PROJECT REPORT ON CLUSTERING AND PCA

SHUBHAM KUMAR 01-10-2023

Contents

CLUSTERING

PCA

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.

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.

1. Information of the dataset

```
<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
```

Table No. 1.1A
Information About the Dataset

We can see that we have 18 variables/columns. Out of 18 variables, 6 variables are float64 datatypes, 7 variables are int64 datatypes and 6 variables are object datatypes

2. First 5 rows of the dataset

	0	1	2	3	4
Timestamp	2020-9-2-17	2020-9-2-10	2020-9-1-22	2020-9-3-20	2020-9-4-15
InventoryType	Format1	Format1	Format1	Format1	Format1
Ad - Length	300	300	300	300	300
Ad- Width	250	250	250	250	250
Ad Size	75000	75000	75000	75000	75000
Ad Type	Inter222	Inter227	Inter222	Inter228	Inter217
Platform	Video	App	Video	Video	Web
Device Type	Desktop	Mobile	Desktop	Mobile	Desktop
Format	Display	Video	Display	Video	Video
Available_Impressions	1806	1780	2727	2430	1218
Matched_Queries	325	285	356	497	242
Impressions	323	285	355	495	242
Clicks	1	1	1	1	1
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	0.0035	0.0028	0.002	0.0041
CPM	0.0	0.0	0.0	0.0	0.0
CPC	0.0	0.0	0.0	0.0	0.0

Table 1.1B
First Five rows of Dataset

3. Last 5 rows of the dataset

	23061	23062	23063	23064	23065
Timestamp	2020-9-13-7	2020-11-2-7	2020-9-14-22	2020-11-18-2	2020-9-14-0
InventoryType	Format5	Format5	Format5	Format4	Format5
Ad - Length	720	720	720	120	720
Ad- Width	300	300	300	600	300
Ad Size	216000	216000	216000	72000	216000
Ad Type	Inter220	Inter224	Inter218	inter230	Inter221
Platform	Web	Web	App	Video	App
Device Type	Mobile	Desktop	Mobile	Mobile	Mobile
Format	Video	Video	Video	Video	Video
Available_Impressions	1	3	2	7	2
Matched_Queries	1	2	1	1	2
Impressions	1	2	1	1	2
Clicks	1	1	1	1	1
Spend	0.07	0.04	0.05	0.07	0.09
Fee	0.35	0.35	0.35	0.35	0.35
Revenue	0.0455	0.026	0.0325	0.0455	0.0585
CTR	NaN	NaN	NaN	NaN	NaN
CPM	NaN	NaN	NaN	NaN	NaN
CPC	NaN	NaN	NaN	NaN	NaN

Table 1.1C
Last five rows of Dataset

4. Shape of the dataset: - Number of Rows and Columns in a Dataset

```
Number of rows in dataset is  23066
Number of columns in a dataset is 19
```

5. Statistical Summary of the dataset [describe( )]

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Timestamp	23066	2018	2020-11-13-22	13	NaN	NaN	NaN	NaN	NaN	NaN	NaN
InventoryType	23066	7	Format4	7165	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Ad - Length	23066.0	NaN	NaN	NaN	385.163097	233.651434	120.0	120.0	300.0	720.0	728.0
Ad- Width	23066.0	NaN	NaN	NaN	337.896037	203.092885	70.0	250.0	300.0	600.0	600.0
Ad Size	23066.0	NaN	NaN	NaN	96674.468048	61538.329557	33600.0	72000.0	72000.0	84000.0	216000.0
Ad Type	23066	14	Inter224	1658	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Platform	23066	3	Video	9873	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Device Type	23066	2	Mobile	14806	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Format	23066	2	Video	11552	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Available_Impressions	23066.0	NaN	NaN	NaN	2432043.665872	4742887.764666	1.0	33672.25	483771.0	2527711.75	27592861.0
Matched_Queries	23066.0	NaN	NaN	NaN	1295099.143241	2512969.861258	1.0	18282.5	258087.5	1180700.0	14702025.0
Impressions	23066.0	NaN	NaN	NaN	1241519.518859	2429399.961091	1.0	7990.5	225290.0	1112428.5	14194774.0
Clicks	23066.0	NaN	NaN	NaN	10678.518816	17353.409363	1.0	710.0	4425.0	12793.75	143049.0
Spend	23066.0	NaN	NaN	NaN	2706.625689	4067.927273	0.0	85.18	1425.125	3121.4	26931.87
Fee	23066.0	NaN	NaN	NaN	0.335123	0.031963	0.21	0.33	0.35	0.35	0.35
Revenue	23066.0	NaN	NaN	NaN	1924.252331	3105.23841	0.0	55.365375	926.335	2091.33815	21276.18
CTR	18330.0	NaN	NaN	NaN	0.073661	0.07516	0.0001	0.0026	0.08255	0.13	1.0
CPM	18330.0	NaN	NaN	NaN	7.672045	6.481391	0.0	1.71	7.66	12.51	81.56
CPC	18330.0	NaN	NaN	NaN	0.351061	0.343334	0.0	0.09	0.16	0.57	7.26

Table 1.1D

Descriptive Statistical Summary of Dataset

- By using Describe ( ) function we get 5 point summary of the dataset i.e., minimum, 25%, 50%, 75% and maximum . We can see that the scale of variables is different.
- Also we get the statistical summary of the data i.e., mean, median and mode.
- We can also see that different variables have different scale. So we need to preform Scaling.
- We can also see that there are NAN values in some of the variables which tell us that there are some null entries in the dataset which we need to check.

6. **Check for Missing values in the Dataset**

```
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
```

Table 1.1E
Missing Value in Dataset

- There are certain null entries present in the 3 variables/columns i.e., CTR, CPM, CPC.
- We need to treat these Missing values
- Before treating a missing value, we need to check if our dataset contains duplicates.

7. **Check for Duplicates**

There are no duplicated data in the dataset.

```
Dupliacted rows/data in a dataset = 0
```

Table 1.1F
Duplicated rows in a Dataset

Part 1.2 - Clustering: Treat missing values in CPC, CTR and CPM using the formula given.

Treating Missing Value

it is given that we need to certain formulas while imputing these missing Values i.e.,

1. CPM = (Total Campaign Spend / Number of Impressions) * 1,000
2. CPC = Total Cost (spend) / Number of Clicks.
3. CTR = Total Measured Clicks / Total Measured Ad Impressions x 100

Here we have defined a function for each variable i.e., CPM, CPC, CTR. For imputing the missing values, we have used Lamda function.

Summary of the dataset after imputing the missing values

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	385.163	233.651	120.000	120.000	300.000	720.000	728.000
Ad- Width	23066.0	337.896	203.093	70.000	250.000	300.000	600.000	600.000
Ad Size	23066.0	79339.343	12407.166	65520.000	72000.000	72000.000	84000.000	102000.000
Available_Impressions	23066.0	1607252.772	2125527.927	1.000	33672.250	483771.000	2527711.750	6268771.000
Matched_Queries	23066.0	799538.035	1026036.789	1.000	18282.500	258087.500	1180700.000	2924326.250
Impressions	23066.0	753611.989	980256.808	1.000	7990.500	225290.000	1112428.500	2769085.500
Clicks	23066.0	8306.828	9574.779	1.000	710.000	4425.000	12793.750	30919.375
Spend	23066.0	2166.060	2425.190	0.000	85.180	1425.125	3121.400	7675.730
Fee	23066.0	0.350	0.000	0.350	0.350	0.350	0.350	0.350
Revenue	23066.0	1449.389	1646.894	0.000	55.365	926.335	2091.338	5145.297
CTR	23066.0	7.993	9.137	0.011	0.397	7.096	12.760	200.000
CPM	23066.0	8.448	8.955	0.000	1.917	8.372	12.841	715.000
CPC	23066.0	0.316	0.309	0.000	0.092	0.186	0.476	7.264

Table 1.2A

Descriptive Statistic Summary of Dataset after treating Missing Values

Before treating the missing values maximum of the CTR, CPM, CPC is 1, 81.56, 7.26 and after treating the missing values the maximum values has changed to 200, 715, 7.26 respectively.

Also we can see that 75% of dataset in column CTR, CPM, CPC is below 12.760, 12.841, 0.476 whereas the maximum value is 200, 715, 7.264 which gives us the indication of the outliers in the dataset.

Part 1.3 - Clustering: Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an analyst your judgement may be different from another analyst).

Checking for Outliers: -

Here to check the outliers we have plotted the boxplot of all the numeric variables

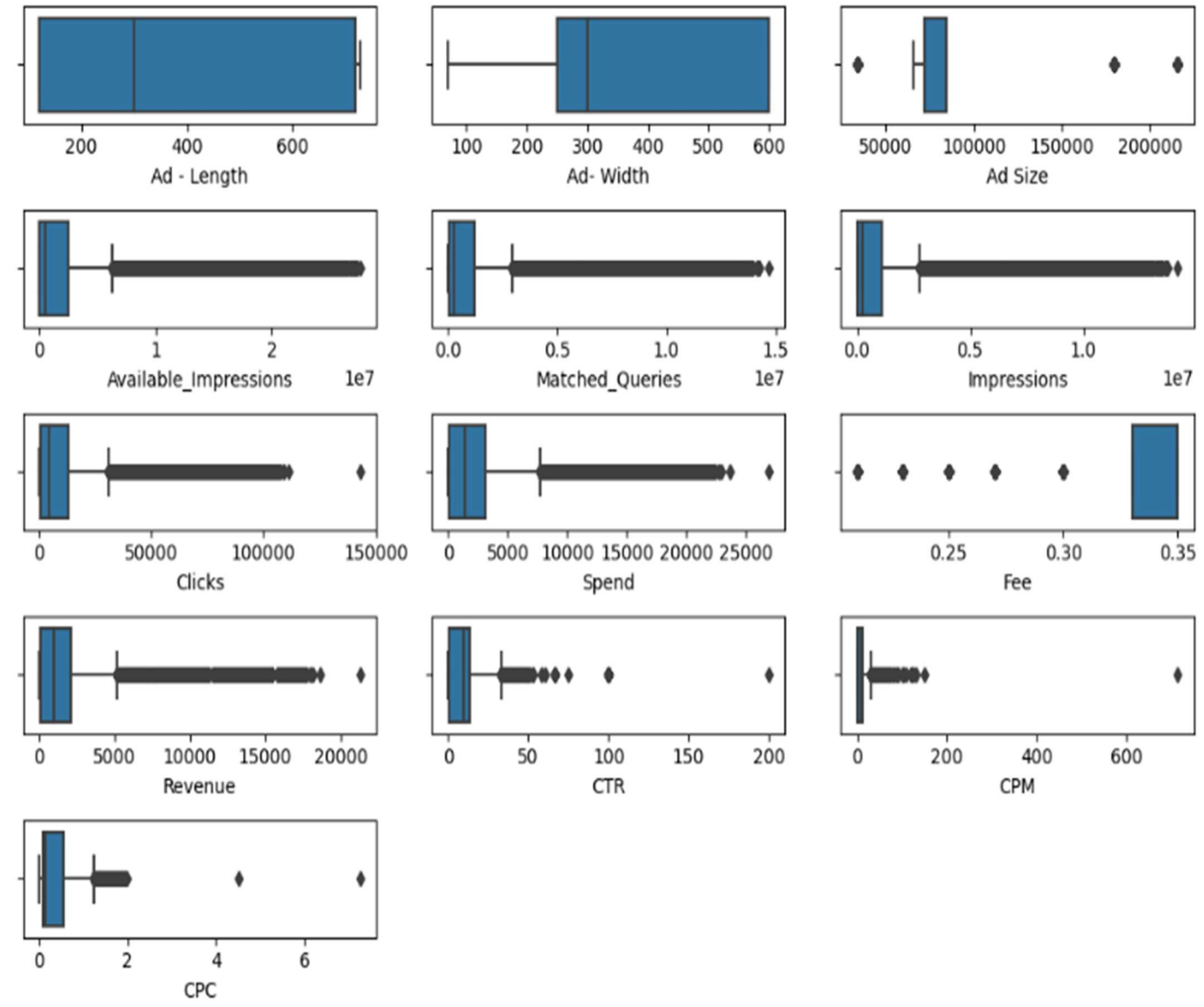


Fig. 1.3A
Boxplot of all the Variables/Columns of the Dataset

As from the figure 1.2A we can see that there are outliers in our dataset. All the numeric variables except Ad – Length and Ad – Width have outliers present in them.

There are two ways to check Outliers: -

1. Using Inter Quartile Range (IQR)
2. Using +- 3 times the Standard Deviation

Interquartile Range is used when we have extreme outliers whereas standard deviation is used when we don't have extreme outliers.

In Descriptive Summary of the data we can see there is large difference in the values between 75% data and Maximum. So we will use IQR technique in order to find the Outliers.

IQR = Q3 – Q1 ; Q3= 75% of data Q1= 25% of data

IQR technique define Upper Fence and Lower Fence with respect to data.

