Department of Computer Science & Engineering

Final Year B. Tech. (CSE) – I : 2023-24

5CS462 : PE5 - Data Mining Lab

Assignment No. 9: Mini Project

Team Members:
2020BTECS00101 Kranti Subhash Bharti
2020BTECS00092 Vishal Shrirang Madle

Problem Statement ID: 14
Problem Statement: **Mining of customer behaviour of any retail shop.**

Customer buying behaviour Analysis is a detailed analysis of a company’s ideal customers. It helps a business to better understand its customers. It makes it easier for them to modify products according to the specific needs, behaviours and concerns of different types of customers.

This analysis helps a business modify its product based on its target customers from different customer segments. For example, instead of spending money to market a new product to every customer in the company’s database, a company can analyse which customer segment is most likely to buy the product and then market the product only on that particular segment.

Dataset Used: marketing_campaign.csv for customer-personality-analysis
Dataset link: dataset

Attributes Types

- People
- Products
- Promotions
- Place

Target

Need to perform clustering to summarize customer segments to analyse the behaviour.

Implementation: Notebook

Analysis Steps:

1) Import required libraries

2) import dataset

3) Data Cleaning

4) Checking the correlation between the attributes

5) Data Preprocessing

6) Visualization

7) Analyse the Correlations and Distributions

- analysis of the correlation between marital status and expenses with respect to education
- analysis of the correlation between marital status and expenses
- distribution of expenses with respect to marital status
- distribution of expenses with respect to education
- distribution of number of total expenses with respect to education
- distribution of age with respect to marital status
- distribution of income with respect to marital status
- analysis of the distribution of people according to marital status
- analysis of the distribution of people according to education
- distribution of expenses based on education

- income based on education level

8) Kmean clustering

9) Final Observation

OBSERVATIONS: Based on information in implementation we can divide the customer into 2 parts:-

Highly Active Customer:- These customers belong to cluster one.

Least Active Customer:- These customers belong to cluster two.

1. Characteristics of Highly Active Customer

- In terms of Education

  Highly Active Customers are from PG background
- In terms of Marital_status

  The number of people in the relationship is approx. two times of single people
- In terms of Income

  Income of Highly active customer are little less as compare to least active customer.
- In terms of Kids

  Highly active customer have more number of children as compare to other customer (avg. of 1 child).
- In terms of Expenses

  Expenses of Highly Active customer are less as compare to least.

  These customer spent avg. of approx. 100-200 unit money.
- In terms of Age

  Age of these customer are between 25 to 75.

Maximum customer age are between 40 to 50.

- In terms of day_engaged

  Highly Active customer are more loyal as they engaged with company for longer period of time.

2. Characteristics of Least Active Customer

- In terms of Education

  Least Active Customer are from UG backgroud
- In terms of Marital_status

  Number of people in relationship are approx. equal to single people
- In terms of Income

  Income of Least active customer are very less or say negligible.
- In terms of Kids

  Only few of these customer have child.
- In terms of Expenses

  Expenses of Least Active customer are very less or say negligible.
- In terms of Age

  Age of these customer are between 15 to 30.
- In terms of day_engaged

  Least Active customer are not much enrolled with company for longer time

Implementation Screen Shots:

