SIDDAGANGA INSTITUTE OF TECHNOLOGY, TUMAKURU- 3

(An Autonomous institution affiliated to Visvesvaraya Technological University- Belagavi, Approved by AICTE,
Accredited by NAAC with 'A' Grade, Awarded Diamond College Rating by QS I-GAUGE & ISO 9001:2015 certified)

MINI PROJECT REPORT

ON

"CUSTOMER SEGMENTATION AND PRODUCT RECOMENDATION"

submitted in the partial fulfilment of the requirements for VI semester,
Bachelor of Engineering in Computer Science and Engineering

By

Rachaith 1SI18CS082

Abhishek KM 1SI19CS001

Faizan Qadri 1SI19CS040

Under the guidance of

Dr.Shobha K
Assistant Professor, Department of CSE

Department of Computer Science and Engineering

(Program Accredited by NBA)

Academic Year: 2021-22

Siddaganga Institute of Technology, Tumakuru-3

(An Autonomous institution affiliated to Visvesvaraya Technological University- Belagavi, Approved by AICTE, Accredited by NAAC with 'A' Grade, Awarded Diamond College Rating by QS I-GAUGE & ISO 9001:2015 certified)

Department of Computer Science and Engineering

(Program Accredited by NBA)

CERTIFICATE

This is to certify that the mini project entitled "Customer Segmentation and Product Recomendation" is a bonafide work carried out by **Rachaith (1SI18CS082), Abhishek K M (1SI19CS001) and Faizan Qadri (1SI19CS040)** of VI semester **Computer Science and Engineering**, **SIDDAGANGA INSTITUTE OF TECHNOLOGY** during the academic year 2021-2022.

Signature of the Guide
Dr.Shobha K
Assistant Professor

Signature of the Convener
Mrs. Thejaswini S M.Tech
Assistant Professor

Signature of the HOD

Dr. A S Poornima
Prof. and Head, Dept. of CSE

Name of the Examiners: **Signature with Date**

1. **Prof.**
2. **Prof.**

ACKNOWLEDGEMENT

The satisfaction euphoric that accompany the success of any work would be incomplete unless we mention the name of the people who made it possible, whose constant suggestions, guidance, encouragement helped us in completing our project work and served our effort with success.

We express my deep sense of gratitude and respectful Pranams at the lotus feet of **Dr. Sree Sree Sivakumara Swamigalu,** Revered Founder President and **Sree Sree Siddalinga Swamigalu,** Revered President, Siddaganga Education Society, for their blessings and for providing us with an opportunity to fulfil our most cherishable desires of reaching our goal.

We would like to express our profound gratitude to **Dr. M. N. Channabasappa**, Director, SIT, Tumakuru and **Dr. S V Dinesh,** Principal, SIT, Tumakuru, for giving all the facilities to pursue the Bachelor's degree successfully.

We would like to thank **Dr. A S Poornima,** Professor and Head, Department of CSE, SIT, Tumakuru for her kind consent and whole hearted co-operation.

We take this opportunity to express our sincere gratitude to our guide **Dr.Shobha K** Assistant Professor, Department of CSE, SIT, Tumakuru for her kind support, guidance and encouragement throughout the course of this work. It was a great privilege and honour to work and study under her guidance.

We sincerely thank all the faculty members of Computer Science and Engineering department for their support and help in enriching our knowledge. Also, we extend our gratitude to the non-teaching staff of the department.

Finally, we would like to present our deepest gratitude to our parents and friends for their moral support and their encouragement, which motivated towards successful completion of our project.

ABSTRACT

Management of customer relationship has always played a vital role to provide business intelligence to organizations to build, manage and develop valuable long term customer relationships. The importance of treating customers as an organizations main asset is increasing in value in present day and era. Organizations are investing rapidly in the development of customer acquisition, maintenance and development strategies.

Customer segmentation helps in figuring out the customers who vary in terms of preferences, expectations, desires and attributes. The main purpose of performing customer segmentation is to group consumers with similar interest so that the marketing team can converge in an effective marketing plan. Clustering is an iterative process of knowledge discovery from vast amounts of raw and unorganized data. Clustering is a type of exploratory data mining that is used in many applications, such as machine learning, classification and pattern recognition.

By using clustering techniques, customers with similar means are clustered together.Customer segmentation helps the marketing team to recognize and expose different customer segments that think differently and follow different purchasing strategies,then recommending items to the users of the individual clusters based on purchase history and similarity of ratings provided by other users who bought items to that of a particular customer.

A model based collaborative filtering technique is chosen in this work,as it helps in recommending products for a particular user by identifying patterns based on preferences from multiple user data.

Table of Contents

CHAPTER 1

INTRODUCTION

Over the years, the increasing competition between businesses and the availability of large-scale historical data has resulted in the extensive use of data mining techniques to discover important and strategic information that is hidden in the information of organizations (Lilien and Kotler, 2013,p.456) .

Data mining is the process of extracting logical information from a dataset and presenting it in a human-accessible way for decision support.Data mining techniques distinguish areas such as statistics, artificial intelligence, machine learning and data systems.

Data mining applications include but are not limited to bioinformatics, weather forecasting, fraud detection, financial analysis and customer segmentation (Nantel 2014, p. 97). The key to this paper is to identify customer segments in the commercial business using a data mining method. Customer division is the division of the customer base of the business into groups called customer segments such that each customer segment consists of customers who share similar market characteristics. These distinctions are based on factors that can directly or indirectly influence themarket or business such as product preferences or expectations, locations, behavior and so on (Wells,2015, b, p. 197).

The importance of customer segmentation includes, the ability of a business to customize market plans that will be appropriate for each segment of its customers; support for business decisions based on a risky environment such as debt relations with their customers; Identification of products related to individual components and how to manage demand and supply power; reveals the interdependence and interaction between consumers, between products, or between customers and products that the business may not be aware of; the ability to predict customer decline, and which customers are most likely to have problems and raise other market research questions and provide clues to finding solutions (Reynolds et al., 2002, p. 687).

