A PROJECT REPORT

on

“CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING”

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR DEGREE IN
COMPUTER SCIENCE AND ENGINEERING

BY

SANDESH GHIMIRE
20051753

UNDER THE GUIDANCE OF
MR. AJAY ANAND

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA - 751024
May 2023

Acknowledgement

We are profoundly grateful to **MR. AJAY ANAND** of **School of Computer Science** for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

SANDESH GHIMIRE

ABSTRACT

Large amounts of data are frequently generated by our surroundings, and it is crucial to analyze this data. The company strategy must be adjusted to the circumstances in the contemporary age of innovation when everyone is striving to outperform one another. Because so many potential customers are unclear about what to buy and what not to buy, today's business relies on new ideas. Businesses can't assess their target market on their own. In order to improve decision-making, machine learning is used to find hidden patterns in data using a variety of techniques. Comparing data points from several groups is a key component of the machine learning process known as clustering. Applications include image processing, pattern identification, market research, medical data, search engine optimization, and others are among them. Our study is about customer segmentation, which is a subset of market research. The division of consumers into groups according to shared traits is known as customer segmentation.

Businesses must categorize their customers in the current environment according to their age, gender, location, and other traits. This allows businesses to focus on certain clients who are most likely to purchase their goods. If they can effectively implement machine learning to enhance their operations, they will have a competitive advantage over their rivals. The major objective of this project is to use the K-means algorithm to categorize clients according to their qualities. The information from the various clusters will finally reveal which group the new client will belong to, with the mean value serving as the primary indicator.

Keywords: K-Means algorithm, Customer segmentation, Mall Customers, Normalization, Elbow method, Python, Machine Learning.

Table of Content

Introduction..2
Literature Review..3
2.1 Customer Segmentation..3
2.2 Clustering and K-Means Algorithm..3
Requirement Specifications..4
3.1 Project Planning..4
3.2 Project Analysis(SRS)..6
3.3 System Design..8
3.3.1 Design Constraints..8
3.3.2 System Architecture..8
Block Diagram..8
Class Diagram..9
Implementation..10
4.1 Algorithm..10
4.2 Libraries Used..10
4.3 Platform Used..12
4.4 Methodology..12
4.5 Review and Testing..16
4.6 Result Analysis..17
Standards Adopted..18
5.1 Design Standards..18
5.2 Coding Standards..19
5.3 Testing Standards..19
Conclusion and Future Scope..20
6.1 Conclusion..20
6.2 Future Scope..20
References..21
Individual contribution report..22
Plagiarism Report..26

Introduction

As competition among businesses continues to intensify, customer segmentation has become a critical tool for businesses to better understand their customers and tailor marketing programs to meet their needs. Data mining techniques, such as clustering algorithms, can be used to extract meaningful and strategic information from organizational databases. One popular clustering algorithm is the K-means algorithm, which groups data objects or customers into clusters based on their similarities in characteristics such as gender, age, interests, and spending habits.

This report aims to demonstrate the use of the K-means algorithm and the elbow method to identify customer segments from a dataset of customer information. The elbow method is used to determine the optimal number of clusters for the data set. The insights obtained from customer segmentation can help businesses modify their marketing programs, make better decisions, manage product demand and supply, identify potential customers, predict customer defection, and find solutions. By leveraging data mining techniques and customer segmentation, businesses can achieve higher customer satisfaction and loyalty, leading to increased sales and growth.

Literature Review

2.1 Customer Segmentation

In today's highly competitive business world, satisfying customer demands and attracting new customers according to their needs is crucial for enhancing profits and business growth. However, this can be a complex and tedious task since customers may have different demands, tastes, preferences, and other behavioral characteristics. Instead of a "one-size-fits-all" approach, customer segmentation clusters customers into groups that share similar properties or behavioral characteristics. Customer segmentation is a market division strategy that relies on various factors such as geographical, economic, and demographic conditions, as well as behavioral patterns. This technique enables businesses to make better use of their marketing budgets, gain a competitive edge over rivals, demonstrate their knowledge of customer needs, increase marketing efficiency, identify new market opportunities, develop better brand strategies, and improve customer retention. These benefits of customer segmentation have led many businesses to adopt this approach to enhance their performance in the marketplace.

2.2 Clustering and K-Means Algorithm

Clustering techniques are used to group similar data objects together based on their characteristics or behavior. K-means algorithm is a popular centroid-based method used for clustering. Given a dataset D containing n objects, partitioning methods distribute the objects into k clusters, Ci, where each cluster is represented by its centroid or mean value. The similarity between an object and the centroid of a cluster is measured using the Euclidean distance. The initial cluster centres are selected at random from a set of D items by the k-means method, which then iteratively assigns each object to the cluster that has the greatest mean value of the cluster's objects. The cluster means are then updated, and the process is repeated until no further changes occur. This algorithm is used for partitioning data into k clusters, with each cluster being represented by its centroid. This approach is commonly used for customer segmentation, as it allows businesses to identify groups of customers with similar characteristics and tailor their marketing strategies accordingly.

Requirement Specifications

The problem we are working on is to help a retail company segment its customers based on their purchasing behavior and demographics. The company has a large customer base, and they want to gain insights into their customers' behavior using data analytics. The main objective is to personalize marketing campaigns and improve customer experience. We aim to develop a data analytics solution that can segment customers into different groups based on their purchasing history, demographics, and other relevant factors. This will help the company to tailor their marketing campaigns to each customer group's needs and preferences, leading to increased customer satisfaction and improved sales. We will follow the IEEE format to present the System Requirements Specification (SRS) for the project.

3.1 Project Planning

1. **Problem Identification:** The problem is to segment the customers of a mall into different groups based on their spending habits and annual income so that the marketing team can target them more efficiently.

2. **Data Collection:** Data is available in the CSV file named "Mall_Customer.csv" which contains the following columns: CustomerID, Gender, Age, Annual Income (k$), Spending Score (1-100). We will load the data into a pandas DataFrame and perform exploratory data analysis to understand the data.

3. Data Preprocessing:

- Check for missing values and remove the column "CustomerID" as it does not contribute to the analysis.

- Check for data types and convert if necessary.

- Scale the data using either MinMaxScaler or StandardScaler.

- Check for outliers and decide if they need to be removed.

4. Exploratory Data Analysis:

- Plot distribution plots for Age, Annual Income, and Spending Score.

- Plot a count plot for Gender.

- Plot a violin plot to see the distribution of Age, Annual Income, and Spending Score.

- Plot a bar chart to see the number of customers in each age group.

- Plot a scatter plot to see the relation between Annual Income and Spending Score.

