**MAJOR PROJECT REPORT ON CUSTOMER SEGMENTATION USING UNSUPERVISED MACHINE LEARNING**

Submitted in the partial fulfilment of the requirements for the award of the degree of

**“MASTER OF BUSINESS ADMINISTRATION”**

**in**

**BUSINESS ANALYTICS**

**Submitted by**

**RAGHAVAPURAM VIRAT CHARY – 19MBMB30**

**MBA-BA (2019-2021)**

**SCHOOL OF MANAGEMENT STUDIES**

**UNIVERSITY OF HYDERABAD**

**CERTIFICATE**

This is to certify that the Summer Project report on **“CUSTOMER SEGMENTATION USING UNSUPERVISED MACHINE LEARNING”** is a record of bona fide work carried out by **RAGHAVAPURAM VIRAT CHARY – 19MBMB30** as a partial fulfilment of the requirement of MBA (Business Analytics) at School of Management Studies, University of Hyderabad. The report has been prepared under our guidance and is a record of the bona fide work carried out successfully.

**……………………**

(Dr. Suresh Kandulapati)

Guest Lecturer.

University of Hyderabad.

**DECLARATION**

I, **RAGHAVAPURAM VIRAT CHARY – 19MBMB30** hereby solemnly declare that the project titled **“CUSTOMER SEGMENTATION USING UNSUPERVISED MACHINE LEARNING”** submitted to the School of Management Studies, the University of Hyderabad under the guidance of Dr. Suresh Kandulapati, Guest Lecturer, University of Hyderabad is an original work carried out by me.

The work I have presented in the Project report does not breach any existing copyright and no portion of this report is copied from any work done earlier for a degree or otherwise.

****

Raghavapuram Virat Chary – 19MBMB30

**Programme:**

Master of Business Administration

Business Analytics (2019 - 2021)

School of Management Studies.

University of Hyderabad.

**Acknowledgement**

The Success and outcome of this project requires guidance and assistance from few people.

I offer my sincere respect and deep gratitude to our mentor Dr. Suresh Kandulapati for giving all the support and guidance which made me to complete the project duly. I’m extremely thankful to him for providing this opportunity and such great support and guidance during these difficult times.

I also thank my friends and classmates for imparting their valuable suggestions and feedback without which this project couldn’t be accomplished.

I’m also thankful to and fortunate enough to get constant encouragement, support and guidance from all the lecturers and the Dean, School of Management Studies.

Raghavapuram Virat Chary - 19MBMB30

**Programme:**

Master of Business Administration

Business Analytics (2019 - 2021)

School of Management Studies.

University of Hyderabad.

**CONTENTS PAGE NO.**

***ABSTRACT*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (1)**

1. **INTRODUCTION\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (2)**
2. **STATEMENT OF THE PROBLEM \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (3)**
3. **OBJECTIVES OF THE STUDY \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (3)**
4. **METHODOLOGY \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (3)**
   1. **Environment and Libraries used \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (4)**
5. **DATA PREPOSSESSING \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (4 - 7)**
6. **CREATING A MODEL \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (8)**

**6.1 Clustering for segmentation purposes \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (8)**

**6.2 K-means clustering \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (9)**

**6.2.1 K-Means Clustering Flowchart \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (9)**

**6.2.2 Elbow Method \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (10 - 12)**

**6.3 Hierarchical clustering \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (12 - 13)**

**7. ANALYSING THE RESULT \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (14)**

**7.1 Few Takeaways from this project \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (14)**

**8. DISCUSSION AND CONCLUSION \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (15)**

**9. REFERENCES \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (16)**

***ANNEXURE-I \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* (17)**

**INTERNSHIP CERTIFICATE FROM PEACOCK SOLAR \_\_\_\_\_\_\_\_ (18)**

**CUSTOMER SEGMENTATION USING UNSUPERVISED MACHINE LEARNING**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

***ABSTRACT***

The emergence of many business competitors has engendered severe rivalries among competing businesses in gaining new customers and retaining old ones. Due to the preceding, the need for exceptional customer services becomes pertinent, notwithstanding the size of the business. Furthermore, the ability of any business to understand each of its customers’ needs will earn it greater leverage in providing targeted customer services and developing customised marketing programs for the customers. This understanding can be possible through systematic customer segmentation. Each segment comprises customers who share similar market characteristics.

Maintaining customer relationships is a key to business success in today’s competitive environment. But all markets contain many subgroups of customers that behave differently, have different hopes, fears and ambitions, and have different purchasing behaviours. So, each subgroup must be behaved differently in order to build these relationships. On the road to this goal, customer segmentation is the first step. The goal of a segmentation system is to identify groups in which the customers are as much alike as possible and greatly differentiated from customers in other segments. If the segmentation system is well designed, members of a segment have similar interests, attitudes and behaviours, and they will respond similarly to elements of the marketing mix such as pricing, promotion and sales channel. Properly developed, segmentation insights inform a strategic roadmap intended to take advantage of key profit driving opportunities within each unique customer group. This could be shortening customer purchase cycles, driving higher spend, building greater customer loyalty, deepening cross-product penetration or lowering service and support costs

The ideas of Big data and machine learning have fuelled a terrific adoption of an automated approach to customer segmentation in preference to traditional market analyses that are often inefficient especially when the number of customers is too large.

The available clustering models for customer segmentation, in general, and the major models of K-Means and Hierarchical Clustering, in particular, are studied and the virtues and vices of the techniques are pointed out. Finally, the possibility of developing a hybrid solution by the combination of the above two techniques, having the ability to outperform the individual models, is discussed.

