INDEX

LIST OF FIGURES

Problem 1 : 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.

The following three features are commonly used in digital marketing:

CPM = (Total Campaign Spend / Number of Impressions) * 1,000. Note that the Total Campaign Spend refers to the 'Spend' Column in the dataset and the Number of Impressions refers to the 'Impressions' Column in the dataset.

CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column in the dataset and the Number of Clicks refers to the 'Clicks' Column in the dataset.

CTR = Total Measured Clicks / Total Measured Ad Impressions x 100. Note that the Total Measured Clicks refers to the 'Clicks' Column in the dataset and the Total Measured Ad Impressions refers to the 'Impressions' Column in the dataset.

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.

Import some libraries like Numpy, Pandas, Seaborn, Matplotlib, etc. After that, load our data set, austo .csv, and use the head() function to view the Top 5 data and the tail() function to view the bottom 5 dat a. Using the shape function, we can determine that there are 23066 rows and 19 columns. Find out the characteristics of the columns using the info() method. The datatypes for the float64(6), int64(7), and o bject(6) columns are present. There are some null values, but there aren't any duplicate values.

- **head()** it given by default top five data

	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

Ad - gth	Ad-Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0	0.35	0.0	0.0031	0.0	0.0
300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.0	0.35	0.0	0.0035	0.0	0.0
300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0	0.35	0.0	0.0028	0.0	0.0
300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.0	0.35	0.0	0.0020	0.0	0.0
300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.0	0.35	0.0	0.0041	0.0	0.0

Fig 1 : Top five data of dataset Digital_Ads

- **tail()** it given by default bottom five data

	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	

Ad - ngth	Ad-Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
720	300	216000	Inter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
720	300	216000	Inter224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.0260	NaN	NaN	NaN
720	300	216000	Inter218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	NaN	NaN	NaN
120	600	72000	inter230	Video	Mobile	Video	7	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
720	300	216000	Inter221	App	Mobile	Video	2	2	2	1	0.09	0.35	0.0585	NaN	NaN	NaN

Fig: 2. Bottom five data of dataset Digital_Ads

- **shape** it tells numbers of rows and columns in given dataset.

```
(23066, 19)
```

Fig 3 : 23066 rows & 19 columns

- **info()** it tells a concise summary of a DataFrame

```
<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
```

Fig 4 : Information about the Digital_Ads structure and content.

- CTR, CPM and CPC have Null Values.
- No Duplicates 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
```

Fig 5 : Null values of the Digital_Ads.

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

Formula Based Technique to treat missing values

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

Note that the Total Campaign Spend refers to the 'Spend' Column in the dataset and the Number of Impressions refers to the 'Impressions' Column in the dataset.

CPC = Total Cost (spend) / Number of Clicks.

Note that the Total Cost (spend) refers to the 'Spend' Column in the dataset and the Number of Clicks refers to the 'Clicks' Column in the dataset.

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

Note that the Total Measured Clicks refers to the 'Clicks' Column in the dataset and the Total Measured Ad Impressions refers to the 'Impressions' Column in the dataset.

Ad - gth	Ad-Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0	0.35	0.0	0.0031	0.0	0.0
300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.0	0.35	0.0	0.0035	0.0	0.0
300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0	0.35	0.0	0.0028	0.0	0.0
300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.0	0.35	0.0	0.0020	0.0	0.0
300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.0	0.35	0.0	0.0041	0.0	0.0

Ad - gth	Ad-Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
720	300	216000	Inter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	0.010	70.0	0.07
720	300	216000	Inter224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.0260	0.005	20.0	0.04
720	300	216000	Inter218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	0.010	50.0	0.05
120	600	72000	inter230	Video	Mobile	Video	7	1	1	1	0.07	0.35	0.0455	0.010	70.0	0.07
720	300	216000	Inter221	App	Mobile	Video	2	2	2	1	0.09	0.35	0.0585	0.005	45.0	0.09

Fig 6 : 1. Top five data of dataset Digital_Ads after Null Value treatment

```
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
```

Fig 7 : Digital_Ads without Null Value

- **Describe()** it tells summary of the central tendency, dispersion, and shape of the distribution of the data.

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	385.163	233.651	120.00	120.000	300.000	720.000	728.00
Ad- Width	23066.0	337.896	203.093	70.00	250.000	300.000	600.000	600.00
Ad Size	23066.0	96674.468	61538.330	33600.00	72000.000	72000.000	84000.000	216000.00
Available_Impressions	23066.0	2432043.666	4742887.765	1.00	33672.250	483771.000	2527711.750	27592861.00
Matched_Queries	23066.0	1295099.143	2512969.861	1.00	18282.500	258087.500	1180700.000	14702025.00
Impressions	23066.0	1241519.519	2429399.961	1.00	7990.500	225290.000	1112428.500	14194774.00
Clicks	23066.0	10678.519	17353.409	1.00	710.000	4425.000	12793.750	143049.00
Spend	23066.0	2706.626	4067.927	0.00	85.180	1425.125	3121.400	26931.87
Fee	23066.0	0.335	0.032	0.21	0.330	0.350	0.350	0.35
Revenue	23066.0	1924.252	3105.238	0.00	55.365	926.335	2091.338	21276.18
CTR	23066.0	0.059	0.073	0.00	0.002	0.005	0.122	1.00
CPM	23066.0	8.397	9.057	0.00	1.750	8.371	13.040	715.00
CPC	23066.0	0.337	0.341	0.00	0.090	0.140	0.550	7.26

Fig 9 : Descriptive statistics of Digital_Ads

3. 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).

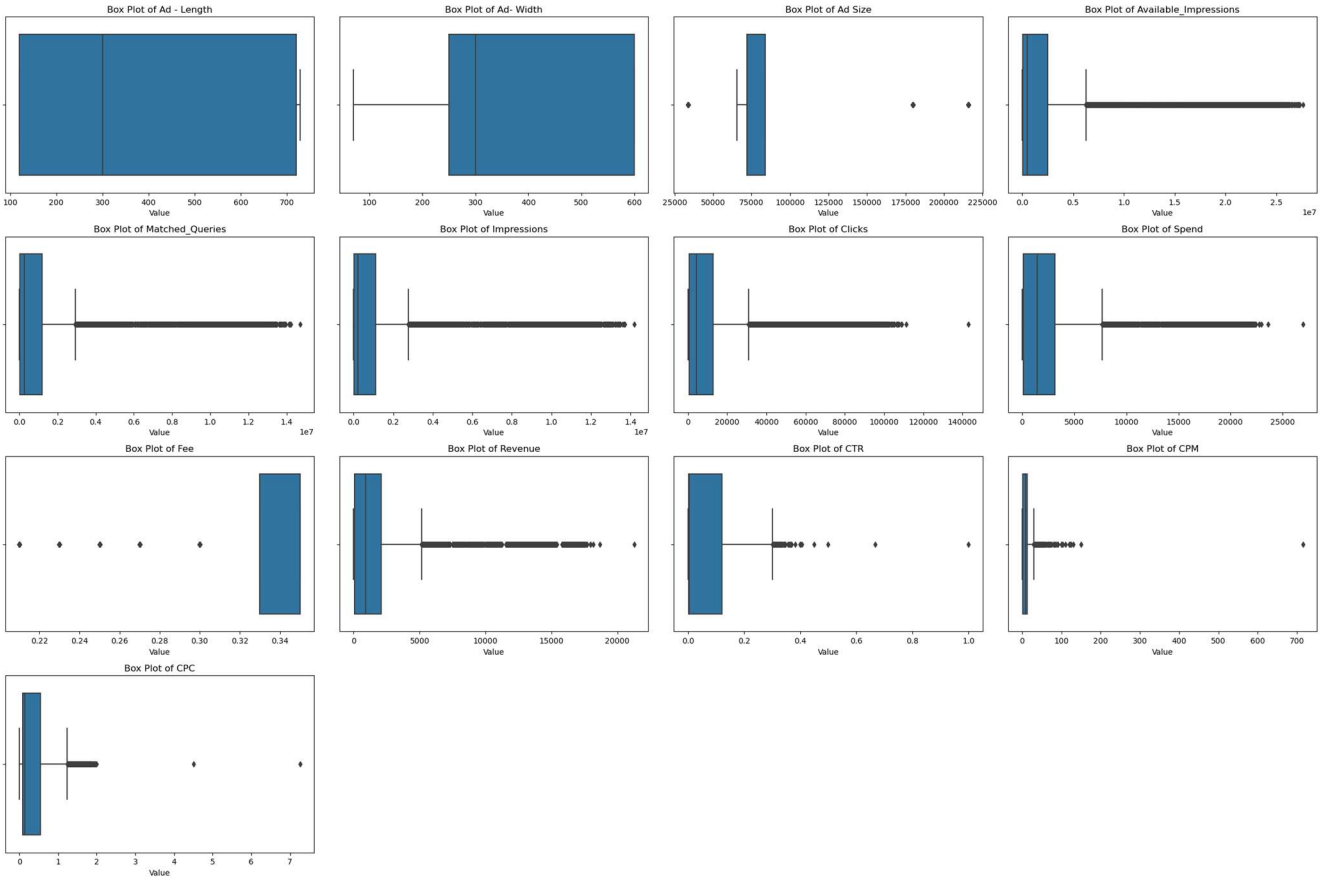


Fig 10 : Before removing outliers of Digital_Ads

- The Digital_Ads dataset has outliers; they had a significant impact on clustering. Hence, it is important to identify and remove outliers before applying the K-means clustering algorithm.

Fig 11 : After removing outliers of Digital_Ads

4. Perform z-score scaling and discuss how it affects the speed of the algorithm.

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	300.0	250.0	75000.0	33672.25	18282.5	7990.5	710.0	85.18	0.35	55.365375	0.0031	1.75	0.09
1	300.0	250.0	75000.0	33672.25	18282.5	7990.5	710.0	85.18	0.35	55.365375	0.0035	1.75	0.09
2	300.0	250.0	75000.0	33672.25	18282.5	7990.5	710.0	85.18	0.35	55.365375	0.0028	1.75	0.09
3	300.0	250.0	75000.0	33672.25	18282.5	7990.5	710.0	85.18	0.35	55.365375	0.0021	1.75	0.09
4	300.0	250.0	75000.0	33672.25	18282.5	7990.5	710.0	85.18	0.35	55.365375	0.0041	1.75	0.09

Fig 12 : Top five data of dataset of scaled_data after perform z-score scaling

Insights: Z-score scaling (standardization) is a common data preprocessing technique used to scale features in a dataset. It transforms each feature to have a mean of 0 and a standard deviation of 1, making the data more suitable for certain machine learning algorithms.

Overall, the performance and effectiveness of the K-Means clustering technique may be improved by z-score scaling. It promotes faster convergence and ensures that all characteristics are handled equally

during the clustering process. Additionally, the standardized feature scales may enhance the stability and quality of the generated clusters.

5. Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.

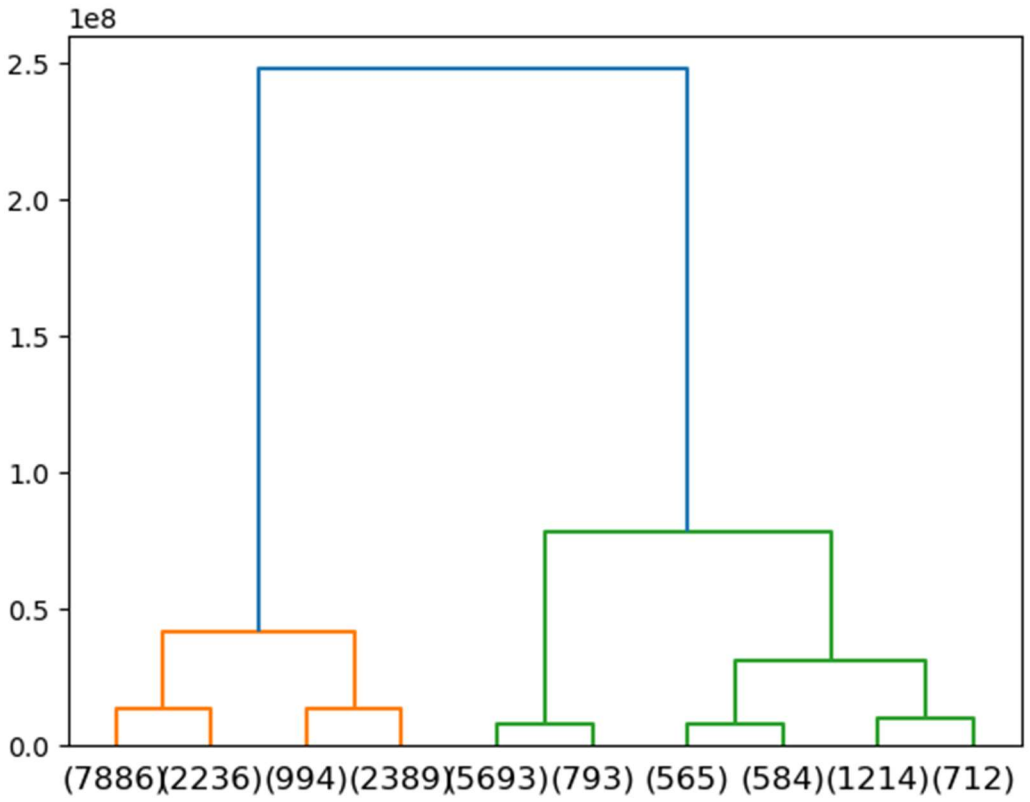


Fig 13 : Dendrogram

The dendrogram resulting from hierarchical clustering with Ward linkage and Euclidean distance represents the process of forming clusters by progressively merging data points or clusters. Each data point starts as an individual cluster, and during the clustering process, similar points or clusters are merged together.

The vertical axis of the dendrogram represents the distance (similarity) between data points or clusters. The longer the branches on the dendrogram, the greater the distance between the clusters being merged. Conversely, shorter branches indicate that the clusters being merged are more similar or closer to each other.

The dendrogram's horizontal axis represents the individual data points or clusters. Each data point is initially shown at the bottom of the dendrogram as its own cluster. As the clustering algorithm progresses, similar data points or clusters are joined together, moving upwards in the dendrogram.

Clusters are merged based on the Ward linkage method, which aims to minimize the variance within the resulting clusters. The Euclidean distance is used as the distance metric to determine the similarity between data points.

