# DATA MINING

# PROJECT REPORT

## Contents:

Part 1 - Clustering: Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.

Part 1 - Clustering: Treat missing values in CPC, CTR and CPM using the formula given.

Part 1 - Clustering: Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an analyst your judgement may be different from another analyst).

Part 1 - Clustering: Perform z-score scaling and discuss how it affects the speed of the algorithm.

Part 1 - Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance, and identify optimum number of clusters

Part 1 - Clustering: Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.

Part 1 - Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters.

Part 1 - Clustering: Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots].

Part 1 - Clustering: Conclude the project by providing summary of your learnings.

PCA: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.

Part 2 - PCA: Perform detailed exploratory analysis by creating certain questions like (i) which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio? (Example Questions). Pick 5 variables out of the given 24 variables below for EDA: No_HH, TOT_M, TOT_F, M_06, F_06, M_SC, F_SC, M_ST, F_ST, M_LIT, F_LIT, M_ILL, F_ILL, TOT_WORK_M, TOT_WORK_F, MAINWORK_M, MAINWORK_F, MAIN_CL_M, MAIN_CL_F, MAIN_AL_M, MAIN_AL_F, MAIN_HH_M, MAIN_HH_F, MAIN_OT_M, MAIN_OT_F

Part 2 - PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?

Part 2 - PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.

Part 2 - PCA: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.

Part 2 - PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.

Part 2 - PCA: Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.

Part 2 - PCA: Write linear equation for first PC.

# Part 1:

**Clustering: Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.**

**Ads_Data. Head ( ): -**

| Timestamp | Inventory Type | Ad-Length | Ad-Width | Ad Size | Ad Type | Platform | Device Type | Format |
|---|---|---|---|---|---|---|---|---|
| 2020-9-2-17 | Format1 | 300 | 250 | 75000 | Inter222 | Video | Desktop | Display |
| 2020-9-2-10 | Format1 | 300 | 250 | 75000 | Inter227 | App | Mobile | Video |
| 2020-9-1-22 | Format1 | 300 | 250 | 75000 | Inter222 | Video | Desktop | Display |
| 2020-9-3-20 | Format1 | 300 | 250 | 75000 | Inter228 | Video | Mobile | Video |
| 2020-9-4-15 | Format1 | 300 | 250 | 75000 | Inter217 | Web | Desktop | Video |

| Available_ Impressions | Matched_Queries | Impressions | Clicks | Spend | Fee | Revenue | CTR | CPM | CPC |
|---|---|---|---|---|---|---|---|---|---|
| 1806 | 325 | 323 | 1 | 0 | 0.35 | 0 | 0.0031 | 0 | 0 |
| 1780 | 285 | 285 | 1 | 0 | 0.35 | 0 | 0.0035 | 0 | 0 |
| 2727 | 356 | 355 | 1 | 0 | 0.35 | 0 | 0.0028 | 0 | 0 |
| 2430 | 497 | 495 | 1 | 0 | 0.35 | 0 | 0.002 | 0 | 0 |
| 1218 | 242 | 242 | 1 | 0 | 0.35 | 0 | 0.0041 | 0 | 0 |

**Ads_Data.tail() :**

| Timestamp | InventoryType | Ad-Length | Ad-Width | Ad Size | Ad Type | Platform | Device Type | Format |
|---|---|---|---|---|---|---|---|---|
| 2020-9-13-7 | Format5 | 720 | 300 | 216000 | Inter220 | Web | Mobile | Video |
| 2020-11-2-7 | Format5 | 720 | 300 | 216000 | Inter224 | Web | Desktop | Video |
| 2020-9-14-22 | Format5 | 720 | 300 | 216000 | Inter218 | App | Mobile | Video |
| 2020-11-18-2 | Format4 | 120 | 600 | 72000 | inter230 | Video | Mobile | Video |
| 2020-9-14-0 | Format5 | 720 | 300 | 216000 | Inter221 | App | Mobile | Video |

| Available _Impressions | Matched_Queries | Impressions | Clicks | Spend | Fee | Revenue | CTR | CPM | CPC |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0.07 | 0.35 | 0.0455 | Nan | Nan | Nan |
| 3 | 2 | 2 | 1 | 0.04 | 0.35 | 0.026 | Nan | Nan | Nan |
| 2 | 1 | 1 | 1 | 0.05 | 0.35 | 0.0325 | Nan | Nan | Nan |
| 7 | 1 | 1 | 1 | 0.07 | 0.35 | 0.0455 | Nan | Nan | Nan |
| 2 | 2 | 2 | 1 | 0.09 | 0.35 | 0.0585 | Nan | Nan | Nan |

**Ads_Data.info ( )**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23066 entries, 0 to 23065
Data columns (total 19 columns):
 #   Column                 Non-Null Count Dtype
---  ------                 -------------  -----
 0   Timestamp              23066 non-null object
 1   InventoryType          23066 non-null  object
 2   Ad - Length            23066 non-null  int64
 3   Ad- Width              23066 non-null  int64
 4   Ad Size                23066 non-null  int64
 5   Ad Type                23066 non-null  object
 6   Platform               23066 non-null  object
 7   Device Type            23066 non-null  object
 8   Format                 23066 non-null  object
 9   Available_Impressions  23066 non-null  int64
 10  Matched_Queries        23066 non-null  int64
 11  Impressions            23066 non-null  int64
 12  Clicks                 23066 non-null  int64
 13  Spend                  23066 non-null  float64
 14  Fee                    23066 non-null  float64
 15  Revenue                23066 non-null  float64
 16  CTR                    18330 non-null  float64
 17  CPM                    18330 non-null  float64
 18  CPC                    18330 non-null  float64
dtypes: float64(6), int64(7), object(6)
memory usage: 3.3+ MB
```

**Ads_Data.isnull ( ).sum ( ) – Missing values**

