BUSINESS REPORT
DATA MINING

Table of contents

Problem:- 1

Problem :- 2

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

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

Q.7) Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.

Q.8) Write linear equation for first PC**.**

Problem Statement: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:

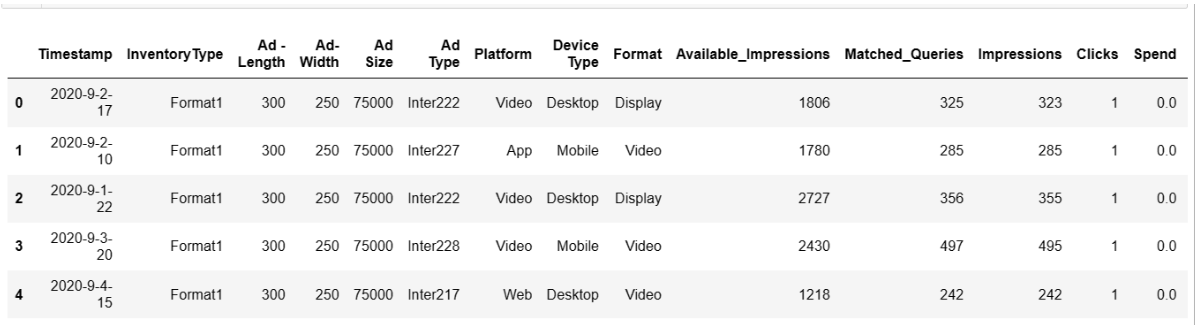
CPM = (Total Campaign Spend / Number of Impressions) * 1,000. Note that the Total Campaign Spend refers to the 'Spend' Column in the dataset and the Number of Impressions refers to the 'Impressions' Column in the dataset.

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.

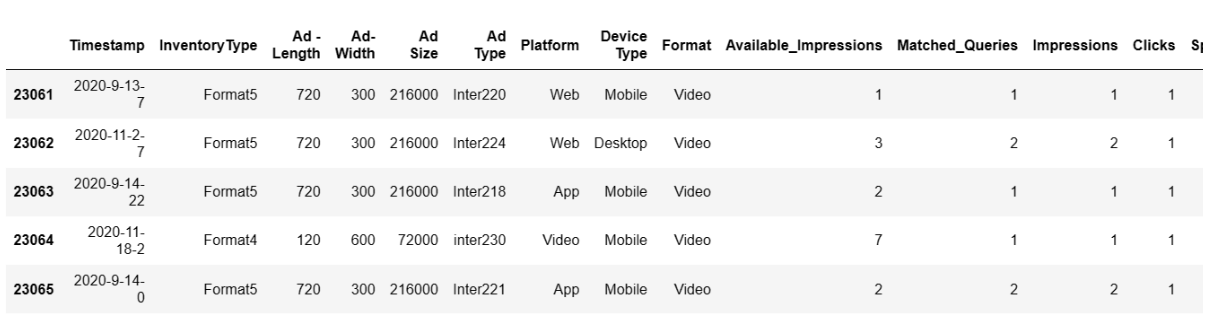
- **Q.1) Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc**.

Ans) Importing the basic libraries and Loading the dataset and getting the Top 5 rows:-



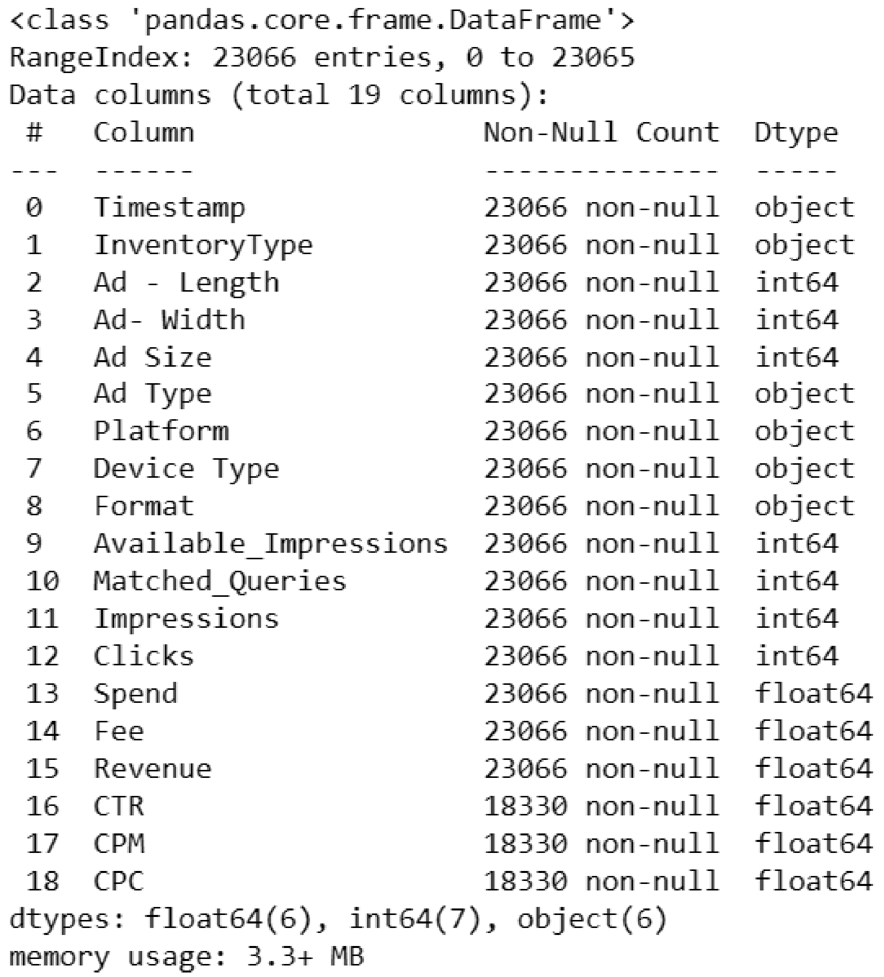
	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	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

Now, printing the last few records :-



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

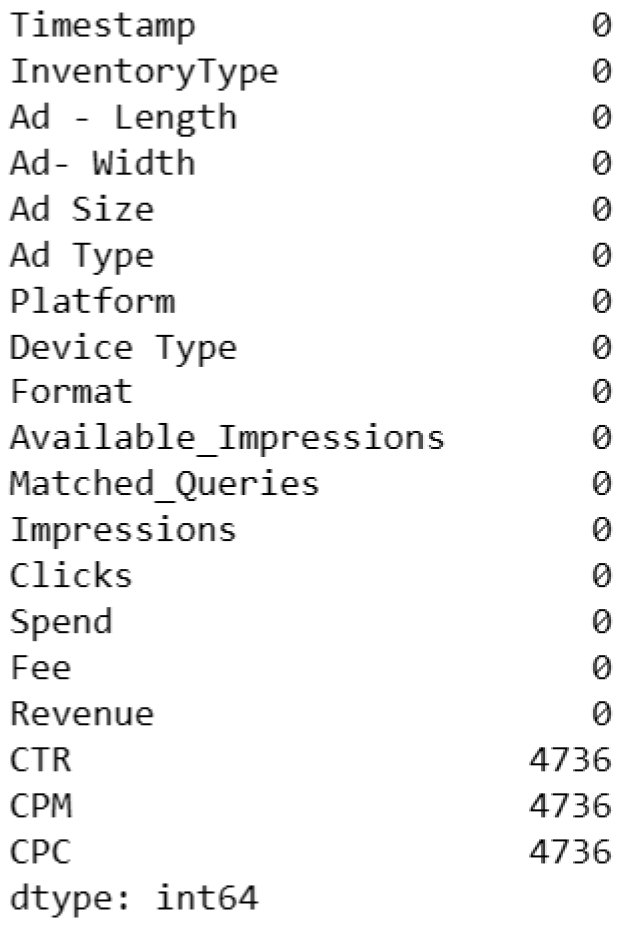
The Summary of the Dataset is as follows :-



```
<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
```

On Checking the duplicate records we have observed that there are no duplicate records present in the dataset

Checking the null-values we have observed :-



```
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
```

From the above result we can observe that CTR, CPM and CPC values are not null

- **Q.2) Treat missing values in CPC, CTR and CPM using the formula given. You may refer to the [Bank_KMeans Solution File] to understand the coding behind treating the missing values using a specific formula. You have to basically create an user defined function and then call the function for imputing.**
-

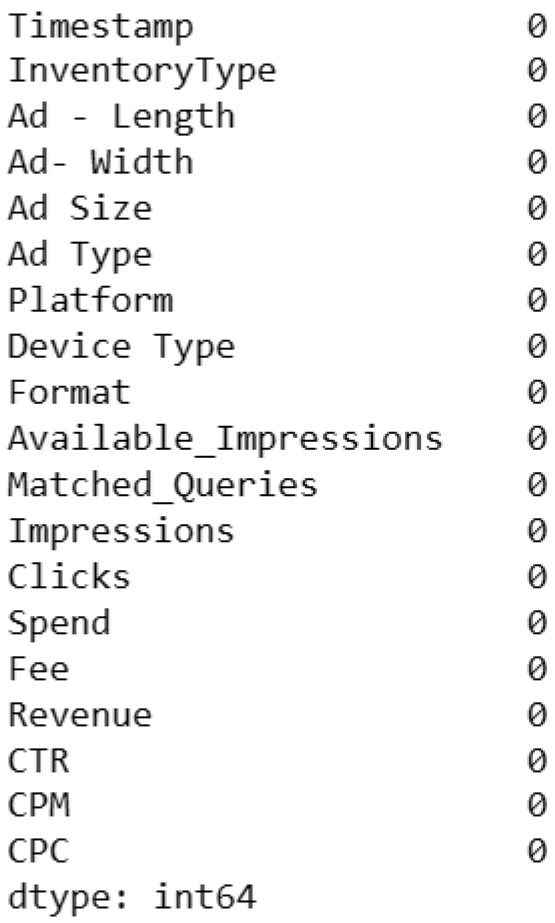
Ans. We can treat the missing values using the formula:-

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.

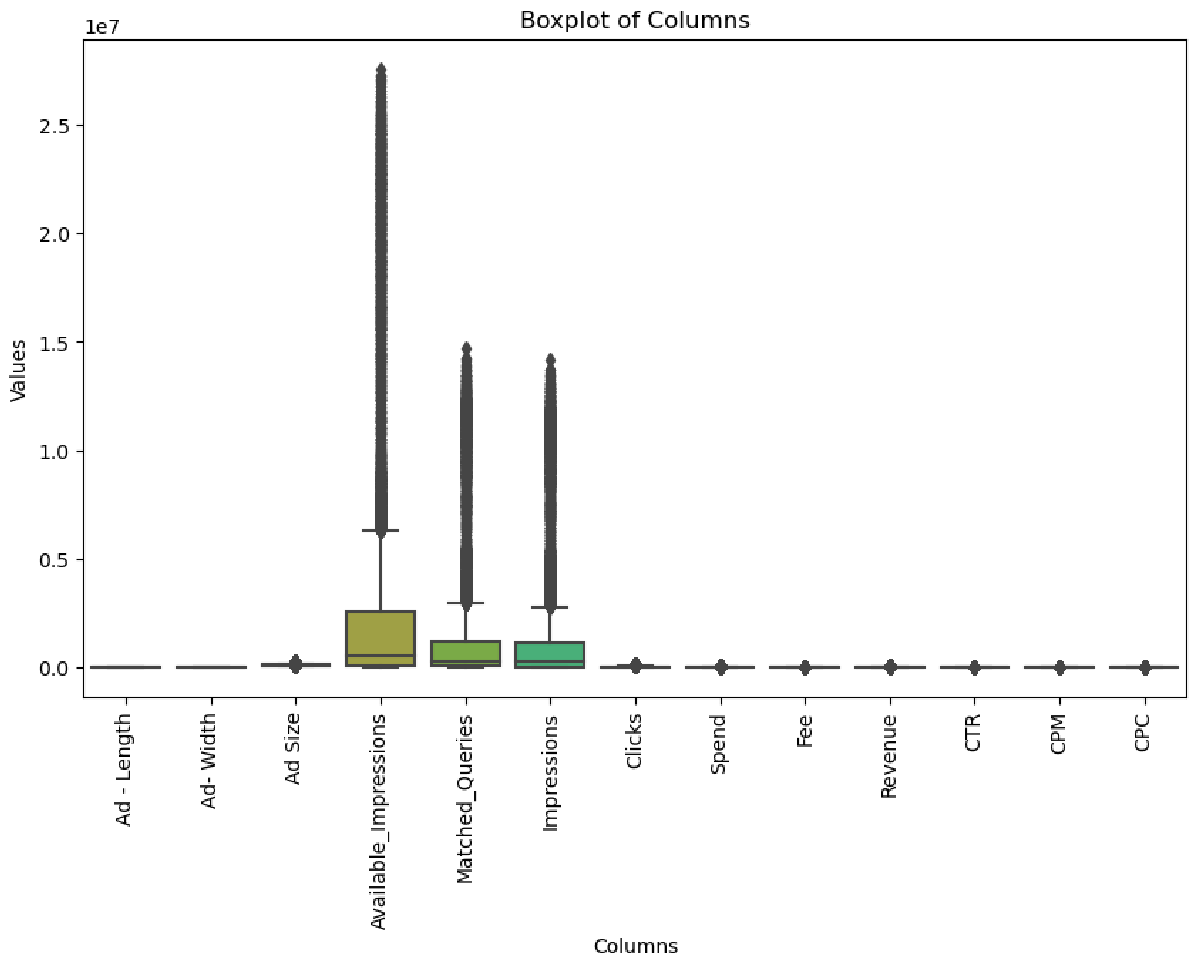
By using the above formula we have treated missing values. Now again checking the null values we have observed :-



```
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
```

Q.3) Check if there are any outliers.?

