Machine Learning 1

Coded Project -Suraj Mishra

Problem Statement:

Clustering

Digital Ads Data:

The ads24x7 is a Digital Marketing company which has now got seed funding of $10 Million. They are expanding their wings in Marketing Analytics. They collected data from their Marketing Intelligence team and now wants you (their newly appointed data analyst) to segment type of ads based on the features provided. Use Clustering procedure to segment ads into homogeneous groups.

PCA:

PCA FH (FT): Primary census abstract for female headed households excluding institutional households (India & States/UTs - District Level), Scheduled tribes - 2011 PCA for Female Headed Household Excluding Institutional Household. The Indian Census has the reputation of being one of the best in the world. The first Census in India was conducted in the year 1872. This was conducted at different points of time in different parts of the country. In 1881 a Census was taken for the entire country simultaneously. Since then, Census has been conducted every ten years, without a break. Thus, the Census of India 2011 was the fifteenth in this unbroken series since 1872, the seventh after independence and the second census of the third millennium and twenty first century. The census has been uninterruptedly continued despite of several adversities like wars, epidemics, natural calamities, political unrest, etc. The Census of India is conducted under the provisions of the Census Act 1948 and the Census Rules, 1990. The Primary Census Abstract which is important publication of 2011 Census gives basic information on Area, Total Number of Households, Total Population, Scheduled Castes, Scheduled Tribes Population, Population in the age group 0-6, Literates, Main Workers and Marginal Workers classified by the four broad industrial categories, namely, (i) Cultivators, (ii) Agricultural Laborers, (iii) Household Industry Workers, and (iv) Other Workers and also Non-Workers. The characteristics of the Total Population include Scheduled Castes, Scheduled Tribes, Institutional and Houseless Population and are presented by sex and rural-urban residence. Census 2011 covered 35 States/Union Territories, 640 districts, 5,924 sub-districts, 7,935 Towns and 6,40,867 Villages.

The data collected has so many variables thus making it difficult to find useful details without using Data Science Techniques. You are tasked to perform detailed EDA and identify Optimum Principal Components that explains the most variance in data. Use Sklearn only.

Contents

Problem 1 - Define the problem and perform Exploratory Data Analysis

- Problem definition - Check shape, Data types, statistical summary - Univariate analysis - Bivariate analysis - Key meaningful observations on individual variables and the relationship between variables

6.5

Problem 1 - Data Preprocessing

- Missing value check and treatment - Outlier Treatment - z-score scaling Note: Treat missing values in CPC, CTR and CPM using the formula given.

2.5

Problem 1 - Hierarchical Clustering

- Construct a dendrogram using Ward linkage and Euclidean distance - Identify the optimum number of Clusters

4

Problem 1 - K-means Clustering

- Apply K-means Clustering - Plot the Elbow curve - Check Silhouette Scores - Figure out the appropriate number of clusters - Cluster Profiling

13

Problem 1 - Actionable Insights & Recommendations

- Extract meaningful insights (atleast 3) from the clusters to identify the most effective types of ads, target audiences, or marketing strategies that can be inferred from each segment. - Based on the clustering analysis and key insights, provide actionable recommendations (atleast 3) to Ads24x7 on how to optimize their digital marketing efforts, allocate budgets efficiently, and tailor ad content to specific audience segments.

6

Problem 2 - Define the problem and perform Exploratory Data Analysis

- Problem Definition - Check shape, Data types, statistical summary - Perform an EDA on the data to extract useful insights Note: 1. 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 2. Example questions to answer from EDA - (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio?

6.5

Problem 2 - Data Preprocessing

- Check for and treat (if needed) missing values - Check for and treat (if needed) data irregularities - Scale the Data using the z-score method - Visualize the data before and after scaling and comment on the impact on outliers

2.5

Problem 2 - PCA

- Create the covariance matrix - Get eigen values and eigen vectors - Identify the optimum number of PCs - Show Scree plot - Compare PCs with Actual Columns and identify which is explaining most variance - Write inferences about all the PCs in terms of actual variables - Write linear equation for first PC Note: For the scope of this project, take at least 90% explained variance.

Problem 1 - Define the problem and perform Exploratory Data Analysis

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

Head of the Data Frame

	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.0
2	2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0
3	2020-9-3-20	Format1	300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.0
4	2020-9-4-15	Format1	300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.0

Tail of the Data Frame

	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Sp
23061	2020-9-13-7	Format5	720	300	216000	Inter220	Web	Mobile	Video	1	1	1	1	
23062	2020-11-2-7	Format5	720	300	216000	Inter224	Web	Desktop	Video	3	2	2	1	
23063	2020-9-14-22	Format5	720	300	216000	Inter218	App	Mobile	Video	2	1	1	1	
23064	2020-11-18-2	Format4	120	600	72000	inter230	Video	Mobile	Video	7	1	1	1	
23065	2020-9-14-0	Format5	720	300	216000	Inter221	App	Mobile	Video	2	2	2	1	

The Data Frame has following Data Types

```
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)
```

Summary of data types:

- Float64: 6 columns
- Int64: 7 columns
- Object: 6 columns

Checking for Null 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
```

Minimum values for several variables are 0. There are no negative Values. The CTR, CPM and CPC are derived fields and have missing values .

Descriptive statistics of the data frame

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revi
count	23066.000000	23066.000000	23066.000000	2.306600e+04	2.306600e+04	2.306600e+04	23066.000000	23066.000000	23066.000000	23066.00
mean	385.163097	337.896037	96674.468048	2.432044e+06	1.295099e+06	1.241520e+06	10678.518816	2706.625689	0.335123	1924.25
std	233.651434	203.092885	61538.329557	4.742888e+06	2.512970e+06	2.429400e+06	17353.409363	4067.927273	0.031963	3105.23
min	120.000000	70.000000	33600.000000	1.000000e+00	1.000000e+00	1.000000e+00	1.000000	0.000000	0.210000	0.00
25%	120.000000	250.000000	72000.000000	3.367225e+04	1.828250e+04	7.990500e+03	710.000000	85.180000	0.330000	55.36
50%	300.000000	300.000000	72000.000000	4.837710e+05	2.580875e+05	2.252900e+05	4425.000000	1425.125000	0.350000	926.33
75%	720.000000	600.000000	84000.000000	2.527712e+06	1.180700e+06	1.112428e+06	12793.750000	3121.400000	0.350000	2091.33
max	728.000000	600.000000	216000.000000	2.759286e+07	1.470202e+07	1.419477e+07	143049.000000	26931.870000	0.350000	21276.18

Duplicate Values

There are no missing values as per the information.

Problem 1 - Data Preprocessing

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

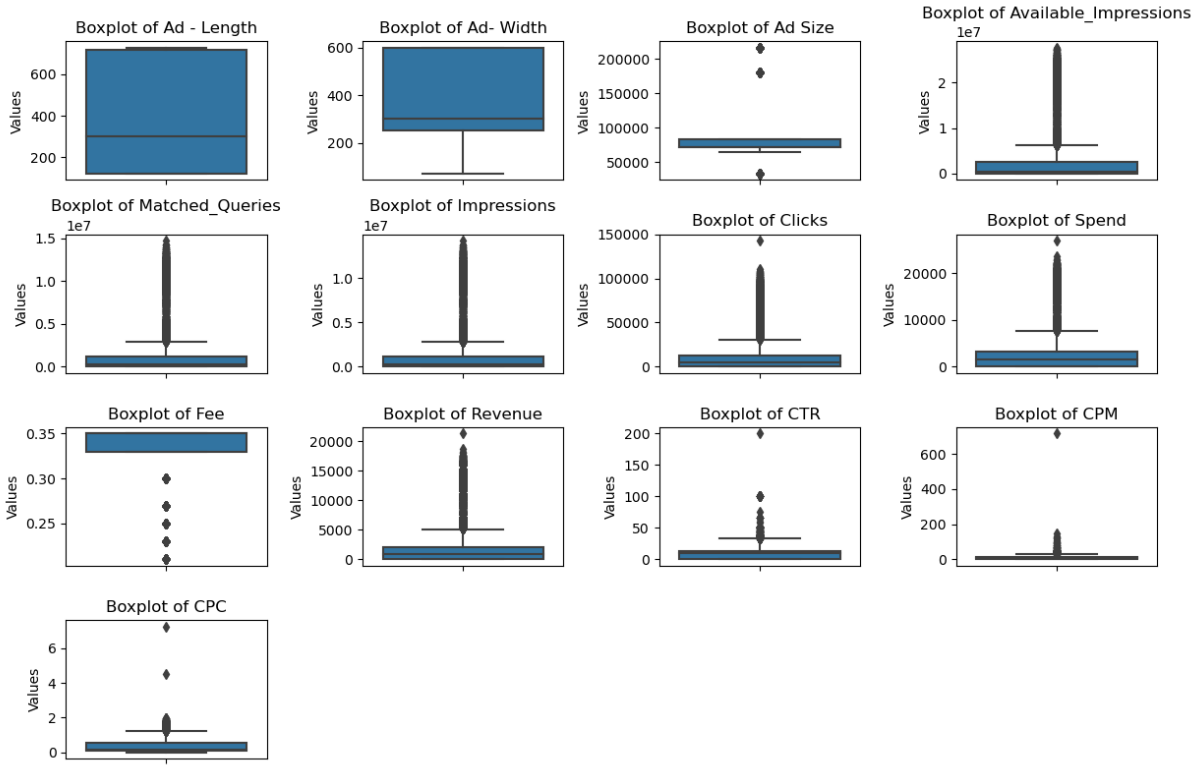
CPM = (Total Campaign Spend / Number of Impressions) * 1,000

CTR = Total Measured Clicks / Total Measured Ad Impressions x 100

CPC = Total Cost / Number of Clicks

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).

Boxplots showing Outliers



the k-means clustering technique is widely used and studied, it may encounter limitations when applied to real-world data. One significant challenge arises from the assumption that all data points can be neatly divided into a predetermined number of clusters. However, in practice, real data often contains noise or outliers, making it difficult to achieve clear and distinct partitions.

This sensitivity to noise can greatly impact the quality of the clustering solution produced by k-means. Therefore, when designing clustering algorithms, it's crucial to consider methods for handling noise and contamination in the data to ensure more robust and reliable partitioning results.

Approaches to reduce noise in data:

1. Outlier Treatment Using IQR Method: This method involves identifying outliers based on the Interquartile Range (IQR) and then treating them accordingly. Outliers are defined as observations that fall below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR. To treat outliers, a function such as "remove outlier" can be employed. This function replaces larger values (beyond the upper whisker) with the 95th percentile value of the distribution, and smaller values (beyond the lower whisker) with the 5th percentile value.

2. Outlier Treatment Using Z-Score Method: Another approach is to identify outliers based on their deviation from the mean in terms of standard deviations (z-scores).

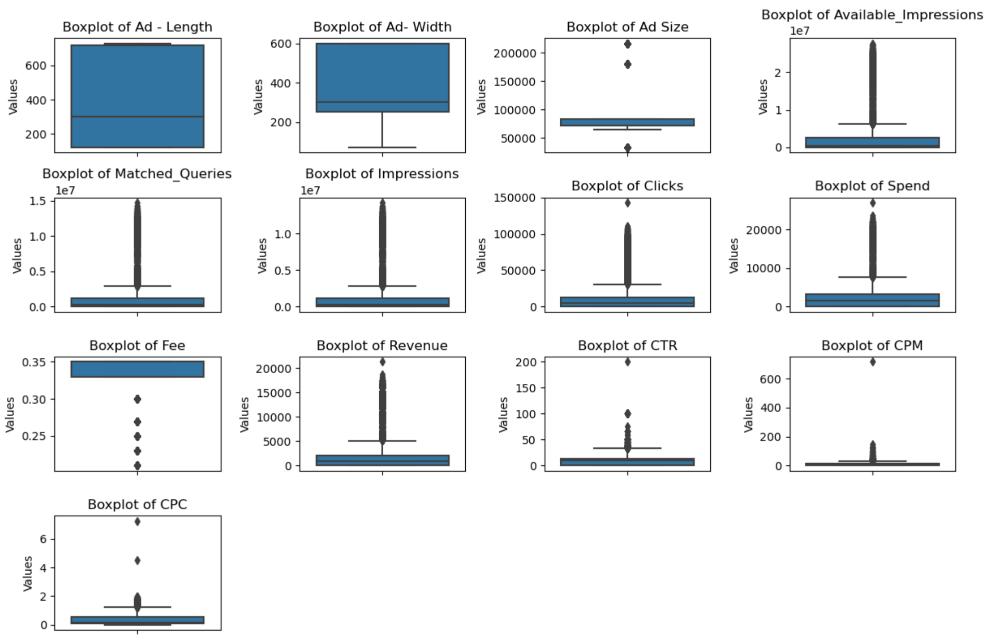
Observations with z-scores beyond a certain threshold (e.g., ±3 standard deviations) are considered outliers and can be replaced or removed.

3. Segmentation Based on EDA Results: Utilizing insights from Exploratory Data Analysis (EDA), data can be segmented into multiple parts based on relevant features. For instance, in Bank Customer Segmentation, high net worth individuals may be separated from low-income individuals. Separate clustering models, such as k-means, can then be applied to each segment individually. This approach is feasible for large datasets with sufficient data points in each segment.

For this particular dataset, we will employ the IQR method to treat outliers and compare the results with a model that does not undergo outlier treatment. The process involves identifying outliers using the IQR technique and then replacing extreme values with appropriate percentiles to mitigate their impact on the analysis.

Refer below table which has outliers removed