```
Timestamp                  0
InventoryType              0
Ad - Length                0
Ad- Width                  0
Ad Size                    0
Ad Type                    0
Platform                   0
Device Type                0
Format                     0
Available_Impressions      0
Matched_Queries            0
Impressions                0
Clicks                     0
Spend                      0
Fee                        0
Revenue                    0
CTR                     4736
CPM                     4736
CPC                     4736
dtype: int64
```

**Number of rows and columns:-**

The dataset contains 23,066 entries with 19 columns.

**Ads_data. Duplicated ( ).sum ( )**

= 0

**#clustering: Treat missing values in CPC, CTR and CPM using the formula given.**

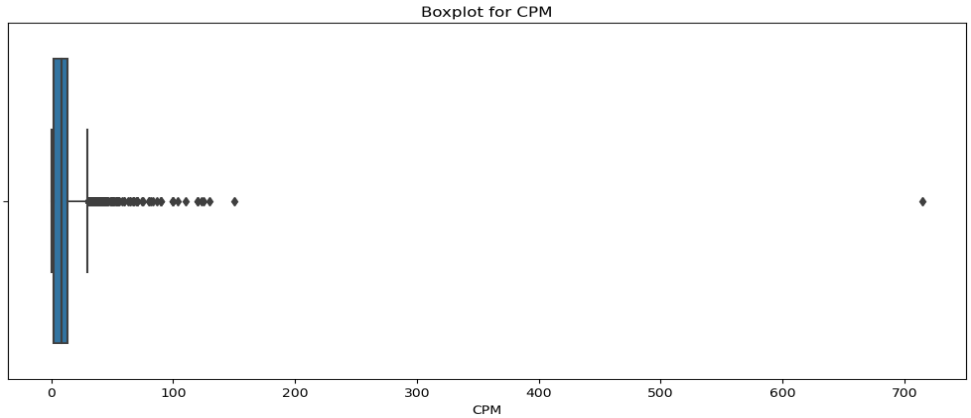
**Output:**

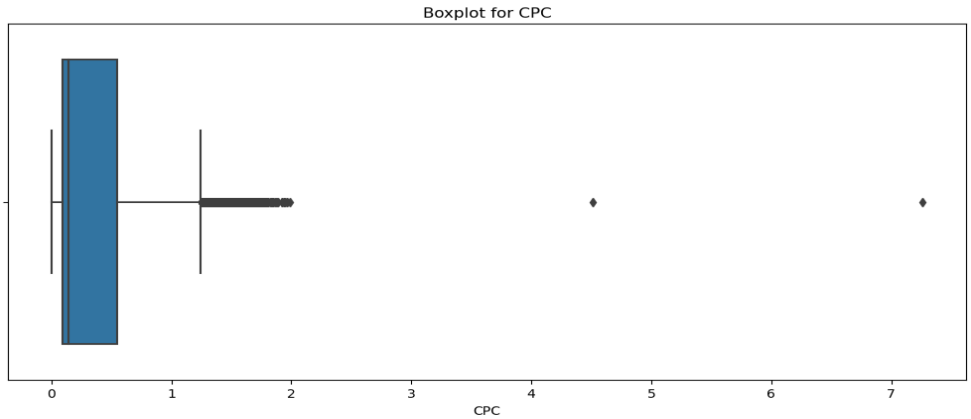
The missing values in the 'CPM', 'CPC', and 'CTR' columns have been successfully treated using the provided formulas. There are no more missing values in these columns.

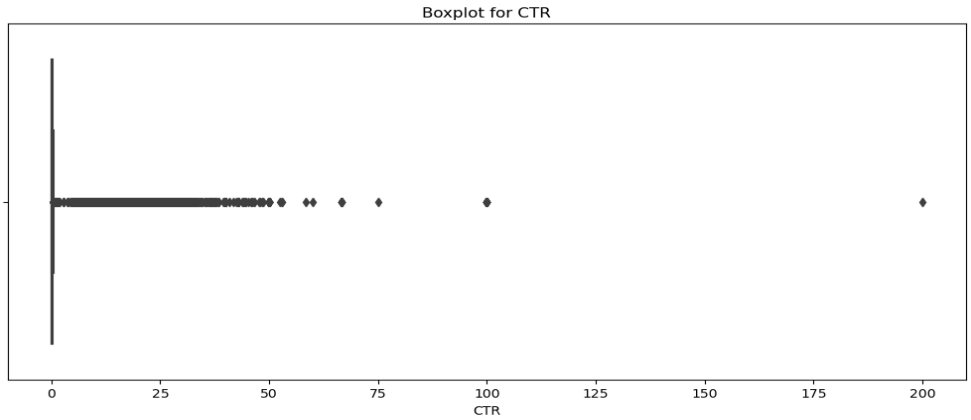
1. CPM:
   CPM= (Total Campaign SpendNumber of Impressions) ×1,000
   CPM= (Number of ImpressionsTotal Campaign Spend) ×1,000
2. CPC:
   CPC=Total Cost (spend) Number of Clicks
   CPC=Number of ClicksTotal Cost (spend)
3. CTR:
   CTR=Total Measured ClicksTotal Measured Ad Impressions×100
   CTR=Total Measured Ad ImpressionsTotal Measured Clicks×100

CPM 0
CPC 0
CTR 0
Dtype: int64

**#clustering: Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an analyst your judgement may be different from another analyst).**




From the boxplots:

CPM: There are potential outliers on the higher end of the value distribution.

CPC: Similarly, there are potential outliers on the higher end.

CTR: We can observe potential outliers both on the lower and higher ends.

Is treating outliers necessary for K-Means clustering?

K-Means clustering is sensitive to outliers. Outliers can heavily influence the centroids' positioning and, consequently, the final clusters. This can lead to misrepresentation of the actual data distribution and formation of clusters that might not be representative of the inherent data structure.

Decision on treating outliers:

While it's generally a good idea to treat outliers before clustering, the method of treatment can vary:

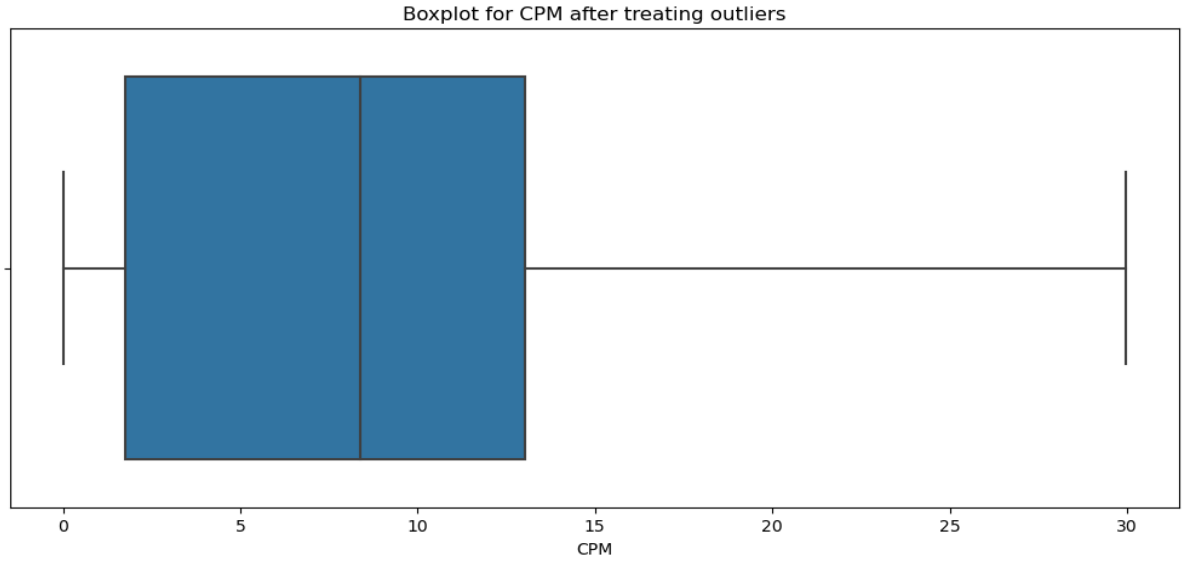
Removing outliers: This is a straightforward approach but can lead to loss of data.

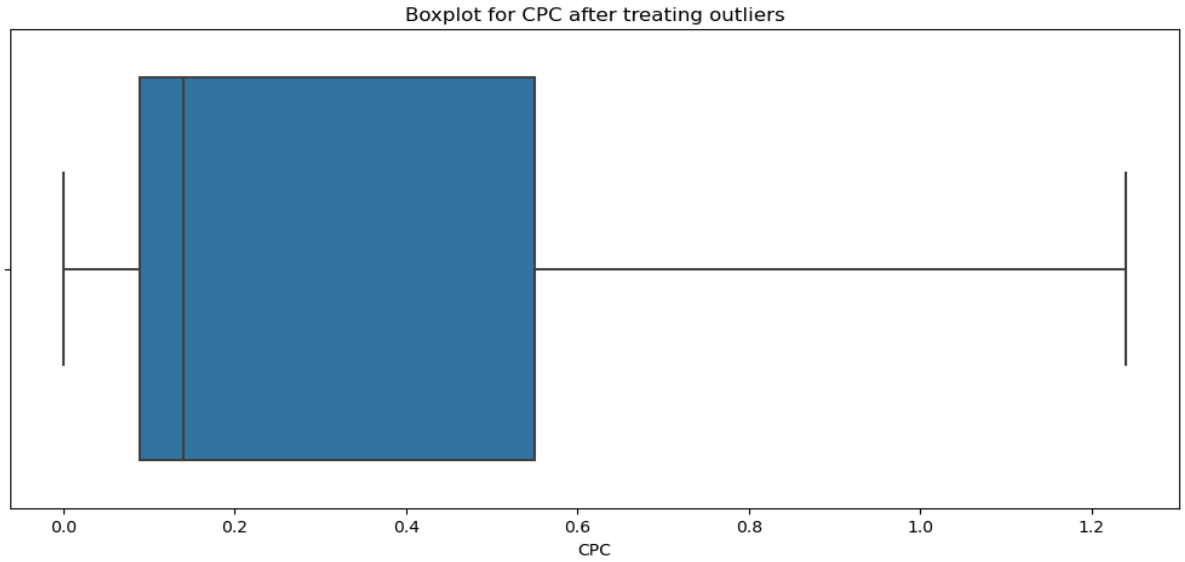
Capping: Outliers can be capped to a certain threshold, preserving the data point but limiting its extreme value.

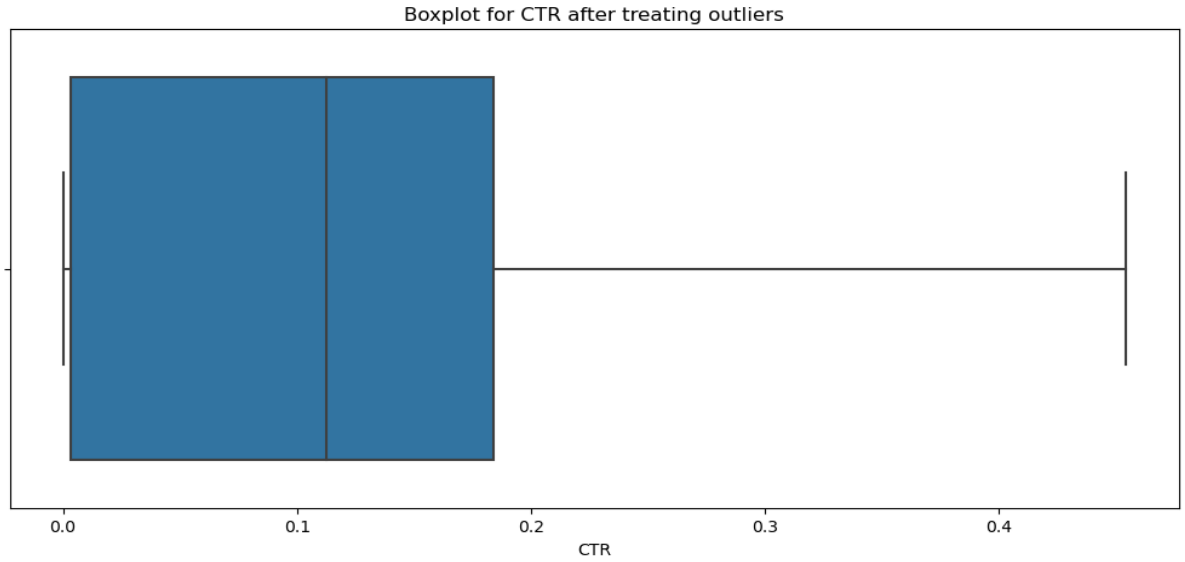
Transformations: Sometimes, applying transformations like log or square root can help in reducing the impact of outliers.

Given the visualization, I'd suggest capping the outliers as removing might result in loss of important data. The capping can be done using the Interquartile Range (IQR) method, where values outside the range [Q1 - 1.5IQR, Q3 + 1.5IQR] can be capped to these boundaries.

The outliers in the columns 'CPM', 'CPC', and 'CTR' have been capped using the IQR method. As observed in the updated boxplots, the data distribution within these columns has been adjusted to reduce the influence of extreme values.


Boxplot for CPM after treating outliers
0
5
10
15
20
25
30
CPM


Boxplot for CPC after treating outliers
0.0
0.2
0.4
0.6
0.8
1.0
1.2
CPC


Boxplot for CTR after treating outliers
0.0
0.1
0.2
0.3
0.4
CTR

**Clustering: Perform z-score scaling and discuss how it affects the speed of the algorithm.**

Z-score scaling (or standardization) means transforming data such that it has a mean of 0 and a standard deviation of 1. This is useful for algorithms like K-Means, which are sensitive to the scale of the features.

The formula for z-score scaling is:

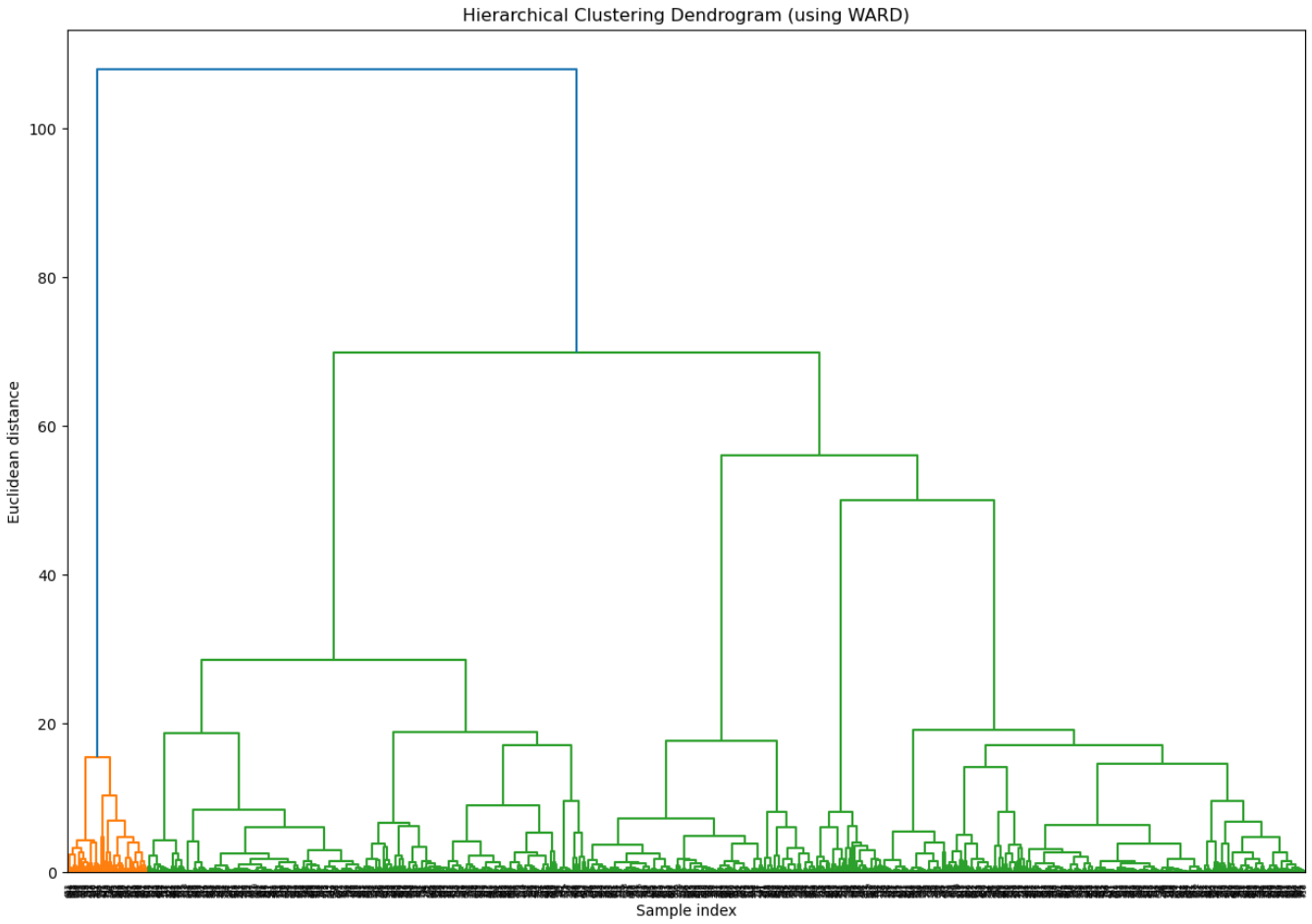
$$Z = x - \text{mean} / \text{standard deviation}$$

| Timestamp | InventoryType | Ad-Length | Ad-Width | Ad Size | Ad Type | Platform | Device Type | Format |
|---|---|---|---|---|---|---|---|---|
| 2020-9-2-17 | Format1 | -0.3645 | -0.4328 | -0.35222 | Inter | 222 | VideoDesktop | Display |
| 2020-9-2-10 | Format1 | -0.3645 | -0.4328 | -0.35222 | Inter | 227 | AppMobile | Video |
| 2020-9-1-22 | Format1 | -0.3645 | -0.4328 | -0.35222 | Inter | 222 | VideoDesktop | Display |
| 2020-9-3-20 | Format1 | -0.3645 | -0.4328 | -0.35222 | Inter | 228 | VideoMobile | Video |
| 2020-9-4-15 | Format1 | -0.3645 | -0.4328 | -0.35222 | Inter | 217 | WebDesktop | Video |

| Available_ Impressions | Matched_Queries | Impressions | Clicks | Spend | Fee | Revenue | CTR | CPM | CPC |
|---|---|---|---|---|---|---|---|---|---|
| -0.51241 | -0.51525 | -0.51092 | -0.61531 | -0.66537 | 0.465447 | -0.61969 | -0.8912 | -1.19456 | -1.04114 |
| -0.51241 | -0.51526 | -0.51093 | -0.61531 | -0.66537 | 0.465447 | -0.61969 | -0.88862 | -1.19456 | -1.04114 |
