**“Customer Segmentation and Analysis”**

***A***

***Project Report***

*submitted in partial fulfillment of the*

*requirements for the award of the degree of*

**BACHELOR OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE & ENGINEERING**

**With Specialization in**

**Business analytics and Optimization**

**by**

|  |  |
| --- | --- |
| **Name** | **Roll No.** |
| **Tarun Kumar** | **R103219031** |
| **Divyanshu Chopra** | **R103219005** |
| **Saumya Sinha** | **R103219015** |
| **Shashwat Mishra** | **R103219017** |

***under the guidance of***

**Ms. Shubhi Sharma**

upes-new-logo

**Department of Informatics**

**School of Computer Science**

**University of Petroleum & Energy Studies**

**Bidholi, Via Prem Nagar, Dehradun, UK**

**May – 2022**

upes-new-logo

**CANDIDATE’S DECLARATION**

I/We hereby certify that the project work entitled **“ Customer segmentation and analysis”** in partial fulfillment of the requirements for the award of the Degree of BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE AND ENGINEERING with specialization in <SPECIALIZATION) and submitted to the Department of Informatics at School of Computer Science, University of Petroleum & Energy Studies, Dehradun, is an authentic record of my/ our work carried out during a period from **January** **2022** to **May**, **2022** under the supervision of **Ms. Shubhi Sharma(*Assistant Professor - Senior Scale)***.

The matter presented in this project has not been submitted by me/ us for the award of any other degree of this or any other University.

**R103219005 – Divyanshu Chopra**

**R103219015 – Saumya Sinha**

**R103219017 – Shashwat Mishra**

**R103219031 – Tarun Kumar**

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date: May 13, 2022 **Ms. Shubhi Sharma**

Project Guide

**Dr. Thipendra Pal Singh**

**Professor and H.O.D**

Department of Informatics

School of Computer Science

University of Petroleum & Energy Studies

Dehradun – 248 001 (Uttarakhand)

**ACKNOWLEDGEMENT**

We wish to express our deep gratitude to our guide **Ms. Shubhi Sharma**, for all advice, encouragement and constant support he has given us through out our project work. This work would not have been possible without his support and valuable suggestions.

We sincerely thank to our respected Head of the Department, **Dr. Thipendra Pal Singh,**, for his great support in doing our minor project in **Area (like Network, Big data etc.)** at **SoCS**.

We are also grateful to **Dr. Manish Madan, Dean SoCS**, for giving us the necessary facilities to carry out our project work successfully.

We would like to thank all our friends for their help and constructive criticism during our project work. Finally we have no words to express our sincere gratitude to our **parents** who have shown us this world and for every support they have given us.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Tarun Kumar** | **Saumya Sinha** | **Shashwat Mishra** | **Divyanshu Chopra** |
| **Roll No.** | **R103219031** | **R103219015** | **R103219017** | **R103219005** |

**ABSTRACT**

Despite the fact that segmentation is a hot topic in marketing research and practice, managers often rely on intuition and traditional segmentation strategies based on socio-demographic factors. Separating business and economy passengers is considered common sense in the airline industry. However, the simplicity of this segmentation logic no longer corresponds to customers' increasingly complicated and varied choices. As a result of the airline industry's liberalization, airlines that rely solely on flight class as a segmentation criterion may not be able to customize their product offerings and marketing policies to the appropriate degree in order to respond to the shifting importance and growing complexity of customer choice drivers, such as flexibility and price. Thus, there is a need to re-evaluate the traditional market segmentation criterion. We show that segmenting into business and leisure does not adequately capture the preference heterogeneity among consumers and leads to a misunderstanding of consumer preferences by evaluating the stated preference data of over 5800 airline passengers. We use latent class modelling to analyze our data and offer a new segmentation strategy: we profile the found segments based on behavioral and socio-demographic attributes. We combine our findings with observable consumer traits to create pronounced fence mechanisms for identifying and addressing customer categories who are open to customized product bundles.

**TABLE OF CONTENTS**

**1 Introduction 1**

**2 Related Work 2**

**(**List pros and corn of various algorithm studied in the project**)**

**3 Problem Statement 2**

**4 Objective 2**

**5 Design 3**

(Include Methodology )

(Define the Algorithm of the project)

**6 Implementation 8**

6.1 Pseudocode 8

6.2 Output Screen 15

6.3 Result Analysis 8

**7 Conclusion and Future Scope 9**

**A APPENDIX I PROJECT CODE 11**

**References**

1. **Introduction**

Customer segmentation can significantly impact client management because it allows a company to appeal to distinct groups of consumers with similar demands and focus on what each type of customer requires at any given time. Depending on the company's resources or goals, specific consumer segments can be targeted, whether large or small.

