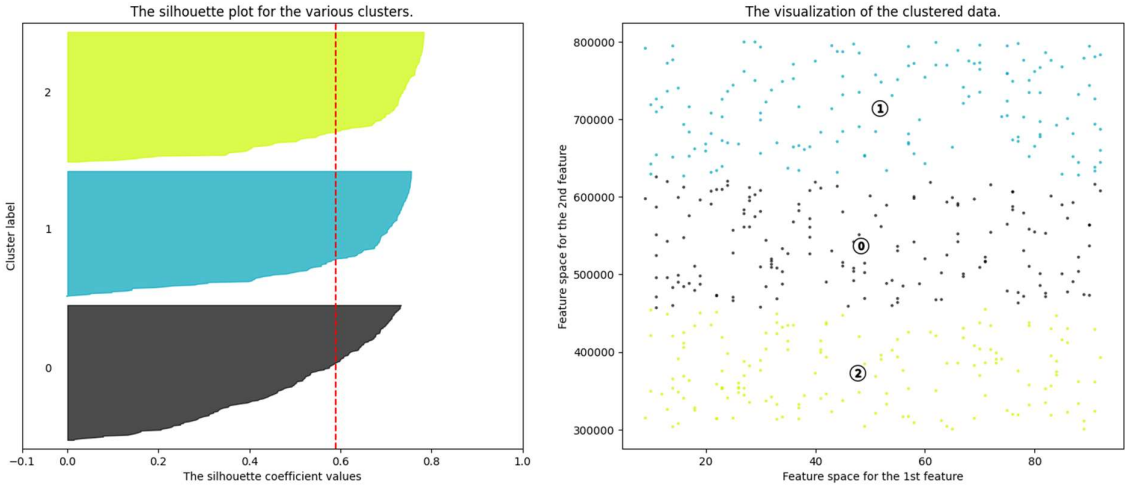
Pin-Code vs Spending score and the results looked like :-

Figure 21 :proves how 6 is the optimal number of clusters.

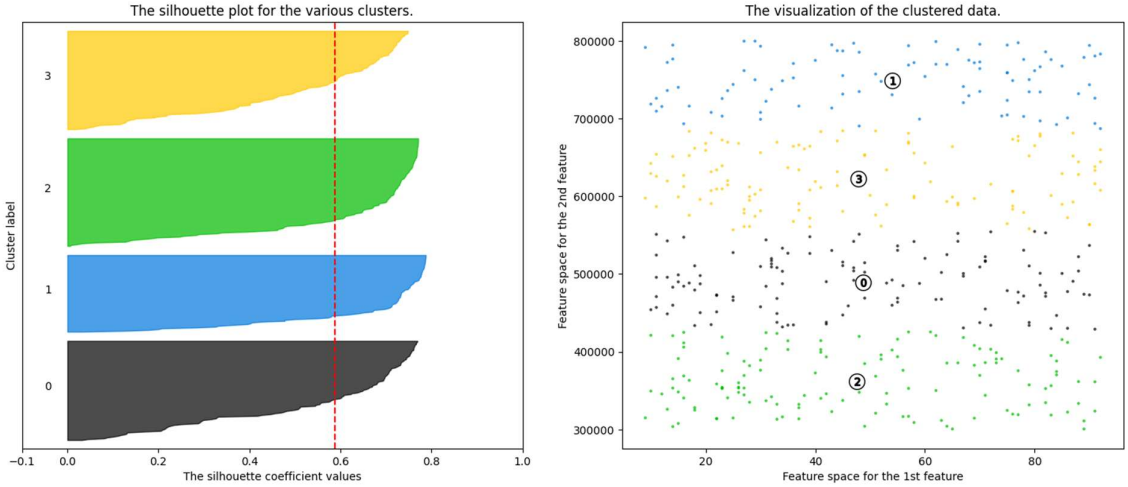
```
For n_clusters = 2 The average silhouette_score is : 0.6332393184473132
For n_clusters = 3 The average silhouette_score is : 0.5888768884614077
For n_clusters = 4 The average silhouette_score is : 0.5869725890224581
For n_clusters = 5 The average silhouette_score is : 0.5738630509939165
For n_clusters = 6 The average silhouette_score is : 0.5851475760744748
```