| -0.51221 | -0.51524 | -0.51091 | -0.61531 | -0.66537 | 0.465447 | -0.61969 | -0.89314 | -1.19456 | -1.04114 |
| -0.51228 | -0.51518 | -0.51085 | -0.61531 | -0.66537 | 0.465447 | -0.61969 | -0.89832 | -1.19456 | -1.04114 |
| -0.51253 | -0.51528 | -0.51095 | -0.61531 | -0.66537 | 0.465447 | -0.61969 | -0.88473 | -1.19456 | -1.04114 |

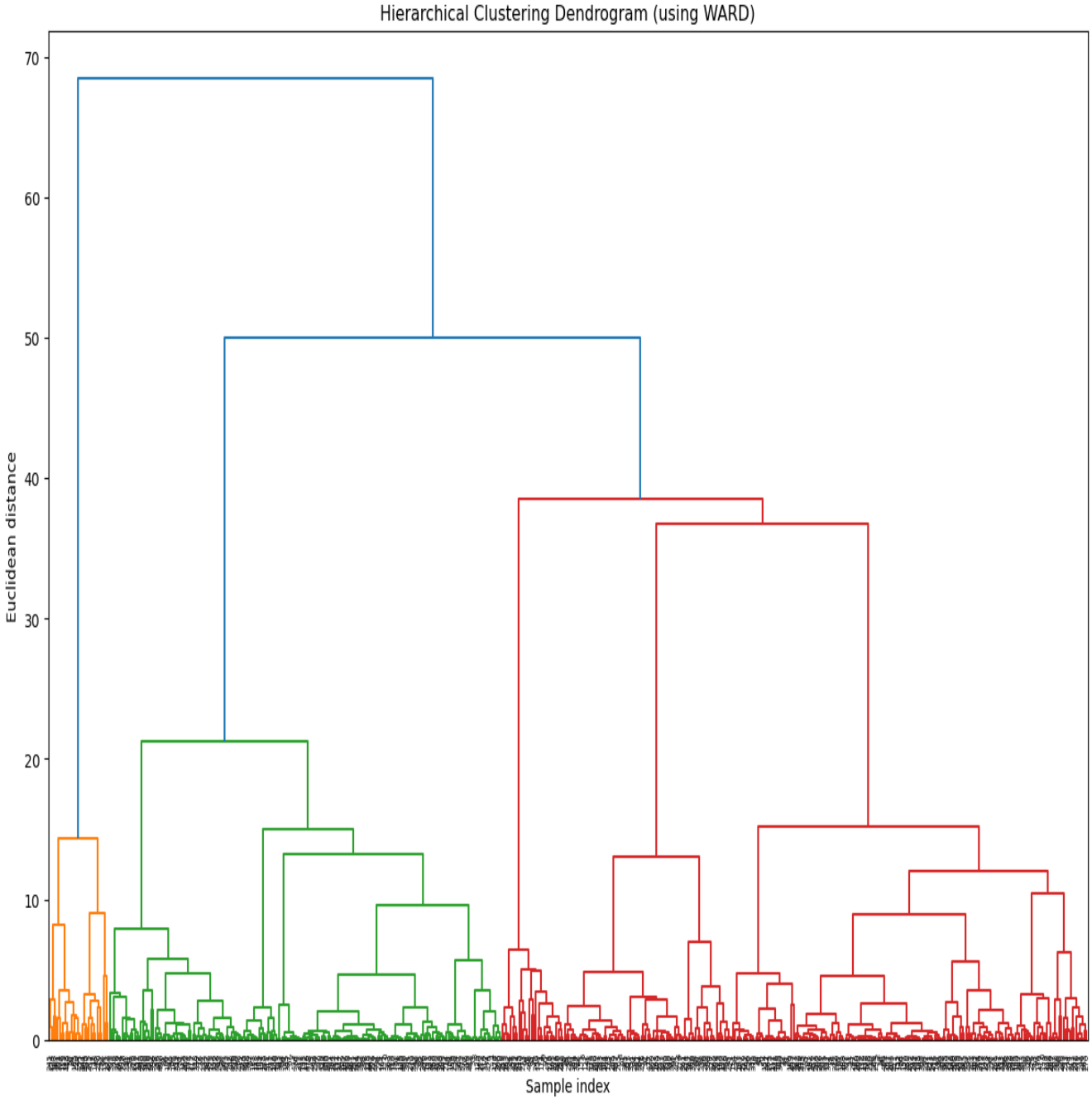
**#clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance, and identify the optimum number of clusters.**

**Sample of 1000:**

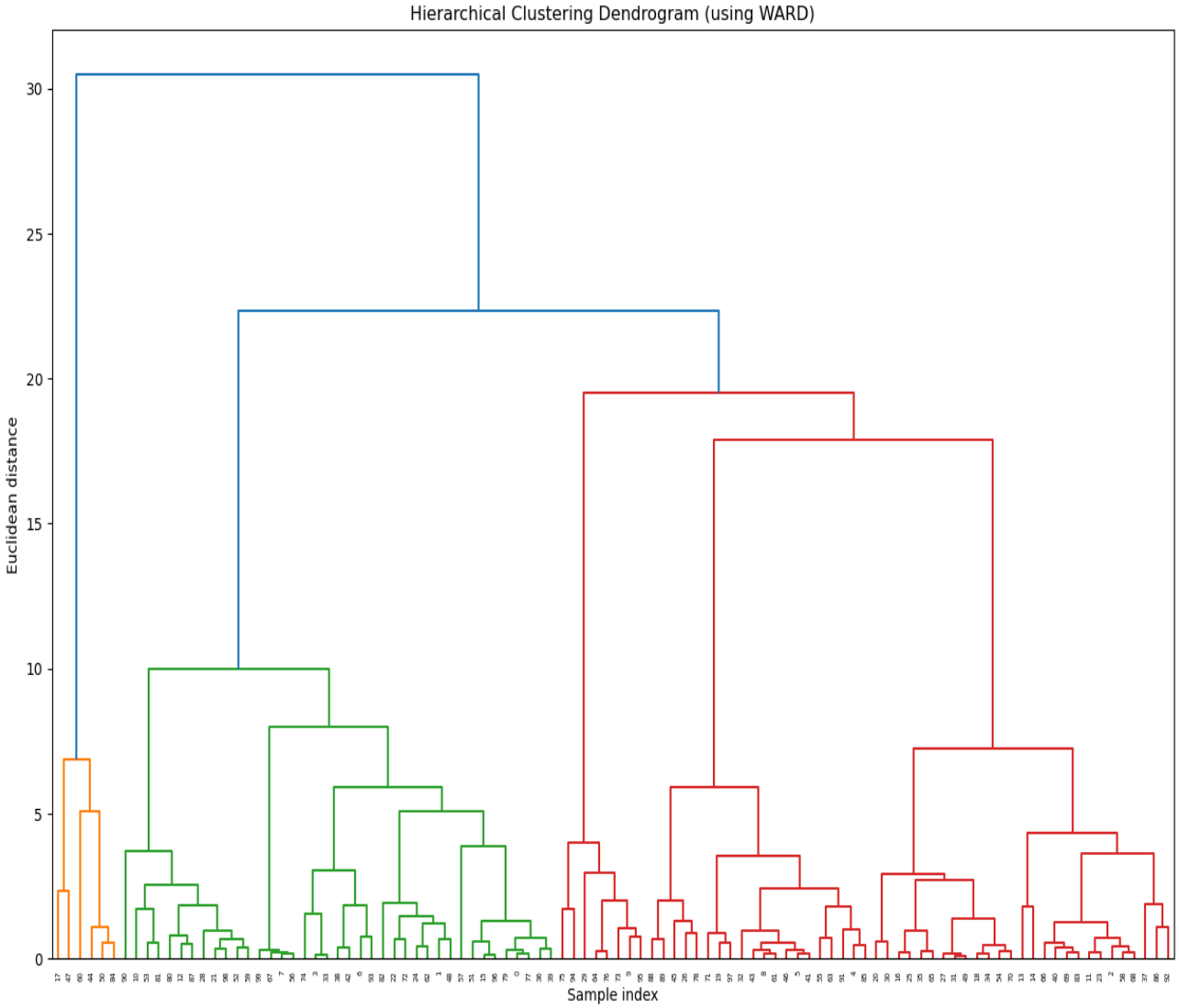



Using hierarchical clustering to construct a Dendrogram with the WARD method and Euclidean distance. This will help us visually assess the number of clusters that might be appropriate for the data. Let's proceed with constructing the Dendrogram.

**Sample of 500**

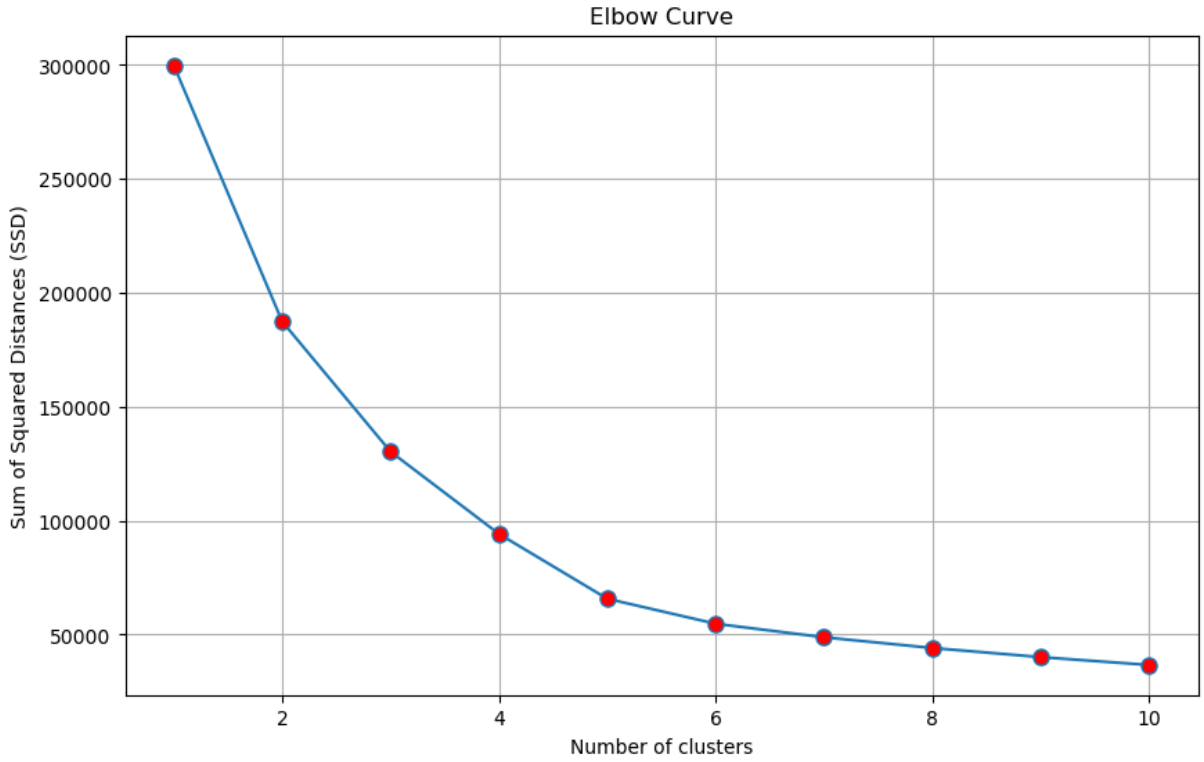


**Sample of 100**



**#clustering: Make Elbow plot (up to n=10) and identify the optimum number of clusters for k-means algorithm.**

The Elbow method is a popular technique to determine the optimal number of clusters for the K-Means algorithm. We will plot the sum of squared distances for different numbers of clusters (from 1 to 10) and look for an "elbow" point, where the rate of decrease sharply changes.

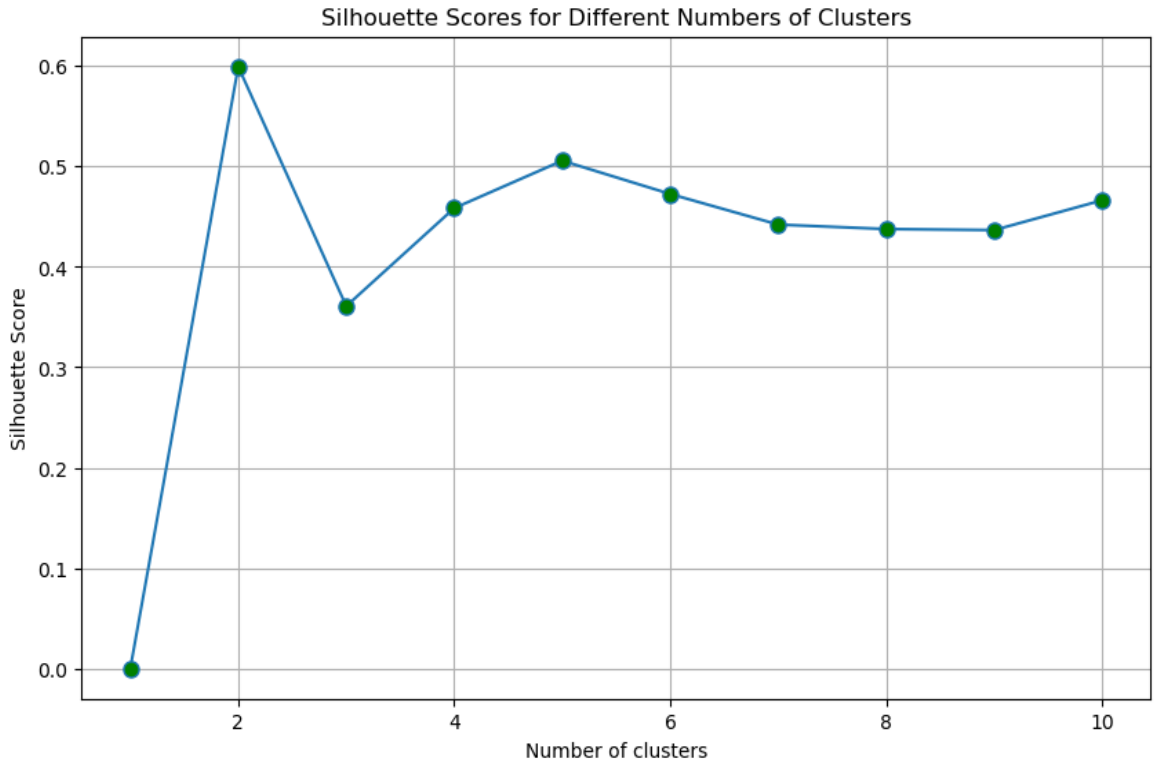



The Elbow curve displays the sum of squared distances (SSD) for different numbers of clusters.

From the plot, there's no clear "elbow" where the rate of decrease sharply changes. However, one could argue that there's a slight bend around 2 or 3 clusters. This means the optimal number of clusters could be 2 or 3, which aligns with our observation from the Dendrogram.

**#clustering: Print silhouette scores for up to 10 clusters and identify the optimum number of clusters.**

The silhouette score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette score ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters.

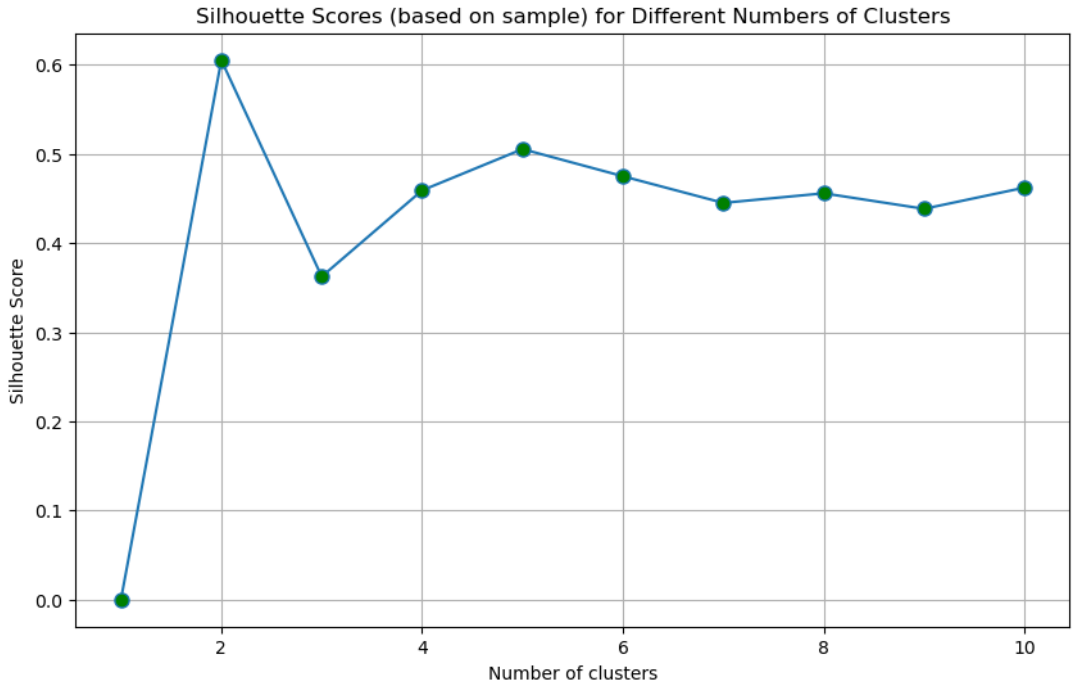



Since we've encountered issues with silhouette scores, we'll rely on our earlier observations from the Dendrogram and elbow plot, which suggested that 2 or 3 clusters might be optimal.

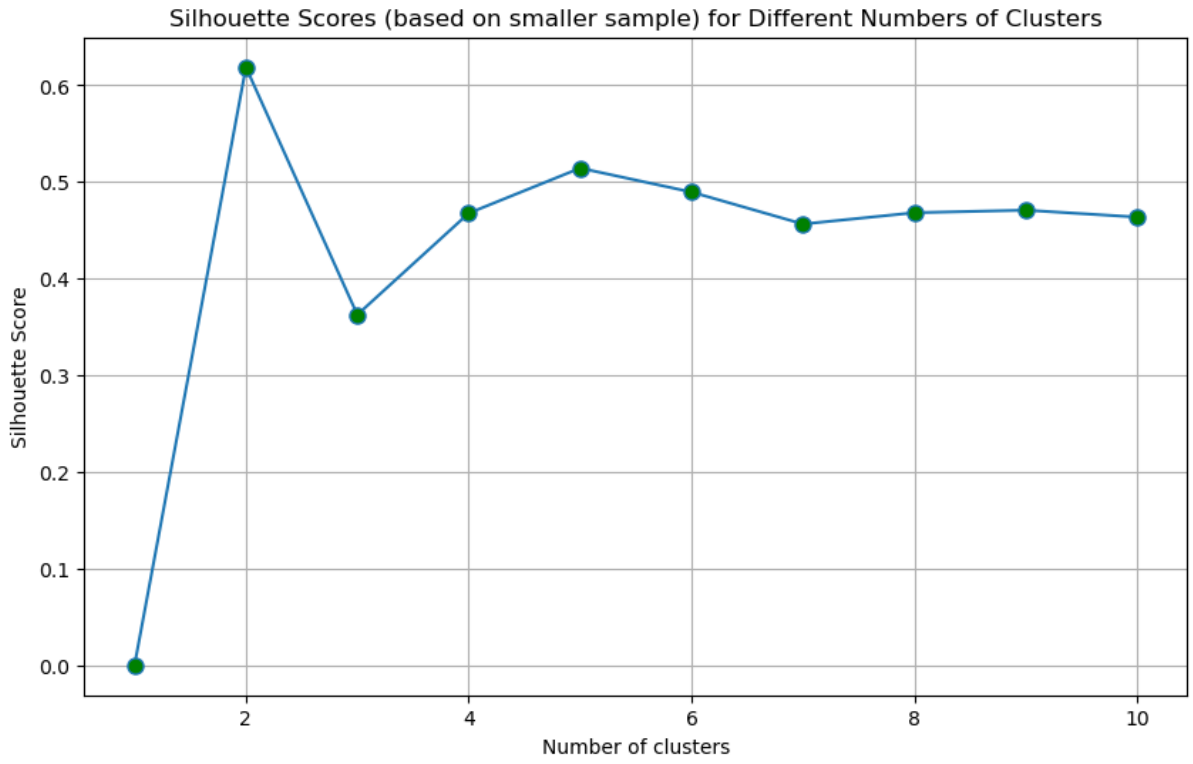
Let's proceed with k=3 (3 clusters) for the K-Means clustering and then profile the ads based on these clusters.

K-Means clustering with k=3 clusters. After obtaining the cluster labels, we'll append these labels to our original dataset. Then, we will profile the ads based on these clusters by grouping the data by clusters and deriving insights on metrics like Clicks, Spend, Revenue, CPM, CTR, and CPC based on Device Type.

**Taking the sample data of 5000.**




**Sample of 2000**



**#clustering: Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots].**

Cluster

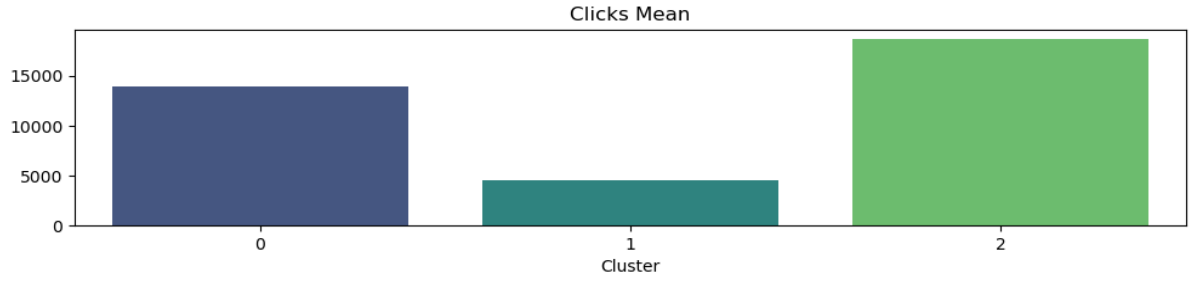
| | |
|---|---|
| 0 | 1 |
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |

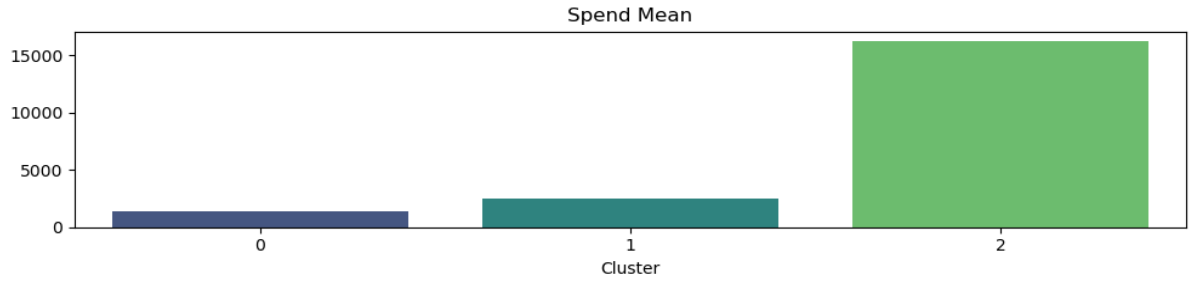
| Cluster | Clicks Mean | Clicks Sum | Spend Mean | Spend Sum | Revenue Mean | Revenue Sum | P_mean | CTR_Mean | CPC_Mean | Desktop | Mobile |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 13981.60859 | 18071 2291 | 1402.6 87308 | 18129 733.46 | 968.67 6961 | 1252 0150 | 13.32 3724 | 0.213 391 | 0.101 41 | 464 0 | 828 5 |
| 1 | 4578.2 73462 | 40197 241 | 2528.0 91157 | 22196 640.36 | 1697.2 86088 | 1490 2170 | 1.718 066 | 0.051 488 | 0.581 295 | 313 3 | 564 7 |
| 2 | 18663.61719 | 25401 183 | 16241.48003 | 22104 654.32 | 12463.24959 | 1696 2480 | 1.680 501 | 0.029 327 | 0.883 705 | 487 | 874 |

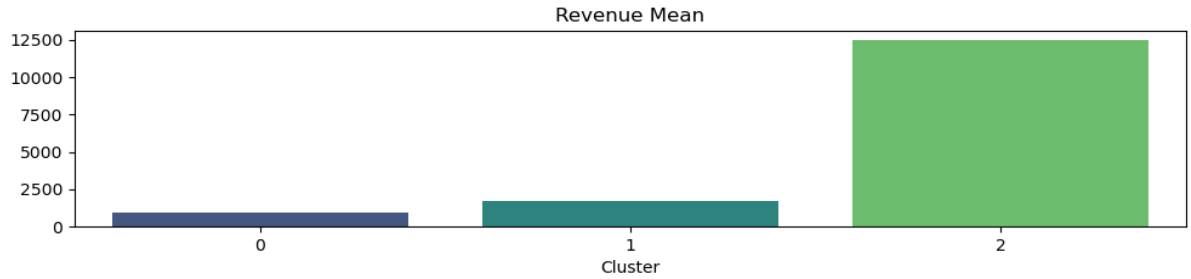
The clustering analysis provided a structured way to segment our ads into meaningful groups. Each cluster or segment has its unique characteristics, behaviours, and performance metrics. By understanding these segments, advertisers and businesses can make informed decisions, optimize their ad spend, and devise strategies tailored to each segment. Whether it's focusing on improving the efficiency of ads in Cluster 1 or capitalizing on the high revenue potential of Cluster 2, these insights pave the way for more targeted and effective advertising strategies.

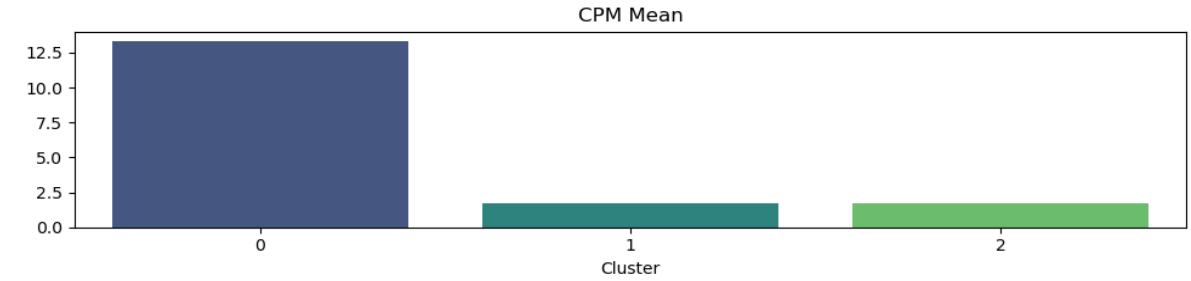
Additionally, the project underscored the importance of data pre-processing, the sensitivity of certain algorithms to outliers and feature scales, and the value of visualizations in understanding and presenting findings.

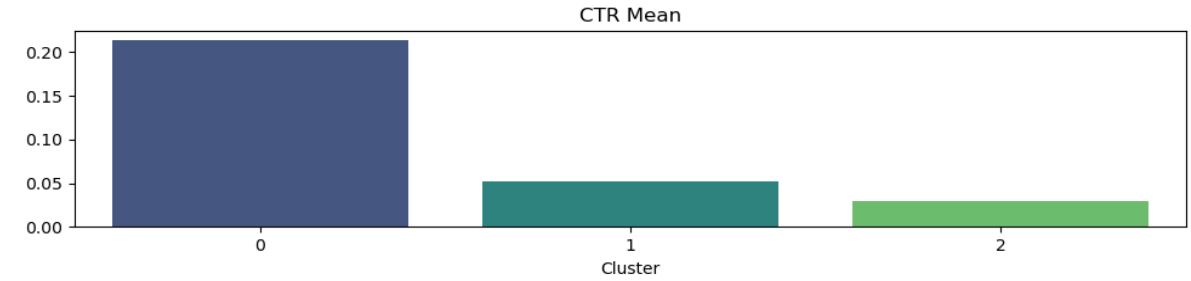
**Bar Plots:**

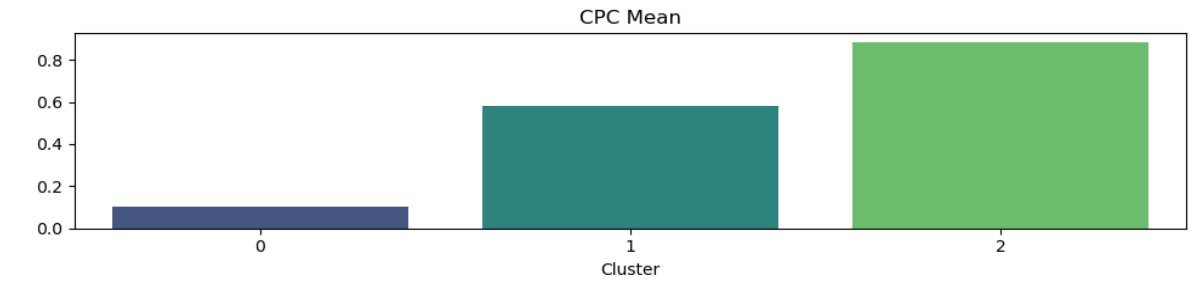

Clicks Mean
15000
10000
5000
0
0
1
2
Cluster


Spend Mean
15000
10000
5000
0
0
1
2
Cluster


Revenue Mean
12500
10000
7500
5000
2500
0
0
1
2
Cluster


CPM Mean
12.5
10.0
7.5
5.0
2.5
0.0
0
1
2
Cluster


CTR Mean
0.20
0.15
0.10
0.05
0.00
0
1
2
Cluster


CPC Mean
0.8
0.6
0.4
0.2
0.0
0
1
2
Cluster

