MACHINE LEARNING PROJECT 1-BUSINESS REPORT

(By Sowmya Subramaniam -PGPDSBA.O.SEP23.B)

PART 1

Clustering:

Digital Ads Data:

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

- Top 5 rows of the Data Frame

	0	1	2	3	4
Timestamp	2020-9-2-17	2020-9-2-18	2020-9-3-16	2020-9-3-2	2020-9-3-13
InventoryType	Format1	Format1	Format6	Format1	Format1
Ad - Length	300	300	336	300	300
Ad- Width	250	250	250	250	250
Ad Size	75000	75000	84000	75000	75000
Ad Type	Inter222	Inter223	Inter217	Inter224	Inter225
Platform	Video	Web	Web	Web	Video
Device Type	Desktop	Mobile	Desktop	Desktop	Mobile
Format	Display	Display	Video	Display	Display
Available_Impressions	1806	1979	1566	643	1550
Matched_Queries	325	384	298	103	347
Impressions	323	380	297	102	345
Clicks	1	0	0	0	0
Spend	0.0	0.0	0.0	0.0	0.0
Fee	0.35	0.35	0.35	0.35	0.35
Revenue	0.0	0.0	0.0	0.0	0.0
CTR	0.0031	0.0	0.0	0.0	0.0
CPM	0.0	0.0	0.0	0.0	0.0
CPC	0.0	NaN	NaN	NaN	NaN

- Bottom 5 rows of the Data Frame

	25852	25853	25854	25855	25856
Timestamp	2020-10-1-5	2020-11-18-2	2020-9-14-0	2020-9-30-4	2020-10-17-3
InventoryType	Format5	Format4	Format5	Format7	Format5
Ad - Length	720	120	720	300	720
Ad- Width	300	600	300	600	300
Ad Size	216000	72000	216000	180000	216000
Ad Type	Inter222	inter230	Inter221	Inter228	Inter225
Platform	Video	Video	App	Video	Video
Device Type	Desktop	Mobile	Mobile	Mobile	Mobile
Format	Video	Video	Video	Display	Display
Available_Impressions	1	7	2	1	1
Matched_Queries	1	1	2	1	1
Impressions	1	1	2	1	1
Clicks	0	1	1	0	0
Spend	0.01	0.07	0.09	0.01	0.01
Fee	0.35	0.35	0.35	0.35	0.35
Revenue	0.0065	0.0455	0.0585	0.0065	0.0065
CTR	NaN	NaN	NaN	NaN	NaN
CPM	NaN	NaN	NaN	NaN	NaN
CPC	NaN	NaN	NaN	NaN	NaN

- Basic info about the Data frame:

```
Data columns (total 19 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Timestamp               25857 non-null  object
 1   InventoryType           25857 non-null  object
 2   Ad - Length             25857 non-null  int64
 3   Ad- Width               25857 non-null  int64
 4   Ad Size                 25857 non-null  int64
 5   Ad Type                 25857 non-null  object
 6   Platform                25857 non-null  object
 7   Device Type             25857 non-null  object
 8   Format                  25857 non-null  object
 9   Available_Impressions   25857 non-null  int64
 10  Matched_Queries         25857 non-null  int64
 11  Impressions             25857 non-null  int64
 12  Clicks                  25857 non-null  int64
 13  Spend                   25857 non-null  float64
 14  Fee                     25857 non-null  float64
 15  Revenue                 25857 non-null  float64
 16  CTR                     19392 non-null  float64
 17  CPM                     19392 non-null  float64
 18  CPC                     18330 non-null  float64
dtypes: float64(6), int64(7), object(6)
```

- Data frame Summary

	count	mean	std	min	25%	50%	75%	max
Ad - Length	25857.0	3.904312e+02	2.306961e+02	120.00	120.0000	300.0000	7.200000e+02	728.00
Ad- Width	25857.0	3.321828e+02	1.942609e+02	70.00	250.0000	300.0000	6.000000e+02	600.00
Ad Size	25857.0	9.968328e+04	6.264069e+04	33600.00	72000.0000	75000.0000	8.400000e+04	216000.00
Available_Impressions	25857.0	2.169621e+06	4.542680e+06	0.00	9133.0000	330968.0000	2.208484e+06	27592861.00
Matched_Queries	25857.0	1.155322e+06	2.407244e+06	0.00	5451.0000	189449.0000	1.008171e+06	14702025.00
Impressions	25857.0	1.107525e+06	2.326648e+06	0.00	2558.0000	162162.0000	9.496930e+05	14194774.00
Clicks	25857.0	9.525881e+03	1.672169e+04	0.00	305.0000	3457.0000	1.068100e+04	143049.00
Spend	25857.0	2.414473e+03	3.932835e+03	0.00	36.0300	1173.6600	2.692280e+03	26931.87
Fee	25857.0	3.367289e-01	3.053978e-02	0.21	0.3500	0.3500	3.500000e-01	0.35
Revenue	25857.0	1.716549e+03	2.993025e+03	0.00	23.4200	762.8800	1.749982e+03	21276.18
CTR	19392.0	6.962653e-02	7.497012e-02	0.00	0.0024	0.0077	1.283000e-01	1.00
CPM	19392.0	7.252900e+00	6.538314e+00	0.00	1.6300	3.0350	1.222000e+01	81.56
CPC	18330.0	3.510606e-01	3.433338e-01	0.00	0.0900	0.1600	5.700000e-01	7.26

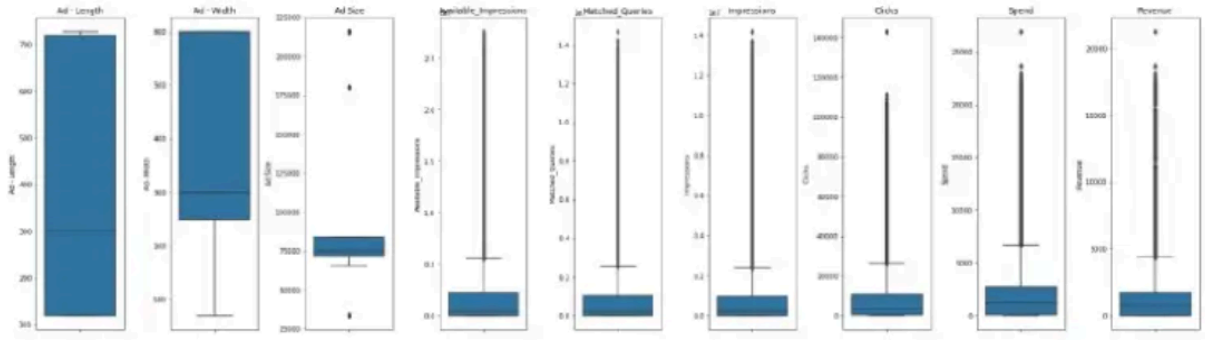
Timestamp	0
InventoryType	0
Ad - Length	0
Ad- Width	0
Ad Size	0
Ad Type	0
Platform	0
Device Type	0
Format	0
Available_Impressions	0
Matched_Queries	0
Impressions	0
Clicks	0
Spend	0
Fee	0
Revenue	0
CTR	6465
CPM	6465
CPC	7527

- Data frame duplicate value check:
  Total duplicate values: 0
  Treat missing values in CPC, CTR and CPM using the formula given