Whatever segmentation models marketers choose, they all require them to create customer groups as a first step in segmenting the customer base. Marketers will usually have a series of levels for each form of segmentation model as a result of this. Marketers can then combine tiers from multiple models to create more defined groupings. Combining the top tier of customers based on an RFM model with a low longevity tier, for example, will result in a sector of highly engaged, freshly acquired customers for marketers.

A customer segmentation study makes use of the massive quantity of data on customers (and potential customers) to accurately identify discrete groups of customers based on demographic, behavioral, and other factors. Because the marketer's goal is to maximize the value (revenue and/or profit) from each client, it's crucial to know how each marketing action will affect the customer ahead of time. Ideally, such "activity-centric" customer segmentation will focus on the long-term customer lifetime value (CLV) impact of a marketing action rather than the short-term value of marketing activities. You can increase customer loyalty by better understanding the customer and thus being able to target them more effectively.  You've enhanced consumer loyalty by interacting with the firm more frequently, even though each basket is smaller.

Customer segmentation is the process of grouping customers together based on common characteristics. These customer segments can help with marketing campaigns, identifying potentially profitable clients, and building customer loyalty. You can develop the correct product, arrange the right distribution and positioning, and match the right sales motion to each consumer after you have all of the segments, while also refining your segments over time.

## When done correctly, it's a model that allows anyone in your firm to understand your clients the right away. All strategic concerns, as well as changes in product, pricing, and packaging over time, cannot be accounted for in segmentation. Instead, it's a continuous effort to identify relevant differences among your clients.

**References:**

[1]Nurma Sari, Juni & Nugroho, Lukito & Ferdiana, Ridi & Santosa, Paulus. (2016). Review on Customer Segmentation Technique on E-commerce. Advanced Science Letters. 22. 3018-3022. 10.1166/asl.2016.7985

[2]Wu, Jing, and Zheng Lin. "Research on customer segmentation model by clustering." Proceedings of the 7th international conference on Electronic commerce. 2005.

[3] <https://archive.ics.uci.edu/ml/datasets/Online+Retail+II>

1. **Related work**

Algorithm used : K- Means clustering

Pros:

* Simple to implement
* Scalable
* Easy to adapt
* Clusters can be of different shapes and sizes

Cons:

* We have to choose K manually
* Sensitive to outliers
* Dependent on initial values
* Trouble clustering data of varying density

1. **Problem statement**

In todays’ world there is loads and loads of data on every customer and every transaction which can ideally help our company in making tremendous improvements in the sales and product improvement department. That’s where customer segmentation comes in and helps companies to club different types of customers from different regions together in a relatable form so that the company can easily identify their customer base and strategize accordingly to improve their sales. Customer segmentation helps in organizing the data to provide insights for companies.

1. **Objective**

Customer segmentation has the ability to help marketers reach out to each customer in the most efficient way possible. Because the marketer's goal is to maximize the value (revenue and/or profit) from each client, it's crucial to know how each marketing action will affect the customer ahead of time. Ideally, such "activity-centric" customer segmentation will focus on the long-term customer lifetime value (CLV) impact of a marketing action rather than the short-term value of a marketing action. As a result, clients must be grouped or segmented based on their CLV.

1. **Design**

**Methodology**

* **Data Gathering:** Online Retail Dataset –II (UCI Machine Learning Repository).

We will sample 10000 rows from the dataset, and we assume that as the whole transactions that the customers do.

For our Second Dataset , we have used 200 entries from Kaggle. This Shopping Mall Dataset collection includes transactions made by Customers with their Monthly Income and Spending Score.

* **Creation of RFM Table:** Recency, Frequency & Monetary Table.

Recency: When a buyer purchases a product for the last time.

Frequency: How frequent the customer buy the product.

Monetary: How much the customer pays for the product.

* **Data Transformation:** Transform data symmetrically, Remove Skewness & Perform Scaling.
* **K-Means Clustering:** Geometrical, Unsupervised Learning Algorithm.

We calculate the distance to each centroid by determining each centroid.

Each data point belongs to a centroid if its distance from the others is the shortest.

It repeats until the following distance total has no significant differences from the previous one.

* **Data Visualization & Detailed Analysis:** Cluster Analysis & Interactive Dashboards along with detailed result analysis.

