

A Mini Project Report

“CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING”

Submitted in partial fulfillment of the course

CSE-1006: Foundation of Data Analytics

Under Guidance of Prof. Deepasikha Mishra

By

20BCE7227 Madhu Sreeja

22BCE9783 Farheen Naz



School of Computer Science & Engineering

VIT-AP UNIVERSITY, INAVOLU, AMARAVATI

INDEX

1. ABSTRACT

2. INTRODUCTION

3. PROBLEM STATEMENT AND OBJECTIVE

4. WORKING WITH DATA SET

5. EXTRACTING DATA

6. DATA CLEANING

7. DATA SORTING

8. PREDICTION/ANALYSIS USING ML TECHNIQUE

9. RESULTS

10. PLOTS

11. CONCLUSION

****Abstract: Unveiling Customer Patterns through K-means Clustering Analysis****

Customer segmentation is a critical aspect of data-driven business strategies, aiming to understand and categorize customers based on shared characteristics. This project employs the K-means clustering algorithm to segment a customer dataset, aiming to enhance targeted marketing efforts and improve overall customer satisfaction.

In the era of data-driven decision-making, businesses are increasingly turning to advanced analytics to understand their customer base. This project focuses on customer segmentation using the K-means clustering algorithm to uncover distinct patterns in a diverse customer dataset.

The project initiates with a comprehensive data extraction process, gathering relevant customer information from various sources. A meticulous data cleaning phase follows, addressing inconsistencies and outliers to ensure the quality of the dataset. Subsequently, the dataset is sorted and prepared for analysis.

The application of the K-means clustering algorithm reveals distinct customer segments based on shared attributes such as purchasing behaviour, demographics, and preferences. The optimal number of clusters is determined through iterative evaluation, ensuring the robustness of the segmentation results. The identified clusters are visualized using plots to provide a clear and accessible representation.

The project's findings contribute valuable insights into customer behaviour and preferences, enabling businesses to tailor marketing strategies and services. The segmented clusters serve as a foundation for targeted promotional campaigns, personalized customer interactions, and product recommendations. This approach not only enhances customer satisfaction but also maximizes the effectiveness of marketing efforts.

In conclusion, this project showcases the power of K-means clustering as an effective tool for customer segmentation, offering businesses a strategic advantage in understanding and catering to the diverse needs of their customer base. As industries continue to evolve, the insights derived from such analyses become imperative for staying competitive and responsive to ever-changing market dynamics.

****Introduction: Customer Segmentation Using K-means Clustering****

In the ever-evolving landscape of business and commerce, understanding and effectively engaging with customers are paramount for sustained success. Customer segmentation, the practice of categorizing customers based on shared characteristics, is a powerful strategy for tailoring marketing efforts, improving customer satisfaction, and optimizing business operations. This project delves into the application of the K-means clustering algorithm, a popular unsupervised machine learning technique, to perform customer segmentation.

**** Background: ****

In a marketplace characterized by diverse consumer behaviours and preferences, a one-size-fits-all approach to marketing and service provision often falls short of maximizing customer satisfaction and loyalty. Customer segmentation offers a solution by identifying distinct groups of customers who share common traits, allowing businesses to tailor their strategies to meet the specific needs of each segment.

**** Methodology: ****

The project begins with the extraction of relevant customer data from various sources. Subsequently, a meticulous data cleaning process is employed to ensure the accuracy and reliability of the dataset. The K-means clustering algorithm is then applied, a technique that partitions the customer data into distinct clusters, each characterized by its own set of defining features.

**** Significance: ****

The significance of this project lies in its potential to revolutionize how businesses interact with their customers. By identifying meaningful customer segments, businesses can tailor marketing campaigns, optimize product offerings, and enhance customer experiences, ultimately leading to improved customer satisfaction and increased profitability.

****Structure of the Report: ****

The report is organized into various sections, including data extraction and cleaning, application of the K-means clustering algorithm, presentation of results, and a comprehensive discussion of the implications for business strategies. Through this project, we aim to contribute insights and methodologies that can be applied across industries to enhance customer-centric approaches and drive business success.

