Business Report

ML 1 Project Coded

Anand B R | Batch PGPDSBA.O.JAN24

Contents

Clustering

1.1 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 DataFrame

```
df.head()
```

	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0	0.35	0.0	0.0031	0.0	0.0
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.0	0.35	0.0	0.0035	0.0	0.0
2	2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0	0.35	0.0	0.0028	0.0	0.0
3	2020-9-3-20	Format1	300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.0	0.35	0.0	0.0020	0.0	0.0
4	2020-9-4-15	Format1	300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.0	0.35	0.0	0.0041	0.0	0.0

Tail of DataFrame

```
df.tail()
```

	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
23061	2020-9-13-7	Format5	720	300	216000	Inter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
23062	2020-11-2-7	Format5	720	300	216000	Inter224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.0260	NaN	NaN	NaN
23063	2020-9-14-22	Format5	720	300	216000	Inter218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	NaN	NaN	NaN
23064	2020-11-18-2	Format4	120	600	72000	inter230	Video	Mobile	Video	7	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
23065	2020-9-14-0	Format5	720	300	216000	Inter221	App	Mobile	Video	2	2	2	1	0.09	0.35	0.0585	NaN	NaN	NaN

Shape of Data set:

Rows: 23066
Columns: 19

Null Values:

```
df.isna().sum()[df.isna().sum() > 0]
```

```
CTR     4736
CPM     4736
CPC     4736
dtype: int64
```

Data Set Information:

```
df.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
```

Duplicated Observations:

```
df.duplicated().sum()
```

```
0
```

Value Counts of Categorical Fields

```
Column InventoryType Value counts
________________________________________
InventoryType
Format4    7165
Format5    4249
Format1    3814
Format3    3540
Format6    1850
Format2    1789
Format7     659
Name: count, dtype: int64
****************************************
Column Ad Type Value counts
________________________________________
Ad Type
Inter224    1658
Inter217    1655
Inter223    1654
Inter219    1650
Inter221    1650
Inter222    1649
Inter229    1648
Inter227    1647
Inter218    1645
inter230    1644
Inter220    1644
Inter225    1643
Inter226    1640
Inter228    1639
Name: count, dtype: int64
****************************************
Column Platform Value counts
________________________________________
Platform
Video    9873
Web      8251
App      4942
Name: count, dtype: int64
****************************************
Column Device Type Value counts
________________________________________
Device Type
Mobile     14806
Desktop     8260
Name: count, dtype: int64
****************************************
Column Format Value counts
________________________________________
Format
Video      11552
Display    11514
Name: count, dtype: int64
****************************************
```

Data Statistics Numerical Columns

```
df_num.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	3.851631e+02	2.336514e+02	120.0000	120.000000	300.00000	7.200000e+02	728.00
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.0000	250.000000	300.00000	6.000000e+02	600.00
Ad Size	23066.0	9.667447e+04	6.153833e+04	33600.0000	72000.000000	72000.00000	8.400000e+04	216000.00
Available_Impressions	23066.0	2.432044e+06	4.742888e+06	1.0000	33672.250000	483771.00000	2.527712e+06	27592861.00
Matched_Queries	23066.0	1.295099e+06	2.512970e+06	1.0000	18282.500000	258087.50000	1.180700e+06	14702025.00
Impressions	23066.0	1.241520e+06	2.429400e+06	1.0000	7990.500000	225290.00000	1.112428e+06	14194774.00
Clicks	23066.0	1.067852e+04	1.735341e+04	1.0000	710.000000	4425.00000	1.279375e+04	143049.00
Spend	23066.0	2.706626e+03	4.067927e+03	0.0000	85.180000	1425.12500	3.121400e+03	26931.87
Fee	23066.0	3.351231e-01	3.196322e-02	0.2100	0.330000	0.35000	3.500000e-01	0.35
Revenue	23066.0	1.924252e+03	3.105238e+03	0.0000	55.365375	926.33500	2.091338e+03	21276.18
CTR	18330.0	7.366054e-02	7.515992e-02	0.0001	0.002600	0.08255	1.300000e-01	1.00
CPM	18330.0	7.672045e+00	6.481391e+00	0.0000	1.710000	7.66000	1.251000e+01	81.56
CPC	18330.0	3.510606e-01	3.433338e-01	0.0000	0.090000	0.16000	5.700000e-01	7.26

Inference:

- Missing Values in Field CTR, CPM, CPC (4736 values missing)
  - Also, the values do not match the formula provided
- No Duplicated Rows
- Column Timestamp is converted to Date Time (for future use case)
- 5 Categorical Columns (excluding Timestamp)
- 13 Numerical Columns

1.2 Treat missing values in CPC, CTR and CPM using the formula given.

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.*

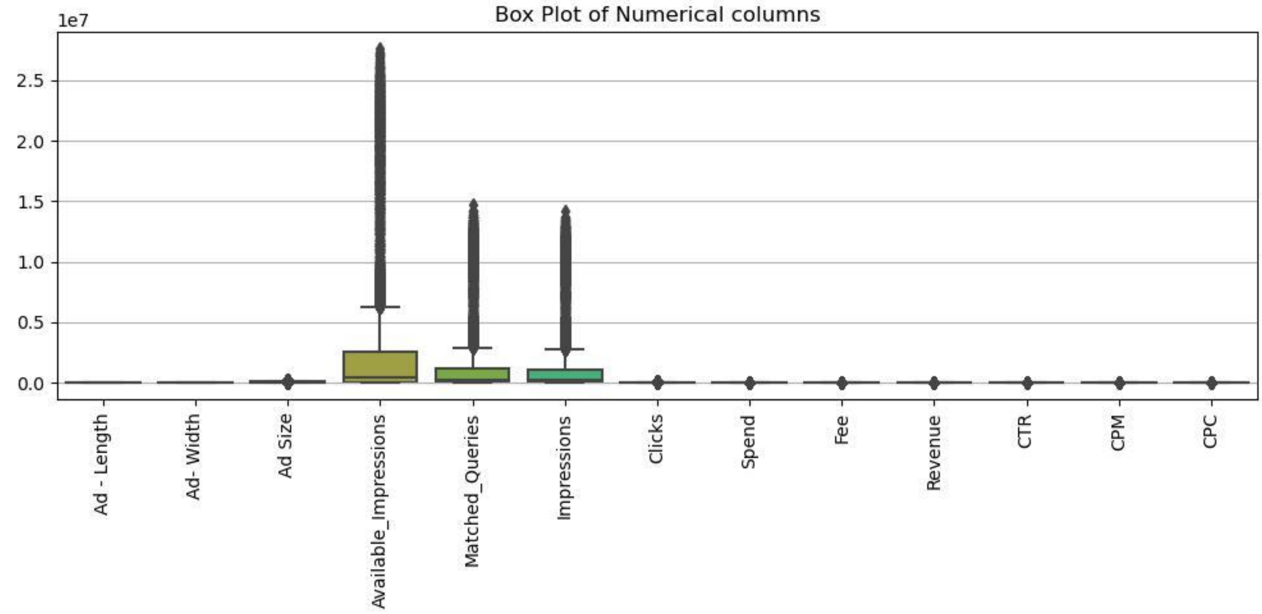
Inference:

- Columns CPM, CPC, CTR values are incorrect and contains Null values.
- Create a class and apply formula to populate the respective fields with appropriate values.
- Check for Null values after applying the functions to the columns.

Timestamp	0
InventoryType	0
Ad-Length	0
Ad-Wdith	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

1.3 Check if there are any outliers.

Columns	Outlier Count
'Ad Size'	8448
'Available_Impressions'	2378
'Matched_Queries'	3192
'Impressions'	3269
'Clicks'	1691
'Spend'	2081
'Fee'	3517
'Revenue'	2325
'CTR'	275
'CPM'	207
'CPC'	585

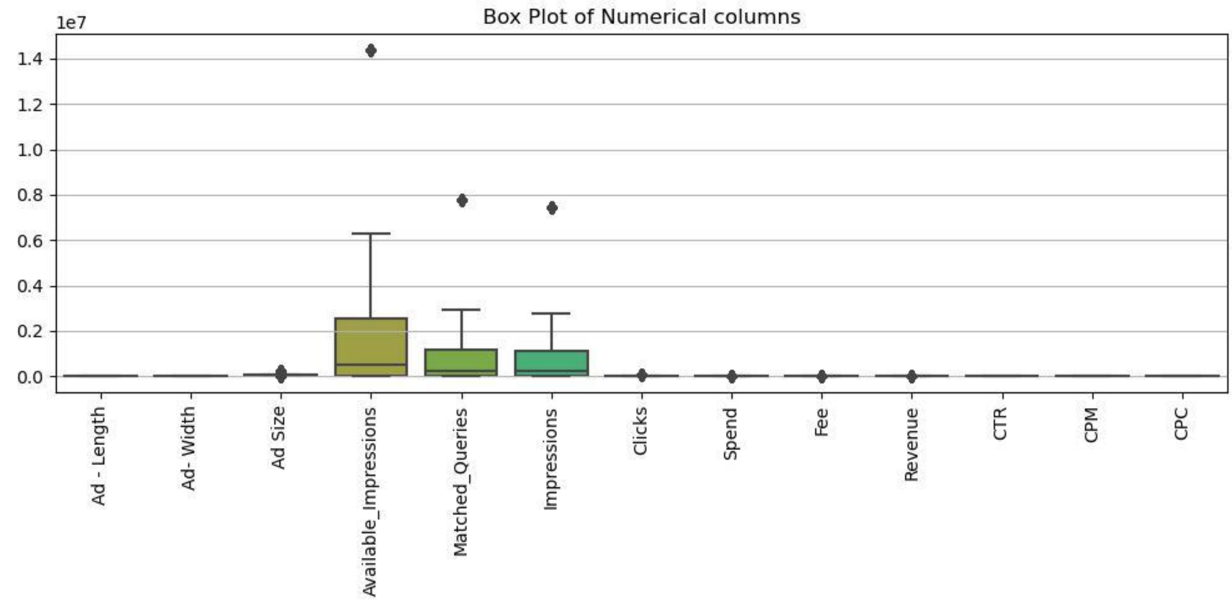


Insights:

- Except Columns 'Ad-Length' & 'Ad-Width' all other columns exhibit Outliers, and the count is displayed in the table above.
- Unsupervised learning is sensitive to Outliers, treatment of Outliers is Recommended.
- Outliers will impact the clustering as algorithms are Distance based. Homogeneity within Clusters v/s Heterogeneity between Clusters
- Outliers can Create additional Clusters lowering overall quality of Cluster analysis.
- In further steps K-means clustering are performed, these Outliers can significantly influence the distance between points, leading incorrect Centroid.

Outlier Treatment:

- 25^{th} and 75^{th} percentile is calculated.
- IQR is the distance between 75^{th} and 25^{th} percentile.
- Lower limit is 1.5* IQR – 25^{th} percentile.
- Upper limit is 1.5 * IQR + 75^{th} percentile.
- Replace all the values below lower limit with 5^{th} percentile.
- Replace all the values above upper limit with 95^{th} percentile.



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

Feature Scaling: With Clustering Techniques relying on distance measure to group based on Homogeneity, fields with larger scales or variances may dominate the distance calculations, leading to biased clustering results. Scaling features to similar range ensures that each feature contributes proportionally to the distance calculations, preventing any single feature dominating the clustering process.

DataFrame after Scaling:

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	-0.364496	-0.432797	-0.359227	-0.569484	-0.567061	-0.563943	-0.719779	-0.722776	0.487214	-0.676118	-0.978830	-1.220346	-1.083011
1	-0.364496	-0.432797	-0.359227	-0.569490	-0.567076	-0.563958	-0.719779	-0.722776	0.487214	-0.676118	-0.973650	-1.220346	-1.083011
2	-0.364496	-0.432797	-0.359227	-0.569269	-0.567049	-0.563931	-0.719779	-0.722776	0.487214	-0.676118	-0.982332	-1.220346	-1.083011
3	-0.364496	-0.432797	-0.359227	-0.569339	-0.566994	-0.563875	-0.719779	-0.722776	0.487214	-0.676118	-0.992329	-1.220346	-1.083011
4	-0.364496	-0.432797	-0.359227	-0.569622	-0.567093	-0.563975	-0.719779	-0.722776	0.487214	-0.676118	-0.965826	-1.220346	-1.083011

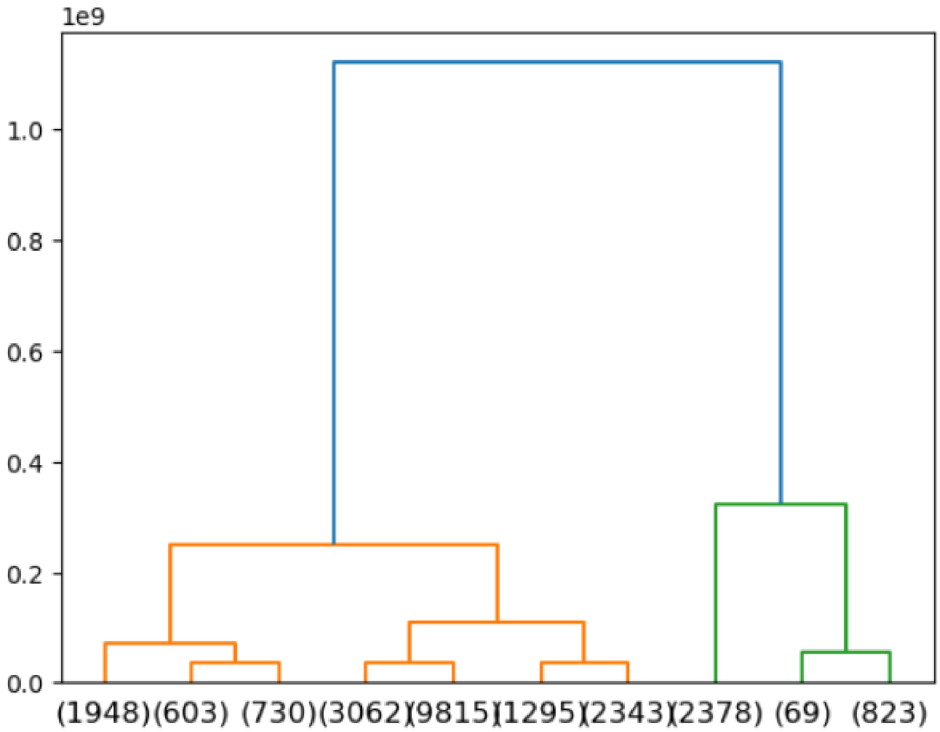
Statistical Summary after Scaling:

```
df100.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	1.281478e-16	1.000022	-1.134891	-1.134891	-0.364496	1.433093	1.467332
Ad- Width	23066.0	-1.182903e-16	1.000022	-1.319110	-0.432797	-0.186599	1.290590	1.290590
Ad Size	23066.0	-6.900268e-17	1.000022	-1.014296	-0.406696	-0.406696	-0.216821	1.871803
Available_Impressions	23066.0	3.943010e-17	1.000022	-0.569906	-0.562047	-0.456997	0.020045	2.782537
Matched_Queries	23066.0	-1.971505e-17	1.000022	-0.567185	-0.560154	-0.467925	-0.113087	2.434028
Impressions	23066.0	0.000000e+00	1.000022	-0.564071	-0.560898	-0.474599	-0.122278	2.403932
Clicks	23066.0	1.971505e-17	1.000022	-0.719779	-0.667456	-0.393291	0.224318	3.018972
Spend	23066.0	-2.365806e-16	1.000022	-0.722776	-0.699432	-0.332218	0.132649	2.812427
Fee	23066.0	1.143473e-15	1.000022	-2.323289	-0.074887	0.487214	0.487214	0.487214
Revenue	23066.0	3.943010e-17	1.000022	-0.676118	-0.656478	-0.347510	0.065763	2.755929
CTR	23066.0	-3.450134e-17	1.000022	-1.016315	-0.984413	0.160779	0.672672	3.133697
CPM	23066.0	-1.380054e-16	1.000022	-1.220346	-0.957898	0.035796	0.736591	3.277915
CPC	23066.0	-7.886020e-17	1.000022	-1.083011	-0.781463	-0.614748	0.752583	3.048123

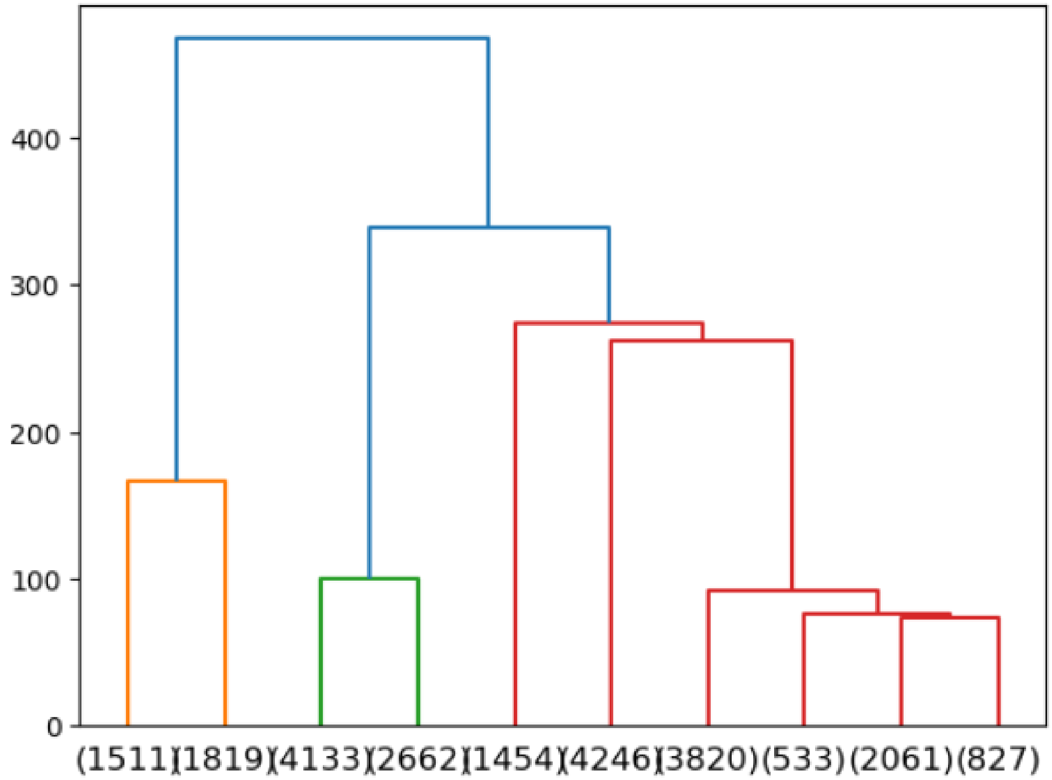
Without Feature Scaling:

```
Duration:  0.04626822471618652 seconds
```



WITH Feature Scaling:

```
Duration: 0.08644652366638184 seconds
```



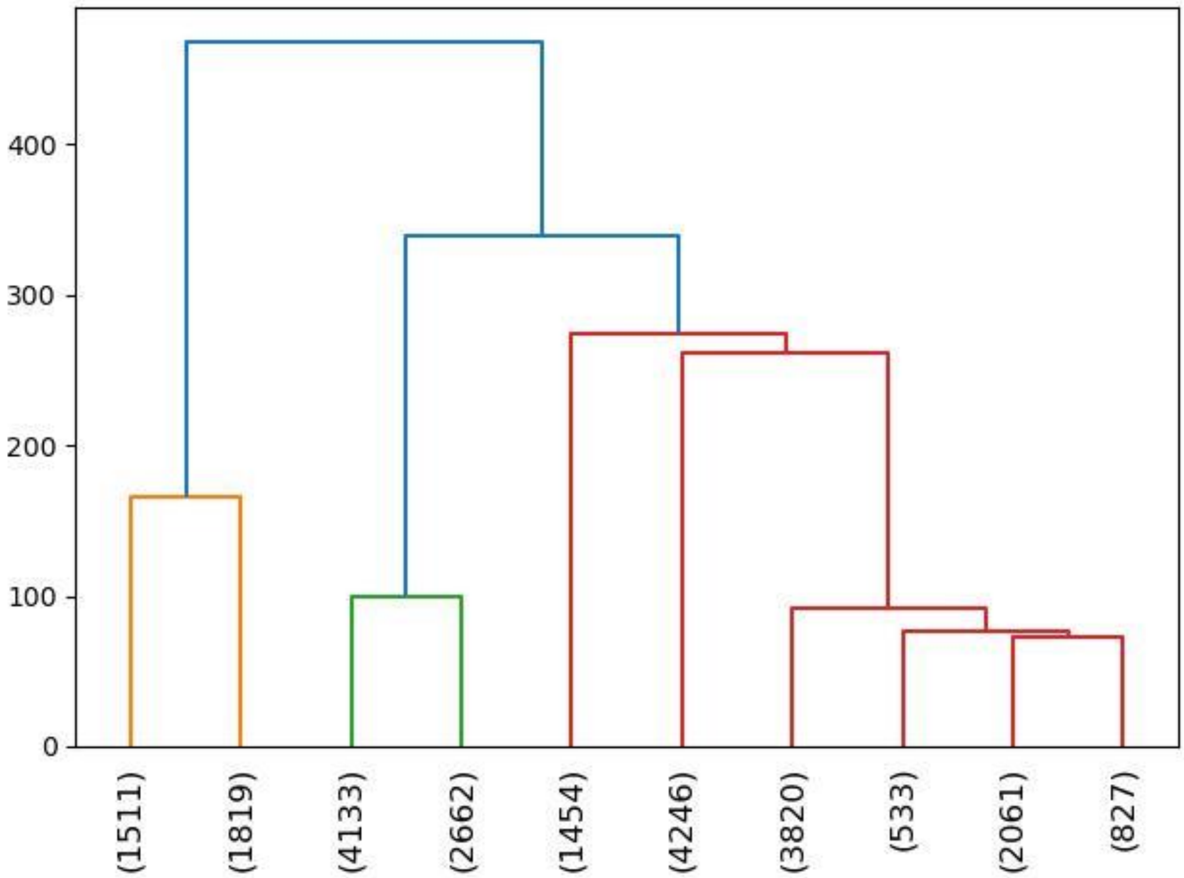
Inference:

- Clustering techniques with distance measure will take similar time complexity O(n) [Time complexity] with or without Feature Scaling. It does not directly affect the Speed of the Algorithm.
- Feature scaling impacts the accuracy of the clustering.
- Feature scaling leads to faster training and fitting the ML models (not very much applicable in clustering techniques)
- Feature Scaling can reduce the Space Complexity and reduce computational efficiency.

1.5 Perform clustering and do the following:

1.5.1 Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.