Companies have to grow their profitability and company through time as a result of fierce competition in the business field to meet customer expectations and attract new clients depending on their desires. It's tough and time-consuming to identify and respond to each customer's needs (Mehrota and Wells, 2007, p. 50). This is owing to the fact that, among other things, clients have a diverse set of aims, interests, and preferences. Customer segmentation, as opposed to a "one-size-fits-all" strategy, divides customers into groups based on comparable characteristics or habits. Customer segmentation is a marketing strategy that divides a market into distinct, homogeneous groups. The data used in the customer segmentation strategy, which divides customers into categories, is based on a number of factors, including regional circumstances, economic patterns, and demographic trends, and behavioral patterns. A client segmentation technique can help a company's marketing resources be better utilized.(Wakefield and Baker , 2009, p. 515).

CHAPTER 2

LITERATURE SURVEY

[1]Title of the work: The Mall Under Attack.

Authors: Wakefield and Baker.

Description: There are essentially three factors that explain the mall's declining role. First, consumers are increasingly busy, have less time for shopping, and therefore reduce the frequency of their visits to the mall. Moreover, too many malls are alike, and customers will go to the shopping center that offers the most product and service variety. Finally, Wakefield and Baker emphasize the fact that fewer consumers are going to the mall because they "enjoy their shopping experience". These factors are driving mall managers to developstrategies to differentiate from the competition. Indeed, in a recent survey (IBM/Retail Council of Canada, 2009), most retailers are shown to base their strategies on special services to enhance customer loyalty.

[2]Title of the work: The Mall Under Attack.

Authors: Stone and Smith.

Description: In spite of the need for a well-grounded segmentation model of shopping malls, the number of empirical studies of shopping malls is very limited. Moreover, shopping behavior research frequently concentrates on individual stores, and not on the mall itself. Furthermore, most research in this area "dates back to the 1970s, 1980s and the early half of the 1990s" (Reynolds et al., 2002, p. 687). Therefore, there is a need for new segmentation analysis approaches applied to shopping centers. The central question in such analysis is that of what should be the base for the classification of shoppers. mid-fifties, segmentation techniques have focused on demographic variables; these studies gave a detailed taxonomy of customers ("who" is buying), but they could not explain why people shop . Since then, "market researchers have made forays into psychology" (Mehrota and Wells, 2007, p. 50).

[3]Title of the work: Psychographic and Demographic Variables.

Author: Wells and Nantel.

Description: Wells (2015, b, p. 197) gave a general definition of "psychographic" (variables) that will be adopted here: "psychographic research can be defined as qualitative research intended to place consumers on psychological—as distinguished from demographic dimensions".

The socio-demographic variables included in the research were the ones suggested by Nantel (2014, p. 97): age, mother tongue, sex, annual income, education level, number of children under eighteen at home and occupation.

[4]Title of the work: Customer classification.

Authors: Puwanenthiren Premkanth

Description: Over the years, the commercial world has become more competitive, as organizations such as these have to meet the needs and desires of their customers, attract new customers, and thus improve their businesses. The task of identifying and meeting the needs and requirements of every customer in the business is very difficult. This is because customers can vary according to their needs, wants, demographics, size and taste, features etc. As it is a bad practice to treat all customers equally in business. This challenge has adopted the concept of customer segmentation or market segmentation, where consumers are divided into subgroups or segments, where members of each subcategory exhibit similar market behaviors or characteristics. Accordingly, customer segmentation is the process of dividing the market into indigenous groups.

[5]Title of the work: Description of the segmentation study.

Authors: Lilien and Kotler.

Description: Lilien and Kotler (2013) proposed that a segmentation study can be described in terms of base variables and descriptive variables.

The base variables of the study (the ones that define the different groups) describe the activities performed by the customers in the mall. The scale used was adapted from Bloch et al. (2004). Seven items were selected from the 13 proposed by Bloch et al. (2004) to represent the activities available in the shopping center where the empirical data was collected. Customers were asked to answer to a series of “yes” or “no” questions related to activities performed in the shopping center during their visit. The activities proposed were: “going to the mall for the exercise”, “talking to other customers”, “going to the bank”, “browsing in the stores without buying”, “taking a snack”, “going to buy a product in a store”, or ‘”making an unplanned purchase”. To complete the study, visitors were also asked about the frequency of visits (seven-point scale from “very seldom” to “very often”), and the amount of their purchases.

CHAPTER 3

PROBLEM STATEMENT AND OBJECTIVES

3.1 PROBLEM STATEMENT

To segment the customers and to recommend the products based on the similarity pattern.

3.2 OBJECTIVES

- To achieve customer segmentation using machine learning algorithm (KMeans Clustering).
- Product recommendations i.e Recommend items to users based on purchase history and similarity of ratings provided by other users who bought items to that of a particular customer.

CHAPTER 4

SYSTEM DESIGN

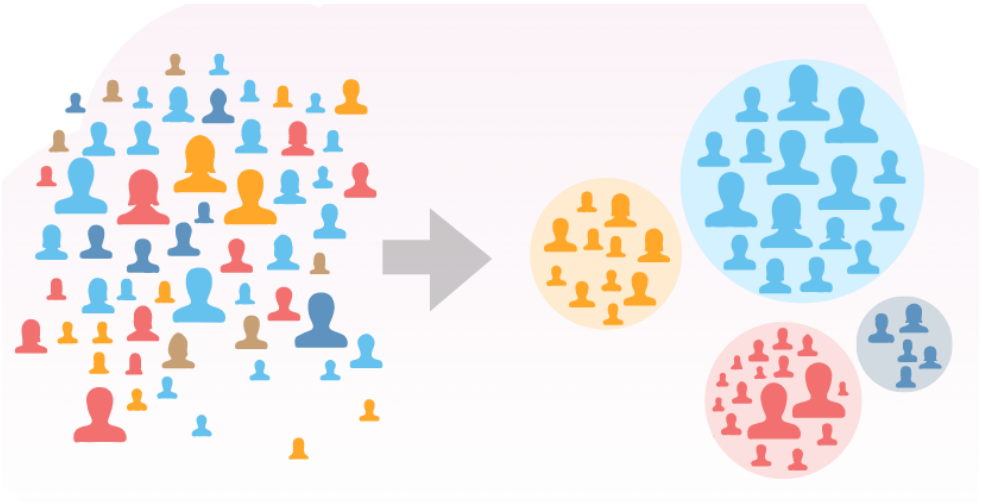

Figure.4.1 Customer segmentation.