- Plot a bar chart to see the number of customers having a particular Spending Score and Annual Income range.

5. Model Building:

- Build a KMeans clustering model to segment the customers into different groups based on their spending habits and annual income.

- Determine the optimal number of clusters using the elbow method.

- Fit the model on the scaled data.

- Predict the clusters for each customer.

6. Model Evaluation:

- Visualize the clusters using scatter plots.

- Analyze the characteristics of each cluster and give them meaningful names.

7. Results:

- Summarize the findings in a report.

- Provide recommendations to the marketing team based on the analysis.

- Discuss the limitations of the model and possible improvements.

8. Deployment:

- Deploy the model as a web application or integrate it into the existing system.

- Provide necessary training to the end-users.

- Monitor the performance of the model and update it if necessary.

3.2 Project Analysis(SRS):

Introduction:

Mall Customer Segmentation Project is a data analytics project that involves customer segmentation based on various factors like annual income, age, and spending score. The project aims to help the mall administration to identify customer patterns, preferences and market effectively based on the specific customer segmentation.

Purpose:

The purpose of this document is to provide a detailed specification of requirements and features that the Mall Customer Segmentation Project must meet.

Scope:

The Mall Customer Segmentation Project aims to achieve the following objectives:

- Segment the customers based on their age, annual income, and spending score.

- Identify the number of customers in each age group and income range.

- Visualize the spending score of customers and understand their spending behavior.

- Evaluate different customer segments and provide insights to the mall administration.

- Develop a customer segmentation model using K-means clustering algorithm.

- Provide a graphical user interface (GUI) to visualize the customer segments.

Functional Requirements:

The following functional requirements must be fulfilled in the Mall Customer Segmentation Project:

1. Import customer data from a CSV file.

2. Display basic information of the dataset, such as size, description, and data types.

3. Pre-process the data by removing unnecessary columns and handling missing values.

4. Visualize the data using different plots, such as distribution plot, count plot, bar plot, and scatter plot.

5. Segment the customers using K-means clustering algorithm.

6. Determine the optimal number of clusters using the elbow method.

7. Plot the clusters and analyze the characteristics of each segment.

8. Provide an option to select the features used for clustering.

9. Save the customer segmentation results to a CSV file.

Non-Functional Requirements:

The following non-functional requirements must be fulfilled in the Mall Customer Segmentation Project:

1. The system should be user-friendly and easy to use.

2. The system should provide accurate results with minimum error.

3. The system should be able to handle a large amount of data.

4. The system should be scalable to accommodate future enhancements.

5. The system should be able to perform the segmentation in a reasonable amount of time.

Assumptions and Dependencies:

The Mall Customer Segmentation Project assumes that the customer data is provided in a CSV file format is accurate and complete. The project also assumes that the K-means clustering algorithm is used for customer segmentation. The project depends on various Python libraries, such as Pandas, Matplotlib, Seaborn, and Scikit-learn.

Constraints:

The Mall Customer Segmentation Project is subject to the following constraints:

1. The project must be developed using the Python programming language.

2. The project must use the K-means clustering algorithm for customer segmentation.

3. The project must be completed within the given time frame.

4. The project must be within the allocated budget

3.3 System Design

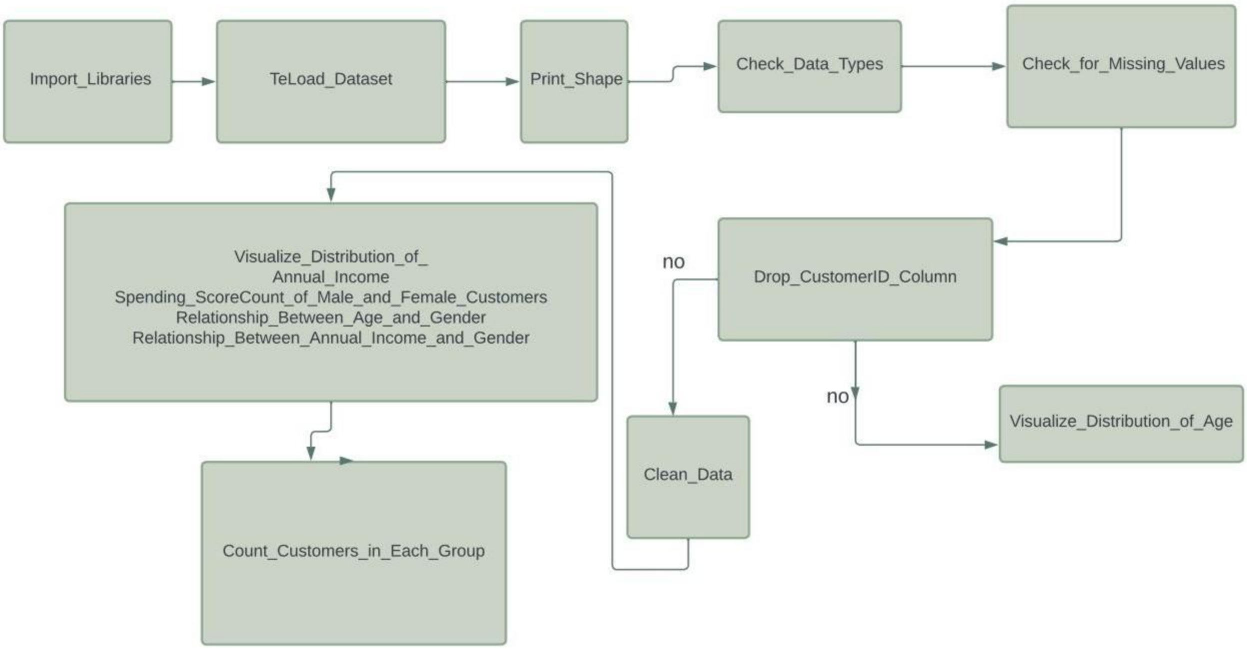
3.3.1 Design Constraints

- The system should have Python and all required libraries installed to execute the code.
- Before running the code, ensure that the 'Mall_Customer.csv' file is present in the 'C:\Users\KIIT\Downloads' path, and the CSV file contains the required columns.
- The code is dependent on external libraries like numpy, pandas, matplotlib, seaborn, and sklearn, which should be imported and installed in the system.
- The system should have enough processing power and memory to handle the data and generate visualizations.
- Before deploying the code in a production environment, it is crucial to review and test it for accuracy and reliability.