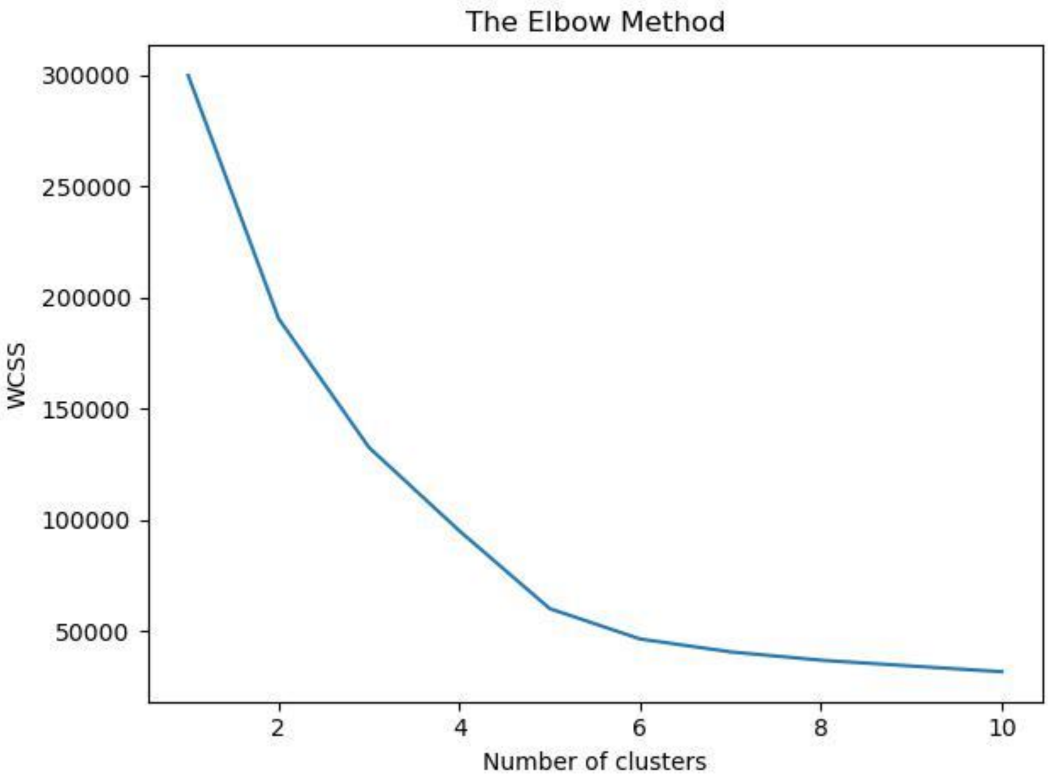
- Calculate Hierarchical clustering for feature scaled arrays (DataFrame) with Ward method and Euclidean Distance



Clusters	Freq
1	3330
2	6795
3	1454
4	4246
5	7241

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

- K-means algorithm Elbow Plot
  WSS (within-cluster sum of squared distances)
  Number of Clusters up to 10

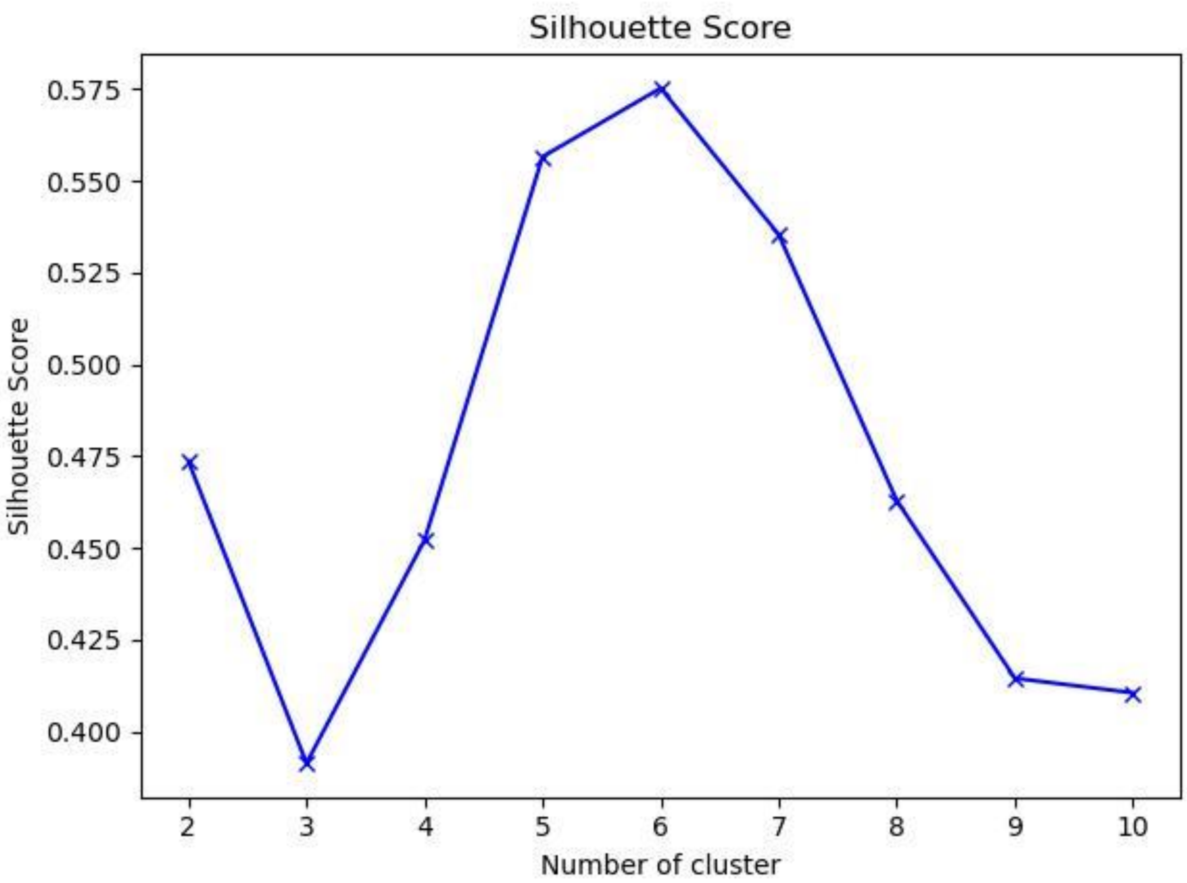


Interpretation:

- The Elbow point is at 6 (further this will be verified using silhouette score)
- Before the Elbow Point: Adding more Clusters leads to significant reduction in WCSS indicating that the clusters are becoming more compact and better capture the structure of Data.
- After the Elbow point: Adding more clusters leads to diminishing returns in terms of WCSS reduction, suggesting additional clusters do not provide much additional explanatory power and may even lead to overfit.

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

	Num of Clusters	Silhouette Score
0	2	0.473675
1	3	0.391359
2	4	0.452141
3	5	0.556591
4	6	0.575260
5	7	0.535362
6	8	0.462904
7	9	0.458505
8	10	0.464285



Inference:

- Silhouette score is maximum at k =6 (6 clusters)
- Choosing 6 clusters signifies appropriate Homogeneity within cluster and Heterogeneity between Clusters.
  silhouette score = 0.5752600558591118 at 6 Clusters

1.5.4 Profile the ads based on optimum number of clusters using silhouette score and your domain understanding.

- Grouping based on Mean values of Clusters.

Clusters	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Freq
0	149.554516	558.206665	76442.560655	4.658225e+04	2.866160e+04	2.125739e+04	6842
1	316.280182	254.538724	79328.337130	9.789532e+06	7.547121e+06	7.435034e+06	1756
2	680.940406	117.924034	71102.789784	1.431922e+07	7.803449e+06	7.473380e+06	1527
3	695.167922	316.803279	215619.849358	2.790594e+05	1.476652e+05	1.267586e+05	4514
4	142.182833	571.179344	76505.233775	8.434057e+05	5.911566e+05	4.987601e+05	1433
5	418.072634	157.144695	57160.817844	2.070385e+06	1.020575e+06	9.809877e+05	6994

Clusters	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	Freq
0	2947.786466	318.920140	0.349670	208.483293	15.520076	14.191189	0.101735	6842
1	8548.277904	4867.490575	0.298565	3334.230211	0.236940	1.377662	0.583717	1756
2	17394.944335	12708.127967	0.250000	9608.073495	0.187388	1.707285	0.890525	1527
3	14758.002437	1224.160505	0.349548	797.231842	12.782331	11.352640	0.094341	4514
4	50588.809491	8960.420656	0.260063	7196.285301	13.768415	15.125511	0.109844	1433
5	3451.112382	1763.331324	0.346617	1157.976570	0.392435	1.794809	0.538987	6994

Cluster 0:

- AD Length to AD width ratio is 0.26, AD Length smaller than AD width.
- AD Size range is 72000 – 216000.
- CTR Mean is 15.52, signifies 15 clicks when AD is shown 100 times. This type of AD has CPM of 14.19 and spend/click is 0.10.
- Standard deviation of CTR, CPM, CPC is 6.16, 4.77, 0.044 which is High.
- This type of AD can generate 0.65% of the Revenue for Total Spend.

Cluster 1:

- AD Length to AD width ratio is 1.24, AD Length is greater than AD width.
- AD Size range is 65520 - 216000.
- CTR Mean is 0.23, signifies less than a click when AD is shown 100 times. This type of AD has CPM of 1.37 and spend/click is 0.5.
- Standard deviation of CTR, CPM, CPC is 0.02, 0.2, 0.12.
- This type of AD can generate 0.68 of the Revenue for Total Spend.

Cluster 2:

- AD Length to AD width ratio is 5.77, AD Length is higher than AD width.
- AD Size range is 65520 - 216000.
- CTR Mean is 0.18, signifies less than a click when AD is shown 100 times. This type of AD has CPM of 1.70 and spend/click is 0.89.
- Standard deviation of CTR, CPM, CPC is 0.02, 0.26, 0.12.
- This type of AD can generate 0.75 of the Revenue for Total Spend.

Cluster 3:

- AD Length to AD width ratio is 2.19, AD Length is greater than AD width.
- AD Size range is 84000 - 216000.
- CTR Mean is 12.78, signifies 12 clicks when AD is shown 100 times. This type of AD has CPM of 11.35 and spend/click is 0.09.
- Standard deviation of CTR, CPM, CPC is 3.6, 3.5, 0.04.
- This type of AD can generate 0.65 of the Revenue for Total Spend.

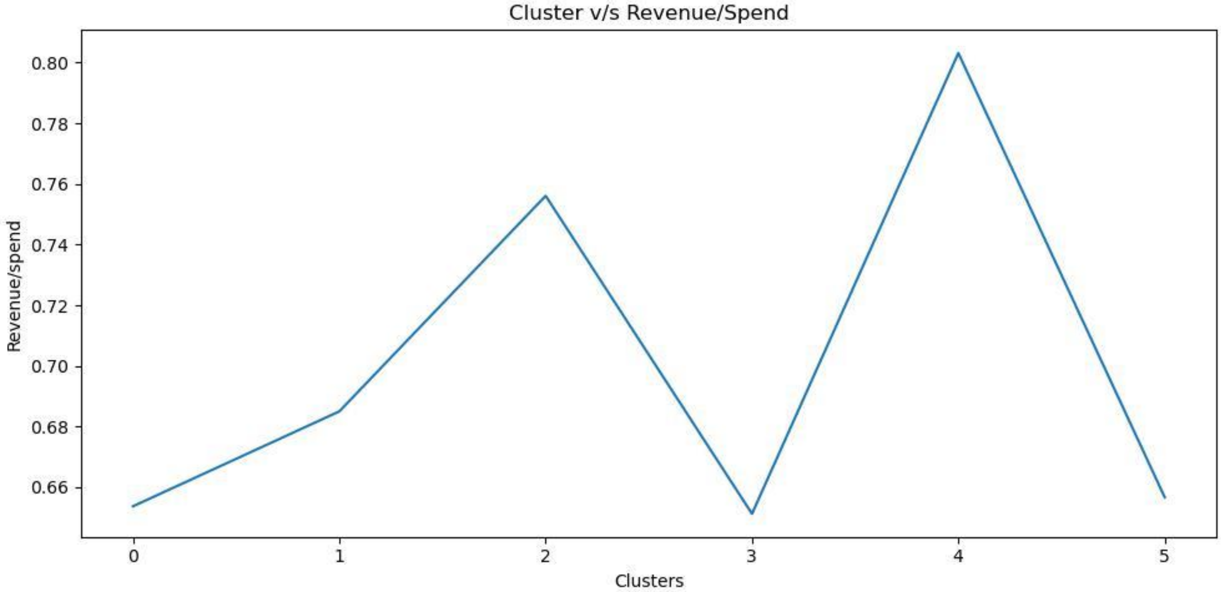
Cluster 4:

- AD Length to AD width ratio is 0.24, AD Length is smaller than AD width.
- AD Size range is 72000 - 216000.
- CTR Mean is 13.76, signifies 14 clicks when AD is shown 100 times. This type of AD has CPM of 15.12 and spend/click is 0.10.
- Standard deviation of CTR, CPM, CPC is 1.1, 3.4, 0.02 – which is Narrow considering the Mean values.
- This type of AD can generate 0.80 of the Revenue for Total Spend.