```
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                     0
CPM                     0
CPC                     0
dtype: int64
```



Perform z-score scaling and discuss how does it affects the performance of the algorithm?

	0	1	2	3	4	5	6 \
0	-0.364496	-0.432797	-0.352218	-0.512407	-0.515248	-0.510918	-0.615311
1	-0.364496	-0.432797	-0.352218	-0.512413	-0.515264	-0.510933	-0.615311
2	-0.364496	-0.432797	-0.352218	-0.512213	-0.515235	-0.510905	-0.615311
3	-0.364496	-0.432797	-0.352218	-0.512276	-0.515179	-0.510847	-0.615311
4	0.364496	0.432797	0.352218	0.512531	0.515281	0.510951	0.615311

	7	8	9	10	11	12
0	-0.665372	0.465447	-0.619693	-0.756200	-0.929911	-0.827068
1	-0.665372	0.465447	-0.619693	-0.750744	-0.929911	-0.827068
2	-0.665372	0.465447	-0.619693	-0.760292	-0.929911	-0.827068
3	-0.665372	0.465447	-0.619693	-0.771205	-0.929911	-0.827068
4	-0.665372	0.465447	-0.619693	-0.742559	-0.929911	-0.827068

Scaling the data using the Z-score scaling method is essential to ensure uniformity and comparability across different attributes present in the dataset. With varying measurement units and scales for each attribute, analyzing the complete dataset can become challenging and prone to errors, especially when using distance-based algorithms like clustering or Principal Component Analysis (PCA).

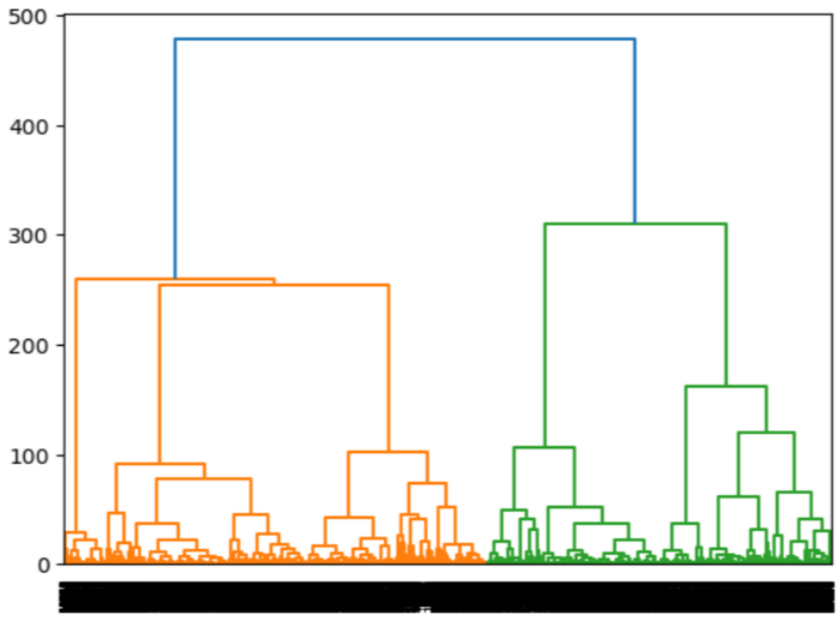
Z-score scaling standardizes the data to a common scale, thereby eliminating the discrepancies arising from different measurement units. This standardization facilitates easier comparison and interpretation of the data. Moreover, scaling enhances the efficiency of algorithms by normalizing the data to a specific range, which accelerates computations during analysis.

In summary, scaling the data using Z-score scaling is crucial for improving the accuracy and efficiency of analyses involving distance-based algorithms like clustering and PCA, by ensuring uniformity and comparability across different attributes.

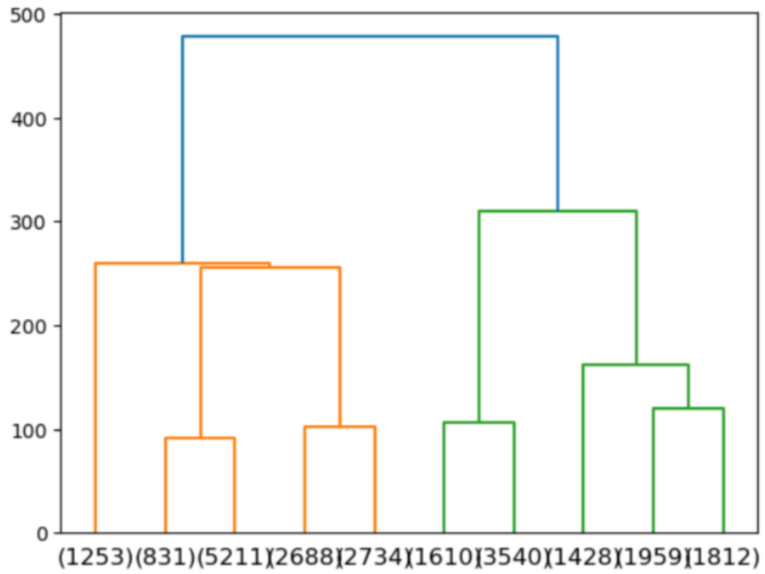
Problem 1 - Hierarchical Clustering

Hierarchical Clustering uses an approach of creating a tree called a 'Dendrogram'.

After performing Hierarchical clustering, we get a figure as shown below.



Upon analysing the provided figure, our objective is to determine the minimum number of clusters necessary to encapsulate a significant portion of the available data.



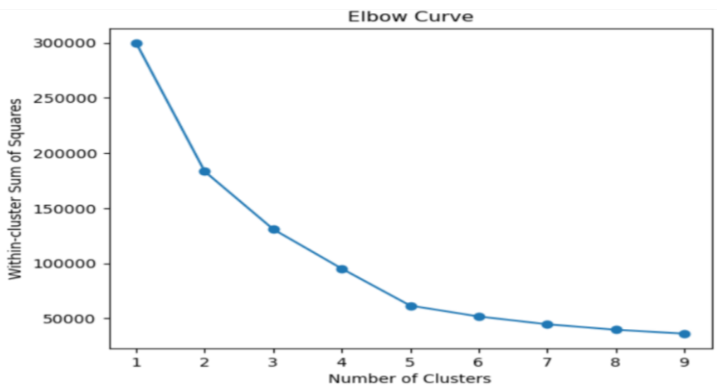
Performing hierarchical clustering involves constructing a dendrogram using the WARD method and Euclidean distance. Utilizing SciPy's cluster hierarchy function.

In a dendrogram, each branch is referred to as a clade, and the terminal end of each clade is known as a leaf. The arrangement of these clades provides insights into the similarity between the leaves. The height of the branching points signifies the degree of similarity or dissimilarity between clusters, with greater heights indicating greater differences.

Based on the reference provided, we observe that the longest branch, denoted in blue, suggests a segmentation into only 2 clusters, which may not be optimal for business purposes. Conversely, segmentation at the tallest red branches, demarcated by the yellow horizontal line, identifies 5 clusters. Alternatively, 3 clusters could also be considered, as indicated by the yellow horizontal line. However, for this dataset, we opt for 5 clusters based on the dendrogram analysis.

Problem 1 - K-means Clustering

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



Based on the Elbow plot depicted above, we can discern the optimal number of clusters. The plot indicates that the optimal number of clusters is 5, as there is minimal difference between the Within-Cluster Sum of Squares (WSS) values when the number of clusters is 5 compared to when it is 6, while the difference between 4 and 5 clusters is more pronounced.

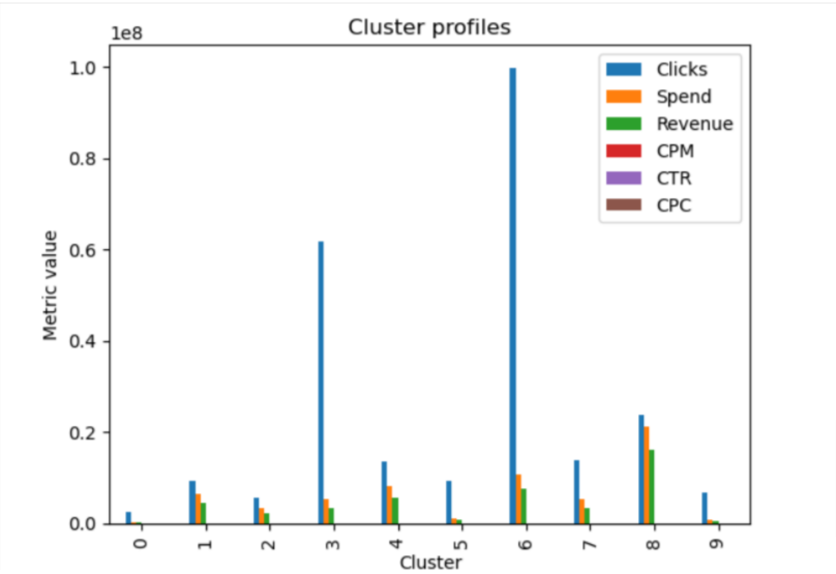
Furthermore, the drop in the WSS value is minimal beyond 5 clusters, reinforcing the conclusion that 5 clusters is the optimal choice.

Print silhouette scores for up to 10 clusters and identify optimum number of clusters.