```
array([[0.00000000e+00, 5.60000000e+01, 0.00000000e+00, 2.00000000e+00],
       [1.00000000e+00, 2.90000000e+01, 0.00000000e+00, 2.00000000e+00],
       [2.00000000e+00, 5.40000000e+01, 0.00000000e+00, 2.00000000e+00],
       ...,
       [4.61250000e+04, 4.61270000e+04, 2.73394790e+02, 1.19800000e+04],
       [4.61230000e+04, 4.61260000e+04, 2.97828390e+02, 1.10860000e+04],
       [4.61280000e+04, 4.61290000e+04, 4.94784625e+02, 2.30660000e+04]])
```

Fig 14 : linkage_matrix

A linkage matrix, also referred to as a linkage array, is a fundamental data structure utilized in hierarchical clustering algorithms. It serves as a record of the clustering process, often represented as a matrix with each row corresponding to a step in the algorithm. Every row in the linkage matrix contains information about the two clusters or data points that were merged at that specific step. Additionally, it includes the distance or similarity measure between these clusters or data points.

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

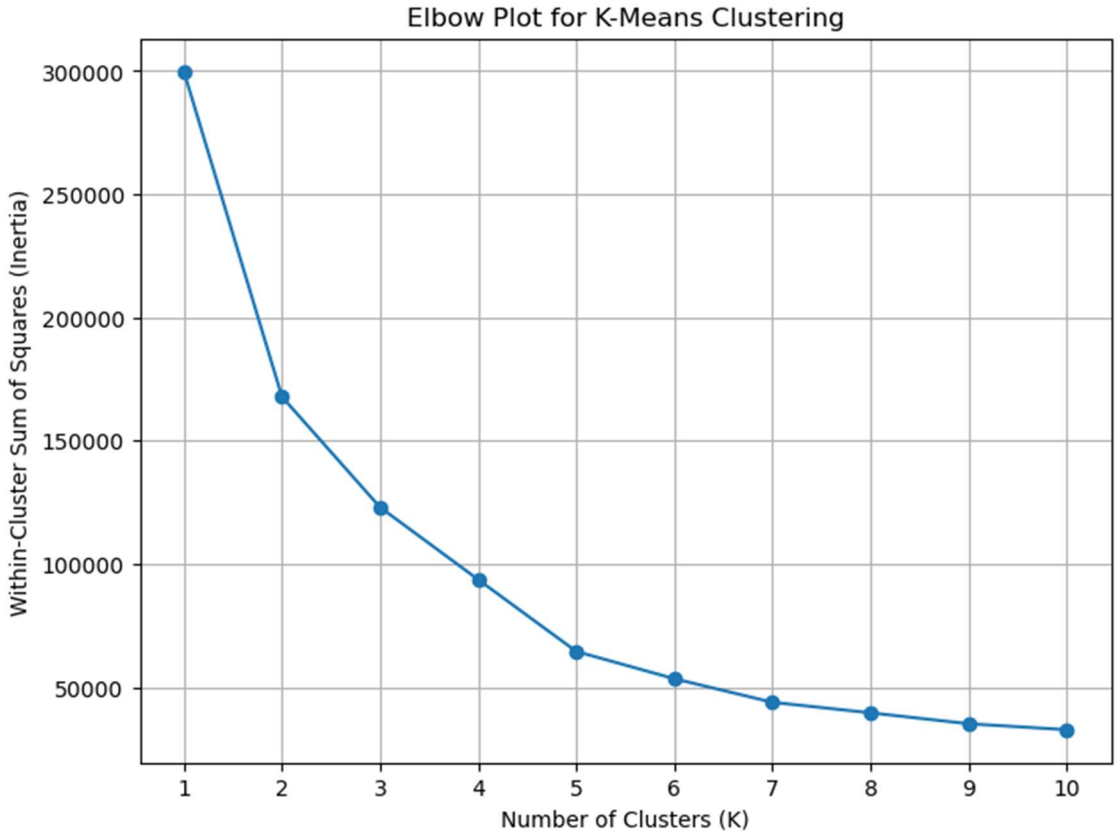


Fig 13 : Elbow Plot for K-Means Clustering

Insigths: The x-axis represents the number of clusters (K), and the y-axis represents the WCSS. The elbow point is identified as the "bend" in the plot, where the WCSS starts to level off. The optimum number of clusters can be chosen based on the location of this elbow point.

- Here K = 1 to 2 is a significant drop, and K = 2 to 3, k=3 to 4 and k=4 to 5 are also a significant drop, but after 5–10 drops, it becomes very graduated.
- So the 5 is a optimal number.

e	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	Clus_kmeans5
0	Inter222	Video	Desktop	Display	33672.25	18282.5	7990.5	710.0	85.18	0.35	55.365375	0.0031	1.75	0.09	1
0	Inter227	App	Mobile	Video	33672.25	18282.5	7990.5	710.0	85.18	0.35	55.365375	0.0035	1.75	0.09	1
0	Inter222	Video	Desktop	Display	33672.25	18282.5	7990.5	710.0	85.18	0.35	55.365375	0.0028	1.75	0.09	1
0	Inter228	Video	Mobile	Video	33672.25	18282.5	7990.5	710.0	85.18	0.35	55.365375	0.0021	1.75	0.09	1
0	Inter217	Web	Desktop	Video	33672.25	18282.5	7990.5	710.0	85.18	0.35	55.365375	0.0041	1.75	0.09	1

Fig 14 : Top five data of dataset Digital_Ads with "Clus_kmeans5"

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

- Silhouette_score is 0.5069189501741144

d Size	Ad Type	Platform	Device Type	Format	Available_Impressions	...	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	Clus_kmeans5	sil_width
5000.0	Inter222	Video	Desktop	Display	33672.25	...	7990.5	710.0	85.18	0.35	55.365375	0.0031	1.75	0.09	1	0.149549
5000.0	Inter227	App	Mobile	Video	33672.25	...	7990.5	710.0	85.18	0.35	55.365375	0.0035	1.75	0.09	1	0.150158
5000.0	Inter222	Video	Desktop	Display	33672.25	...	7990.5	710.0	85.18	0.35	55.365375	0.0028	1.75	0.09	1	0.149086
5000.0	Inter228	Video	Mobile	Video	33672.25	...	7990.5	710.0	85.18	0.35	55.365375	0.0021	1.75	0.09	1	0.147992
5000.0	Inter217	Web	Desktop	Video	33672.25	...	7990.5	710.0	85.18	0.35	55.365375	0.0041	1.75	0.09	1	0.151057

Fig 15 : Top five data of dataset Digital_Ads with sil_width

- Silhouette_samples is -0.07656855547495799