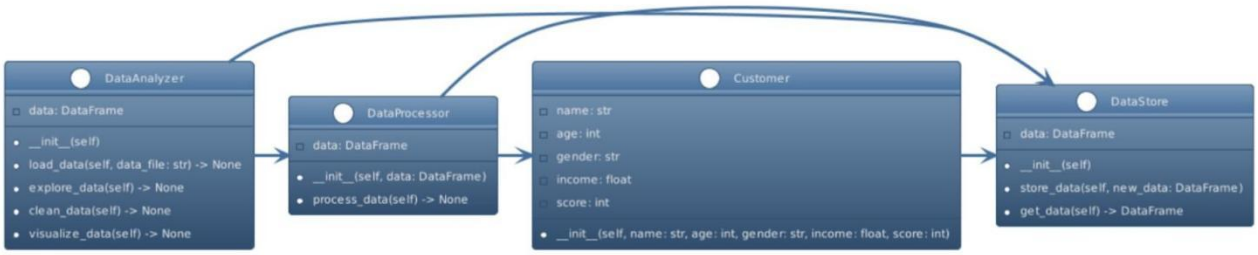
3.3.2 System Architecture

Block Diagram

The system architecture involves importing the necessary libraries and loading the customer data CSV file using pandas. The system performs various data preprocessing and visualization tasks, such as removing unwanted columns and checking for missing values. After the data is preprocessed, the system applies clustering techniques to segment customers into different groups based on their similarities


Import_Libraries
TeLoad_Dataset
Print_Shape
Check_Data_Types
Check_for_Missing_Values
Visualize_Distribution_of_
Annual_Income
Spending_ScoreCount_of_Male_and_Female_Customers
Relationship_Between_Age_and_Gender
Relationship_Between_Annual_Income_and_Gender
no
Drop_CustomerID_Column
no
Visualize_Distribution_of_Age
Clean_Data
Count_Customers_in_Each_Group

Class Diagram :


DataAnalyzer
data: DataFrame
__init__(self)
load_data(self, data_file: str) -> None
explore_data(self) -> None
clean_data(self) -> None
visualize_data(self) -> None
DataProcessor
data: DataFrame
__init__(self, data: DataFrame)
process_data(self) -> None
Customer
name: str
age: int
gender: str
income: float
score: int
__init__(self, name: str, age: int, gender: str, income: float, score: int)
DataStore
data: DataFrame
__init__(self)
store_data(self, new_data: DataFrame)
get_data(self) -> DataFrame

Explanation:

Customer: represents a single customer and stores their information such as name, age, gender, income, and score.

DataStore: represents a class that is responsible for storing and managing the customer data. It has a single attribute, data, which is a Pandas DataFrame that stores the customer data.

DataProcessor: represents a class that is responsible for processing the customer data. It has a single attribute, data, which is a Pandas DataFrame that stores the customer data.

DataAnalyzer: represents a class that is responsible for analyzing the customer data. It has an attribute, data, which is a Pandas DataFrame that stores the customer data. The class has several methods that perform various tasks such as loading the data from a file, exploring the data, cleaning the data, and visualizing the data.

Implementation

4.1 Algorithm

KMeans clustering algorithm is used to find the clusters of given dataset. The K-means algorithm in data mining begins by randomly selecting initial centroids to form clusters. It then proceeds to iteratively calculate and adjust the positions of the centroids until they reach an optimized state, or until the maximum number of iterations is reached. The algorithm terminates when the centroids stabilize, indicating successful clustering, or when the specified number of iterations is completed.
KMeans algorithm utilizes the Euclidean distance metric to determine the similarity between data points and selects the mean value of the data points as the initial cluster centroid. It then iteratively updates the centroid by recalculating the mean of the data points assigned to each cluster and moving the centroid towards the new mean until convergence is achieved.

$$\textbf{Euclidean Distance} = |X - Y| = \sqrt{\sum_{i=1}^{i=n} (x_i - y_i)^2}$$

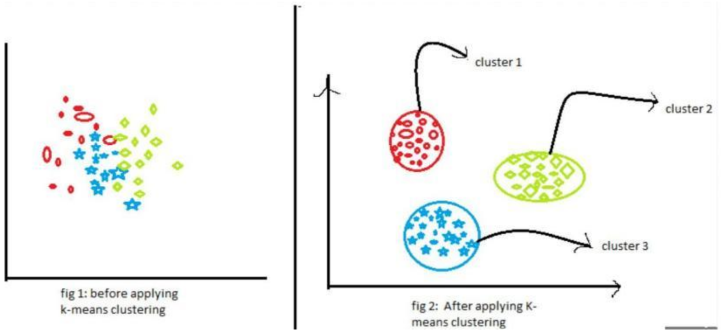


fig 1: before applying k-means clustering

fig 2: After applying K-means clustering

4.2 Libraries Used

NumPy:
NumPy is a Python library for numerical computation that provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays. It offers efficient and optimized computation of mathematical operations on arrays, making it a popular tool for scientific computing, data analysis, and machine learning tasks in Python.

Pandas:

Pandas is a Python library designed for data manipulation and analysis, offering a range of data structures and tools for processing and organizing data in a flexible and efficient manner. It is widely used for tasks such as data cleaning, filtering, transformation, and visualization, making it an essential tool for data analysis and exploration in Python.Pandas provides a high-performance, easy-to-use data structure and data analysis tools for Python, enabling efficient handling and processing of tabular data, time-series, and statistical data sets. Its powerful functionality makes it a popular choice for data scientists and analysts working with data in Python.

Matplotlib:

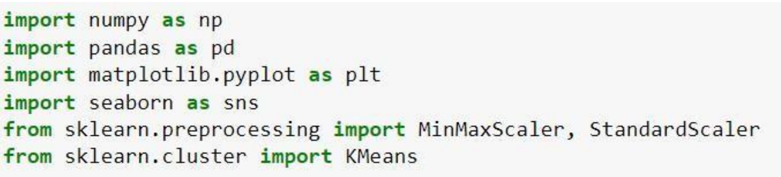
Matplotlib is a Python library used for creating high-quality visualizations and plots in a variety of formats, including interactive and publication-ready graphics. It provides a range of customizable plots, charts, and graphs, making it a powerful tool for data visualization and exploration. Matplotlib is widely used in fields such as data science, engineering, and finance for data analysis and presentation.

Seaborn:

Seaborn is a Python data visualization library based on Matplotlib that provides a higher-level interface for creating informative and attractive statistical graphics. It simplifies the creation of complex visualizations through built-in themes, color palettes, and functions for visualizing data distributions, regression models, and categorical data. Seaborn is widely used in data science, machine learning, and other domains for exploratory data analysis and presentation.