# Part 2

**PCA: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc…**

**Data. Head ( ):**

| State Code | Dist. Code | State | Area Name | No_HH | TOT_M | TOT_F | M_06 | F_06 | M_SC |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | Jammu & Kashmir | Kupwara | 7707 | 23388 | 29796 | 5862 | 6196 | 3 |
| 1 | 2 | Jammu & Kashmir | Badgam | 6218 | 19585 | 23102 | 4482 | 3733 | 7 |
| 1 | 3 | Jammu & Kashmir | Leh(Ladakh) | 4452 | 6546 | 10964 | 1082 | 1018 | 3 |
| 1 | 4 | Jammu & Kashmir | Kargil | 1320 | 2784 | 4206 | 563 | 677 | 0 |
| 1 | 5 | Jammu & Kashmir | Punch | 11654 | 20591 | 29981 | 5157 | 4587 | 20 |

| MARG_CL_0_3_M | MARG_CL_0_3_F | MARG_AL_0_3_M | MARG_AL_0_3_F | MARG_HH_0_3_M | MARG_HH_0_3_F | MARG_OT_0_3_M | MARG_OT_0_3_F | NON_WORK_M | NON_WORK_F |
|---|---|---|---|---|---|---|---|---|---|
| 1150 | 749 | 180 | 237 | 680 | 252 | 32 | 46 | 258 | 214 |
| 525 | 715 | 123 | 229 | 186 | 148 | 76 | 178 | 140 | 160 |
| 114 | 188 | 44 | 89 | 3 | 34 | 0 | 4 | 67 | 61 |
| 194 | 247 | 61 | 128 | 13 | 50 | 4 | 10 | 116 | 59 |
| 874 | 1928 | 465 | 1043 | 205 | 302 | 24 | 105 | 180 | 478 |

**Data.tail( ):**

| State Code | Dist.Code | State | Area Name | No_HH | TOT_M | TOT_F | M_06 | F_06 | M_SC |
|---|---|---|---|---|---|---|---|---|---|
| 34 | 636 | Puducherry | Mahe | 3333 | 8154 | 11781 | 1146 | 1203 | 21 |
| 34 | 637 | Puducherry | Karaikal | 10612 | 12346 | 21691 | 1544 | 1533 | 2234 |
| 35 | 638 | Andaman & Nicobar Island | Nicobars | 1275 | 1549 | 2630 | 227 | 225 | 0 |
| 35 | 639 | Andaman & Nicobar Island | North & Middle Andaman | 3762 | 5200 | 8012 | 723 | 664 | 0 |
| 35 | 640 | Andaman & Nicobar Island | South Andaman | 7975 | 11977 | 18049 | 1470 | 1358 | 0 |

| MARG_CL_0_3_M | MARG_CL_0_3_F | MARG_AL_0_3_M | MARG_AL_0_3_F | MARG_HH_0_3_M | MARG_HH_0_3_F | MARG_OT_0_3_M | MARG_OT_0_3_F | NON_WORK_M | NON_WORK_F |
|---|---|---|---|---|---|---|---|---|---|
| 32 | 47 | 0 | 0 | 0 | 0 | 0 | 0 | 32 | 47 |
| 155 | 337 | 3 | 14 | 38 | 130 | 4 | 23 | 110 | 170 |
| 104 | 134 | 9 | 4 | 2 | 6 | 17 | 47 | 76 | 77 |
| 136 | 172 | 24 | 44 | 11 | 21 | 1 | 4 | 100 | 103 |
| 173 | 122 | 6 | 2 | 17 | 17 | 2 | 4 | 148 | 99 |

**Data.info( ):**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 640 entries, 0 to 639
Data columns (total 61 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   State Code      640 non-null    int64
 1   Dist.Code       640 non-null    int64
 2   State           640 non-null    object
 3   Area Name       640 non-null    object
 4   No_HH           640 non-null    int64
 5   TOT_M           640 non-null    int64
 6   TOT_F           640 non-null    int64
 7   M_06            640 non-null    int64
 8   F_06            640 non-null    int64
 9   M_SC            640 non-null    int64
 10  F_SC            640 non-null    int64
 11  M_ST            640 non-null    int64
 12  F_ST            640 non-null    int64
 13  M_LIT           640 non-null    int64
```

```
 14  F_LIT           640 non-null    int64
 15  M_ILL           640 non-null    int64
 16  F_ILL           640 non-null    int64
 17  TOT_WORK_M      640 non-null    int64
 18  TOT_WORK_F      640 non-null    int64
 19  MAINWORK_M      640 non-null    int64
 20  MAINWORK_F      640 non-null    int64
 21  MAIN_CL_M       640 non-null    int64
 22  MAIN_CL_F       640 non-null    int64
 23  MAIN_AL_M       640 non-null    int64
 24  MAIN_AL_F       640 non-null    int64
 25  MAIN_HH_M       640 non-null    int64
 26  MAIN_HH_F       640 non-null    int64
 27  MAIN_OT_M       640 non-null    int64
 28  MAIN_OT_F       640 non-null    int64
 29  MARGWORK_M      640 non-null    int64
 30  MARGWORK_F      640 non-null    int64
 31  MARG_CL_M       640 non-null    int64
 32  MARG_CL_F       640 non-null    int64
 33  MARG_AL_M       640 non-null    int64
 34  MARG_AL_F       640 non-null    int64
 35  MARG_HH_M       640 non-null    int64
 36  MARG_HH_F       640 non-null    int64
 37  MARG_OT_M       640 non-null    int64
 38  MARG_OT_F       640 non-null    int64
 39  MARGWORK_3_6_M  640 non-null    int64
 40  MARGWORK_3_6_F  640 non-null    int64
 41  MARG_CL_3_6_M   640 non-null    int64
 42  MARG_CL_3_6_F   640 non-null    int64
 43  MARG_AL_3_6_M   640 non-null    int64
 44  MARG_AL_3_6_F   640 non-null    int64
 45  MARG_HH_3_6_M   640 non-null    int64
 46  MARG_HH_3_6_F   640 non-null    int64
 47  MARG_OT_3_6_M   640 non-null    int64
 48  MARG_OT_3_6_F   640 non-null    int64
 49  MARGWORK_0_3_M  640 non-null    int64
 50  MARGWORK_0_3_F  640 non-null    int64
 51  MARG_CL_0_3_M   640 non-null    int64
 52  MARG_CL_0_3_F   640 non-null    int64
 53  MARG_AL_0_3_M   640 non-null    int64
 54  MARG_AL_0_3_F   640 non-null    int64
 55  MARG_HH_0_3_M   640 non-null    int64
 56  MARG_HH_0_3_F   640 non-null    int64
 57  MARG_OT_0_3_M   640 non-null    int64
 58  MARG_OT_0_3_F   640 non-null    int64
 59  NON_WORK_M      640 non-null    int64
 60  NON_WORK_F      640 non-null    int64
dtypes: int64(59), object(2)
memory usage: 305.1+ KB
```