In this project, I worked on a real-world Mall customer dataset and segmented its customer base with Python. After several iterations, five stable clusters or customer segments were identified. The two features considered in the clustering are the “Annual Income (k$)” and “Spending Score (1-100)”.

1. **INTRODUCTION**

Over the years, increased competition among businesses and the availability of large-scale historical data has resulted in widespread use of data mining techniques to find critical and strategic information that is hidden in organizations' information.

Customer segmentation is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits.

Companies employing customer segmentation operate under the fact that every customer is different and that their marketing efforts would be better served if they target specific, smaller groups with messages that those consumers would find relevant and lead them to buy something. Companies also hope to gain a deeper understanding of their customers' preferences and needs with the idea of discovering what each segment finds most valuable to more accurately tailor marketing materials toward that segment.

Customer segmentation relies on identifying key differentiators that divide customers into groups that can be targeted. Information such as a customers' demographics (age, race, religion, gender, family size, ethnicity, income, education level), geography (where they live and work), psychographic (social class, lifestyle and personality characteristics) and behavioural (spending, consumption, usage and desired benefits) tendencies are taken into account when determining customer segmentation practices.

Benefits of customer segmentation include:

1. **Personalisation**
   * Personalisation ensures that you provide exceptional customer experience.
2. **Customer Retention**
   * It is 16 times as costly to build a long-term business relationship with a new customer than simply to cultivate the loyalty of an existing customer.
3. **Better ROI for marketing**
   * Affirmations that right marketing messages are sent to the right people based on their life cycle stage.
4. **Reveal new opportunities**
   * Customer segmentation may reveal new trends about products and it may even give the first mover’s advantage in a product segment.

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, Hierarchical clustering algorithm. 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. Clustering has got immense applications in pattern recognition, image analysis, bioinformatics and so on. In this paper, the k-Means clustering algorithm has been applied in customer segmentation.

1. **STATEMENT OF THE PROBLEM**

Let us consider that we own a mall and want to understand the customers like who can be easily converge [Target Customers] so that the sense can be given to marketing team and plan the strategy accordingly. Malls or shopping complexes are often indulged in the race to increase their customers and hence making huge profits. To achieve this task machine learning is being applied by many stores already.

It is amazing to realize the fact that how machine learning can aid in such ambitions. The shopping complexes make use of their customer’s data and develop Machine Learning models to target the right ones. This not only increases sales but also makes the complexes efficient.

1. **OBJECTIVES OF THE STUDY**

* The main objective of this study is to develop the machine learning clustering models in order to figure out the Target customers for the business.
* To reduce risk in deciding where, when, how, and to whom a product, service, or brand will be marketed.
* To increase marketing efficiency by directing effort specifically toward the designated segment in a manner consistent with that segment's characteristics.

The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business. Since the marketer’s goal is usually to maximize the value (revenue and/or profit) from each customer, it is critical to know in advance how any particular marketing action will influence the customer. Ideally, such “action-centric” customer segmentation will not focus on the short-term value of a marketing action, but rather the long-term customer lifetime value (CLV) impact that such a marketing action will have. Thus, it is necessary to group, or segment, customers according to their CLV.

1. **METHODOLOGY**

This project is a part of Mass customer segmentation data which is collected from a retail store. The data consists of 200 rows and 5 columns. This data contains the basic information (ID, age, gender, income, spending score) about the customers.

Here we have the following features:

1. Customer ID: It is the unique ID given to a customer

2. Gender: Gender of the customer

3. Age: The age of the customer

4. Annual Income(k$): It is the annual income of the customer

5. Spending Score: It is the score (out of 100) given to a customer by the mall authorities, based on the money spent and the behavior of the customer.

* 1. **Environment and Libraries used**

Here, I used Anaconda’s Jupyter Notebook to write the code and used the following libraries.

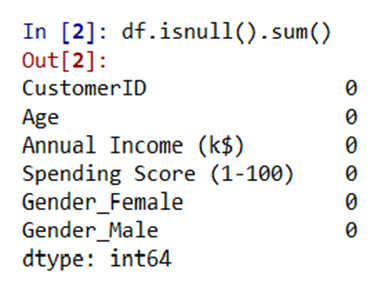
1. scikit-learn
2. seaborn
3. NumPy
4. pandas
5. matplotlib

In this study, we mainly focused on implementing K-means clustering algorithm to find out the optimum number of clusters by using the Elbow method. After several iterations, five stable clusters or customer segments were identified. The main two features considered in the clustering are the “Annual Income (k$)” and “Spending Score (1-100)”. After that, I also implemented the Hierarchical clustering algorithm to identify the optimum number of clusters using Dendrograms.

1. **DATA PRE-PROCESSING**

Before we proceed, we have to make sure that the data is in the correct format so as to be used for modelling later on. I started with loading all the libraries and dependencies. The columns in the dataset are Customer id, Gender, Age, Annual Income (k$) and Spending Score (1-100).

**Checking the null values:**

****

We have no null values in any of the columns. We see that we have only one categorical feature: “Gender” of the customer.

Let’s see the first five rows of the dataset.

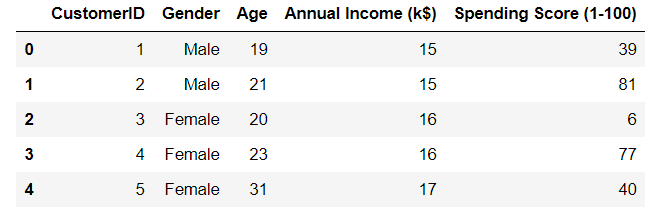


Table 1: Shows the first five rows of the dataset.