Scikit-Learn:

Scikit-learn, also known as sklearn, is a popular Python machine learning library that provides a range of supervised and unsupervised learning algorithms for data mining and analysis. It offers tools for data preprocessing, feature selection, model selection, and evaluation, making it a powerful tool for building and deploying machine learning models. Scikit-learn is widely used in academic research, industry, and other domains for a wide range of applications, including classification, regression, clustering, and dimensionality reduction.



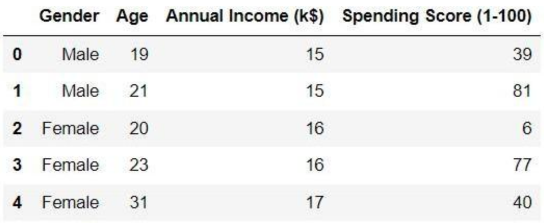
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler, StandardScaler
from sklearn.cluster import KMeans
```

4.3 Platform Used:

Jupyter Notebook is an open-source web application that allows users to create and share documents that contain live code, equations, visualizations, and narrative text. It is a popular platform for Python programming, as it provides an interactive environment for data analysis and exploration. The platform supports a wide range of programming languages and offers features such as code highlighting, autocompletion, and debugging. Jupyter Notebook is widely used in data science, scientific research, and education for sharing and collaborating on data-driven projects.

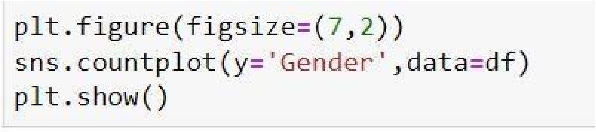
4.4 Methodology

A shopping center store provided the dataset for clustering using the K-means algorithm. The dataset consisting of 200 tuples representing information on 200 consumers was provided by a shopping center store for clustering using the K-means algorithm. The dataset comprises five attributes, including CustomerId, gender, age, yearly income (k$), and spending score rated on a scale of 1-100.

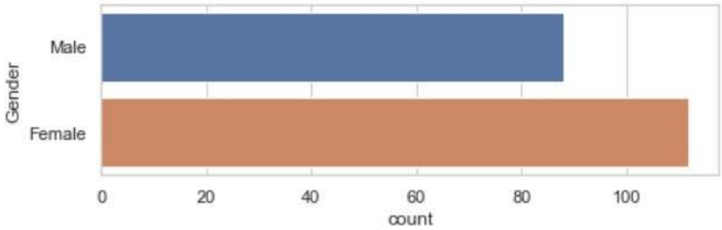


	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	Male	19	15	39
1	Male	21	15	81
2	Female	20	16	6
3	Female	23	16	77
4	Female	31	17	40

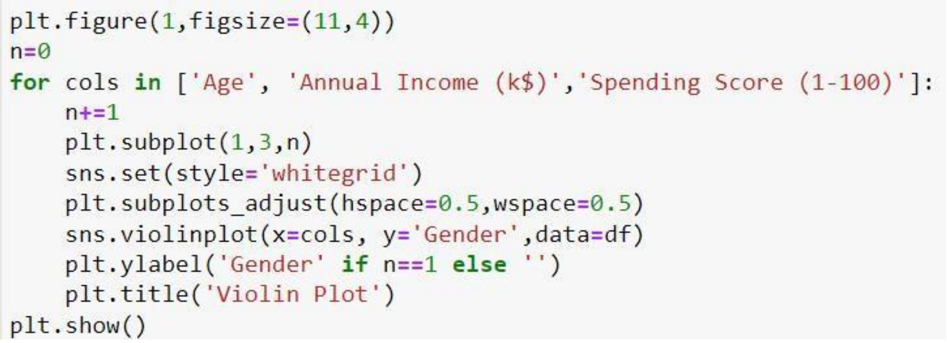
Visualize the gender of customer:



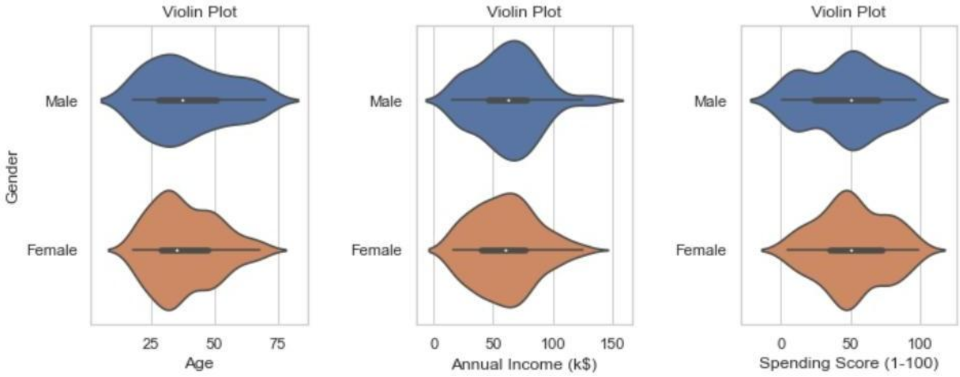
```
plt.figure(figsize=(7,2))
sns.countplot(y='Gender',data=df)
plt.show()
```



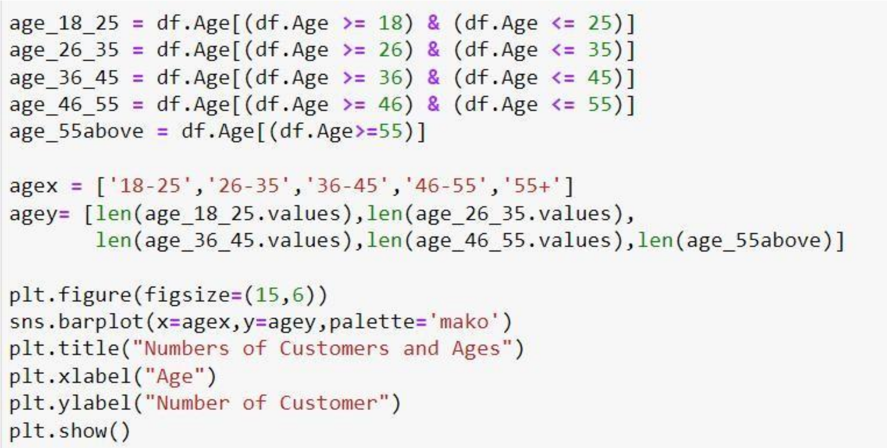
Violin plot of gender and other features:



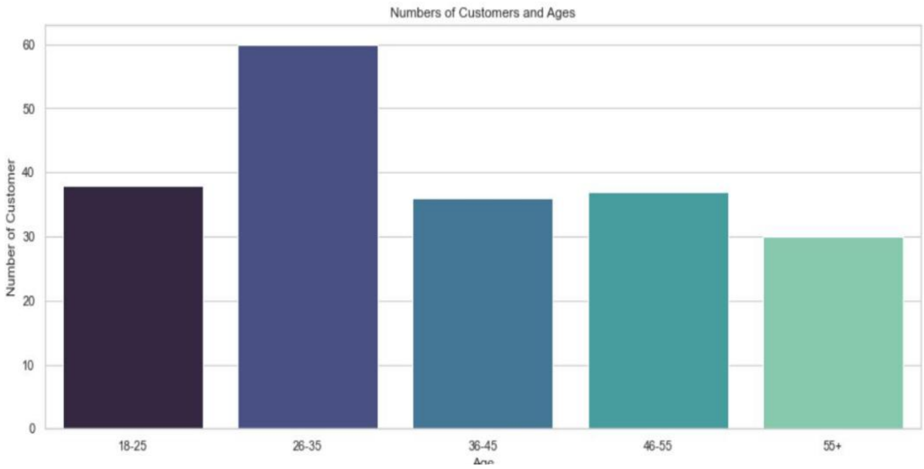
```
plt.figure(1,figsize=(11,4))
n=0
for cols in ['Age', 'Annual Income (k$)','Spending Score (1-100)']:
    n+=1
    plt.subplot(1,3,n)
    sns.set(style='whitegrid')
    plt.subplots_adjust(hspace=0.5,wspace=0.5)
    sns.violinplot(x=cols, y='Gender',data=df)
    plt.ylabel('Gender' if n==1 else '')
    plt.title('Violin Plot')
plt.show()
```



Bar plot of number of customers and their age

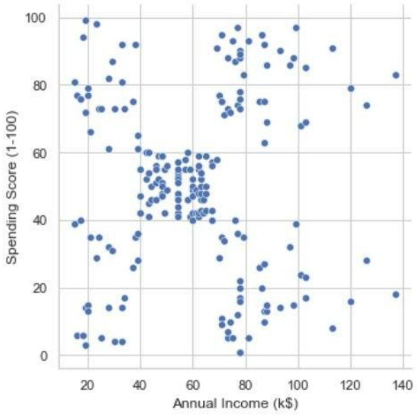


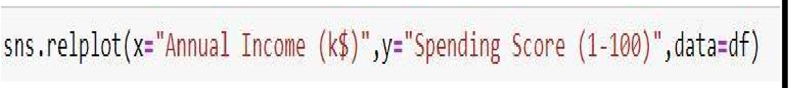
```
age_18_25 = df.Age[(df.Age >= 18) & (df.Age <= 25)]
age_26_35 = df.Age[(df.Age >= 26) & (df.Age <= 35)]
age_36_45 = df.Age[(df.Age >= 36) & (df.Age <= 45)]
age_46_55 = df.Age[(df.Age >= 46) & (df.Age <= 55)]
age_55above = df.Age[(df.Age>=55)]