In this work, the dataset provided by the mall is used for clustering using the K-means algorithm. Five attributes and 200 tuples make up the data set, which represents the information of 200 consumers. The attributes in the data set are CustomerId, gender, age, yearly income (k$), and spending score on a scale of (1-100),timestamp,productId and ratings. Figure 4.1 shows Customers have been divided and grouped into clusters using K -Means.

CHAPTER 5

HIGH LEVEL DESIGN

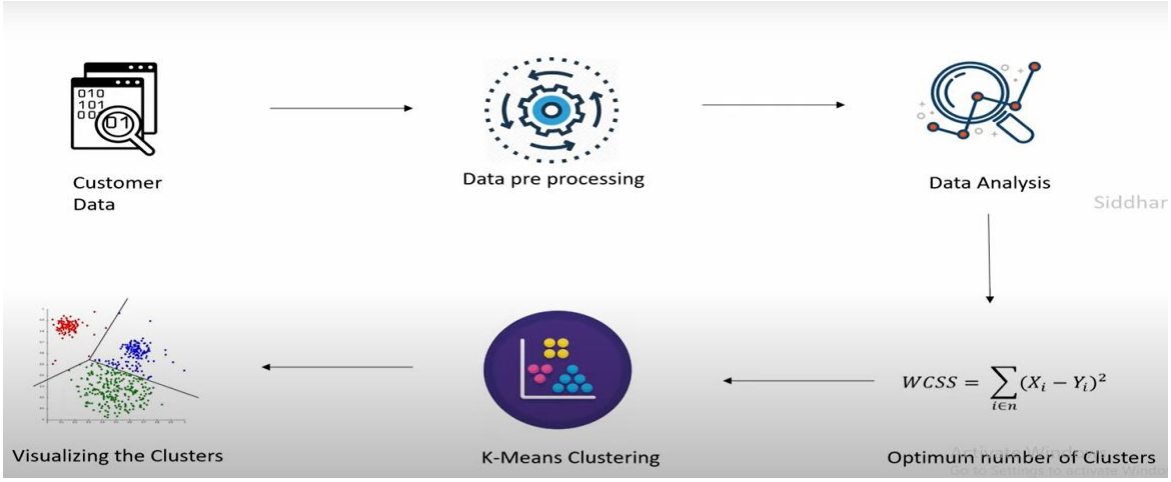

Figure 5 Steps of segmentation.

Collecting the customer data.

The attributes in the data collection are CustomerId, gender, age, yearly income (k$), and spending score on a scale of (1-100),timestamp,productId and ratings,it has 200 rows and 8 columns.

```
In [5]: df.head()
```

Out[5]:

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)	ProductId	Ratings	Timestamp
0	1	Male	19	15	39	B00008NJEP	1	1230249600
1	2	Male	21	15	81	132793040	4	1143072000
2	3	Female	20	16	6	321732944	5	1195516800
3	4	Female	23	16	77	439886341	4	1159833600
4	5	Female	31	17	40	439886341	2	1175731200

Figure 5.1 Dataset.

Figure 5.1 Reprsents dataset of Mall Customers,that can be downloaded from the kaggle website.

https://www.kaggle.com/code/shawamar/product-recommendation-system-for-e-commerce

https://www.kaggle.com/code/kushal1996/customer-segmentation-k-means-analysis

Data pre-processing.

Data preprocessing can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance and is an important step in the data mining process.
Data preprocessing is a process of preparing the raw data and making it usefull for further process.It is the first and crucial step while creating a machine learning model.

Before feeding the data to the k-means clustering algorithm, we need to pre-process the dataset. Implementing the necessary pre-processing for the customer dataset.

```
In [5]: df.columns = ['CustomerID','Gender','Age','Annual Income (k$)','Spending Score (1-100)','ProductId','Ratings','Timestamp']
        df.info()

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 200 entries, 0 to 199
        Data columns (total 8 columns):
         #   Column                  Non-Null Count  Dtype
        ---  ------                  --------------  -----
         0   CustomerID              200 non-null    int64
         1   Gender                  200 non-null    object
         2   Age                     200 non-null    int64
         3   Annual Income (k$)      200 non-null    int64
         4   Spending Score (1-100)  200 non-null    int64
         5   ProductId               200 non-null    object
         6   Ratings                 200 non-null    int64
         7   Timestamp               200 non-null    int64
        dtypes: int64(6), object(2)
        memory usage: 12.6+ KB
```

Figure 5.2 Dataset information.

Figure 5.2 Represents information about the dataset i.e which range of index, memory usage, datatypes etc.

```
In [9]: df.isnull().sum()

Out[9]: CustomerID              0
        Gender                  0
        Age                     0
        Annual Income (k$)      0
        Spending Score (1-100)  0
        ProductId               0
        Ratings                 0
        Timestamp               0
        dtype: int64
```

Figure 5.3 Checking for missing values.

Figure 5.3 shows dataset dosenot have any missing values, if the dataset has any missing values then a method called “Imputation” will be used to replace those missing values with suitable values.

Data analysis.

```
In [7]: df.describe()
```

Out[7]:

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)	Ratings	Timestamp
count	200.000000	200.000000	200.000000	200.000000	200.000000	2.000000e+02
mean	100.500000	38.850000	60.560000	50.200000	3.880000	1.290893e+09
std	57.879185	13.969007	26.264721	25.823522	1.361938	8.954232e+07
min	1.000000	18.000000	15.000000	1.000000	1.000000	1.116893e+09
25%	50.750000	28.750000	41.500000	34.750000	3.000000	1.207678e+09
50%	100.500000	36.000000	61.500000	50.000000	4.000000	1.311379e+09
75%	150.250000	49.000000	78.000000	73.000000	5.000000	1.373436e+09
max	200.000000	70.000000	137.000000	99.000000	5.000000	1.406074e+09

Figure 5.4 Describion the dataset.