```
Number of clusters (K) = 2, Silhouette Score = 0.4103
Number of clusters (K) = 3, Silhouette Score = 0.4167
Number of clusters (K) = 4, Silhouette Score = 0.4491
Number of clusters (K) = 5, Silhouette Score = 0.5069
Number of clusters (K) = 6, Silhouette Score = 0.5037
Number of clusters (K) = 7, Silhouette Score = 0.5046
Number of clusters (K) = 8, Silhouette Score = 0.5153
Number of clusters (K) = 9, Silhouette Score = 0.5034
Number of clusters (K) = 10, Silhouette Score = 0.5004
```

Fig 16 : Silhouette scores for different values of K

- Optimal number of clusters is 5

8. 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].

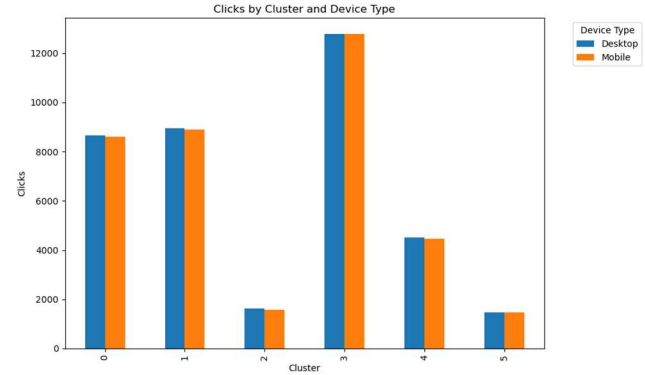


Fig 17 : Clicks by Cluster and Device Type

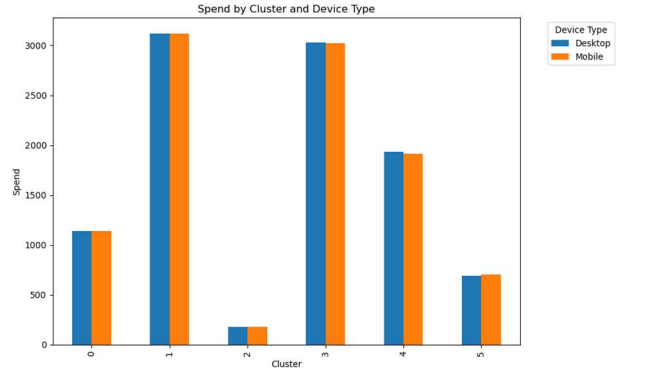


Fig 18 : Spend by Cluster and Device Type

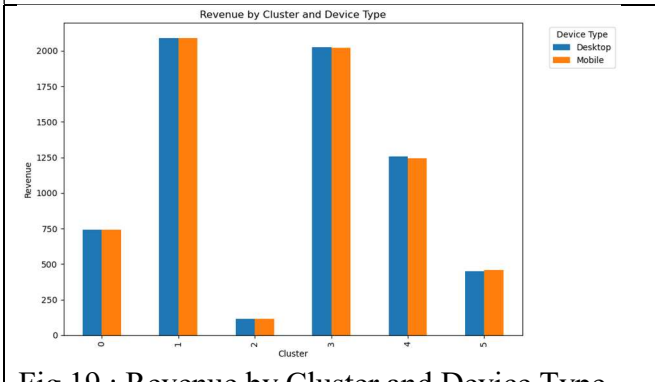


Fig 19 : Revenue by Cluster and Device Type

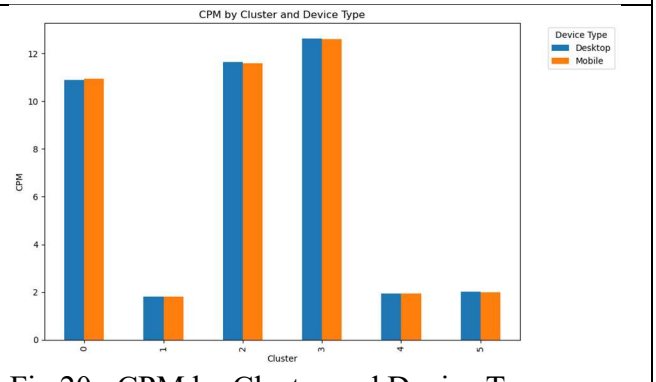


Fig 20 : CPM by Cluster and Device Type

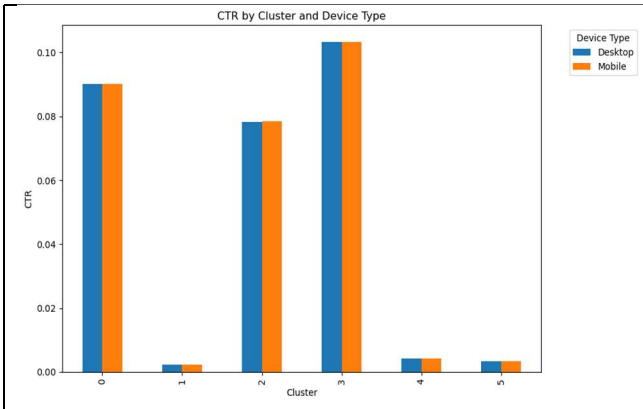


Fig 21 : CTR by Cluster and Device Type

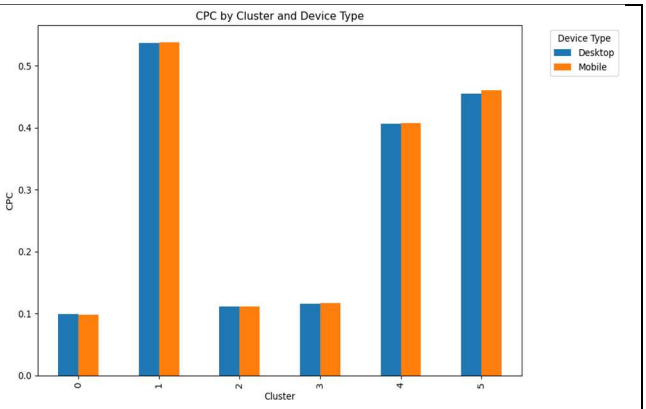


Fig 22 : CPC by Cluster and Device Type

Insights: The bar plots will help you understand how different metrics vary across clusters and Device Types, enabling you to profile the ads based on the clustering results and domain understanding.

- Clicks by Cluster and Device Type: Cluster 3 with Desktop and Mobile has more than 12000, and Cluster 5 with Desktop and Mobile has less than 2000.
- Spend by Cluster and Device Type: Cluster 1 with Desktop and Mobile has more than 3000 and Cluster 2 with Desktop and Mobile has less than 500.
- Revenue by Cluster and Device Type: Cluster 1 with Desktop and Mobile has more than 2000, and Cluster 2 with Desktop and Mobile has less than 250.
- CPM by Cluster and Device Type: Cluster 3 with Desktop and Mobile has more than 12, and Cluster 2 with Desktop and Mobile has less than 2.
- CPM by Cluster and Device Type: Cluster 3 with Desktop and Mobile has more than 0.10, and Cluster 2 with Desktop and Mobile has less than 0.02.
- CPC by Cluster and Device Type: Cluster 1 with Desktop and Mobile has more than 0.5, and Cluster 0 with Desktop and Mobile has approximately 0.1.

9. Clustering: Conclude the project by providing summary of your learnings.

- The dataset consists of 23066 rows and 19 columns. To handle missing values in the 'CPC', 'CTR', and 'CPM' variables, we used a user-defined function that applied appropriate formulas to impute the missing data.
- After handling missing values, we performed outlier detection and found that there are outliers present in some variables of the dataset.
- Next, we visualized the data using a dendrogram by computing linkage with Ward's method. The linkage function was applied on the relevant columns of the data to calculate the distances and

sequentially merge the clusters from 'n' to '1'. This allowed us to observe how the distances change during the clustering process.

- Using the fit-transform function, we stored the data frame in an array for further analysis. With this array, we proceeded to perform k-means clustering. However, before running the k-means algorithm, we needed to determine the optimal number of clusters to use as output.
- To find the optimal number of clusters, we created an elbow plot by plotting the within-cluster sum of squares (WSS) values for different values of 'k' (number of clusters). The elbow plot indicated that the WSS values significantly dropped when moving from k=1 to k=2, and again from k=2 to k=3 and k=3 to k=4. However, beyond k=4, the drop in WSS values reduced, suggesting that k=5 is the optimal number of clusters.
- Based on this analysis, we proceeded with k-means clustering using 5 clusters to segment the data and identify distinct patterns within the dataset."

Problem 2 : 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.

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

- **head()** it given by default top five data

	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	

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
749	180	237	680	252	32	46	258	214
715	123	229	186	148	76	178	140	160
188	44	89	3	34	0	4	67	61
247	61	128	13	50	4	10	116	59
1928	465	1043	205	302	24	105	180	478

Fig 23 : Top five data of dataset Digital_Ads

- **tail()** it given by default bottom five data

	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	

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
47	0	0	0	0	0	0	32	47
337	3	14	38	130	4	23	110	170
134	9	4	2	6	17	47	76	77
172	24	44	11	21	1	4	100	103
122	6	2	17	17	2	4	148	99

- Fig: 24. Bottom five data of dataset Digital_Ads

- **shape** it tells numbers of rows and columns in given dataset.