agex = ['18-25','26-35','36-45','46-55','55+']
agey= [len(age_18_25.values),len(age_26_35.values),
       len(age_36_45.values),len(age_46_55.values),len(age_55above)]

plt.figure(figsize=(15,6))
sns.barplot(x=agex,y=agey,palette='mako')
plt.title("Numbers of Customers and Ages")
plt.xlabel("Age")
plt.ylabel("Number of Customer")
plt.show()
```



In order to ensure reliable and usable data for analysis, data cleaning must be performed on datasets containing null values, duplicates, or other noisy data. Once the data has been cleaned, it can be visualized through gender-specific comparisons of annual income and spending scores using one of five different types of plots. These plots illustrate different customer behaviors related to annual income and spending scores and can be used to identify different groups of customers.

.





```
sns.relplot(x="Annual Income (k$)",y="Spending Score (1-100)",data=df)
```

Elbow Method

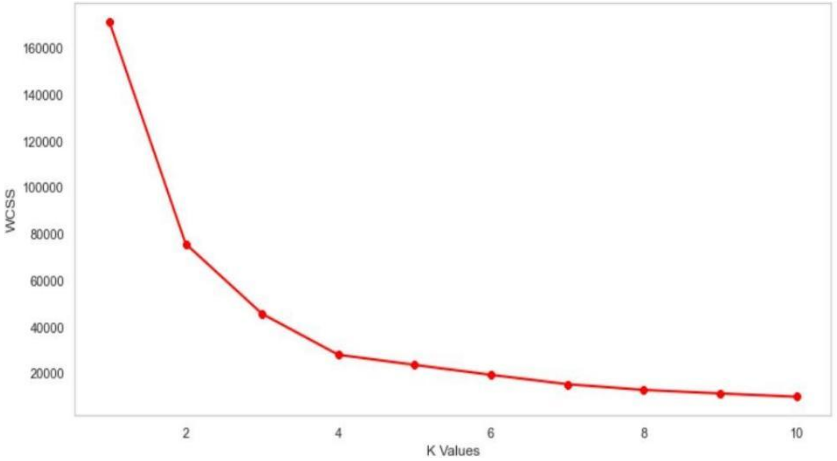
The elbow method is a popular technique used to determine the optimal number of clusters in K-means clustering. It works by plotting the sum of squared distances (SSE) between the data points and their assigned cluster centroids for a range of values of K, the number of clusters. The point on the plot where the SSE begins to level off is called the elbow point, and it indicates the optimal number of clusters for the data. The elbow method works by comparing the reduction in SSE with the increasing number of clusters, and selecting the value of K that provides a good balance between low SSE and the number of clusters required.

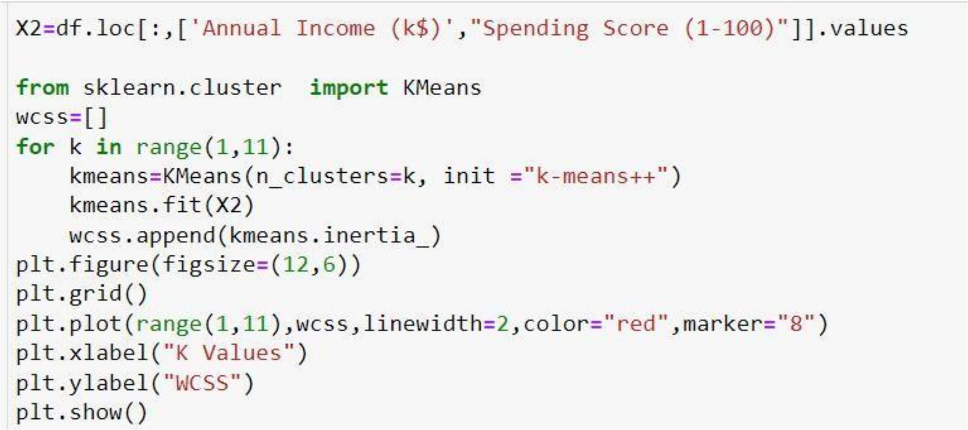
The elbow method uses the following formula to calculate the sum of squared distances (SSE) between the data points and their assigned cluster centroids for a given value of K:

SSE = Σ_i=1^k Σ_x∈C_i (dist(x, c_i))^2

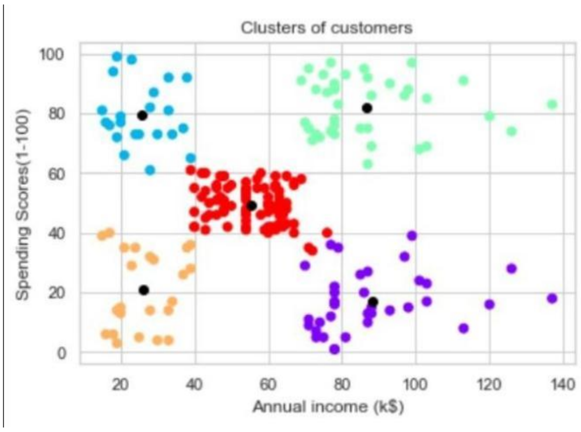
Where:

k is the number of clusters
C_i is the i-th cluster
x is a data point in cluster C_i
c_i is the centroid of cluster C_i
dist(x, c_i) is the Euclidean distance between data point x and centroid c_i.



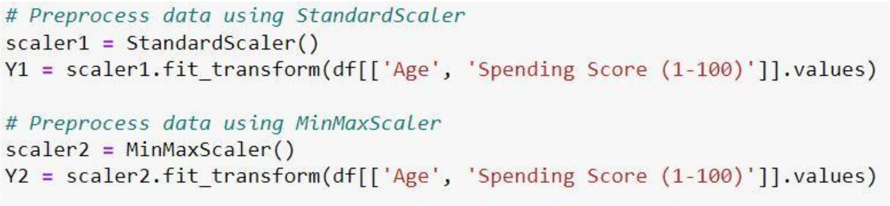