```
For n_clusters=2, the silhouette score is 0.38572769619101077
For n_clusters=3, the silhouette score is 0.3825486036570082
For n_clusters=4, the silhouette score is 0.44534519247649795
For n_clusters=5, the silhouette score is 0.5240888532385488
For n_clusters=6, the silhouette score is 0.5221533662938636
For n_clusters=7, the silhouette score is 0.5165635029478517
For n_clusters=8, the silhouette score is 0.4797524035378018
For n_clusters=9, the silhouette score is 0.431966512420492
For n_clusters=10, the silhouette score is 0.4363637504360103
```

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.]



When grouping the data based on clusters and examining the mean values of certain fields, we can infer the following about the different clusters formed:

- **Cluster 0:** These are small-size budgeted ads with average revenue and lower user clicks. They tend to have low CPM and lower CTR ratios.

- **Cluster 1:** This cluster consists of the lowest budgeted ads with the lowest revenue and user clicks, similar to Cluster 0. However, they have a higher CPM ratio.
- **Cluster 2:** These are large-size budgeted ads with the highest revenue and highest user clicks. They tend to have low CTR and CPM ratios.
- **Cluster 3:** These ads are small-size budgeted with lower revenue but higher user clicks compared to other clusters. They also exhibit a higher CPM ratio.
- **Cluster 4:** Medium-size budgeted ads with average revenue, highest user clicks, and high CPM and CTR ratios.

Conclude the project by providing summary of your learnings.

1. **Cluster 1**: The CTR ratio is notably high for Cluster 1, indicating good user engagement with the ads. However, despite the high CPM, the revenue generated is comparatively lower. Additionally, the CPC for Cluster 1 is also lower. It's noteworthy that the spend on desktop ads within Cluster 1 needs to be increased as the CTR and CPC are nearly the same as mobile, yet the revenue is higher for desktop.
2. **Cluster 2**: This cluster exhibits the highest revenue among all clusters. Desktop ads contribute a larger share to the revenue compared to mobile ads within Cluster 2. Furthermore, the CPC is higher for mobile ads in Cluster 2 compared to desktop ads. It's worth mentioning that Cluster 2 has the highest CPC with the lowest CTR, suggesting fewer user clicks relative to the ad spend. Despite this, they remain the most profitable ads, indicating the continuation of the strategy with minor adjustments, particularly for mobile devices.
3. **Cluster 0**: In this cluster, the CTR is remarkably low, while the CPC is considerably high. Despite these metrics, the revenues remain average. This suggests a need to revisit the strategy for Cluster 0 to reduce the CPC and increase the CTR for better performance.
4. **Cluster 3**: With the highest CTR and the lowest CPC, Cluster 3 seems promising. However, the revenue generated is less compared to the spend. This indicates a need for optimization and cost-cutting measures in Cluster 3 to improve profitability.
5. **Cluster 4**: Ranking as the second most profitable cluster after Cluster 2, Cluster 4 has a remarkably low CPC compared to Cluster 2. This suggests a potential opportunity to allocate more budget to Cluster 4, given its profitability and lower CPC compared to Cluster 2.

By analysing each cluster's performance metrics and revenue generation, strategic insights can be derived to optimize ad campaigns and allocate resources effectively for maximum return on investment.

PCA

PCA:

PCA FH (FT): Primary census abstract for female headed households excluding institutional households (India & States/UTs - District Level), Scheduled tribes - 2011 PCA for Female Headed Household Excluding Institutional Household. The Indian Census has the reputation of being one of the best in the world. The first Census in India was conducted in the year 1872. This was conducted at different points of time in different parts of the country. In 1881 a Census was taken for the entire country simultaneously. Since then, Census has been conducted every ten years, without a break. Thus, the Census of India 2011 was the fifteenth in this unbroken series since 1872, the seventh after independence and the second census of the third millennium and twenty first century. The census has been uninterruptedly continued despite of several adversities like wars, epidemics, natural calamities, political unrest, etc. The Census of India is conducted under the provisions of the Census Act 1948 and the Census Rules, 1990. The Primary Census Abstract which is important publication of 2011 Census gives basic information on Area, Total Number of Households, Total Population, Scheduled Castes, Scheduled Tribes Population, Population in the age group 0-6, Literates, Main Workers and Marginal Workers classified by the four broad industrial categories, namely, (i) Cultivators, (ii) Agricultural Laborers, (iii) Household Industry Workers, and (iv) Other Workers and also Non-Workers. The characteristics of the Total Population include Scheduled Castes, Scheduled Tribes, Institutional and Houseless Population and are presented by sex and rural-urban residence. Census 2011 covered 35 States/Union Territories, 640 districts, 5,924 sub-districts, 7,935 Towns and 6,40,867 Villages.

Problem 2 - Define the problem and perform Exploratory Data Analysis

- Problem Definition - Check shape, Data types, statistical summary - Perform an EDA on the data to extract useful insights

Head of Data Frame

	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180	
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123	
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	...	114	188	44	
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	...	194	247	61	
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	...	874	1928	465	

5 rows × 61 columns

Tail of Data Frame

	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_
635	34	636	Puducherry	Mahe	3333	8154	11781	1146	1203	21	...	32	47	0	
636	34	637	Puducherry	Karaikal	10612	12346	21691	1544	1533	2234	...	155	337	3	
637	35	638	Andaman & Nicobar Island	Nicobars	1275	1549	2630	227	225	0	...	104	134	9	
638	35	639	Andaman & Nicobar Island	North & Middle Andaman	3762	5200	8012	723	664	0	...	136	172	24	
639	35	640	Andaman & Nicobar Island	South Andaman	7975	11977	18049	1470	1358	0	...	173	122	6	

5 rows × 61 columns

Data Types