If the Dataframe contains numerical data, Figure 5.4 describe() function gives the description of mall dataset information for each column:

count - The number of not-empty values.
mean - The average (mean) value.
std - The standard deviation.
min - the minimum value.
25% - The 25% percentile*.
50% - The 50% percentile*.
75% - The 75% percentile*.
max - the maximum value.

*Percentile meaning: how many of the values are less than the given percentile

Data is analysed by heatmap,barplot,pairplot etc , then we analyse the data of two particular columns (ex:- Spending score and annual income).

1.Heatmap

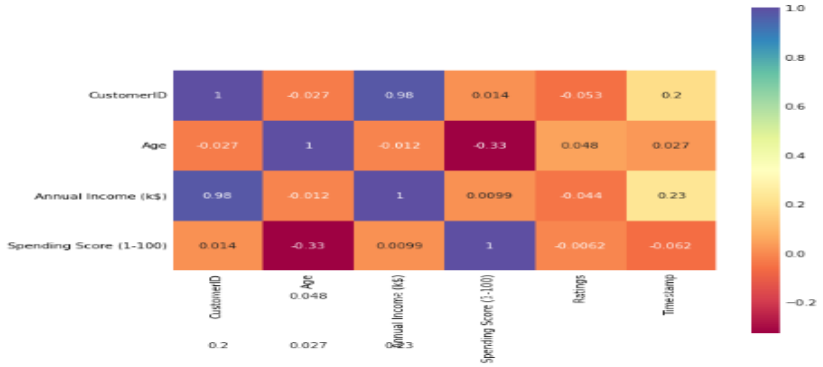

Figure 5.5 Heatmap

Correlation matrix is a very useful tool to analyze the relationship between features.
In Figure 5.4 It can be seen that there is a strong correlation between Spending score and Annual income,their correlation is 0.99.
That means If you have a higher salary, you are more likely to have a higher-level spending.

2.Barplot

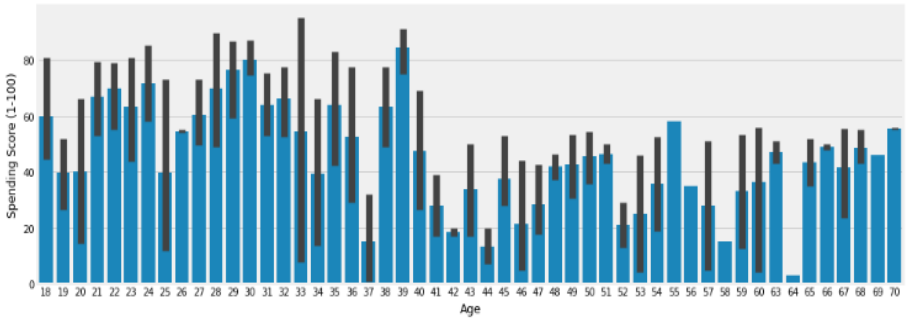

Figure 5.6 Barplot

Barplot is a graph that represents category of data within the rectangular boxes.
Figure 5.6 shows comparision between age and spending score.
In the figure 5.6, it can be visualized that age group of 20-40 has high spending score.

3.Pairplot

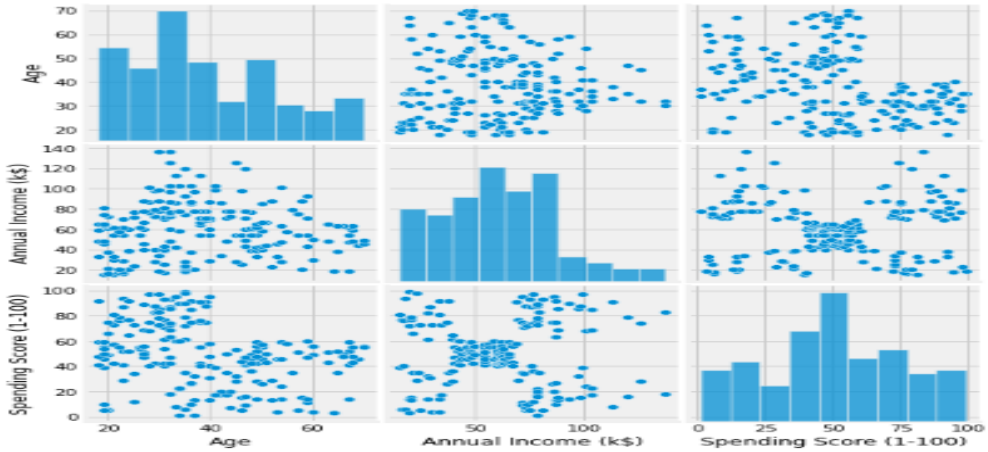

Figure 5.7 Pairplot

Pairplot is a matrix of reprsentation that let us understand the pairwise relationships between different variables in a dataset.

For the dataset considered in this work,following observations are made using pairplot.

- In the age column, most people belong to 20-50 age.
- In the annual income column, most people have 45k-50k.
- Maximum spending score is 50,that can be seen in spending score column.

Elbow Method.

In this method Within Cluster Sum of Squared Errors (WSS) for different values of "***k***" will be calculated and will be chosen, the "***k***" for which WSS first starts to diminish, hence name elbow method. In the plot of WSS-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.