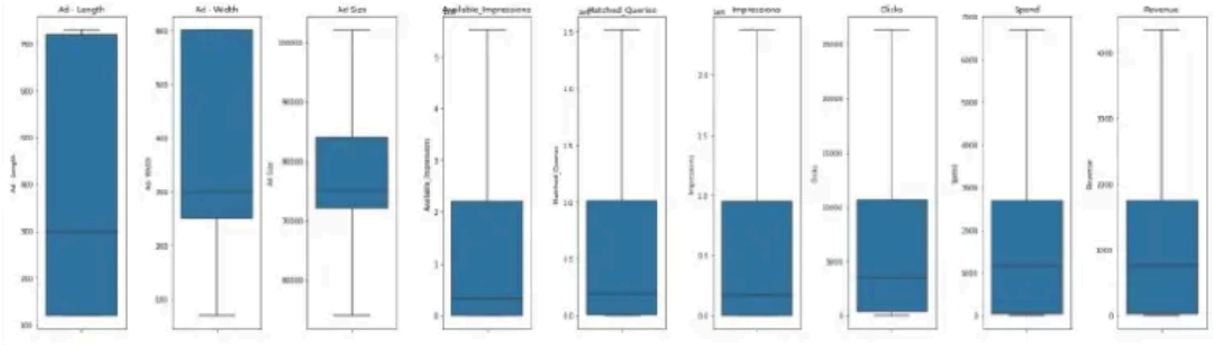
```
Timestamp                0
InventoryType            0
Ad - Length              0
Ad- Width                0
Ad Size                  0
Ad Type                  0
Platform                 0
Device Type              0
Format                   0
Available_Impressions    0
Matched_Queries          0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

There are no outliers in data_cluster

- Check if there are any outliers. Do you think treating outliers is necessary for K-
  Means clustering? Based on your judgment decide whether to treat outliers and if yes, which method to employ.



After treating outliers



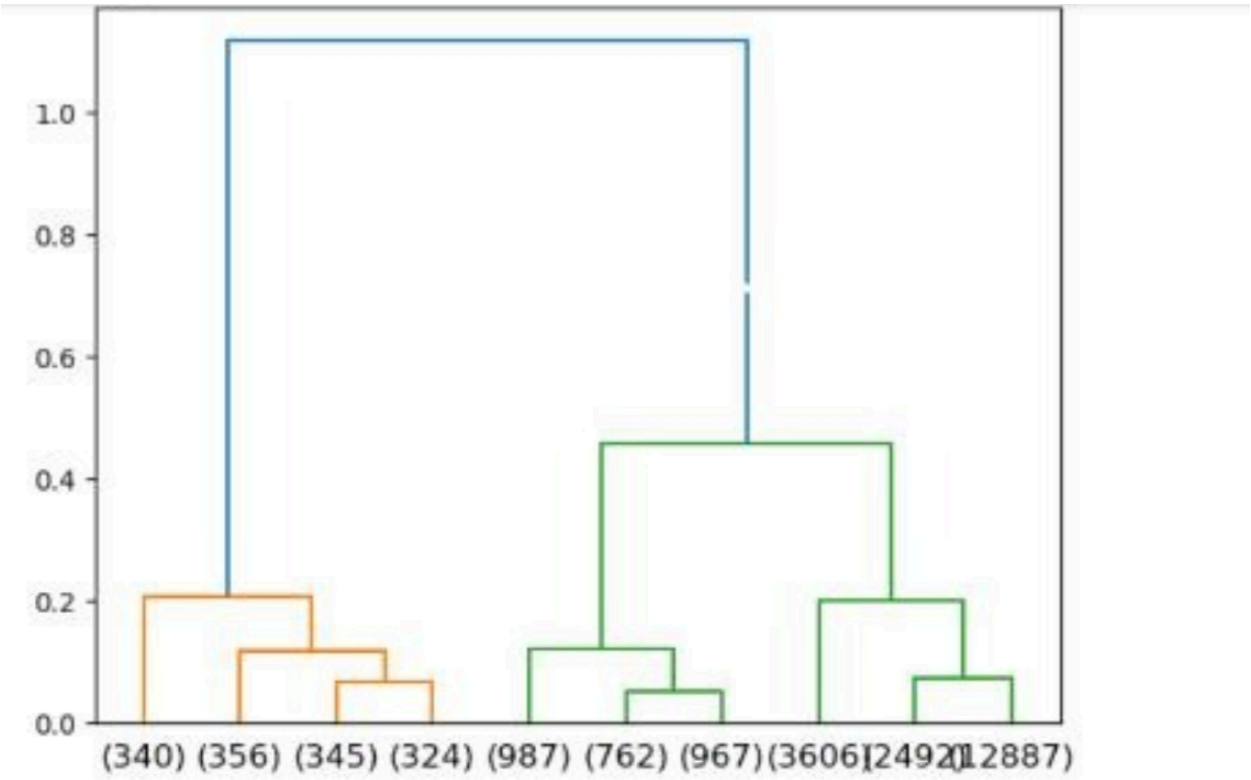
- Yes we need to treat outliers as k means clustering is sensitive to outliers.

- 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	Revenue
0	-0.392000	-0.423062	-0.161806	-0.714953	-0.744816	-0.735050	-0.821889	-0.844382	-0.841307
1	-0.392000	-0.423062	-0.161806	-0.714862	-0.744749	-0.734983	-0.822006	-0.844382	-0.841307
2	-0.235948	-0.423062	0.424415	-0.715079	-0.744846	-0.735081	-0.822006	-0.844382	-0.841307
3	-0.392000	-0.423062	-0.161806	-0.715566	-0.745066	-0.735313	-0.822006	-0.844382	-0.841307
4	-0.392000	-0.423062	-0.161806	-0.715088	-0.744791	-0.735024	-0.822006	-0.844382	-0.841307

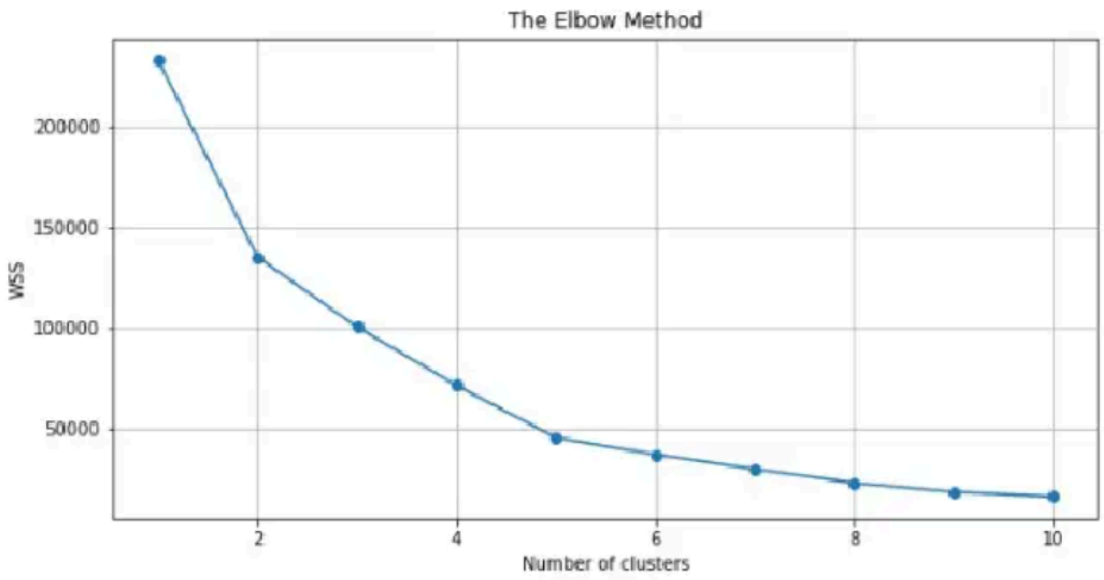
Without scaling the data , the algorithm may be biased towards a higher value

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



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

The WSS value for 1 cluster is 232713.00000000006
The WSS value for 2 clusters is 135274.9268314021
The WSS value for 3 clusters is 100590.2395311129
The WSS value for 4 clusters is 71656.59481682391
The WSS value for 5 clusters is 45771.31324276951
The WSS value for 6 clusters is 37438.815811017026
The WSS value for 7 clusters is 30149.7112338386
The WSS value for 8 clusters is 23382.874391416677
The WSS value for 9 clusters is 18790.993332464503
The WSS value for 10 clusters is 16544.499210561502



from the Dendrogram we can say optimum number of clusters: ' 5 '

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

The Silhouette scores for 2 clusters is 0.43093038125940913
The Silhouette scores for 3 clusters is 0.4169029019588384
The Silhouette scores for 4 clusters is 0.4859045662423113
The Silhouette scores for 5 clusters is 0.5484421685630947
The Silhouette scores for 6 clusters is 0.5554079926857388
The Silhouette scores for 7 clusters is 0.5882973964429631
The Silhouette scores for 8 clusters is 0.6005106775133303