I dropped the Customer ID column as that does now seems relevant to the context. Also, I plotted the age frequency of customers.

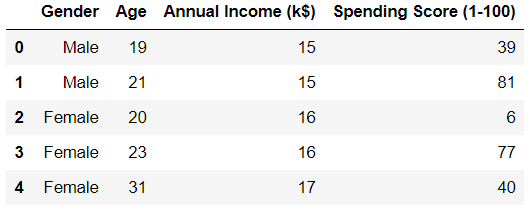


Table 2: Shows only the required columns of the dataset.

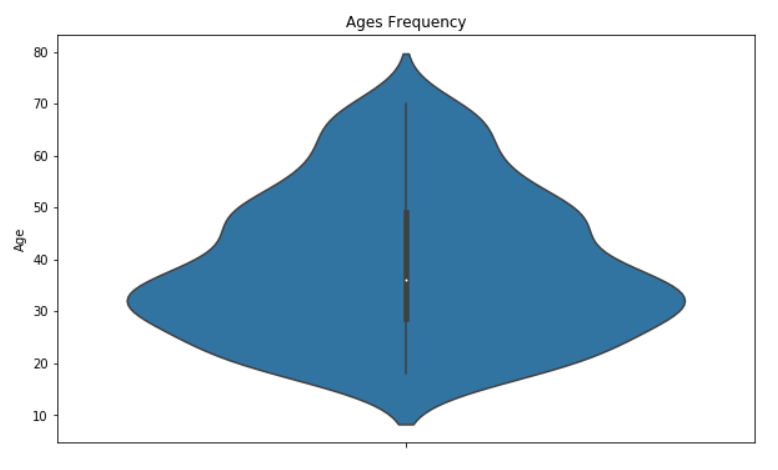


Figure 1: Shows the Age frequency of the customers.

I made a bar plot to check the distribution of male and female population in the dataset. The female population clearly outweighs the male counterpart.

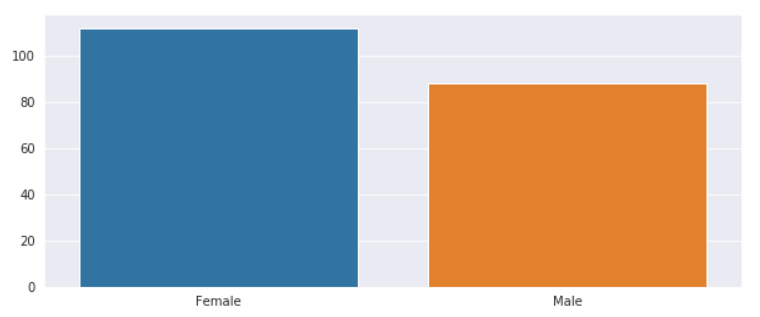


Figure 2: Shows the distribution of male and female population in the dataset.

Next, I made a bar plot to check the distribution of number of customers in each age group. Clearly the 26–35 age group outweighs every other age group.

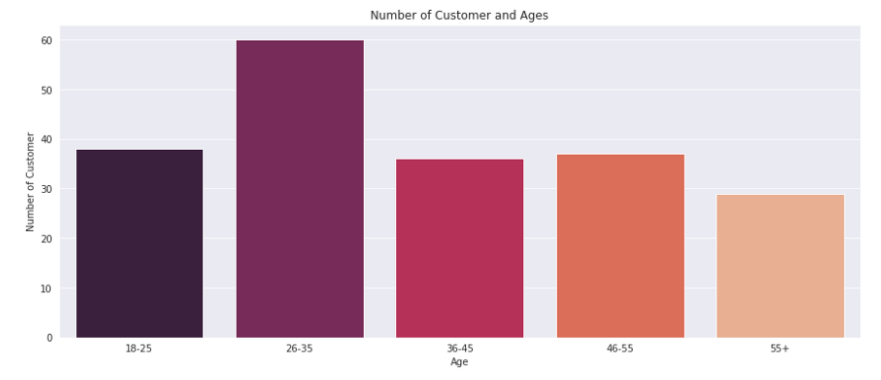


Figure 3: Shows the distribution of number of customers in each age group.

I continued with making a bar plot to visualize the number of customers according to their spending scores. The majority of the customers have spending score in the range 41–60.

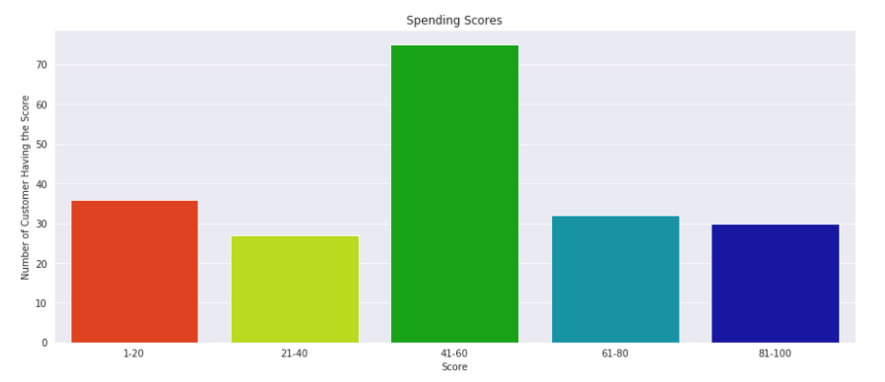


Figure 4: Shows the number of customers according to their spending scores.

Also, I made a bar plot to visualize the number of customers according to their annual income. The majority of the customers have annual income in the range 60000 and 90000.