Clusters	WCSS_3
2	180000
4	70000
6	40000
8	30000
10	20000

Table 5.1 shows different wcss values for different "***k***" values

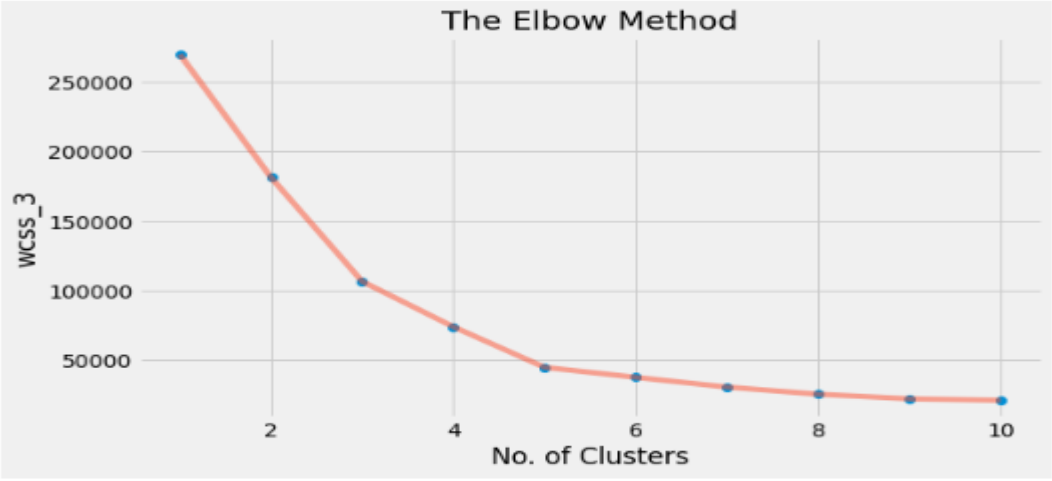

Figure 5.8 Elbow curve.

Figure 5.8 shows after 5 the drop is minimal, so we take 5 to be the number of clusters.

Validating the optimum number of clusters by silhouette analysis.

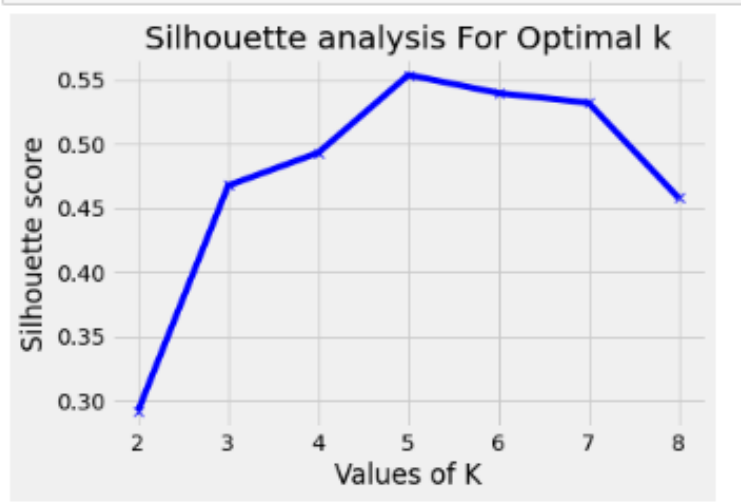

Figure 5.9 Silhoutte analysis

Value of K	Silhoutte score
2	0.29
3	0.47
4	0.50
5	0.55
6	0.54
7	0.53
8	0.46

Table 5.2 Silhoutte scores for different values of "***k***".

Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other.

In the figure 5.9 The Silhouette score is maximized at 5,Table 5.2 shows 0.55 is the value for 5,Hence the optimal number of clusters are 5.

K-means clustering algorithm.

Once, we have the optimum number of clusters, we can feed that data into K-means clustering algorithm, so that it can group the data depending on the similarities.

1. Pick the number of clusters for the dataset ("***k***")
2. Randomly select a point as the centroid of each cluster
3. Assign each data point to the nearest centroid (can use a measurement like the Euclidean distance)
4. Compute the centroid of the clusters again by finding a point in the cluster equidistant from all the data points
5. Once again, find the points nearest to the new centroids for each cluster
6. Repeat steps 3–5 until the position of the centroids doesn't change

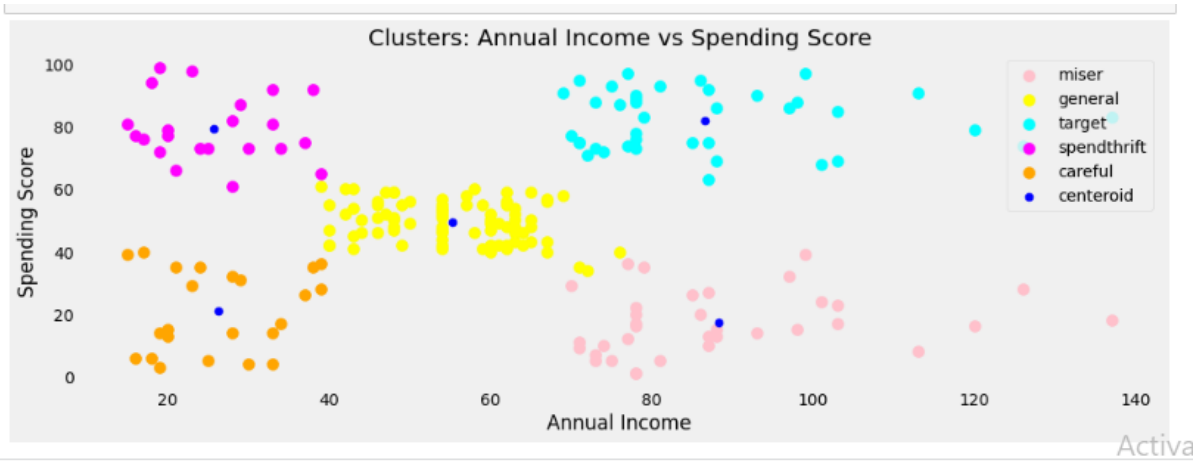

Figure 5.10 shows 6 different clusters have been formed from the data.

1. Miser:- Most Income and Less Spending Score.
2. General :- Average Income and Average Spending Score.
3. Target :- High Income with High Spending Score.
4. Carefull :- Less Income with Low Spending Score.
5. Spendthrift :- High Spending but Less Income.

Recommendation

Recommending items to the users of the individual clusters based on purchase history and similarity of ratings provided by other users.

Model based collaborative filtering is chosen,as it helps in making predicting products for a particular user by identifying patterns based on multiple user data.