The Silhouette scores for 9 clusters is 0.6298955511943023
The Silhouette scores for 10 clusters is 0.6296839903311501

from the Silhouette scores we can say optimum number of clusters: ' 5 '

- 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.]

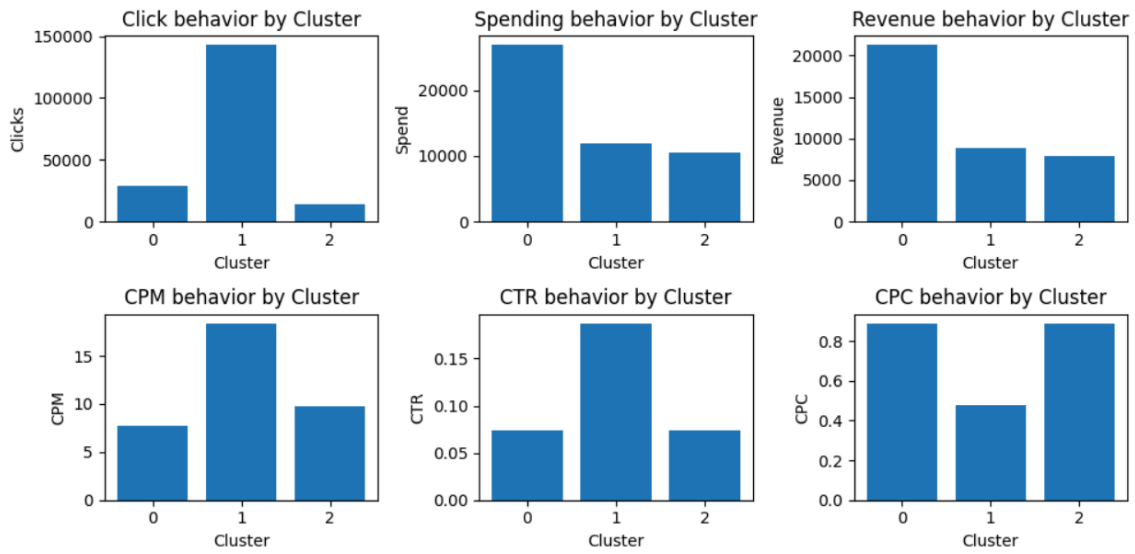
	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Revenue	k_means_cluster_5	sil_width
0	300.0	250.0	75000.0	1806.0	325.0	323.0	1.0	0.0	0.0	1	0.484134
1	300.0	250.0	75000.0	1979.0	384.0	380.0	0.0	0.0	0.0	1	0.484118
2	336.0	250.0	84000.0	1556.0	298.0	297.0	0.0	0.0	0.0	1	0.455931
3	300.0	250.0	75000.0	643.0	103.0	102.0	0.0	0.0	0.0	1	0.484184
4	300.0	250.0	75000.0	1550.0	347.0	345.0	0.0	0.0	0.0	1	0.484141
5	300.0	250.0	75000.0	2641.0	493.0	491.0	0.0	0.0	0.0	1	0.484063
6	300.0	250.0	75000.0	469.0	104.0	103.0	0.0	0.0	0.0	1	0.484184
7	300.0	250.0	75000.0	1244.0	154.0	153.0	0.0	0.0	0.0	1	0.484173
8	300.0	250.0	75000.0	1961.0	287.0	287.0	0.0	0.0	0.0	1	0.484134
9	300.0	250.0	75000.0	1670.0	223.0	223.0	0.0	0.0	0.0	1	0.484155

- **Total Count per cluster:**

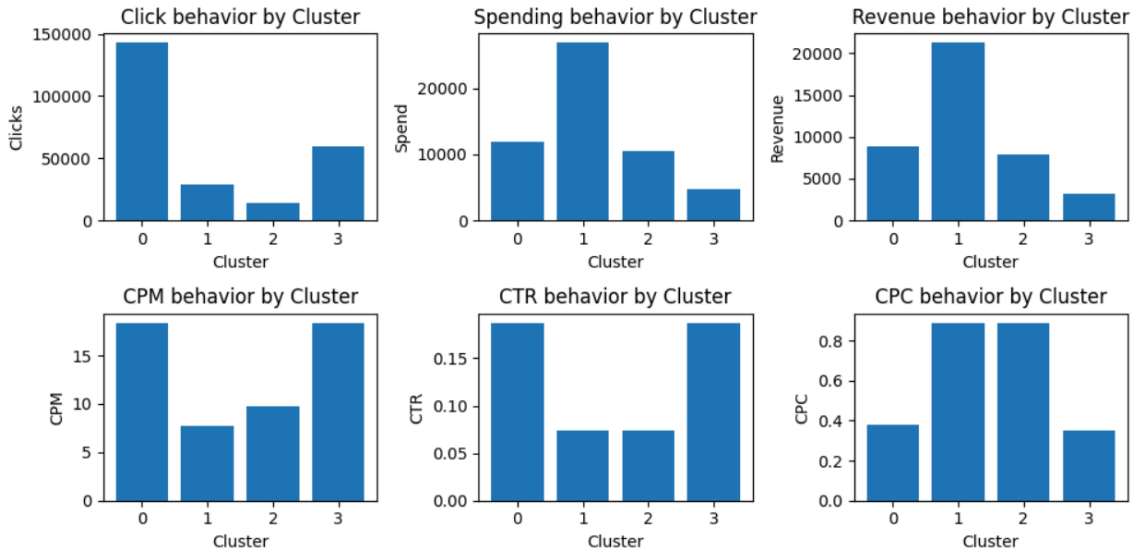
Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Revenue	sil_width	cluster count
460.364417	201.664679	73127.421894	5.161947e+06	2.450391e+06	2.324038e+06	11031.213212	5291.014889	3522.500595	0.585120	4193
190.589008	486.999124	75315.305452	7.423864e+04	3.971523e+04	3.467679e+04	1208.494964	166.682201	108.343391	0.625923	9134
442.612982	122.553151	61203.386642	1.939653e+06	9.174353e+05	8.754002e+05	3559.829163	1600.696988	1042.279585	0.501689	5315
693.438349	303.413000	101136.338946	2.174275e+05	1.177387e+05	1.002682e+05	11424.665037	1045.018470	680.529256	0.583809	5523
142.957447	572.281324	73925.531915	7.561640e+05	5.324346e+05	4.491000e+05	25720.598109	5734.283874	3785.384493	0.720601	1692

- **Cluster profiling**

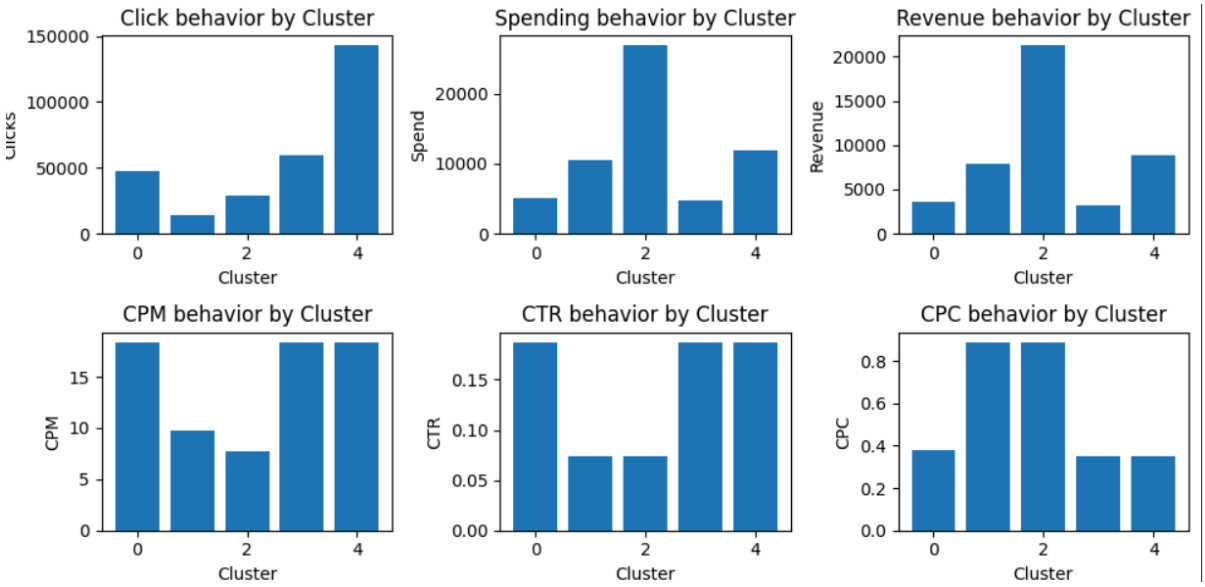
  k=3


Click behavior by Cluster
Clicks
150000
100000
50000
0
0
1
2
Cluster
Spending behavior by Cluster
Spend
20000
10000
0
0
1
2
Cluster
Revenue behavior by Cluster
Revenue
20000
15000
10000
5000
0
0
1
2
Cluster
CPM behavior by Cluster
CPM
15
10
5
0
0
1
2
Cluster
CTR behavior by Cluster
CTR
0.15
0.10
0.05
0.00
0
1
2
Cluster
CPC behavior by Cluster
CPC
0.8
0.6
0.4
0.2
0.0
0
1
2
Cluster

k=4


Click behavior by Cluster
Clicks
150000
100000
50000
0
0
1
2
3
Cluster
Spending behavior by Cluster
Spend
20000
10000
0
0
1
2
3
Cluster
Revenue behavior by Cluster
Revenue
20000
15000
10000
5000
0
0
1
2
3
Cluster
CPM behavior by Cluster
CPM
15
10
5
0
0
1
2
3
Cluster
CTR behavior by Cluster
CTR
0.15
0.10
0.05
0.00
0
1
2
3
Cluster
CPC behavior by Cluster
CPC
0.8
0.6