Lower Fence = Q1 – 1.5*IQR Upper Fence = Q3 + 1.5*IQR

The values below Lower Fence and above Upper Fence are treated as Outliers.

```
Ad - Length :- Lower Fence = -780.0 & Upper Fence =  1620.0
Min Value for Ad - Length is  120 whereas Max value for Ad - Length is  728
Number of Outliers = 0

Ad- Width :- Lower Fence = -275.0 & Upper Fence =  1125.0
Min Value for Ad- Width is  70 whereas Max value for Ad- Width is  600
Number of Outliers = 0

Ad Size :- Lower Fence = 54000.0 & Upper Fence =  102000.0
Min Value for Ad Size is  33600 whereas Max value for Ad Size is  216000
Number of Outliers = 4908

Available_Impressions :- Lower Fence = -3707387.0 & Upper Fence =  6268771.0
Min Value for Available_Impressions is  1 whereas Max value for Available_Impressions is  27592861
Number of Outliers = 2378

Matched_Queries :- Lower Fence = -1725343.75 & Upper Fence =  2924326.25
Min Value for Matched_Queries is  1 whereas Max value for Matched_Queries is  14702025
Number of Outliers = 3192

Impressions :- Lower Fence = -1648666.5 & Upper Fence =  2769085.5
Min Value for Impressions is  1 whereas Max value for Impressions is  14194774
Number of Outliers = 3269

Clicks :- Lower Fence = -17415.625 & Upper Fence =  30919.375
Min Value for Clicks is  1 whereas Max value for Clicks is  143049
Number of Outliers = 1691

Spend :- Lower Fence = -4469.15 & Upper Fence =  7675.73
Min Value for Spend is  0.0 whereas Max value for Spend is  26931.87
Number of Outliers = 2081

Fee :- Lower Fence = 0.3 & Upper Fence =  0.38
Min Value for Fee is  0.21 whereas Max value for Fee is  0.35
Number of Outliers = 0

Revenue :- Lower Fence = -2998.5938 & Upper Fence =  5145.2973
Min Value for Revenue is  0.0 whereas Max value for Revenue is  21276.18
Number of Outliers = 2325

CTR :- Lower Fence = -19.5431 & Upper Fence =  33.2788
Min Value for CTR is  0.01 whereas Max value for CTR is  200.0
Number of Outliers = 275

CPM :- Lower Fence = -15.1903 & Upper Fence =  29.9814
Min Value for CPM is  0.0 whereas Max value for CPM is  715.0
Number of Outliers = 207

CPC :- Lower Fence = -0.595 & Upper Fence =  1.231
Min Value for CPC is  0.0 whereas Max value for CPC is  7.26
Number of Outliers = 585
```

Table 1.3A
Upper Fence/Lower Fence & Min/Max & number of Outliers in each variables

Treating these outliers: -

For Treating these outliers, we have defined a function ‘remove_outliers’ : -

- Where all the values which are above the Upper Fence are replaced with Upper Fence value of that column.
- Where all the values which are below Lower Fence are replaced with Lower Fence value of that column.

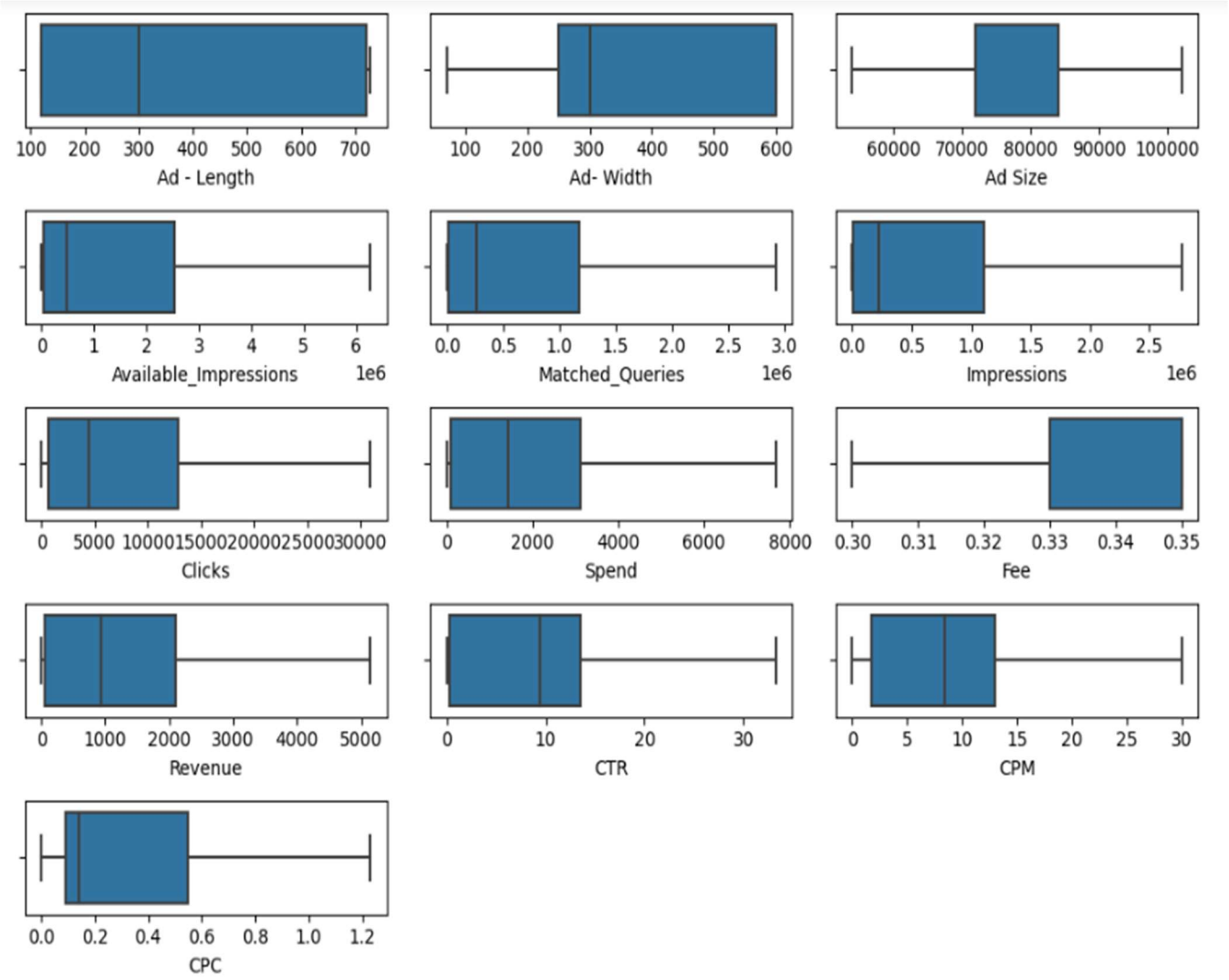


Fig. 1.3B
Boxplot Plot after Treating Outliers

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	385.16	233.65	120.00	120.00	300.00	720.00	728.00
Ad- Width	23066.0	337.90	203.09	70.00	250.00	300.00	600.00	600.00
Ad Size	23066.0	76576.84	15381.32	54000.00	72000.00	72000.00	84000.00	102000.00
Available_Impressions	23066.0	1607252.77	2125527.93	1.00	33672.25	483771.00	2527711.75	6268771.00
Matched_Queries	23066.0	799538.04	1026036.79	1.00	18282.50	258087.50	1180700.00	2924326.25
Impressions	23066.0	753611.99	980256.81	1.00	7990.50	225290.00	1112428.50	2769085.50
Clicks	23066.0	8306.83	9574.78	1.00	710.00	4425.00	12793.75	30919.38
Spend	23066.0	2166.06	2425.19	0.00	85.18	1425.12	3121.40	7675.73
Fee	23066.0	0.34	0.02	0.30	0.33	0.35	0.35	0.35
Revenue	23066.0	1449.39	1646.89	0.00	55.37	926.34	2091.34	5145.30
CTR	23066.0	8.22	8.25	0.01	0.27	9.39	13.47	33.28
CPM	23066.0	8.22	6.88	0.00	1.75	8.37	13.04	29.98
CPC	23066.0	0.33	0.32	0.00	0.09	0.14	0.55	1.23

Table 1.3B
Descriptive Statistic Summary of Dataset after treating Outliers

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

As we can see in the descriptive statistics of the dataset different variables have different scale, so we need to do scaling of our dataset in order to bring our dataset to same level.

Scaling is import for clustering. If scaling is not done than the model will consider the column with high scale as important data and cluster as per that column.

For scaling of dataset we have used 'sklearn' library

From sklearn.preprocessing we have imported StandardScaler ( ) to perform scaling

StandardScaler function reduce the scale of all the column which has Mean 0 & Standard Deviation 1

	0	1	2	3	4	5	6	7	8	9	...	23056	23057
Ad - Length	-0.364496	-0.364496	-0.364496	-0.364496	-0.364496	-0.364496	-0.364496	-0.364496	-0.364496	-0.364496	...	-1.134891	-1.134891
Ad- Width	-0.432797	-0.432797	-0.432797	-0.432797	-0.432797	-0.432797	-0.432797	-0.432797	-0.432797	-0.432797	...	1.290590	1.290590
Ad Size	-0.102518	-0.102518	-0.102518	-0.102518	-0.102518	-0.102518	-0.102518	-0.102518	-0.102518	-0.102518	...	-0.297564	-0.297564
Available_Impressions	-0.755333	-0.755345	-0.754900	-0.755040	-0.755610	-0.755952	-0.755620	-0.755542	-0.755523	-0.755328	...	-0.756182	-0.756180
Matched_Queries	-0.778949	-0.778988	-0.778919	-0.778781	-0.779030	-0.779203	-0.779069	-0.779073	-0.779132	-0.778962	...	-0.779264	-0.779264
Impressions	-0.768478	-0.768516	-0.768445	-0.768302	-0.768560	-0.768742	-0.768601	-0.768607	-0.768668	-0.768490	...	-0.768805	-0.768805
Clicks	-0.867488	-0.867488	-0.867488	-0.867488	-0.867488	-0.867384	-0.867488	-0.867488	-0.867488	-0.867488	...	-0.867488	-0.867488
Spend	-0.893170	-0.893170	-0.893170	-0.893170	-0.893170	-0.893170	-0.893166	-0.893170	-0.893170	-0.893170	...	-0.893129	-0.893141
Fee	0.535724	0.535724	0.535724	0.535724	0.535724	0.535724	0.535724	0.535724	0.535724	0.535724	...	0.535724	0.535724
Revenue	-0.880093	-0.880093	-0.880093	-0.880093	-0.880093	-0.880093	-0.880093	-0.880093	-0.880093	-0.880093	...	-0.880054	-0.880066
CTR	-0.958836	-0.953835	-0.962218	-0.971871	-0.946281	-0.617714	-0.936366	-0.934530	-0.907258	-0.957389	...	3.035808	3.035808
CPM	-1.194498	-1.194498	-1.194498	-1.194498	-1.194498	-1.194498	-1.187303	-1.194498	-1.194498	-1.194498	...	3.162718	3.162718
CPC	-1.042561	-1.042561	-1.042561	-1.042561	-1.042561	-1.042561	-1.010972	-1.042561	-1.042561	-1.042561	...	-0.726667	-0.821435

Table 1.4A
Dataset after Scaling

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	0.0	1.0	-1.1349	-1.1349	-0.3645	1.4331	1.4673
Ad- Width	23066.0	-0.0	1.0	-1.3191	-0.4328	-0.1866	1.2906	1.2906
Ad Size	23066.0	0.0	1.0	-1.4678	-0.2976	-0.2976	0.4826	1.6529
Available_Impressions	23066.0	0.0	1.0	-0.7562	-0.7403	-0.5286	0.4331	2.1932
Matched_Queries	23066.0	0.0	1.0	-0.7793	-0.7614	-0.5277	0.3715	2.0709
Impressions	23066.0	0.0	1.0	-0.7688	-0.7607	-0.5390	0.3661	2.0561
Clicks	23066.0	-0.0	1.0	-0.8675	-0.7934	-0.4054	0.4686	2.3617
Spend	23066.0	-0.0	1.0	-0.8932	-0.8580	-0.3055	0.3939	2.2719
Fee	23066.0	0.0	1.0	-2.2224	-0.5675	0.5357	0.5357	0.5357
Revenue	23066.0	0.0	1.0	-0.8801	-0.8465	-0.3176	0.3898	2.2442
CTR	23066.0	0.0	1.0	-0.9950	-0.9642	0.1415	0.6358	3.0358
CPM	23066.0	0.0	1.0	-1.1945	-0.9403	0.0221	0.7009	3.1627
CPC	23066.0	0.0	1.0	-1.0426	-0.7591	-0.6024	0.6830	2.8461

Table 1.4B
Descriptive Statistic Summary of Scaled Dataset

We can see that after scaling the numeric dataset all the variables have been scaled to same scale of which Mean is 0 and Standard Deviation 1.

When all the variables are scaled to same level, the model give equal weightage to each variables and also increase the speed of the model

Part 1.5 - Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.