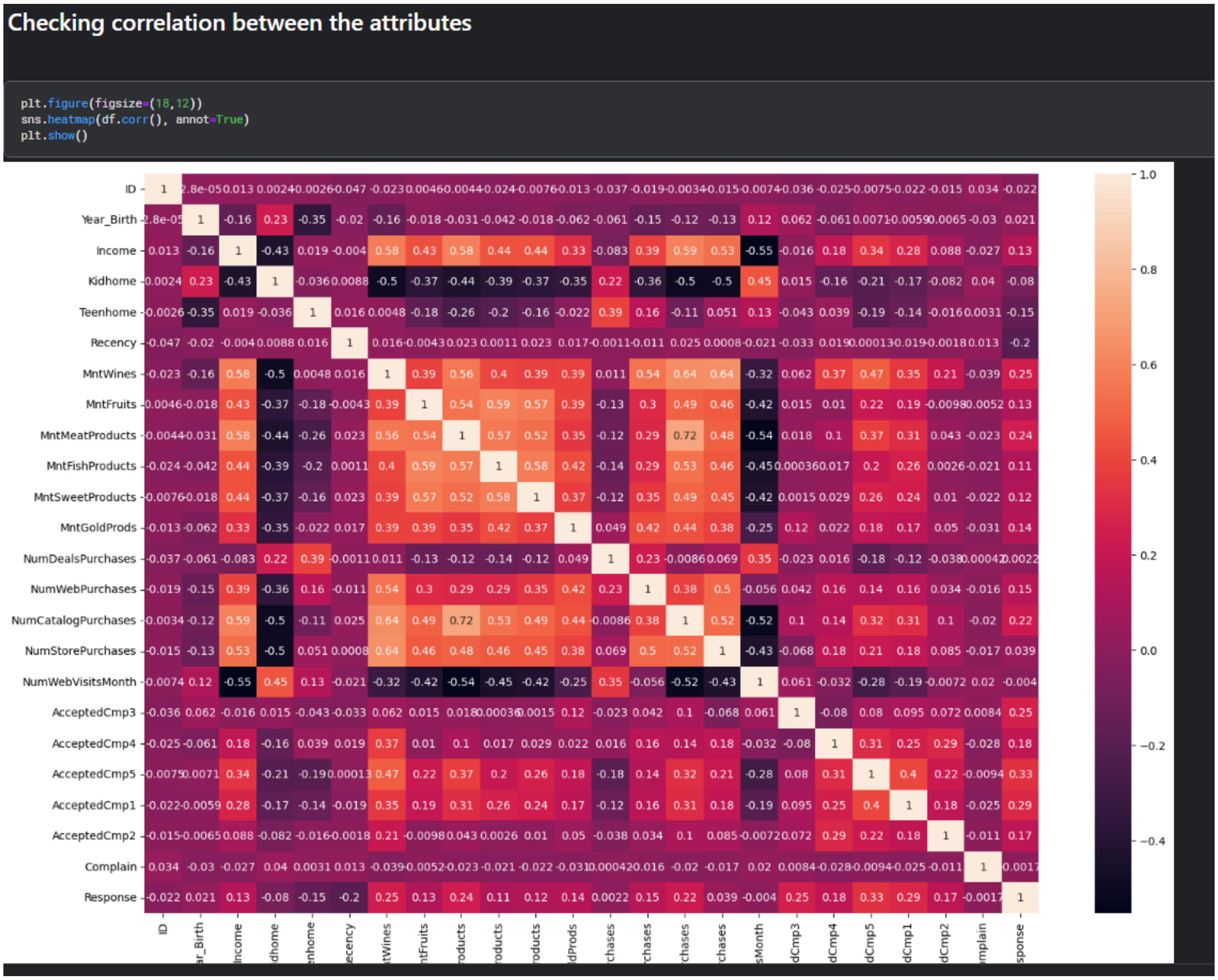
IMPORTING DATASET

[2]:

```
# Import the Dataset
df = pd.read_csv("../input/customer-personality-analysis/marketing_campaign.csv", sep="\t")
df.head()
```

[2]:

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumW
0	5524	1957	Graduation	Single	58138.0	0	0	04-09-2012	58	635	...	7
1	2174	1954	Graduation	Single	46344.0	1	1	08-03-2014	38	11	...	5
2	4141	1965	Graduation	Together	71613.0	0	0	21-08-2013	26	426	...	4
3	6182	1984	Graduation	Together	26646.0	1	0	10-02-2014	26	11	...	6
4	5324	1981	PhD	Married	58293.0	1	0	19-01-2014	94	173	...	5