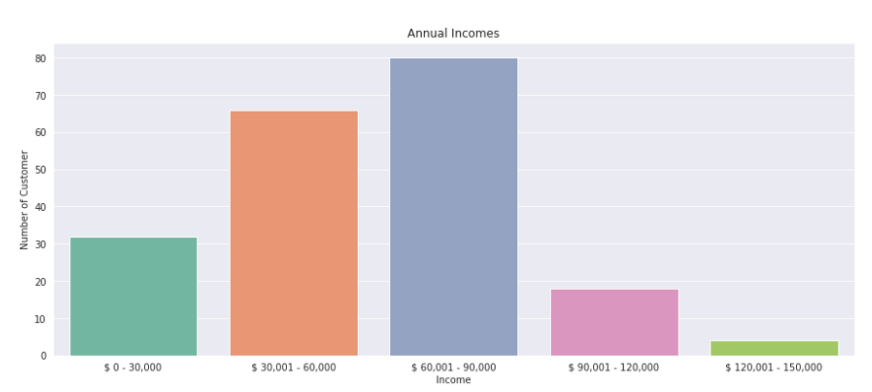


Figure 5: Shows the number of customers according to their annual income.

Once, I got the x, y, and z coordinates of all the five clusters that are created to visualize the centroids of the graph using a scatterplot. As I was comparing 3 attributes at the same time, I had to use a 3D plot.

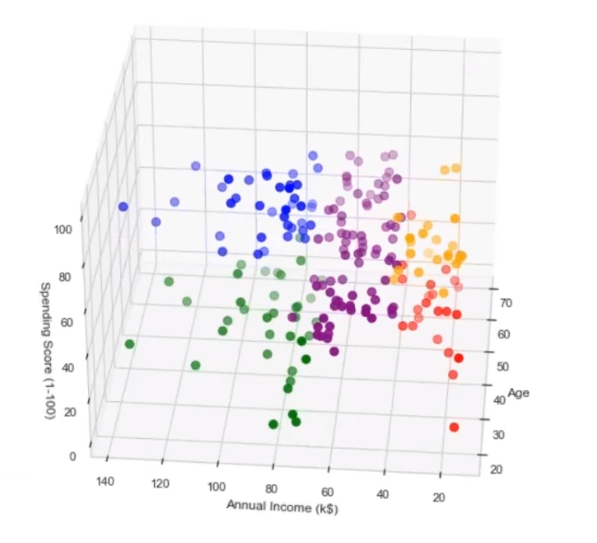


Figure 6: Shows the centroids of the ‘Age’, ‘Annual Income’, and ‘Spending Score’ using a scatterplot.

1. **CREATING A MODEL**

**6.1.** **Clustering for segmentation purposes**

Clustering techniques reveal internally homogeneous and externally heterogeneous groups. Customers vary in terms of behavior, needs, wants and characteristics and the main goal of clustering techniques is to identify different customer types and segment the customer base into clusters of similar profiles so that the process of target marketing can be executed more efficiently. Both, hierarchical and non-hierarchical clustering algorithms are widely used in customer segmentation, most prominent among them being K-Means and Agglomerative Hierarchical Clustering.

In K-Means has been used as part of the clustering approach. Also implemented K-Means for customer segmentation on the dataset. Although, hierarchical clustering algorithm seems unsuitable to many, we have used it for intelligent customer segmentation for our research and have made use of it for applying clustering algorithms on the transaction data from a supermarket. K-means and Hierarchical Clustering algorithms are useful for clustering data and find extensive usage in customer segmentation. Hence, they will be our main focus of interest.