Hierarchical Clustering is a clustering technique which it starts with n clusters (n=sample size /population) and then iteratively combines the closest clusters until a stopping criterion is reached.

To perform hierarchical clustering we need to import it from 'scipy' library i.e., scipy.cluster.hierarchy

Scipy.cluster,hierarchy we need to import two import function: -

1. Linkage
2. Dendrogram

Linkage is used in order to provide the linking technique in order to form a cluster.

Dendrogram is a tree like structure which summarizes the process of clustering

It is given that we have to use WARS and Euclidean distance method to construct Dendrogram.

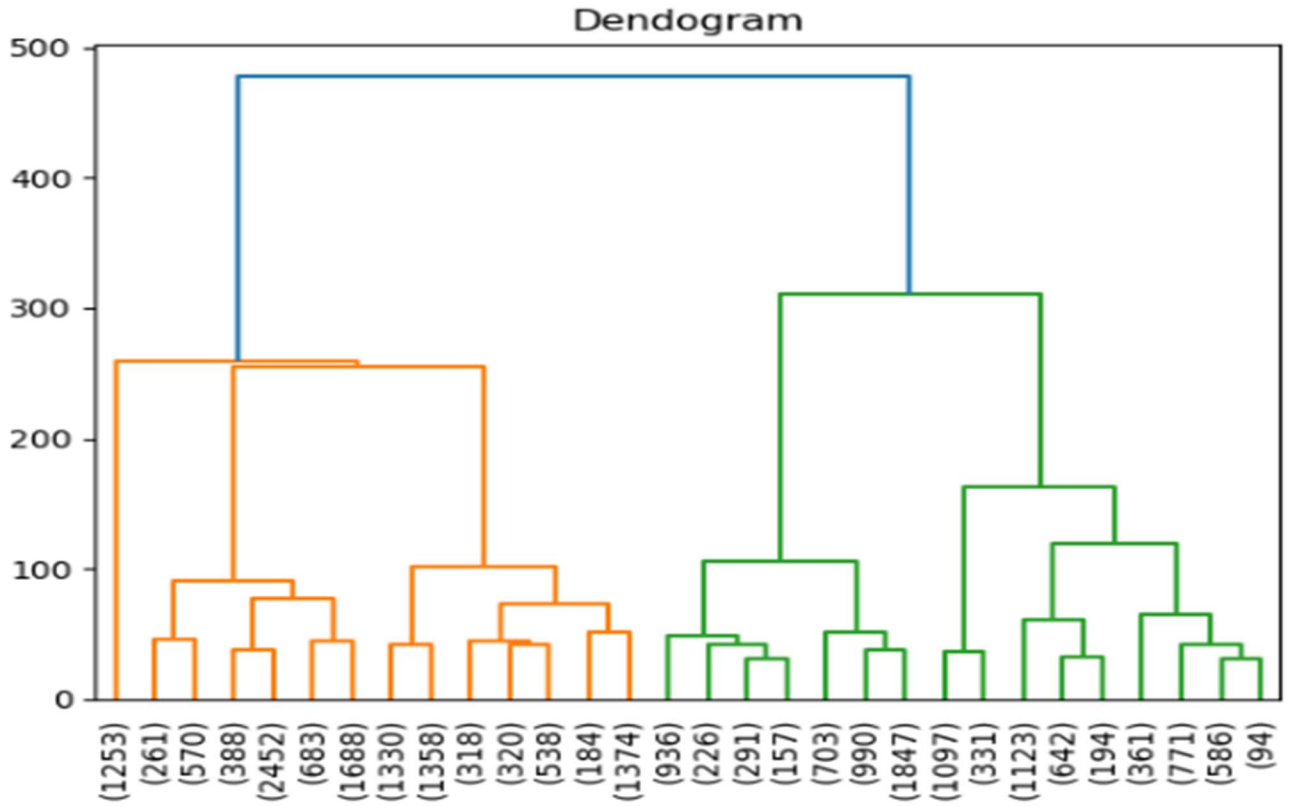


Fig 1.5A
Dendrogram with 30 Clusters

In above Dendrogram plot with 30 clusters where each color (Blue, Orange, Green) represent a cluster and each color have several sub clusters.

- Blue Represents 02 clusters
- Orange represents 13 clusters
- Green represents 15 clusters

We can also divide the cluster based on the cutoff distance.

For example, we want to set the cutoff distance to be 240 or 280

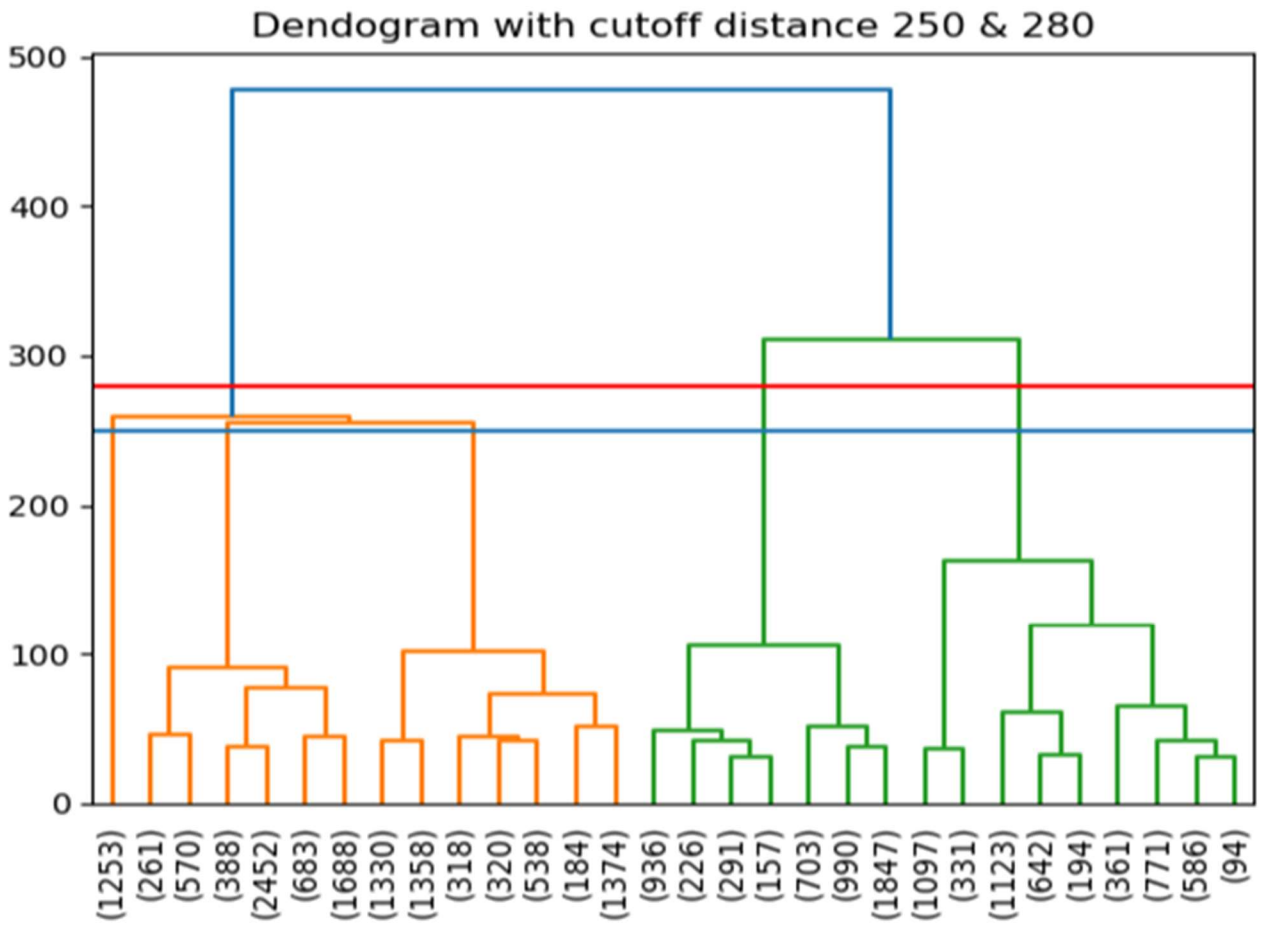


Fig 1.5B

Dendrogram with Cutoff Distance 250 & 280

The number of clusters formed is equal to numbers of lines crosses when a straight line is drawn from that cutoff distance. i.e.,

- Cutoff distance 250: - 5 clusters
- Cutoff distance 280: - 3 clusters

Part 1.6 - Clustering: Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.

In k-means clustering algorithm the clusters are homogeneous within the cluster and highly heterogeneous between the clusters. Here we need to pre-specify a desired number of clusters

Here it is specified to check the Within Sum of Squares of cluster ranging from 1 to 10

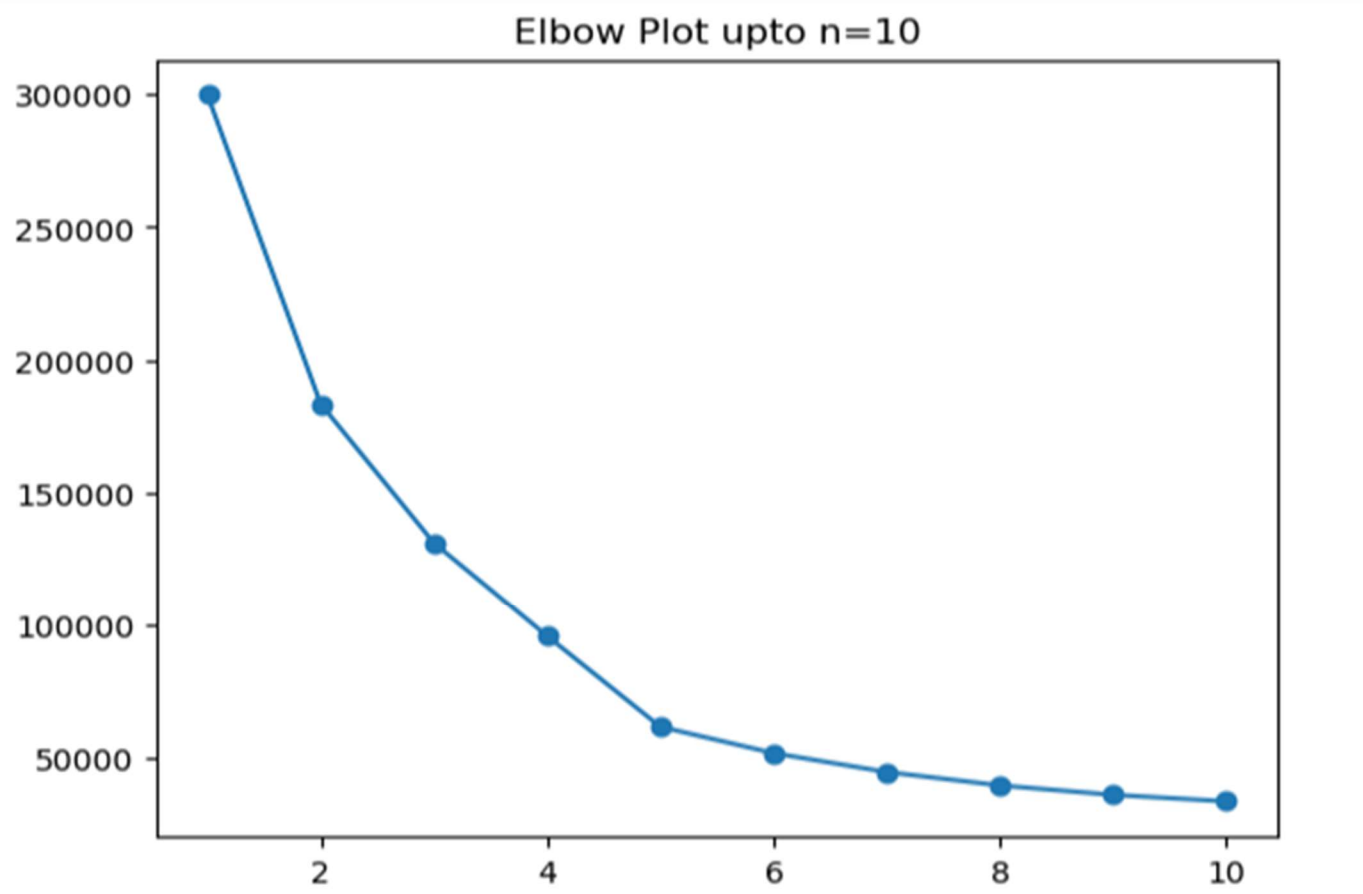


Fig 1.6A
Elbow Plot(n=10)

From the plot it can identified that elbow is forming at 5 , hence we can say that optimum number cluster after performing k means clustering is 5 for the given dataset.

Part 1.7 - Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters.

Silhouette score is used to analyze if our observation/records/rows has been clustered to correct cluster i.e., cluster centroid nearest to its observation.
If silhouette score is positive value than the mapping of observation is correct.
If silhouette score is negative value than mapping of observation is incorrect.