```
(640, 61)
```

Fig 25 : 640 rows & 61 columns

- **info()** it tells a concise summary of a DataFrame
- **Describe()** it tells summary of the central tendency, dispersion, and shape of the distribution of the data.

Insights: There are Rows 640 and columns 61. no duplicate values and no null values.

2. PCA: Perform detailed Exploratory analysis by creating certain questions like (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

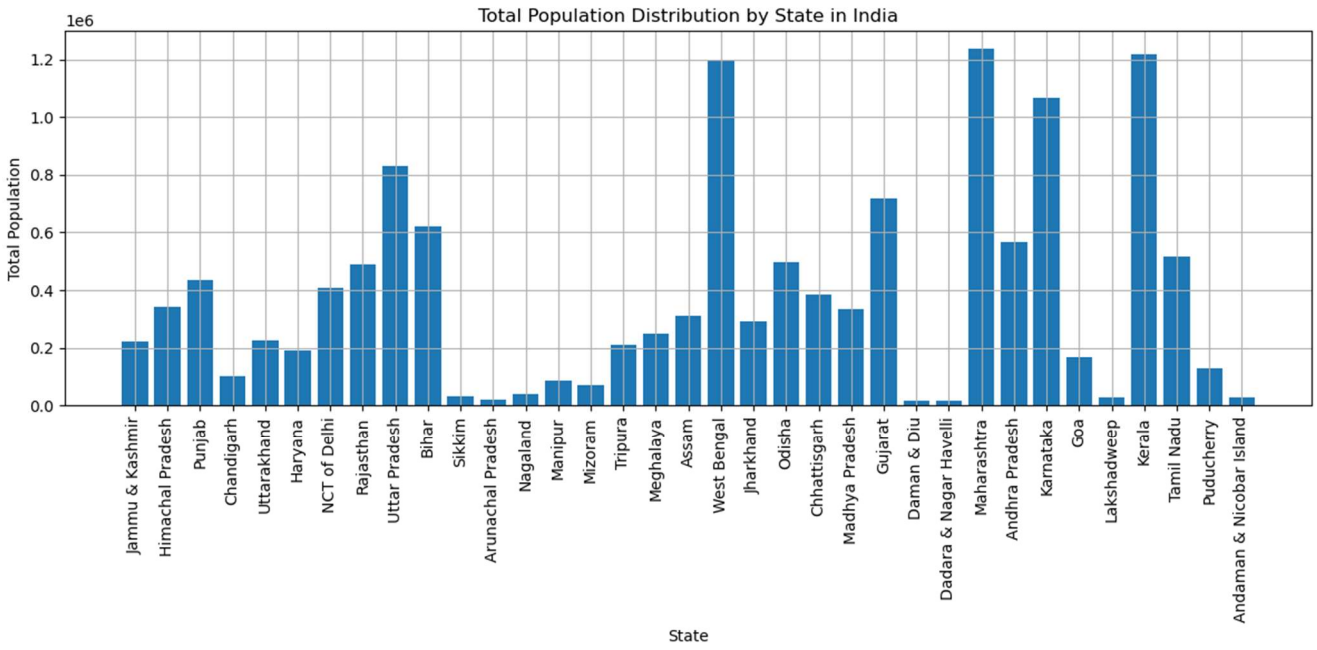


Fig 26 : Total Population Distribution by State in India

- State with the highest population: Maharashtra

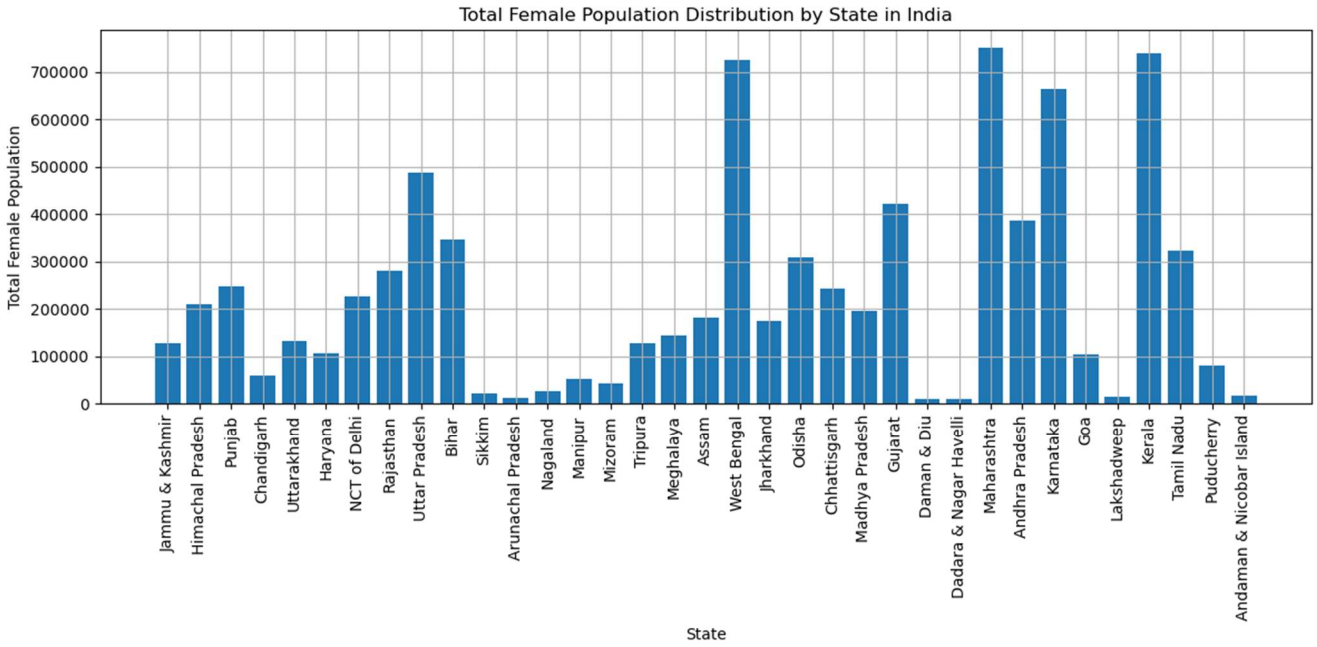


Fig 27 : Total Female Population Distribution by State in India

- State with the highest total female population: Maharashtra

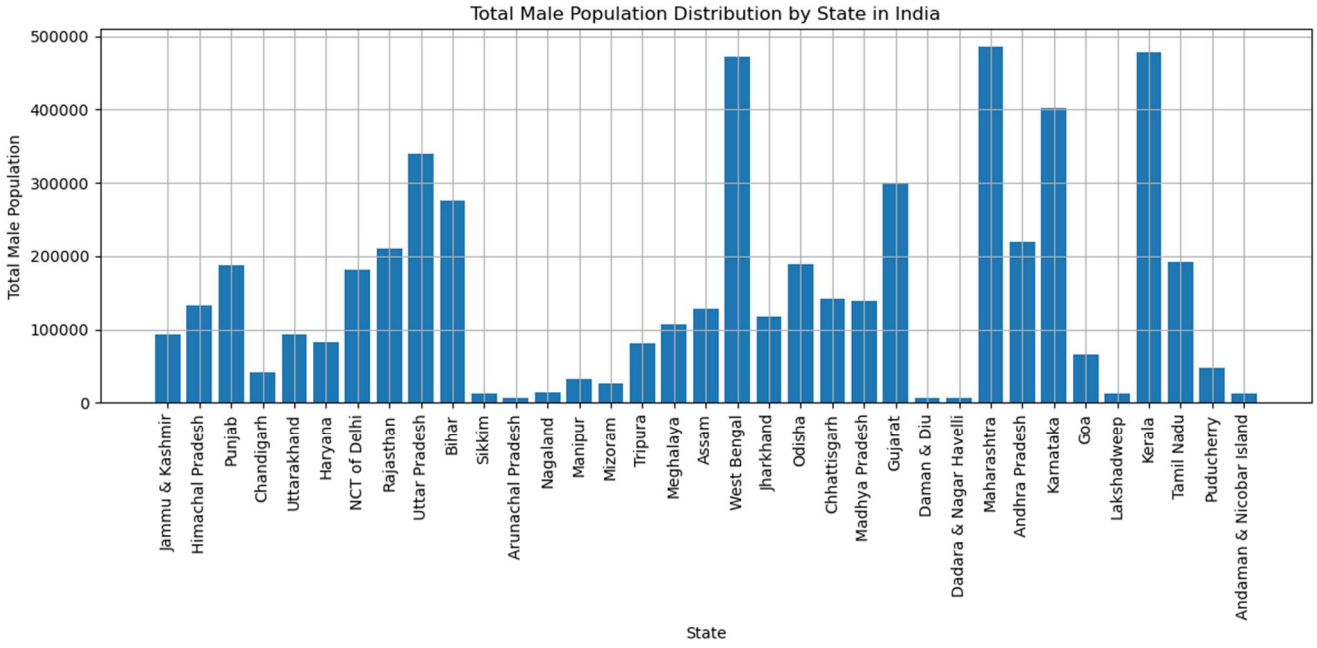


Fig 28 : Total Male Population Distribution by State in India

- State with the highest total male population: Maharashtra

For EDA we 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

So the 5 variables are: No_HH, TOT_M, TOT_F,TOT_WORK_M, and TOT_WORK_F

Now Plot Histograms and Box plots for each of the five variables individually.

Univariate Analysis:

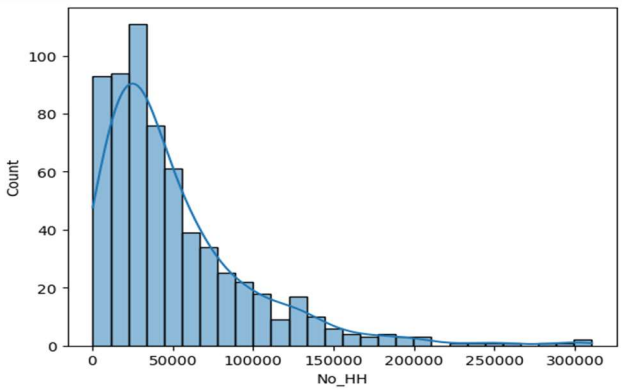


Fig 29 : Histogram of No_HH

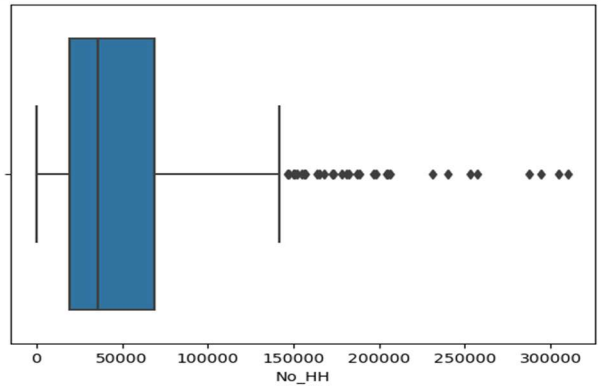


Fig 30 : Box Plot of No_HH

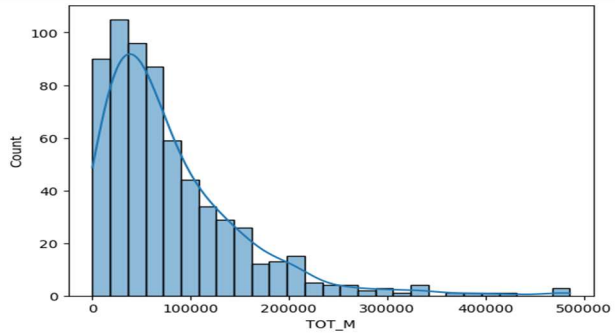


Fig 31 : Histogram of TOT_M

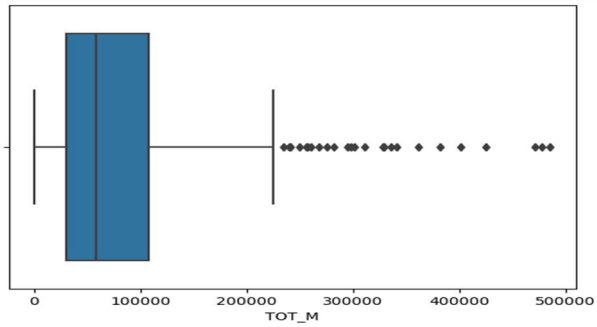


Fig 32 : Box Plot of TOT_M

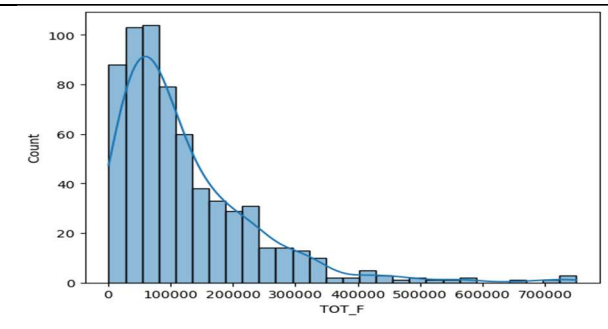


Fig 33 : Histogram of TOT_F

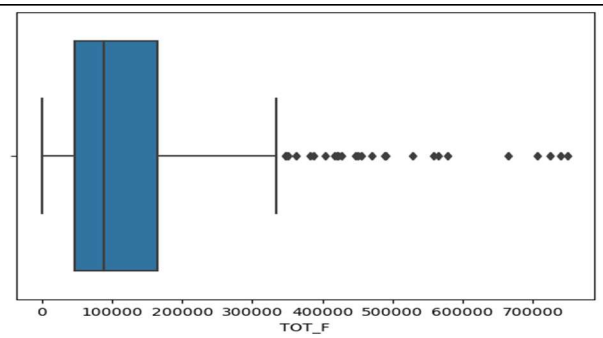


Fig 34 : Box Plot of TOT_F

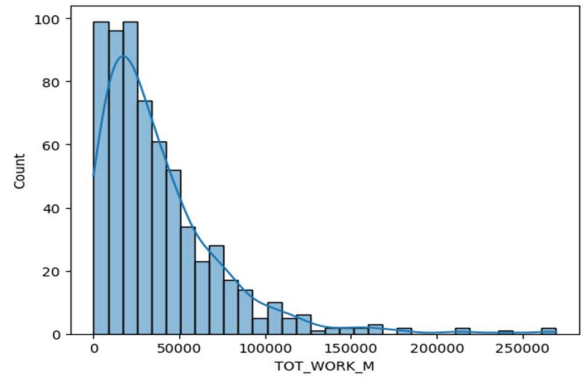


Fig 35 : Histogram of TOT_WORK_M

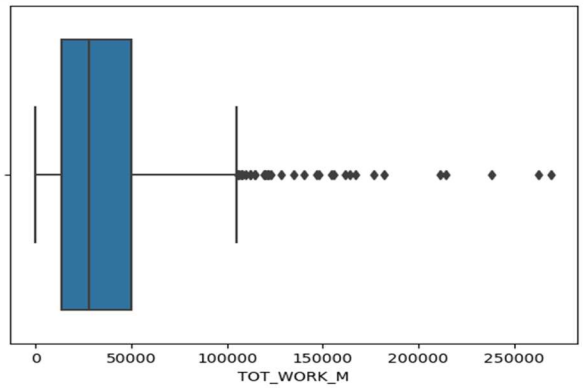


Fig 36 : Box Plot of TOT_WORK_M

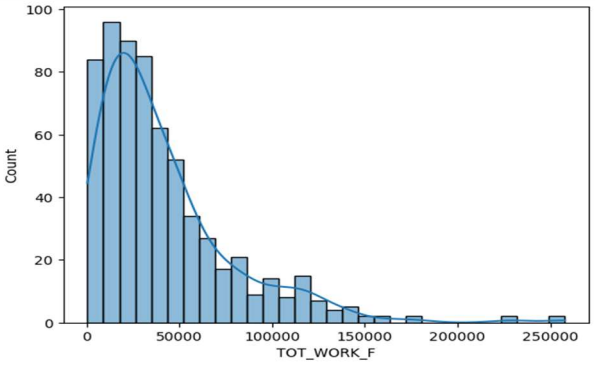


Fig 37 : Histogram of TOT_WORK_F

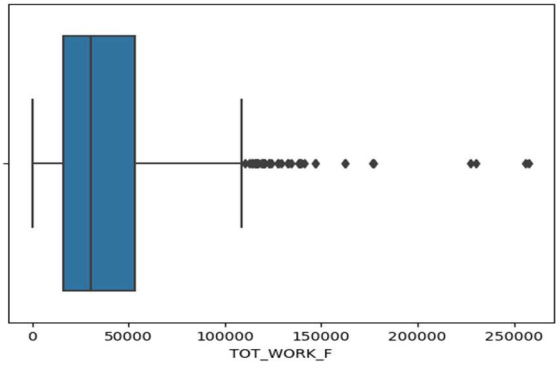


Fig 38 : Box Plot of TOT_WORK_F

- According to Univariate analysis, all five variables are left skewed and all have outliers.

Bivariate Analysis:

Now plot Scatterplots for No_HH, TOT_M, TOT_F, TOT_WORK_M, and TOT_WORK_F

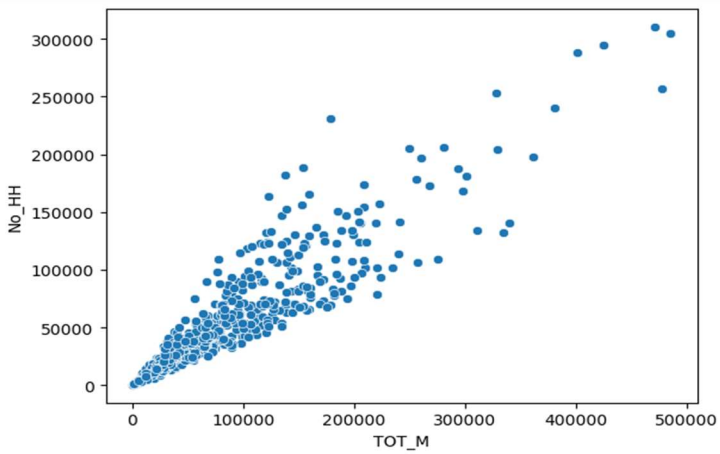


Fig 38 : Scatterplots for TOT_M

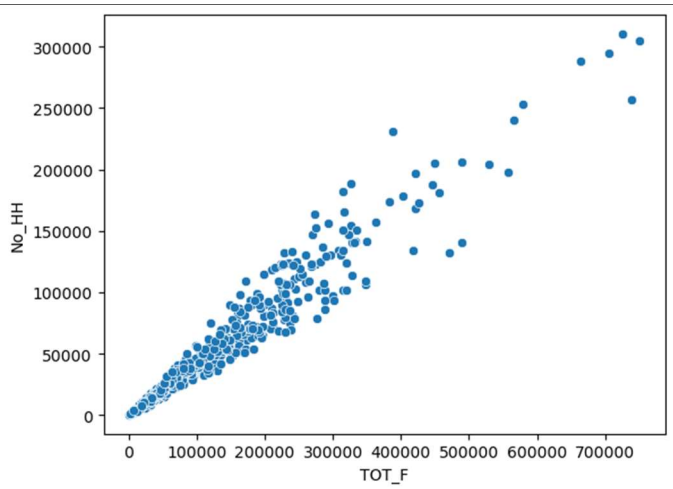


Fig 39 : Scatterplots for TOT_F

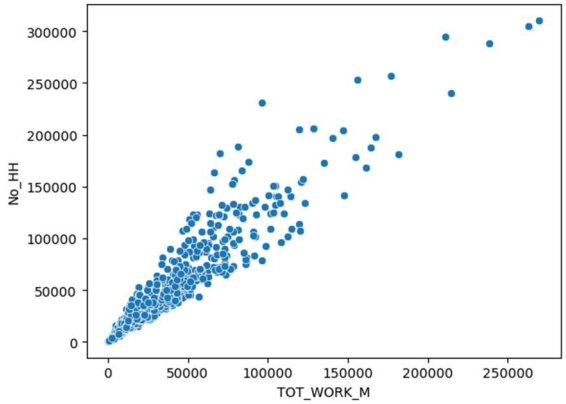


Fig 40 : Scatterplots for TOT_WORK _M

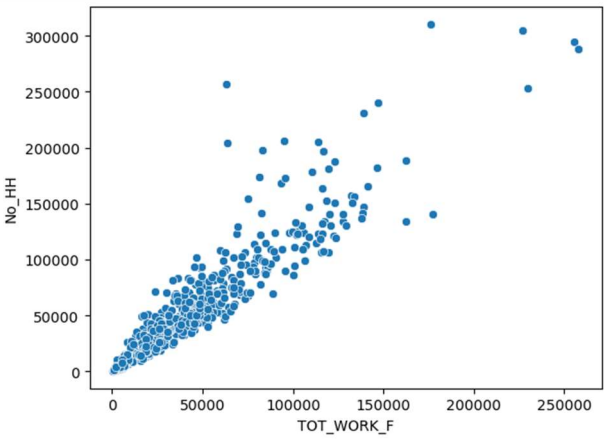


Fig 41 : Scatterplots for TOT_WORK _F