```
X2=df.loc[:,['Annual Income (k$)',"Spending Score (1-100)"]].values

from sklearn.cluster  import KMeans
wcss=[]
for k in range(1,11):
    kmeans=KMeans(n_clusters=k, init ="k-means++")
    kmeans.fit(X2)
    wcss.append(kmeans.inertia_)
plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,11),wcss,linewidth=2,color="red",marker="8")
plt.xlabel("K Values")
plt.ylabel("WCSS")
plt.show()
```

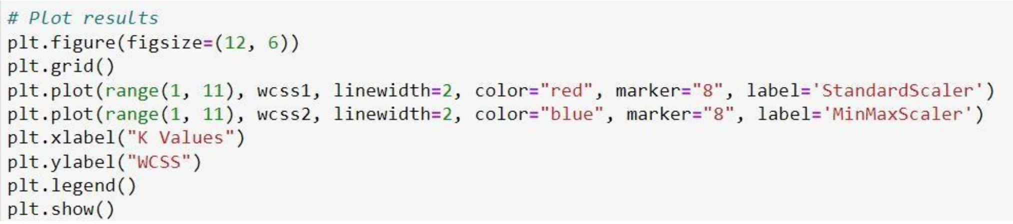


4.5 Review and Testing

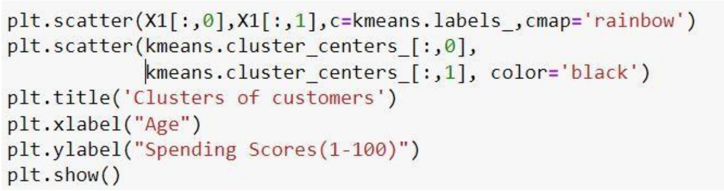
Further exploring the data, we see some possible deviations in values ans result like small change in one feature affect another feature largely no before clustering we tried to process data again using standardScalar and minMaxScalar.



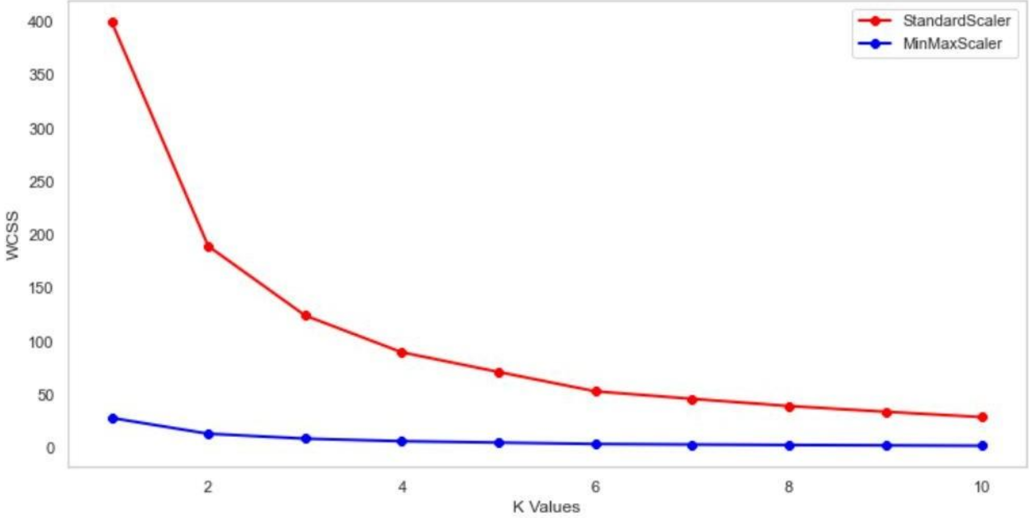
```
# Preprocess data using StandardScaler
scaler1 = StandardScaler()
Y1 = scaler1.fit_transform(df[['Age', 'Spending Score (1-100)']].values)