**Data.isnull ( ).sum (  ):  - No missing values.**

```
State Code        0
Dist.Code         0
State             0
Area Name         0
No_HH             0
                 ..
MARG_HH_0_3_F     0
MARG_OT_0_3_M     0
MARG_OT_0_3_F     0
NON_WORK_M        0
NON_WORK_F        0
Length: 61, dtype: int64
```

**Data. Duplicated ( ).sum ( ) – No duplicated values**

0

**Part 2 - PCA: Perform detailed exploratory analysis by creating certain questions like**

**(i) which state has highest gender ratio and which has the lowest?**

**(ii) Which district has the highest & lowest gender ratio?**

**(Example Questions). Pick 5 variables out of the given 24 variables below for EDA: No_HH, TOT_M, TOT_F, M_06, F_06, M_SC, F_SC, M_ST, F_ST, M_LIT, F_LIT, M_ILL, F_ILL, TOT_WORK_M, TOT_WORK_F, MAINWORK_M, MAINWORK_F, MAIN_CL_M, MAIN_CL_F, MAIN_AL_M, MAIN_AL_F, MAIN_HH_M, MAIN_HH_F, MAIN_OT_M, MAIN_OT_F**

**Output**: Gender ratio as Total Females/Total Males

('Andhra Pradesh',

1.8950931296262146,

'Lakshadweep',

1.1519925134523903,

'Krishna',

2.28324963845265,

'Lakshadweep',

1.1519925134523903)

State with the highest gender ratio: Andhra Pradesh with a gender ratio of approximately 1.895, meaning there are approximately 1.895 females for every male in the state.

State with the lowest gender ratio: Lakshadweep with a gender ratio of approximately 1.152, meaning there are approximately 1.152 females for every male in the state.

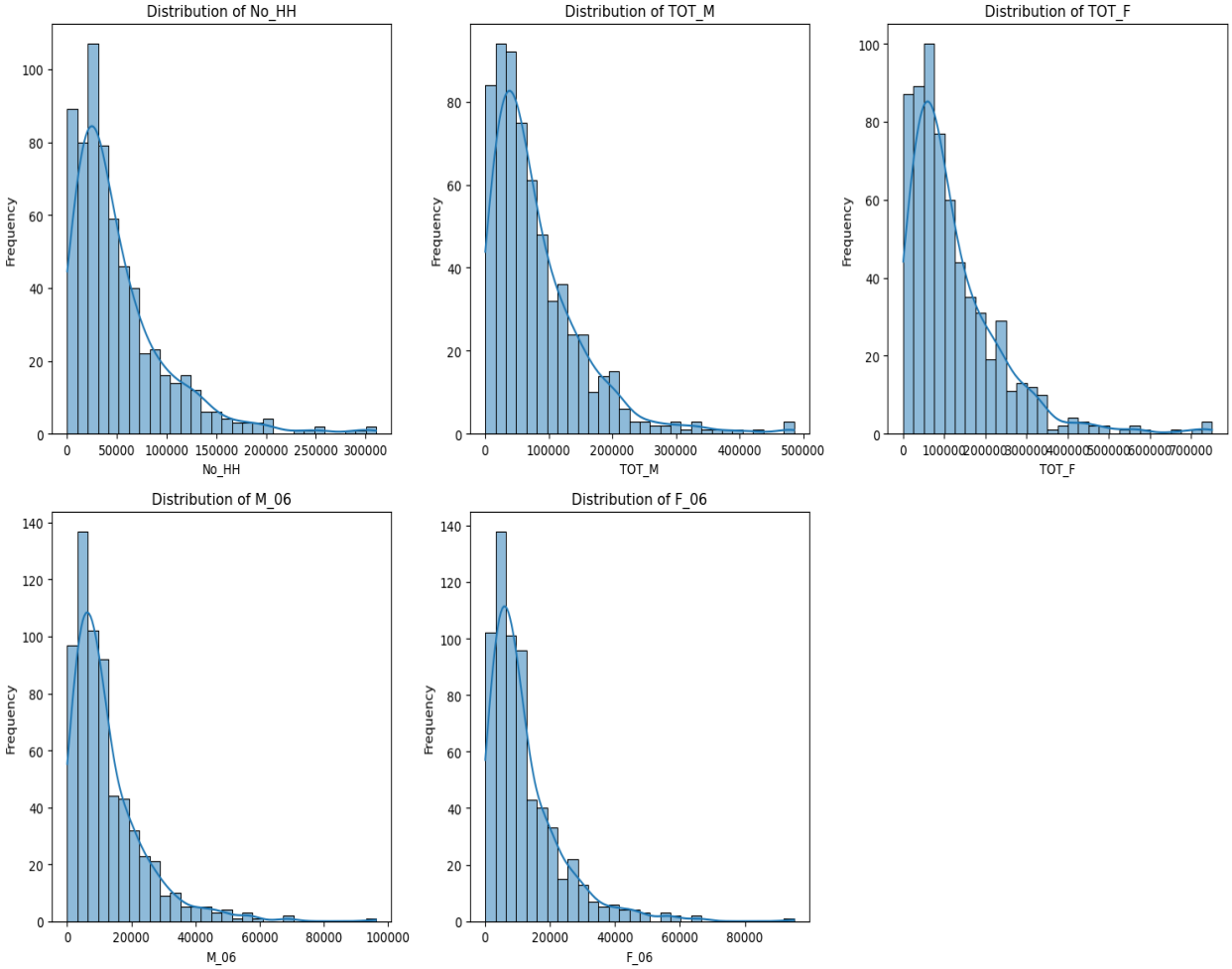
District with the highest gender ratio: Krishna district (located in Andhra Pradesh) with a gender ratio of approximately 2.283 indicating there are about 2.283females for every male in this district.

District with the lowest gender ratio: Lakshadweep with a gender ratio of approximately 1.152.

**(ii) Exploratory Data Analysis (EDA) on the following 5 variables out of the given 24 variables**:

```
# Variables for EDA

Variables = ['No_HH', 'TOT_M', 'TOT_F', 'M_06', 'F_06']
```



No_HH (Number of Households): The distribution is right-skewed, indicating that most districts have a lower number of households, while a few districts have a very high number of households.

TOT_M (Total Males): The distribution is right-skewed, which means that most districts have a lower male population, with a few exceptions having a high male population.

TOT_F (Total Females): This distribution also appears right-skewed, similar to the male population distribution.

M_06 (Males below 6 years): Again, the distribution is right-skewed, indicating that most districts have a lower number of males under the age of 6 years.

F_06 (Females below 6 years): This distribution is consistent with the M_06 distribution, showing a right-skewed pattern.

**Part 2 - PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?**

I choose not to treat outliers for this case. However, let's discuss the potential implications of this decision.

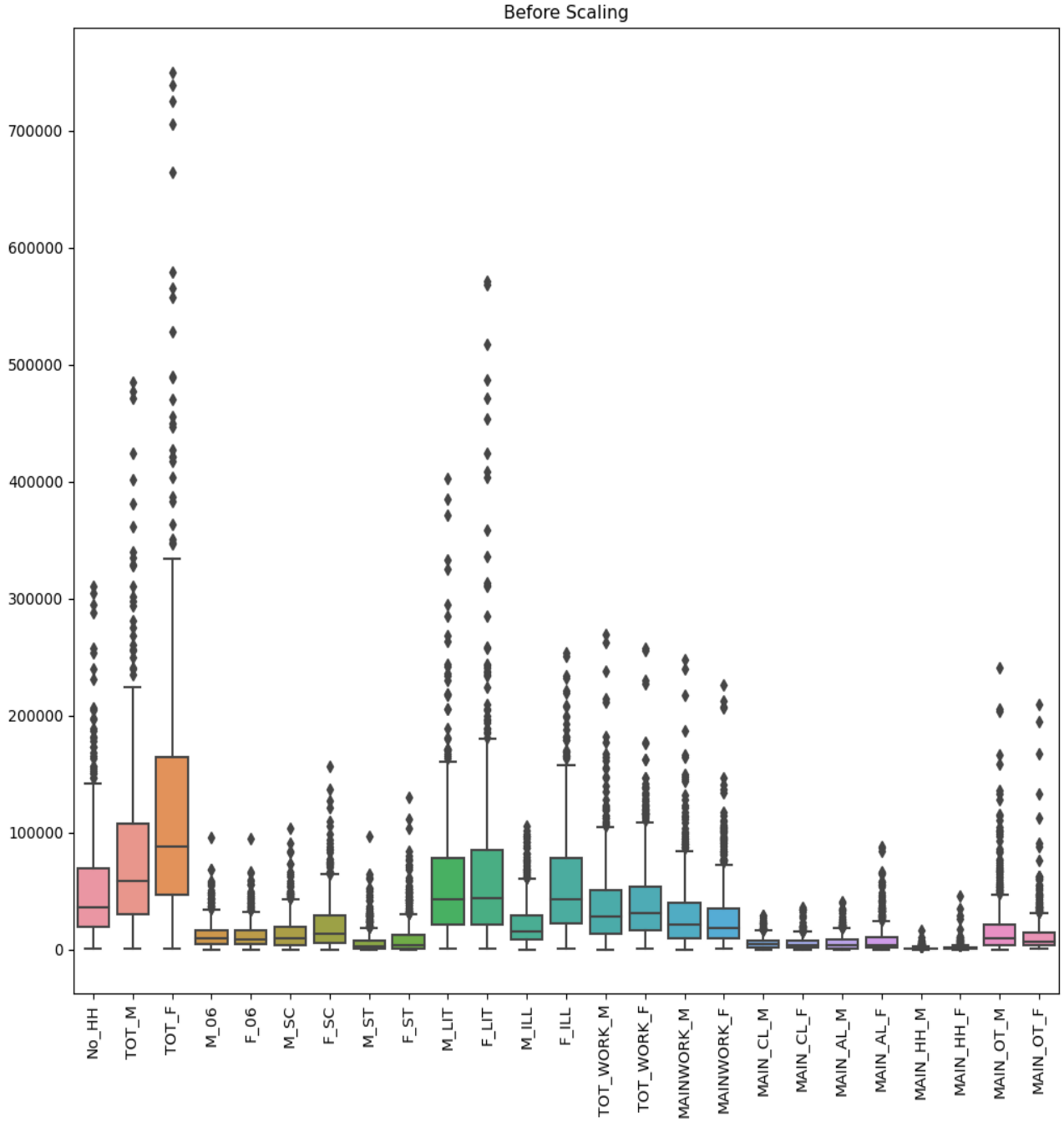
### Think that treating outliers is required:

Treating outliers is essential when they might significantly influence the results, especially in algorithms sensitive to outliers, like PCA. Outliers can distort the principal components, making them more aligned with the outliers than with the overall data distribution. This can lead to misleading interpretations.