- According to Bivariate Analysis, all five variables are Positively Co-related to each other.

2. (i) Which state has highest gender ratio and which has the lowest?

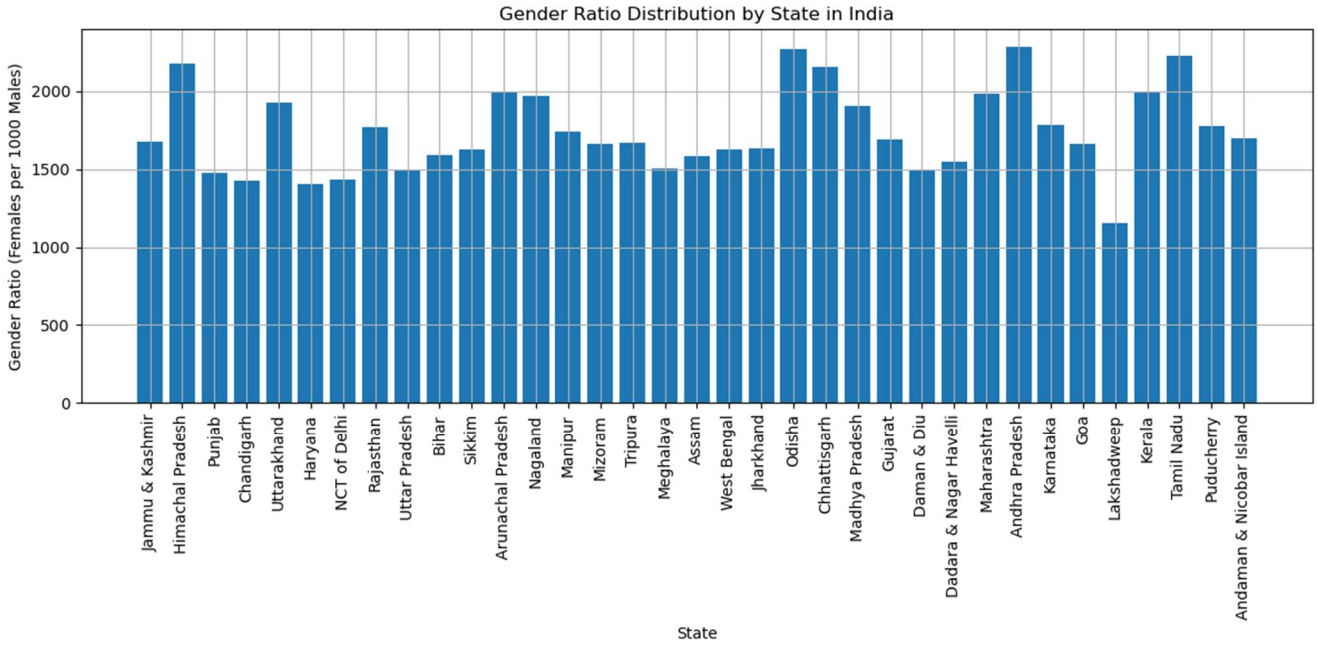


Fig 42 : Gender Ratio Distribution by State in India

- State with the highest gender ratio: Andhra Pradesh
- State with the lowest gender ratio: Lakshadweep

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

- District with the highest gender ratio: Krishna
- District with the lowest gender ratio: Lakshadweep

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

Outliers have a significant impact on clustering. So it is important to identify and remove outliers before applying the K-means Clustering algorithm.

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

Drop unnecessary features. like State Code, Dist.Code, State, Area Name, Gender_Ratio, and Total_Population

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MAR
0	7707	23388	29796	5862	6196	3	0	1999	2598	13381	...	1150	749	180	237	
1	6218	19585	23102	4482	3733	7	6	427	517	10513	...	525	715	123	229	
2	4452	6546	10964	1082	1018	3	6	5806	9723	4534	...	114	188	44	89	
3	1320	2784	4206	563	677	0	0	2666	3968	1842	...	194	247	61	128	
4	11654	20591	29981	5157	4587	20	33	7670	10843	13243	...	874	1928	465	1043	

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
749	180	237	680	252	32	46	258	214
715	123	229	186	148	76	178	140	160
188	44	89	3	34	0	4	67	61
247	61	128	13	50	4	10	116	59
1928	465	1043	205	302	24	105	180	478

Fig 43 : Top five data of dataset PCA_India_cleaned

Plot individual box plots for each numeric column with increased figure sizes in multiple lines

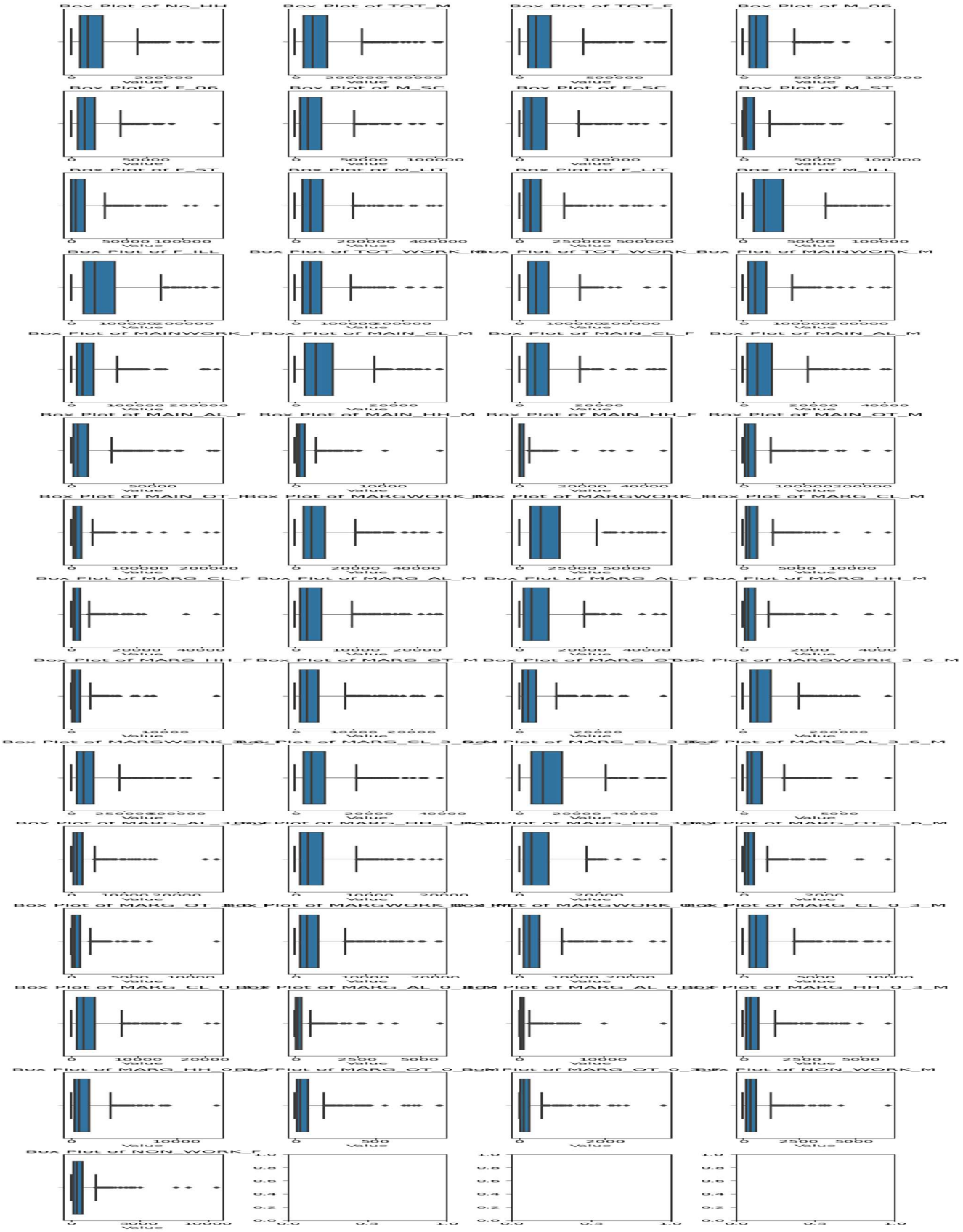


Fig 43 : Box Plot with outliers

Perform z-score scaling

Fig 44 : Box Plot with outliers

Conclusion: We can say that Scaling has no impact on outliers, according to before and after outlier trimming and visualization of box plots.

- For further processing, we treat the outliers.

To treat outliers lets define a function 'treat_outlier'.

For the higher outliers we will treat it to get it at 95 percentile value.

Lower level outliers will be treated to get it at 5 percentile value.

Compare box plots before and after treating outliers

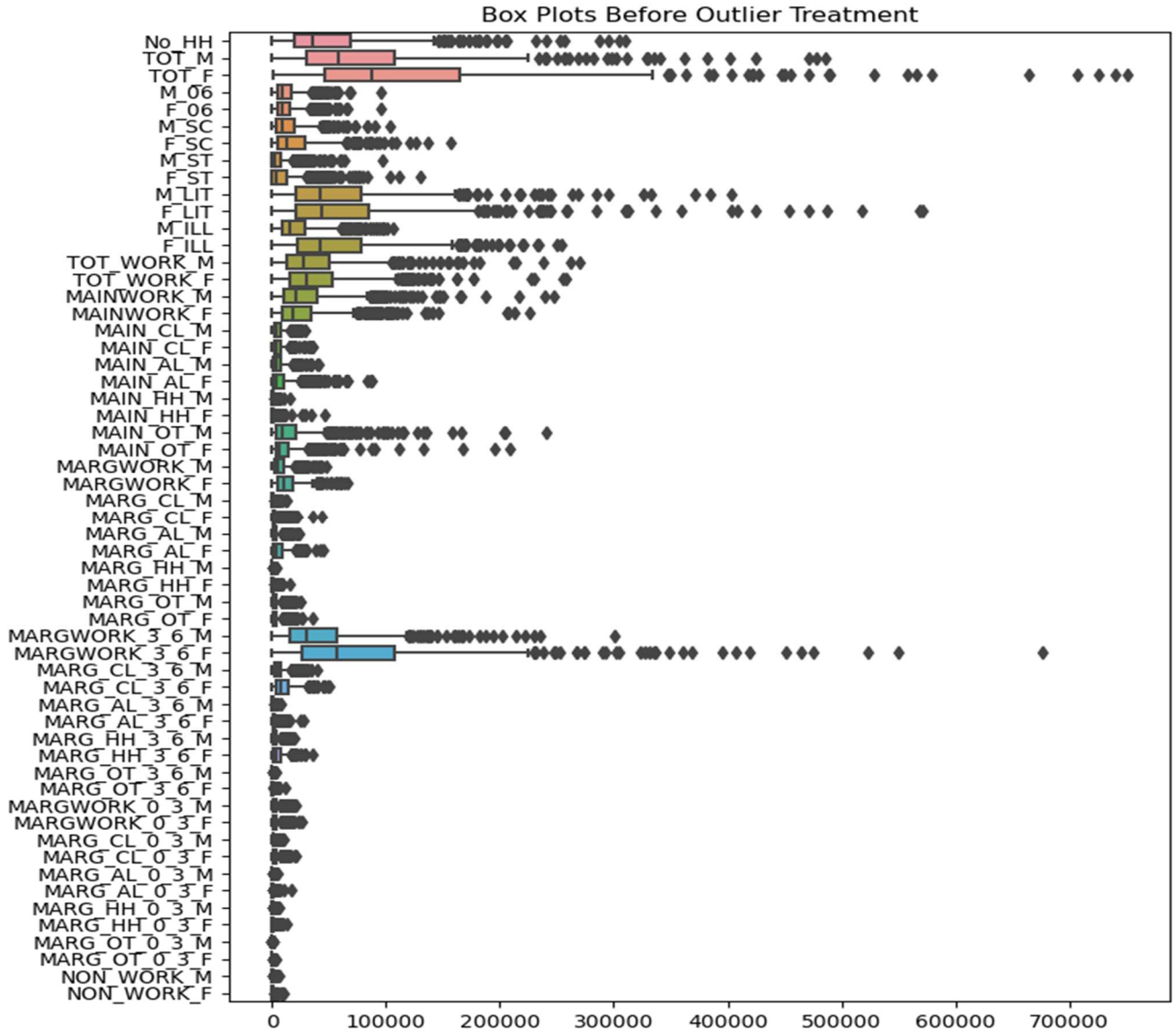


Fig 45 : Box Plots Before Outlier Treatment

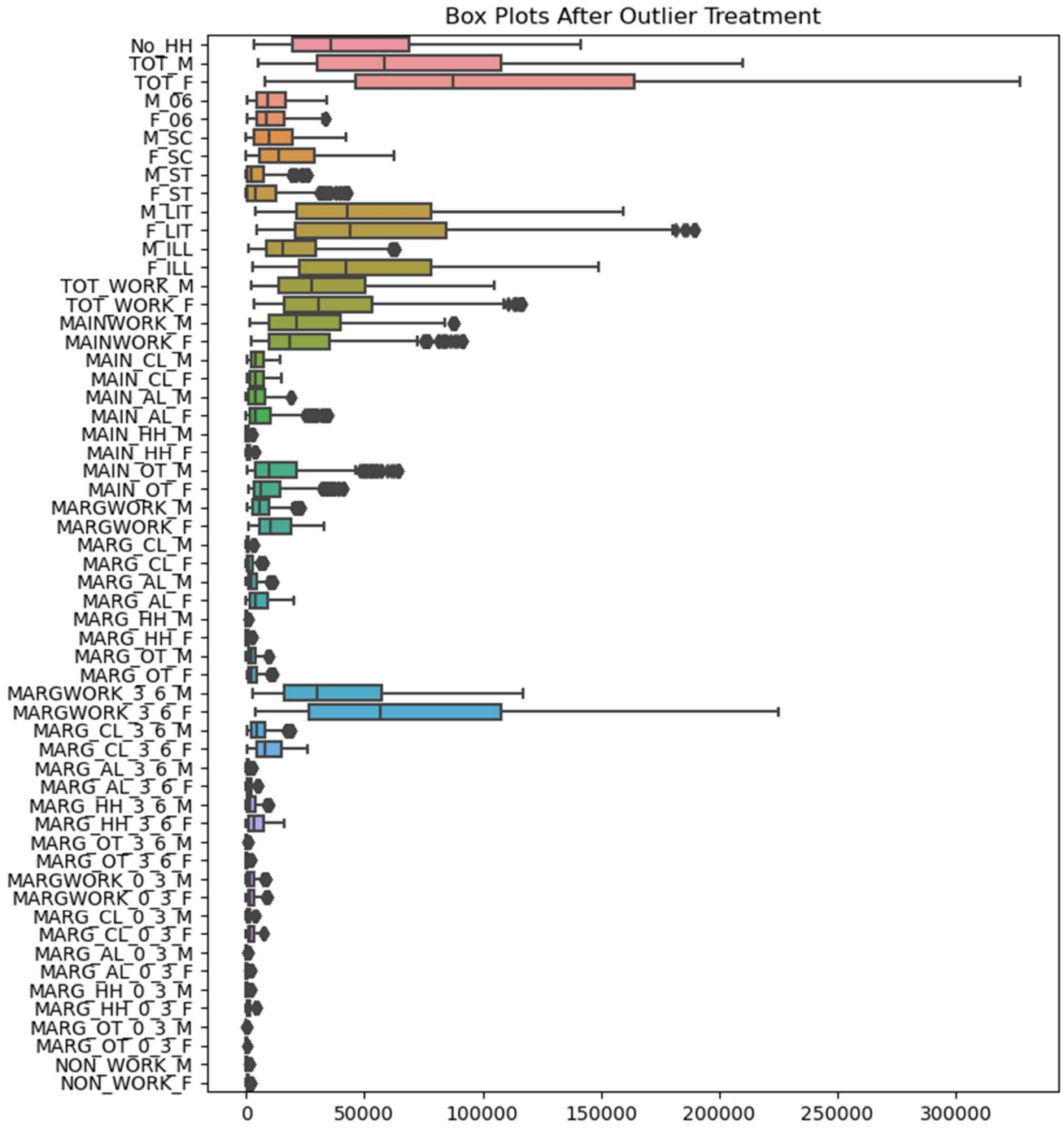


Fig 46 : Box Plots After Outlier Treatment

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