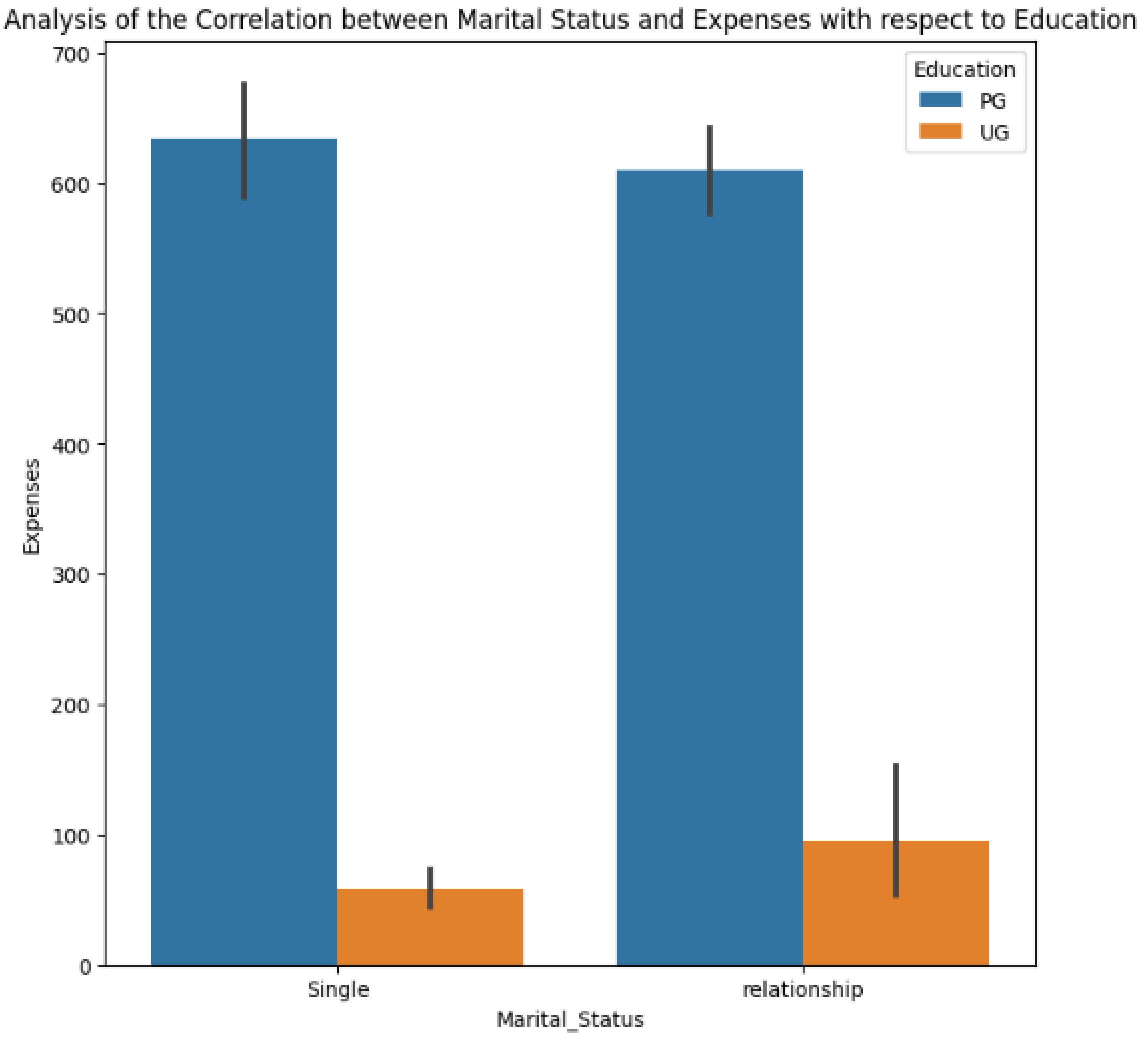
Checking correlation between the attributes

```
plt.figure(figsize=(18,12))
sns.heatmap(df.corr(), annot=True)
plt.show()
```

ANALYSIS OF THE CORRELATION BETWEEN MARITAL STATUS AND EXPENSES WITH RESPECT TO EDUCATION

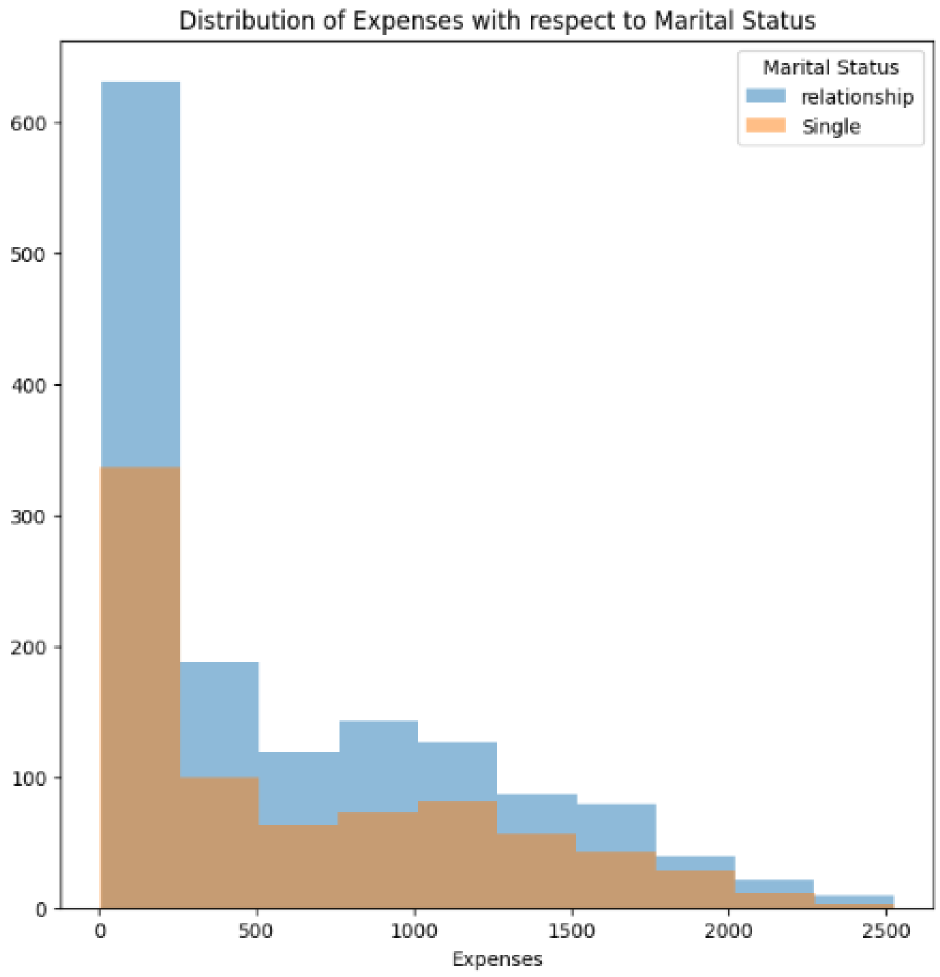
8]:

```
plt.figure(figsize=(8,8))
sns.barplot(x=df['Marital_Status'], y=df['Expenses'], hue = df["Education"])
plt.title("Analysis of the Correlation between Marital Status and Expenses with respect to E
n")
plt.show()
```