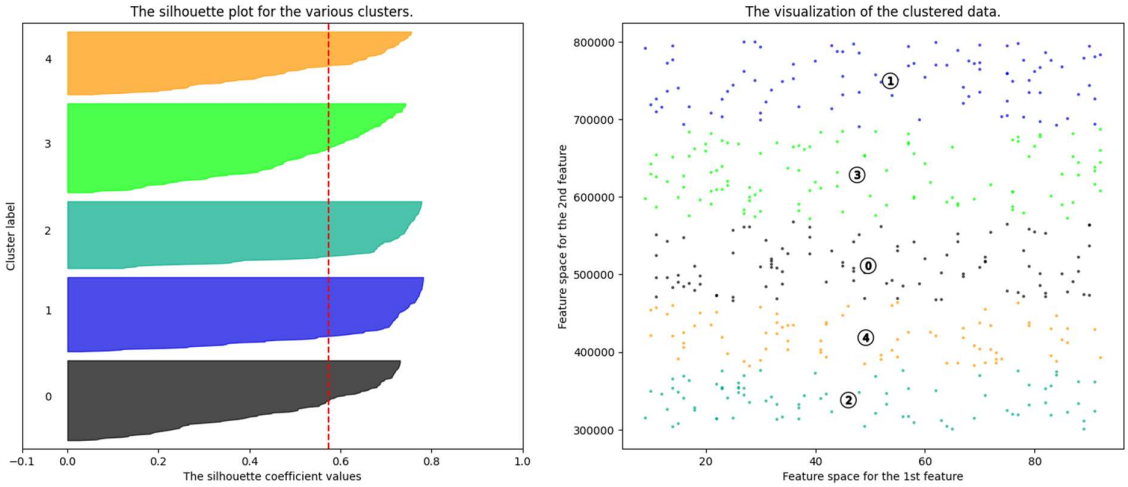
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3