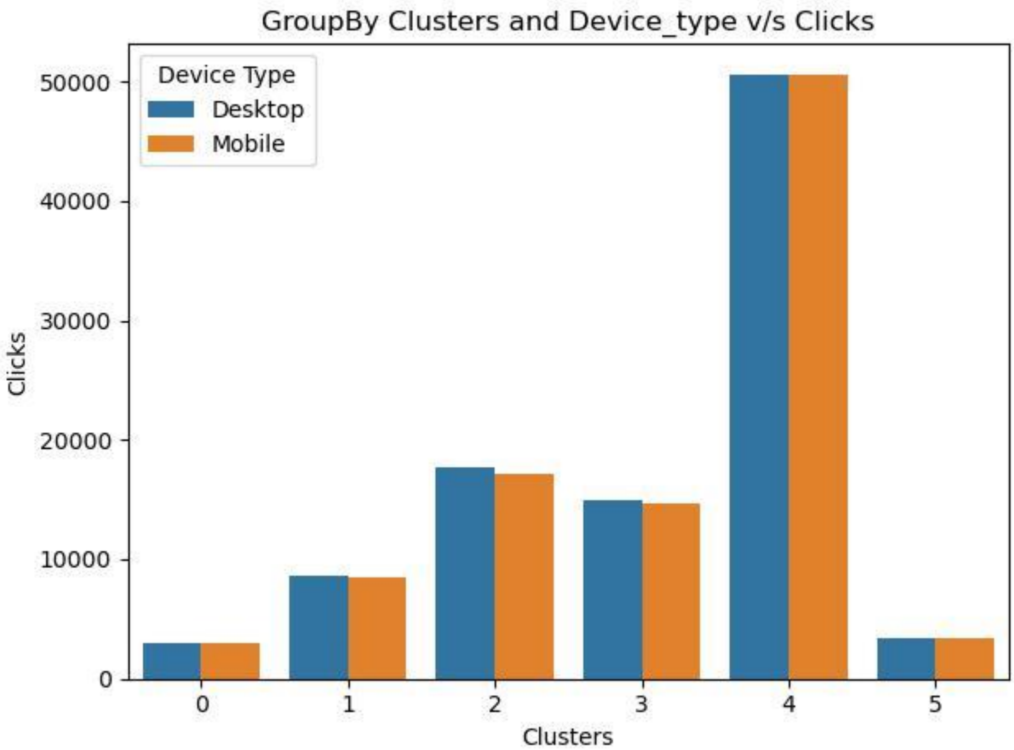
Cluster 5:

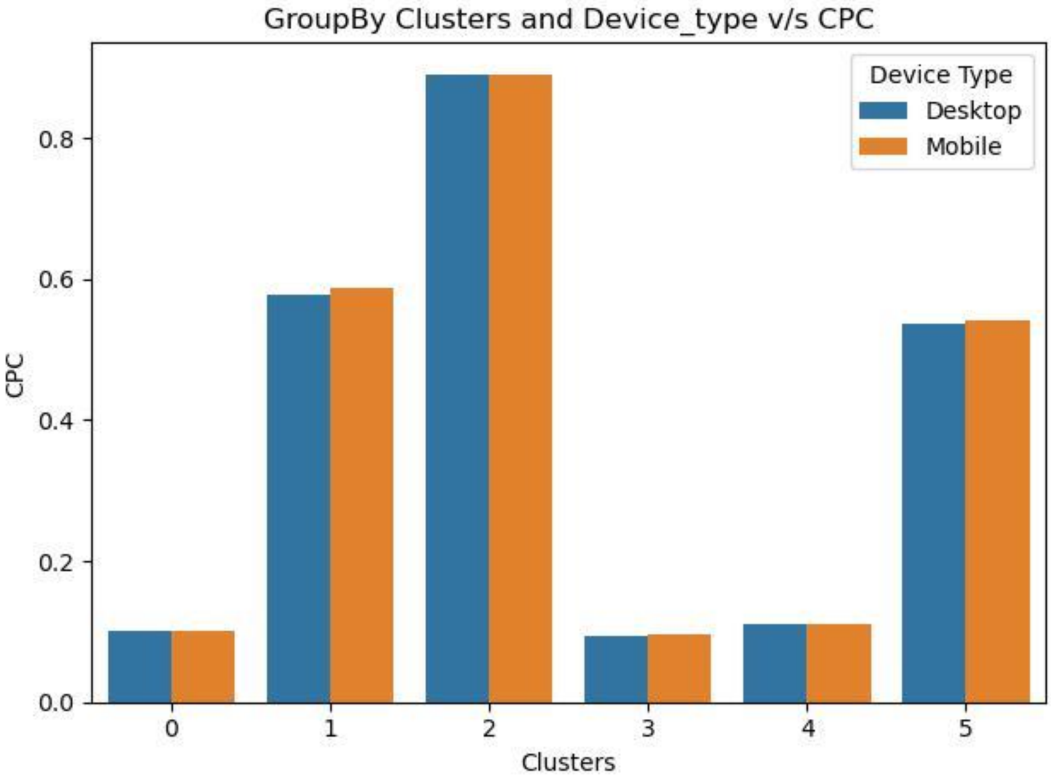
- AD Length to AD width ratio is 2.66, AD Length is greater AD width.
- AD Size range is 33600 - 216000.
- CTR Mean is 0.39, signifies less than a click when AD is shown 100 times. This type of AD has CPM of 1.7 and spend/click is 0.53.
- Standard deviation of CTR, CPM, CPC is 0.29, 0.64, 0.23.
- This type of AD can generate 0.65 of the Revenue for Total Spend.

Cluster V/s Revenue/Spend

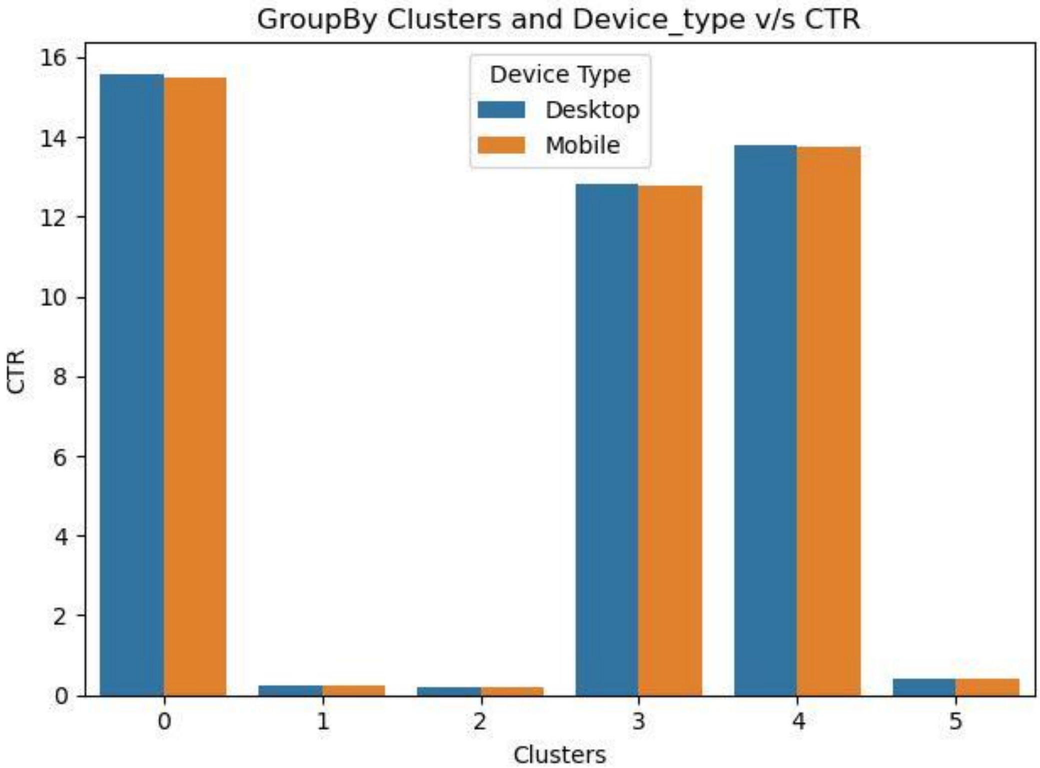


- Clicks based on Device Type in each Clusters.




GroupBy Clusters and Device_type v/s CPC
Device Type
Desktop
Mobile
CPC
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
Clusters

-


GroupBy Clusters and Device_type v/s CTR
Device Type
Desktop
Mobile
CTR
0
2
4
6
8
10
12
14
16
0
1
2
3
4
5
Clusters

Inference:

Cluster 4:

- generating a greater Revenue/Spend 0.8 whose AD Length is less than AD width.
- If AD Length is lesser than AD width, then maintain the AD size around 76505.
    - Maintain the AD size around 76505.
    - Maintain the Ratio of the AD Length 0.25 times of AD Width
- Mean Fee payable 0.26

Cluster 2:

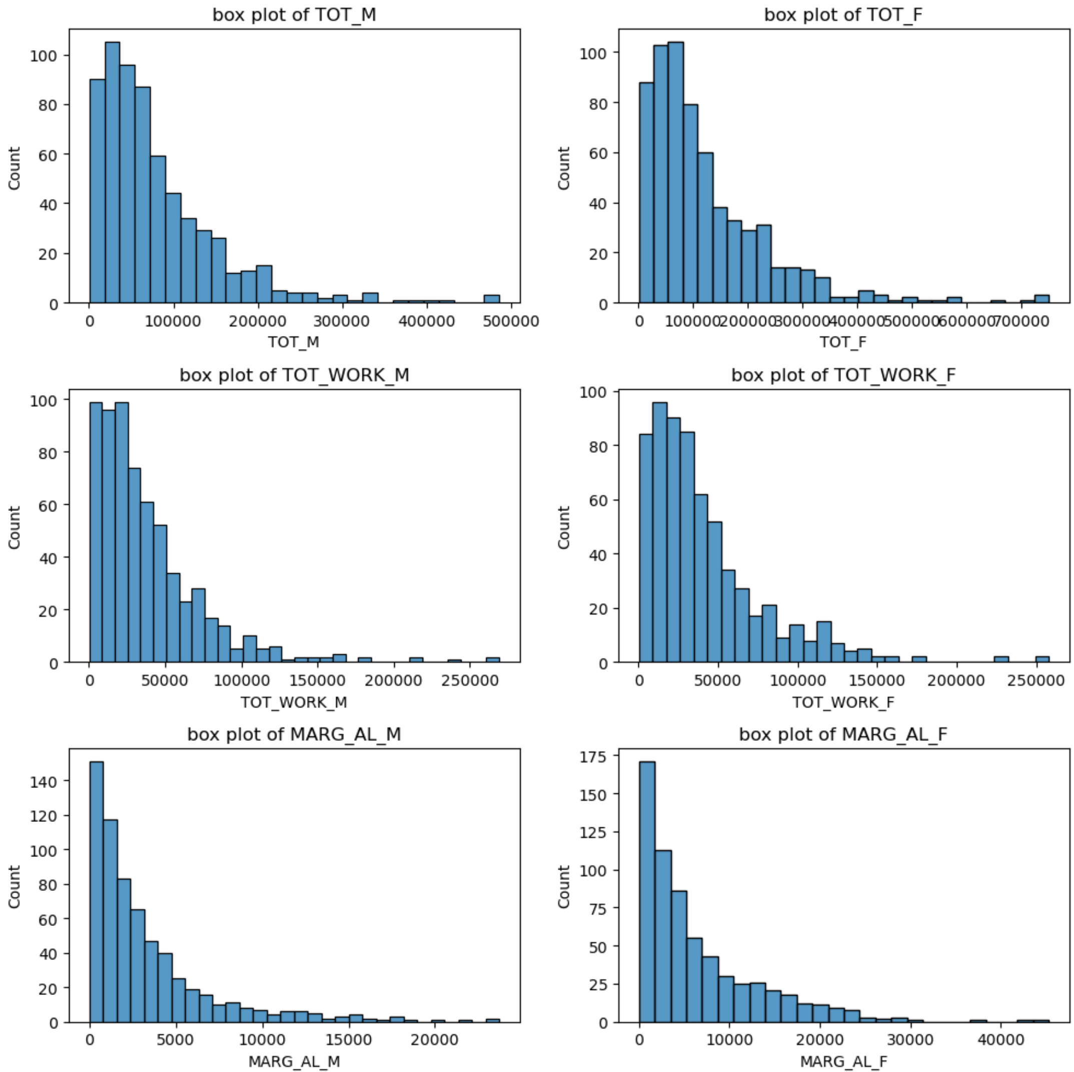
- ranks 2 with 0.75 whose AD Length is higher than its AD width.
- If AD Length is greater than AD Width.
    - maintain the AD size around 71102.
    - Maintain the Ratio of the Ad Length 5 times AD Width
- Mean Fee Payable is 0.25