Ans. We can check the presence of outlier in the dataset using Box-plot. The below figure shows the presence of outliers



Q.4) 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).

Ans. Yes, outlier treatment is necessary for K-Means Clustering. On the basis of my judgement I have decided to treat outliers using IQR method

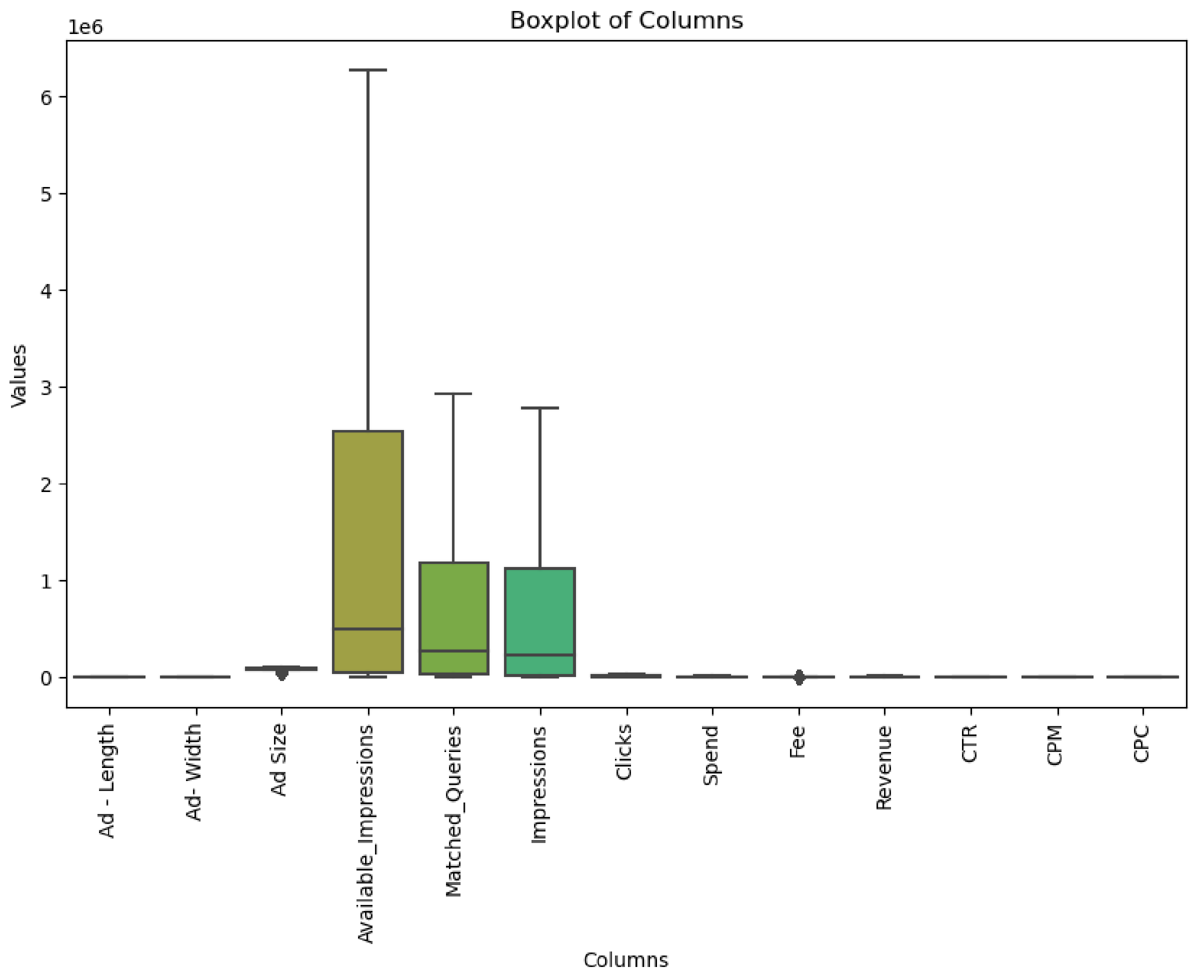
The IQR method is basically

Q1 = column.quantile(0.25)

Q3 = column.quantile(0.75)

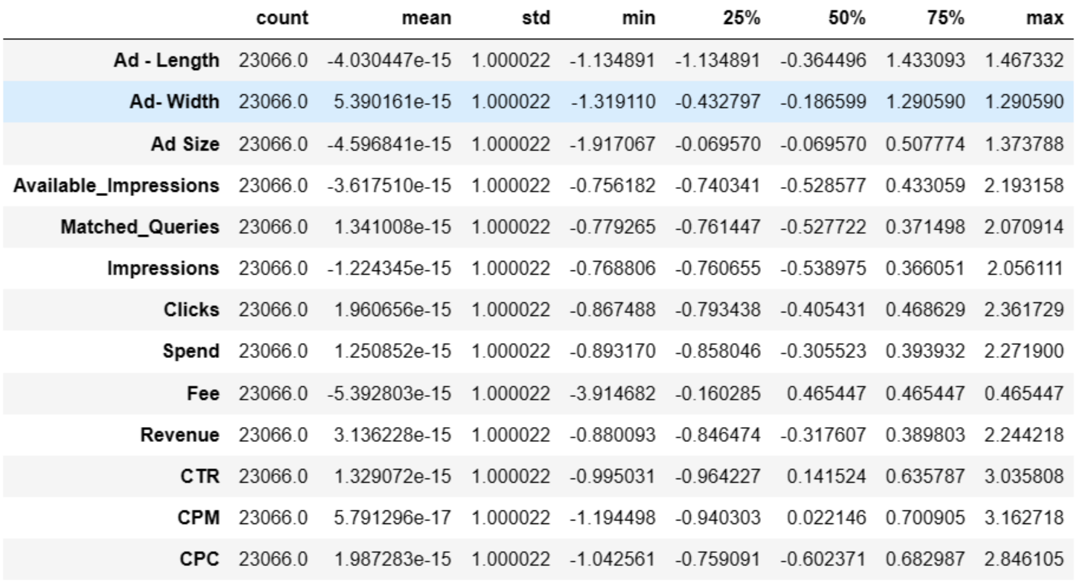
IQR = Q3 - Q1

Thus, using above formula we have treated outliers. Below diagram shows the box plot of treated outliers



Q.5) Perform z-score scaling and discuss how it affects the speed of the algorithm

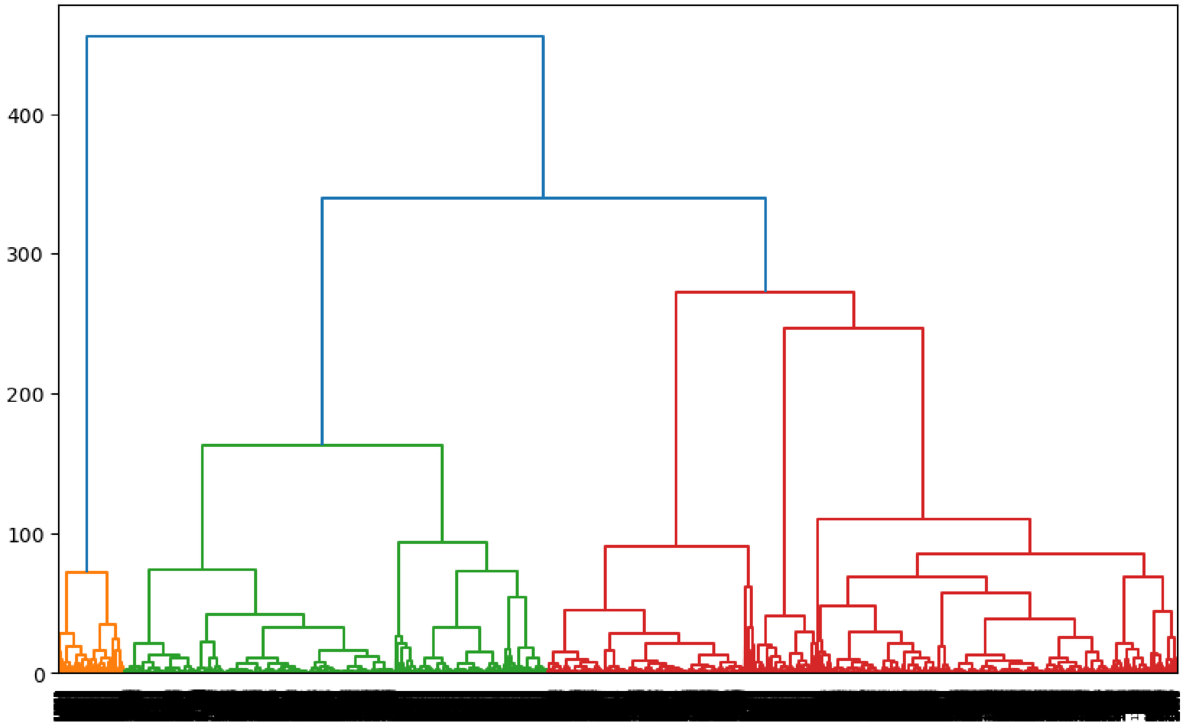
Ans. By performing the z-score scaling in the given dataset we have observed that scaling increases the memory usages which impacted algorithm performance as well as it involves extra computation which affects overall execution time. Below data represent the dataset after scaling



	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	-4.030447e-15	1.000022	-1.134891	-1.134891	-0.364496	1.433093	1.467332
Ad- Width	23066.0	5.390161e-15	1.000022	-1.319110	-0.432797	-0.186599	1.290590	1.290590
Ad Size	23066.0	-4.596841e-15	1.000022	-1.917067	-0.069570	-0.069570	0.507774	1.373788
Available_Impressions	23066.0	-3.617510e-15	1.000022	-0.756182	-0.740341	-0.528577	0.433059	2.193158
Matched_Queries	23066.0	1.341008e-15	1.000022	-0.779265	-0.761447	-0.527722	0.371498	2.070914
Impressions	23066.0	-1.224345e-15	1.000022	-0.768806	-0.760655	-0.538975	0.366051	2.056111
Clicks	23066.0	1.960656e-15	1.000022	-0.867488	-0.793438	-0.405431	0.468629	2.361729
Spend	23066.0	1.250852e-15	1.000022	-0.893170	-0.858046	-0.305523	0.393932	2.271900
Fee	23066.0	-5.392803e-15	1.000022	-3.914682	-0.160285	0.465447	0.465447	0.465447
Revenue	23066.0	3.136228e-15	1.000022	-0.880093	-0.846474	-0.317607	0.389803	2.244218
CTR	23066.0	1.329072e-15	1.000022	-0.995031	-0.964227	0.141524	0.635787	3.035808
CPM	23066.0	5.791296e-17	1.000022	-1.194498	-0.940303	0.022146	0.700905	3.162718
CPC	23066.0	1.987283e-15	1.000022	-1.042561	-0.759091	-0.602371	0.682987	2.846105

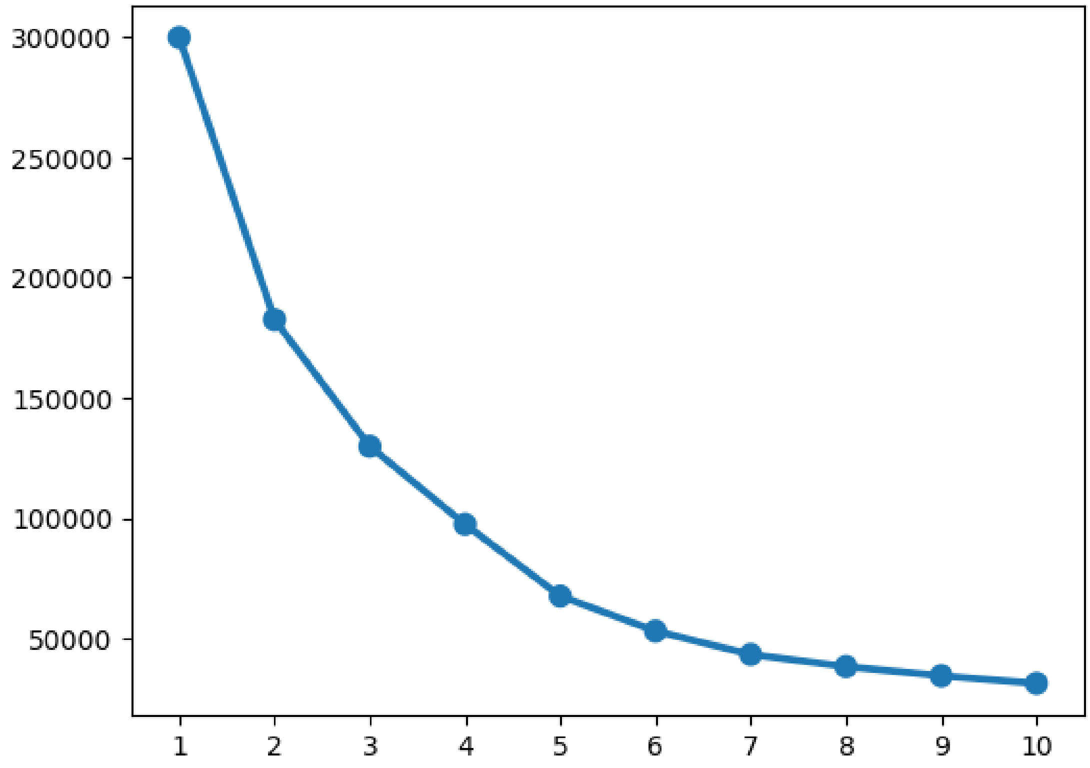
Q.6) Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.