**6.2. K-Means Clustering**

**6.2.1 K-Means Clustering Flowchart:**

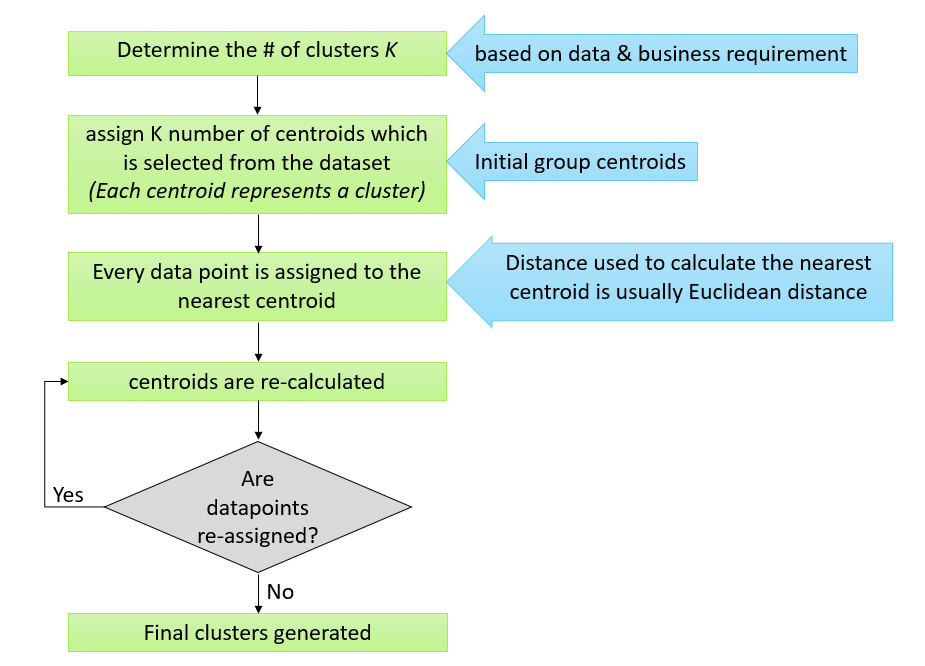
****

Figure 7: Flow chart of K-means Clustering

K-Means is one of the most widely used clustering algorithms, and is simple and efficient. The aim of K-means algorithm is to divide M points in N dimensions into K clusters (assume k centroids) fixed a priori. These centroids should be placed in a wise fashion so that the results are optimal which otherwise can differ if locations of the centroids change. So, they should be placed as far as possible from each other. Each data point is then taken and associated with the nearest centroid until no data points are pending. This way an early grouping is done and at this point, k new centroids have to be recalculated as these will be the centres of the clusters formed earlier. After having calculated these centroids, the data points are then allocated to the clusters to the nearest centroids. In this iteration, the centroids change their position stepwise until no further modifications have to be done and the location of the centroids remain intact. The K-Means algorithm is relatively simple. The ‘K’ cluster points, which will be the centroids, are placed in the space among the data points. Each data point is assigned to the centroid for which the distance is the least. After each data object has been assigned, centroids of the new groups are re-calculated. The above two steps are repeated until the movement of the centroid ceases. This means that the objective function of having the least squared error is completed and it cannot be improved further. Hence, we get K clusters as a result.

K-means algorithm aims at minimizing an objective function, which here, is the squared-error. It is an indicator of the distance of the data points from their respective cluster centres. The process in this algorithm always terminates but the relevance or the optimal configuration cannot be guaranteed even when the condition on the objective function is met. The algorithm is also sensitive to the selection of the initial random cluster centres. That is why it runs multiple times to reduce this effect but for a large number of data points, it tends to perform very well even though it is iterative.

Here we apply K-Means Clustering algorithm on a relatively small dataset and the results are depicted. The dataset is based on customer information for a mall and has 5 attributes named Customer ID, Gender, Age, Annual Income (k$) and Spending Score (1-100). It consists of 200 observations, each of which refers to a unique customer and the spending scores are decided and calculated by the company, based on their spending habits. Hence annual income and spending scores are the key indicators in this data. The age attribute of the customers can also be experimented with, to analyze which age group works best for a business. Any business would always keep the monetary values of any customer as top indicators. Thus, the annual income and spending scores of the customers will be best suited for clustering. As K-means algorithm requires the number of clusters as input, below we will use the elbow method to get the optimal number of clusters which can be formed. It works on the principal that after a certain number of ‘K’ clusters, the difference in SSE (Sum of Squared Errors) starts to decrease and diminishes gradually. Here, the WCSS(Within-Cluster-Sum-of-Squared-errors) metric is used as an indicator of the same. Hence, the ‘K’ value, specifies the number of clusters. In Figure 8., it can be observed that an elbow point occurs at K=5. After K=5, the difference in WCSS is not so visible. Hence, we will choose to have 5 clusters and provide the same as input to the K-Means algorithm.

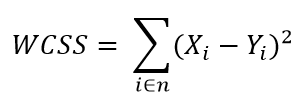
**6.2.2. ELBOW METHOD**

Calculate the Within Cluster Sum of Squared Errors (WCSS) for different values of k, and choose the k for which WCSS first starts to diminish. In the plot of WCSS-versus k, this is visible as an elbow.

The steps can be summarized in the below steps:

1. Compute K-Means clustering for different values of K by varying K from 1 to 10 clusters.
2. For each K, calculate the total within-cluster sum of square (WCSS).
3. Plot the curve of WCSS vs the number of clusters K.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

Next, I plotted Within Cluster Sum of Squares (WCSS) against the number of clusters (K Value) to figure out the optimal number of clusters value. WCSS measures sum of distances of observations from their cluster centroids which is given by the below formula.



where *Yi* is centroid for observation *Xi*. The main goal is to maximize number of clusters and in limiting case each data point becomes its own cluster centroid.

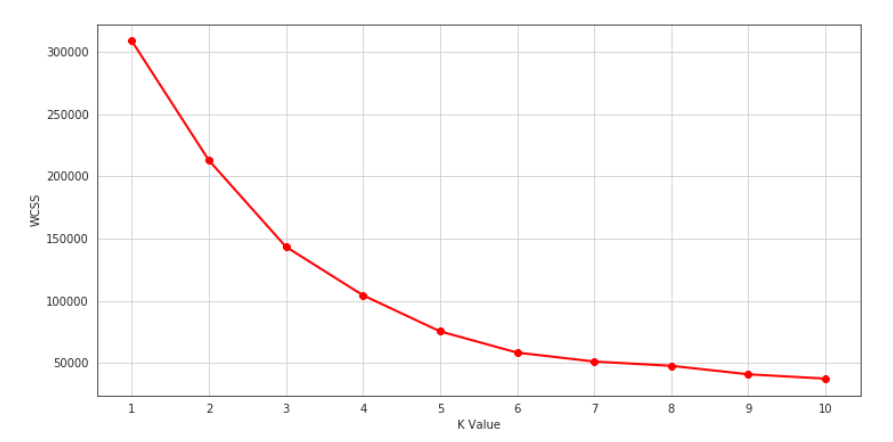
****

Figure 8. Finding the optimal number of clusters using Elbow method for K-means Clustering.