**** Problem Statement: ****

In the era of data abundance, businesses face the challenge of extracting meaningful insights from vast amounts of customer data to optimize their marketing and operational strategies. Traditional, one-size-fits-all approaches are becoming less effective, and there is a growing need for businesses to adopt more personalized and targeted approaches to meet customer expectations. The project aims to address this challenge by implementing customer segmentation using K-means clustering to efficiently group customers based on shared characteristics, enabling businesses to tailor their offerings and strategies to specific customer segments.

**** Objective: ****

The "Customer Segmentation using K-means Clustering" project is designed to achieve the following objectives:

1. Segmentation: Utilize the K-means clustering algorithm to categorize customers into distinct segments. By grouping customers with similar behaviours and attributes, businesses can create targeted marketing campaigns and personalized experiences.
2. Feature Selection: Identify and analyse key features that contribute significantly to the clustering process. This involves understanding the relevance of various customer attributes such as purchasing frequency, recency, monetary value, and demographic information.
3. Optimization of K: Experiment with different values of K (number of clusters) to find the optimal configuration that best represents the underlying patterns in the data. Utilize metrics like the silhouette score to evaluate the quality of the segmentation.
4. Visualization: Develop visual representations of the customer segments to enhance interpretability. Graphical representations, such as scatter plots or heatmaps, can help businesses and stakeholders understand the relationships and distinctions among the identified clusters.
5. Customer Profiling: Create detailed profiles for each customer segment, highlighting their unique characteristics, preferences, and behaviours. This information is invaluable for crafting targeted marketing messages and tailoring products or services to meet specific segment needs.

6. Predictive Analysis: Explore the potential for predictive analysis within each segment. This involves forecasting future behaviours, such as future purchases or response to marketing campaigns, based on historical data and segment characteristics.

7. Implementation Strategy: Provide a roadmap for the integration of customer segmentation insights into business operations. This includes recommendations for adapting marketing strategies, refining product offerings, and enhancing customer service based on the identified segments.

By achieving these objectives, the project aims to equip businesses with the tools and insights needed to transition from a generic approach to a more personalized and customer-centric model, ultimately improving customer satisfaction, loyalty, and overall business performance.

**** WORKING WITH DATA SET: ****

1. Dataset Description:

The dataset includes the following features:

1. Customer ID
2. Customer Gender
3. Customer Age
4. Annual Income of the customer (in Thousand Dollars)
5. Spending score of the customer (based on customer behaviour and spending nature)

2. Data Cleaning and Preprocessing

3. Exploratory Data Analysis (EDA)

4. Data Transformation for Clustering (K-means):

Prior to applying K-means, numerical features were standardized to have a mean of 0 and a standard deviation of 1. This transformation ensured that all features contributed equally to the clustering process.

5. Data Splitting:

The dataset was split into a training set (80%) and a test set (20%) to evaluate the performance of the K-means model. The split was randomized, maintaining a representative distribution of subscription types across both sets.

6. Validation of Segments:

Segment validation involved cross-validating the clustering results with a holdout dataset. The stability of segments was confirmed by observing consistent patterns in customer behaviour across different time periods.

7. Interpretation of Results

EXTRACTING DATA:

Data Source:

The customer data used in this project was sourced from Kaggle's "Mall Customer Segmentation Data" dataset, available at Kaggle Dataset Link (<https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python>). This dataset contains information about customers in a mall, including their age, annual income, and spending score.

Data Collection Methods:

The dataset was downloaded from Kaggle as a CSV file. It was originally created for a customer segmentation tutorial and includes features relevant to understanding customer behaviour in a mall setting.

Timeframe:

The dataset represents a snapshot of customer information at a single point in time. The exact timeframe is not specified, but for the purpose of this project, the focus is on understanding current customer characteristics rather than changes over time.

Data Format:

The downloaded dataset is in CSV format and contains columns such as 'Customer ID,' 'Gender,' 'Age,' 'Annual Income,' and 'Spending Score.' These attributes are considered key for customer segmentation based on demographic and spending behaviour.

Ethical Considerations:

Since the dataset is publicly available on Kaggle and has been shared for educational purposes, there are no specific ethical concerns. However, it's crucial to respect Kaggle's terms of use and any licensing restrictions associated with the dataset.