Ans. By constructing a Dendrogram using WARD and Euclidean distance of the scaled data is shown below :-



Q.7) Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm

Ans. The Elbow plot (up to n=10) is shown below :-



We can use WSS method for checking the optimal no of clusters. On using this formula we can observe that from clusters 5 to 6 the values reduces as compared to clusters 1, 2, 3, 4, 5. Hence, we can use 5 clusters as optimal no of clusters

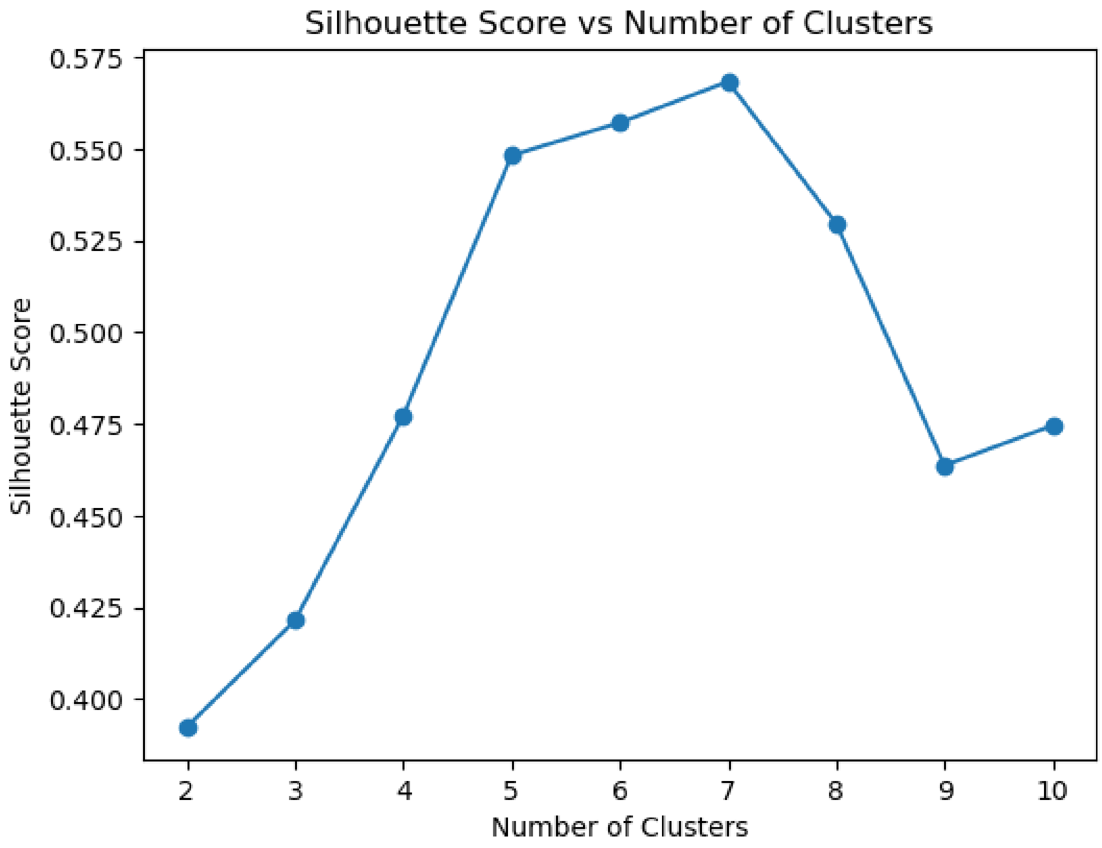
Q.7) Print silhouette scores for up to 10 clusters and identify optimum number of clusters

Ans. The silhouette scores for up to 10 clusters is as follows :-

```
For n_clusters=2, the silhouette score is 0.3923066069287274
For n_clusters=3, the silhouette score is 0.4213647016545103
For n_clusters=4, the silhouette score is 0.47726717100695615
For n_clusters=5, the silhouette score is 0.5483112473610738
For n_clusters=6, the silhouette score is 0.5572454232383197
For n_clusters=7, the silhouette score is 0.5684396102017544
For n_clusters=8, the silhouette score is 0.5296758658996598
For n_clusters=9, the silhouette score is 0.46191222457922787
For n_clusters=10, the silhouette score is 0.4610009389013054
```

The value of silhouette Score is usually ranges from -1 to 1

The diagram for silhouette score is as follows:-



Q.9) **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].**

Q.10) Conclude the project by providing summary of your learnings.

Ans. The summary of the learnng are as follows :-

i) The dataset contains 23066 rows and 19 columns

ii) We have identified the missing values in the dataset and learn how to treat those missing values

iii) We have observed that outliers is present in our dataset and treated them using IQR method

iv) We have done the scaling of the dataset using z-score method and also learn the scaling impact on algorithm

v) We have also plotted Dendrogram and Elbow plot which helps us to calculate the no of clusters

vi) We get the 5 optimal clusters from the dataset

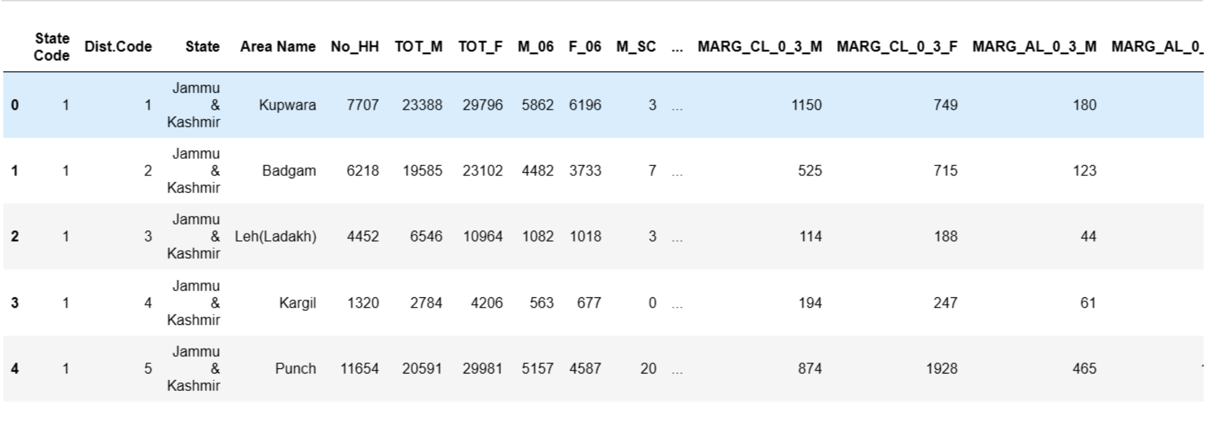
Problem Statement:2

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.

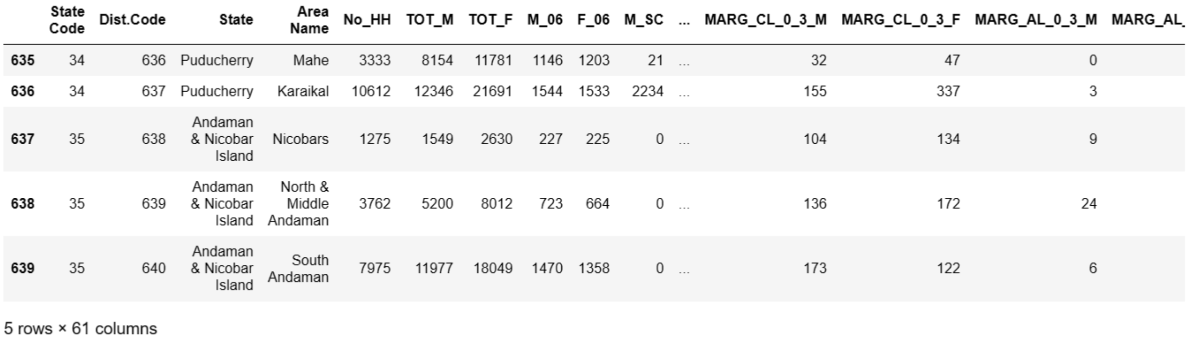
Q.1) **Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.**

Ans. Importing the basic libraries and Loading the dataset and getting the 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_
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	