**Algorithm**

Data pre-processing for K-Means Clustering.

- Importing the required libraries

- detecting and handling outliers

\* Exploratory Data Analysis and Visualizations of the dataset.

- univariate analysis

- bivariate analysis

- multivariate analysis

\* Building a K-Means Clustering model from scratch.

- Standardizing the variables

- building the clustering model (K-means)

\* Visualizations, interpretations, and analysis of the clusters built.

- scatter plots

- heatmaps

- elbow method graph

**Final code:**

**Program 1:**

!pip install xlrd

import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

from google.colab import files

uploaded = files.upload()

df = pd.read\_excel('online\_retail\_II.xlsx')

df = df[df['Customer ID'].notna()]

df\_fix = df.sample(10000, random\_state = 42)

from datetime import datetime

df\_fix['InvoiceDate'] = pd.to\_datetime(df.InvoiceDate, format='%d-%m-%y %H:%M:%S')

df\_fix["TotalSum"] = df\_fix["Quantity"] \* df\_fix["Price"]

import datetime

snapshot\_date = max(df\_fix.InvoiceDate) + datetime.timedelta(days=1)

customers = df\_fix.groupby(['Customer ID']).agg({

    'InvoiceDate': lambda x: (snapshot\_date - x.max()).days,

    'Invoice': 'count',

    'TotalSum': 'sum'})

customers.rename(columns = {'InvoiceDate': 'Recency',

                            'Invoice': 'Frequency',

                            'TotalSum': 'MonetaryValue'}, inplace=True)

customers

from scipy import stats

customers\_fix = pd.DataFrame()

customers\_fix["Recency"] = stats.boxcox(customers['Recency'])[0]

customers\_fix["Frequency"] = stats.boxcox(customers['Frequency'])[0]

customers\_fix["MonetaryValue"] = pd.Series(np.cbrt(customers['MonetaryValue'])).values

customers\_fix.tail()

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

scaler.fit(customers\_fix)

customers\_normalized = scaler.transform(customers\_fix)

print(customers\_normalized.mean(axis = 0).round(2))

print(customers\_normalized.std(axis = 0).round(2))

df1 = pd.DataFrame(customers\_normalized)

df1

**Modelling**

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.cluster import KMeans

sse = {}

for k in range(1, 11):

    kmeans = KMeans(n\_clusters=k, random\_state=42)

    kmeans.fit(customers\_normalized)

    sse[k] = kmeans.inertia\_

plt.title('The Elbow Method')

plt.xlabel('k')

plt.ylabel('SSE')

sns.pointplot(x=list(sse.keys()), y=list(sse.values()))

plt.show()

model = KMeans(n\_clusters=3, random\_state=42)

model.fit(customers\_normalized)

model.labels\_.shape

customers["Cluster"] = model.labels\_

customers.groupby('Cluster').agg({

    'Recency':'mean',

    'Frequency':'mean',

    'MonetaryValue':['mean', 'count']}).round(2)

df\_normalized = pd.DataFrame(customers\_normalized, columns=['Recency', 'Frequency', 'MonetaryValue'])

df\_normalized['ID'] = customers.index

df\_normalized['Cluster'] = model.labels\_

df\_nor\_melt = pd.melt(df\_normalized.reset\_index(),

                      id\_vars=['ID', 'Cluster'],

                      value\_vars=['Recency','Frequency','MonetaryValue'],

                      var\_name='Attribute',

                      value\_name='Value')

df\_nor\_melt.head()

sns.lineplot('Attribute', 'Value', hue='Cluster', data=df\_nor\_melt)

**Program 2:**

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.cluster import KMeans

import warnings

warnings.filterwarnings('ignore')

from google.colab import files

uploaded = files.upload()

dframe = pd.read\_excel("Dataset.xlsx")

dframe.head()

dframe.describe()

sns.distplot(dframe['Monthly\_Income\_Rs']);

dframe.columns

columns = ['Age', 'Monthly\_Income\_Rs','Spending\_Score']

for i in columns:

    plt.figure()

    sns.distplot(dframe[i])

sns.kdeplot(dframe['Monthly\_Income\_Rs'],shade=True,hue=dframe['Gender']);

columns = ['Age', 'Monthly\_Income\_Rs','Spending\_Score']

for i in columns:

    plt.figure()

    sns.kdeplot(dframe[i],shade=True,hue=dframe['Gender'])

columns = ['Age', 'Monthly\_Income\_Rs','Spending\_Score']