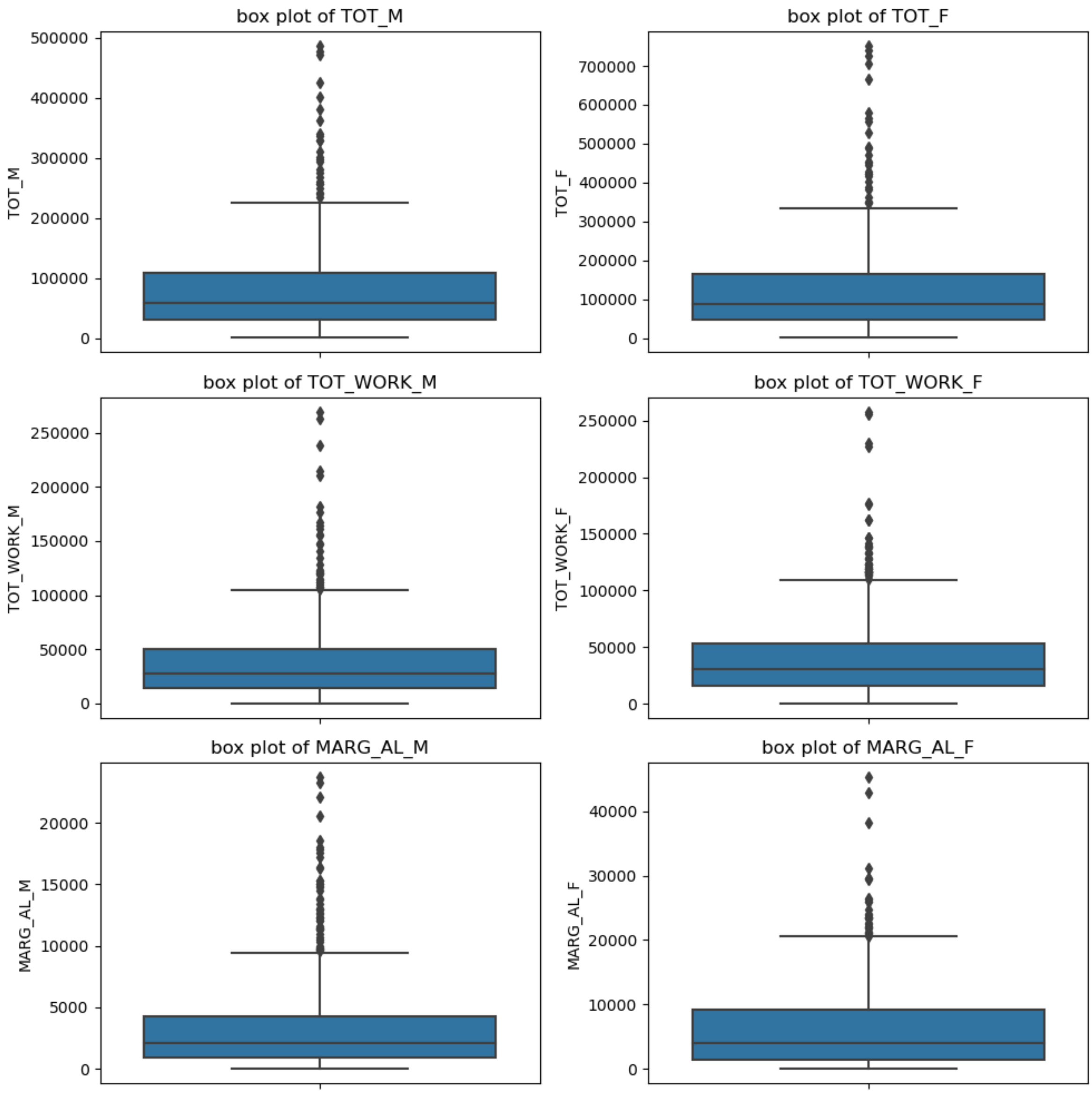
Parameters compared in different device types is Similar.

PCA

2.1 Define the problem and perform Exploratory Data Analysis

- Data has 640 rows and 61 Columns.
    - 2 columns are Object.
    - 2 columns are Categorical (Dist. code and State code)
    - 57 columns are Numerical.
        - 6 columns for EDA





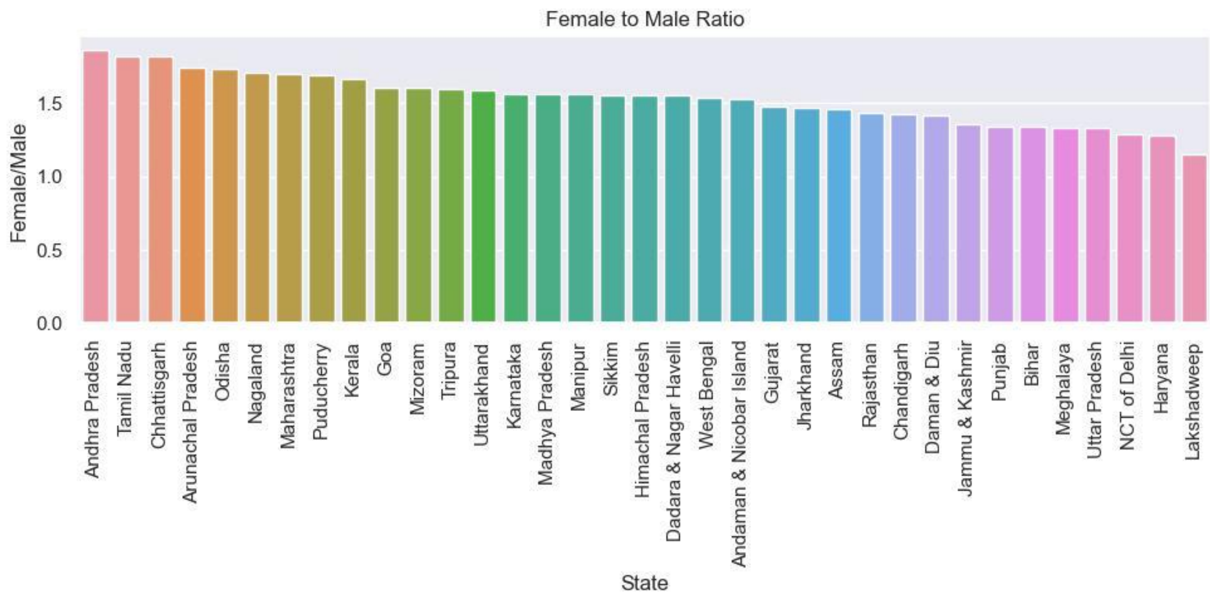
(i) Which state has the highest gender ratio, and which has the lowest?

- State with Lowest Gender Ratio – Lakshadweep. For every 1000 Male there is 1151 Female.

	TOT_F	TOT_M	Female/Male
State			
Lakshadweep	14772	12823	1.151993
Haryana	1498873	1167816	1.283484

- State with Highest Gender Ratio – Andhra Pradesh. For every 1000 Male there is 1862 Female.

State	TOT_F	TOT_M	Female/Male
Andhra Pradesh	6097235	3274363	1.862113
Tamil Nadu	5610310	3074009	1.825079



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

Dist.Code	TOT_F	TOT_M	Female/Male
547	314182	137603	2.283250
398	86272	38026	2.268763

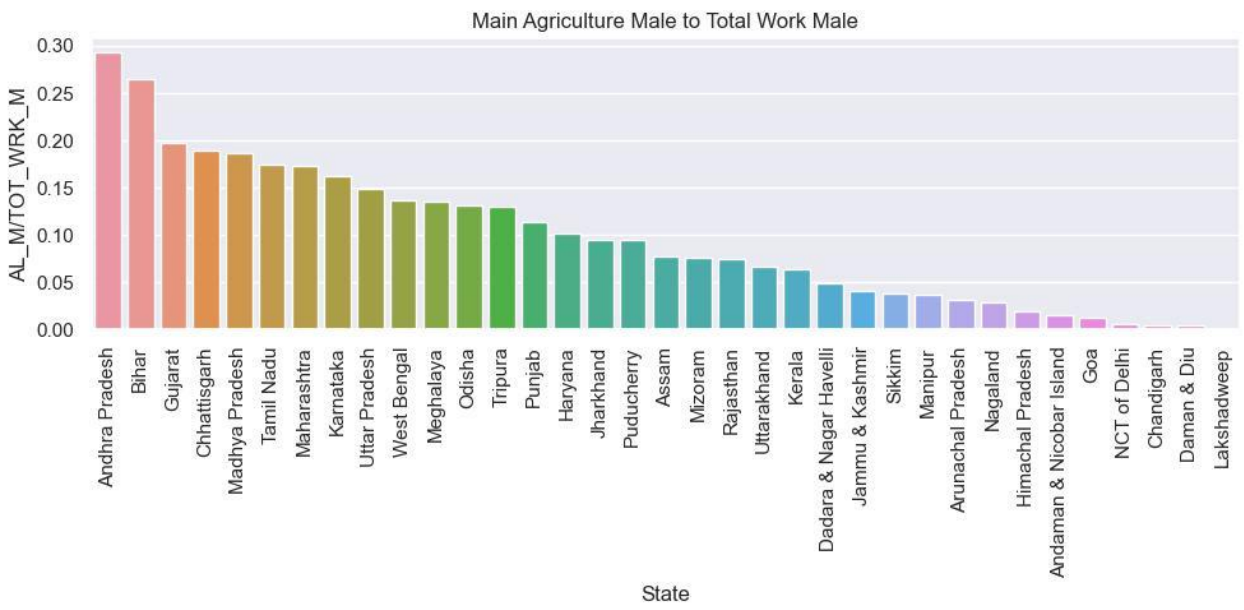
Dist.Code	TOT_F	TOT_M	Female/Male
587	14772	12823	1.151993
2	23102	19585	1.179576

- District with highest Female to Male ratio:
  - State: Andhra Pradesh, Area Name: Krishna

- District with Lowest Female to Male ratio:
    - State: Lakshadweep, Area Name: Lakshadweep

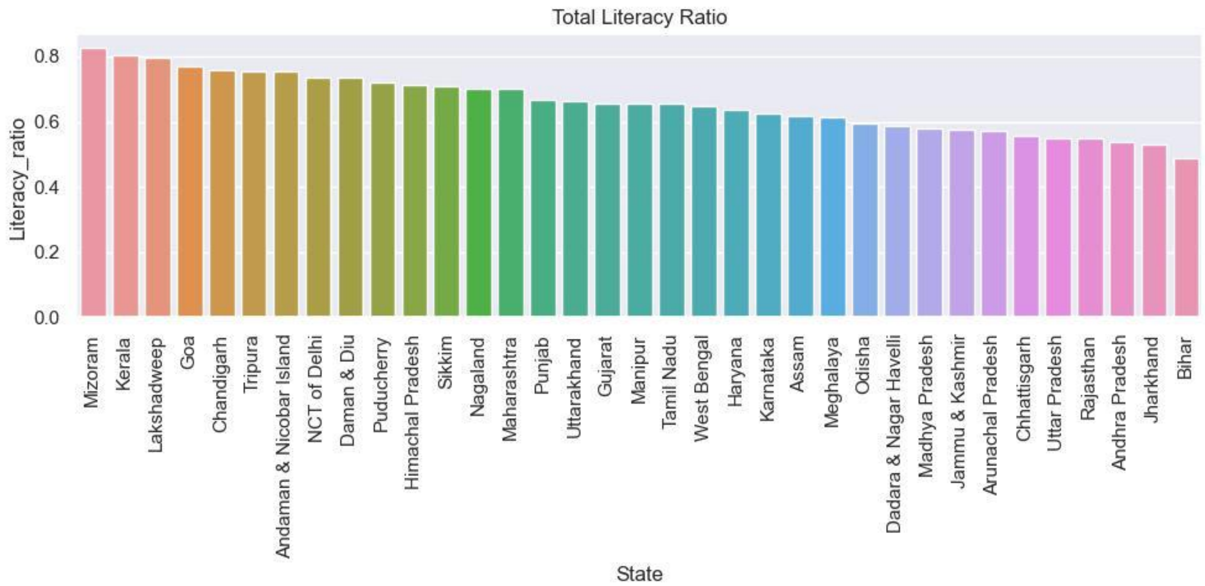
(iii) Which state has highest ratio of Main Agricultural Labourers Population Male v/s Total Worker Population Male.
Andra Pradesh has highest Ratio of Male as Main Agricultural Labor of Total Worker Population.