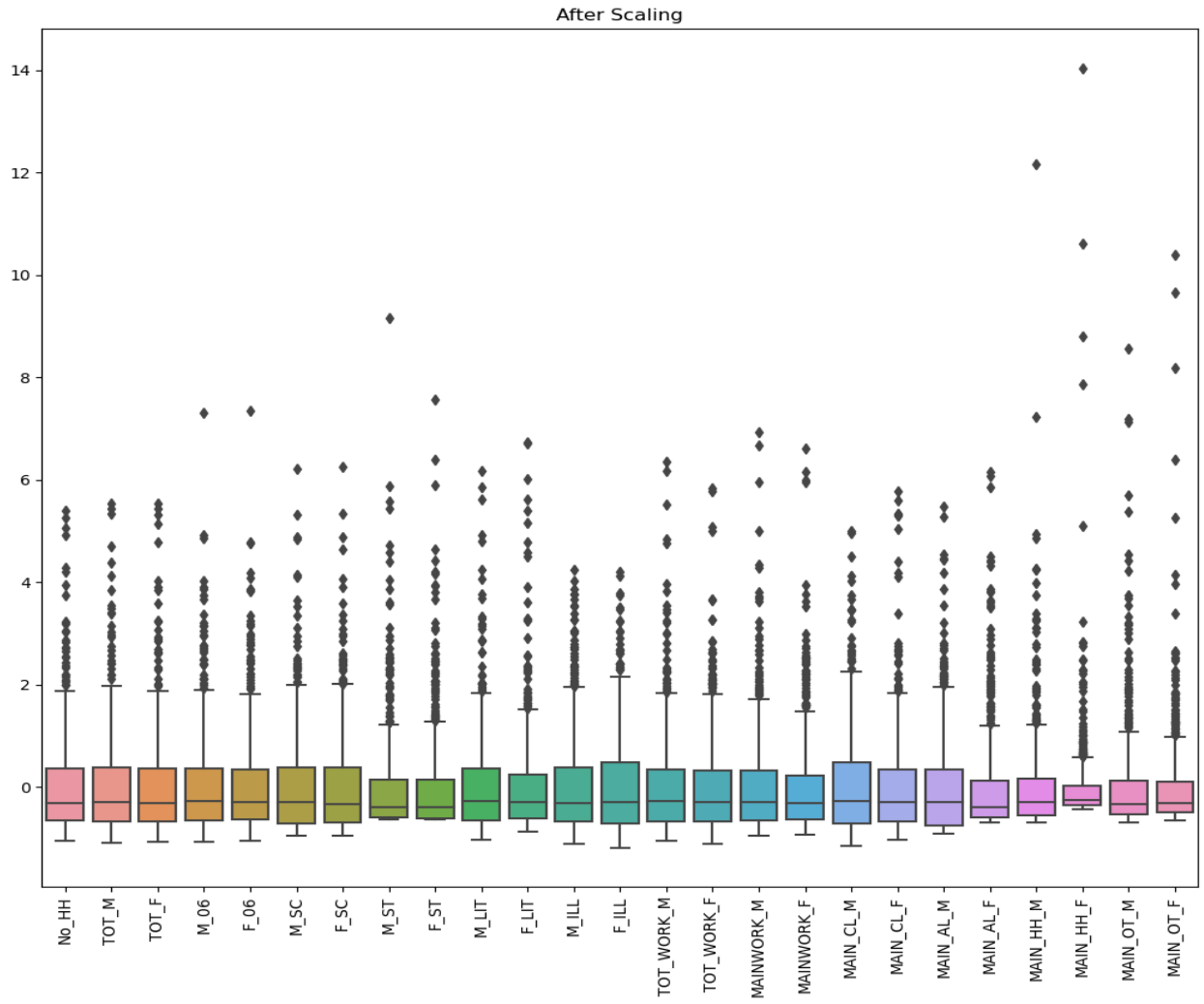
However, it's also important to determine if these outliers are genuine or due to data errors. In real-world scenarios, especially with census data, extreme values might represent actual observations and not anomalies. If that's the case, removing them might lead to a loss of essential information.

**#PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.**

Before Scaling: The boxplots show the range and presence of outliers for each variable in its original scale. We can see that many variables have outliers, especially on the upper end.



After Scaling: After applying the z-score method, the data for each variable has been standardized to have a mean of 0 and a standard deviation of 1. The outliers are still present, but their relative positions have shifted due to the standardized scale.




**Does scaling have any impact on outliers?**

Scaling does not eliminate outliers; it merely changes the scale of the data. As we can observe, the outliers' relative positions remain consistent before and after scaling. The primary purpose of scaling in PCA is to ensure that each variable has equal weight, especially when variables have different units or scales. In our case, even though the units are the same (population counts), the magnitudes differ, so scaling is essential for PCA.