# Preprocess data using MinMaxScaler
scaler2 = MinMaxScaler()
Y2 = scaler2.fit_transform(df[['Age', 'Spending Score (1-100)']].values)
```



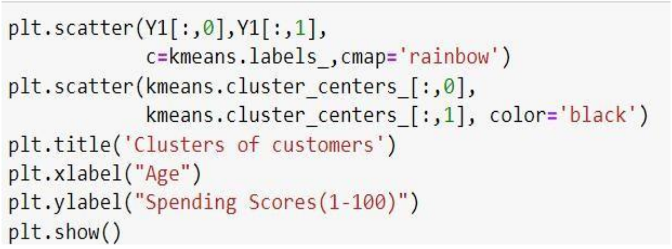
```
# Plot results
plt.figure(figsize=(12, 6))
plt.grid()
plt.plot(range(1, 11), wcss1, linewidth=2, color="red", marker="8", label='StandardScaler')
plt.plot(range(1, 11), wcss2, linewidth=2, color="blue", marker="8", label='MinMaxScaler')
plt.xlabel("K Values")
plt.ylabel("WCSS")
plt.legend()
plt.show()
```



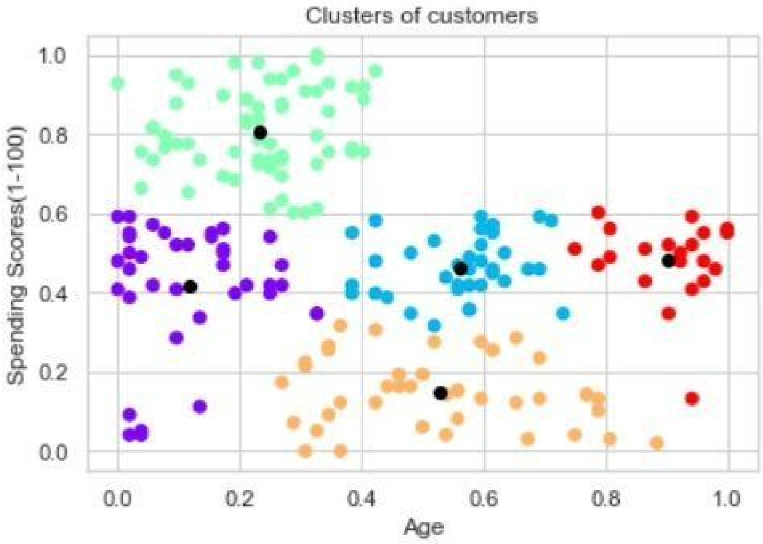
```
plt.scatter(X1[:,0],X1[:,1],c=kmeans.labels_,cmap='rainbow')
plt.scatter(kmeans.cluster_centers_[:,0],
            kmeans.cluster_centers_[:,1], color='black')
plt.title('Clusters of customers')
plt.xlabel("Age")
plt.ylabel("Spending Scores(1-100)")
plt.show()
```



After standardization