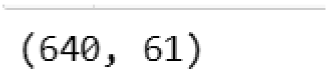
The last few records are as follows :-



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

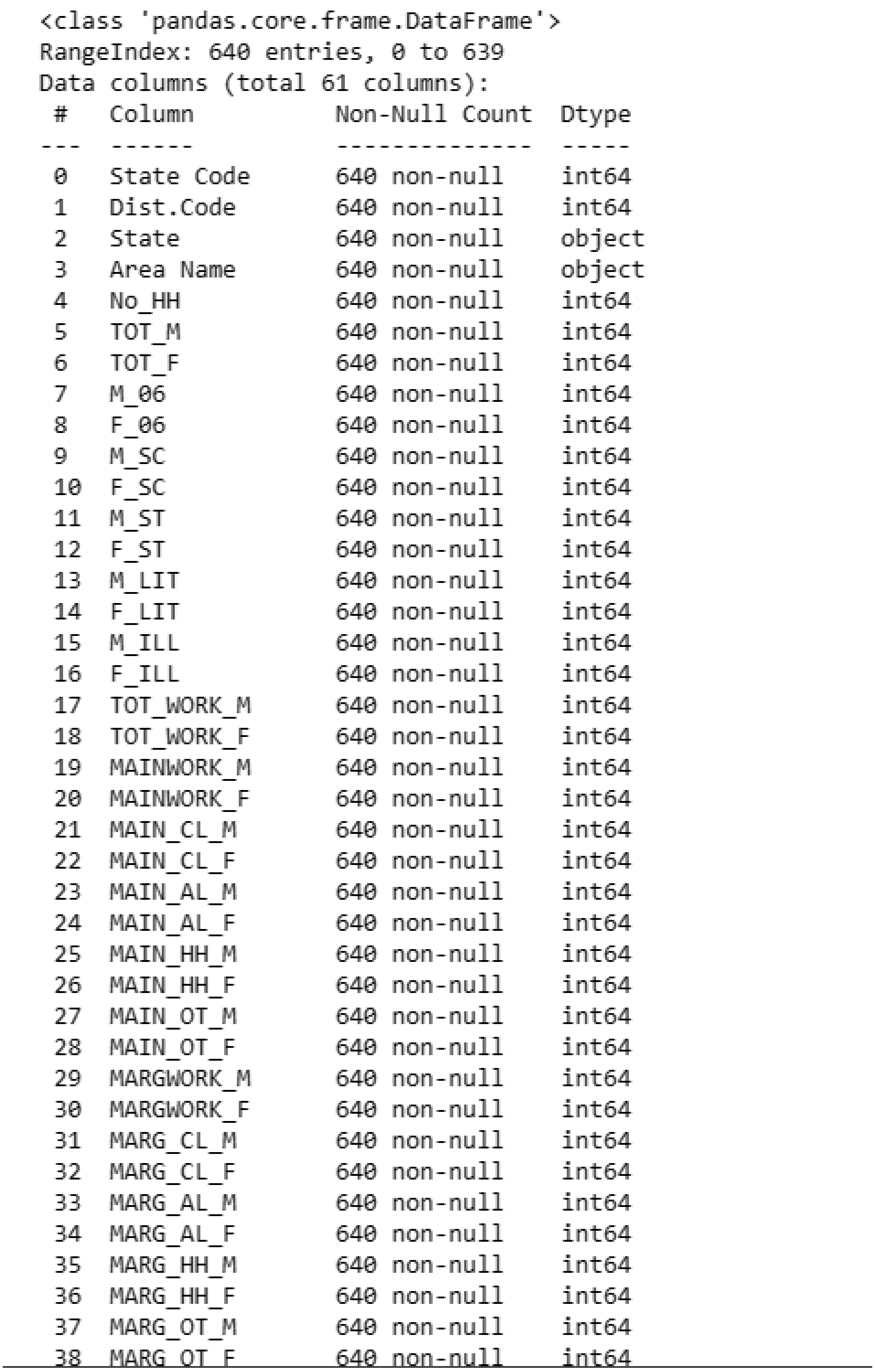
5 rows × 61 columns

There are 640 rows and 61 columns in the dataset

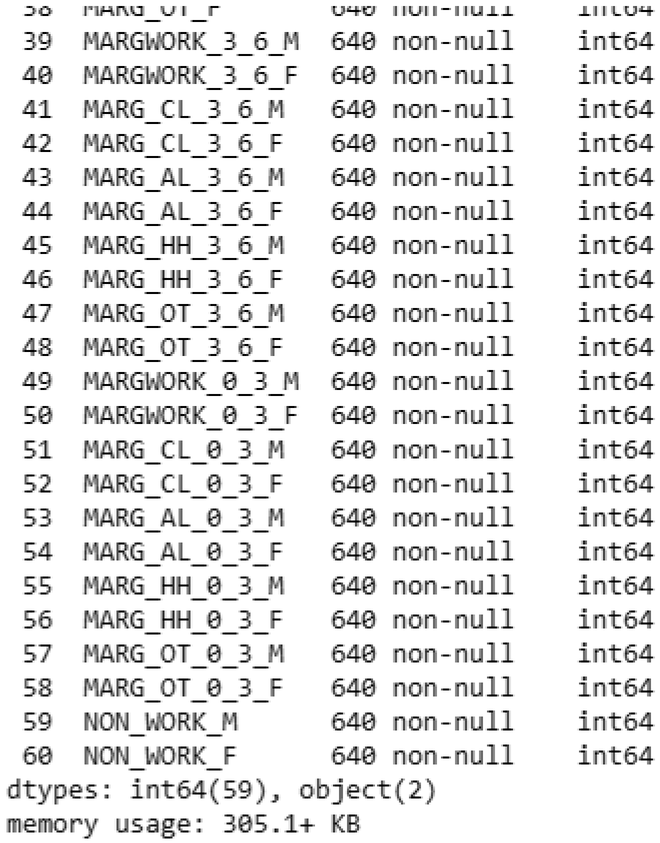


```
(640, 61)
```

The Summarization of the dataset is as follows :-



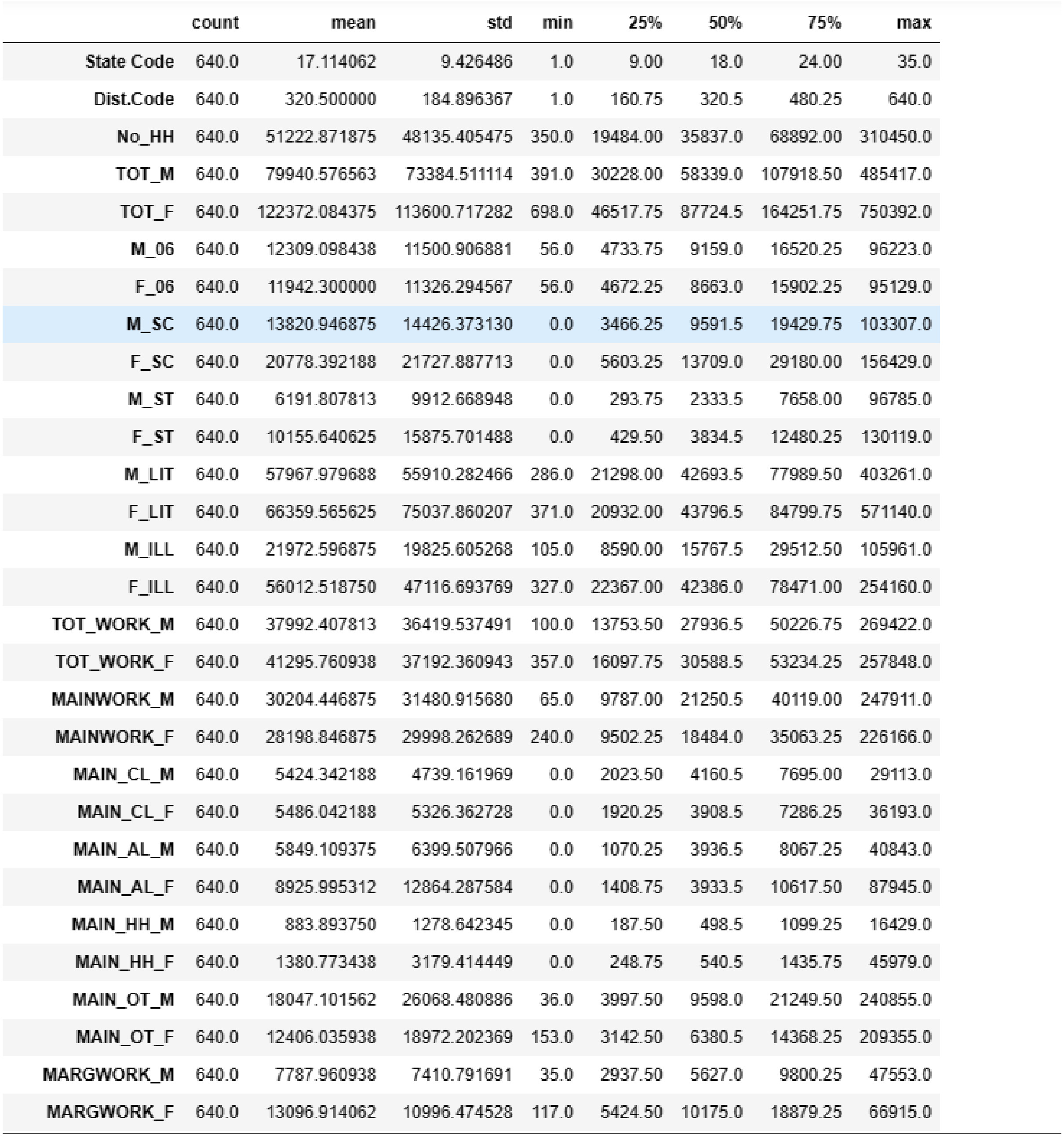
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 640 entries, 0 to 639
Data columns (total 61 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   State Code      640 non-null    int64 
 1   Dist.Code       640 non-null    int64 
 2   State           640 non-null    object
 3   Area Name       640 non-null    object
 4   No_HH           640 non-null    int64 
 5   TOT_M           640 non-null    int64 
 6   TOT_F           640 non-null    int64 
 7   M_06            640 non-null    int64 
 8   F_06            640 non-null    int64 
 9   M_SC            640 non-null    int64 
 10  F_SC            640 non-null    int64 
 11  M_ST            640 non-null    int64 
 12  F_ST            640 non-null    int64 
 13  M_LIT           640 non-null    int64 
 14  F_LIT           640 non-null    int64 
 15  M_ILL           640 non-null    int64 
 16  F_ILL           640 non-null    int64 
 17  TOT_WORK_M      640 non-null    int64 
 18  TOT_WORK_F      640 non-null    int64 
 19  MAINWORK_M      640 non-null    int64 
 20  MAINWORK_F      640 non-null    int64 
 21  MAIN_CL_M       640 non-null    int64 
 22  MAIN_CL_F       640 non-null    int64 
 23  MAIN_AL_M       640 non-null    int64 
 24  MAIN_AL_F       640 non-null    int64 
 25  MAIN_HH_M       640 non-null    int64 
 26  MAIN_HH_F       640 non-null    int64 
 27  MAIN_OT_M       640 non-null    int64 
 28  MAIN_OT_F       640 non-null    int64 
 29  MARGWORK_M      640 non-null    int64 
 30  MARGWORK_F      640 non-null    int64 
 31  MARG_CL_M       640 non-null    int64 
 32  MARG_CL_F       640 non-null    int64 
 33  MARG_AL_M       640 non-null    int64 
 34  MARG_AL_F       640 non-null    int64 
 35  MARG_HH_M       640 non-null    int64 
 36  MARG_HH_F       640 non-null    int64 
 37  MARG_OT_M       640 non-null    int64 
 38  MARG_OT_F       640 non-null    int64 
```



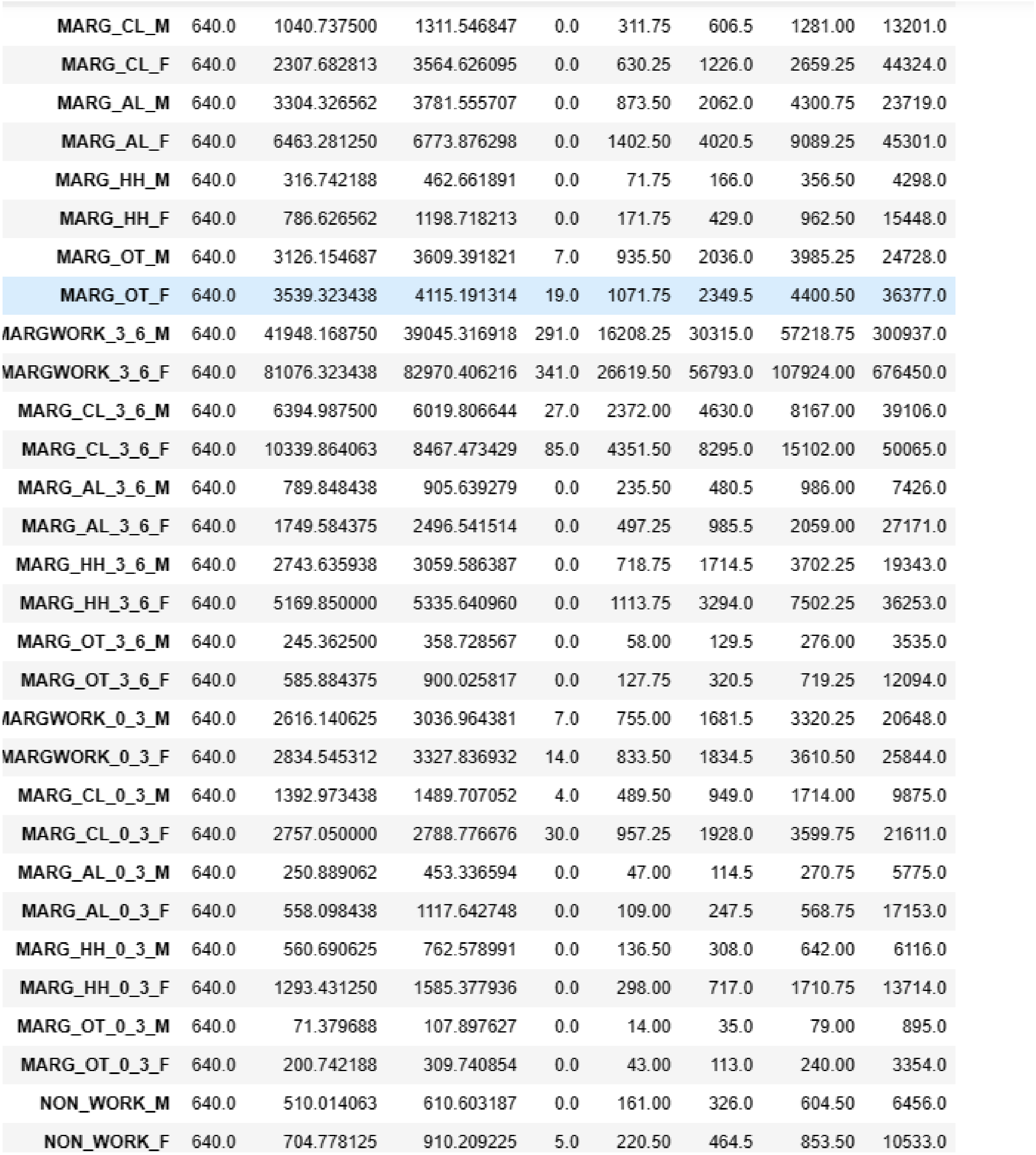
```
 38  MARG_OT_F       640 non-null    int64
 39  MARGWORK_3_6_M  640 non-null    int64
 40  MARGWORK_3_6_F  640 non-null    int64
 41  MARG_CL_3_6_M   640 non-null    int64
 42  MARG_CL_3_6_F   640 non-null    int64
 43  MARG_AL_3_6_M   640 non-null    int64
 44  MARG_AL_3_6_F   640 non-null    int64
 45  MARG_HH_3_6_M   640 non-null    int64
 46  MARG_HH_3_6_F   640 non-null    int64
 47  MARG_OT_3_6_M   640 non-null    int64
 48  MARG_OT_3_6_F   640 non-null    int64
 49  MARGWORK_0_3_M  640 non-null    int64
 50  MARGWORK_0_3_F  640 non-null    int64
 51  MARG_CL_0_3_M   640 non-null    int64
 52  MARG_CL_0_3_F   640 non-null    int64
 53  MARG_AL_0_3_M   640 non-null    int64
 54  MARG_AL_0_3_F   640 non-null    int64
 55  MARG_HH_0_3_M   640 non-null    int64
 56  MARG_HH_0_3_F   640 non-null    int64
 57  MARG_OT_0_3_M   640 non-null    int64
 58  MARG_OT_0_3_F   640 non-null    int64
 59  NON_WORK_M      640 non-null    int64
 60  NON_WORK_F      640 non-null    int64
dtypes: int64(59), object(2)
memory usage: 305.1+ KB
```

On checking duplicate and null values we have observe that there are no duplicate and no null values present in the dataset

The description of the dataset is as follows :-



	count	mean	std	min	25%	50%	75%	max
State Code	640.0	17.114062	9.426486	1.0	9.00	18.0	24.00	35.0
Dist.Code	640.0	320.500000	184.896367	1.0	160.75	320.5	480.25	640.0
No_HH	640.0	51222.871875	48135.405475	350.0	19484.00	35837.0	68892.00	310450.0
TOT_M	640.0	79940.576563	73384.511114	391.0	30228.00	58339.0	107918.50	485417.0
TOT_F	640.0	122372.084375	113600.717282	698.0	46517.75	87724.5	164251.75	750392.0
M_06	640.0	12309.098438	11500.906881	56.0	4733.75	9159.0	16520.25	96223.0