DISTRIBUTION OF EXPENSES WITH RESPECT TO MARITAL STATUS

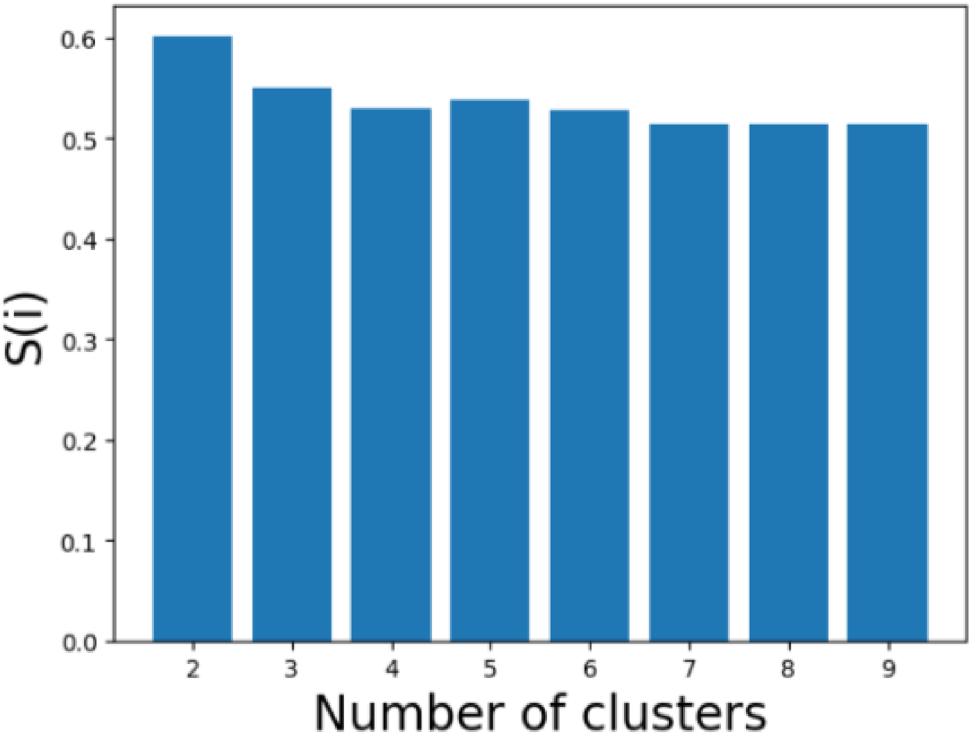
```
plt.figure(figsize=(8,8))
plt.hist("Expenses", data = df[df["Marital_Status"] == "relationship"], alpha = 0.5, label = "relati
onship")
plt.hist("Expenses", data = df[df["Marital_Status"] == "Single"], alpha = 0.5, label = "Single")
plt.title("Distribution of Expenses with respect to Marital Status")
plt.xlabel("Expenses")
plt.legend(title = "Marital Status")
plt.show()
```