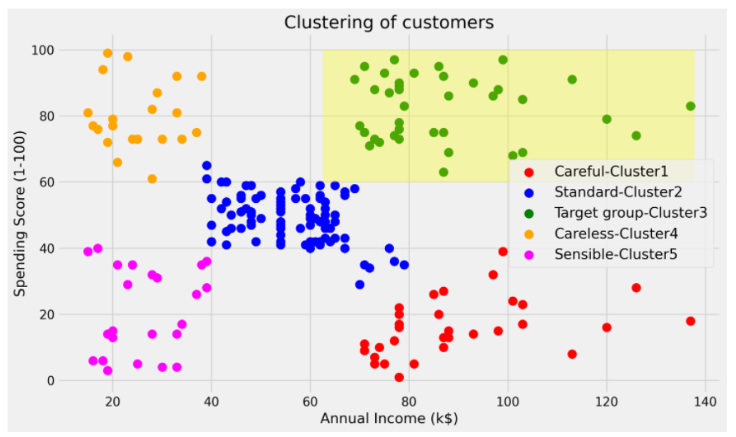


Figure 9. Clusters formed as a result of applying K-Means Clustering on the dataset taken for study

As shown in Figure 9, the scatter plot of the clusters is created with Annual income plotted against X-axis and Spending Score against Y-axis. The data points under each cluster are represented using distinct colours and the centroids are also highlighted, as shown above.

**6.3 Hierarchical Clustering**

Hierarchical clustering is a method of cluster analysis which builds a hierarchy of data points as they move into a cluster or out of it.

**Basically, there are two types of hierarchical cluster analysis strategies –**

1. **Agglomerative Clustering:** Also known as bottom-up approach or hierarchical agglomerative clustering (HAC). A structure that is more informative than the unstructured set of clusters returned by flat clustering. This clustering algorithm does not require us to prespecify the number of clusters. Bottom-up algorithms treat each data as a singleton cluster at the outset and then successively agglomerates pairs of clusters until all clusters have been merged into a single cluster that contains all data.
2. **Divisive Clustering:** Also known as top-down approach. This algorithm also does not require to prespecify the number of clusters. Top-down clustering requires a method for splitting a cluster that contains the whole data and proceeds by splitting clusters recursively until individual data have been splitted into singleton cluster.

The algorithm for the agglomerative hierarchical clustering approach proceeds by taking each observation in a cluster of its own. A pair of clusters with the shortest distance between them is chosen. The above two clusters are replaced with a new cluster by merging the original clusters in the previous step. Previous two steps are repeated until only one cluster remains and that cluster will contain all the observations. One major advantage of Hierarchical clustering is that we do not need to know the exact number of clusters beforehand and we can choose the formation of the clusters as they merge. The dataset containing 200 mall customers and their information is taken, same as used for the K-means algorithm, and agglomerative hierarchical clustering is applied. We can visualize the cluster formation in the form of a tree-diagram called a dendrogram, which allows us to illustrate the hierarchical organization of entities just like in Figure 10.

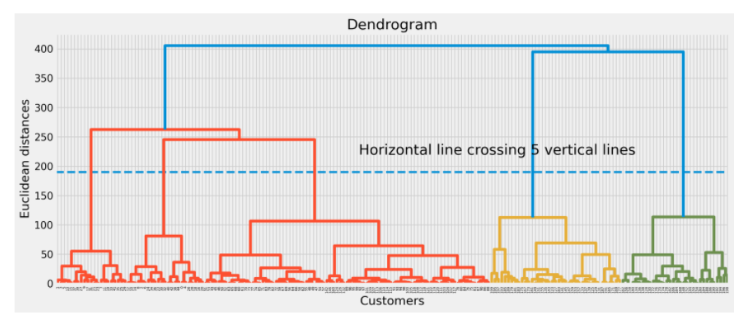


Figure 10. Visualization of the formation of clusters in the studied dataset with the help of a dendrogram.

We have the option of choosing the required number of clusters from the dendrogram itself by selecting the range of maximum distance and then placing a cut-off line at that position. It simply indicates that the distance between the formed clusters is maximum and distinction can be made among them. Hence, according to Figure 10, for satisfactory results, we can choose five clusters (K=5). The clusters are depicted as follows in Figure 11.

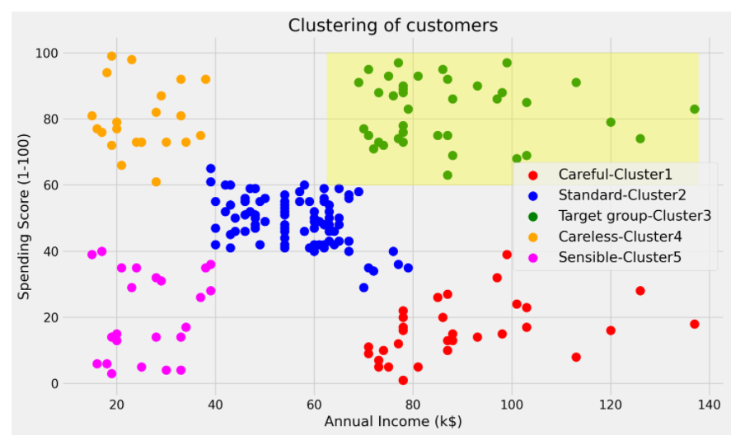


Figure 11. Clusters formed as a result of applying Hierarchical Clustering on the dataset taken for study.

Each color represents a different cluster and data points are plotted against Annual Income on the X-axis and the Spending Score on Y-axis. Hierarchical clustering has been extensively used for segmentation purposes due to its ability to produce results in a visual way [24]. It helps in determining the number of clusters for any analysis. It can be used for varied datasets like categorical, spatial and time series with numerical data set being the most common as it consists of data as just real numbers.

The main advantage of Hierarchical clustering is that the output is in the form of a hierarchy(dendrogram) which tells us exactly at which point the clusters merged or split. Hence it is easy to choose and decide on the number of clusters that we wish to take by looking at the dendrogram. However, for a large number of observations its computational speed is very low as compared to the non-hierarchical methods of clustering.