F_06	640.0	11942.300000	11326.294567	56.0	4672.25	8663.0	15902.25	95129.0
M_SC	640.0	13820.946875	14426.373130	0.0	3466.25	9591.5	19429.75	103307.0
F_SC	640.0	20778.392188	21727.887713	0.0	5603.25	13709.0	29180.00	156429.0
M_ST	640.0	6191.807813	9912.668948	0.0	293.75	2333.5	7658.00	96785.0
F_ST	640.0	10155.640625	15875.701488	0.0	429.50	3834.5	12480.25	130119.0
M_LIT	640.0	57967.979688	55910.282466	286.0	21298.00	42693.5	77989.50	403261.0
F_LIT	640.0	66359.565625	75037.860207	371.0	20932.00	43796.5	84799.75	571140.0
M_ILL	640.0	21972.596875	19825.605268	105.0	8590.00	15767.5	29512.50	105961.0
F_ILL	640.0	56012.518750	47116.693769	327.0	22367.00	42386.0	78471.00	254160.0
TOT_WORK_M	640.0	37992.407813	36419.537491	100.0	13753.50	27936.5	50226.75	269422.0
TOT_WORK_F	640.0	41295.760938	37192.360943	357.0	16097.75	30588.5	53234.25	257848.0
MAINWORK_M	640.0	30204.446875	31480.915680	65.0	9787.00	21250.5	40119.00	247911.0
MAINWORK_F	640.0	28198.846875	29998.262689	240.0	9502.25	18484.0	35063.25	226166.0
MAIN_CL_M	640.0	5424.342188	4739.161969	0.0	2023.50	4160.5	7695.00	29113.0
MAIN_CL_F	640.0	5486.042188	5326.362728	0.0	1920.25	3908.5	7286.25	36193.0
MAIN_AL_M	640.0	5849.109375	6399.507966	0.0	1070.25	3936.5	8067.25	40843.0
MAIN_AL_F	640.0	8925.995312	12864.287584	0.0	1408.75	3933.5	10617.50	87945.0
MAIN_HH_M	640.0	883.893750	1278.642345	0.0	187.50	498.5	1099.25	16429.0
MAIN_HH_F	640.0	1380.773438	3179.414449	0.0	248.75	540.5	1435.75	45979.0
MAIN_OT_M	640.0	18047.101562	26068.480886	36.0	3997.50	9598.0	21249.50	240855.0
MAIN_OT_F	640.0	12406.035938	18972.202369	153.0	3142.50	6380.5	14368.25	209355.0
MARGWORK_M	640.0	7787.960938	7410.791691	35.0	2937.50	5627.0	9800.25	47553.0
MARGWORK_F	640.0	13096.914062	10996.474528	117.0	5424.50	10175.0	18879.25	66915.0



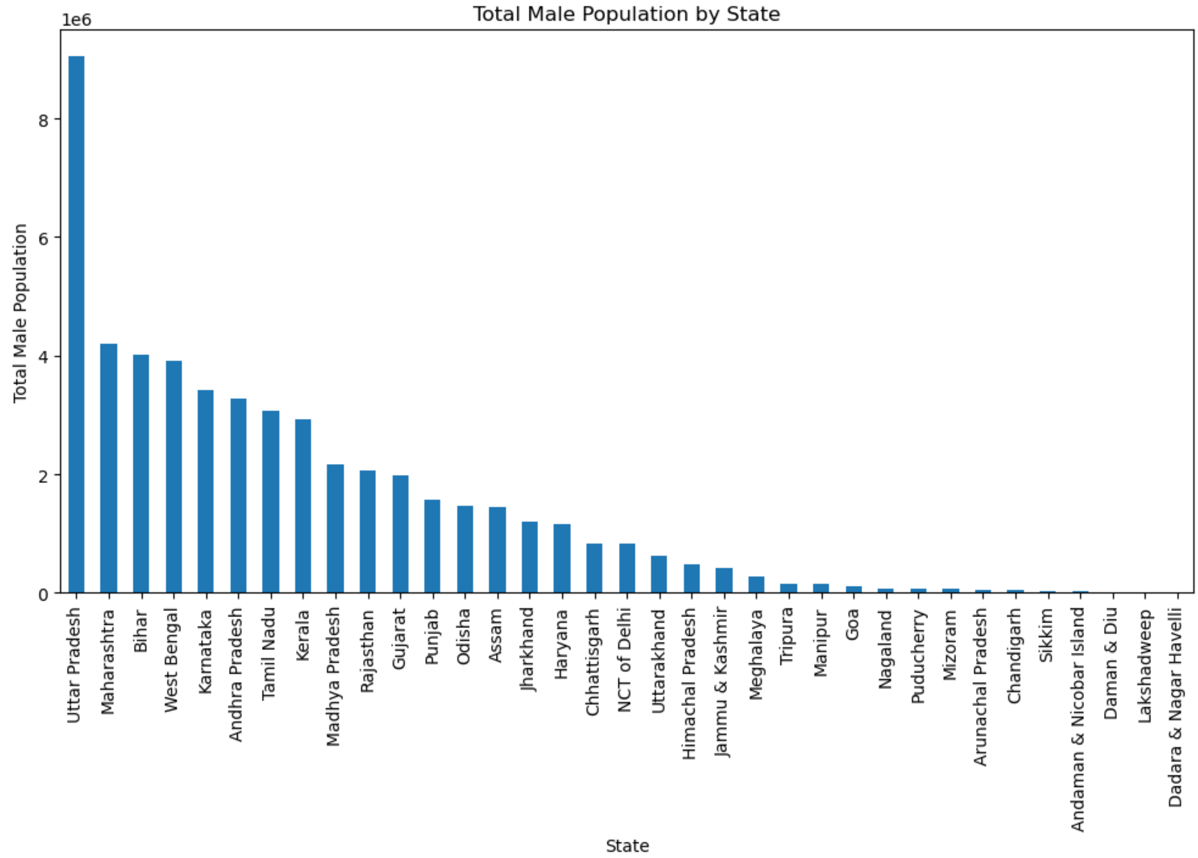
MARG_CL_M	640.0	1040.737500	1311.546847	0.0	311.75	606.5	1281.00	13201.0
MARG_CL_F	640.0	2307.682813	3564.626095	0.0	630.25	1226.0	2659.25	44324.0
MARG_AL_M	640.0	3304.326562	3781.555707	0.0	873.50	2062.0	4300.75	23719.0
MARG_AL_F	640.0	6463.281250	6773.876298	0.0	1402.50	4020.5	9089.25	45301.0
MARG_HH_M	640.0	316.742188	462.661891	0.0	71.75	166.0	356.50	4298.0
MARG_HH_F	640.0	786.626562	1198.718213	0.0	171.75	429.0	962.50	15448.0
MARG_OT_M	640.0	3126.154687	3609.391821	7.0	935.50	2036.0	3985.25	24728.0
MARG_OT_F	640.0	3539.323438	4115.191314	19.0	1071.75	2349.5	4400.50	36377.0
ARGWORK_3_6_M	640.0	41948.168750	39045.316918	291.0	16208.25	30315.0	57218.75	300937.0
MARGWORK_3_6_F	640.0	81076.323438	82970.406216	341.0	26619.50	56793.0	107924.00	676450.0
MARG_CL_3_6_M	640.0	6394.987500	6019.806644	27.0	2372.00	4630.0	8167.00	39106.0
MARG_CL_3_6_F	640.0	10339.864063	8467.473429	85.0	4351.50	8295.0	15102.00	50065.0
MARG_AL_3_6_M	640.0	789.848438	905.639279	0.0	235.50	480.5	986.00	7426.0
MARG_AL_3_6_F	640.0	1749.584375	2496.541514	0.0	497.25	985.5	2059.00	27171.0
MARG_HH_3_6_M	640.0	2743.635938	3059.586387	0.0	718.75	1714.5	3702.25	19343.0
MARG_HH_3_6_F	640.0	5169.850000	5335.640960	0.0	1113.75	3294.0	7502.25	36253.0
MARG_OT_3_6_M	640.0	245.362500	358.728567	0.0	58.00	129.5	276.00	3535.0
MARG_OT_3_6_F	640.0	585.884375	900.025817	0.0	127.75	320.5	719.25	12094.0
ARGWORK_0_3_M	640.0	2616.140625	3036.964381	7.0	755.00	1681.5	3320.25	20648.0
MARGWORK_0_3_F	640.0	2834.545312	3327.836932	14.0	833.50	1834.5	3610.50	25844.0
MARG_CL_0_3_M	640.0	1392.973438	1489.707052	4.0	489.50	949.0	1714.00	9875.0
MARG_CL_0_3_F	640.0	2757.050000	2788.776676	30.0	957.25	1928.0	3599.75	21611.0
MARG_AL_0_3_M	640.0	250.889062	453.336594	0.0	47.00	114.5	270.75	5775.0
MARG_AL_0_3_F	640.0	558.098438	1117.642748	0.0	109.00	247.5	568.75	17153.0
MARG_HH_0_3_M	640.0	560.690625	762.578991	0.0	136.50	308.0	642.00	6116.0
MARG_HH_0_3_F	640.0	1293.431250	1585.377936	0.0	298.00	717.0	1710.75	13714.0
MARG_OT_0_3_M	640.0	71.379688	107.897627	0.0	14.00	35.0	79.00	895.0
MARG_OT_0_3_F	640.0	200.742188	309.740854	0.0	43.00	113.0	240.00	3354.0
NON_WORK_M	640.0	510.014063	610.603187	0.0	161.00	326.0	604.50	6456.0
NON_WORK_F	640.0	704.778125	910.209225	5.0	220.50	464.5	853.50	10533.0