for i in columns:

    plt.figure()

    sns.boxplot(data=dframe,x='Gender',y=dframe[i])

dframe['Gender'].value\_counts(normalize=True)

sns.scatterplot(data=dframe, x='Monthly\_Income\_Rs',y='Spending\_Score' )

sns.pairplot(dframe,hue='Gender')

dframe.groupby(['Gender'])['Age', 'Monthly\_Income\_Rs',

       'Spending\_Score'].mean()

dframe.corr()

sns.heatmap(dframe.corr(),annot=True,cmap='coolwarm')

sns.scatterplot(data=dframe, x='Monthly\_Income\_Rs',y='Spending\_Score' )

sns.pairplot(dframe,hue='Gender')

dframe.groupby(['Gender'])['Age', 'Monthly\_Income\_Rs',

       'Spending\_Score'].mean()

dframe.corr()

sns.heatmap(dframe.corr(),annot=True,cmap='coolwarm')

clustering1 = KMeans(n\_clusters=3)

clustering1.fit(dframe[['Monthly\_Income\_Rs']])

clustering1.labels\_

dframe['Income Cluster'] = clustering1.labels\_

dframe.head()

dframe['Income Cluster'].value\_counts()

clustering1.inertia\_

intertia\_scores=[]

for i in range(1,11):

    kmeans=KMeans(n\_clusters=i)

    kmeans.fit(dframe[['Monthly\_Income\_Rs']])

    intertia\_scores.append(kmeans.inertia\_)

intertia\_scores

plt.plot(range(1,11),intertia\_scores)

dframe.columns

dframe.groupby('Income Cluster')['Age', 'Monthly\_Income\_Rs',

       'Spending\_Score'].mean()

clustering2 = KMeans(n\_clusters=5)

clustering2.fit(dframe[['Monthly\_Income\_Rs','Spending\_Score']])

dframe['Spending and Income Cluster'] =clustering2.labels\_

dframe.head()

intertia\_scores2=[]

for i in range(1,11):

    kmeans2=KMeans(n\_clusters=i)

    kmeans2.fit(dframe[['Monthly\_Income\_Rs','Spending\_Score']])

    intertia\_scores2.append(kmeans2.inertia\_)

plt.plot(range(1,11),intertia\_scores2)

centers =pd.DataFrame(clustering2.cluster\_centers\_)

centers.columns = ['x','y']

plt.figure(figsize=(10,8))

plt.scatter(x=centers['x'],y=centers['y'],s=100,c='black',marker='\*')

sns.scatterplot(data=dframe, x ='Monthly\_Income\_Rs',y='Spending\_Score',hue='Spending and Income Cluster',palette='tab10')

plt.savefig('clustering\_bivaraiate.png')

pd.crosstab(dframe['Spending and Income Cluster'],dframe['Gender'],normalize='index')

dframe.groupby('Spending and Income Cluster')['Age', 'Monthly\_Income\_Rs',

       'Spending\_Score'].mean()

from sklearn.preprocessing import StandardScaler

scale = StandardScaler()

dframe.head()

dff = pd.get\_dummies(dframe,drop\_first=True)

dff.head()

dff.columns

dff = dff[['Age', 'Monthly\_Income\_Rs', 'Spending\_Score','Gender\_Male']]

dff.head()

dff = scale.fit\_transform(dff)

dff = pd.DataFrame(scale.fit\_transform(dff))

dff.head()

intertia\_scores3=[]

for i in range(1,11):

    kmeans3=KMeans(n\_clusters=i)

    kmeans3.fit(dff)

    intertia\_scores3.append(kmeans3.inertia\_)

plt.plot(range(1,11),intertia\_scores3)

dframe

dframe.to\_csv('Clustering.csv')

1. **Implementation**
   1. **Pseudocode**

1. Choose the number of clusters(K) and obtain the data points

2. Place the centroids c\_1, c\_2, ..... c\_k randomly

3. Repeat steps 4 and 5 until convergence or until the end of a fixed number of iterations

4. for each data point x\_i:

- find the nearest centroid(c\_1, c\_2 .. c\_k)