Data Quality and Limitations:

The dataset appears to have good quality, with no obvious missing values or anomalies. However, it's important to conduct thorough exploratory data analysis (EDA) to identify potential outliers or peculiarities that might impact the segmentation analysis.

Data Pre-processing Steps:

Pre-processing steps included checking for missing values, exploring the distribution of key variables, and normalizing or scaling features if necessary. For instance, 'Annual Income' might need scaling to ensure equal weight with other features during clustering.

Data Extraction Tools:

No specialized data extraction tools were required since the dataset was directly downloaded from Kaggle. Standard data manipulation and analysis tools such as Python and Pandas were used for initial exploration and preprocessing.

Data Validation:

To validate the data, exploratory data analysis (EDA) techniques were applied, including visualizations and statistical summaries. Any unexpected patterns or discrepancies were addressed through further investigation and cleaning.

Importing required libraries:

Importing the Dependencies


```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
```


Reading dataset:

Data Collection & Analysis

```
[ ] # loading the data from csv file to a Pandas DataFrame
customer_data = pd.read_csv('/content/Mall_Customers.csv')
```

```
▶ # first 5 rows in the dataframe
customer_data.head()
```



	CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Checking shape and datatypes:

The shape function is used to check the number of rows and columns present in the dataset.

```
[ ] # finding the number of rows and columns
customer_data.shape

(200, 5)
```

The info or dtypes function is used to check what all the data types are in a data-frame.

```
[ ] # getting some informations about the dataset
customer_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 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
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

DATA CLEANING:

Looking for null or missing values:

The `isnull()` function is used to detect missing values.

```
[ ] # checking for missing values
customer_data.isnull().sum()
```

```
CustomerID      0
Gender          0
Age            0
Annual Income (k$)  0
Spending Score (1-100)  0
dtype: int64
```

The output result shows that there are no missing values.

DATA SORTING:

Choosing the Annual Income Column & Spending Score column

```
[ ] X = customer_data.iloc[:,[3,4]].values
```

```
▶ print(X)
```

```
[[ 15  39]
 [ 15  81]
 [ 16   6]
 [ 16  77]
 [ 17  40]
 [ 17  76]
 [ 18   6]
 [ 18  94]
 [ 19   3]
 [ 19  72]
 [ 19  14]
 [ 19  99]
 [ 20  15]
 [ 20  77]
 [ 20  13]
 [ 20  79]
 [ 21  35]
 [ 21  66]
 [ 23  29]
 [ 23  98]
 [ 24  35]
 [ 24  73]]
```



[25 5]
[25 73]
[28 14]
[28 82]
[28 32]
[28 61]
[29 31]
[29 87]
[30 4]
[30 73]
[33 4]
[33 92]
[33 14]
[33 81]
[34 17]
[34 73]
[37 26]
[37 75]
[38 35]
[38 92]
[39 36]
[39 61]
[39 28]
[39 65]
[40 55]
[40 47]



[40 42]
[42 52]
[42 60]
[43 54]
[43 60]
[43 45]
[43 41]
[44 50]
[44 46]
[46 51]
[46 46]
[46 56]
[46 55]
[47 52]
[47 59]
[48 51]
[48 59]
[48 50]
[48 48]
[48 59]
[48 47]
[49 55]
[49 42]
[50 49]
[50 56]
[54 47]



[54	54]
[54	53]
[54	48]
[54	52]
[54	42]
[54	51]
[54	55]
[54	41]
[54	44]
[54	57]
[54	46]
[57	58]
[57	55]
[58	60]
[58	46]
[59	55]
[59	41]
[60	49]
[60	40]
[60	42]
[60	52]
[60	47]
[60	50]
[61	42]
[61	49]
[62	41]



[62	48]
[62	59]
[62	55]
[62	56]
[62	42]
[63	50]
[63	46]
[63	43]
[63	48]
[63	52]
[63	54]
[64	42]
[64	46]
[65	48]
[65	50]
[65	43]
[65	59]
[67	43]
[67	57]
[67	56]
[67	40]
[69	58]
[69	91]
[70	29]
[70	77]
[71	35]



[71 95]
[71 11]
[71 75]
[71 9]
[71 75]
[72 34]
[72 71]
[73 5]
[73 88]
[73 7]
[73 73]
[74 10]
[74 72]
[75 5]
[75 93]
[76 40]
[76 87]
[77 12]
[77 97]
[77 36]
[77 74]
[78 22]
[78 90]
[78 17]
[78 88]
[78 20]



[78 76]
[78 16]
[78 89]
[78 1]
[78 78]
[78 1]
[78 73]
[79 35]
[79 83]
[81 5]
[81 93]
[85 26]
[85 75]
[86 20]
[86 95]
[87 27]
[87 63]
[87 13]
[87 75]
[87 10]
[87 92]