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

Ans. By performing the exploratory analysis we can get the answer of following questions (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio? As follows

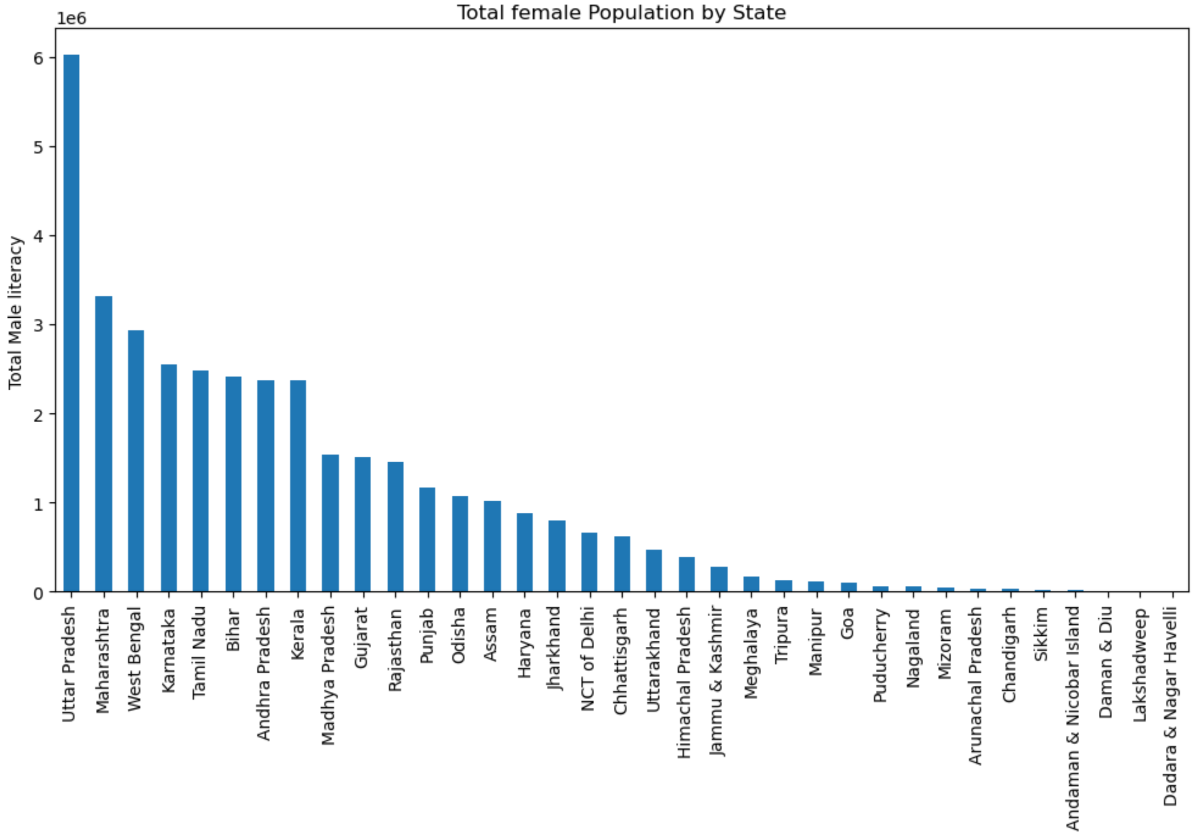
```
State with the highest gender ratio: Andhra Pradesh
State with the lowest gender ratio: Lakshadweep
District with the highest gender ratio: ('Andhra Pradesh', 547)
District with the lowest gender ratio: ('Lakshadweep', 587)
```

The 5 variables are :- **TOT_M, M_LIT, NON_WORK_M, NON_WORK_F, F_SC**



The above bar plot between total male population vs state helps us to know the following question

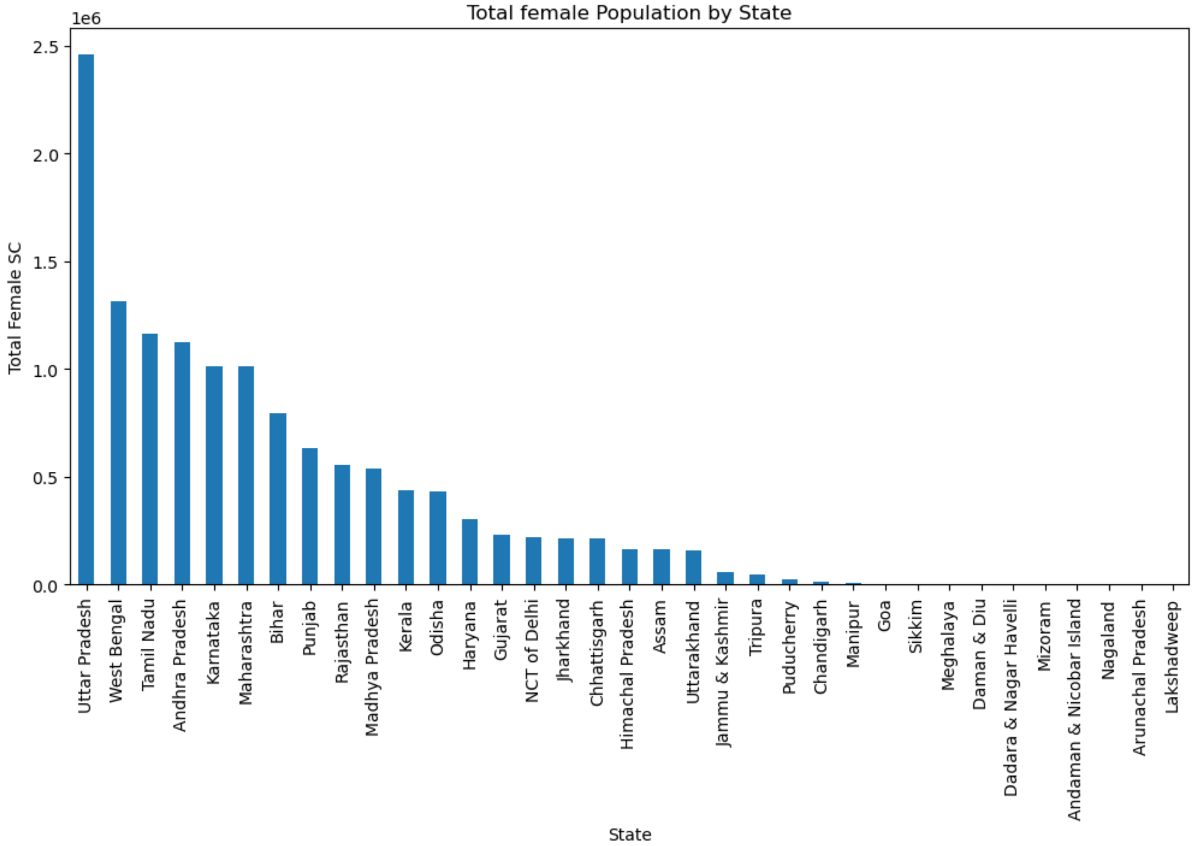
i) Which state has highest male population ?
ii) Which state has lowest male population ?



The above bar plot between total male literacy vs state helps us to know the following question

i) Which state has highest male literacy ?

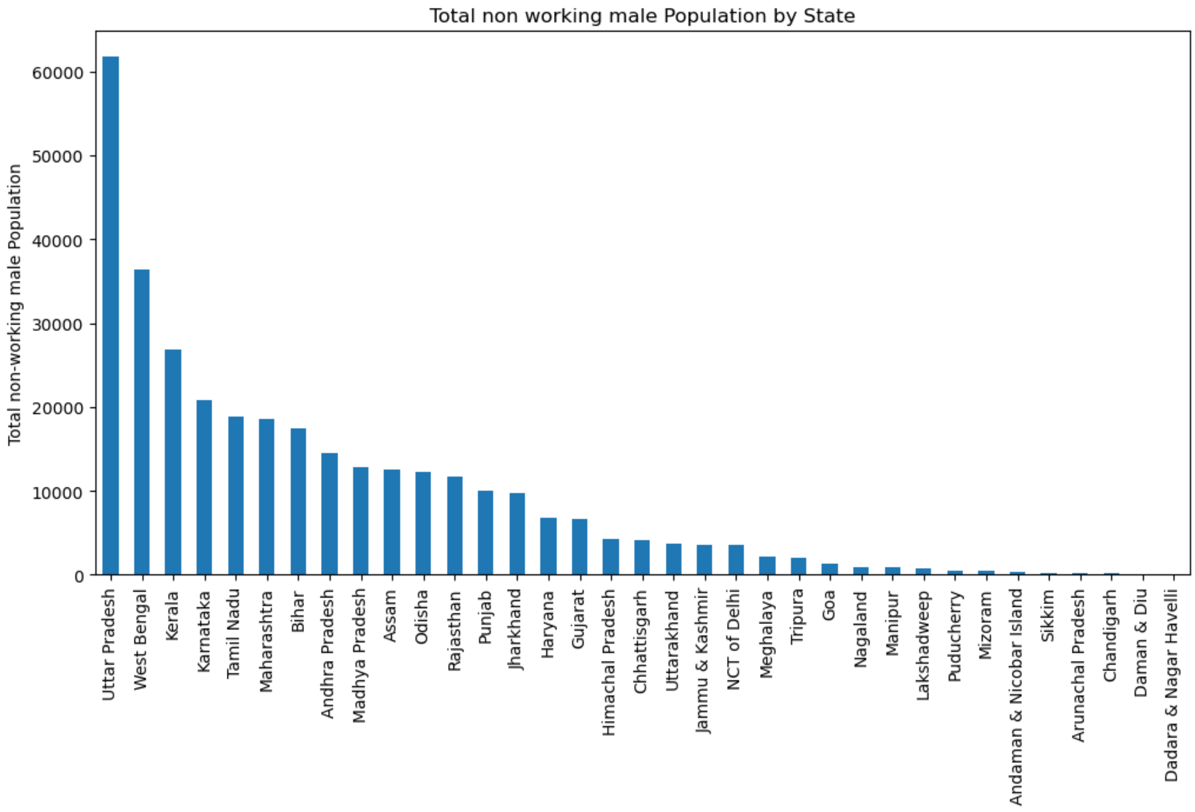
ii) Which state has lowest male literacy ?



The above bar plot between total female SC vs state helps us to know the following question

i) Which state has highest female SC ?

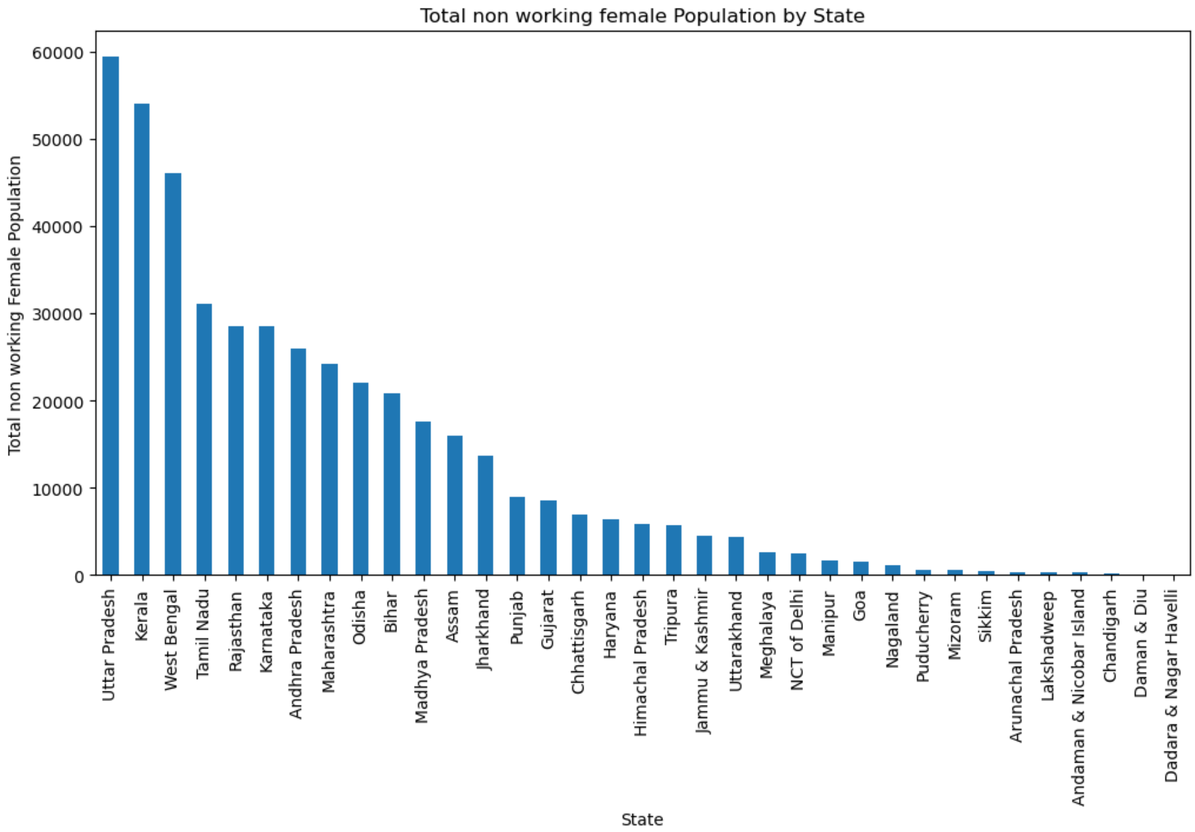
ii) Which state has lowest female SC ?



The above bar plot between total non-working male vs state helps us to know the following question

i) Which state has highest non-working male ?

ii) Which state has lowest non-working male ?



The above bar plot between total non- working female vs state helps us to know the following question

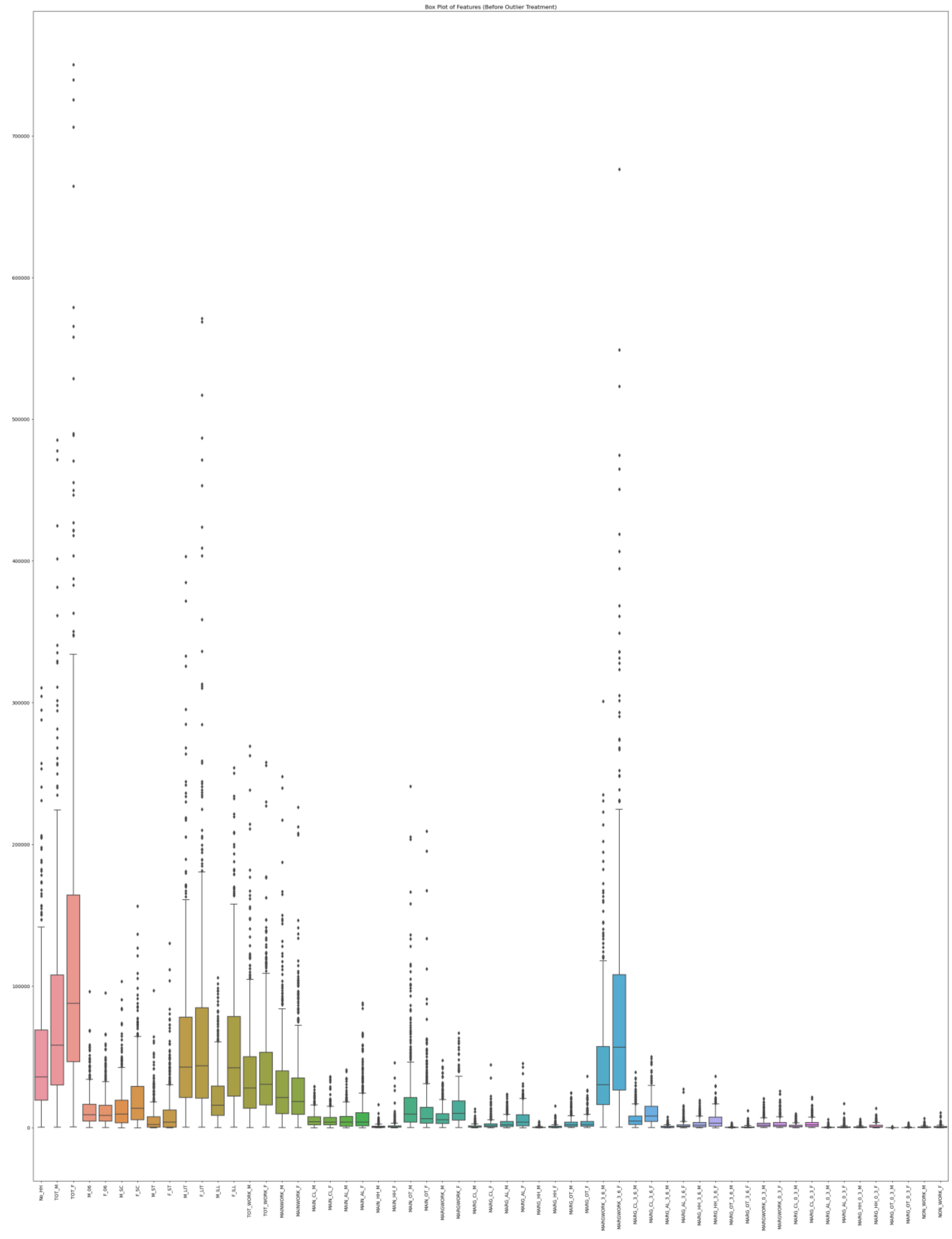
i) Which state has highest non-working female ?

ii) Which state has lowest non-working female ?

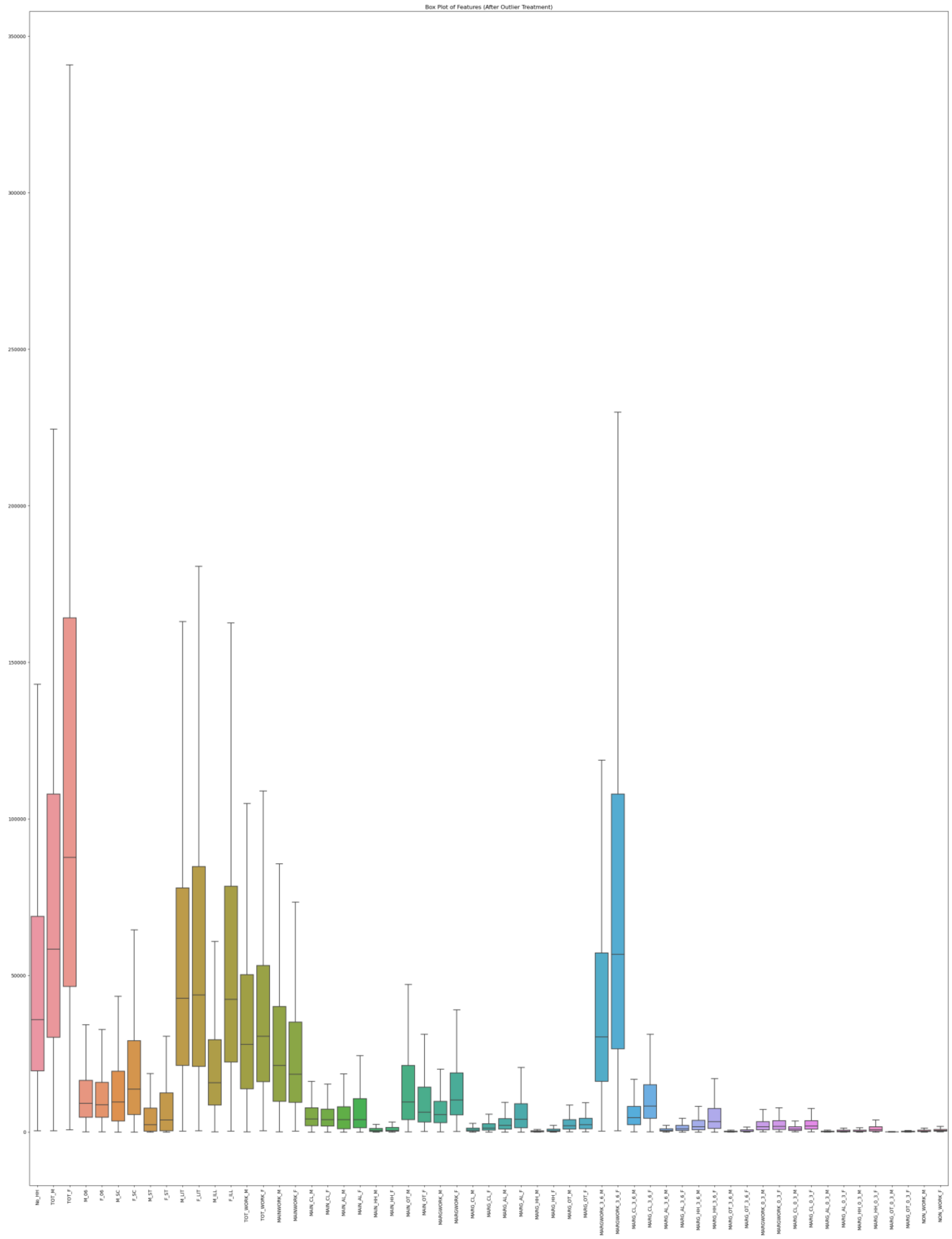
Q.3) We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?

Ans. Yes, according to me the treating outliers in this case is necessary we will treat them using IQR method

Before treating outliers :-



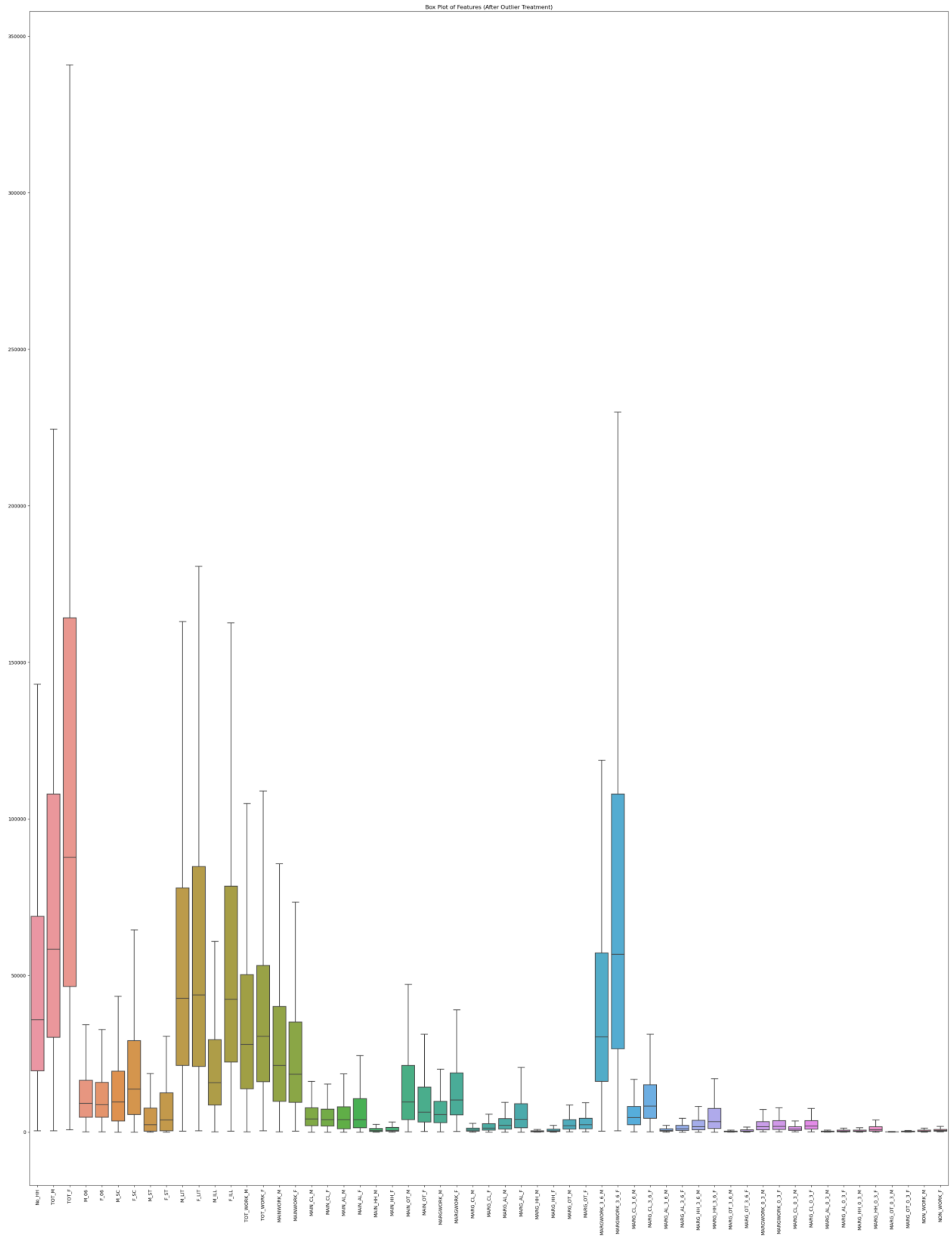
After outliers treatment :-



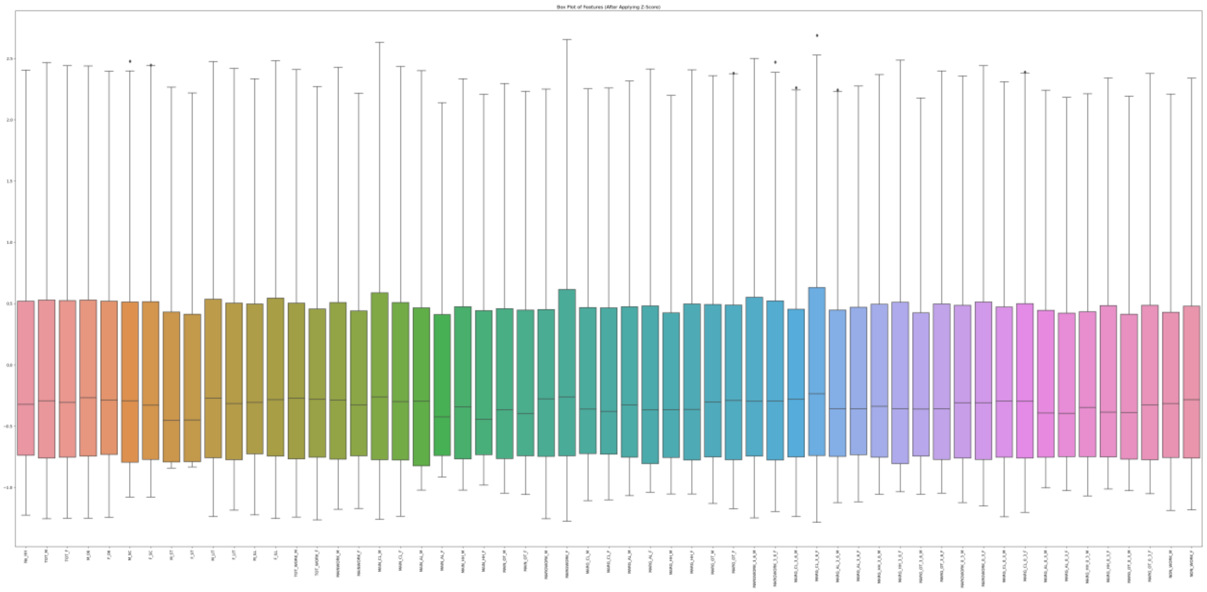
Q.4) Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.

Ans. Scale the data using Z score method and then plot then using box plot is as follows

Before z- score

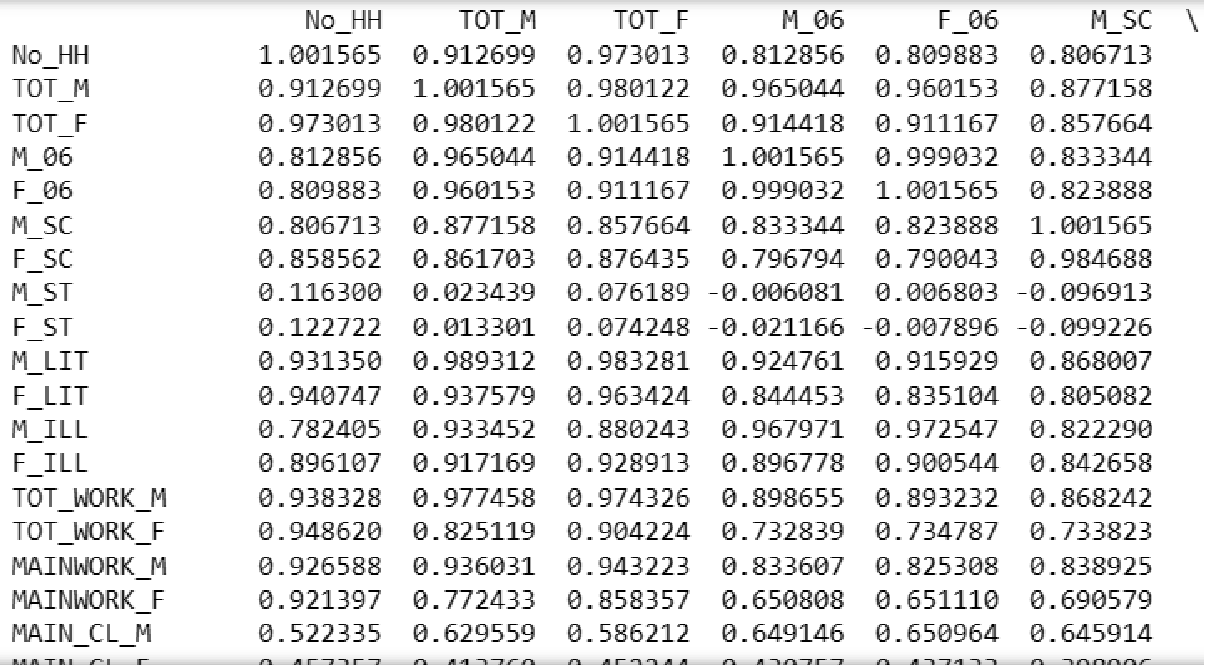


After applying z- score



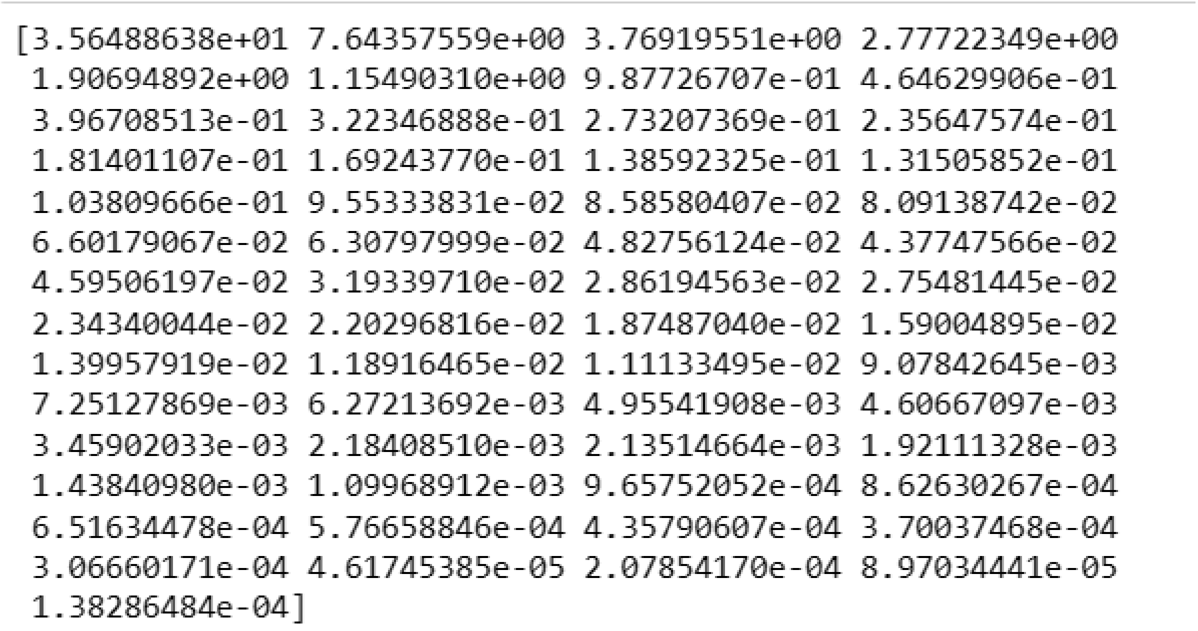
Q.5) Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.

Ans. After performing the required steps for PCA the covariance matrix is as follows



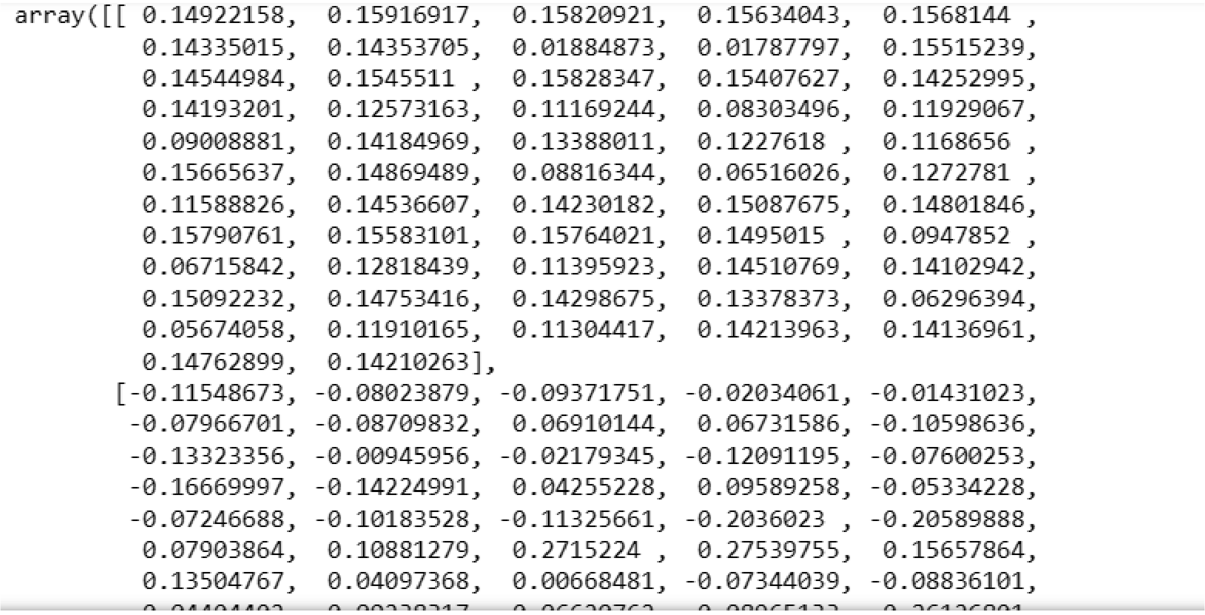
```
                 No_HH     TOT_M     TOT_F      M_06      F_06      M_SC  \
No_HH         1.001565  0.912699  0.973013  0.812856  0.809883  0.806713
TOT_M         0.912699  1.001565  0.980122  0.965044  0.960153  0.877158
TOT_F         0.973013  0.980122  1.001565  0.914418  0.911167  0.857664
M_06          0.812856  0.965044  0.914418  1.001565  0.999032  0.833344
F_06          0.809883  0.960153  0.911167  0.999032  1.001565  0.823888
M_SC          0.806713  0.877158  0.857664  0.833344  0.823888  1.001565
F_SC          0.858562  0.861703  0.876435  0.796794  0.790043  0.984688
M_ST          0.116300  0.023439  0.076189 -0.006081  0.006803 -0.096913
F_ST          0.122722  0.013301  0.074248 -0.021166 -0.007896 -0.099226
M_LIT         0.931350  0.989312  0.983281  0.924761  0.915929  0.868007
F_LIT         0.940747  0.937579  0.963424  0.844453  0.835104  0.805082
M_ILL         0.782405  0.933452  0.880243  0.967971  0.972547  0.822290
F_ILL         0.896107  0.917169  0.928913  0.896778  0.900544  0.842658
TOT_WORK_M    0.938328  0.977458  0.974326  0.898655  0.893232  0.868242
TOT_WORK_F    0.948620  0.825119  0.904224  0.732839  0.734787  0.733823
MAINWORK_M    0.926588  0.936031  0.943223  0.833607  0.825308  0.838925
MAINWORK_F    0.921397  0.772433  0.858357  0.650808  0.651110  0.690579
MAIN_CL_M     0.522335  0.629559  0.586212  0.649146  0.650964  0.645914
MAIN_CL_F     0.457357  0.413760  0.452344  0.430757  0.437133  0.398006
```

The eigen values of the dataset is as follows :-