```
covariance_matrix:
[[1.00156495 0.91288416 0.97304568 ... 0.64390892 0.76800425 0.79093271]
 [0.91288416 1.00156495 0.97984644 ... 0.72559871 0.85906948 0.77715037]
 [0.97304568 0.97984644 1.00156495 ... 0.70486133 0.83628756 0.80301997]
 ...
 [0.64390892 0.72559871 0.70486133 ... 1.00156495 0.75100717 0.70813034]
 [0.76800425 0.85906948 0.83628756 ... 0.75100717 1.00156495 0.9048239 ]
 [0.79093271 0.77715037 0.80301997 ... 0.70813034 0.9048239  1.00156495]]
```

Fig 47 : Covariance_matrix

```
Eigenvalues:
[3.51964508e+01 7.75864164e+00 3.85313758e+00 2.93088251e+00
 2.01945117e+00 1.20283006e+00 1.05103883e+00 4.73410095e-01
 3.80311737e-01 3.18502503e-01 2.82337363e-01 2.11384169e-01
 1.80828864e-01 1.58534721e-01 1.34716174e-01 1.18114933e-01
 1.03095461e-01 8.94103559e-02 8.17461188e-02 7.41246556e-02
 6.82600138e-02 5.76145806e-02 4.88597660e-02 3.95906909e-02
 3.58001574e-02 2.79156242e-02 2.45501882e-02 2.36365599e-02
 2.13936483e-02 1.85082625e-02 1.52904953e-02 1.40346774e-02
 1.15247344e-02 1.13716166e-02 9.12573518e-03 7.91717986e-03
 6.59703591e-03 5.48569071e-03 4.70879023e-03 3.35011299e-03
 2.82906202e-03 2.38708937e-03 2.09550435e-03 1.32016186e-03
 1.01120293e-03 9.02948881e-04 8.47903528e-04 6.75743915e-04
 5.93576716e-04 5.58134687e-04 4.66754410e-04 2.92512691e-04
 2.31094543e-04 1.87648731e-04 1.20499123e-04 1.09839330e-04
 8.69325448e-05]
```

Fig 48 : Eigenvalues