```
[ ] [ 88 86]
    [ 88 15]
    [ 88 69]
    [ 93 14]
    [ 93 90]
    [ 97 32]
    [ 97 86]
    [ 98 15]
    [ 98 88]
    [ 99 39]
    [ 99 97]
    [101 24]
    [101 68]
    [103 17]
    [103 85]
    [103 23]
    [103 69]
    [113  8]
    [113 91]
    [120 16]
    [120 79]
    [126 28]
    [126 74]
    [137 18]
    [137 83]]
```

PREDICTION/ANALYSIS USING ML TECHNIQUE:

- Machine learning is a broader field that encompasses various techniques and algorithms, including supervised learning, unsupervised learning, and reinforcement learning.
- K-means is a specific algorithm used in machine learning, particularly in the field of unsupervised learning and clustering.
- So, K-means is a machine learning technique.
- In K-means clustering, the algorithm partitions a dataset into K clusters, where each data point belongs to the cluster with the nearest mean.
- It is commonly used for grouping similar data points together based on their features.

Choosing the optimum number of clusters:

1. WCSS stands for "Within-Cluster Sum of Squares," and it is a method used in the context of K-means clustering to evaluate the performance of the clustering algorithm. Specifically, WCSS measures the compactness of the clusters.

For each cluster, WCSS calculates the sum of the squared distances between each data point within the cluster and the centroid of that cluster. It represents how tightly grouped the data points within each cluster are.

The main goal of K-means is to minimize the WCSS across all clusters. In other words, the algorithm aims to find cluster centroids in such a way that the sum of squared distances of each point in the cluster to its centroid is minimized. The “init” argument is a method for initializing the centroid.

WCSS -> Within Clusters Sum of Squares

```
[ ] # finding wcss value for different number of clusters

wcss = []

for i in range(1,11):
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
    kmeans.fit(X)

    wcss.append(kmeans.inertia_)
```

2. The elbow method is a technique used to determine the optimal number of clusters (K) in K-means clustering.

To apply this method, the K-means algorithm is run for a range of K values, and the Within-Cluster Sum of Squares (WCSS) is calculated for each iteration.

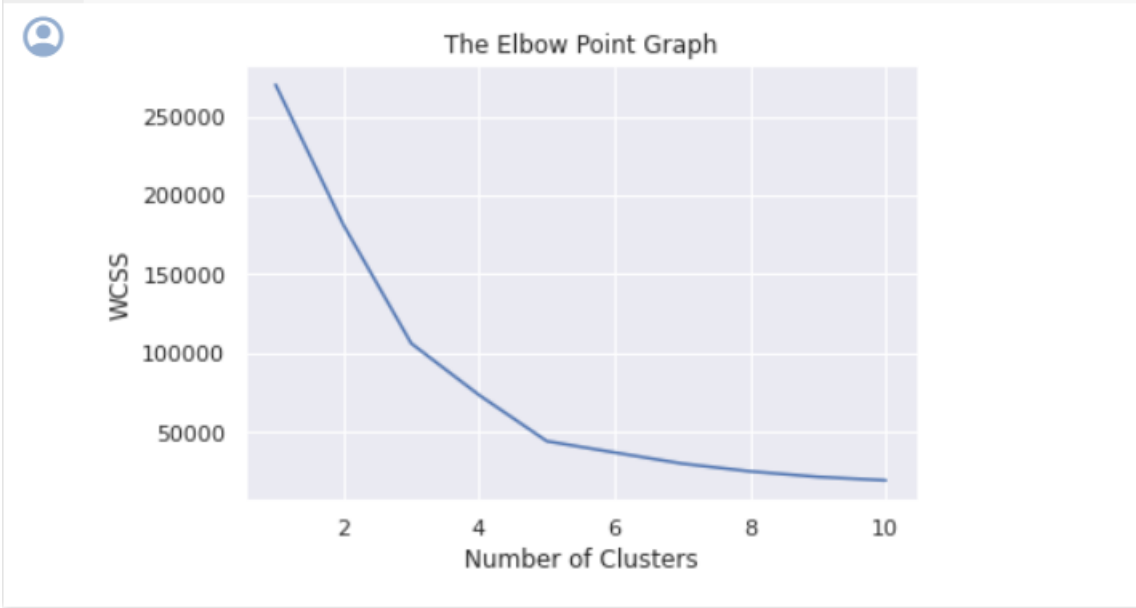
The results are then plotted on a graph with K on the x-axis and WCSS on the y-axis. The key step is to identify the "elbow" point in the graph, where the rate of WCSS decrease slows down.

This point signifies the optimal K, as it represents a balance between reducing WCSS and avoiding excessive complexity in the number of clusters.

The elbow method serves as a visual aid in determining a suitable number of clusters for a given dataset in K-means clustering.