0.4
0.2
0.0
0
1
2
3
Cluster

k=5



Observations

1.The clusters 0 and 3 contain ads that have higher mean length than other clusters.

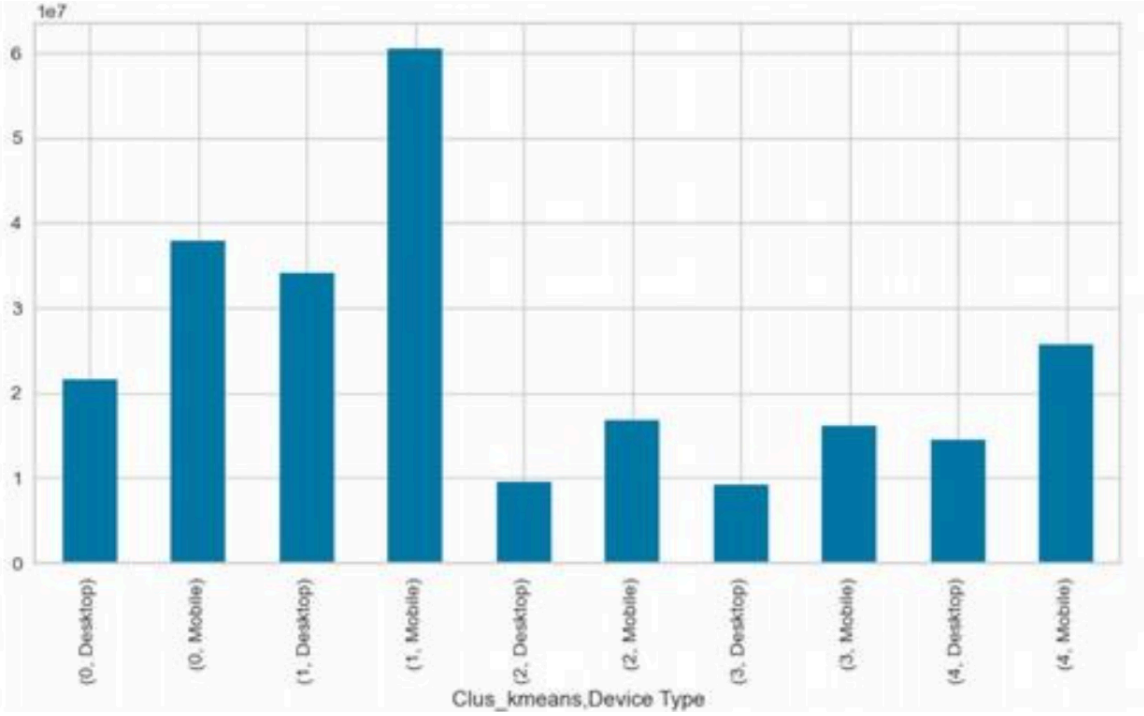
2.The clusters 1 and 2 have ads whose mean width is considerably more than the other clusters.

3.Cluster 4 has minimum ad size

4.There is not much difference in Fee, but cluster 3 has very high mean spend and mean revenue compared to the others

5.Available impression is maximum for cluster 3

- Comparison of Clusters according to device type (x-axis) and total clicks (y-axis)



Observations:

- The Mobile segment within Cluster 1 has the maximum number of clicks followed by Mobile segment within Cluster 0.
- Only for Cluster 3, desktop segment shows considerable number of clicks.
- Only for Cluster 1, desktop segment shows considerable number of clicks

SUMMARY

- The dataset consists of 25,857 rows and 19 columns.
- Missing values in the CPC, CTR, and CPM columns have been addressed using specified formulas through a user-defined function.
- Outliers have been identified in the variables.
- The dendrogram serves as a visualization, and linkage is employed to compute distances and merge clusters from n to 1.
- Ward's method is utilized to create linkage, and the linkage function is applied to the relevant columns in the data.
- The resulting linkage stores distances at which n clusters are sequentially merged into a single cluster.
- The transformed data frame is now stored in an array, allowing for the execution of the k-means algorithm.
- Before running the k-means algorithm, it is essential to determine the optimal number of clusters.
- The elbow plot, based on the within-cluster sum of squares (wss) values, is examined. Observations from the plot indicate a significant drop in wss values as we move from k=1 to k=2 and from k=2 to k=3, continuing until k=4. However, beyond k=4, the drop in values diminishes noticeably.
- In summary, the wss does not show a substantial reduction beyond 4 clusters, suggesting that 4 is the optimal number of clusters.

PART 2

PCA

- Read the data and perform basic checks like checking head, info, summary,nulls, and duplicates,

Checking top 5 rows:

	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_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180	237	680	252	32	
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123	229	186	148	76	
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6548	10904	1082	1018	3	...	114	188	44	89	3	34	0	
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	...	194	247	61	128	13	50	4	