State	TOT_WORK_M	MAIN_AL_M	AL_M/TOT_WRK_M
Andhra Pradesh	1674517	490307	0.292805
Bihar	1524553	403261	0.264511
Gujarat	1057781	208481	0.197093
Chhattisgarh	398935	75308	0.188773
Madhya Pradesh	1004639	187847	0.186980



(iv) Which state has highest Literacy Rate (both Male and Female combined)
- Mizoram has highest Literacy Ratio

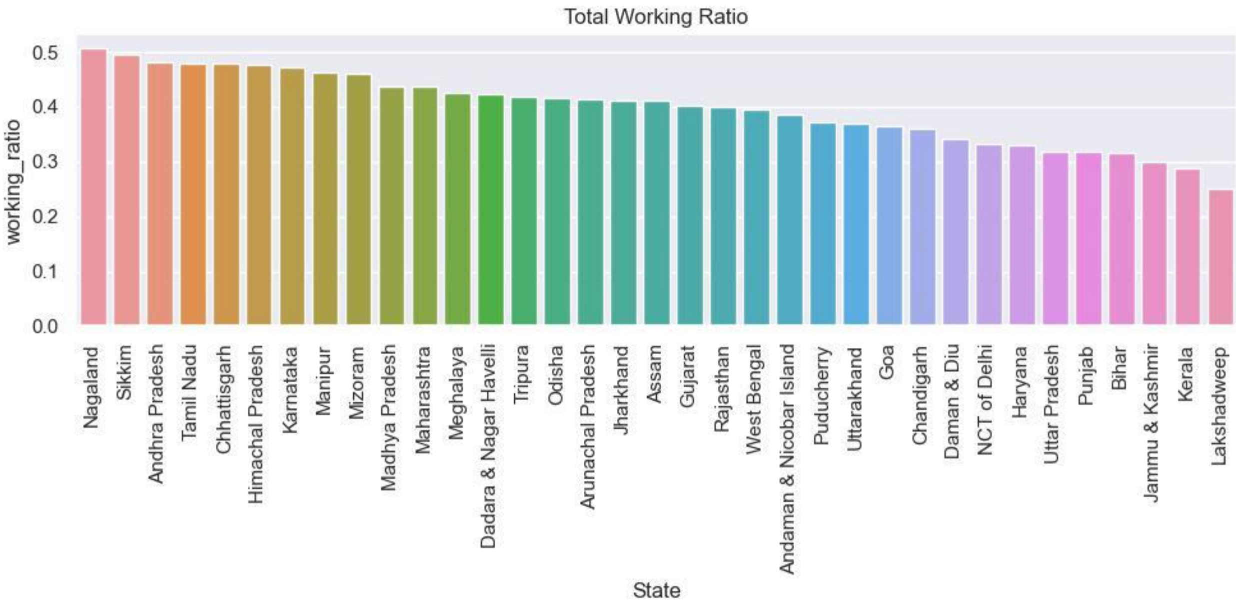
State	TOT_M	TOT_F	M_LIT	F_LIT	TOTAL_POP	TOTAL_LIT	Literacy_ratio
Mizoram	59534	95463	48512	79412	154997	127924	0.825332
Kerala	2919825	4856357	2370331	3878204	7776182	6248535	0.803548
Lakshadweep	12823	14772	10601	11334	27595	21935	0.794890
Goa	118979	191393	99381	139749	310372	239130	0.770463
Chandigarh	41753	59644	33552	43438	101397	76990	0.759293



(v) Which state has highest ratio of Working Population to Total Population (including Male and Female).

- Nagaland has the Highest Working ration (including Male and Female).

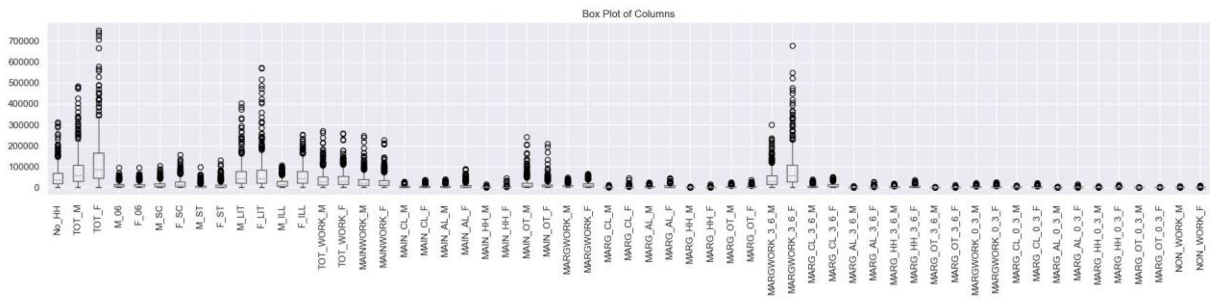
State	TOT_M	TOT_F	TOT_WORK_M	TOT_WORK_F	TOTAL_POP	TOTAL_WORK	working_ratio
Nagaland	73506	125935	30889	70104	199441	100993	0.506380
Sikkim	26664	41518	13608	20161	68182	33769	0.495277
Andhra Pradesh	3274363	6097235	1674517	2833719	9371598	4508236	0.481053
Tamil Nadu	3074009	5610310	1724274	2441679	8684319	4165953	0.479710
Chhattisgarh	838404	1526592	398935	732456	2364996	1131391	0.478390



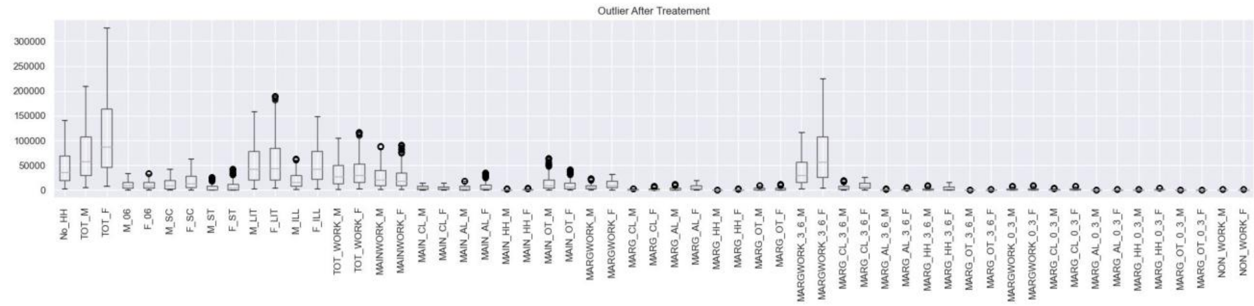
2.2 - Data Preprocessing

- No Missing Values
- Outliers:
    - There are few columns with Outliers.
    - Unsupervised learning is sensitive to Outliers.
    - Outliers are treated with 95^{th} and 5^{th} percentile value.
        - If Value > 95^{th} percentile then replace it with 95^{th} percentile
        - If value < 5^{th} percentile

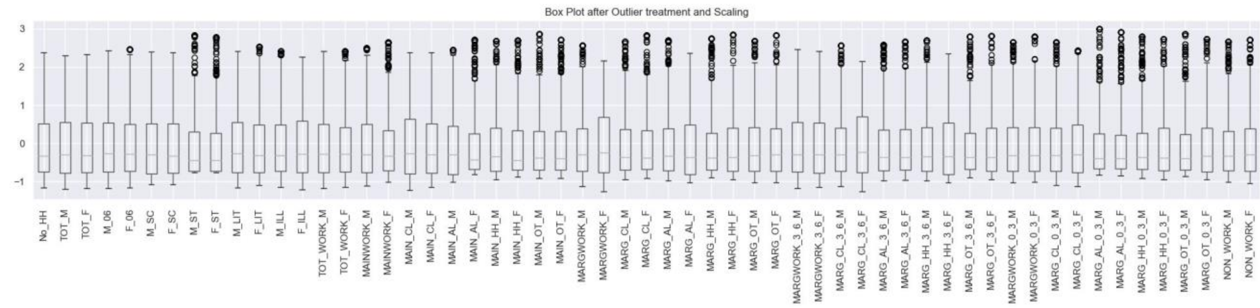
Before treating:



After treating:



After Scaling:



Outlier Treatment:

- PCA is sensitive to Extreme values because it involves calculating covariance or correlation Matrices. Outliers significantly influence these calculations, leading misleading principal components.
- Outliers can affect the interpretation of principal components.
- Outliers can inflate the eigen values associated with principal components.

Feature Scaling:

- PCA aims to maximize the variance of the data along principal components. If columns are on different scales, those with larger variance will dominate the principal components – which can result in neglecting the contribution of features with smaller variances.
- PCA relies on distance between data points. Features on different scale will contribute unequally to the distance calculation. Scaling ensures the distances are computed accurately and that principal components reflect the structure of the data.

Box Plot after feature scaling displays all features in the same scale.

2.3 PCA

KMO Test:

The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.

If MSA is less than 0.5, PCA is not recommended since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

Kmo_Model = 0.93

Inference:

- Given strong KMO value, the resulting factors or components from the analysis should be reliable and meaningful.
- KMO test measures the suitability of data for factor analysis.
- Since data is highly suitable, we can proceed with PCA to reduce dimensionality.

Bartletts Test of Sphericity:

Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.

- H0: All variables in the data are uncorrelated.
- Ha: At least one pair of variables in the data are correlated

If the null hypothesis cannot be rejected, then PCA is not advisable.

If the p-value is small, then we can reject the null hypothesis and agree that there is at least one pair of variables in the data which are correlated hence PCA is recommended.