```
<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
```

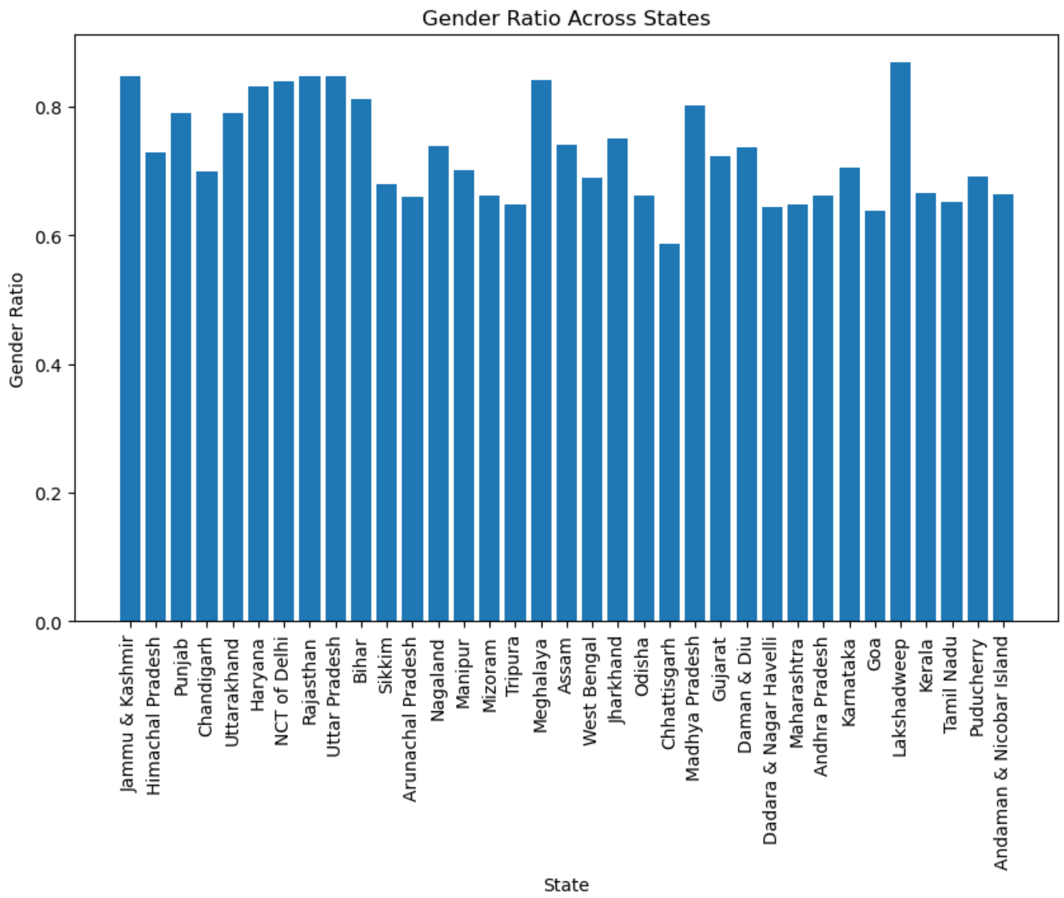
The data set has 240 rows and 61 columns.

There are no null values or duplicate values in the data set.

There are very few records for Lakshadweep, Chandigarh and Dadara & Nagar Haveli There are 2 datatypes present in the data(int64 =59 and object =2.)

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

Which state has highest gender ratio and which has the lowest?



Based on the analysis of the provided dataset and the accompanying graph depicting the gender ratio across states, several observations can be made.

Andhra Pradesh emerges with the highest gender ratio among states, standing at 1.89, indicating a greater proportion of males to females within the population. Conversely, Lakshadweep exhibits the lowest gender ratio among states, with a ratio of 1.15, suggesting a relatively higher proportion of females to males.

Furthermore, to delve deeper into the gender distribution, a dataset was constructed to examine the gender ratio at the district level. By sorting this dataset, additional insights were gleaned. Notably, Badgam, located in Jammu and Kashmir, emerges with the lowest gender ratio, excluding Lakshadweep. Conversely, Krishna district in

Andhra Pradesh showcases the highest gender ratio, indicating a different demographic profile compared to Badgam.

These findings shed light on the gender distribution patterns across different geographical regions, highlighting areas with notable variations in gender ratios. Such insights are invaluable for understanding demographic trends and informing targeted interventions aimed at addressing gender disparities.

Problem 2 - Data Preprocessing