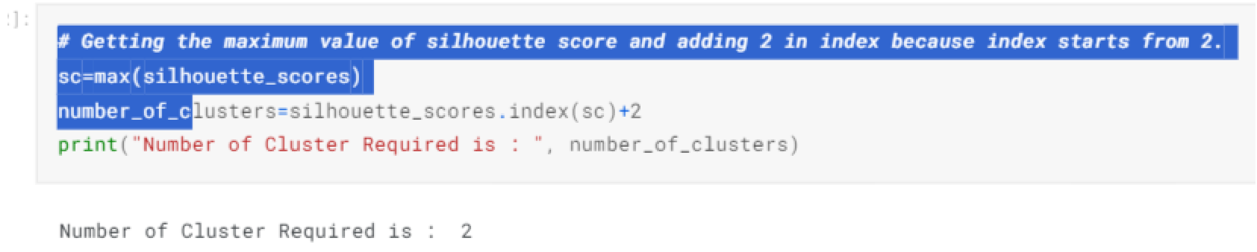


Silhouette Score

```python
from sklearn.metrics import silhouette_score
```

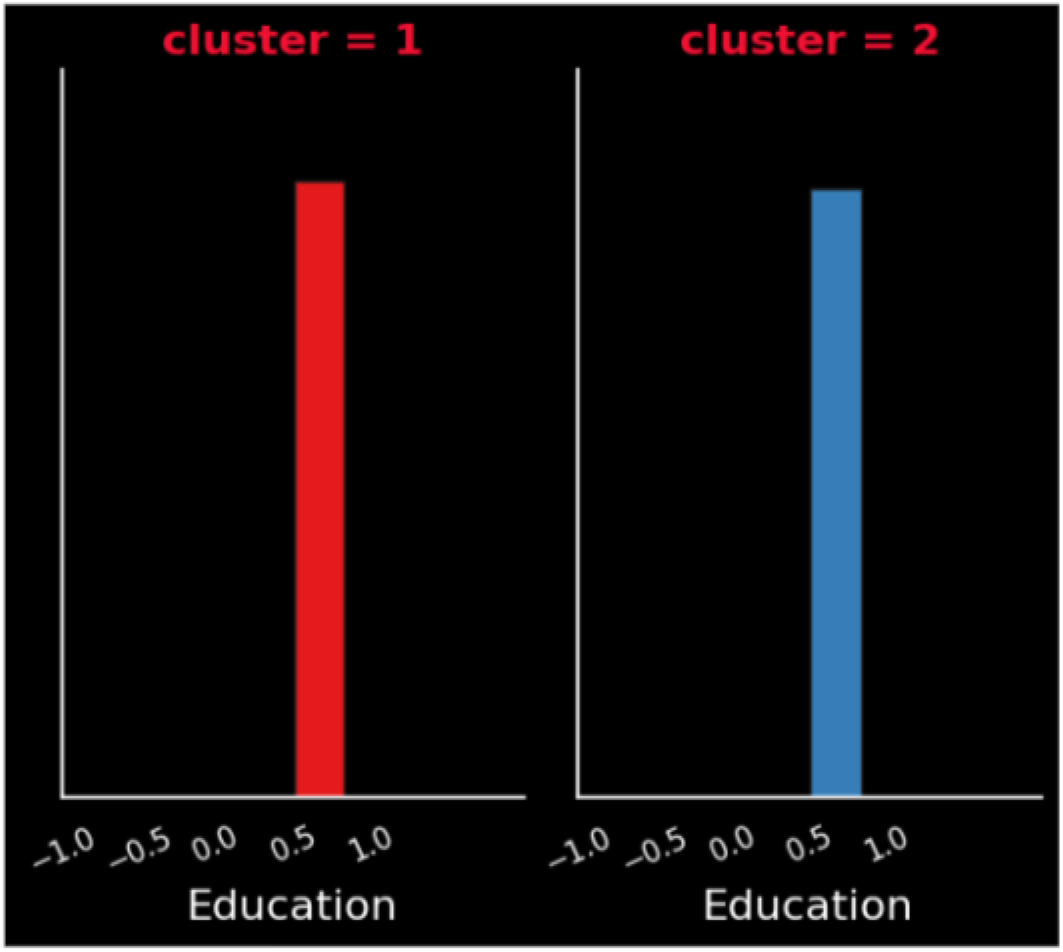
```python
silhouette_scores = []
for i in range(2,10):
    m1=KMeans(n_clusters=i, random_state=42)
    c = m1.fit_predict(df1)
    silhouette_scores.append(silhouette_score(df1, m1.fit_predict(df1)))
plt.bar(range(2,10), silhouette_scores)
plt.xlabel('Number of clusters', fontsize = 20)
plt.ylabel('S(i)', fontsize = 20)
plt.show()
```