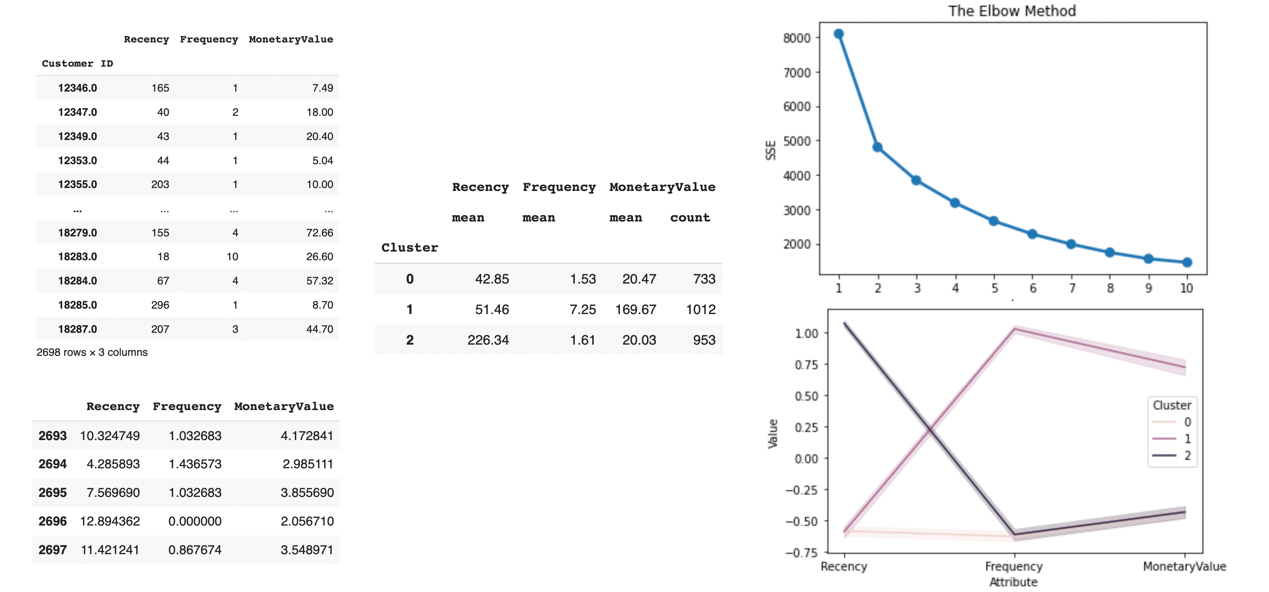
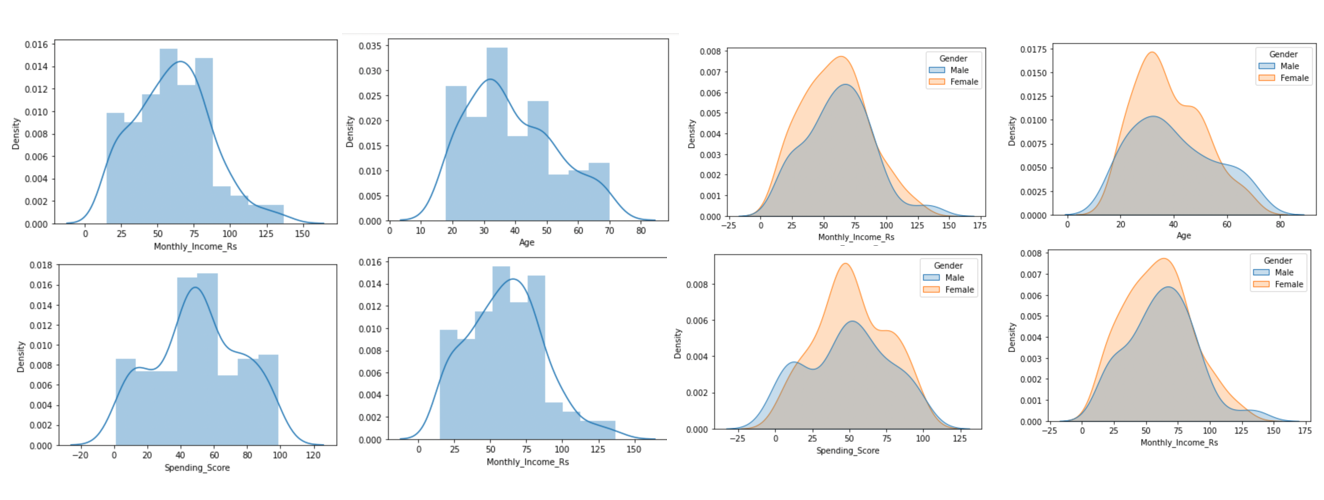