```
[3.56488638e+01 7.64357559e+00 3.76919551e+00 2.77722349e+00
 1.90694892e+00 1.15490310e+00 9.87726707e-01 4.64629906e-01
 3.96708513e-01 3.22346888e-01 2.73207369e-01 2.35647574e-01
 1.81401107e-01 1.69243770e-01 1.38592325e-01 1.31505852e-01
 1.03809666e-01 9.55333831e-02 8.58580407e-02 8.09138742e-02
 6.60179067e-02 6.30797999e-02 4.82756124e-02 4.37747566e-02
 4.59506197e-02 3.19339710e-02 2.86194563e-02 2.75481445e-02
 2.34340044e-02 2.20296816e-02 1.87487040e-02 1.59004895e-02
 1.39957919e-02 1.18916465e-02 1.11133495e-02 9.07842645e-03
 7.25127869e-03 6.27213692e-03 4.95541908e-03 4.60667097e-03
 3.45902033e-03 2.18408510e-03 2.13514664e-03 1.92111328e-03
 1.43840980e-03 1.09968912e-03 9.65752052e-04 8.62630267e-04
 6.51634478e-04 5.76658846e-04 4.35790607e-04 3.70037468e-04
 3.06660171e-04 4.61745385e-05 2.07854170e-04 8.97034441e-05
 1.38286484e-04]
```

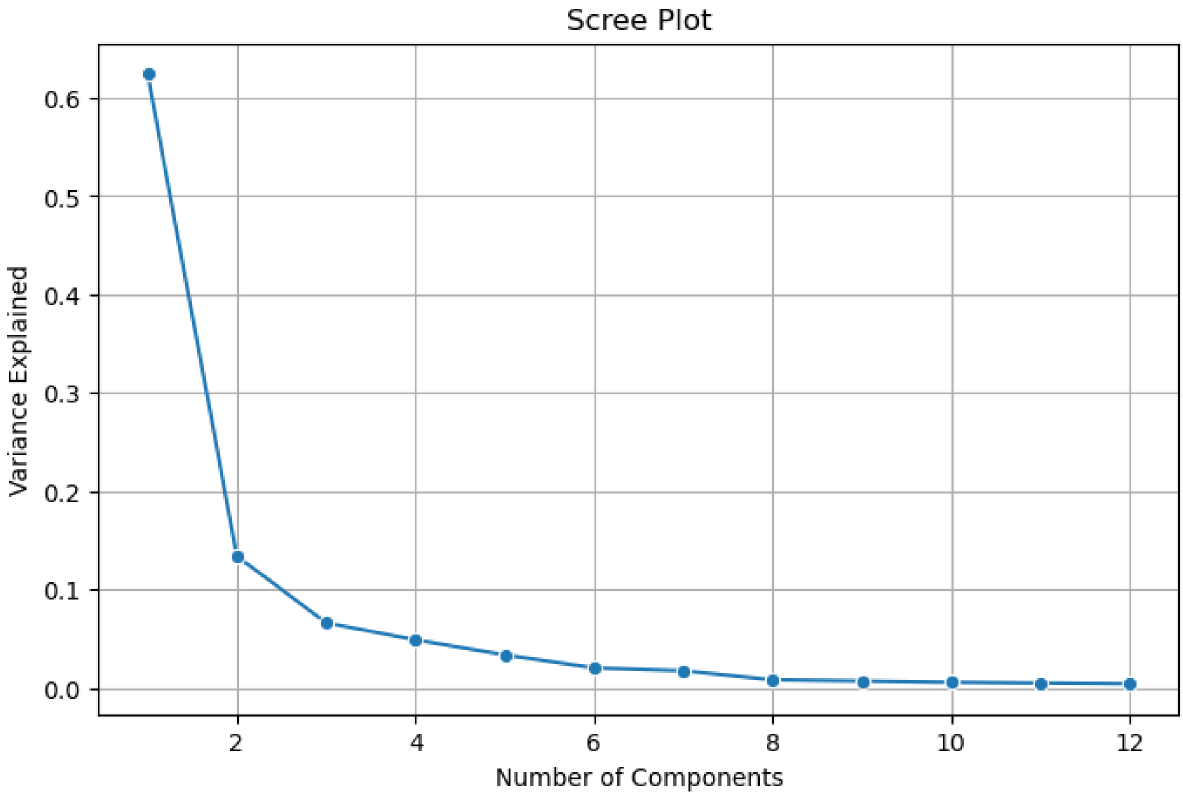
The eigen vectors of the dataset is as follows :-