To check silhouette score we need to import 'sklearn.metrices' library

```
For n_clusters=2, the silhouette score is 0.38572769619101077
For n_clusters=3, the silhouette score is 0.3825486036570082
For n_clusters=4, the silhouette score is 0.45324270552598256
For n_clusters=5, the silhouette score is 0.5240956940501831
For n_clusters=6, the silhouette score is 0.5221533662938636
For n_clusters=7, the silhouette score is 0.5165635029478517
For n_clusters=8, the silhouette score is 0.47972249893837277
For n_clusters=9, the silhouette score is 0.4320636564025043
For n_clusters=10, the silhouette score is 0.43124854581084165
```

Table 1.7

Silhouette Scores (n_clusters= 2 to 10)

Silhouette score is maximum for 5 clusters. So we will consider 5 clusters.

Part 1.8 - Clustering: Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots].

We have a new column 'Clusters' to our dataset.
Observations:-

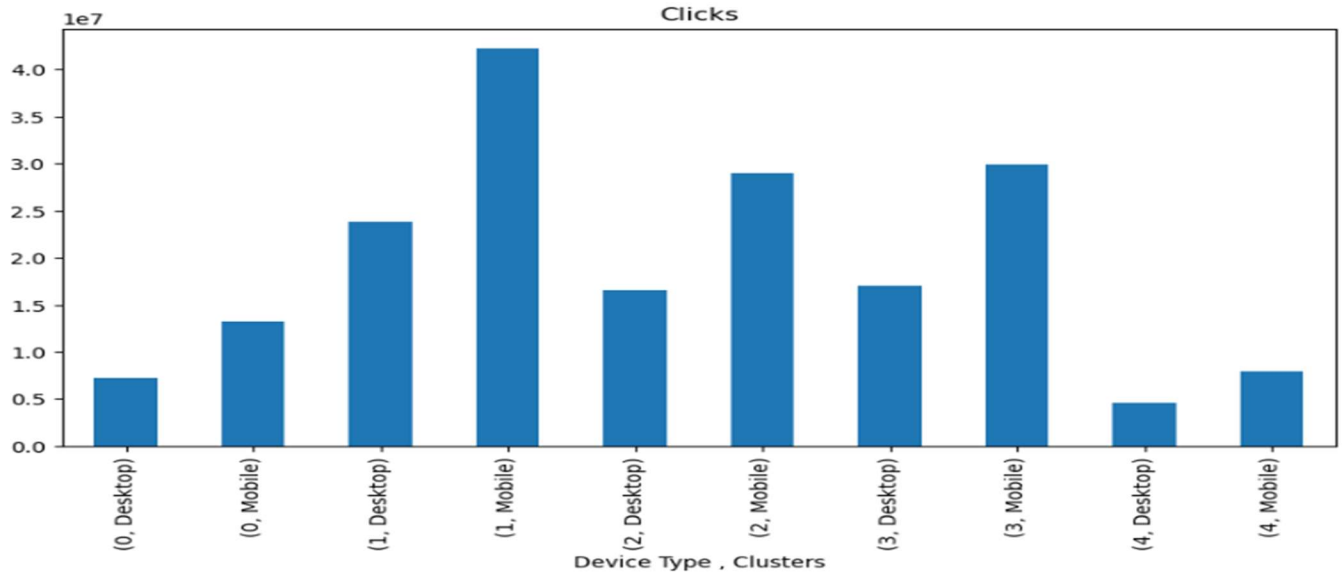


Fig 1.8A

Bar Plot of Clicks in each Cluster on Desktop & Mobile

- **In Desktop as well as Mobile Cluster 1 has maximum Clicks**
- **In Desktop as well as Mobile Cluster 4 has minimum Clicks**

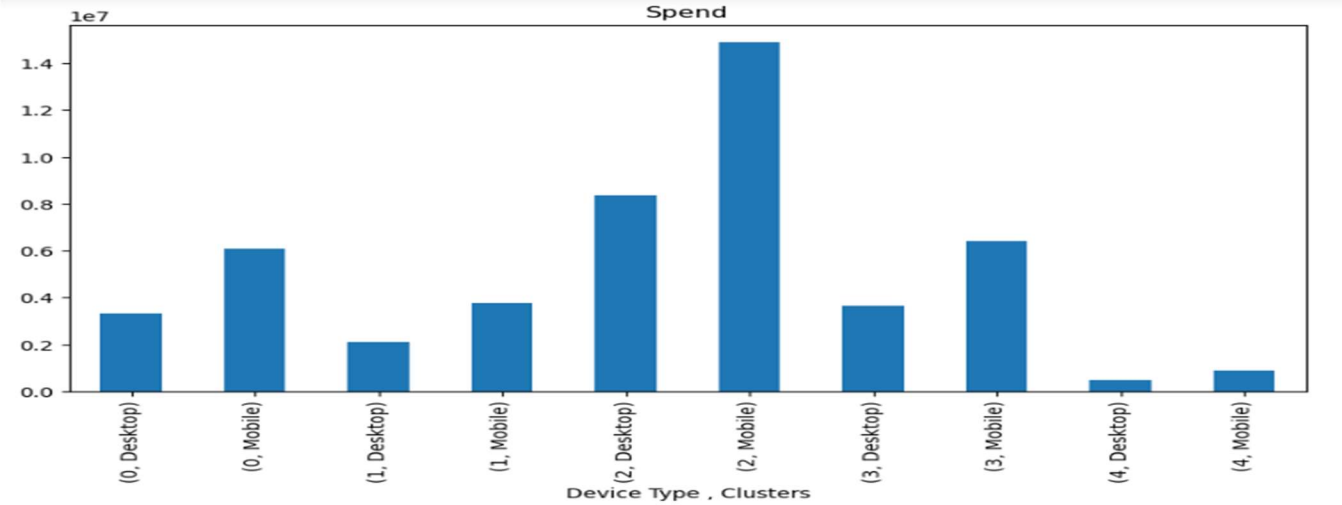


Fig 1.8B

Bar Plot of Spends in each Cluster on Desktop & Mobile

- **In Desktop as well as mobile Cluster 2 has maximum Spends**
- **In Desktop as well as mobile Cluster 4 has minimum Spends**

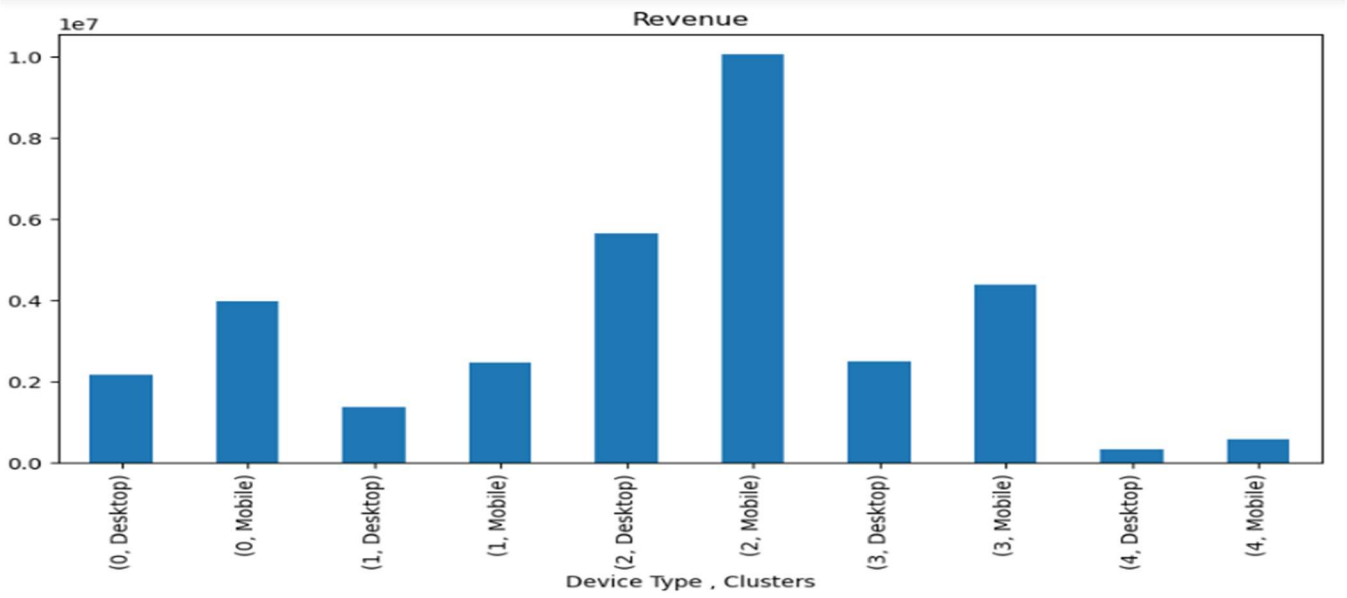


Fig 1.8C

Bar Plot of Revenue in each Cluster on Desktop & Mobile

- **In Desktop as well as mobile Cluster 2 has maximum Revenue**
- **In Desktop as well as mobile Cluster 4 has minimum Revenue**

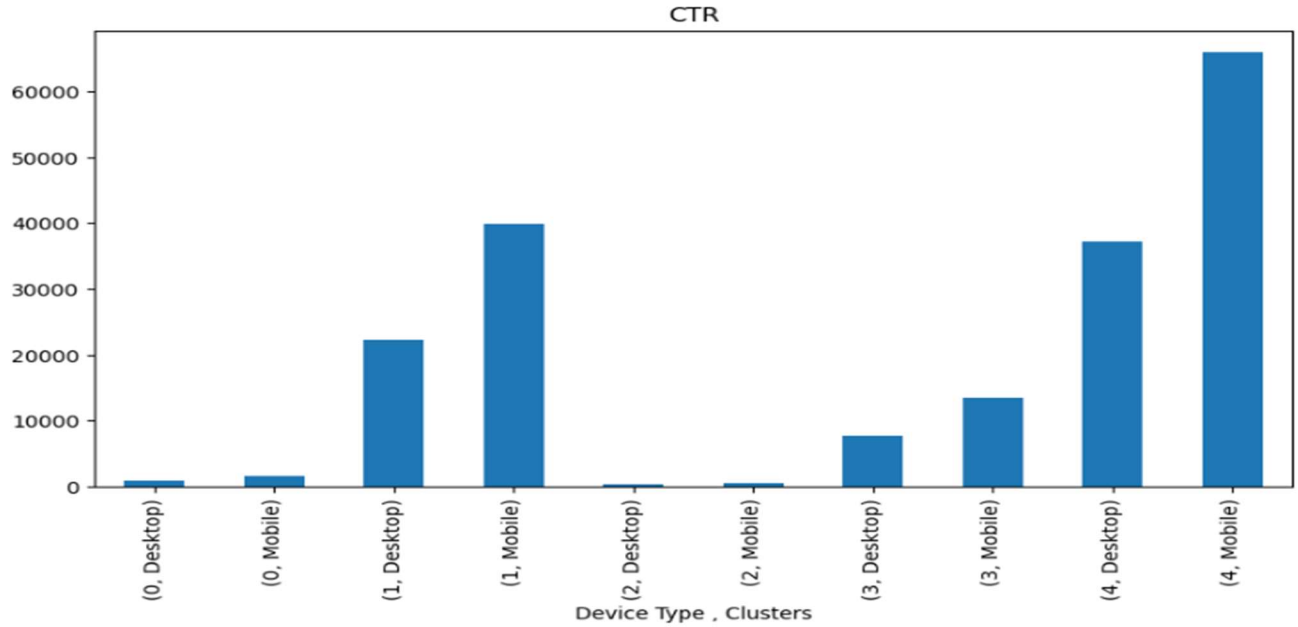


Fig 1.8D

Count Plot of CTR in each Clusters

- **In Desktop as well as mobile Cluster 4 has maximum CTR**
- **In Desktop as well as mobile Cluster 2 has minimum CTR**