1. **ANALYSING THE RESULT**

We can see that the mall customers can be broadly grouped into 5 groups based on their purchases made in the mall.

In cluster 5(Pink coloured) we can see people have low annual income and low spending scores, this is quite reasonable as people having low salaries prefer to buy less, in fact, these are the wise people who know how to spend and save money. The shops/mall will be least interested in people belonging to this cluster.

In cluster 4(Orange coloured) we can see that people have low income but higher spending scores, these are those people who for some reason love to buy products more often even though they have a low income. Maybe it’s because these people are more than satisfied with the mall services. The shops/malls might not target these people that effectively but still will not lose them.

In cluster 2(Blue coloured) we see that people have average income and an average spending score, these people again will not be the prime targets of the shops or mall, but again they will be considered and other data analysis techniques may be used to increase their spending score.

In cluster 3(Green-coloured) we see that people have high income and high spending scores, this is the ideal case for the mall or shops as these people are the prime sources of profit. These people might be the regular customers of the mall and are convinced by the mall’s facilities.

In cluster 1(Red coloured) we see that people have high income but low spending scores, this is interesting. Maybe these are the people who are unsatisfied or unhappy by the mall’s services. These can be the prime targets of the mall, as they have the potential to spend money. So, the mall authorities will try to add new facilities so that they can attract these people and can meet their needs.

Finally, based on our machine learning technique we may deduce that to increase the profits of the mall, the mall authorities should target people belonging to cluster 1 and cluster 2 and should also maintain its standards to keep the people belonging to cluster 3 and cluster 4 happy and satisfied.

To conclude, I would like to say that it is amazing to see how machine learning can be used in businesses to enhance profit.

* 1. **FEW TAKEAWAYS FROM THIS PROJECT**
* Identified 5 customer clusters/segments that can be aid the sales & marketing teams on strategy accordingly.
* Range of spending score is more than the annual income range.
* Female population clearly outweigh their male counterpart.
* 26–35 age group outweighs every other age group.
* Majority of the customers have spending score in the range 41–60.
* Majority of the customers have annual income in the range 60000 and 90000.

1. **DISCUSSION AND CONCLUSION**

Due to increasing commercialization, consumer data is increasing exponentially. When dealing with this large magnitude of data, organizations need to make use of more efficient clustering algorithms for customer segmentation. These clustering models need to possess the capability to process this enormous data effectively. Each of the above discussed clustering algorithms come with their own set of merits and demerits. The computational speed of K-Means clustering algorithm is relatively better as compared to the hierarchical clustering algorithms as the latter require the calculation of the full proximity matrix after each iteration.

K-Means clustering gives better performance for a large number of observations while hierarchical clustering has the ability to handle fewer data points. The major hindrance produces itself in the form of selecting the numbers of clusters ‘K’ for the K-Means process, which have to be provided as an input to this non-hierarchical clustering algorithm. This limitation does not exist in the case of hierarchical clustering since it does not require any cluster centres as input. It depends on the user to choose the cluster groups as well as their number. Hierarchical clustering also gives better results as compared to K-Means when a random dataset is used. The output or results obtained when using hierarchical clustering are in the form of dendrograms but the output of K-Means consists of flat-structured clusters which may be difficult to analyze. As the value of k increases, the quality(accuracy) of hierarchical clustering improves when compared to K-Means clustering. As such, partitioning algorithms like K-Means are suitable for large datasets while hierarchical clustering algorithms are more suitable for small datasets.

Both K-Means and Hierarchical clustering have drawbacks that make them unsuitable when used individually. For business use, data visualization forms a major part of efficient data analysis and hierarchical clustering aids in doing so. Furthermore, when the performance aspect is taken into account, K-Means tends to deliver better results. With the vices and virtues of the two techniques pointed out, it comes to light that an amalgam of the best of these algorithms could outperform the individual models. In summary, different clustering algorithms, owing to their properties towards different kinds of data can be used in succession such that the advantages of these techniques could be harnessed in full. However, the selection process of these suitable techniques as well as their judicious implementation could require a considerable time investment in studying and processing the data along with an adequate understanding of the goals and requirements.

1. **REFERENCES**

**[**1] E. Ngai, L. Xiu and D. Chau, “Application of data mining techniques in customer relationship management: A literature review and classification”, Expert Systems with Applications, vol. 36, no. 2, pp. 2592-2602, 2009.

[2] J. Peppard, “Customer Relationship Management (CRM) in financial services”, European Management Journal, vol. 18, no. 3, pp. 312-327, 2000.

[3] A. Ansari and A. Riasi, “Taxonomy of marketing strategies using bank customers clustering”, International Journal of Business and Management, vol. 11, no. 7, pp. 106-119, 2016.

[4] M. Ghzanfari, et al., “Customer segmentation in clothing exports based on clustering algorithm”, Iranian Journal of Trade Studies, vol. 14, no. 56, pp. 59-86, 2010.

[5] C. Rygielski, J. Wang and D. Yen, “Data mining techniques for customer relationship management”, Technology in Society, vol. 24, no. 4, pp. 483-502, 2002.

[6] J. Lee and S. Park, “Intelligent profitable customers segmentation system based on business intelligence tools”, Expert Systems with Applications, vol. 29, no. 1, pp. 145-152, 2005.

[7] D. A. Kandeil, A. A. Saad and S. M. Youssef, “A two-phase clustering analysis for B2B customer segmentation”, in International Conference on Intelligent Networking and Collaborative Systems, Salerno, 2014, pp. 221-228.