4	1	5	Jammu & Kashmir	Punch	11654	20691	28981	5157	4587	20	...	874	1928	465	1043	205	302	24	

5 rows × 61 columns

PARS_OT_0_3_M	PARS_OT_0_3_F	NON_WORK_M	NON_WORK_F
32	46	258	214
70	178	140	160
0	4	67	61
4	10	116	59
24	105	180	478

- Shape of the dataset

  There are 640 rows and 61 columns

- Check the data types and duplicate values

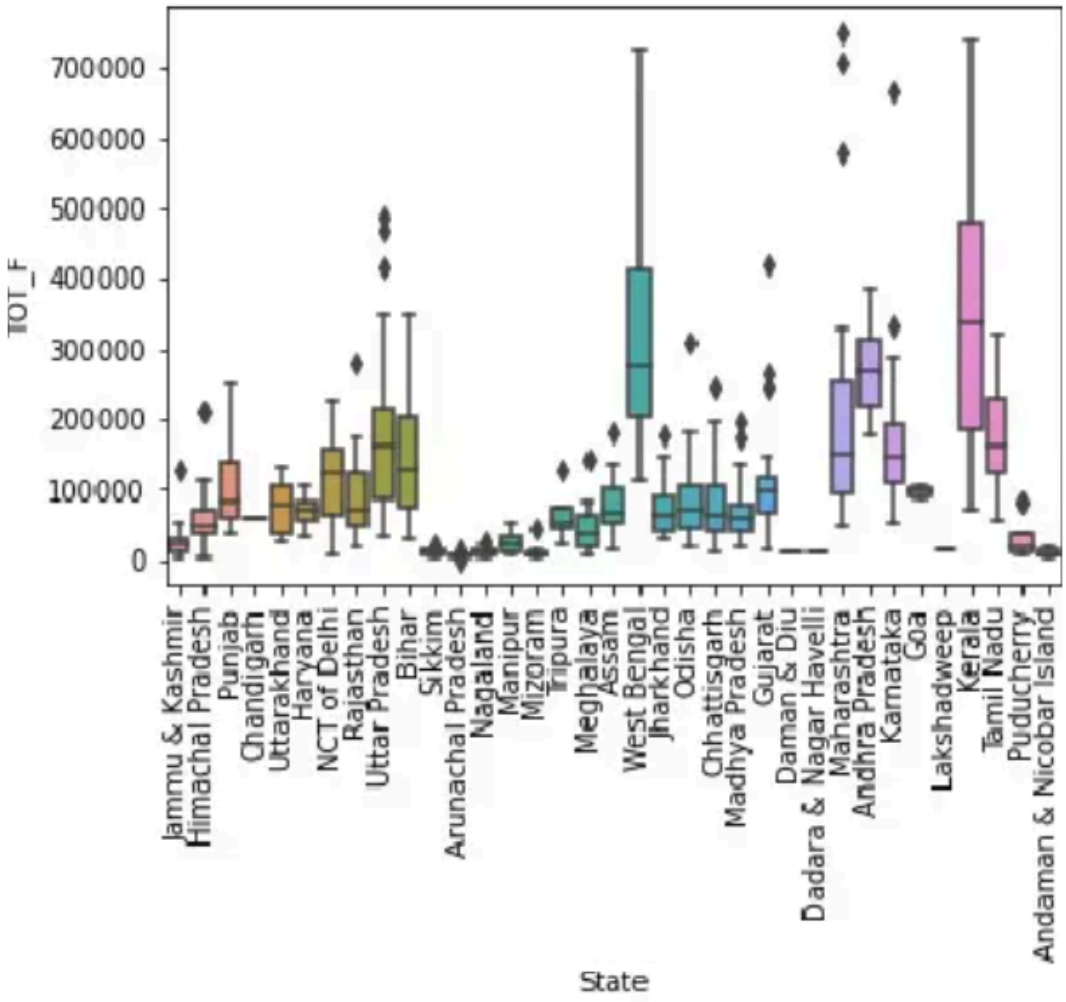
```
dtypes: int64(59), object(2)
memory usage: 305.1+ KB
```

59 of 61 columns are int data type and 2 columns are categorical object data type. And no null values.

- Exploratory analysis

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

II)State with the highest female population



III)State with the Highest male population

State	TOT_M	TOT_F	
Andaman & Nicobar Island	1549	2630	1
	5200	8012	1
Odisha	34009	62403	1
	37237	75437	1
	38026	86272	1
	39761	65519	1
	39932	82084	1
	41774	66198	1
	42947	69119	1
	54805	95812	1

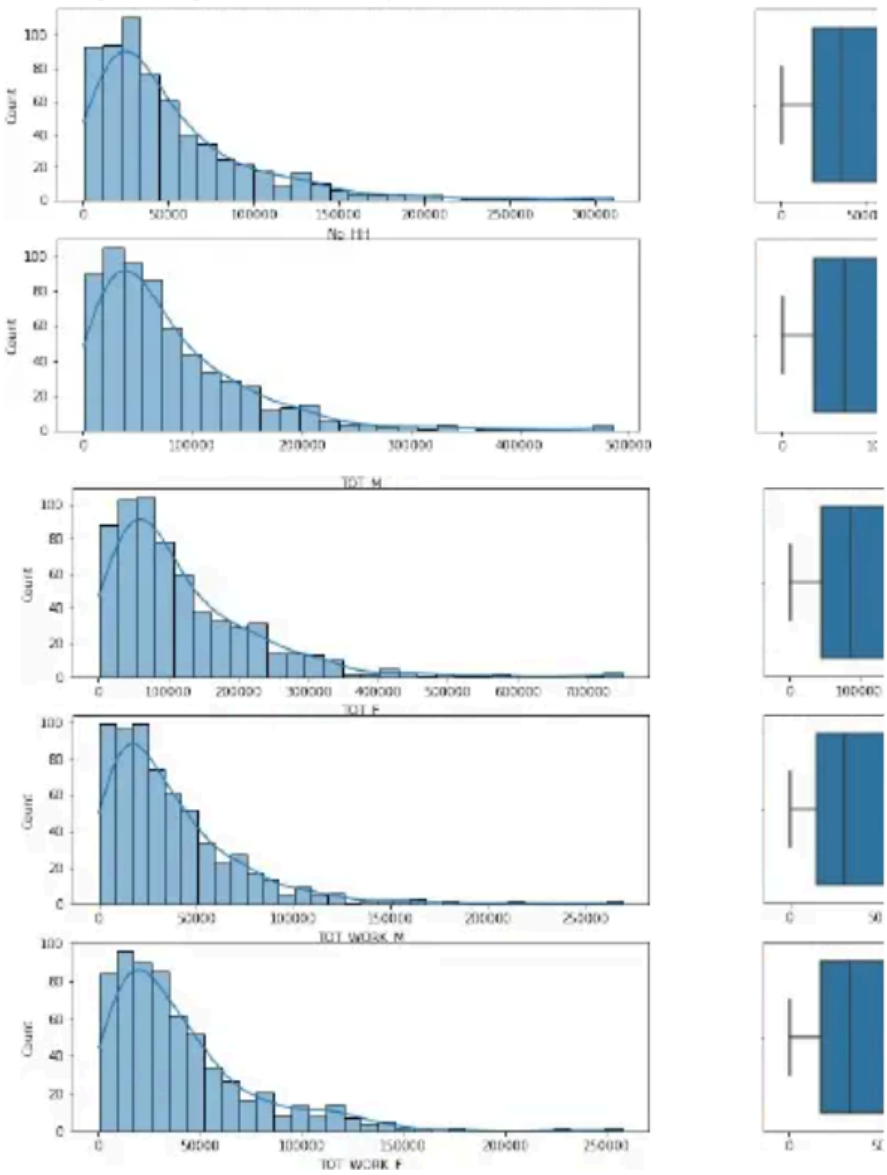
Name: TOT_F, dtype: int64

State	TOT_M	TOT_F	
Jammu & Kashmir	39014	52278	1
Jharkhand	27945	39517	1
Jammu & Kashmir	93269	128379	1
Jharkhand	18340	29128	1
	22652	36131	1
	22949	34664	1
	23764	36432	1
	24195	38510	1
	27527	36861	1
West Bengal	471482	725514	1

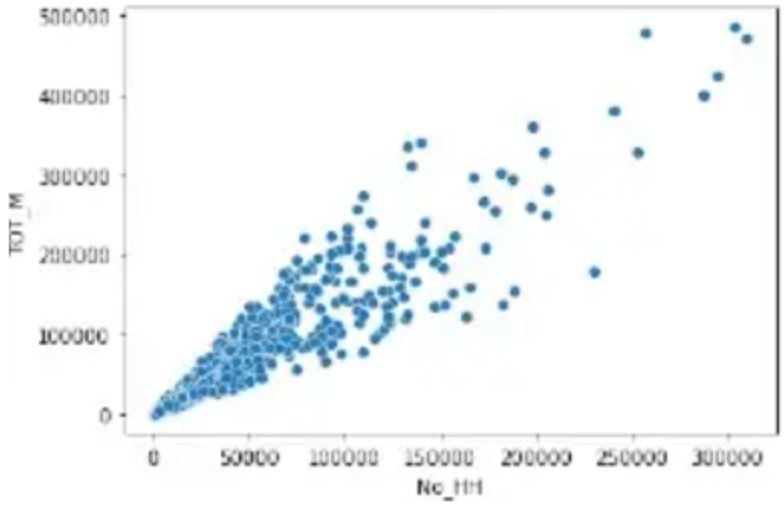
- **For EDA - Variables considered:**

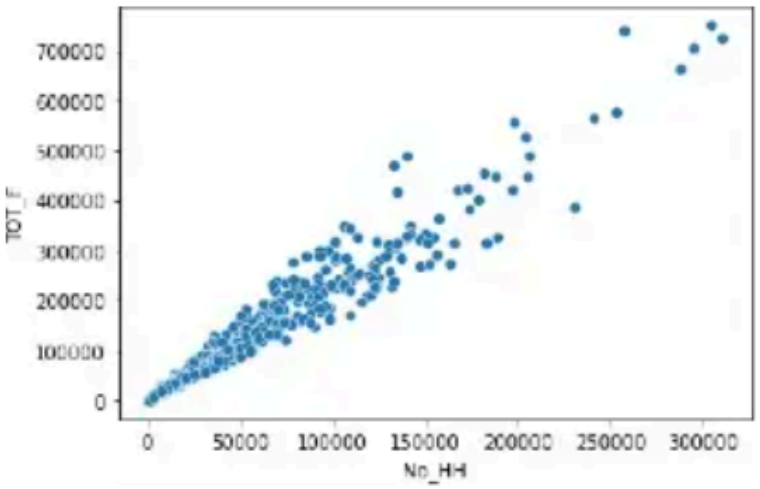
No_HH TOT_M TOT_F TOT_WORK_M TOT_WORK_F
No of Household
Total population Male
Total population Female
Total Worker Population Male
Total Worker Population Female

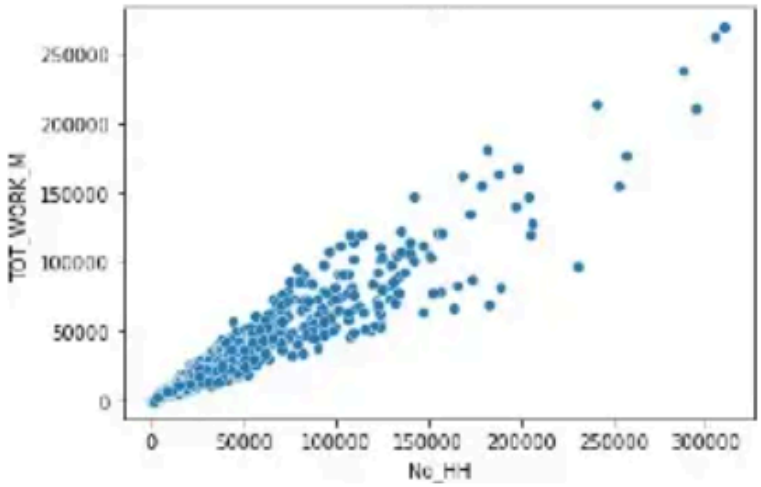
- **Univariate Analysis:**

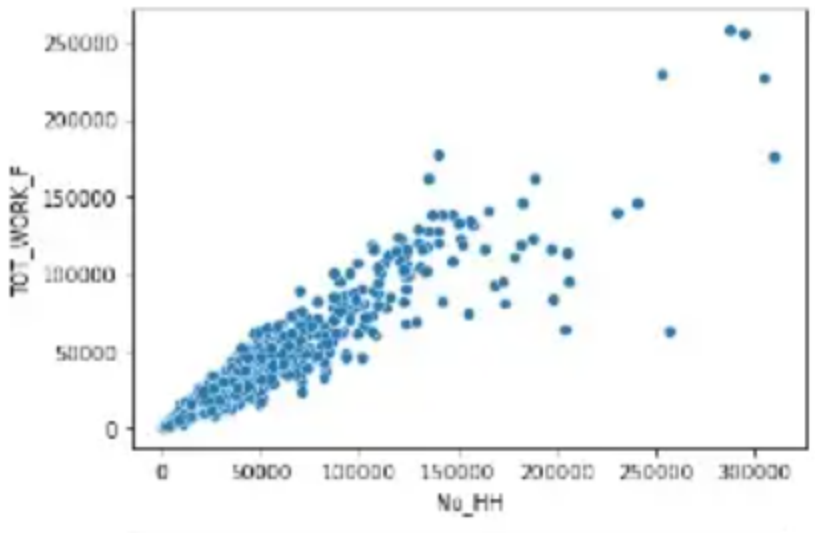


- **Bivariate Analysis**


500000
400000
300000
200000
100000
0
TOT_M
0
50000
100000
150000
200000
250000
300000
No_HH


700000
600000
500000
400000
300000
200000
100000
0
TOT_F
0
50000
100000
150000
200000
250000
300000
No_HH


250000
200000
150000
100000
50000
0
TOT_WORK_M
0
50000
100000
150000
200000
250000
300000
No_HH

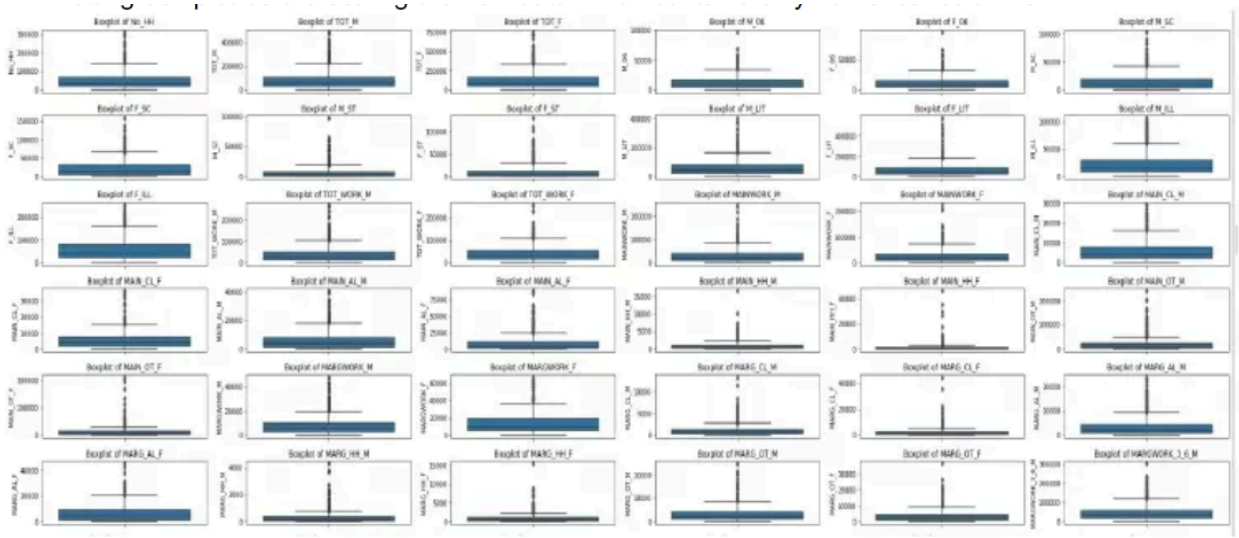


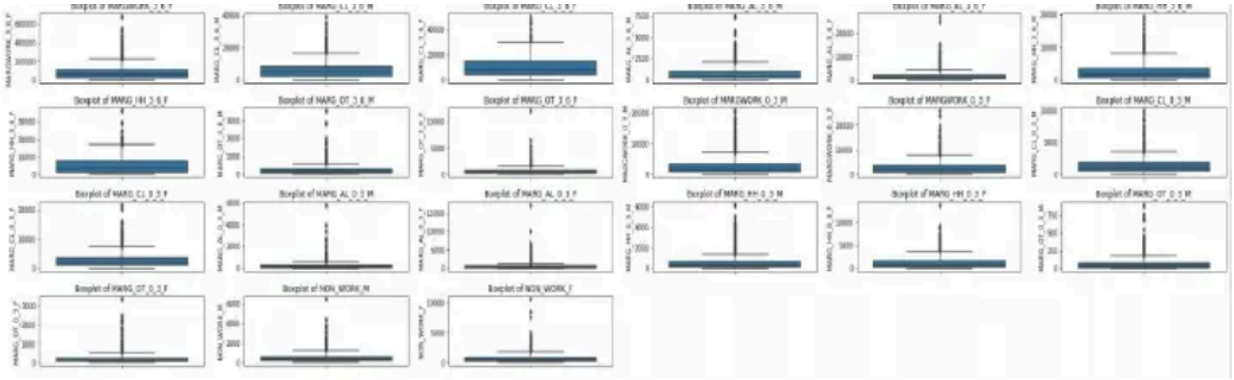
- Checking for outliers and treating them

	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	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
0	7707	23388	29796	5862	6196	3	0	1999	2598	13381	...	1150	749	180	237	680	252	32	46	259	214
1	6218	19585	23102	4482	3733	7	6	427	517	10513	...	525	715	123	229	186	148	76	178	140	160
2	4452	6546	10964	1082	1018	3	6	5806	9723	4534	...	114	188	44	89	3	34	0	4	57	61
3	1320	2784	4206	563	677	0	0	2666	3968	1842	...	194	247	61	128	13	50	4	10	115	89
4	11654	20891	29981	5157	4687	20	33	7670	10843	13243	...	874	1928	465	1043	206	302	24	105	180	478