**Part 2 - PCA: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get Eigen values and Eigen vector.**

```
(array([ 1.62107515e+01+0.00000000e+00j,  2.42307151e+00+0.00000000e+00
j,
         2.00578300e+00+0.00000000e+00j,  1.34445040e+00+0.00000000e+00
j,
         8.20116560e-01+0.00000000e+00j,  7.20513809e-01+0.00000000e+00
j,
         5.05666322e-01+0.00000000e+00j,  3.07288778e-01+0.00000000e+00
j,
         3.01513233e-01+0.00000000e+00j,  1.66973671e-01+0.00000000e+00
j,
         7.47965844e-02+0.00000000e+00j,  6.11833157e-02+0.00000000e+00
j,
         4.73587939e-02+0.00000000e+00j,  1.62702744e-02+0.00000000e+00
j,
         1.09624342e-02+0.00000000e+00j,  9.37136341e-03+0.00000000e+00
j,
         6.91859593e-03+0.00000000e+00j,  3.02374971e-03+0.00000000e+00
j,
         1.58187256e-03+0.00000000e+00j,  5.93806200e-04+0.00000000e+00
j,
         9.34086028e-04+0.00000000e+00j, -3.91161215e-16+0.00000000e+00
j,
         3.47186015e-16+0.00000000e+00j,  1.13968974e-16+8.58639751e-17
j,
         1.13968974e-16-8.58639751e-17j]),
 array([[-2.38709793e-01+0.00000000e+00j,  1.62290697e-02+0.00000000e+0
0j,
         -8.76647456e-02+0.00000000e+00j,  1.18008259e-01+0.00000000e+0
0j,
         -8.76388277e-02+0.00000000e+00j, -6.21462065e-02+0.00000000e+0
0j,
          6.79659636e-02+0.00000000e+00j,  6.43699653e-02+0.00000000e+0
0j,
         -7.64918910e-03+0.00000000e+00j,  1.42143998e-01+0.00000000e+0
0j,
          4.68119251e-01+0.00000000e+00j,  3.26275798e-01+0.00000000e+0
0j,
          4.95961590e-02+0.00000000e+00j, -4.14820126e-01+0.00000000e+0
0j,
         -2.96419040e-02+0.00000000e+00j,  7.89959648e-02+0.00000000e+0
0j,
         -2.07197943e-01+0.00000000e+00j,  4.54316275e-01+0.00000000e+0
0j,
         -3.12838099e-01+0.00000000e+00j,  1.56622125e-01+0.00000000e+0
0j,
         -3.99463042e-02+0.00000000e+00j,  1.34055290e-13+0.00000000e+0
0j,
          9.94161931e-14+0.00000000e+00j,  3.65566735e-14-6.31841784e-1
4j,
          3.65566735e-14+6.31841784e-14j],
        [-2.41900082e-01+0.00000000e+00j, -9.40391173e-02+0.00000000e+0
0j,
```

```
         -1.33624380e-02+0.00000000e+00j, -1.32280714e-01+0.00000000e+0
0j,
         -3.11677026e-02+0.00000000e+00j, -3.14874634e-02+0.00000000e+0
0j,
          8.71613874e-02+0.00000000e+00j, -1.01066088e-02+0.00000000e+0
0j,
         -6.93858859e-03+0.00000000e+00j,  3.87168511e-02+0.00000000e+0
0j,
         -4.94496518e-02+0.00000000e+00j, -3.07176442e-02+0.00000000e+0
0j,
          5.95675463e-02+0.00000000e+00j,  1.38190960e-01+0.00000000e+0
0j,
          4.30457789e-02+0.00000000e+00j,  3.17408168e-02+0.00000000e+0
0j,
          4.10847222e-01+0.00000000e+00j,  9.04784530e-02+0.00000000e+0
0j,
         -1.05775343e-01+0.00000000e+00j,  2.03781233e-01+0.00000000e+0
0j,
         -2.00916820e-01+0.00000000e+00j,  2.23398130e-01+0.00000000e+0
0j,
          5.68749561e-01+0.00000000e+00j, -2.15989241e-01-1.77344167e-0
1j,
         -2.15989241e-01+1.77344167e-01j],
        [-2.44573206e-01+0.00000000e+00j, -4.63346143e-02+0.00000000e+0
0j,
         -5.32336244e-02+0.00000000e+00j, -3.28898712e-02+0.00000000e+0
0j,
         -7.51896600e-02+0.00000000e+00j, -2.94719538e-02+0.00000000e+0
0j,
          1.12553106e-01+0.00000000e+00j,  5.18361136e-02+0.00000000e+0
0j,
          1.93301052e-02+0.00000000e+00j,  1.17650186e-01+0.00000000e+0
0j,
          2.24822988e-01+0.00000000e+00j,  5.99477032e-02+0.00000000e+0
0j,
          2.23860615e-02+0.00000000e+00j, -8.78418174e-02+0.00000000e+0
0j,
          4.02321496e-02+0.00000000e+00j, -1.20768101e-02+0.00000000e+0
0j,
          1.39949897e-01+0.00000000e+00j, -2.38616462e-01+0.00000000e+0
0j,
          2.73125049e-01+0.00000000e+00j, -1.53357266e-01+0.00000000e+0
0j,
          1.93670639e-01+0.00000000e+00j, -6.24629108e-01+0.00000000e+0
0j,
          4.32178176e-01+0.00000000e+00j, -4.51452466e-01+3.59371661e-0
2j,
         -4.51452466e-01-3.59371661e-02j],
        [-2.20937829e-01+0.00000000e+00j, -1.00523300e-01+0.00000000e+0
0j,
          9.64525135e-02+0.00000000e+00j, -2.84972993e-01+0.00000000e+0
0j,
         -4.92306785e-02+0.00000000e+00j, -6.19096038e-02+0.00000000e+0
0j,
          2.84862949e-01+0.00000000e+00j,  9.98996559e-02+0.00000000e+0
0j,
          7.47134132e-02+0.00000000e+00j,  2.25354026e-02+0.00000000e+0
0j,
```

```
         -9.08884876e-02+0.00000000e+00j, -2.67718922e-01+0.00000000e+0
0j,
         -1.00687801e-01+0.00000000e+00j,  8.05590460e-02+0.00000000e+0
0j,
          1.68646026e-01+0.00000000e+00j,  4.50738922e-02+0.00000000e+0
0j,
         -3.06087683e-01+0.00000000e+00j,  3.96541911e-01+0.00000000e+0
0j,
          1.62909246e-01+0.00000000e+00j, -5.10308421e-01+0.00000000e+0
0j,
         -2.86655327e-01+0.00000000e+00j, -3.91272068e-13+0.00000000e+0
0j,
         -3.35124296e-13+0.00000000e+00j, -2.70086509e-14+1.61134627e-1
3j,
         -2.70086509e-14-1.61134627e-13j],
        [-2.20547952e-01+0.00000000e+00j, -9.33814815e-02+0.00000000e+0
0j,
          9.52827253e-02+0.00000000e+00j, -2.88517949e-01+0.00000000e+0
0j,
         -4.27193129e-02+0.00000000e+00j, -6.51777783e-02+0.00000000e+0
0j,
          2.99099279e-01+0.00000000e+00j,  9.46146237e-02+0.00000000e+0
0j,
          7.87682167e-02+0.00000000e+00j, -5.63856609e-03+0.00000000e+0
0j,
         -5.78434692e-02+0.00000000e+00j, -2.53379596e-01+0.00000000e+0
0j,
         -1.69040195e-01+0.00000000e+00j,  4.07160232e-02+0.00000000e+0
0j,
          6.07248610e-02+0.00000000e+00j, -3.21219364e-02+0.00000000e+0
0j,
         -4.07797574e-01+0.00000000e+00j, -2.99793716e-01+0.00000000e+0
0j,
         -2.93739530e-01+0.00000000e+00j,  4.46648368e-01+0.00000000e+0
0j,
          3.08454009e-01+0.00000000e+00j,  3.39414869e-13+0.00000000e+0
0j,
          3.22966719e-13+0.00000000e+00j,  7.70209119e-15-1.41183426e-1
3j,
          7.70209119e-15+1.41183426e-13j],
        [-2.14817474e-01+0.00000000e+00j, -8.22317092e-02+0.00000000e+0
0j,
          1.76882764e-01+0.00000000e+00j, -2.67974025e-02+0.00000000e+0
0j,
          3.41962968e-02+0.00000000e+00j, -7.91540731e-02+0.00000000e+0
0j,
         -4.76369673e-01+0.00000000e+00j,  3.94286838e-01+0.00000000e+0
0j,
         -8.84910084e-03+0.00000000e+00j,  2.70203331e-02+0.00000000e+0
0j,
         -2.28867926e-01+0.00000000e+00j, -6.97256240e-02+0.00000000e+0
0j,
         -3.27155284e-02+0.00000000e+00j,  1.40477445e-02+0.00000000e+0
0j,
         -1.07209848e-02+0.00000000e+00j, -1.25225218e-01+0.00000000e+0
0j,
          1.14736376e-01+0.00000000e+00j,  3.96477665e-01+0.00000000e+0
0j,
```

```
          6.53456295e-02+0.00000000e+00j,  1.18355848e-02+0.00000000e+0
0j,
          5.24763961e-01+0.00000000e+00j,  5.26102737e-14+0.00000000e+0
0j,
          2.34116303e-14+0.00000000e+00j, -3.45787520e-14-1.63959270e-1
4j,
         -3.45787520e-14+1.63959270e-14j],
        [-2.17988628e-01+0.00000000e+00j, -3.24130673e-02+0.00000000e+0
0j,
          1.55443948e-01+0.00000000e+00j,  6.18350213e-02+0.00000000e+0
0j,
          1.04259772e-02+0.00000000e+00j, -1.00183717e-01+0.00000000e+0
0j,
         -4.50805495e-01+0.00000000e+00j,  4.48552947e-01+0.00000000e+0
0j,
          2.44650725e-02+0.00000000e+00j,  1.02647469e-01+0.00000000e+0
0j,
          1.47541079e-02+0.00000000e+00j,  2.06439000e-04+0.00000000e+0
0j,
         -2.56832210e-02+0.00000000e+00j, -9.59978372e-02+0.00000000e+0
0j,
         -2.58684474e-02+0.00000000e+00j,  8.55819465e-02+0.00000000e+0
0j,
         -1.64646069e-01+0.00000000e+00j, -4.13259044e-01+0.00000000e+0
0j,
         -5.55084583e-02+0.00000000e+00j, -6.35108422e-03+0.00000000e+0
0j,
         -5.27862504e-01+0.00000000e+00j, -5.39535047e-14+0.00000000e+0
0j,
         -2.07978421e-14+0.00000000e+00j,  3.28092250e-14+1.58347960e-1
4j,
          3.28092250e-14-1.58347960e-14j],
        [-3.51272864e-02+0.00000000e+00j,  4.78461556e-01+0.00000000e+0
0j,
         -3.49090415e-01+0.00000000e+00j, -3.35599493e-01+0.00000000e+0
0j,
          1.39877719e-01+0.00000000e+00j, -6.28134009e-02+0.00000000e+0
0j,
         -1.04477565e-01+0.00000000e+00j,  3.81513852e-02+0.00000000e+0
0j,
         -1.62273992e-02+0.00000000e+00j,  4.65885208e-02+0.00000000e+0
0j,
         -1.05820175e-01+0.00000000e+00j, -6.57225175e-02+0.00000000e+0
0j,
         -5.80151458e-02+0.00000000e+00j, -8.55603901e-02+0.00000000e+0
0j,
         -1.27481309e-01+0.00000000e+00j,  6.63806581e-01+0.00000000e+0
0j,
          2.23694346e-02+0.00000000e+00j,  1.54517859e-02+0.00000000e+0
0j,
          6.71535802e-02+0.00000000e+00j,  2.22927383e-02+0.00000000e+0
0j,
          9.05162467e-02+0.00000000e+00j,  2.47535960e-14+0.00000000e+0
0j,
          2.10281631e-14+0.00000000e+00j, -9.67390520e-15-1.31353627e-1
4j,
         -9.67390520e-15+1.31353627e-14j],
```

```
        [-3.56584288e-02+0.00000000e+00j,  4.88306682e-01+0.00000000e+0
0j,
         -3.50657157e-01+0.00000000e+00j, -3.06504948e-01+0.00000000e+0
0j,
          1.32026114e-01+0.00000000e+00j, -7.58177727e-02+0.00000000e+0
0j,
         -9.75274163e-02+0.00000000e+00j,  6.27952256e-02+0.00000000e+0
0j,
          3.38962757e-02+0.00000000e+00j,  4.45930049e-02+0.00000000e+0
0j,
          6.43370801e-02+0.00000000e+00j,  3.78233329e-02+0.00000000e+0
0j,
          5.52785812e-02+0.00000000e+00j,  4.05633409e-02+0.00000000e+0
0j,
          1.58556675e-01+0.00000000e+00j, -6.73370022e-01+0.00000000e+0
0j,
         -2.18606466e-02+0.00000000e+00j, -1.07732568e-02+0.00000000e+0
0j,
         -7.09343280e-02+0.00000000e+00j, -2.33139093e-02+0.00000000e+0
0j,
         -9.27374295e-02+0.00000000e+00j, -2.49863597e-14+0.00000000e+0
0j,
         -2.05865366e-14+0.00000000e+00j,  9.88186874e-15+1.30116376e-1
4j,
          9.88186874e-15-1.30116376e-14j],
        [-2.40051149e-01+0.00000000e+00j, -1.09986449e-01+0.00000000e+0
0j,
         -7.76706996e-02+0.00000000e+00j, -8.11420277e-02+0.00000000e+0
0j,
         -6.95207195e-02+0.00000000e+00j,  6.83852637e-03+0.00000000e+0
0j,
          5.43394286e-02+0.00000000e+00j,  5.49152595e-03+0.00000000e+0
0j,
         -2.54220667e-02+0.00000000e+00j,  2.03074693e-01+0.00000000e+0
0j,
          1.25114071e-02+0.00000000e+00j, -8.25836103e-02+0.00000000e+0
0j,
          8.31932251e-02+0.00000000e+00j,  3.16134861e-01+0.00000000e+0
0j,
          2.29326619e-01+0.00000000e+00j,  9.69233198e-02+0.00000000e+0
0j,
          4.46663806e-01+0.00000000e+00j,  1.16864493e-01+0.00000000e+0
0j,
         -1.53744073e-01+0.00000000e+00j,  2.47770907e-01+0.00000000e+0
0j,
         -2.06964090e-01+0.00000000e+00j, -1.70202845e-01+0.00000000e+0
0j,
         -4.33319622e-01+0.00000000e+00j,  1.64558151e-01+1.35115194e-0
1j,
          1.64558151e-01-1.35115194e-01j],
        [-2.25970766e-01+0.00000000e+00j, -1.25864235e-01+0.00000000e+0
0j,
         -1.87476114e-01+0.00000000e+00j, -2.01677716e-02+0.00000000e+0
0j,
         -1.57693833e-01+0.00000000e+00j,  2.65050831e-02+0.00000000e+0
0j,
          1.15439514e-01+0.00000000e+00j,  2.86759884e-02+0.00000000e+0
0j,
```

```
         -5.98994561e-02+0.00000000e+00j,  4.07601638e-01+0.00000000e+0
0j,
          2.10331628e-01+0.00000000e+00j, -6.25981241e-02+0.00000000e+0
0j,
         -2.32709764e-01+0.00000000e+00j, -1.47671528e-01+0.00000000e+0
0j,
         -2.11275559e-01+0.00000000e+00j, -1.09663205e-01+0.00000000e+0
0j,
          1.81679253e-01+0.00000000e+00j, -2.35048847e-01+0.00000000e+0
0j,
          3.02773552e-01+0.00000000e+00j, -1.63189673e-01+0.00000000e+0
0j,
          1.71678223e-01+0.00000000e+00j,  4.12592744e-01+0.00000000e+0
0j,
         -2.85471134e-01+0.00000000e+00j,  2.98202581e-01-2.37379491e-0
2j,
          2.98202581e-01+2.37379491e-02j],
        [-2.18424186e-01+0.00000000e+00j, -3.79126499e-02+0.00000000e+0
0j,
          1.69578418e-01+0.00000000e+00j, -2.60808272e-01+0.00000000e+0
0j,
          8.06884038e-02+0.00000000e+00j, -1.35836259e-01+0.00000000e+0
0j,
          1.69385144e-01+0.00000000e+00j, -5.28963074e-02+0.00000000e+0
0j,
          4.60096922e-02+0.00000000e+00j, -4.29381405e-01+0.00000000e+0
0j,
         -2.18321447e-01+0.00000000e+00j,  1.19193016e-01+0.00000000e+0
0j,
         -1.41242321e-02+0.00000000e+00j, -3.80019336e-01+0.00000000e+0
0j,
         -4.87391051e-01+0.00000000e+00j, -1.55845222e-01+0.00000000e+0
0j,
          2.61112985e-01+0.00000000e+00j,  5.33604105e-03+0.00000000e+0
0j,
          4.20467741e-02+0.00000000e+00j,  5.55566773e-02+0.00000000e+0
0j,
         -1.60033545e-01+0.00000000e+00j, -6.03533780e-02+0.00000000e+0
0j,
         -1.53653736e-01+0.00000000e+00j,  5.83517880e-02+4.79114107e-0
2j,
          5.83517880e-02-4.79114107e-02j],
        [-2.29798146e-01+0.00000000e+00j,  8.87357981e-02+0.00000000e+0
0j,
          1.70224774e-01+0.00000000e+00j, -4.71800194e-02+0.00000000e+0
0j,
          6.98565250e-02+0.00000000e+00j, -1.13270253e-01+0.00000000e+0
0j,
          8.75226821e-02+0.00000000e+00j,  7.93102100e-02+0.00000000e+0
0j,
          1.42001492e-01+0.00000000e+00j, -3.65484245e-01+0.00000000e+0
0j,
          2.07086206e-01+0.00000000e+00j,  2.44230451e-01+0.00000000e+0
0j,
          4.24586570e-01+0.00000000e+00j,  2.33900539e-02+0.00000000e+0
0j,
          4.33478356e-01+0.00000000e+00j,  1.45531390e-01+0.00000000e+0
0j,
```

```
          4.80845774e-02+0.00000000e+00j, -2.00978420e-01+0.00000000e+0
0j,
          1.76322262e-01+0.00000000e+00j, -1.09856851e-01+0.00000000e+0
0j,
          1.93535587e-01+0.00000000e+00j,  2.59069301e-01+0.00000000e+0
0j,
         -1.79248928e-01+0.00000000e+00j,  1.87243075e-01-1.49051915e-0
2j,
          1.87243075e-01+1.49051915e-02j],
        [-2.41797250e-01+0.00000000e+00j, -7.90420705e-02+0.00000000e+0
0j,
         -8.95299212e-02+0.00000000e+00j, -2.30256537e-02+0.00000000e+0
0j,
          5.75280609e-03+0.00000000e+00j, -1.73816018e-02+0.00000000e+0
0j,
         -6.46786018e-02+0.00000000e+00j, -1.52599101e-01+0.00000000e+0
0j,
         -1.34253603e-01+0.00000000e+00j,  3.16676979e-02+0.00000000e+0
0j,
         -9.54910810e-02+0.00000000e+00j,  1.69015580e-01+0.00000000e+0
0j,
          1.80782307e-01+0.00000000e+00j,  2.65381062e-01+0.00000000e+0
0j,
         -2.68063980e-01+0.00000000e+00j,  3.37536732e-02+0.00000000e+0
0j,
         -3.11967576e-02+0.00000000e+00j, -1.67570098e-01+0.00000000e+0
0j,
         -5.86621472e-01+0.00000000e+00j, -5.13085199e-01+0.00000000e+0
0j,
          1.72287086e-01+0.00000000e+00j, -2.16103569e-13+0.00000000e+0
0j,
         -4.11946048e-13+0.00000000e+00j,  7.89391541e-14+2.40056001e-1
3j,
          7.89391541e-14-2.40056001e-13j],
        [-2.25614542e-01+0.00000000e+00j,  1.77987002e-01+0.00000000e+0
0j,
         -5.57167231e-02+0.00000000e+00j,  2.01428618e-01+0.00000000e+0
0j,
         -7.35201847e-02+0.00000000e+00j,  8.83568983e-02+0.00000000e+0
0j,
          3.62445204e-02+0.00000000e+00j,  9.96060743e-02+0.00000000e+0
0j,
          1.33318720e-01+0.00000000e+00j, -2.01631931e-01+0.00000000e+0
0j,
          2.39639785e-01+0.00000000e+00j,  1.87582990e-02+0.00000000e+0
0j,
          5.34007342e-02+0.00000000e+00j,  5.80659808e-01+0.00000000e+0
0j,
         -4.82422294e-01+0.00000000e+00j, -2.17699067e-02+0.00000000e+0
0j,
         -2.02039213e-01+0.00000000e+00j,  1.17522694e-01+0.00000000e+0
0j,
          2.66464651e-01+0.00000000e+00j,  1.78490272e-01+0.00000000e+0
0j,
         -4.86722502e-02+0.00000000e+00j,  5.33477098e-14+0.00000000e+0
0j,
          1.28100978e-13+0.00000000e+00j, -3.86504725e-14-7.86027908e-1
4j,
```

```
         -3.86504725e-14+7.86027908e-14j],
        [-2.35359144e-01+0.00000000e+00j, -8.16965531e-02+0.00000000e+0
0j,
         -1.38074688e-01+0.00000000e+00j,  5.19484729e-02+0.00000000e+0
0j,
         -2.06298288e-02+0.00000000e+00j,  1.97344204e-02+0.00000000e+0
0j,
         -9.11874930e-02+0.00000000e+00j, -2.49550664e-01+0.00000000e+0
0j,
         -2.06688021e-01+0.00000000e+00j,  6.67349980e-02+0.00000000e+0
0j,
         -2.71175782e-01+0.00000000e+00j,  9.93626515e-02+0.00000000e+0
0j,
          1.69087279e-01+0.00000000e+00j, -5.12132633e-02+0.00000000e+0
0j,
          6.56693027e-02+0.00000000e+00j, -2.81257090e-02+0.00000000e+0
0j,
         -2.14823097e-01+0.00000000e+00j,  6.93016002e-03+0.00000000e+0
0j,
          2.03140758e-01+0.00000000e+00j,  1.32542975e-01+0.00000000e+0
0j,
         -4.29071217e-02+0.00000000e+00j, -2.69387064e-01+0.00000000e+0
0j,
         -2.50488915e-01+0.00000000e+00j,  5.12975847e-01+0.00000000e+0
0j,
          5.12975847e-01-0.00000000e+00j],
        [-2.13711260e-01+0.00000000e+00j,  1.67367642e-01+0.00000000e+0
0j,
         -1.09949099e-01+0.00000000e+00j,  3.16826986e-01+0.00000000e+0
0j,
         -1.01977154e-01+0.00000000e+00j,  1.31880849e-01+0.00000000e+0
0j,
          1.19371039e-02+0.00000000e+00j, -2.84960681e-02+0.00000000e+0
0j,
          4.03682070e-02+0.00000000e+00j, -2.09294334e-01+0.00000000e+0
0j,
         -7.62154731e-02+0.00000000e+00j, -2.17463173e-01+0.00000000e+0
0j,
         -1.55131479e-01+0.00000000e+00j, -1.21491903e-01+0.00000000e+0
0j,
          1.18421364e-01+0.00000000e+00j, -7.43087099e-03+0.00000000e+0
0j,
          6.97527671e-02+0.00000000e+00j, -1.95543150e-02+0.00000000e+0
0j,
         -9.88614714e-02+0.00000000e+00j, -6.18216442e-02+0.00000000e+0
0j,
          7.54427449e-03+0.00000000e+00j,  3.16434343e-01+0.00000000e+0
0j,
          1.83645062e-01+0.00000000e+00j, -2.09323890e-02+2.20966574e-0
1j,
         -2.09323890e-02-2.20966574e-01j],
        [-1.42611235e-01+0.00000000e+00j,  1.70396692e-01+0.00000000e+0
0j,
          3.94896798e-01+0.00000000e+00j, -2.19112149e-01+0.00000000e+0
0j,
          1.86011693e-02+0.00000000e+00j,  2.88436132e-01+0.00000000e+0
0j,
```