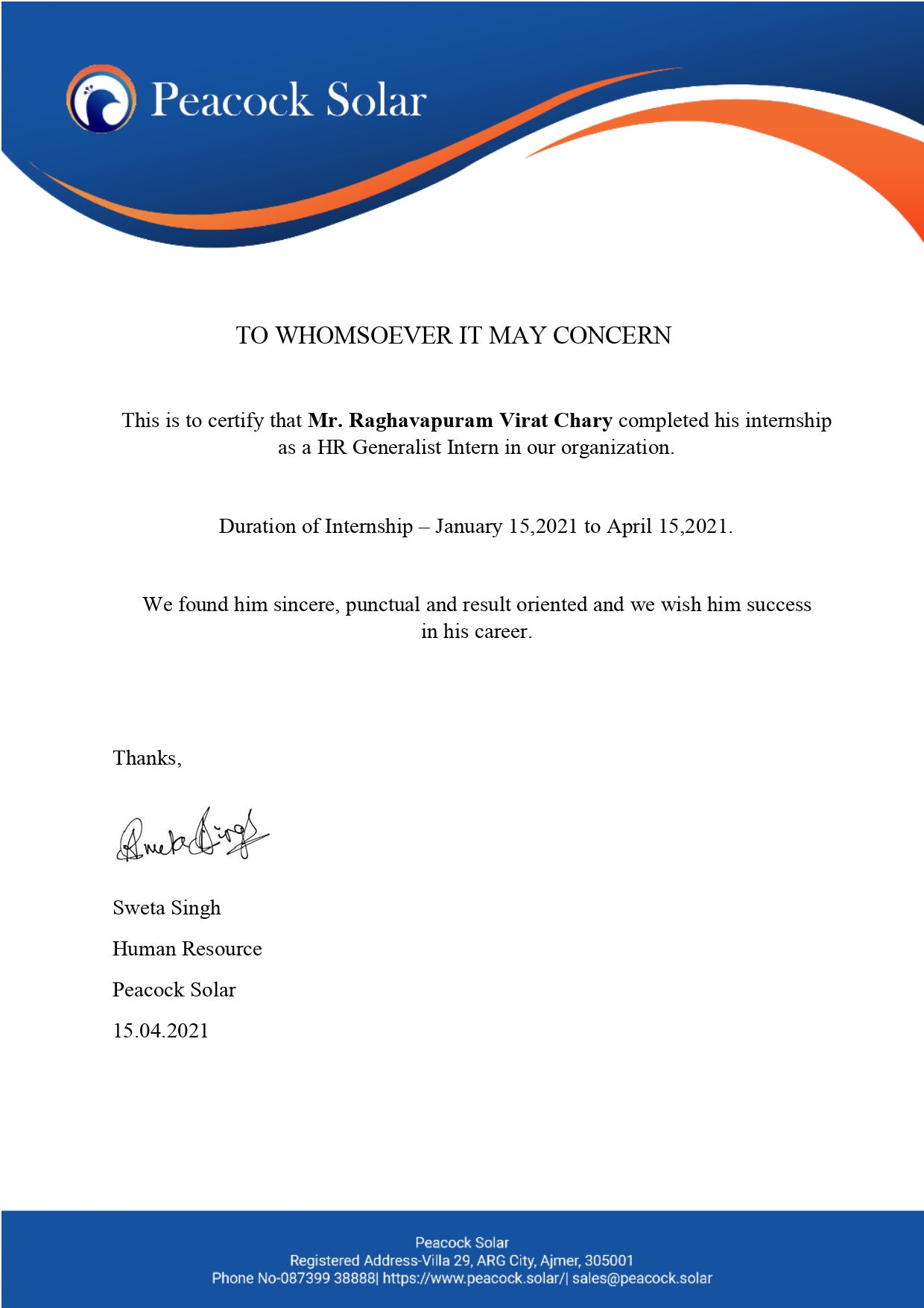
[8] R. Swift, Accelerating Customer Relationships: Using CRM and Relationship Technologies, 1st ed. Upper Saddle River, N.J.: Prentice Hall PTR, 2000.

[9] A. Parvatiyar and J. N. Sheth, “Customer relationship management: Emerging practice, process and discipline”, Journal of Economic and Social Research, vol. 3, no. 2, pp. 1-34, 2001.

[10] A. Riasi, “Barriers to international supply chain management in Iranian flower industry”, Management Science Letters, vol. 5, no. 4, pp. 363-368, 2015.

**ANNEXURE-I**

|  |  |
| --- | --- |
| **Figure/Table** | **Name of the Figure/Table** |
| **Table 1** | **Shows the first five rows of the dataset.** |
| **Table 2** | **Shows only the required columns of the dataset.** |
| **Figure 1** | **Shows the Age frequency of the customers.** |
| **Figure 2** | **Shows the distribution of male and female population in the dataset.** |
| **Figure 3** | **Shows the distribution of number of customers in each age group.** |
| **Figure 4** | **Shows the number of customers according to their spending scores.** |
| **Figure 5** | **Shows the number of customers according to their annual income.** |
| **Figure 6** | **Shows the centroids of the ‘Age’, ‘Annual Income’, and ‘Spending Score’ using a scatterplot.** |
| **Figure 7** | **Flow chart of K-means Clustering** |
| **Figure 8** | **Finding the optimal number of clusters using Elbow method for K-means Clustering.** |
| **Figure 9** | **Clusters formed as a result of applying K-Means Clustering on the dataset taken for study** |
| **Figure 10** | **Visualization of the formation of clusters in the studied dataset with the help of a dendrogram.** |
| **Figure 11** | **Clusters formed as a result of applying Hierarchical Clustering on the dataset taken for study.** |

****