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
635	3333	8154	11781	1146	1235	21	30	0	0	6916	...	32	47	0	0	0	0	0	0	32	47
636	10612	12046	21691	1544	1530	2204	4165	0	0	10292	...	155	337	3	14	38	130	4	23	110	170
637	1275	1549	2630	227	225	0	0	1012	1750	1187	...	104	134	9	4	2	6	17	47	75	77
638	3762	5280	8012	723	654	0	0	28	50	4206	...	136	172	24	44	11	21	1	4	100	103
639	7975	11977	18049	1470	1358	0	0	161	264	10095	...	173	122	6	2	17	17	2	4	148	99

640 rows × 57 columns

- Plotting box plots before scaling data





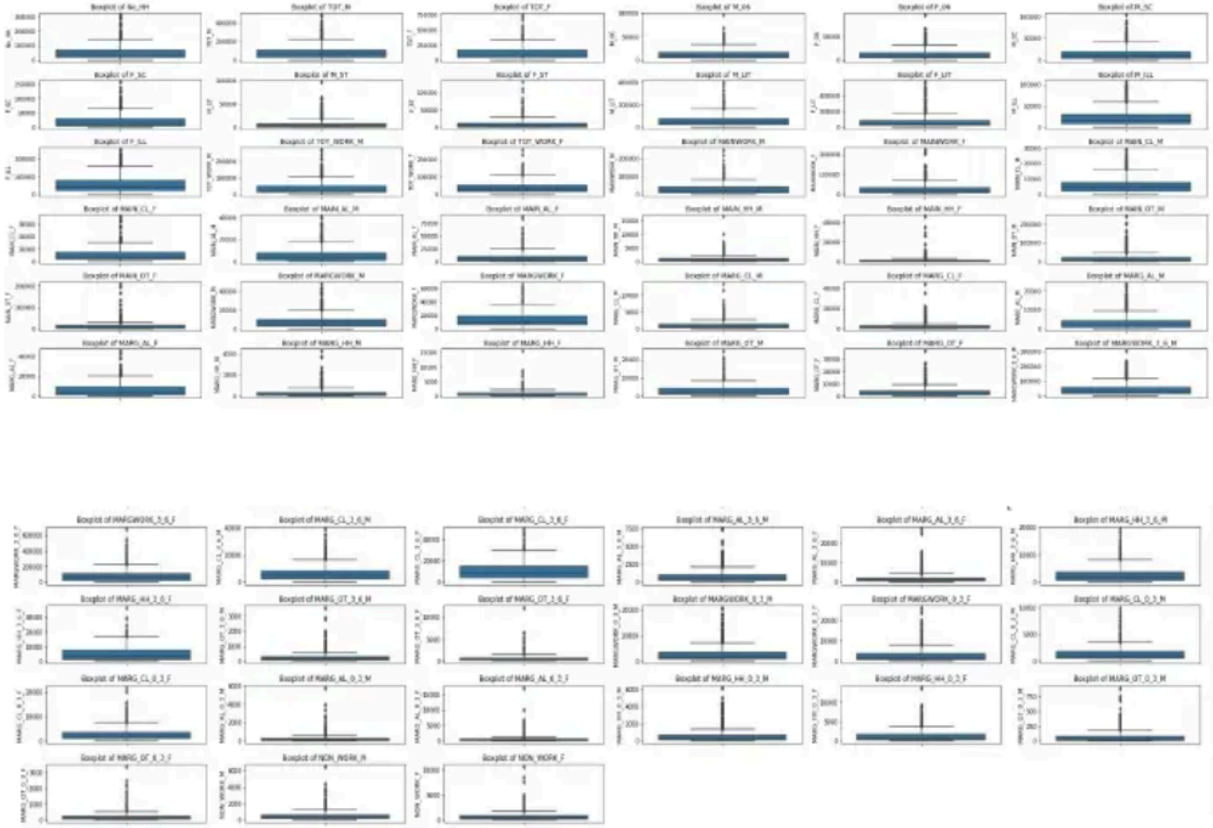
- **Scaling the data set using the Z score and checking for top 5 rows of the scaled dataset :**

	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	MARG_HH_0_3_M	MARG_HH
0	-0.904738	-0.771236	-0.815563	-0.561012	-0.507738	-0.958575	-0.957049	-0.423306	-0.478423	-0.798097	...	-0.163229	-0.720610	-0.159494	-0.287524	0.156577	-0
1	-0.935695	-0.823100	-0.874534	-0.681096	-0.725367	-0.958297	-0.956772	-0.582014	-0.607607	-0.849434	...	-0.583103	-0.732811	-0.282327	-0.294588	-0.491731	-0
2	-0.972412	-1.000919	-0.981466	-0.976856	-0.965262	-0.958575	-0.956772	-0.038951	-0.027273	-0.956457	...	-0.859212	-0.921931	-0.459727	-0.420050	-0.731894	-0
3	-1.037530	-1.052224	-1.041001	-1.022118	-0.995393	-0.958783	-0.957049	-0.355965	-0.390060	-1.004643	...	-0.905468	-0.900758	-0.419198	-0.385127	-0.718770	-0
4	-0.822678	-0.809361	-0.813933	-0.622359	-0.649908	-0.957395	-0.955529	0.149238	0.043330	-0.800566	...	-0.348645	-0.297513	0.472670	0.434200	-0.468796	-0

5 rows × 57 columns

MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
-0.365258	-0.499977	-0.413053	-0.539614
0.042855	-0.073481	-0.606455	-0.598988
-0.662068	-0.635680	-0.726103	-0.707839
-0.624966	-0.616294	-0.645791	-0.710038
-0.439461	-0.309346	-0.540895	-0.249344

- Checking for outliers of the scaled data



- Covariance matrix -Eigen values and eigen vectors

Extracting eigen vectors and looking at PCA components

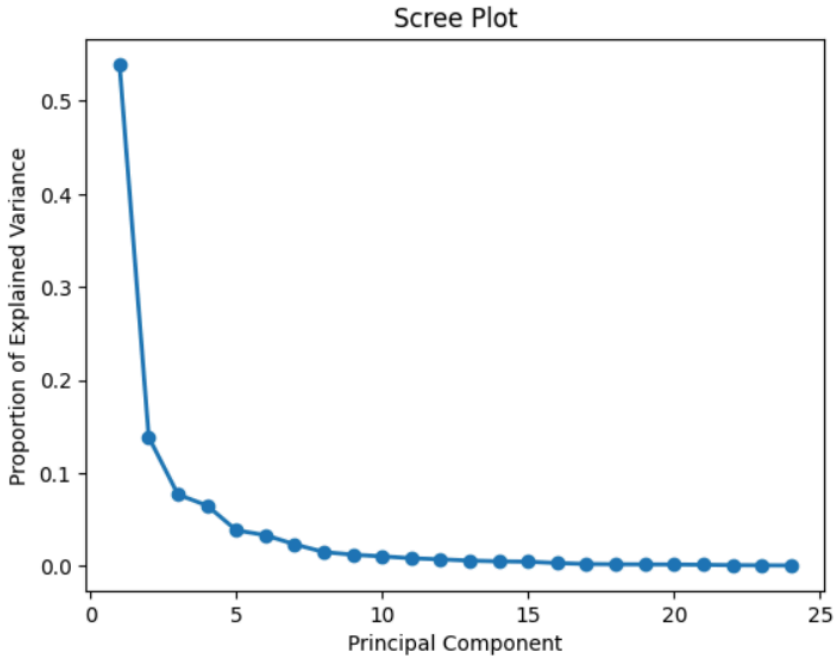
```
array([[ 0.15602058,  0.16711763,  0.16555318, ...,  0.13219224,
         0.15037558,  0.1310662 ],
       [-0.12634653, -0.08967655, -0.10491237, ...,  0.05081332,
        -0.06536455, -0.07384742],
       [-0.00269025,  0.05669762,  0.03874947, ..., -0.07871987,
         0.11182732,  0.1025525 ],
       ...,
       [ 0.        ,  0.37643683,  0.15058437, ...,  0.03363703,
        -0.07959556, -0.02552519],
       [-0.        ,  0.2448199 ,  0.09383958, ..., -0.02638552,
        -0.01672564,  0.03567243],
       [-0.        , -0.09325898, -0.0110033 , ...,  0.01165739,
        -0.01279215, -0.00377366]])
```