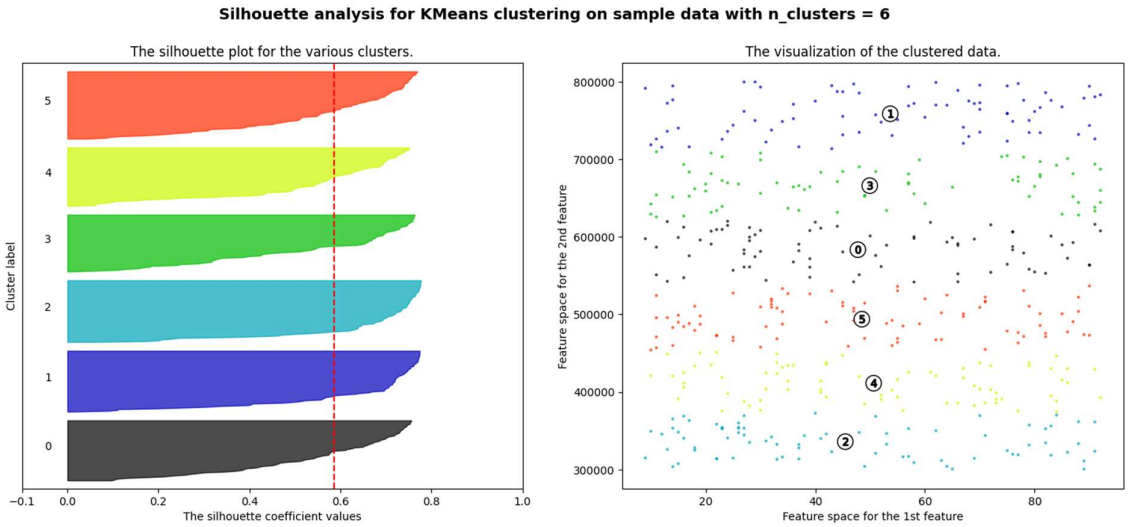


Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



Silhouette analysis for KMeans clustering on sample data with n_clusters = 5





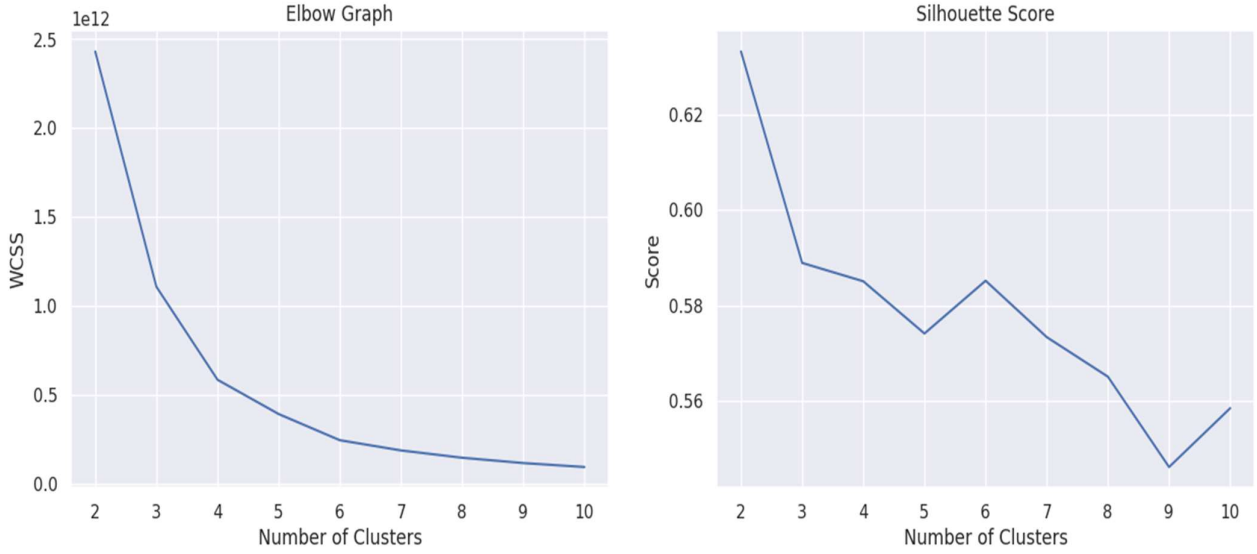


Figure 22: Elbow Method and Silhouette Score for optimal number of Clusters

From the elbow graph and silhouette method we can say that the optimum number of clusters here are 6.