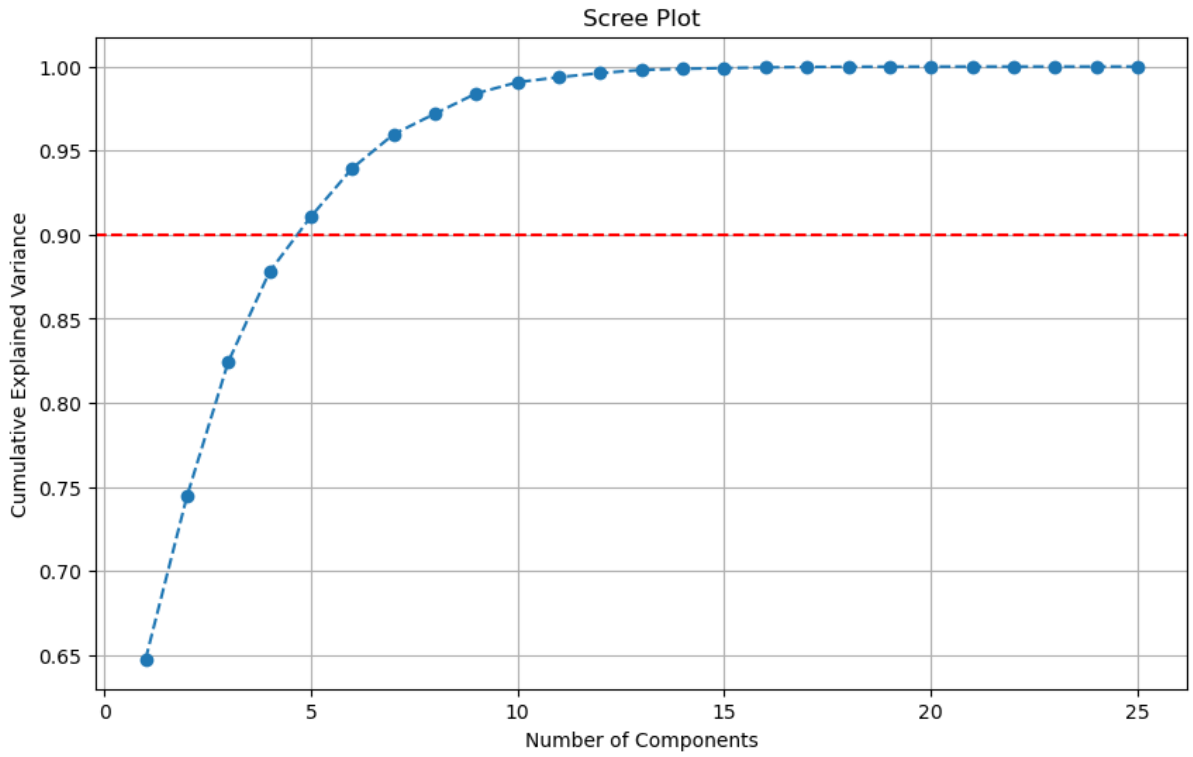
```
         -2.32930642e-01+0.00000000e+00j, -3.18076692e-01+0.00000000e+0
0j,
         -5.54065911e-01+0.00000000e+00j, -5.41535912e-02+0.00000000e+0
0j,
          3.22587536e-01+0.00000000e+00j, -2.70243438e-01+0.00000000e+0
0j,
          4.39802918e-02+0.00000000e+00j, -3.54291761e-02+0.00000000e+0
0j,
         -5.47832305e-03+0.00000000e+00j, -1.75651198e-02+0.00000000e+0
0j,
         -3.14030050e-02+0.00000000e+00j,  1.26161837e-02+0.00000000e+0
0j,
          4.25212923e-02+0.00000000e+00j,  3.11405714e-02+0.00000000e+0
0j,
         -1.71865444e-02+0.00000000e+00j,  4.05537419e-02+0.00000000e+0
0j,
          3.77087996e-02+0.00000000e+00j, -7.72237901e-02-2.34833477e-1
4j,
         -7.72237901e-02+2.34833477e-14j],
        [-1.10946487e-01+0.00000000e+00j,  3.19063353e-01+0.00000000e+0
0j,
          2.69405829e-01+0.00000000e+00j, -2.67476427e-02+0.00000000e+0
0j,
         -2.48559509e-01+0.00000000e+00j,  6.61810749e-01+0.00000000e+0
0j,
          1.01629745e-01+0.00000000e+00j,  6.16041808e-02+0.00000000e+0
0j,
          2.49696078e-01+0.00000000e+00j,  1.82878223e-01+0.00000000e+0
0j,
         -2.81843303e-01+0.00000000e+00j,  3.01694672e-01+0.00000000e+0
0j,
         -6.42589220e-02+0.00000000e+00j, -5.52487009e-02+0.00000000e+0
0j,
          4.96872036e-02+0.00000000e+00j, -5.94101242e-04+0.00000000e+0
0j,
          2.43886582e-02+0.00000000e+00j, -2.08980717e-02+0.00000000e+0
0j,
         -4.22301070e-02+0.00000000e+00j, -2.44936274e-02+0.00000000e+0
0j,
          9.72581221e-04+0.00000000e+00j, -5.61847234e-02+0.00000000e+0
0j,
         -3.26072287e-02+0.00000000e+00j,  3.71666512e-03-3.92338762e-0
2j,
          3.71666512e-03+3.92338762e-02j],
        [-1.72790685e-01+0.00000000e+00j,  2.40585193e-01+0.00000000e+0
0j,
          2.56907430e-01+0.00000000e+00j,  1.81596063e-01+0.00000000e+0
0j,
          9.36149711e-02+0.00000000e+00j, -4.37726183e-01+0.00000000e+0
0j,
          5.57412156e-02+0.00000000e+00j, -2.38725403e-01+0.00000000e+0
0j,
         -1.87622687e-01+0.00000000e+00j,  1.19148943e-01+0.00000000e+0
0j,
         -1.63568570e-01+0.00000000e+00j,  3.88808500e-01+0.00000000e+0
0j,
         -4.87240421e-01+0.00000000e+00j,  1.81488328e-01+0.00000000e+0
0j,
```

```
          1.46829356e-01+0.00000000e+00j,  7.27389347e-03+0.00000000e+0
0j,
         -3.37957352e-02+0.00000000e+00j, -4.57046068e-03+0.00000000e+0
0j,
          8.72802469e-02+0.00000000e+00j,  3.16532387e-02+0.00000000e+0
0j,
         -1.27273569e-02+0.00000000e+00j,  5.47615793e-02+0.00000000e+0
0j,
          5.09199232e-02+0.00000000e+00j, -1.04278829e-01-3.50486022e-1
4j,
         -1.04278829e-01+3.50486022e-14j],
        [-1.34034463e-01+0.00000000e+00j,  3.63978211e-01+0.00000000e+0
0j,
          1.48355950e-01+0.00000000e+00j,  4.20676847e-01+0.00000000e+0
0j,
         -1.21286896e-01+0.00000000e+00j, -2.76204686e-01+0.00000000e+0
0j,
          9.57699129e-02+0.00000000e+00j, -1.31861508e-01+0.00000000e+0
0j,
          1.54250464e-01+0.00000000e+00j,  1.44185464e-01+0.00000000e+0
0j,
         -1.09114861e-01+0.00000000e+00j, -4.84288957e-01+0.00000000e+0
0j,
          2.81224568e-01+0.00000000e+00j, -1.77997034e-01+0.00000000e+0
0j,
         -6.03799930e-02+0.00000000e+00j, -3.39825503e-02+0.00000000e+0
0j,
          6.77737446e-02+0.00000000e+00j, -6.02650179e-03+0.00000000e+0
0j,
         -8.29333264e-02+0.00000000e+00j, -3.64984033e-02+0.00000000e+0
0j,
          2.29788009e-02+0.00000000e+00j, -1.35697938e-01+0.00000000e+0
0j,
         -7.87533236e-02+0.00000000e+00j,  8.97652889e-03-9.47580727e-0
2j,
          8.97652889e-03+9.47580727e-02j],
        [-1.84217020e-01+0.00000000e+00j, -1.61006580e-01+0.00000000e+0
0j,
          3.24970967e-02+0.00000000e+00j, -5.40157997e-02+0.00000000e+0
0j,
          4.09028148e-01+0.00000000e+00j,  9.38716756e-02+0.00000000e+0
0j,
         -3.11567095e-01+0.00000000e+00j, -4.83130764e-01+0.00000000e+0
0j,
          6.15544504e-01+0.00000000e+00j,  1.00270489e-01+0.00000000e+0
0j,
          1.29432985e-01+0.00000000e+00j, -7.11279322e-02+0.00000000e+0
0j,
         -1.22163619e-01+0.00000000e+00j, -2.12305691e-02+0.00000000e+0
0j,
          1.48706567e-02+0.00000000e+00j,  8.31194577e-03+0.00000000e+0
0j,
         -4.04917402e-02+0.00000000e+00j,  1.55592402e-02+0.00000000e+0
0j,
          1.55045130e-02+0.00000000e+00j,  7.33949897e-03+0.00000000e+0
0j,
         -4.21967190e-03+0.00000000e+00j,  1.09415403e-02+0.00000000e+0
0j,
```

```
          1.01739650e-02+0.00000000e+00j, -2.08352465e-02-7.87943825e-1
5j,
         -2.08352465e-02+7.87943825e-15j],
        [-1.28232746e-01+0.00000000e+00j, -3.99156365e-03+0.00000000e+0
0j,
         -3.03101749e-02+0.00000000e+00j,  2.71384805e-01+0.00000000e+0
0j,
          7.79900106e-01+0.00000000e+00j,  2.67565417e-01+0.00000000e+0
0j,
          3.05301846e-01+0.00000000e+00j,  2.58489458e-01+0.00000000e+0
0j,
         -2.20450555e-01+0.00000000e+00j,  7.61942644e-02+0.00000000e+0
0j,
         -5.23794311e-02+0.00000000e+00j, -5.20654304e-02+0.00000000e+0
0j,
          7.98133336e-03+0.00000000e+00j, -3.51301316e-02+0.00000000e+0
0j,
         -3.00347596e-04+0.00000000e+00j, -1.58822979e-02+0.00000000e+0
0j,
          2.33783422e-02+0.00000000e+00j, -9.07325324e-04+0.00000000e+0
0j,
         -1.34501015e-02+0.00000000e+00j, -8.17135231e-03+0.00000000e+0
0j,
          1.09252604e-03+0.00000000e+00j, -3.35378063e-02+0.00000000e+0
0j,
         -1.94639193e-02+0.00000000e+00j,  2.21855314e-03-2.34195002e-0
2j,
          2.21855314e-03+2.34195002e-02j],
        [-2.06845216e-01+0.00000000e+00j, -1.80799828e-01+0.00000000e+0
0j,
         -3.03194913e-01+0.00000000e+00j,  6.06377832e-02+0.00000000e+0
0j,
         -7.13386533e-02+0.00000000e+00j,  7.42473635e-02+0.00000000e+0
0j,
         -6.61758244e-02+0.00000000e+00j, -1.61236472e-01+0.00000000e+0
0j,
         -1.33006924e-01+0.00000000e+00j,  5.62678274e-02+0.00000000e+0
0j,
         -3.52318061e-01+0.00000000e+00j,  7.71628689e-02+0.00000000e+0
0j,
          3.21802226e-01+0.00000000e+00j, -9.89173679e-02+0.00000000e+0
0j,
          4.35254575e-02+0.00000000e+00j, -3.29653452e-02+0.00000000e+0
0j,
         -2.43433953e-01+0.00000000e+00j,  6.43426989e-03+0.00000000e+0
0j,
          2.15400599e-01+0.00000000e+00j,  1.46270290e-01+0.00000000e+0
0j,
         -4.53598120e-02+0.00000000e+00j,  2.23072022e-01+0.00000000e+0
0j,
          2.07422985e-01+0.00000000e+00j, -4.24781197e-01-1.19360856e-1
3j,
         -4.24781197e-01+1.19360856e-13j],
        [-1.94392948e-01+0.00000000e+00j, -7.10691453e-02+0.00000000e+0
0j,
         -3.44997380e-01+0.00000000e+00j,  1.77742891e-01+0.00000000e+0
0j,
```

```
        -1.39919207e-01+0.00000000e+00j,  1.65169406e-01+0.00000000e+0
0j,
        -1.25758584e-01+0.00000000e+00j, -1.62606062e-02+0.00000000e+0
0j,
        -7.39194336e-02+0.00000000e+00j, -4.92807235e-01+0.00000000e+0
0j,
         4.13810025e-02+0.00000000e+00j, -9.14433977e-02+0.00000000e+0
0j,
        -4.19273317e-01+0.00000000e+00j, -5.00085914e-02+0.00000000e+0
0j,
         2.14286331e-01+0.00000000e+00j,  1.41211274e-02+0.00000000e+0
0j,
         5.35714616e-02+0.00000000e+00j, -2.08132585e-02+0.00000000e+0
0j,
        -8.59730399e-02+0.00000000e+00j, -6.47565256e-02+0.00000000e+0
0j,
        -4.10836268e-03+0.00000000e+00j, -2.00126803e-01+0.00000000e+0
0j,
        -1.16145102e-01+0.00000000e+00j,  1.32385506e-02-1.39748845e-0
1j,
         1.32385506e-02+1.39748845e-01j]]))
```

**Part 2 - PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.**




The Scree plot showcases the cumulative explained variance against the number of components. The red dashed line represents the 90% variance threshold.

num_components = np.where (cumulative_variance >= 0.9)[0][0] + 1

Num_components From the plot, we observe that we need 5 principal components to capture at least 90% of the total variance in the dataset.

**#Part 2 - PCA: Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.**

```
PC1          M_ST
PC2          F_ST
PC3     MAIN_CL_M
PC4     MAIN_AL_F
PC5     MAIN_HH_F
dtype: object
```

Here are the variables that contribute the most to each of the first 5 principal components:

PC1 (Principal Component 1): The variable M_ST (Male belonging to Scheduled Tribes) contributes the most.

PC2: The variable F_ST (Female belonging to Scheduled Tribes) contributes the most.

PC3: The variable MAIN_CL_M (Main workers who are Cultivators - Male) contributes the most.

PC4: The variable MAIN_AL_F (Main workers in Household industries - Female) contributes the most.

PC5: The variable MAIN_HH_F (Main workers in other services - Female) contributes the most.

This means that these variables have the highest weights in their respective principal components and hence explain the most variance in that direction.

**Inferences:**

PC1: This component is mainly influenced by the male population belonging to Scheduled Tribes. It could represent the variance in the distribution of the ST male population across districts.

PC2: This component captures the variance related to the female population of Scheduled Tribes. It can provide insights into the districts with higher or lower representation of ST females.

PC3: This component captures the variance related to male main workers who are cultivators. It could help identify districts with more agricultural activities by males.

PC4: Representing the variance associated with female main workers in household industries, this component can provide insights into the districts with more home-based industries or businesses run by females.

PC5: This component captures the variance related to female main workers in other services. It might represent the diversity of professions among females in various districts.

**PCA: Write linear equation for first PC.**

```
'PC1 = -0.2387+0.0000j * No_HH + -0.2419+0.0000j * TOT_M + -0.2446+0.00
00j * TOT_F + -0.2209+0.0000j * M_06 + -0.2205+0.0000j * F_06 + -0.2148
+0.0000j * M_SC + -0.2180+0.0000j * F_SC + -0.0351+0.0000j * M_ST + -0.
0357+0.0000j * F_ST + -0.2401+0.0000j * M_LIT + -0.2260+0.0000j * F_LIT
+ -0.2184+0.0000j * M_ILL + -0.2298+0.0000j * F_ILL + -0.2418+0.0000j *
TOT_WORK_M + -0.2256+0.0000j * TOT_WORK_F + -0.2354+0.0000j * MAINWORK_
M + -0.2137+0.0000j * MAINWORK_F + -0.1426+0.0000j * MAIN_CL_M + -0.110
9+0.0000j * MAIN_CL_F + -0.1728+0.0000j * MAIN_AL_M + -0.1340+0.0000j *
MAIN_AL_F + -0.1842+0.0000j * MAIN_HH_M + -0.1282+0.0000j * MAIN_HH_F +
-0.2068+0.0000j * MAIN_OT_M + -0.1944+0.0000j * MAIN_OT_F
```

PC1 = -0.2387 \times \text{No_HH} - 0.2419 \times \text{TOT_M} - 0.2446 \times \text{TOT_F} \ldots - 0.1944 \times \text{MAIN_OT_F}

This equation represents how the first principal component is derived from the original variables. The weights (coefficients) in this equation come from the first eigenvector, and they indicate the importance and direction (positive or negative) of each variable in forming PC1.