P_value = 0

2.3.1 Create the Covariance Matrix:

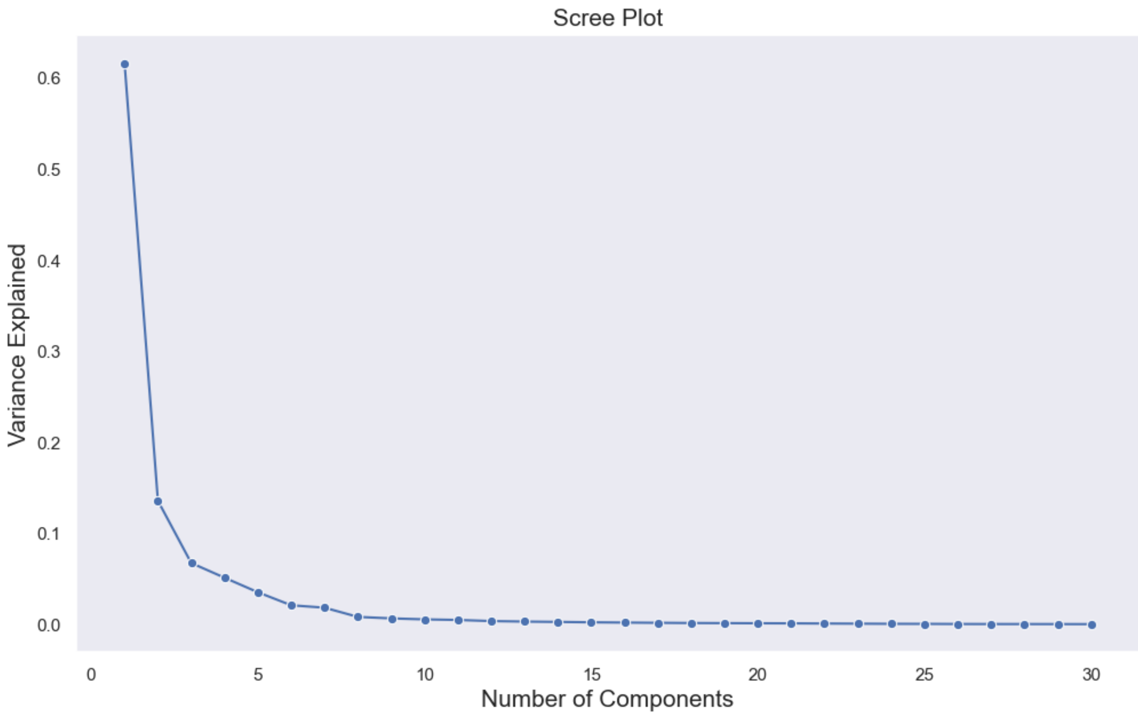
- Initially considering 30 components/dimensions.

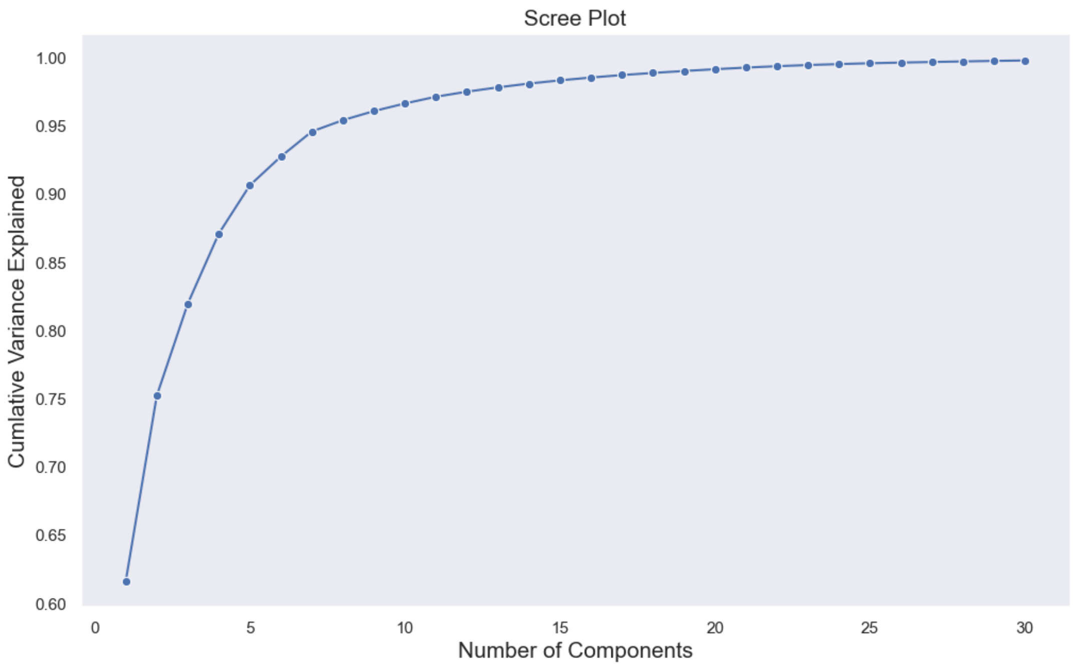
Cumulative Variance Explained in Percentage: [61.65 75.24 81.99 87.13 90.66 92.77 94.61 95.44 96.11 96.66 97.16 97.53

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]

- Calculating the Cumulative sum of the Eigen Values
    - 6 principal components cover 92.77% oof the Variance.
    - Dataset can be effectively reduced in dimensionality without losing significant information.
    - Remaining components account approx. 7.23% of total variance, this relatively small amount might represent noise, less significant patterns.
    - Representing high-dimensional data in lower-dimensional space can help visualize complex patterns.





Re-model with 6 Components

Covariance Matrix:

```
array([[-5.47, -5.49, -7.25, ..., -7.47, -7.56, -7.17],
       [ 0.36, -0.06, -0.18, ..., -0.8 , -0.84, -1.17],
       [-1.49, -1.93, -0.43, ..., -0.98, -0.94, -0.98],
       [-1.13, -1.55, -0.11, ..., -1.03, -0.78, -0.63],
       [ 0.37,  0.01,  0.56, ...,  0.13,  0.01,  0.15],
       [-0.49,  0.92,  0.17, ...,  0.11, -0.26, -0.39]])
```

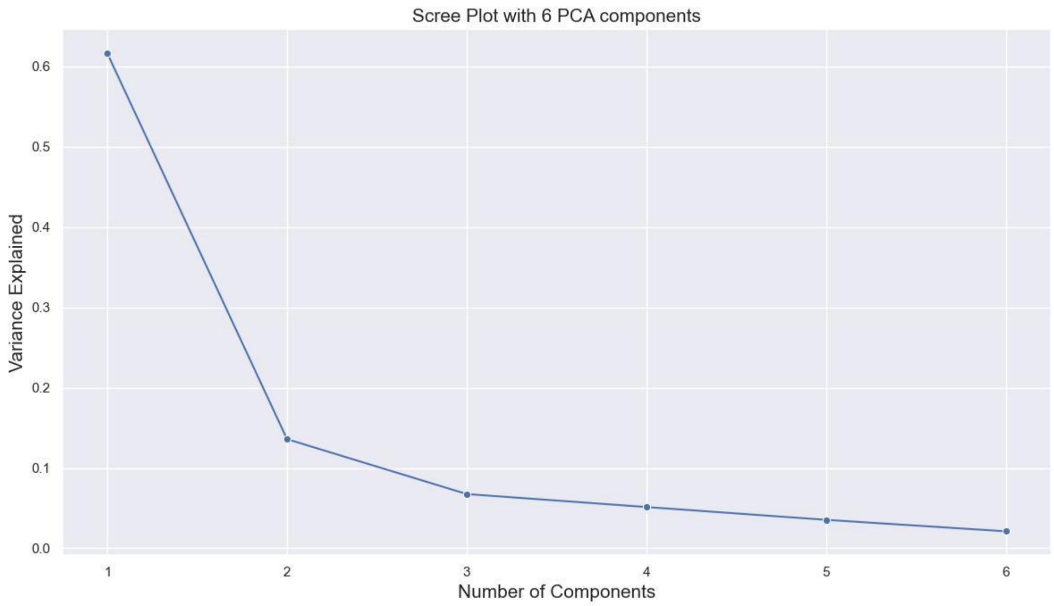
Eigen Values:

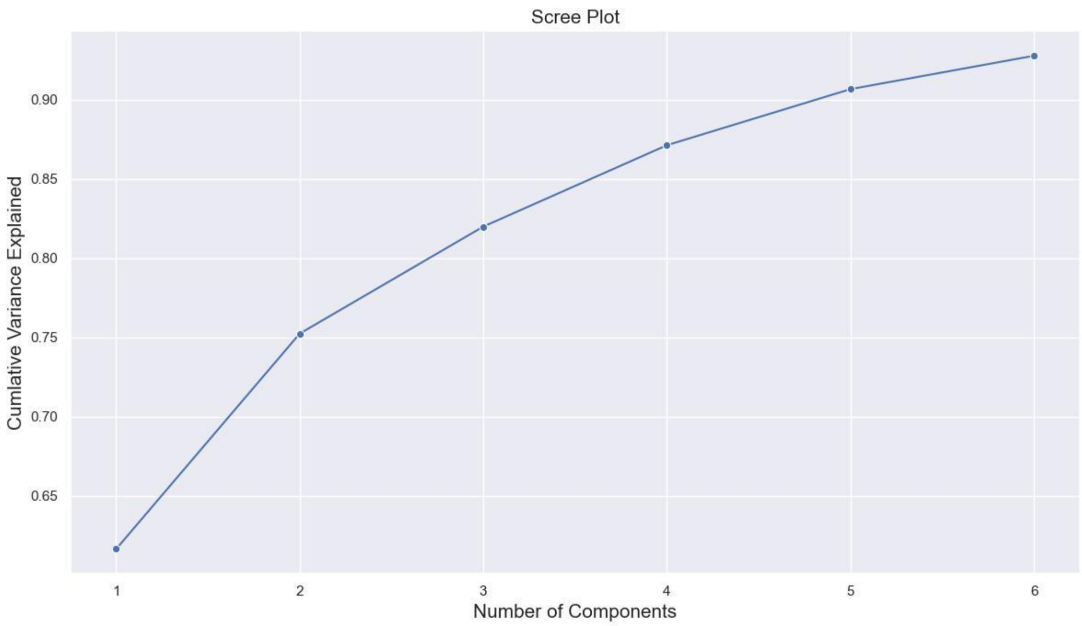
Array ([35.19645077, 7.75864164, 3.85313758, 2.93088251, 2.01945117, 1.20283006])

PC1: 35.19
PC2: 7.75
PC3: 3.85
PC4: 2.93
PC5: 2.01
PC6: 1.20

Maximum Variance explained by PC1 35.19

Cumulative Variance Explained in Percentage: [61.65 75.24 81.99 87.13 90.66 92.77]





<u>Percentage of variance explained by each PC:</u>

- Variance Explained in Percentage: [0.62 0.14 0.07 0.05 0.04 0.02]
- 62% of total variance explained by PC1.
- 14% of total variance explained by PC2.
- 7% of total variance explained by PC3.
- 5% of total variance explained by PC4.
- 4% of total variance explained by PC5
- 2% of total variance explained by PC6.