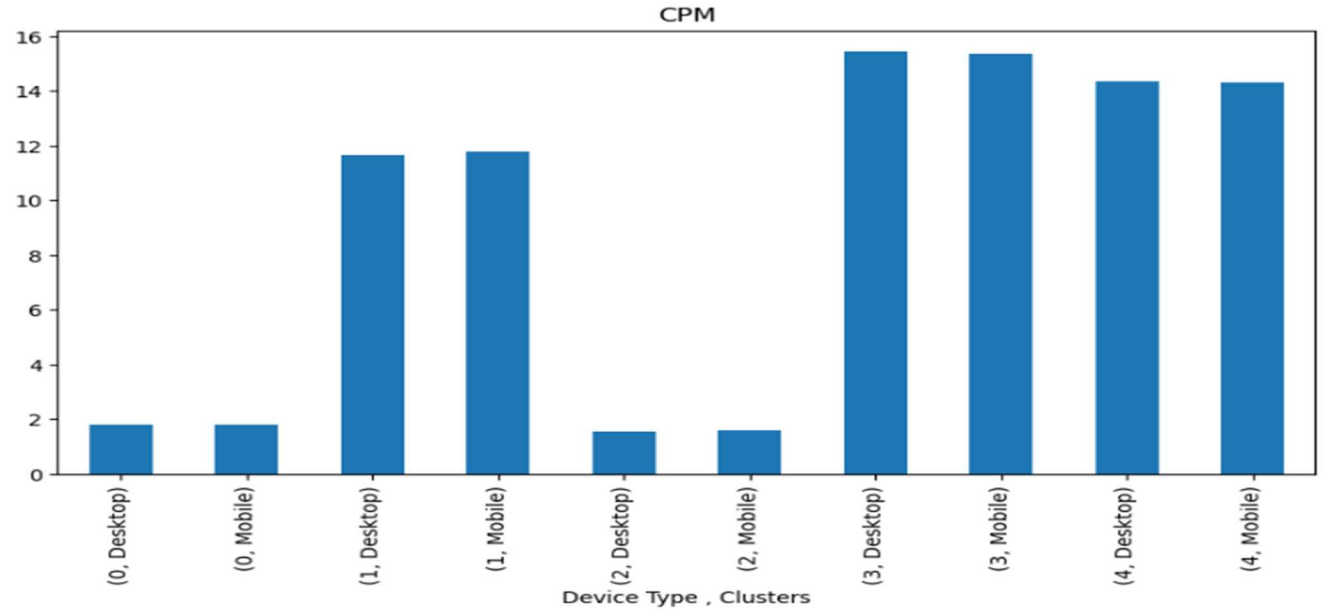


Fig 1.8E

Count Plot of CPM in each Clusters

- **In Desktop as well as mobile Cluster 3 has maximum CPM**
- **In Desktop as well as mobile Cluster 2 has minimum CPM**

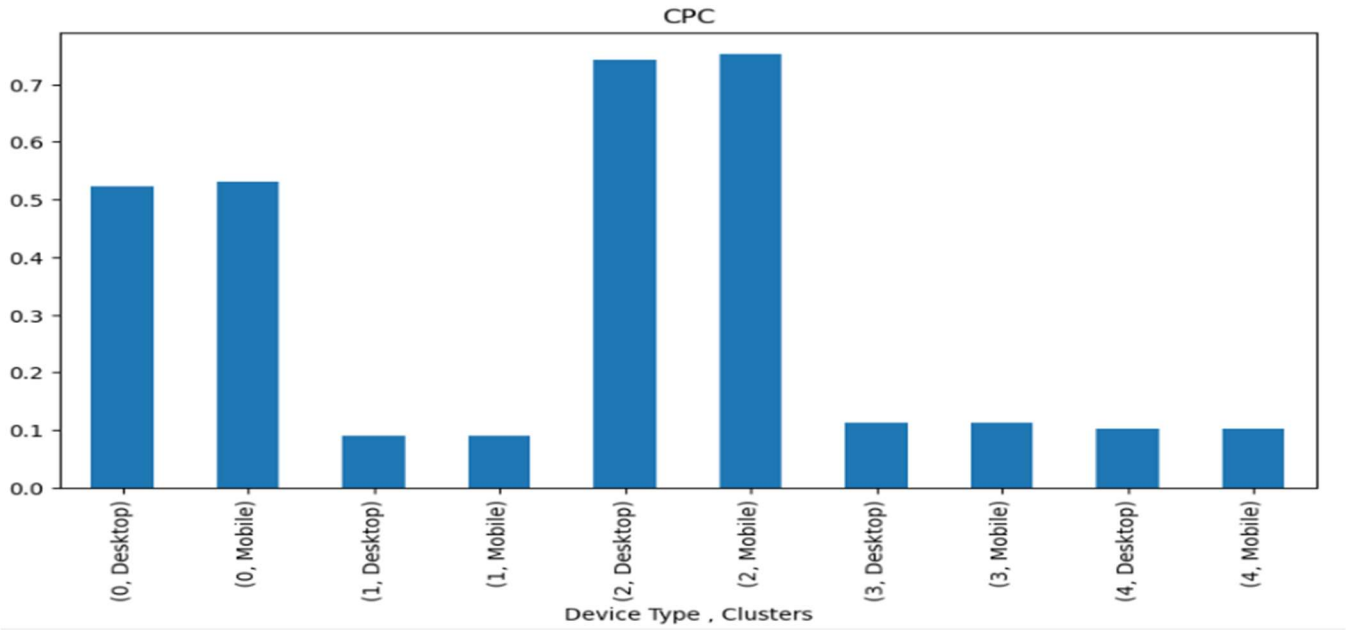


Fig 1.8F

Count Plot of CPC in each Cluster

- **In Desktop as well as mobile Cluster 2 has maximum CPC**
- **In Desktop as well as mobile Cluster 1 has minimum CPC**

Part 1.9 - Clustering: Conclude the project by providing summary of your learnings.

Clustering is an Unsupervised Learning technique where unlabeled data or data points are divided into different clusters such that the similar data points falls in the same cluster than those which differ from the others.

Basically dataset is divided into the clusters such that each cluster is highly homogeneous within the cluster and heterogeneous between the clusters.

There are mainly two type of Clustering Techniques: -

1. K-means Clustering: - It works better when dataset is large. Mostly we prefer to use K-means clustering as it form good clusters which are highly homogeneous within the cluster and heterogeneous between the clusters. Hereby we need to pre-define the number of clusters.

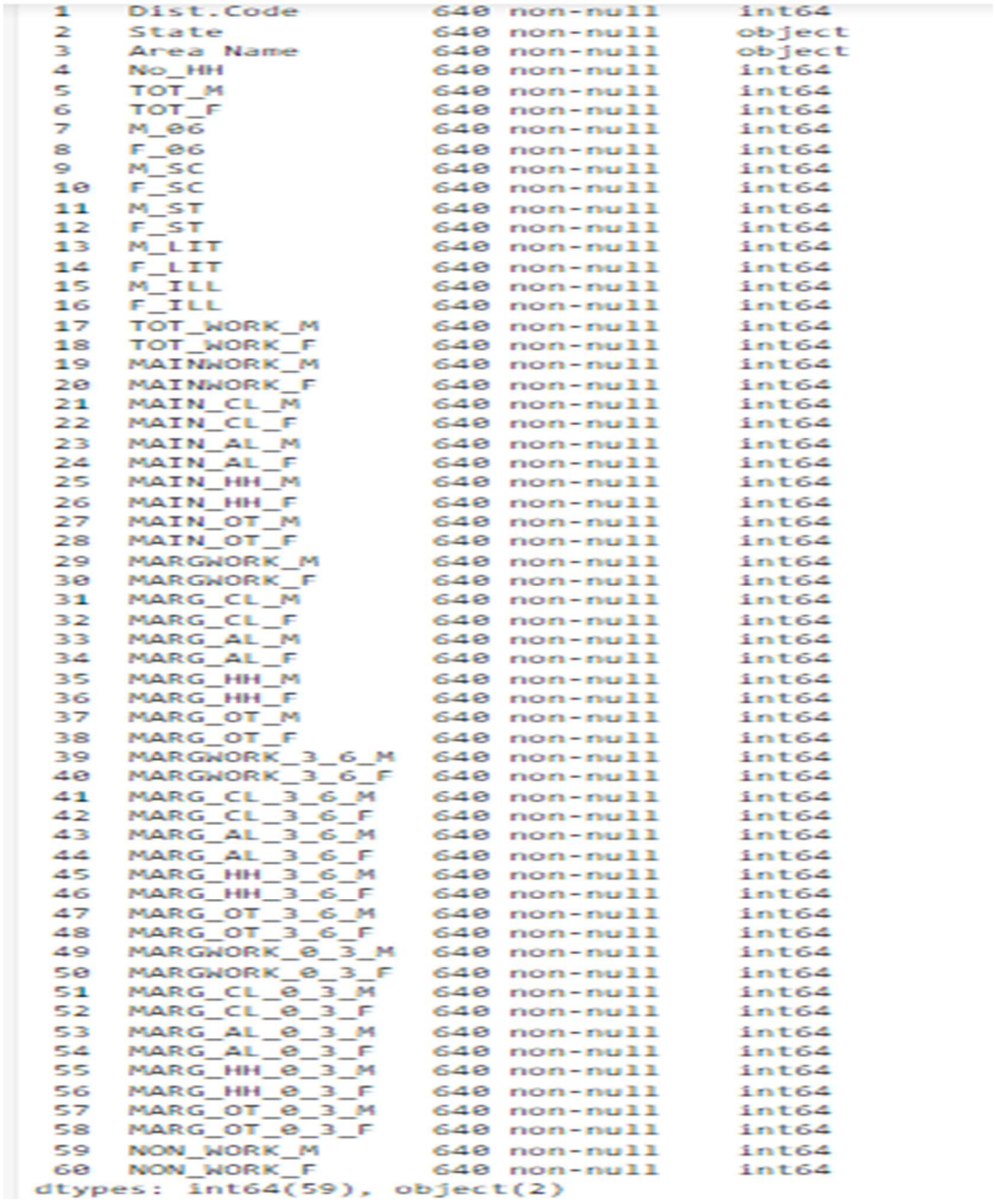
2. Hierarchical Clustering: - It is works better when dataset is not too large. It starts with n clusters where each row/record is a cluster and two closet point forms a cluster, then again two closest points or cluster & point for a cluster, and so on.

Conclusion from the Dataset: -

1. The dataset has 25857 rows and 19 columns
2. The missing values in CPC, CTR and CPM are treated by using the formulae given and writing a user-defined function, and calling it.
3. We check for outliers; we can see there are outliers in the variables.
4. Dendrogram is the visualization and linkage is for computing the distances and merging the clusters from n to 1.
5. The output of Linkage is visualized by Dendrogram
6. We will create linkage using Ward's method and run linkage function on the usable columns of the data
7. Then we used k-means clustering in order to decide a number of cluster
8. We ploted an elbow plot (n=10) and tried to find the number of clusters i.e., 5
9. Then we calculated Silhouette Score of these 10 k_means clusters and found the largest value of silhouette score to be when n_clusters=5. Hence we divided our dataset into 5 Clusters
10. Then we compared the dataset based on the clusters and found some meaningful inferences from the data

Part 2.1 - PCA: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.

1. Information of the dataset



1	Dist.Code	640	non-null	int64
2	State	640	non-null	object
3	Area Name	640	non-null	object
4	No_HH	640	non-null	int64
5	TOT_M	640	non-null	int64
6	TOT_F	640	non-null	int64
7	M_06	640	non-null	int64
8	F_06	640	non-null	int64
9	M_SC	640	non-null	int64
10	F_SC	640	non-null	int64
11	M_ST	640	non-null	int64
12	F_ST	640	non-null	int64
13	M_LIT	640	non-null	int64
14	F_LIT	640	non-null	int64
15	M_ILL	640	non-null	int64
16	F_ILL	640	non-null	int64
17	TOT_WORK_M	640	non-null	int64
18	TOT_WORK_F	640	non-null	int64
19	MAINWORK_M	640	non-null	int64
20	MAINWORK_F	640	non-null	int64
21	MAIN_CL_M	640	non-null	int64
22	MAIN_CL_F	640	non-null	int64
23	MAIN_AL_M	640	non-null	int64
24	MAIN_AL_F	640	non-null	int64
25	MAIN_HH_M	640	non-null	int64
26	MAIN_HH_F	640	non-null	int64
27	MAIN_OT_M	640	non-null	int64
28	MAIN_OT_F	640	non-null	int64
29	MARGWORK_M	640	non-null	int64
30	MARGWORK_F	640	non-null	int64
31	MARG_CL_M	640	non-null	int64
32	MARG_CL_F	640	non-null	int64
33	MARG_AL_M	640	non-null	int64
34	MARG_AL_F	640	non-null	int64
35	MARG_HH_M	640	non-null	int64
36	MARG_HH_F	640	non-null	int64
37	MARG_OT_M	640	non-null	int64
38	MARG_OT_F	640	non-null	int64
39	MARGWORK_3_6_M	640	non-null	int64
40	MARGWORK_3_6_F	640	non-null	int64
41	MARG_CL_3_6_M	640	non-null	int64
42	MARG_CL_3_6_F	640	non-null	int64
43	MARG_AL_3_6_M	640	non-null	int64
44	MARG_AL_3_6_F	640	non-null	int64
45	MARG_HH_3_6_M	640	non-null	int64
46	MARG_HH_3_6_F	640	non-null	int64
47	MARG_OT_3_6_M	640	non-null	int64
48	MARG_OT_3_6_F	640	non-null	int64
49	MARGWORK_0_3_M	640	non-null	int64
50	MARGWORK_0_3_F	640	non-null	int64
51	MARG_CL_0_3_M	640	non-null	int64
52	MARG_CL_0_3_F	640	non-null	int64
53	MARG_AL_0_3_M	640	non-null	int64
54	MARG_AL_0_3_F	640	non-null	int64
55	MARG_HH_0_3_M	640	non-null	int64
56	MARG_HH_0_3_F	640	non-null	int64
57	MARG_OT_0_3_M	640	non-null	int64
58	MARG_OT_0_3_F	640	non-null	int64
59	NON_WORK_M	640	non-null	int64
60	NON_WORK_F	640	non-null	int64

dtypes: int64(59), object(2)

Table 2.1A

Information of PCA India Data Census

We can see that we have 61 variables /columns. Out of 61 variables, 59 variables are int64 datatypes and 2 variables are object datatypes.