```
array([[ 0.14922158,  0.15916917,  0.15820921,  0.15634043,  0.1568144 ,
         0.14335015,  0.14353705,  0.01884873,  0.01787797,  0.15515239,
         0.14544984,  0.1545511 ,  0.15828347,  0.15407627,  0.14252995,
         0.14193201,  0.12573163,  0.11169244,  0.08303496,  0.11929067,
         0.09008881,  0.14184969,  0.13388011,  0.1227618 ,  0.1168656 ,
         0.15665637,  0.14869489,  0.08816344,  0.06516026,  0.1272781 ,
         0.11588826,  0.14536607,  0.14230182,  0.15087675,  0.14801846,
         0.15790761,  0.15583101,  0.15764021,  0.1495015 ,  0.0947852 ,
         0.06715842,  0.12818439,  0.11395923,  0.14510769,  0.14102942,
         0.15092232,  0.14753416,  0.14298675,  0.13378373,  0.06296394,
         0.05674058,  0.11910165,  0.11304417,  0.14213963,  0.14136961,
         0.14762899,  0.14210263],
       [-0.11548673, -0.08023879, -0.09371751, -0.02034061, -0.01431023,
        -0.07966701, -0.08709832,  0.06910144,  0.06731586, -0.10598636,
        -0.13323356, -0.00945956, -0.02179345, -0.12091195, -0.07600253,
        -0.16669997, -0.14224991,  0.04255228,  0.09589258, -0.05334228,
        -0.07246688, -0.10183528, -0.11325661, -0.2036023 , -0.20589888,
         0.07903864,  0.10881279,  0.2715224 ,  0.27539755,  0.15657864,
         0.13504767,  0.04097368,  0.00668481, -0.07344039, -0.08836101,
```

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

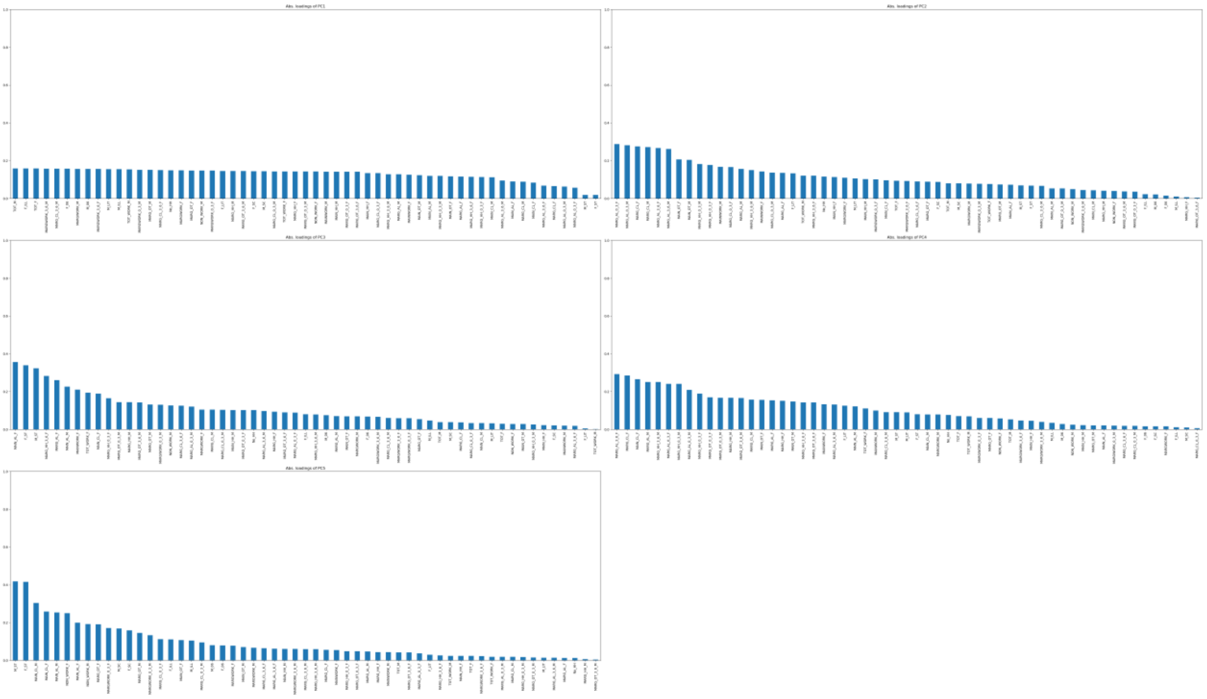
Ans. The Scree plot for optimal no of PCs is as follows :-



We will take 5 optimal no of Pc's as 5 Pc's are enough to explain the 90% of the variance

Q.7) Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.

Ans. By comparing the PCs with actual columns we will get following result :-



From above graph we can observe that most variance explain by PC1

Q.8) Write linear equation for first PC

Ans. The linear equation for first PC is as follows :-

$PC1=a1x1+a2x2+...an*xn$