```
array([3.18135647e+01, 7.86942415e+00, 4.15340812e+00, 3.66879058e+00,
       2.20652588e+00, 1.93827502e+00, 1.17617374e+00, 7.51159086e-01,
       6.17053743e-01, 5.28300887e-01, 4.29831189e-01, 3.53440201e-01,
       2.96163013e-01, 2.81275560e-01, 1.92158325e-01, 1.36267920e-01,
       1.13389199e-01, 1.06303946e-01, 9.72885376e-02, 8.01062194e-02,
       5.76089954e-02, 4.43955966e-02, 3.78910846e-02, 2.96360194e-02,
       2.70797618e-02, 2.34458139e-02, 1.45111511e-02, 1.09852268e-02,
       9.31507853e-03, 8.13540203e-03, 7.89250253e-03, 5.02601514e-03,
       2.59771182e-03, 1.06789820e-03, 7.13559124e-04, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31])
```

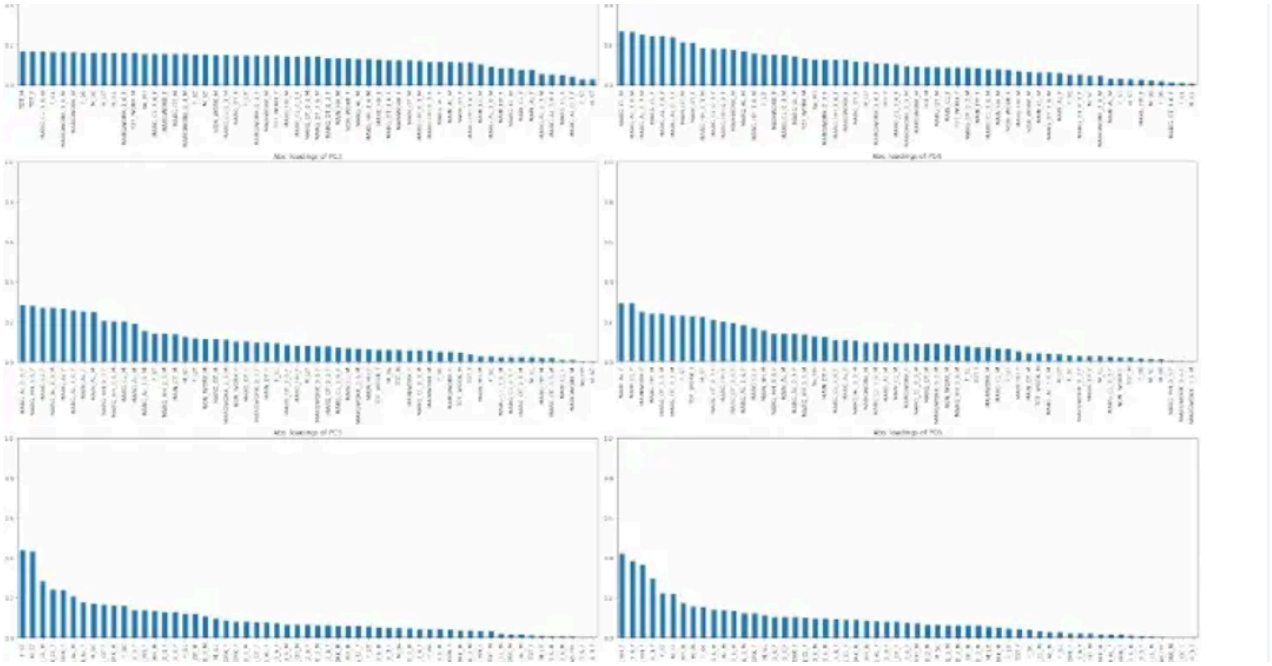
- Check the explained variance for each PC
  Explained variance=(eigen value of each PC)/(sum of eigen values of all PCs)