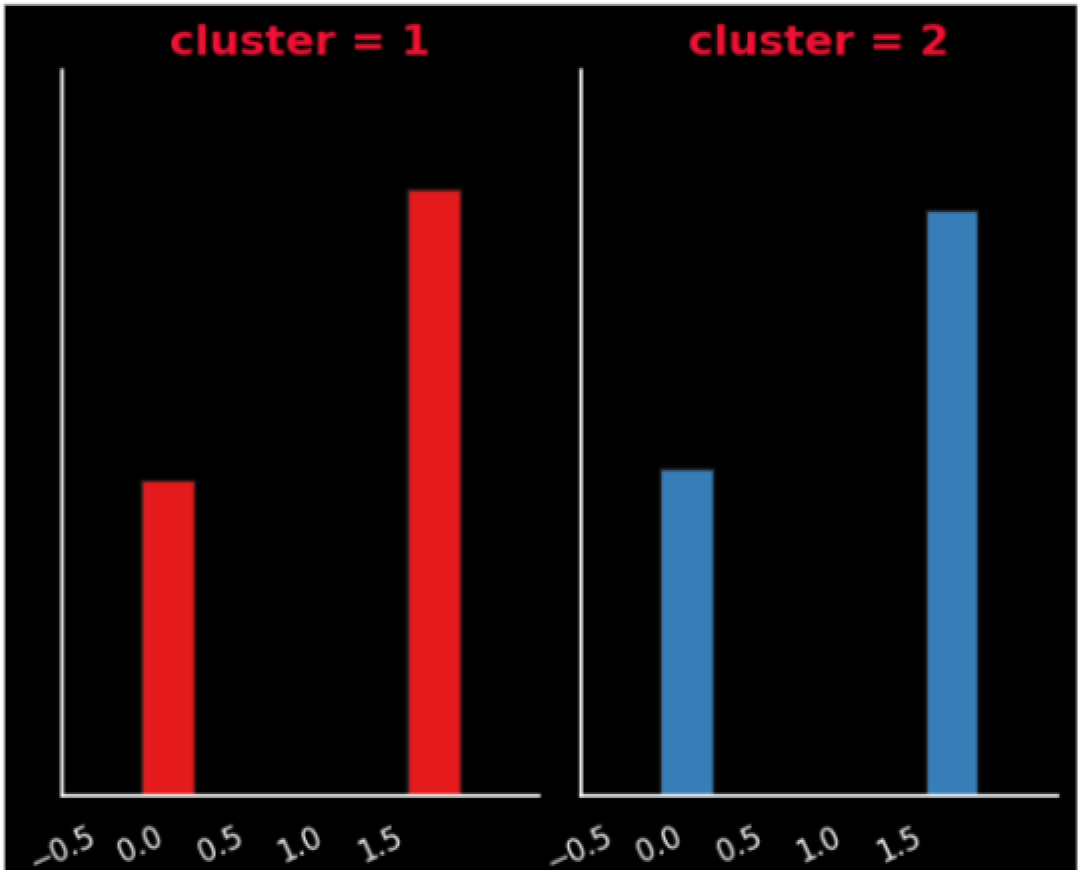


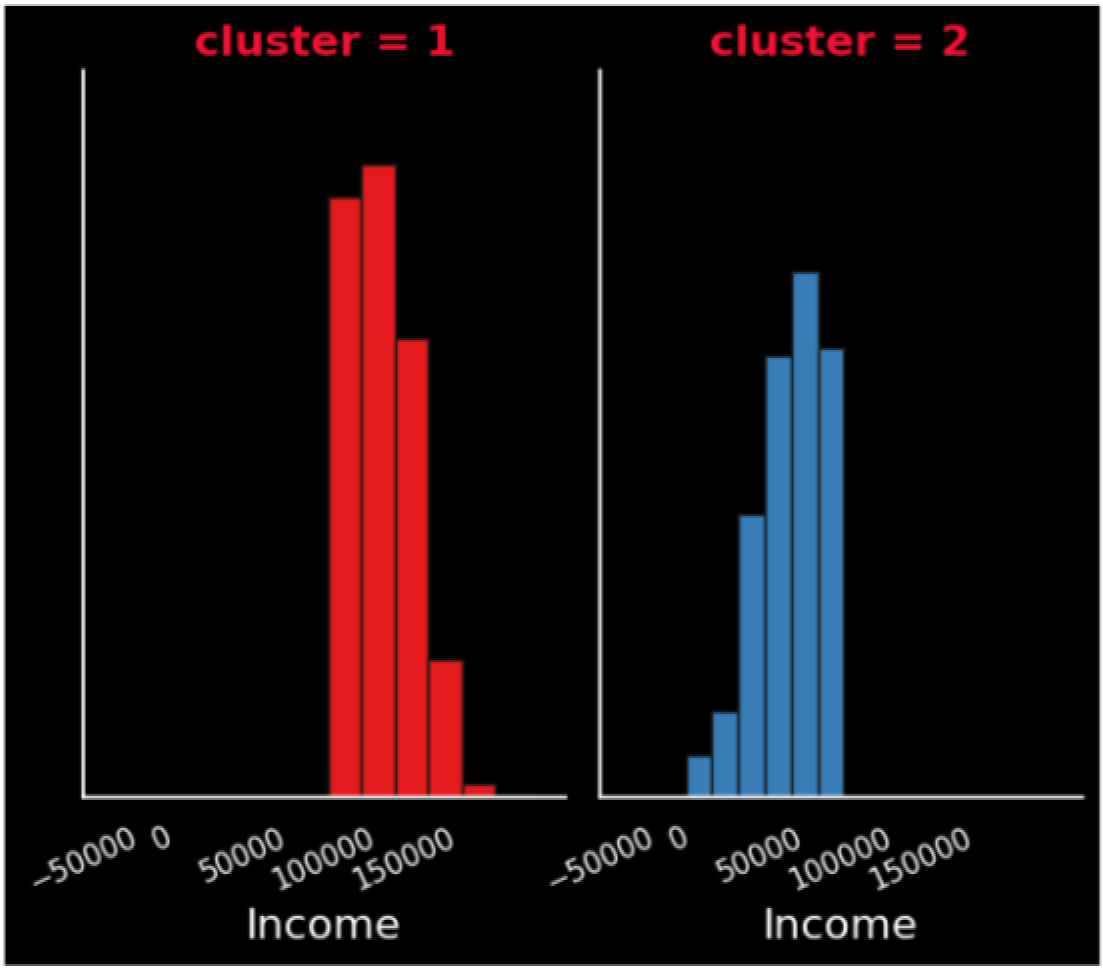


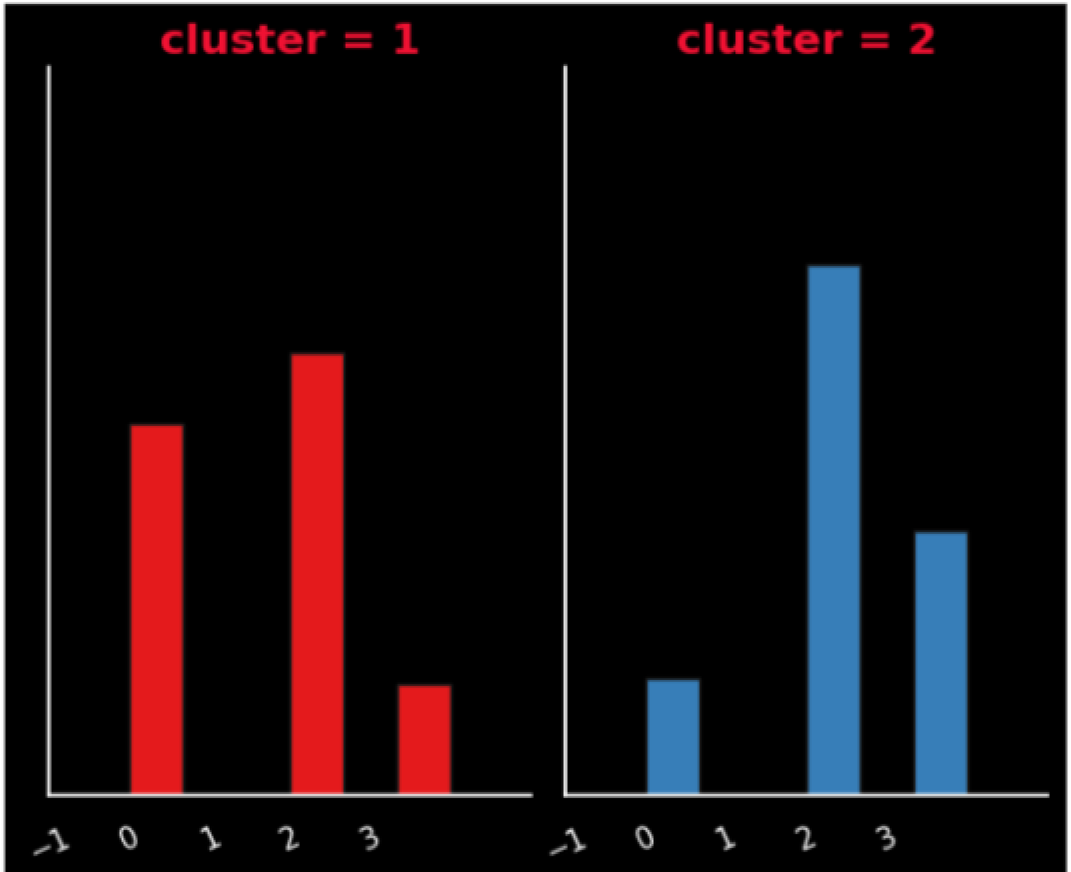
```python
# Getting the maximum value of silhouette score and adding 2 in index because index starts from 2.
sc=max(silhouette_scores)
number_of_clusters=silhouette_scores.index(sc)+2
print("Number of Cluster Required is : ", number_of_clusters)
```