Utility Matrix : An utlity matrix consists of all possible user-item preferences (ratings) details represented as a matrix. The utility matrix is sparce as none of the users would buy all teh items in the list, hence, most of the values are unknown.

```
In [38]: ratings_utility_matrix = data1.pivot_table(values='Ratings', index='CustomerID', columns='ProductId', fill_value=0)
         ratings_utility_matrix.head()
```

Out[38]:

ProductId	132793040	439886341	511189877	528881469	594481902	594511488	594514681	594514789	594549558	743610431	...	B0000DZEZ9	B0001LS0ZU
CustomerID													
2.0	4	0	0	0	0	0	0	0	0	0	...	0	0
6.0	0	4	0	0	0	0	0	0	0	0	...	0	0
8.0	0	0	5	0	0	0	0	0	0	0	...	0	0
10.0	0	0	5	0	0	0	0	0	0	0	...	0	0
12.0	0	0	5	0	0	0	0	0	0	0	...	0	0

5 rows × 26 columns

Figure 5.11 Utility matrix.

Figure 5.11 shows as expected, the utility matrix obtained above is sparce, it has been filled up with values 0.

Singular Value Decomposition

One of the popular algorithms to factorize a matrix is the singular value decomposition (SVD) algorithm.Singular value decomposition (SVD) is a matrix factorization method that generalizes the eigendecomposition of a square matrix (n x n) to any matrix (n x m).

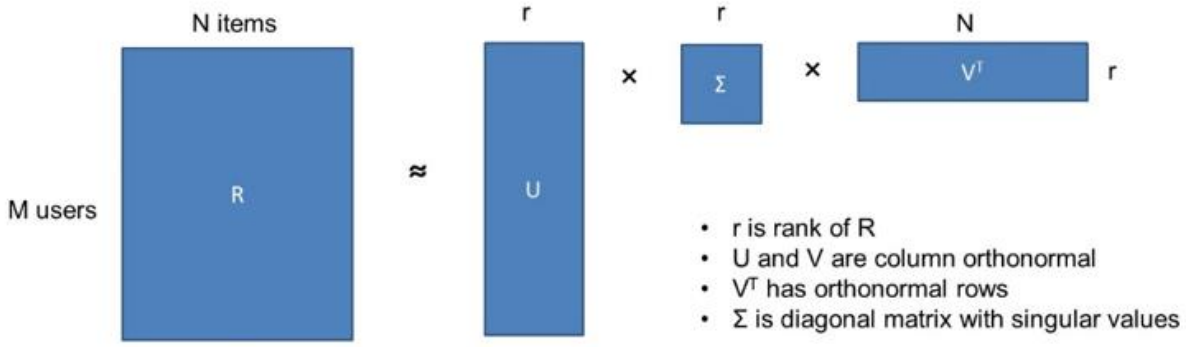

Figure 5.12 shows Singular value decomposition in recommended systems.

```
In [43]: from sklearn.decomposition import TruncatedSVD
         SVD = TruncatedSVD(n_components=10)
         decomposed_matrix = SVD.fit_transform(X)
         decomposed_matrix.shape

Out[43]: (26, 10)
```

Figure 5.13 Decomposing the matrix.

After decomposing the matrix,we will get correlation matrix,This matrix gives the correlation for all the items with the item purchased by the customer based on ratings by the other customers who bought the same product.

```
In [44]: correlation_matrix = np.corrcoef(decomposed_matrix)
         correlation_matrix.shape

Out[44]: (26, 26)
```

Figure 5.14 shows Correlating the decomposed matrix.

Isolating Product ID # B0001LSDUC from the Correlation Matrix

Assuming the customer buys Product ID # B0001LSDUC (randomly chosen)

```
X.index[19]
```

```
'B0001LSDUC'
```

```
i = "B0001LSDUC"

product_names = list(X.index)
product_ID = product_names.index(i)
product_ID
```

```
19
```

Figure 5.15 shows ProductID and CustomerID.

```
Recommend.remove(i)
```

Figure 5.17 shows removing the item that was already bought by the customer

```
Recommend = list(X.index[correlation_product_ID > 0.50])

# Removes the item already bought by the customer
Recommend.remove(i)

Recommend[0:5]
```

```
['594511488', '594549558', '743610431', 'B00005T3BD', 'B00008NJEP']
```

Figure 5.18 shows **top 5 highly correlated products in sequence**

CHAPTER 6

TOOLS AND TECHNOLOGIES

REQUIREMENTS

1. Jupyter Notebook by Anaconda-Anaconda is a conditional free and open-source distribution of the Python and R programming languages for scientific computing (data science ,machine learning applications, large-scale data processing ,predictive analytics ,etc.),that aims to simplify package management and deployment.

2. The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. Its uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

3. As a server-client application, the Jupyter Notebook App allows you to edit and run your notebooks via a web browser. The application can be executed on a PC without Internet access, or it can be installed on a remote server, where you can access it through the Internet. Its two main components are the kernels and a dashboard.

4. Python(3.6 32 bit) – To code Machine Learning algorithms.

5. IDE – VSCode/Atom.

6. Python Libraries are

    - Numpy - Used for working with arrays
    - Pandas - Data analysis and modifications of tabular data in dataframes.
    - Matplotlib - Used to create 2D graphs and plots.
    - Seaborn - Data visualization library.
    - Sklearn - Provides selection of efficient tools for machine learning and statistical modeling including classification,regression etc.

CHAPTER 7

IMPLEMENTATION

DATASET

It contains basic informations like :-
customer_id,age,gender,spending_score,annual_income,timestamp,product_id and ratings.

First five attributes have been used for segmentation and timestamp,product_id,ratings have been used for product recommendation.

K Means Algorithm

INPUT:

E = set of e data points
K = number of clusters
I = Iterations desired // this is necessary as full convergence is extremely costly.