Check for and treat (if needed) missing values - Check for and treat (if needed) data irregularities - Scale the Data using the z-score method - Visualize the data before and after scaling and comment on the impact on outliers.

Null Values

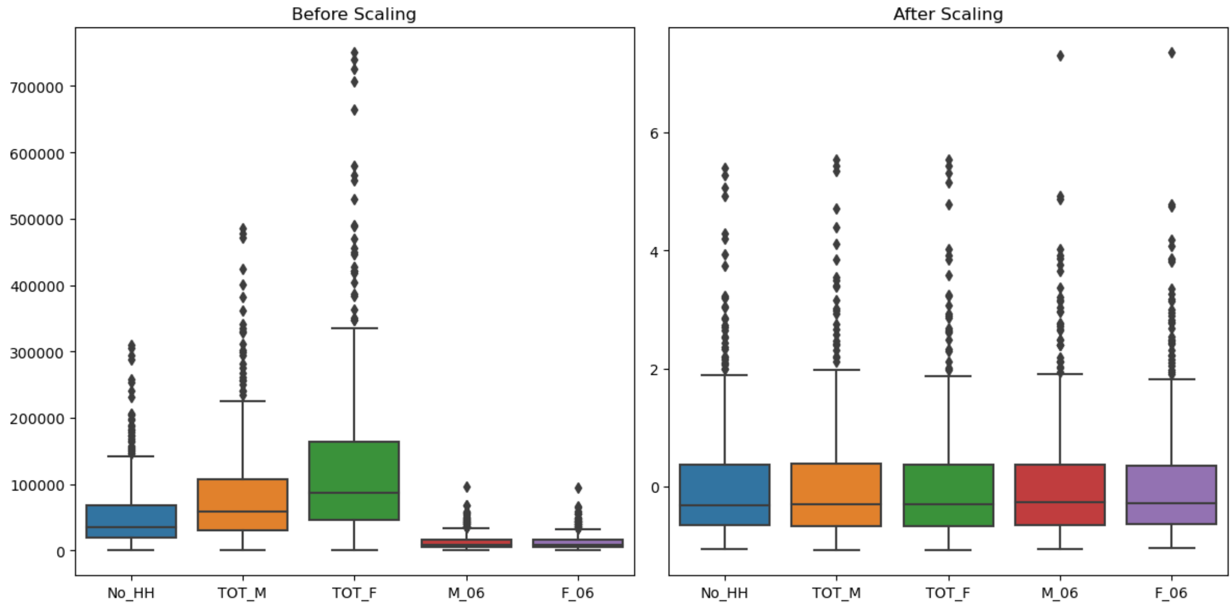
```
Missing values:
 State Code      0
Dist.Code       0
State           0
Area Name       0
No_HH           0
               ..
MARG_HH_0_3_F   0
MARG_OT_0_3_M   0
MARG_OT_0_3_F   0
NON_WORK_M      0
NON_WORK_F      0
Length: 61, dtype: int64
```

Scale the data using the Zscore method

	No_HH	TOT_M	TOT_F	M_06	F_06
0	-0.904738	-0.771236	-0.815563	-0.561012	-0.507738
1	-0.935695	-0.823100	-0.874534	-0.681096	-0.725367
2	-0.972412	-1.000919	-0.981466	-0.976956	-0.965262
3	-1.037530	-1.052224	-1.041001	-1.022118	-0.995393
4	-0.822676	-0.809381	-0.813933	-0.622359	-0.649908
...	...	...	...	...	...
635	-0.995677	-0.978990	-0.974268	-0.971387	-0.948916
636	-0.844340	-0.921822	-0.886965	-0.936754	-0.919757
637	-1.038465	-1.069066	-1.054885	-1.051356	-1.035331
638	-0.986758	-1.019276	-1.007472	-1.008195	-0.996541
639	-0.899166	-0.926854	-0.919050	-0.943193	-0.935220

640 rows × 5 columns

Visualize



The scaling has an impact on the outliers. Due to scaling the range of values for the feature is reduced drastically while keeping the outliers in the dataset.

Problem 2 - PCA

- Create the covariance matrix - Get eigen values and eigen vectors - Identify the optimum number of PCs - Show Scree plot - Compare PCs with Actual Columns and identify which is explaining most variance - Write inferences about all the PCs in terms of actual variables - Write linear equation for first PC Note: For the scope of this project, take at least 90% explained variance.

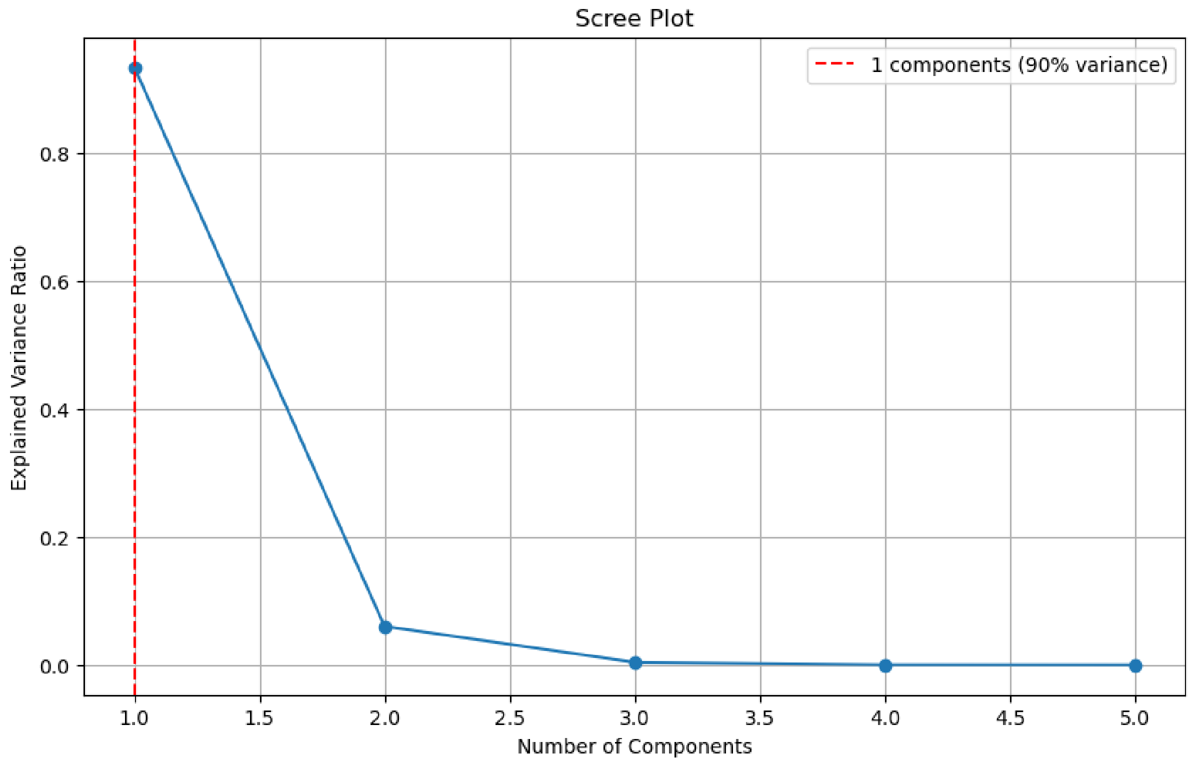
Eigen Values