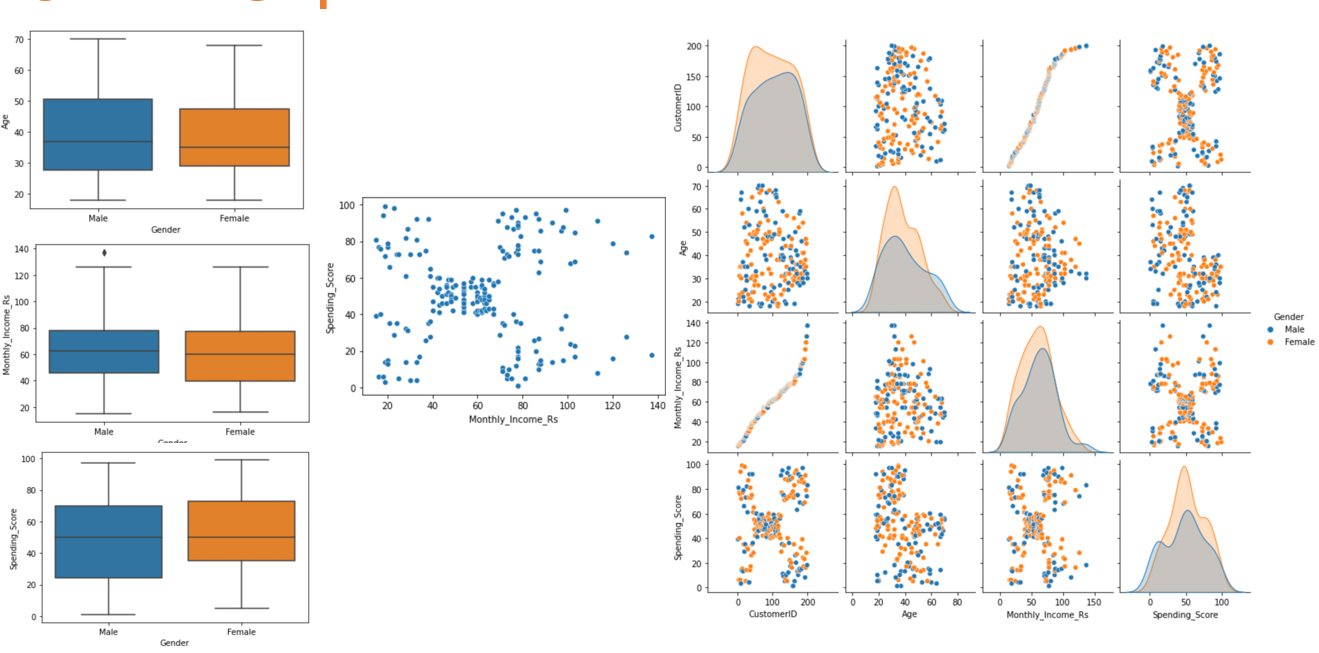
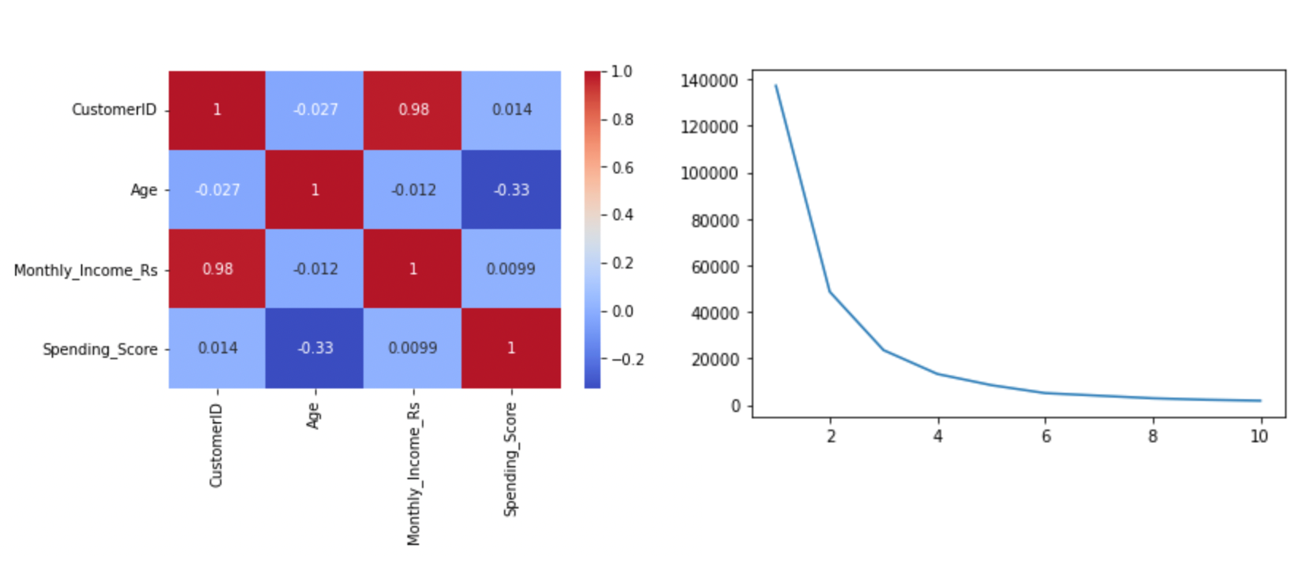
- assign the point to that cluster