OUTPUT:

C = set of c cluster centroids
L = set of distances, l from e to assigned centroid.

for c in C:

Randomly assign centroid c to be at some e.

for e in E:

Calculate distance from e to all centroids c.
Assign each e to centroid c with min. distance. Store in L.

```
i = 0.
minDistance = Inf

while i < I:
    for c in C:
        Compute the average location of all e assigned to cluster c.
        Reassign centroid c to new location.
    for e in E:
        Calculate distance from e to all centroids c.
    if  minDistance != l:
            Assign each e to centroid c with min. distance = L.
    else:
      end
    return assignments
```

```
from sklearn.cluster import KMeans
wcss_1 = []
for i in range(1, 11):
    km = KMeans(n_clusters = i, init = 'k-means++', max_iter = 300, n_init = 10, random_state = 0)
    km.fit(x1)
    wcss_1.append(km.inertia_)
```

```
kmeans_3=KMeans(n_clusters=5,init='k-means++',max_iter=300,n_init=10,random_state=0)
y_kmeans_3=kmeans_3.fit_predict(x3)
y_kmeans_3
labels_3 = kmeans_3.labels_
centroids_3 = kmeans_3.cluster_centers_
```

Figure 7.1 Implementation of K Means algorithm.

From figure 5.8 and 5.9, it can be inferred that 5 clusters have been got from elbow curve and silhouette,so K = 5.

In the figure 7.1, the algorithm is implemented with K = 5,then the dataset will be seperated into 5 different group of clusters.

The algorithm stops when the

- Centroids of newly formed clusters donot change.
- Points remain in the same cluster.
- Maximum number of iterations are reached.

CHAPTER 8

RESULTS

```
X.index[19]
```

```
'B0001LSDUC'
```

Table 8.1

Table 8.1 shows productID B0001LSDUC, that was already bought by the 19th customer.

```
Recommend = list(X.index[correlation_product_ID > 0.50])

# Removes the item already bought by the customer
Recommend.remove(i)

Recommend[0:5]
```

Figure 8.1 Recommendation

From the figure 8.1, Top 5 products to be displayed in the below table 8.2 by the recommendation system to the above 19th customer based on the purchase history of other customers in the website.

CustomerID	New ProductID
40	1327930404
	439886341
	594482127
	594514681
	777700018

Table 8.1 shows the Customer ID and top 5 prodcuts to that customer.

CHAPTER 9

SNAPSHOTS

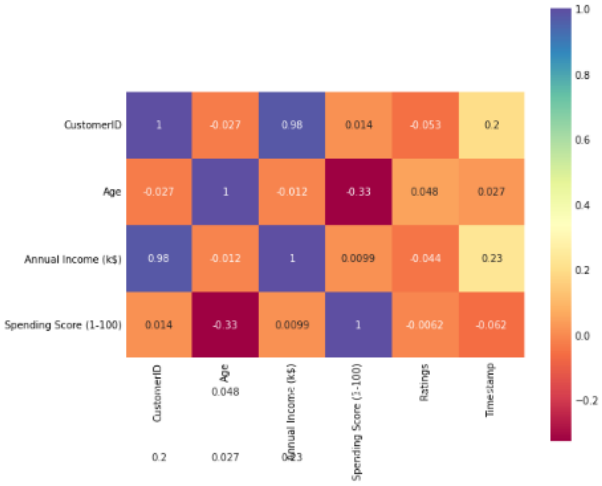

Figure 9.1 Heatmap

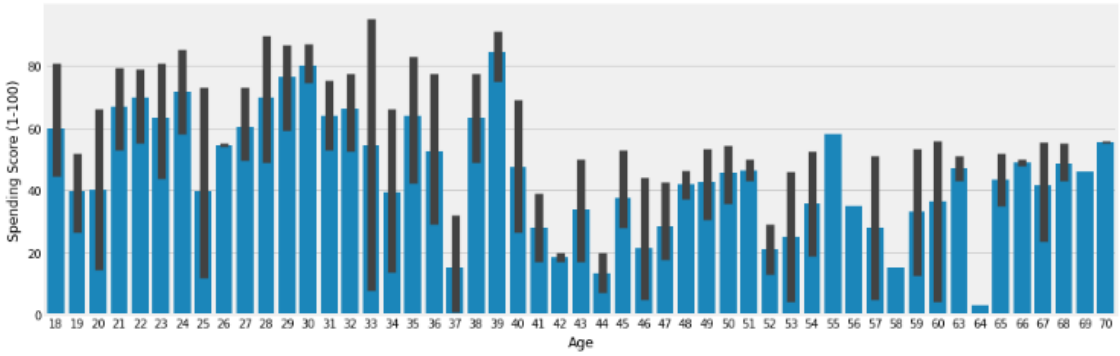

Figure 9.2 Barplot

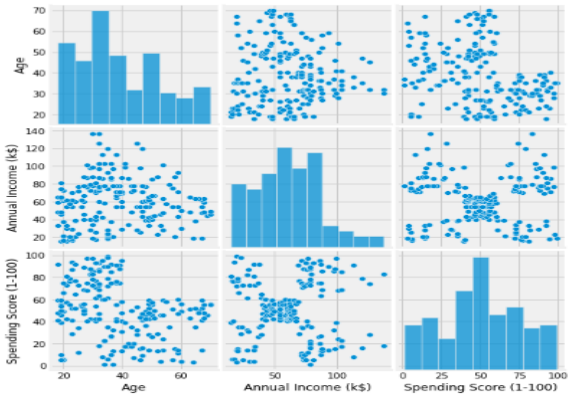

Figure 9.3 Pairplot.

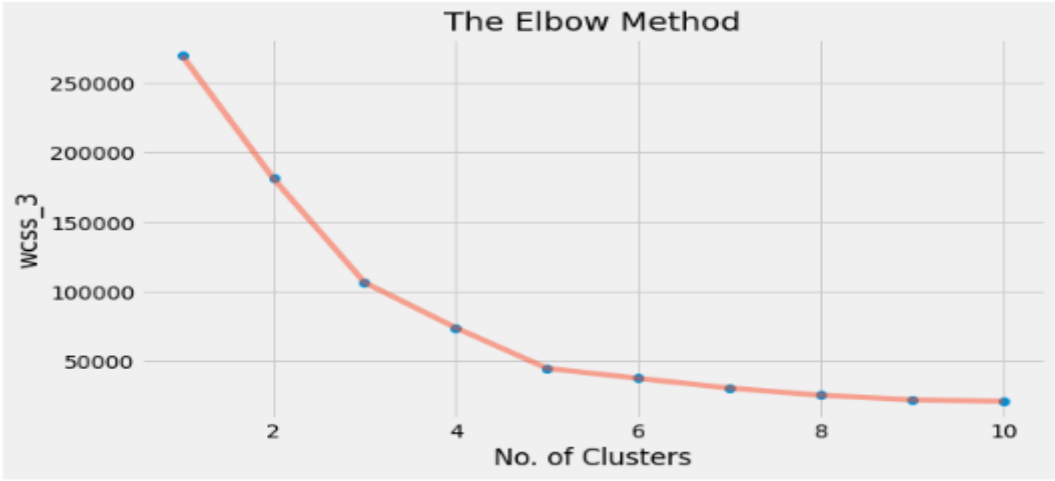

Figure 9.4 Elbow Curve.