The final scatter plot clusters look like –

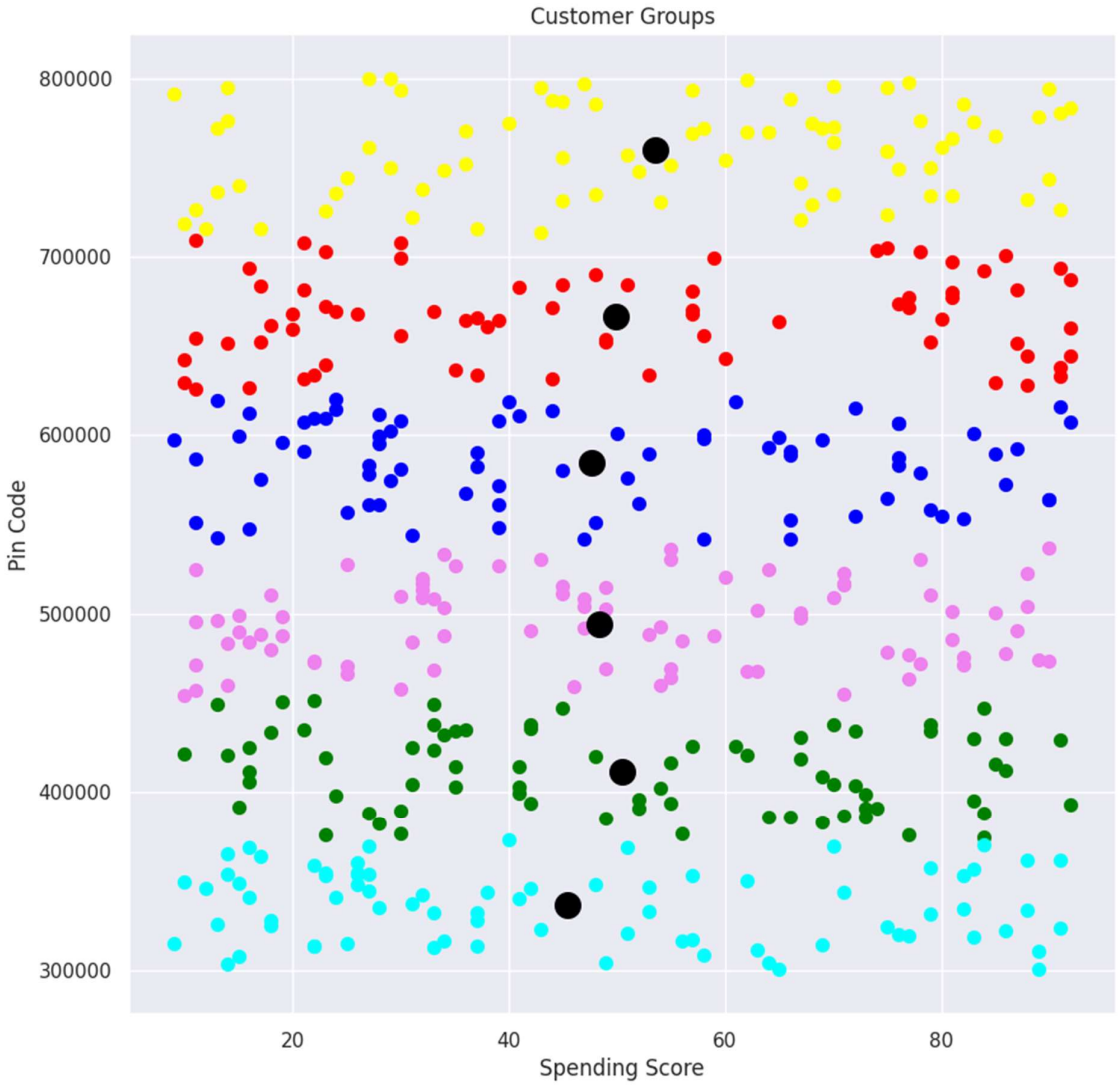


Figure 23: Final scatter plot

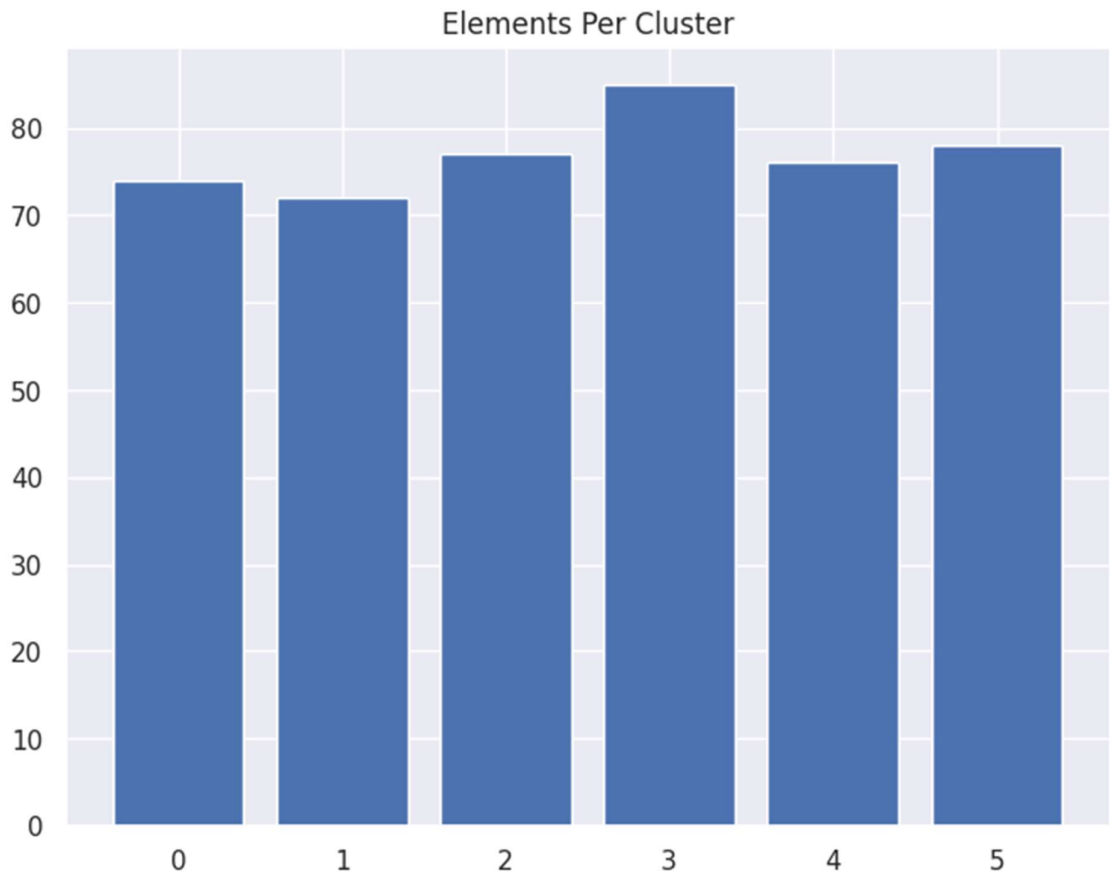


Figure 24: Elements per cluster

Here is the information provided in a better language:

• 0.0 - Green Cluster: Customers from this region form a relatively concentrated cluster, indicating their significant investment in buying products from this business.

• 1.0 - Red Cluster: This cluster is the least concentrated, suggesting that customers from this region are less interested in buying products from this business compared to customers from other regions.

• 2.0 - Yellow Cluster: This cluster is the third most concentrated, indicating that customers from this region are quite invested in buying products from this business.

• 3.0 - Violet Cluster: Most customers from this region are highly invested in buying products from this business, making it the most concentrated cluster.

• 4.0 - Blue Cluster: Customers from this region form a fairly concentrated cluster, suggesting their significant investment in buying products from this business.

• 5.0 - Cyan Cluster: This group of customers, residing in this region, forms the second most concentrated cluster, indicating their strong investment in buying products from this business.

By analyzing the concentration and investment levels of customers in each cluster, the business can effectively target specific regions and customer segments to maximize engagement and drive sales.

Comparative STUDY

There have been many research works done in the field of 'Customer Segmentation', each of them having several advantages and disadvantages. A comparative study of some of them with the research work mentioned in this paper has been explained in Table 4.