```
Eigenvectors:
[[ 0.15023744  0.16052153  0.15955834 ...  0.14148045  0.14763628
   0.14086398]
 [-0.11528684 -0.07678683 -0.09111385 ...  0.0370082  -0.05070556
  -0.04588453]
 [ 0.10318023 -0.02970629  0.03432408 ... -0.10348503 -0.13764487
  -0.04249229]
 ...
 [-0.00593624 -0.04996136  0.0350455  ...  0.00897393  0.02677304
   0.00181857]
 [ 0.01726872 -0.02644952  0.00758134 ...  0.01022161 -0.08216062
   0.00725763]
 [ 0.00104469 -0.00707503  0.03342686 ... -0.00884273  0.07502936
  -0.00443845]]
```

Fig 49 : Eigenvectors

	0	1	2	3	4	5	
Principal Component	PC1	PC2	PC3	PC4	PC5	PC6	
Explained Variance	0.616517	0.135904	0.067493	0.051339	0.035374	0.021069	
Most Influential Variable	MARG_OT_M	MARG_AL_0_3_M	MARG_OT_3_6_M	MARG_OT_3_6_F	MARG_OT_3_6_F	F_ST	
Loadings	[0.8913076255492959, 0.4471221135545423, 0.313...	[-0.6839576012301155, -0.21388464655389436, -0...	[0.6121332181258797, -0.0827449122762405, 0.06...	[0.4427078726870926, 0.140776687303528, 0.1322...	[-0.09136713149763623, -0.15309921045839497, -...	[-0.38036863400160537, -0.21347432456281587, -...	[0.698220 -0.31992618

50	51	52	53	54	55	56
PC51	PC52	PC53	PC54	PC55	PC56	PC57
0.000008	0.000005	0.000004	0.000003	0.000002	0.000002	0.000002
MARG_CL_F	MARGWORK_3_6_M	MARG_AL_M	MAIN_OT_F	MARG_CL_F	MARG_OT_F	MAIN_HH_F
-0.1646845994114152, -0.9854161172255576, 0.3...	[0.08968656772209985, 0.2639838883416747, -0.0...	[-0.2009561054382648, -0.24629825508844666, 0....	[-0.1245907546924603, -0.24331224720089545, 0....	[-0.03521767620014727, -0.1391640544389626, 0....	[0.10244942446085425, -0.07367338793316812, 0....	[0.006197812416218915, -0.01970701580157716, 0...

Fig 50 : Explained Variance and Loadings

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