```
Eigenvalues:
[6.95495130e+00 2.60273066e+00 1.61839101e+00 6.93356489e-01
 4.32367815e-01 3.14770773e-01 1.60098486e-01 1.13701628e-01
 6.98689613e-02 2.94153034e-02 1.05869964e-02 1.94454700e-04
 1.29742820e-04]
```

Eigen Vectors

```
Eigenvectors:
[[ 1.18809362e-01 -2.31809100e-01 -1.14317007e-01  3.57325084e-01
   3.60815462e-01  3.59724633e-01  1.13798016e-01  3.16716970e-01
  -2.79473590e-01  3.15381633e-01 -2.80450757e-01 -2.66069848e-01
   3.02544530e-01]
 [-1.06216001e-01  3.28710634e-01  2.18224473e-01 -1.55363769e-02
   2.96038999e-03 -6.05222833e-03  4.71780611e-01  3.27743391e-01
  -3.61187974e-01  3.31901073e-01  3.14065521e-01  3.49184688e-01
  -2.03921405e-01]
 [ 7.00418360e-01 -3.55305351e-01  5.18121342e-01 -7.09716301e-02
  -8.13653182e-02 -8.33065035e-02  2.74906529e-01 -1.28065095e-02
   1.16961648e-01 -1.94174397e-02 -1.28094435e-03 -3.93706350e-02
  -6.91294781e-02]
 [-4.20850029e-02  3.01578365e-01  6.30133860e-01  2.60696934e-01
   2.87852922e-01  2.99833815e-01 -4.22439570e-01 -1.19998226e-01
   3.49067501e-02 -1.17193947e-01  1.68601546e-01  9.42587687e-02
   1.56106645e-01]
 [ 3.39392922e-01 -2.79635051e-01 -3.74282405e-01  8.35797771e-02
   5.54866113e-02  5.66745388e-02 -2.93545538e-01  1.01335228e-02
  -1.14566538e-01  1.09630243e-02  5.05133198e-01  5.15040792e-01
   1.82507812e-01]
 [ 1.14112592e-02  1.49824042e-01  1.45785411e-01 -3.62088807e-01
  -2.45371671e-01 -2.39847422e-01 -6.61872639e-02  1.08484706e-01
  -1.32871863e-01  1.09005178e-01 -1.83168840e-01  1.21765680e-01
   7.83110692e-01]
 [-1.60761994e-01 -4.69111492e-03 -7.02742286e-02  1.30278350e-01
   1.61824488e-01  1.70126199e-01  4.93069871e-01 -2.91773354e-02
   6.82227502e-01 -6.57539208e-02  2.66360991e-02  2.93601712e-01
   3.10570389e-01]
 [ 2.82556843e-02  1.22847559e-01 -7.38621664e-02 -4.86898147e-02
  -1.97363201e-02 -1.73770082e-02  1.49409390e-01 -1.46151293e-02
   5.10733238e-02 -1.95909267e-02  6.82615392e-01 -6.51994967e-01
```

Identify the optimum number of PCs - Show Scree plot



The optimal number of Principal Components can be determined by examining the cumulative sum of explained variances. By assessing when the cumulative explained variance reaches 0.85, we can identify the optimal number of Principal Components.

This approach ensures that at least 85% of the variance in the data is explained by the Principal Components, providing a balance between dimensionality reduction and information retention.

Compare PCs with Actual Columns and identify which is explaining most variance - Write inferences about all the PCs in terms of actual variables