2. First 5 rows of the dataset

	0	1	2	3	4	5	6	7	8	9
State Code	1	1	1	1	1	1	1	1	1	1
Dist.Code	1	2	3	4	5	6	7	8	9	10
State	Jammu & Kashmir	Jammu & Kashmir	Jammu & Kashmir	Jammu & Kashmir	Jammu & Kashmir	Jammu & Kashmir	Jammu & Kashmir	Jammu & Kashmir	Jammu & Kashmir	Jammu & Kashmir
Area Name	Kupwara	Badgam	Leh(Ladakh)	Kargil	Punch	Rajouri	Kathua	Baramula	Bandipore	Srinagar
No_HH	7707	6218	4452	1320	11654	16345	12510	9414	3814	15095
...	...	...	...	...	...	...	...	...	...	...
MARG_HH_0_3_F	252	148	34	50	302	256	120	265	234	22
MARG_OT_0_3_M	32	76	0	4	24	19	11	50	49	16
MARG_OT_0_3_F	46	178	4	10	105	71	19	94	144	134
NON_WORK_M	258	140	67	116	180	283	198	246	140	246
NON_WORK_F	214	160	61	59	478	835	139	198	243	247

61 rows × 10 columns

Table 2.1B
First 5 rows of the dataset

3. Last 5 rows of the dataset

	630	631	632	633	634	635	636	637	638	639
State Code	33	33	33	34	34	34	34	35	35	35
Dist.Code	631	632	633	634	635	636	637	638	639	640
State	Tamil Nadu	Tamil Nadu	Tamil Nadu	Puducherry	Puducherry	Puducherry	Puducherry	Andaman & Nicobar Island	Andaman & Nicobar Island	Andaman & Nicobar Island
Area Name	Krishnagiri	Coimbatore	Tiruppur	Yanam	Puducherry	Mahe	Karaikal	Nicobars	North & Middle Andaman	South Andaman
No_HH	65952	133255	98258	2219	37786	3333	10612	1275	3762	7975
...	...	...	...	...	...	...	...	...	...	...
MARG_HH_0_3_F	1139	664	499	11	503	0	130	6	21	17
MARG_OT_0_3_M	76	37	17	0	11	0	4	17	1	2
MARG_OT_0_3_F	223	225	171	1	46	0	23	47	4	4
NON_WORK_M	509	563	310	8	327	32	110	76	100	148
NON_WORK_F	793	1202	875	18	388	47	170	77	103	99

61 rows × 10 columns

Table 2.1C
Last 5 rows of the dataset

4. Shape of the dataset: - Number of Rows and Columns in a Dataset

```
Number of rows in dataset is  640
Number of columns in a dataset is 61
```

5. Statistical summary of the Dataset

	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	98785.0
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

Table 2.1D/1

	count	mean	std	min	25%	50%	75%	max
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
MARGWORK_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
MARGWORK_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

Table 2.1D/2

Descriptive Statistical Summary of Dataset

- By using Describe ( ) function we get 5 point summary of the dataset i.e., minimum, 25%, 50%, 75% and maximum . We can see that the scale of variables is different.
- Also we get the statistical summary of the data i.e., mean, median and mode.
- We can also see that different variables have different scale. So we need to preform Scaling.

- We can also see that there are NAN values in some of the variables which tell us that there are some null entries in the dataset which we need to check.

6. Missing Value in a Dataset

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

Table 2.1E
Missing Values in a Dataset

7. Check for Duplicates

```
Total Number od Duplicated data in a dataset =  0
```

Part 2.2

PCA: 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

For Exploratory Data Analysis (EDA) we have Selected No_HH, TOT_M, TOT_F, M_06, F_06 and made a new dataset 'data_eda'

Gender ratio = Total Number of Female/Total Number of Male

We have added a New Variable Gender Ratio in the "data_eda" to perform EDA

Data_eda dataset

	State	Area Name	No_HH	TOT_M	TOT_F	TOT_WORK_M	TOT_WORK_F	Gender_ratio
0	Jammu & Kashmir	Kupwara	7707	23388	29796	6723	3752	1.273987
1	Jammu & Kashmir	Badgam	6218	19585	23102	6982	4200	1.179576
2	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	2775	4800	1.674916
3	Jammu & Kashmir	Kargil	1320	2784	4206	1002	1118	1.510776
4	Jammu & Kashmir	Punch	11654	20591	29981	5717	7692	1.456024
...	...	...	...	...	...	...	...	...
635	Puducherry	Mahe	3333	8154	11781	3808	1328	1.444812
636	Puducherry	Karaikal	10612	12346	21691	6458	5286	1.756925
637	Andaman & Nicobar Island	Nicobars	1275	1549	2630	715	1031	1.697870
638	Andaman & Nicobar Island	North & Middle Andaman	3762	5200	8012	2707	2174	1.540769
639	Andaman & Nicobar Island	South Andaman	7975	11977	18049	6345	5278	1.506972