Table 4. Comparative study of some of the research works done on 'Customer Segmentation'

PAPER	WORK	TARGET	GENERAL METHODOLOGY	REMARKS	DRAWBACKS
Jain, Murty, and Flynn (1999)	**Data Clustering: A Review**	Survey and overview of data clustering techniques and algorithms	The paper provides a comprehensive review of various clustering methods, including partitioning, hierarchical, density-based, grid-based, and model-based approaches.	Highly influential and widely cited. Offers a valuable resource for understanding different clustering techniques used.	The paper does not delve into specific applications or comparative evaluations of the methods discussed.
Xu and Wunsch (2005)	**Clustering in Data Mining: A Survey**	Survey of clustering techniques in data mining	The paper covers a broad range of clustering methods, including partitioning, hierarchical, density-based, grid-based, model-based, and others. It also discusses evaluation measures and clustering in specific domains.	Provides a comprehensive overview of clustering techniques and their applications in data mining.	The survey does not extensively cover recent advancements and emerging clustering methods.
MacQueen (1967)	**Some Methods for Classification and Analysis of**	Introduces the k-means clustering algorithm	The paper presents the k-means algorithm for partitioning multivariate data into clusters based on similarity. It discusses	Pioneering work that introduced the k-means algorithm, which has become a	The k-means algorithm is sensitive to initial centroid selection and can converge to suboptimal