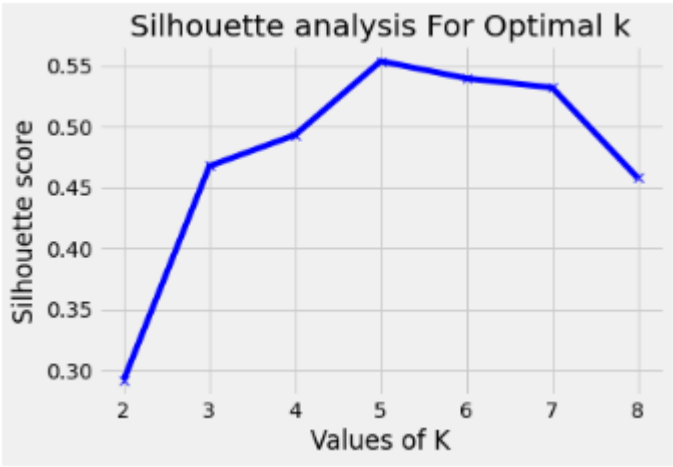

Figure 9.4 Silhoutte analysis.

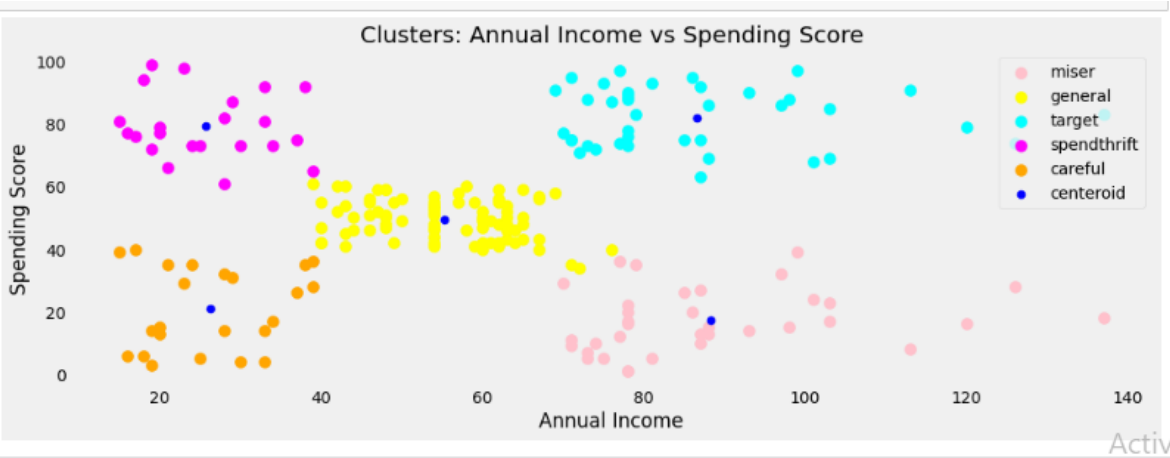

Figure 9.5 K means.

CHAPTER 10

CONCLUSION

This is a machine learning project, Where we have chosen mall customers dataset that dataset is pre-processed,analyzed using heatmap,barplot and pairplot,analyzed dataset is segmented into group of clusters using K-Means and the number of clusters are validated by silhouette method. Then data of each cluster is extracted and products have been recommend to the particular customer in that cluster based on their informations using model based collaborative filtering.

This study demonstrates that client segmentation in shopping malls is achievable despite the fact that this form of machine learning application is highly useful in the market, a manager can concentrate all of his or her attention on each cluster that has been discovered and meet all of their requirements.

Mall managers can be able to understand what customers require and, more importantly, how to meet those needs. analyze their purchasing habits, and establish frequent encounters with customers that make them feel comfortable in order to satisfy their demands.

A well developed recommendation system will help businesses improve their shopper's experience on website and result in better customer acquisition and retention.

REFERENCES

[1] Wakefield, K.L, Baker, J., 2009. Excitement at the mall: determinants and efects on shopping response. Journal of Retailing 74 (4), 515–539

[2] Stone, G.P., 2044. City shoppers and urban identification: observations on the social psychology of city life. American Journal of Sociology 60, 36–45
Smith, W.R., 2006. Product differentiation and market segmentation as alternative marketing strategies. Journal of Marketing (July), 3–8

[3] Wells, W.D., 2015a. Psychographics: a critical review. Journal of Marketing Research 12, 196–213
Nantel, J., 2014. La Segmentation. In: G. Morin (Ed.), Gestion du Marketing. Montreal, Que., Canada, pp. 85–119.

[4] Puwanenthiren Premkanth, - Market Classification and Its Impact on Customer Satisfaction and Special Reference to the Commercial Bank of Ceylon PLC.‖ Global Journal of Management and Business Publisher Research: Global Magazenals Inc. (USA). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1.

[5] Lilien, G.L, Kotler, P., 2013. Marketing Decision Making: A Model Building Approach. Harper & Row Publishers, New York.

[6] **https://www.kaggle.com/shawamar/product-recommendation-system-for-e-commerce**

[7] **https://www.enjoyalgorithms.com/blog/k-means-clustering-algorithm**

[8]https://scikitlearn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

[9] **https://www.analyticsvidhya.com/blog/2021/05/k-means-clustering-with-mall-customer-segmentation-data-full-detailed-code-and-explanation/**

[10] https://www.ijert.org/customer-segmentation-using-k-means-clustering

[11] https://brilliant.org/wiki/k-means-clustering/