```
array([5.57260632e-01, 1.37844354e-01, 7.27529548e-02, 6.42641771e-02,
       3.86504944e-02, 3.39516923e-02, 2.06023855e-02, 1.31576386e-02,
       1.08085894e-02, 9.25395468e-03, 7.52911540e-03, 6.19101667e-03,
       5.18772384e-03, 4.92694855e-03, 3.36593119e-03, 2.38692984e-03,
       1.98617593e-03, 1.86206747e-03, 1.70414955e-03, 1.40317638e-03,
       1.00910494e-03, 7.77653131e-04, 6.63717190e-04, 5.19117774e-04,
       4.74341222e-04, 4.10687364e-04, 2.54183814e-04, 1.92422147e-04,
       1.63167083e-04, 1.42503342e-04, 1.38248605e-04, 8.80379297e-05,
       4.55026824e-05, 1.87057826e-05, 1.24990208e-05, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33])
```

- Optimum number of PC using scree plot



- Comparison of PCs with actual columns to find the variance



- Linear equation for the first principal component PC1

$$PC1 = 0.259 \times No_H H - 0.215 \times TOT_M - 0.215 \times TOT_F + 0.253 \times M_{06} + 0.254 \times F_{06} - 0.246 \times M_S C - 0.245 \times F_S C - 0.251 \times M_S T - 0.252 \times F_S T$$

$$- 0.195 \times M_L IT - 0.195 \times F_L IT + 0.195 \times M_I LL + 0.196 \times F_I LL - 0.202 \times TOT_W ORK_M - 0.202 \times TOT_W ORK_F -$$

$$0.198 \times MAINWORK_M - 0.198 \times MAINWORK_F - 0.196 \times MAIN_C L_M$$

$$- 0.195 \times MAIN_A L_F - 0.195 \times MAIN_H H_M - 0.195 \times MAIN_H H_F - 0.195 \times MAIN_O T_M - \quad 0.195 \times MAIN_O T_F$$

- SUMMARY-INFERENCES

1. Scaling can affect the outliers by changing their values, but it does not remove them
2. From the scree plot, we can see that the first 5 principal components explain a significant amount of variance in the data, with the first PC explaining the most variance.
3. The linear equation represents the combination of the original variables that make up the first principal component, where the

coefficients represent the weights assigned to each variable in the linear combination.

4. The sign of the coefficient indicates the direction of the relationship between the variable and the PC1 score. In this case, positive coefficients indicate a positive relationship with PC1, while negative coefficients indicate a negative relationship.
5. The first principal component as representing a combination of variables that are related to overall population and household size, with a negative influence from the number of male and female population and a positive influence from the number of households, population in the age group 0-6, and illiterate population. This component also shows a negative influence from the number of Scheduled Castes and Scheduled Tribes population and a negative influence from the number of workers, with a particularly strong negative influence from the number of cultivators.
6. Overall, PC1 represents a contrast between larger households with higher proportions of young and illiterate population, and a smaller population of Scheduled Castes and Scheduled Tribes, and fewer workers, particularly cultivators.
7. Overall, PC1 represents a contrast between larger households with higher proportions of young and illiterate population, and a smaller population of Scheduled Castes and Scheduled Tribes, and fewer workers, particularly cultivators.