	Multivariate Observations		the algorithm's steps and provides a mathematical formulation.	fundamental and widely used clustering method.	solutions. It assumes equal-sized and spherical clusters.
Tan, Steinbach, and Kumar (2006)	**Introduction to Data Mining**	Textbook introducing data mining	The book covers a wide range of topics in data mining, including preprocessing, classification, clustering, association analysis, and more. It explains the concepts, techniques, and algorithms with examples and case studies.	Offers a comprehensive introduction to data mining, suitable for beginners and students. Includes practical examples and exercises.	The book focuses on providing an overview and introduction to various topics, so it may lack in-depth coverage and advanced techniques.
Cormac and Eleni (2017)	**Integration of Machine Learning Techniques to Evaluate Dynamic Customer Segmentation Analysis for Mobile Customers**	Mobile providers aiming to enhance customer satisfaction, loyalty, and retention rates.	Uses machine learning algorithms, specifically the C.5 algorithm within naive Bayesian modeling, to segment customers based on billing and socio-demographic aspects, and applies decision tree rules models for churn prediction.	Highlights the significance of customer relationship management and data mining techniques, such as RFM analysis and customer demographic profiles, in telecom churn prediction.	Further research required to assess scalability and practical implementation challenges in real-world telecom scenarios.
K.R.Kashwan and C.M.Velu (2013)	**Customer Segmentation Using Clustering and Data Mining Techniques**	Aims to segment customers based on sales data records and evaluate the accuracy of the model in predicting sales statistics.	The authors employed k-means clustering algorithm and ANOVA analysis to divide 2,138 customers into 4 groups, based on their similarities, and compared predicted sales statistics with actual sales data.	The model showed high accuracy in predicting sales and provided valuable insights for market forecasting and planning.	The paper lacks specific details on the dataset used and the specific techniques employed in the SPSS tool.
Chinedu and Simeon (2015)	**Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted**	To segment customers in a retail business based on their market characteristics using the K-Means	Developed a MATLAB program to apply the K-Means algorithm on a dataset of 100 training patterns, considering the average amount of goods purchased and the average number of	Customer segmentation allows businesses to provide targeted customer services and develop customized	The paper does not discuss the evaluation of the segmented clusters or provide a validation dataset for assessing the generalization

	Customer Services	clustering algorithm.	customer visits per month. Four customer clusters were identified with 95% accuracy.	marketing programs. The use of automated approaches like K-Means clustering is more efficient for large customer datasets compared to traditional market analyses.	performance of the approach. The study is limited to a specific retail business and dataset.
Tushar, Suraj ,Vishal and Tanupriya (2018)	**Customer Segmentation using K-means Clustering**	Determines customer segments based on similar behaviours and patterns, helping businesses make better decisions for targeting specific customer groups and improving marketing strategies.	The authors collect a dataset from a local retail shop, apply feature scaling, and then implement the k-means, agglomerative, and mean shift clustering algorithms to form clusters based on customer shopping behavior. They compare and analyze the results to identify distinct customer segments.	The study successfully segments customers into clusters, such as Careless, Careful, Standard, Target, Sensible, High buyers, and frequent/occasional visitors, providing valuable insights for businesses to customize marketing strategies and cater to different customer segments.	The paper does not discuss the evaluation metrics used to assess the effectiveness of the clustering algorithms, and it does not provide insights into the real-world application or validation of the identified customer segments.
Research work in this paper	**Customer Segmentation using K-Means Algorithm**	To analyze customer segments and evaluate the effectiveness of K-means clustering.	Applies K-means algorithm for behavioral and geographical segmentation of customers using Python libraries (numpy, pandas, matplotlib, seaborn, and scikit-learn) with visualization capabilities (Matplotlib, seaborn), evaluation metrics (elbow graph and silhouette analysis),	The findings help improve marketing strategies by targeting specific customer segments and demonstrate the effectiveness of machine learning in customer segmentation.	Challenges with data quality, sensitivity to initial parameters, interpretation difficulties, scalability limitations with larger datasets. Additionally, the assumption of spherical and equally sized clusters may not always hold.

			and interpretation of clustering results.		