```
Principal Component 1:
  No_HH: 0.42857823607763434
  TOT_M: 0.4593623855765384
  TOT_F: 0.45632420259992235
  M_06: 0.44584832674987984
  F_06: 0.4453025779555879
```

In the majority of cases, PC1 accounts for the most variance compared to the actual columns.

- PC5 explains the highest variance for the following variables: M_ST, F_ST, MAIN_CL_M, MAIN_CL_F, MARG_OT_F, MARGWORK_0_3_F, NON_WORK_M, and NON_WORK_F.
- PC4 explains the highest variance for the following variables: MARG_CL_F, MARG_AL_M, MARG_AL_3_6_F, MARG_HH_3_6_M, MARG_HH_0_3_M, MARG_HH_0_3_F, MARG_OT_0_3_M, and MARG_OT_3_F.
- PC3 explains the highest variance for the following variables: TOT_WORK_F, MAINWORK_F, MAIN_AL_M, MAIN_AL_F, MARG_AL_F, and MARG_HH_3_6_F.
- PC2 explains the highest variance for the following variables: MAIN_OT_M, MAIN_OT_F, MARG_CL_M, MARG_AL_3_6_M, MARG_AL_0_3_M, and MARG_AL_0_3_F.

Write linear equation for first PC

```
Linear equation for the first PC:
0.43 * No_HH + 0.46 * TOT_M + 0.46 * TOT_F + 0.45 * M_06 + 0.45 * F_06
```