```
plt.scatter(Y1[:,0],Y1[:,1],
            c=kmeans.labels_,cmap='rainbow')
plt.scatter(kmeans.cluster_centers_[:,0],
            kmeans.cluster_centers_[:,1], color='black')
plt.title('Clusters of customers')
plt.xlabel("Age")
plt.ylabel("Spending Scores(1-100)")
plt.show()
```



4.6 Result Analysis:

This analysis is based on data without standardization as cluster after standardization looks more disturbed ans inefficient to analyze. The spending habits and yearly earnings of mall shoppers can be clustered into five different groups. The green group, with high earnings and high spending scores, is the most profitable target for the mall. The purple group, with high earnings and low spending, presents an opportunity for the mall to improve its offerings to attract them. The red group, with average earnings and expenditures, is not a significant source of revenue for the mall. The blue group, with low earnings but high spending, may be satisfied with the mall's services and thus compelled to spend money. The orange group, with low earnings and poor spending habits, should not be a priority for the mall. By analyzing this data, marketing strategies can be tailored to specific customer groups based on their spending habits and earnings. Promotions and discounts may be used to entice customers with lower incomes and spending scores, while higher-income customers may be attracted by the variety of products available. Cluster analysis may help identify what types of products each customer group desires, allowing for more targeted marketing efforts. Clusters 3 and 4 are the potential target groups in this scenario.

Standards Adopted

5.1 Design Standards

When designing a customer segmentation project using K-Means algorithm, it is important to follow certain design standards to ensure that the project is well-structured and effective. Here are some important design standards that we have considered in our project

- Define the objectives: Before starting the project, it is essential to define the objectives of the segmentation. This includes understanding the business problem and the goals of the project, and ensuring that the segmentation is aligned with these objectives.
- Determine the data requirements: It is important to identify the data sources required for the project, and ensure that the data is relevant, accurate and reliable. This involves understanding the data structure and format, and performing data cleaning and preprocessing as needed.
- Selecting the appropriate K value: K-Means algorithm requires the specification of the number of clusters (K) to be generated. It is important to determine the appropriate K value by analyzing the data and experimenting with different values.
- Choosing an appropriate feature scaling method: The selection of an appropriate feature scaling method can have a significant impact on the performance of the algorithm. Standardization and Normalization is used to avoid biasing the results in favor of certain features.
- Select appropriate distance metric: The distance metric is used to calculate the distance between points in the feature space. Euclidean distance is the most commonly used metric for K-Means algorithm which is applied in our project

5.2 Coding Standards

When coding a customer segmentation project using K-Means algorithm, it is important to follow certain coding standards to ensure that the code is well-structured, readable, and maintainable. Here are some important coding standards to consider:

- Use modular programming: Break the code into smaller functions or modules that are easier to understand and test. This also help in with code reusability.
- Use descriptive variable names: Choose descriptive and meaningful variable names that accurately reflect their purpose.
- Use comments: Add comments to the code to explain what the code does and how it works.
- Use consistent coding style: Use consistent coding style throughout the code to improve readability and maintainability.

5.3 Testing Standards

When testing a customer segmentation project using K-Means algorithm, it is important to follow certain testing standards to ensure that the results are reliable and accurate. Here are some important testing standards that we have considered:

- Use a representative sample: Use a representative sample of the data to ensure that the results are not biased.
- Use appropriate evaluation metrics: Appropriate evaluation metrics such as within-cluster sum of square(wcss) is used to evaluate the performance of the algorithm.
- Perform sensitivity analysis: Perform sensitivity analysis to determine the robustness of the algorithm to different parameter values and data variations.
- Perform cross-validation: Perform cross-validation to ensure that the algorithm is generalizable to new data.

Interpret the results: Interpret the results of the segmentation to gain insights and make data-driven decisions

Conclusion and Future Scope

6.1 Conclusion

In conclusion, the application of k-means clustering in customer segmentation has proven to be a useful technique for businesses seeking to gain a competitive edge in the marketplace. By dividing customers into distinct segments based on similarities in their characteristics and behaviors, businesses can tailor their marketing strategies to meet the unique needs and preferences of each segment. In this project, we have used two popular data normalization techniques, Standard Scalar and MinMax Scalar, to preprocess the data before applying k-means clustering. The results of our analysis have shown that both normalization techniques have improved the accuracy and effectiveness of the clustering algorithm. The elbow method has also been employed to determine the optimal number of clusters, which has resulted in a more meaningful and interpretable segmentation of customers. Overall, the use of k-means clustering with appropriate data preprocessing techniques has the potential to offer businesses valuable insights into their customer base and guide strategic decision-making for long-term success.

6.2 Future Scope

The future scope of customer segmentation using K-means clustering is vast. With the increasing amount of data available, businesses can use more advanced machine learning techniques and big data technologies to extract more valuable insights from customer data. Some of the future directions for this area of research include the use of deep learning algorithms for more accurate clustering, the integration of multiple data sources for a more comprehensive view of customers, and the incorporation of real-time data for faster and more effective decision-making. Additionally, businesses can explore the use of other clustering algorithms or combination of algorithms to achieve better results in customer segmentation. Moreover, the application of customer segmentation is not limited to businesses; it can be applied in various domains such as healthcare, education, and social media analysis, providing opportunities for further research and innovation

References

[1] “Customer segmentation based on survival character,” IEEE, Jul. 2003.

[2] “Customer Segmentation Using K Means Clustering,” Towards Data Science, Apr. 2019.

[3] Ruhul Reddy, “Who’s who: Understanding your business with customer segmentation,” INTERCOM.

[4] Kristen Baker, “The Ultimate Guide to Customer Segmentation: How to Organize Your Customers to Grow Better,” Hunspot.

[5] Tim Ehrens, “customer segmentation,” TechTarget.

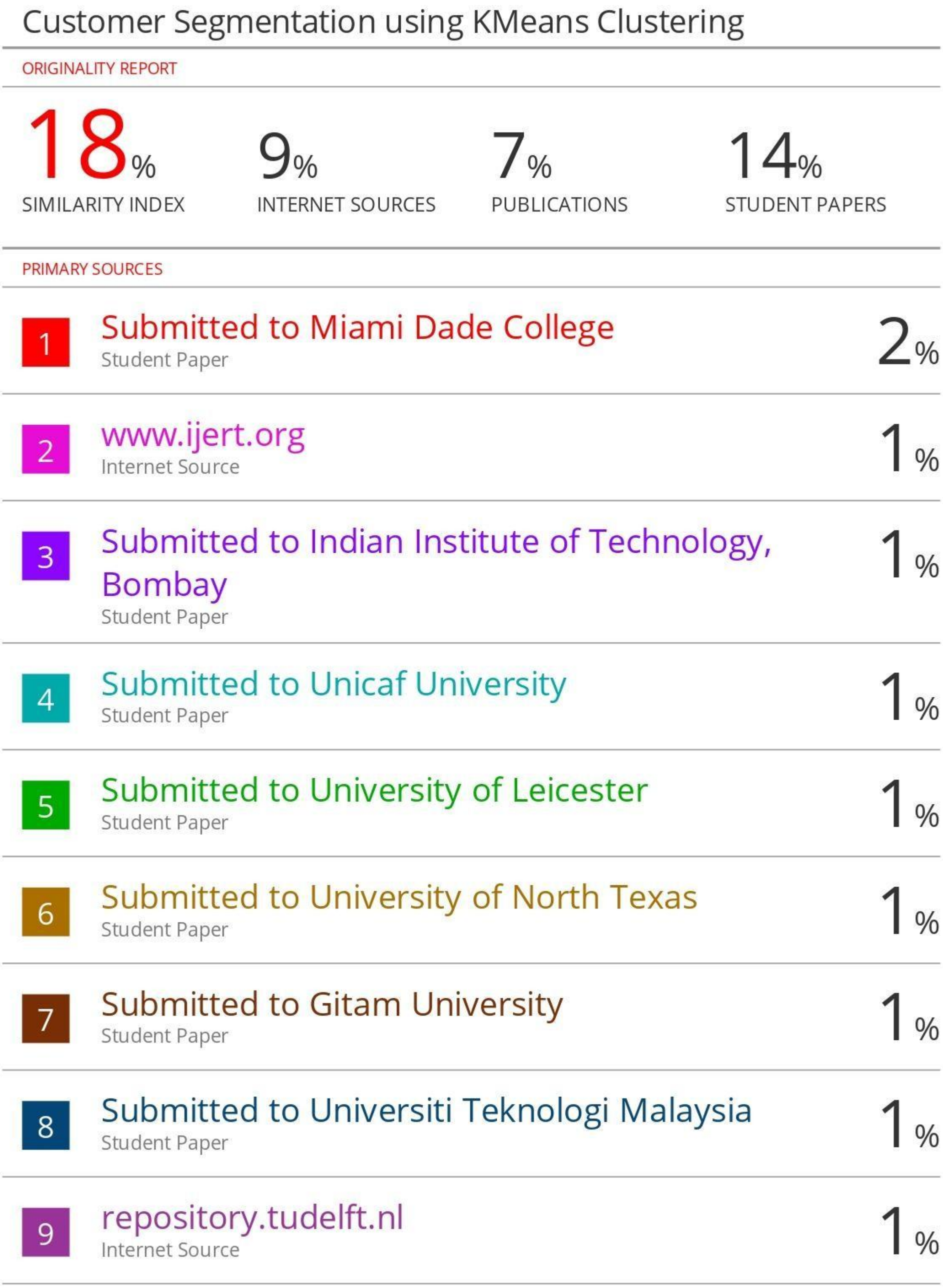
[6] V.Vijilesh, “CUSTOMER SEGMENTATION USING MACHINE LEARNING,” International Research Journal of Engineering and Technology (IRJET), vol. 08, no. 05, May 2021.

[7] Expert Systems with Applications, vol. 100, Feb. 2018, "Retail Business Analytics: Customer Visit Segmentation Using Market Basket Data."

[8] “Cluster analysis.”, Wikipedia.

[9] https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python?resource=download

Plagiarism Report



Customer Segmentation using KMeans Clustering

ORIGINALITY REPORT

18%	9%	7%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Miami Dade College Student Paper	2%
2	www.ijert.org Internet Source	1%
3	Submitted to Indian Institute of Technology, Bombay Student Paper	1%
4	Submitted to Unicaf University Student Paper	1%
5	Submitted to University of Leicester Student Paper	1%
6	Submitted to University of North Texas Student Paper	1%
7	Submitted to Gitam University Student Paper	1%
8	Submitted to Universiti Teknologi Malaysia Student Paper	1%
9	repository.tudelft.nl Internet Source	1%