FUTURE RESEARCH DIRECTIONS

In order to enhance the project, there are certain areas that can be focused on for future improvements. Alternative methods for determining the number of clusters, such as the gap statistic or hierarchical clustering, could be explored to obtain more accurate estimates. Additionally, incorporating additional relevant features into the clustering analysis can provide a more comprehensive understanding of customer behavior. Considering alternative clustering algorithms like DBSCAN, hierarchical clustering, or Gaussian mixture models can be beneficial for datasets that do not conform to the assumptions of K-means. Further, evaluating the clustering results using metrics like purity or completeness, and utilizing visualization techniques, can ensure the meaningfulness and usefulness of the results while also identifying any anomalies or outliers.

CONCLUSION

Most of the research work of the clustering algorithm was done using Python libraries such as numpy, pandas, matplotlib, seaborn, and scikit-learn which is advantageous for accurate data analysis. Moreover, powerful visualization capabilities with tools like Matplotlib and seaborn, and the ability to evaluate clustering quality with metrics like silhouette analysis.
However, there are also disadvantages to consider, including challenges related to data quality, sensitivity to initial parameters, interpretation difficulties, and scalability limitations with larger datasets. Being aware of these factors is crucial for ensuring accurate and reliable analysis results and making informed decisions based on the findings. Additionally, our code only considers two features (behavioral segmentation and geographical segmentation), which may not be sufficient to capture the complexity of customer behavior. It assumes that clusters are spherical and equally sized, which may not be true in all cases, especially when clusters have complex shapes or different sizes.
By leveraging the capabilities of the libraries mentioned, it becomes easier to analyze and interpret large datasets, identify patterns, and make data-driven decisions. Overall, this paper serves as a practical guide for anyone interested in applying clustering techniques to their data analysis tasks. It demonstrates the importance of selecting the right libraries, understanding their functionalities, and interpreting the results accurately.

REFERENCES

T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An efficient K-means clustering algorithm,” IEEE Trans. Pattern Analysis and Machine Intelligence,vol. 24, pp. 881-892, 2002.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1(14), 281-297.

Jain, A., Murty, M.N., and Flynn, P.J. (1999). "Data Clustering: A Review." ACM Computing Surveys, 31(3), 264-323.

Kishana R. Kashwan and C.M. Velu,” Customer Segmentation Using Clustering and Data Mining Techniques” in International Journal of Computer Theory and Engineering · January 2013.

Xu, R., and Wunsch, D. (2005). "Clustering in Data Mining: A Survey." IEEE Transactions on Knowledge and Data Engineering, 16(3), 303-316. doi: 10.1109/TKDE.2004.68

Tan, P.N., Steinbach, M., and Kumar, V. (2006). "Introduction to Data Mining." Addison-Wesley, Boston, MA.

Cormac Dullaghan and Eleni Rozaki,”Integration of Machine Learning Techniques to Evaluate Dynamic Customer Segmentation Analysis for Mobile Customer”, School of Computing, National College of Ireland, Dublin, Ireland , International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.7, No.1, January 2017

MacKay and David, “An Example Inference Task: Clustering,”Information Theory, Inference and Learning Algorithms, Cambridge University Press, pp. 284-292, 2003.

Tushar Kansal, Suraj Bahuguna ,Vishal Singh, Tanupriya Choudhury, “Customer Segmentation using K-means Clustering”, International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS),2018.

Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance kalu Electrical/Electronics & Computer Engineering Department, University of Uyo, Uyo, Akwa Ibom State, Nigeria “Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services”, IJARAI,Year: 2015.

S. Dasgupta and Y. Freund, “Random Trees for Vector Quantization,”IEEE Trans. on Information Theory, vol. 55, pp. 3229-3242, 2009.

M. Mahajan, P. Nimbhorkar, and K. Varadarajan, “The Planar K-Means Problem is NP-Hard,” LNCS, Springer, vol. 5431, pp. 274-285, 2009.

A.Vattani, “K-means exponential iterations even in the plane,”Discrete and Computational Geometry, vol. 45, no. 4, pp. 596-616, 2011.