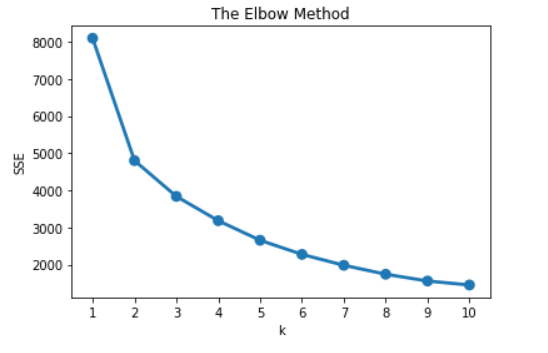
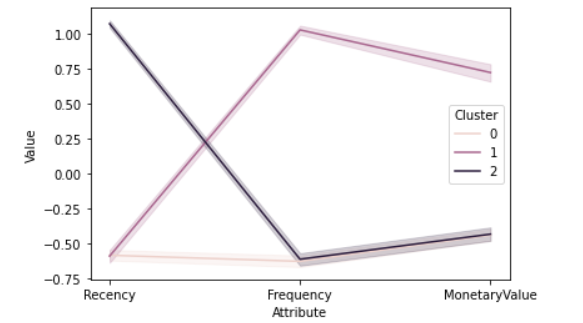
5. for each cluster j = 1..k