```
Cumulative Variance Explained in Percentage: [ 61.65  75.24  81.99  87.13  90.66  92.77  94.61  95.44  96.11  96.66
  97.16  97.53  97.85  98.12  98.36  98.57  98.75  98.9   99.05  99.18
  99.3   99.4   99.48  99.55  99.61  99.66  99.71  99.75  99.79  99.82
  99.84  99.87  99.89  99.91  99.93  99.94  99.95  99.96  99.97  99.97
  99.98  99.98  99.99  99.99  99.99  99.99  99.99 100.   100.   100.
 100.   100.   100.   100.   100.   100.   100.  ]
```

Fig 51 : Cumulative Variance Explained in Percentage

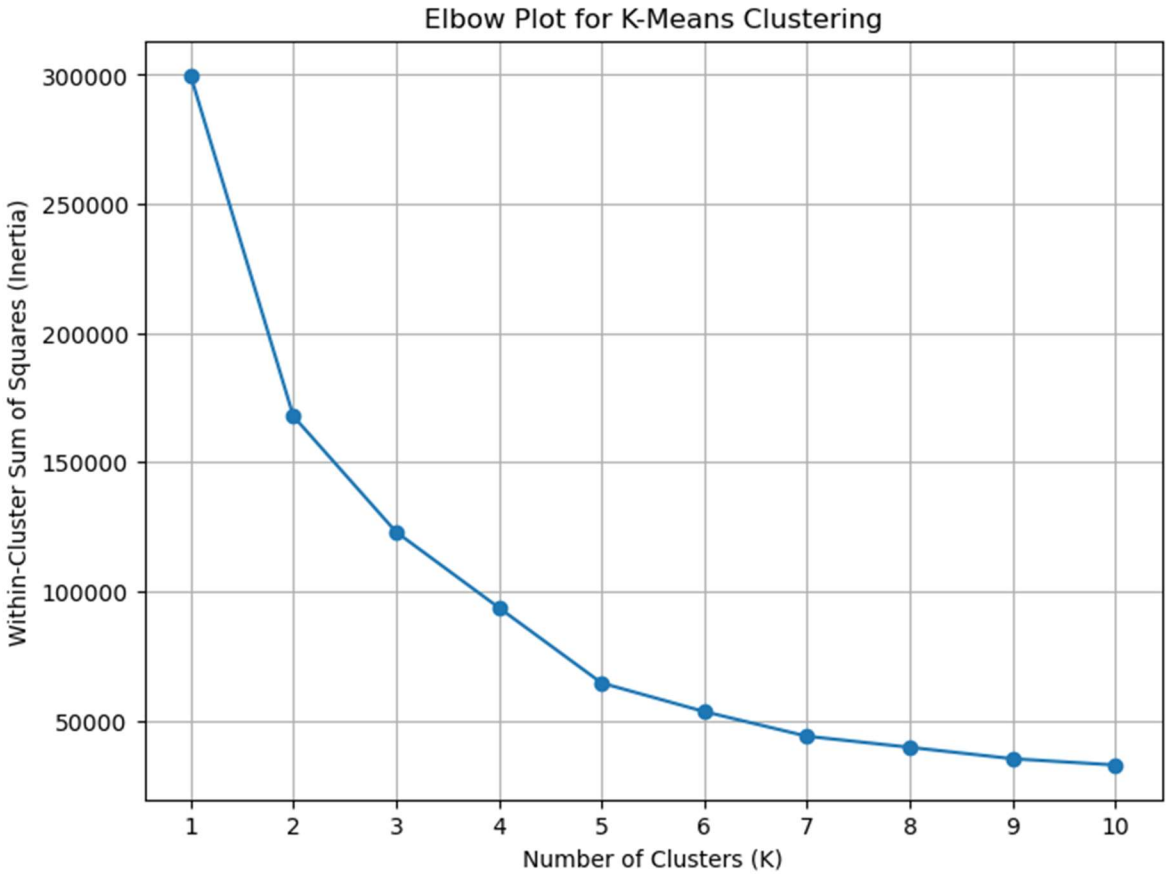


Fig 52 : Scree Plot

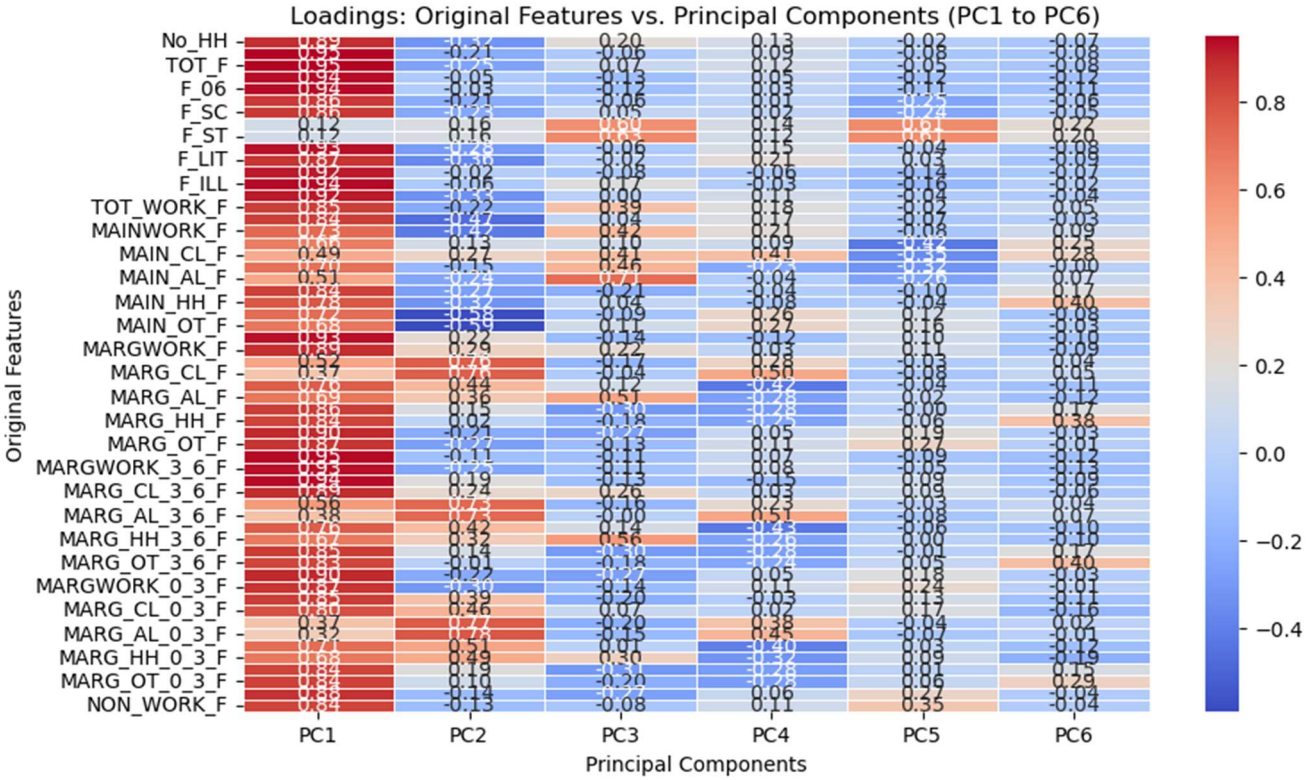
- Number of PCs to explain at least 90% variance: 5
-

7. 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.

	0	1	2	3	4	5	
Principal Component	PC1	PC2	PC3	PC4	PC5	PC6	
Explained Variance	0.616517	0.135904	0.067493	0.051339	0.035374	0.021069	
Most Influential Variable	TOT_M	MARG_AL_0_3_F	MAIN_AL_F	MARG_AL_3_6_F	F_ST	MARG_OT_3_6_F	
Loadings	[0.8913076255492959, -0.32112387165568484, 0.2...	[0.9523196349961867, -0.21388464655389436, -0....	[0.9466053020185625, -0.2537915993179117 3, 0.0...	[0.9350644173605095, -0.04938523181587537, -0....	[0.936991326356999, -0.03345853526890607, -0.1...	[0.8573479585881317, -0.21104317017131127, -0....	[0.8582 -0.2336

49	50	51	52	53	54	55	56
50	PC51	PC52	PC53	PC54	PC55	PC56	PC57
01	0.000008	0.000005	0.000004	0.000003	0.000002	0.000002	0.000002
_M	MARG_OT_F	MARG_HH_F	MARG_AL_F	MARG_HH_M	MARG_AL_M	MARG_OT_M	MARGWORK_M
55, 16, 2...	[0.32401882433500795, 0.7814343838588058, -0.1...	[0.7138742034956348, 0.5139539847458224, 0.009...	[0.676848004307706, 0.48877919398738073, 0.297...	[0.8360780269778895, 0.1863024352109115, -0.30...	[0.8393553547119904, 0.10308390285620143, -0.2...	[0.8758758091829465, -0.14123699204034063, -0....	[0.835698059777317, -0.12780831942991136, -0.0...

Fig 53 : variance_df.T



8 PCA: Write linear equation for first PC.

To write the linear equation for the first Principal Component (PC1), we need to consider the loadings of each original feature on PC1. The loadings represent the correlation between the original features and the first Principal Component. These loadings are stored in the first column of the loadings array obtained from PCA.

Let's assume the original features are represented by X1, X2, X3, ..., Xn, and the loadings for PC1 are represented by L1, L2, L3, ..., Ln. The linear equation for PC1 can be written as follows:

PC1 = L1 * X1 + L2 * X2 + L3 * X3 + ... + Ln * Xn

Here, PC1 is the value of the first Principal Component, X1, X2, X3, ..., Xn are the original features, and L1, L2, L3, ..., Ln are the loadings of each original feature on PC1.

Since the loadings array contains the loadings for each original feature on PC1, we can write the linear equation directly from the loadings array.