640 rows × 8 columns

Table 2.2A
Data_eda Dataset for EDA

Univariate Analysis

1. Total Number of Households

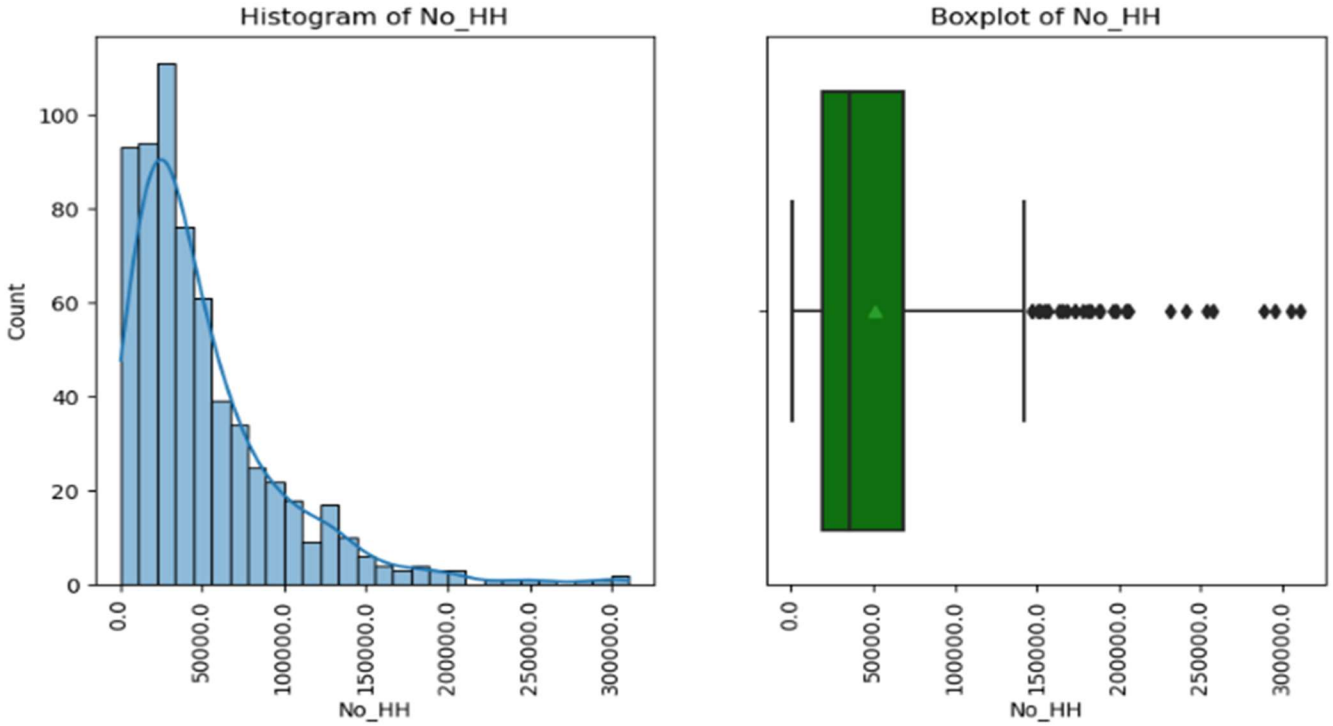


Fig. 2.2A
Histogram and Boxplot of Total Households

2. Total Number of Male

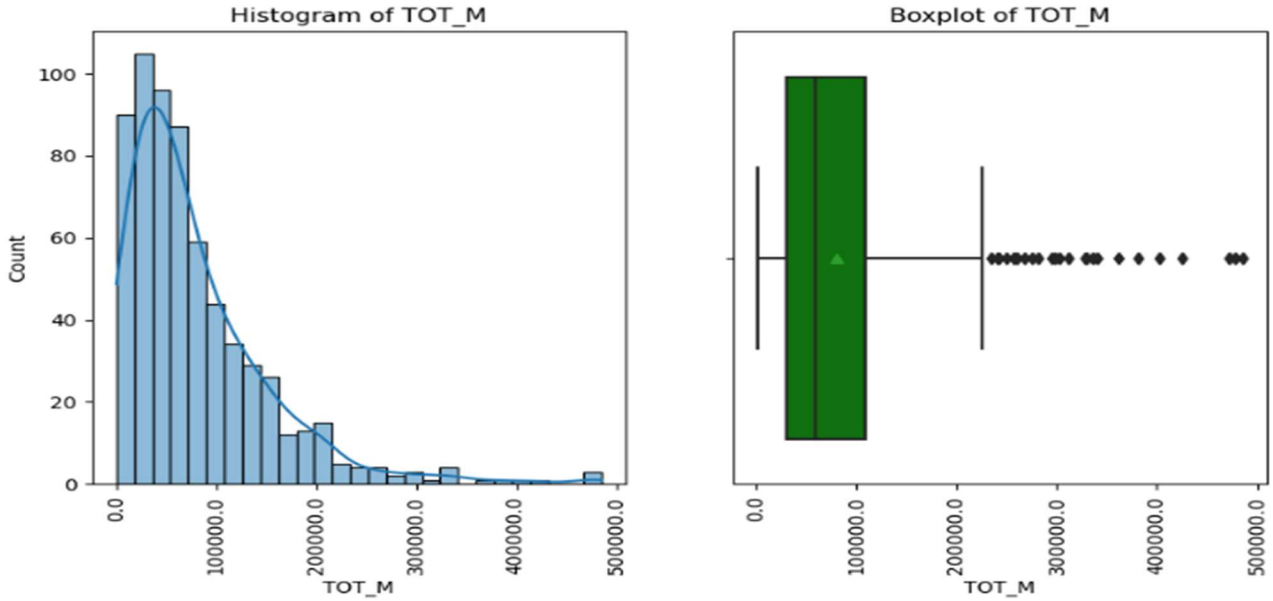


Fig. 2.2B
Histogram and Boxplot of Total Number of Male

3. Total Number of Females

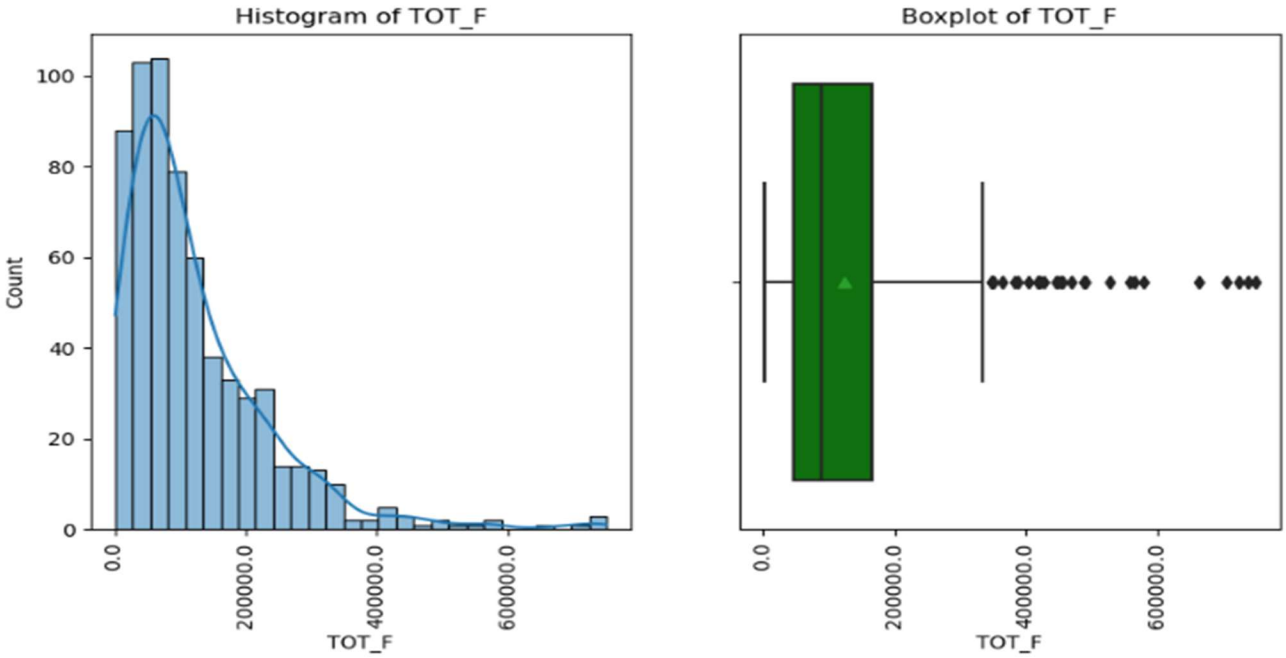


Fig. 2.2C
Histogram and Boxplot of Total Number of Female

4. Total Working Male

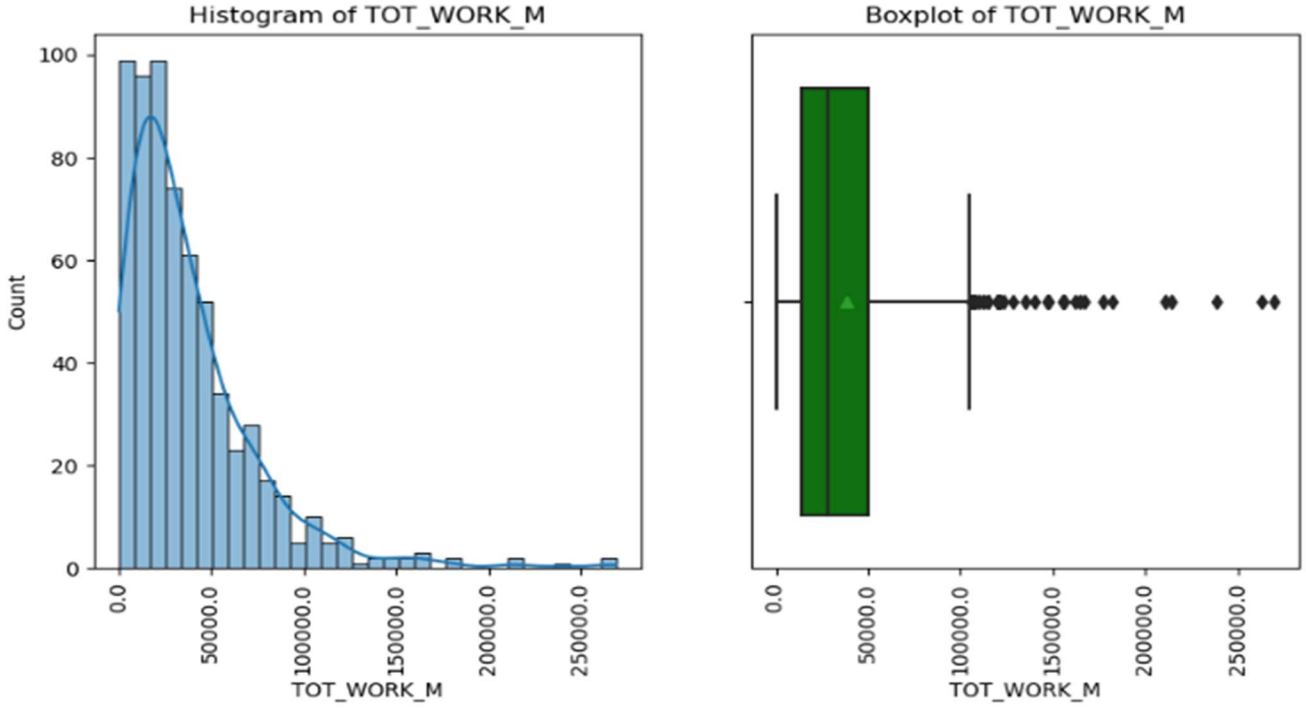


Fig. 2.2D
Histogram and Boxplot Total working Male

5. Total Working Female

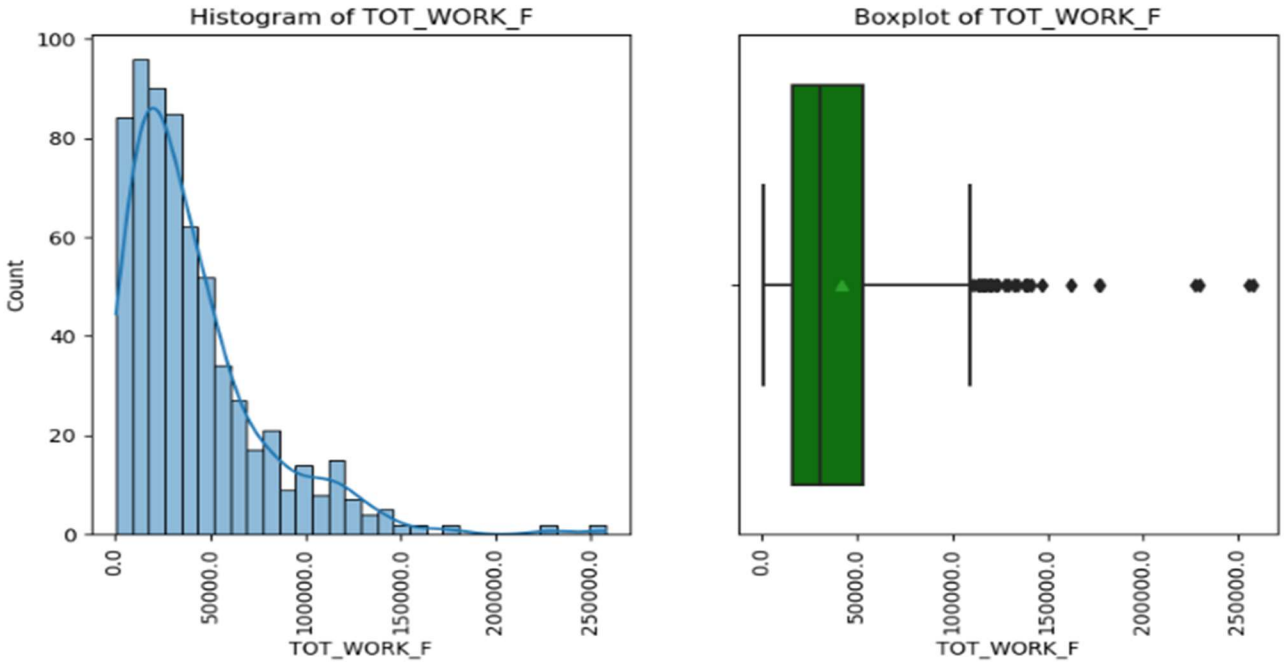


Fig. 2.2E
Histogram and Boxplot of Total working Female

Bi-Variate Analysis

Maximum Number of District/Area Name in a State

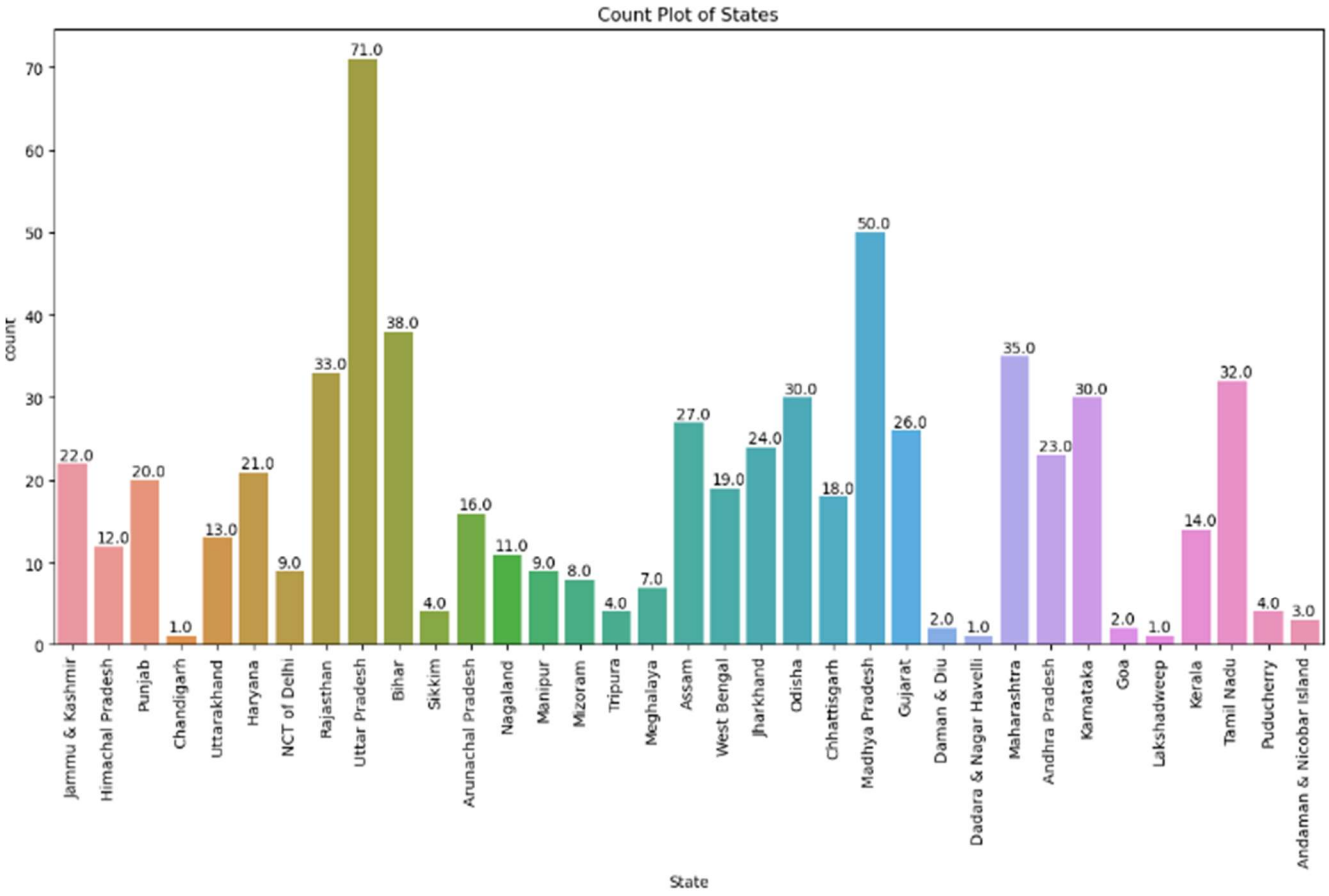


Fig.2.2F
Count plot of District in each State

- **Uttar Pradesh has maximum number of Districts (Area Name) i.e., 71**
- **Chandigarh, Dadar & Nagar Haveli and Lakshadweep has minimum number of District (Area name) i.e., 1**