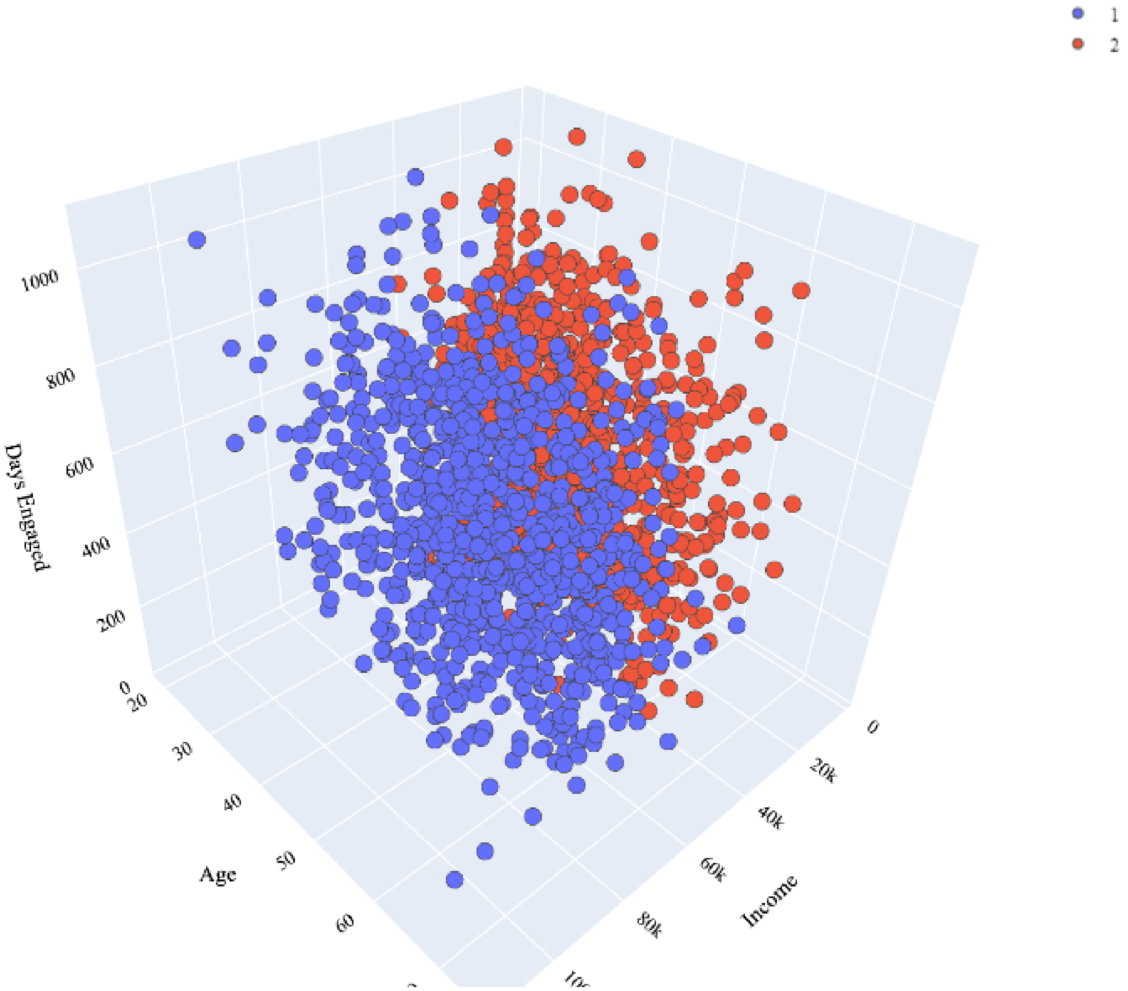
```
Number of Cluster Required is :  2
```


cluster = 1
cluster = 2
−1.0
−0.5
0.0
0.5
1.0
Education
−1.0
−0.5
0.0
0.5
1.0
Education


cluster = 1
cluster = 2
−0.5
0.0
0.5
1.0
1.5
−0.5
0.0
0.5
1.0
1.5


cluster = 1
cluster = 2
−50000
0
50000
100000
150000
Income
−50000
0
50000
100000
150000
Income


cluster = 1
cluster = 2
−1
0
1
2
3
−1
0
1
2
3


1
2
Days Engaged
0
200
400
600
800
1000
Age
20
30
40
50
60
Income
0
20k
40k
60k
80k

1.Characteristics of Highly Active Customer

In terms of Education

Highly Active Customer are from PG background

In terms of Marital_status

Number of people in relationship are approx. two times of single people

In terms of Income

Income of Highly active customer are little less as compare to least active customer.

In terms of Kids

Highly active customer have more number of children as compare to other customer (avg. of 1 child).

In terms of Expenses

Expenses of Highly Active customer are less as compare to least. These customer spent avg. of approx. 100-2

In terms of Age

Age of these customer are between 25 to 75. Maximum customer age are between 40 to 50.

In terms of day_engaged

Highly Active customer are more loyal as they engaged with company for longer period of time.

2.Characteristics of Least Active Customer

In terms of Education

Least Active Customer are from UG backgroud

In terms of Marital_status

Number of people in relationship are approx. equal to single people

In terms of Income

Income of Least active customer are very less or say negligible.

In terms of Kids

Only few of these customer have child.

In terms of Expenses

Expenses of Least Active customer are very less or say negligible.

In terms of Age

Age of these customer are between 15 to 30.

In terms of day_engaged

Least Active customer are not much enrolled with company for longer time