```
# plot an elbow graph

sns.set()
plt.plot(range(1,11), wcss)
plt.title('The Elbow Point Graph')
plt.xlabel('Number of Clusters')
plt.ylabel('WCSS')
plt.show()
```



Optimum Number of Clusters = 5

Training the K-Means clustering model:

Let's train the model on the dataset using five clusters.

[illegible]

5 Clusters - 0, 1, 2, 3, 4

These values above denote the cluster assignment for each data point.

Visualizing all the clusters:

```
# plotting all the clusters and their Centroids

plt.figure(figsize=(8,8))
plt.scatter(X[Y==0,0], X[Y==0,1], s=50, c='green', label='Cluster 1')
plt.scatter(X[Y==1,0], X[Y==1,1], s=50, c='red', label='Cluster 2')
plt.scatter(X[Y==2,0], X[Y==2,1], s=50, c='yellow', label='Cluster 3')
plt.scatter(X[Y==3,0], X[Y==3,1], s=50, c='violet', label='Cluster 4')
plt.scatter(X[Y==4,0], X[Y==4,1], s=50, c='blue', label='Cluster 5')

# plot the centroids
plt.scatter(kmeans.cluster_centers_[0,0], kmeans.cluster_centers_[0,1], s=100, c='cyan', label='Centroids')

plt.title('Customer Groups')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.show()
```

Plots:



Result:

The analysis reveals five customer clusters for targeted marketing in the retail store.

1. Cluster 1 (green): Moderate income, Moderate spending. Cautious spenders.
2. Cluster 2 (yellow): High income, high spending. Profitable. Targeting with discounts can boost spending and maximize profit.
3. Cluster 3 (red): Higher income but lower spending. Potential for increased profit by addressing potential dissatisfaction with store services.
4. Cluster 4 (purple): Low income, low spending. Expected due to limited purchasing power.
5. Cluster 5 (blue): Low income, high spending. Suggests satisfaction with store services despite lower earnings.

Conclusion:

In conclusion, the K-means clustering analysis conducted in the "Customer Segmentation" project has provided a comprehensive understanding of customer behaviour and preferences. The identified clusters offer clear demarcations in income and spending patterns, enabling businesses to strategically target and engage with specific customer segments.

The K-means clustering methodology, as demonstrated in this project, serves as a powerful tool for businesses seeking to understand and respond effectively to the diverse needs and behaviours of their customer base, ultimately fostering improved customer relationships and maximizing overall business success.

This segmentation facilitates personalized marketing strategies, allowing for more effective communication and product/service customization. As businesses navigate the competitive landscape, the insights gained from K-means clustering become pivotal in optimizing marketing efforts, resource allocation, and overall business strategy.

The project underscores the significance of data-driven approaches in enhancing customer segmentation, leading to improved decision-making and, ultimately, heightened customer satisfaction and business success.