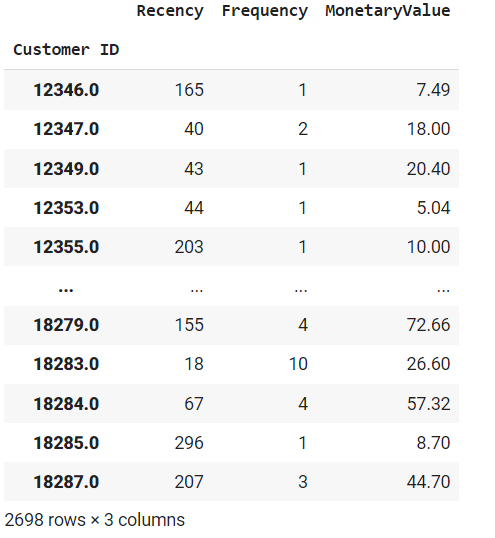
- new centroid = mean of all points assigned to that cluster

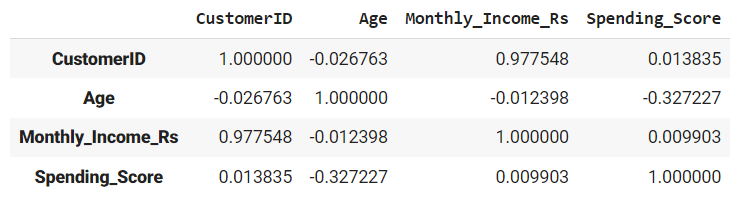
6. End

* 1. **Output screen**

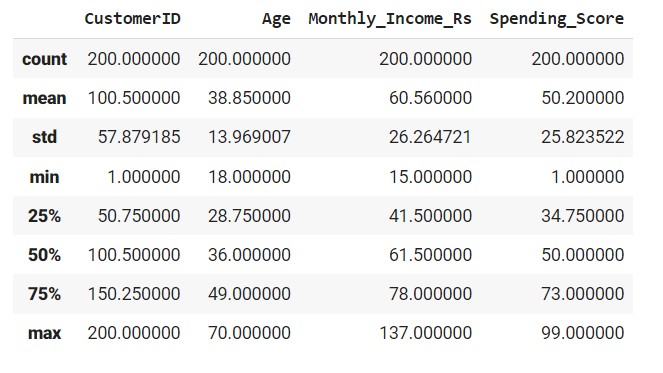
**** ****  

**** ****

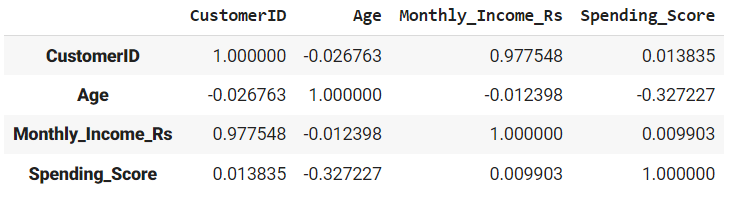
**** Data Table

****

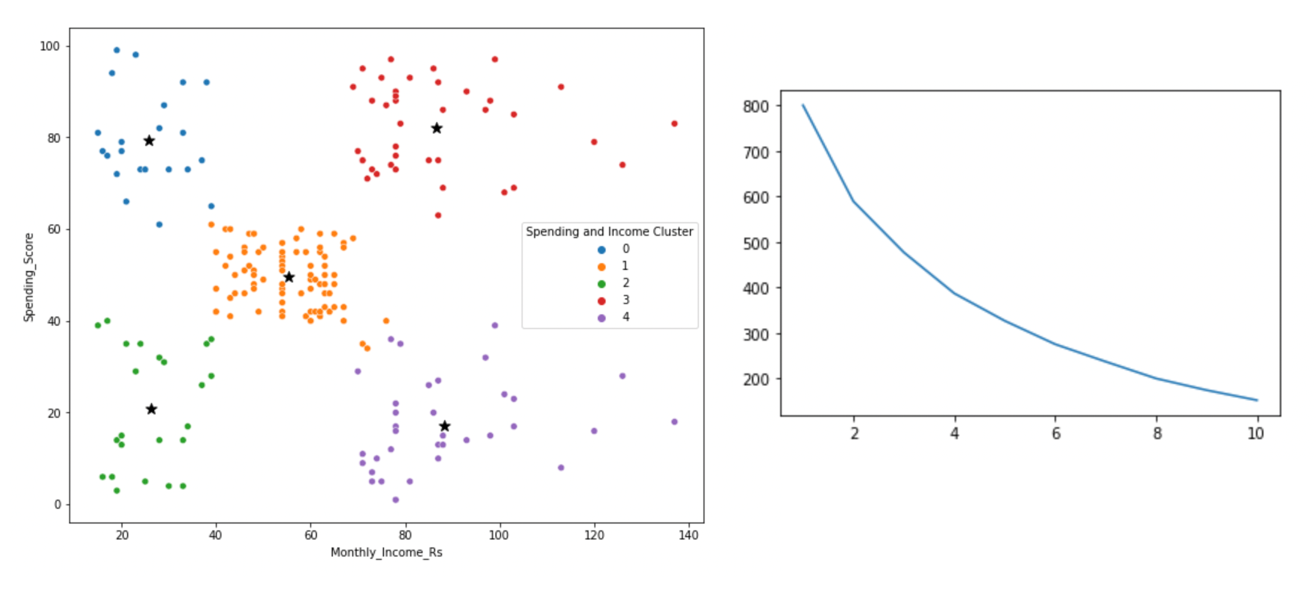
Error

****

**Description**

****

**Correlation**

* 1. **Result analysis**

1. **Conclusion**

K means clustering is quite possibly the most famous clustering calculations and normally the primary thing practitioner apply while addressing clustering task to find out about the structure of the dataset. The objective of K means is to group informative items into particular non-overlapping subgroups. One of the significant utilization of K means clustering is division of clients to get a superior comprehension of them which could be utilized to build the income of the organization.

The general population and customer population have been thought about and segmented utilizing an Unsupervised learning algorithm . We had the option to figure out which clustering have more customers and which are likely groups to have probable customers. We have likewise utilized supervised learning to foresee a potential future customer in view of segment information.

1. **Future scope**

Customer segmentation is only effective if it is done collaboratively from the beginning. My team's success hinged on Marketing, Sales, Product, Finance, and Analytics buy-in, as segmentation evolved from a business need to a functional tool. After all, different client needs imply different features to design, varied distribution and positioning in the market, and different sales movements – all of which are explained as a result of segmentation.

Without extensive study, stakeholder feedback, and iteration, getting segmentation right is unlikely. Finally, alignment and adoption must go hand in hand so that all teams may benefit from this new framework's understanding of the business.