ADDITIONAL READING

Jayant Tikmani, Sudhanshu Tiwari, Sujata Khedkar “Telecom customer segmentation based on cluster analysisAn Approach to Customer Classification using k-means”, IJIRCCE,Year: 2015.

M. Inaba, N. Katoh, and H. Imai, “Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering,” in Proc.10th ACM Symposium on Computational Geometry, 1994, pp.332-339.

D. Aloise, A. Deshpande, P. Hansen, and P. Popat, “NP-hard Euclidean sumof-squares clustering,” Machine Learning, vol. 75, pp. 245-249, 2009

SulekhaGoyat“The basis of market segmentation: a critical review of literature”, EJBM,Year: 2011.

Vaishali R. Patel and Rupa G. Mehta “Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm”, IJCSI,Year: 2011.

Tanupriya Choudhury, Vivek Kumar, Darshika Nigam, Intelligent Classification & Clustering Of Lung & Oral Cancer through Decision Tree & Genetic Algorithm, International Journal of Advanced Research in Computer Science and Software Engineering,2015.

Tanupriya Choudhury, Vivek Kumar, Darshika Nigam, An Innovative and Automatic Lung and Oral Cancer Classification Using Soft Computing Techniques, International Journal of Computer Science & Mobile Computing,2015.

Puwanenthiren Premkanth, ―Market Segmentation and Its Impact on Customer Satisfaction with Especial Reference to Commercial Bank of Ceylon PLC.‖ Global Journal of Management and Business Research Publisher: Global Journals Inc. (USA). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1.

B. Huang, M. T. Kechadi, and B. Buckley, “Customer churn prediction in telecommunications, “Expert Systems with Applications, vol. 39, no. 1, pp. 1414– 1425.

M. Farhadloo, R. A. Patterson, and E. Rolland, “Modelling customer satisfaction from unstructured data using a Bayesian approach,” Decision Support Systems, vol. 90, pp. 1–11.

L. Luan and H. Shu, “Integration of data mining techniques to evaluate promotion for mobile customers' data traffic in data plan,” 2016 13th International Conference on Service Systems and Service Management (ICSSSM), 2016. 82.

A. A. Khan, S. Jamwal, and M. Sepehri, “Applying Data Mining to Customer Churn Prediction in an Internet Service Provider,” International Journal of Computer Applications, vol. 9, no. 7, pp. 8–14.

Y. Huang and T. Kechadi, “An effective hybrid learning system for telecommunication churn prediction,” Expert Systems with Applications, vol. 40, no. 14, pp. 5635–5647.

C. Elkan, “Using the triangle inequality to accelerate K-means,” in Proc. the 12th International Conference on Machine Learning (ICML), 2003.

H. Zha, C. Ding, M. Gu, X. He, and H. D. Simon, “Spectral Relaxation for K-means Clustering,” Neural Information Processing Systems, Vancouver, Canada, vol.14, pp. 1057-1064, 2001

KEY TERMS AND DEFINITIONS

Customer Segmentation : It is the process of dividing a customer base into distinct groups based on shared characteristics, behaviors, or preferences.

Unsupervised Machine Learning: It is a type of machine learning where the model learns patterns and relationships in data without being explicitly trained on labeled examples. It discovers hidden structures or groups within the data without prior knowledge or guidance.

Clustering: It is a technique used in unsupervised learning to group similar data points together based on their features or attributes. The goal is to maximize similarity within each cluster and maximize dissimilarity between different clusters.

K-Means Algorithm: It is a popular clustering algorithm that aims to partition a given dataset into k clusters. It iteratively assigns data points to the nearest cluster centroid and updates the centroids based on the mean of the assigned points until convergence is achieved.

Elbow Method: Elbow method is used for finding optimal value of K for K-means clustering algorithm.

Silhouette Score: Silhouette score is used to calculate the quantitative measure of the quality of the clusters.

Behavioral Segmentation: It is a type of customer segmentation based on buying habits, usage patterns, loyalty, and engagement.

Geographical Segmentation: It is a segmentation approach that divides customers based on their geographic location or physical proximity.