Eigen Vectors:

	PC1	PC2	PC3	PC4	PC5	PC6
No_HH	0.150237	-0.115287	0.103180	0.074622	-0.015401	-0.064114
TOT_M	0.160522	-0.076787	-0.029706	0.050540	-0.054964	-0.076640
TOT_F	0.159558	-0.091114	0.034324	0.067355	-0.032032	-0.072525
M_06	0.157613	-0.017730	-0.065176	0.027072	-0.081270	-0.108668
F_06	0.157938	-0.012012	-0.060278	0.014851	-0.076328	-0.100658
M_SC	0.144513	-0.075767	-0.031688	0.007019	-0.176904	-0.056666
F_SC	0.144668	-0.083882	0.024432	0.012197	-0.165584	-0.048954
M_ST	0.020579	0.057669	0.304044	0.080970	0.429280	0.199007
F_ST	0.020071	0.056227	0.318858	0.070049	0.429934	0.183270
M_LIT	0.156657	-0.102100	-0.028268	0.087068	-0.026587	-0.075214
F_LIT	0.146500	-0.131014	-0.009020	0.125210	0.023254	-0.085913
M_ILL	0.155355	-0.008298	-0.038494	-0.036156	-0.100376	-0.065371
F_ILL	0.159196	-0.021189	0.088247	-0.018157	-0.109074	-0.017846
TOT_WORK_M	0.155016	-0.119898	0.002007	0.066352	-0.030359	-0.039786
TOT_WORK_F	0.142567	-0.080485	0.196666	0.102831	-0.015676	0.046374
MAINWORK_M	0.142307	-0.167457	0.021732	0.097093	-0.047882	-0.024588
MAINWORK_F	0.122715	-0.150881	0.215881	0.122300	-0.053621	0.084241
MAIN_CL_M	0.110994	0.045027	0.053099	0.052491	-0.298306	0.227653
MAIN_CL_F	0.082170	0.095243	0.209970	0.238135	-0.248924	0.257364
MAIN_AL_M	0.118710	-0.054905	0.231973	-0.134189	-0.228158	-0.002503
MAIN_AL_F	0.085496	-0.087595	0.363843	-0.024065	-0.180063	0.061499
MAIN_HH_M	0.142096	-0.098005	-0.106057	-0.021517	-0.069997	0.156218
MAIN_HH_F	0.131505	-0.115298	0.019318	-0.045637	-0.027614	0.361935
MAIN_OT_M	0.120682	-0.208063	-0.048392	0.153090	0.082603	-0.077461
MAIN_OT_F	0.115319	-0.211979	0.054925	0.157541	0.111242	-0.029449
MARGWORK_M	0.157409	0.080308	-0.071343	-0.072740	0.070861	-0.087868
MARGWORK_F	0.149269	0.105613	0.113688	0.019583	0.079530	-0.077898
MARG_CL_M	0.087209	0.273126	-0.087794	0.163200	-0.022243	0.036596
MARG_CL_F	0.061758	0.271822	-0.021352	0.294797	-0.056268	0.042011
MARG_AL_M	0.128042	0.157085	0.058896	-0.247792	-0.031582	-0.099072
MARG_AL_F	0.115583	0.129719	0.261036	-0.161233	0.012909	-0.112396
MARG_HH_M	0.144243	0.054458	-0.153784	-0.165794	-0.002601	0.151729
MARG_HH_F	0.141142	0.008095	-0.093288	-0.147843	0.039335	0.348818
MARG_OT_M	0.150881	-0.075929	-0.140024	0.030448	0.136228	-0.027513
MARG_OT_F	0.146784	-0.097861	-0.068724	0.066070	0.190008	-0.011439
MARGWORK_3_6_M	0.159369	-0.041186	-0.058535	0.038466	-0.066057	-0.111087
MARGWORK_3_6_F	0.157407	-0.088268	-0.054826	0.045982	-0.033444	-0.116369
MARG_CL_3_6_M	0.158435	0.066959	-0.064243	-0.085584	0.064360	-0.082874
MARG_CL_3_6_F	0.149881	0.087303	0.133955	0.020142	0.066179	-0.055292
MARG_AL_3_6_M	0.094111	0.263357	-0.080724	0.134875	-0.019035	0.040684
MARG_AL_3_6_F	0.064124	0.263713	-0.001050	0.296008	-0.059771	0.060172
MARG_HH_3_6_M	0.128741	0.149872	0.069237	-0.249295	-0.039978	-0.094177

MARG_HH_3_6_F	0.113282	0.115664	0.284826	-0.152415	0.001375	-0.093203
MARG_OT_3_6_M	0.144083	0.050806	-0.152920	-0.164305	-0.003943	0.157772
MARG_OT_3_6_F	0.140007	-0.001906	-0.089153	-0.140497	0.036863	0.368728
MARGWORK_0_3_M	0.150922	-0.080127	-0.138536	0.029343	0.124310	-0.026712
MARGWORK_0_3_F	0.146724	-0.108280	-0.071731	0.063933	0.166101	-0.004758
MARG_CL_0_3_M	0.143658	0.139366	-0.102500	-0.014628	0.091696	-0.104014
MARG_CL_0_3_F	0.134757	0.166220	0.035642	0.013616	0.117067	-0.145761
MARG_AL_0_3_M	0.062955	0.275625	-0.102208	0.222787	-0.029411	0.013711
MARG_AL_0_3_F	0.054616	0.280543	-0.073987	0.261182	-0.047490	-0.005591
MARG_HH_0_3_M	0.120330	0.184515	0.005010	-0.232525	0.018822	-0.113667
MARG_HH_0_3_F	0.114088	0.175477	0.151726	-0.187199	0.064907	-0.177600
MARG_OT_0_3_M	0.140928	0.066885	-0.156288	-0.164833	0.005318	0.134923
MARG_OT_0_3_F	0.141480	0.037008	-0.103485	-0.162535	0.043115	0.261812
NON_WORK_M	0.147636	-0.050706	-0.137645	0.035397	0.188305	-0.033770
NON_WORK_F	0.140864	-0.045885	-0.042492	0.063011	0.249229	-0.032293

2.3.2 Compare PCs with Actual Columns and identify which is explaining most variance: (considering highlighted fields in the plot)

PC1:

No_HH	0.150237
TOT_M	0.160522
TOT_F	0.159558
M_06	0.157613
F_06	0.157938
M_LIT	0.156657
F_LIT	0.146500
M_ILL	0.155355
F_ILL	0.159196
TOT_WORK_M	0.155016
MARGWORK_M	0.157409
MARGWORK_F	0.149269
MARG_OT_M	0.150881
MARGWORK_3_6_M	0.159369
MARGWORK_3_6_F	0.157407
MARG_CL_3_6_M	0.158435
MARG_CL_3_6_F	0.149881
MARGWORK_0_3_M	0.150922
MARG_CL_0_3_M	0.143658

- Total population of Male and Female
- Male and Female Illiteracy and illiteracy.

PC1 can be interpreted as capturing a mix of demographic attributes, educational levels, and economic activities, providing a holistic overview of the socioeconomic landscape represented by dataset. It represents population size, gender distribution, educational levels, age group of population between 0-6 Male and Female.
It also represents Marginal Cultivator Male and Female for 3 to 6 Months and Marginal Worker Population Male and Female.

PC2:

MAINWORK_M	-0.167457
MAIN_OT_M	-0.208063
MAIN_OT_F	-0.211979
MARG_CL_M	0.273126
MARG_AL_3_6_M	0.263357
MARG_CL_0_3_F	0.166220
MARG_AL_0_3_M	0.275625
MARG_AL_0_3_F	0.280543

- Negative loading for Main work Male indicating inverse relationship with this component.
- Marginal Agriculture Labor 0 to 3 months Male and Female
- Marginal Agriculture Labor Male 3 to 6 months.

PC2 represent a contrast between engagement in main work activities and engagement in Marginal work, particularly cultivation-related activities. Entities with higher involvement in marginal work, especially for short durations, contribute positively to PC2. It captures variations in engagement in different types of economic activities, particularly main work, and Marginal work. Provides insights into economic diversity and labor market dynamics.

PC3:

TOT_WORK_F	0.196666
MAINWORK_F	0.215881
MAIN_AL_M	0.231973
MAIN_AL_F	0.363843
MARG_AL_F	0.261036
MARG_HH_3_6_F	0.284826
MARG_OT_0_3_M	-0.156288

- Main Agriculture population Female & Male.
- Main and total workforce Female
- Main Agriculture Labor and Marginal Agriculture Labor Female.

PC3 represent combination of factors related to work engagement for Main work and Marginal Agriculture labor Female. Focuses on Total and Main work for Females. It captures variations in work engagement, particularly females, and highlights the importance of both main work and marginal work activities shaping workforce dynamics within dataset. It provides valuable insights in to gender specific work patterns.

PC4:

MARG_CL_F	0.294797
MARG_AL_M	-0.247792
MARG_HH_M	-0.165794
MARG_AL_3_6_F	0.296008
MARG_HH_3_6_M	-0.249295
MARG_OT_3_6_M	-0.164305
MARG_HH_0_3_M	-0.232525
MARG_HH_0_3_F	-0.187199

- Marginal Agriculture Labor for 3 to 6 Months Female.
- Marginal Cultivator Female.

PC4 represent engagement in Marginal Cultivation work Female and engagement in other types of Marginal work activities largely Male. It Captures variations in engagement in different types of Marginal work activities, particularly focusing on contract between engagement in marginal cultivation work by females and engagement in other types of Marginal work activities, particularly by males and within Households. It provides valuable insights into the diversity of economic activities within the dataset, facilitating further analysis and decision-making in various domains such as labor economics, gender studies, and public policy.

PC5:

M_SC	-0.176904
F_SC	-0.165584
M_ST	0.429280
F_ST	0.429934

MAIN_CL_M	-0.298306
MARG_OT_F	0.190008
MARGWORK_0_3_F	0.166101
NON_WORK_M	0.188305
NON_WORK_F	0.249229

- Population Male and Female in Scheduled Tribe
- Nonworking Male and Female

PC5 represents contrast between Scheduled Castes and tribes with Scheduled Tribes contributing positively and Scheduled Castes contributing negatively. It Captures Non-Working Population both Male and Female. Focuses on contrast between Scheduled Castes and Scheduled Tribes, gender specific work patterns and the size of the non-working population. It Provides valuable insights into socio-economic disparities and labor market dynamics within the dataset.

PC6:

MAIN_CL_F	0.257364
MAIN_HH_M	0.156218
MAIN_HH_F	0.361935
MARG_HH_F	0.348818
MARG_OT_3_6_F	0.368728
MARG_OT_0_3_F	0.261812

- Main Household Male and Female Population
- Marginal Others Female for 0 to 6 Months

PC6 represent a combination of factors related to household work and Marginal activities, particularly focusing on the engagement of females in these activities. Entities with higher engagement in main cultivation work by females, household work by both males and females, and marginal household work and other types of marginal work by females contribute positively to PC6. It captures variations in engagement in household work and marginal work activities, particularly focusing on the roles of females within households. It provides valuable insights into gender-specific work patterns and household-level economic activities within the dataset, facilitating further analysis and decision-making in various domains such as gender studies, labor economics, and public policy.

2.3.3 Write Linear Equation for first PC

Linear Equation =

0.15 * No_HH + 0.16 * TOT_M + 0.16 * TOT_F + 0.16 * M_06 + 0.16 * F_06 + 0.14 * M_SC + 0.14 * F_SC + 0.02 * M_ST + 0.02 * F_ST + 0.16 * M_LIT + 0.15 * F_LIT + 0.16 * M_ILL + 0.16 * F_ILL + 0.16 * TOT_WORK_M + 0.14 * TOT_WORK_F + 0.14 * MAINWORK_M + 0.12 * MAINWORK_F + 0.11 * MAIN_CL_M + 0.08 * MAIN_CL_F + 0.12 * MAIN_AL_M + 0.09 * MAIN_AL_F + 0.14 * MAIN_HH_M + 0.13 * MAIN_HH_F + 0.12 * MAIN_OT_M + 0.12 * MAIN_OT_F + 0.16 * MARGWORK_M + 0.15 * MARGWORK_F + 0.09 * MARG_CL_M + 0.06 * MARG_CL_F + 0.13 * MARG_AL_M + 0.12 * MARG_AL_F + 0.14 * MARG_HH_M + 0.14 * MARG_HH_F + 0.15 * MARG_OT_M + 0.15 * MARG_OT_F + 0.16 * MARGWORK_3_6_M + 0.16 * MARGWORK_3_6_F + 0.16 * MARG_CL_3_6_M + 0.15 * MARG_CL_3_6_F + 0.09 * MARG_AL_3_6_M + 0.06 * MARG_AL_3_6_F + 0.13 * MARG_HH_3_6_M + 0.11 * MARG_HH_3_6_F + 0.14 * MARG_OT_3_6_M + 0.14 * MARG_OT_3_6_F + 0.15 * MARGWORK_0_3_M + 0.15 * MARGWORK_0_3_F + 0.14 * MARG_CL_0_3_M + 0.13 * MARG_CL_0_3_F + 0.06 * MARG_AL_0_3_M + 0.05 * MARG_AL_0_3_F + 0.12 * MARG_HH_0_3_M + 0.11 * MARG_HH_0_3_F + 0.14 * MARG_OT_0_3_M + 0.14 * MARG_OT_0_3_F + 0.15 * NON_WORK_M + 0.14 * NON_WORK_F