Gender Ratio in Different States

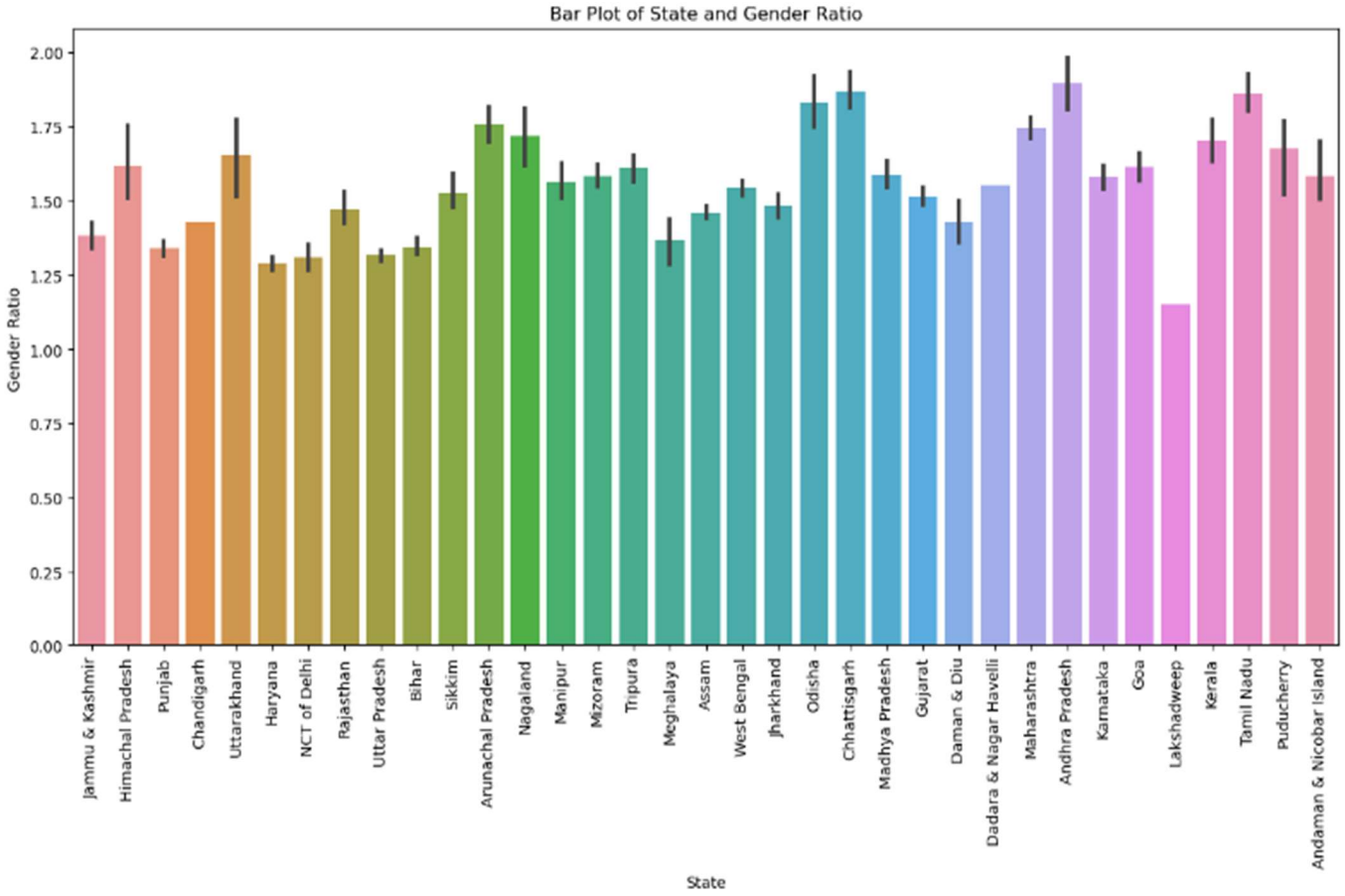


Fig. 2.2G
Bar Plot of State and Gender Ratio

- **State with maximum Gender Ratio (Total Female/Total Male)**

```
State
Andhra Pradesh      1.895093
```

- **State with Minimum Gender Ratio (Total Female/Total Male)**

```
State
Lakshadweep      1.151993
```

State with Maximum/Minimum Average Male Population

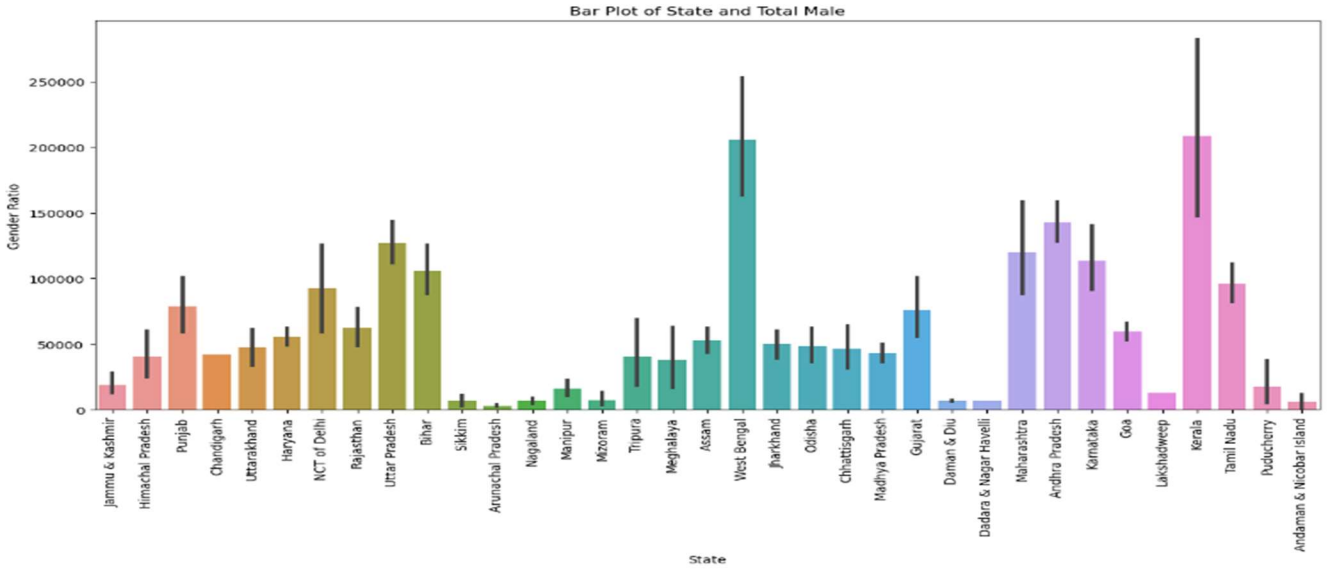


Fig. 2.2 H

Bar Plot of average Total Male in each State

From Above Bar plot we can see that: -

- **Maximum average Male Population is in Kerala**
- **Minimum average Male Population is in Arunachal Pradesh**

State with Maximum/Minimum Average Female Population

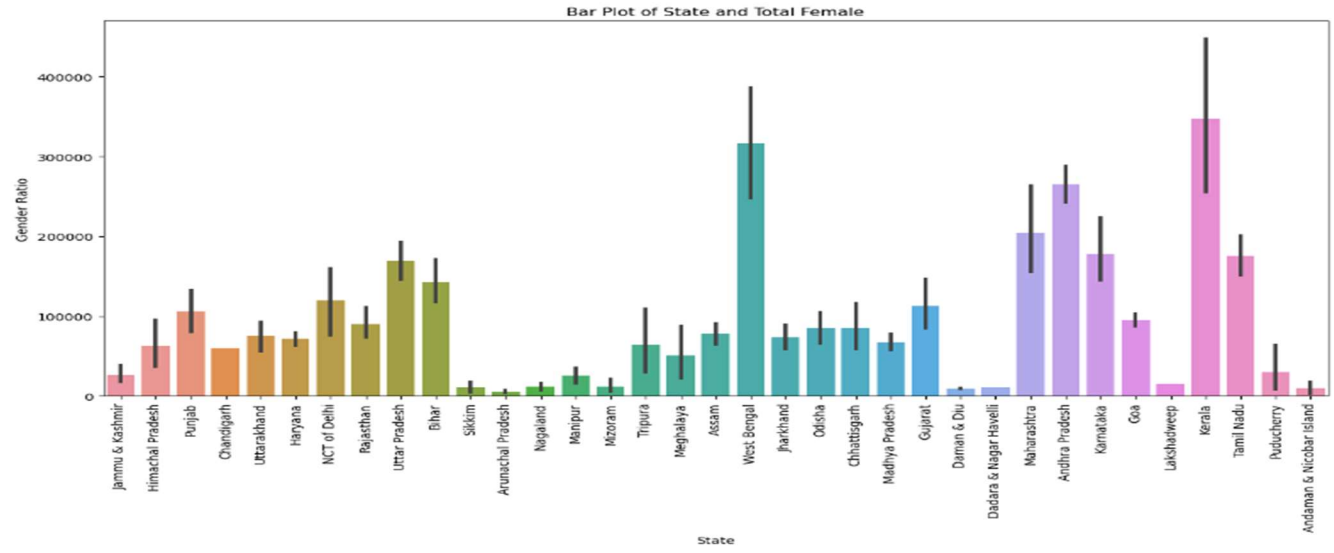


Fig. 2.2 I

Bar Plot of average Total Female in each State

From Above Bar plot we can see that: -

- **Maximum average Female Population is in West Bengal**
- **Minimum average Female Population is in Arunachal Pradesh**

State with Maximum/Minimum Average Working Male Population

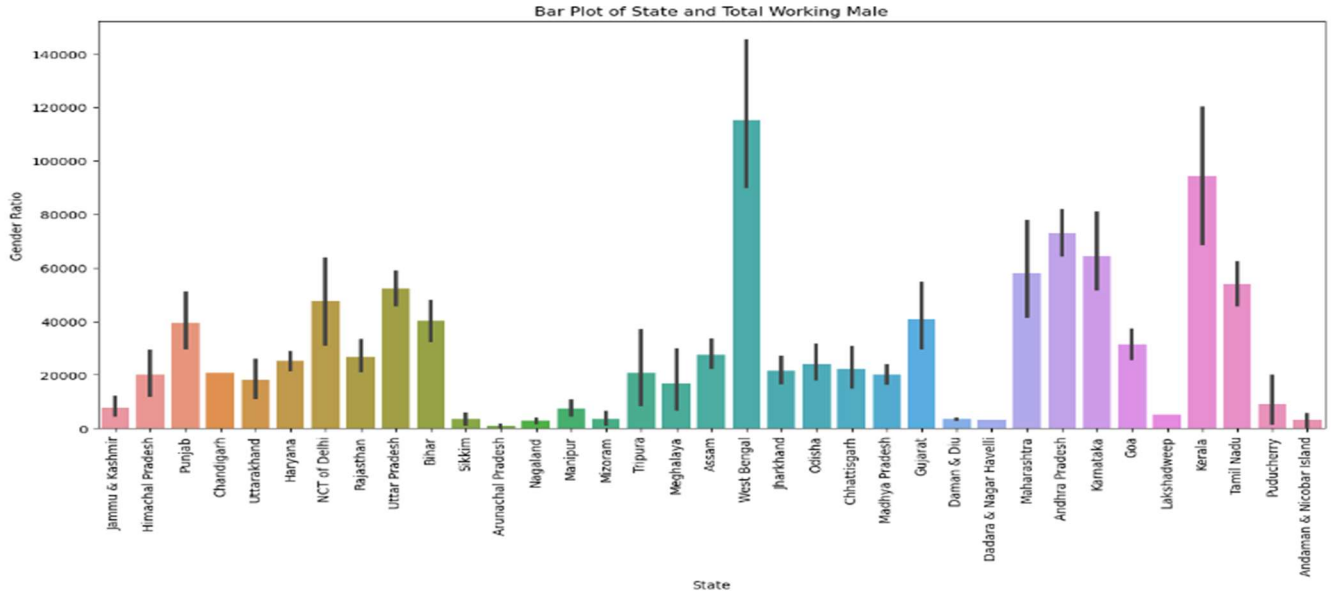


Fig. 2.2 J

Bar Plot of average Total Working Male

From Above Bar plot we can see that: -

- **Maximum working Male Population is in West Bengal**
- **Minimum working Male Population is in Arunachal Pradesh**

State with Maximum/Minimum Average working Female Population

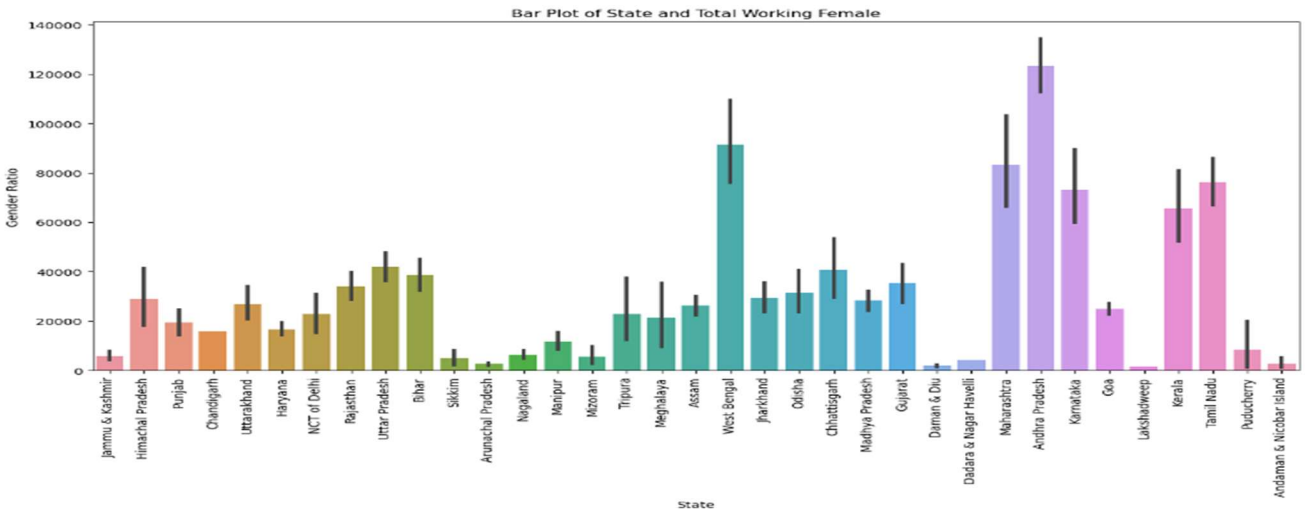


Fig. 2.2 K

Bar Plot of Total Working Male

From Above Bar plot we can see that: -

- **Maximum working Female Population is in Andhra Pradesh**
- **Minimum working Female Population is in Lakshadweep**

Heat Map/Correlation Map of all Numeric Variables

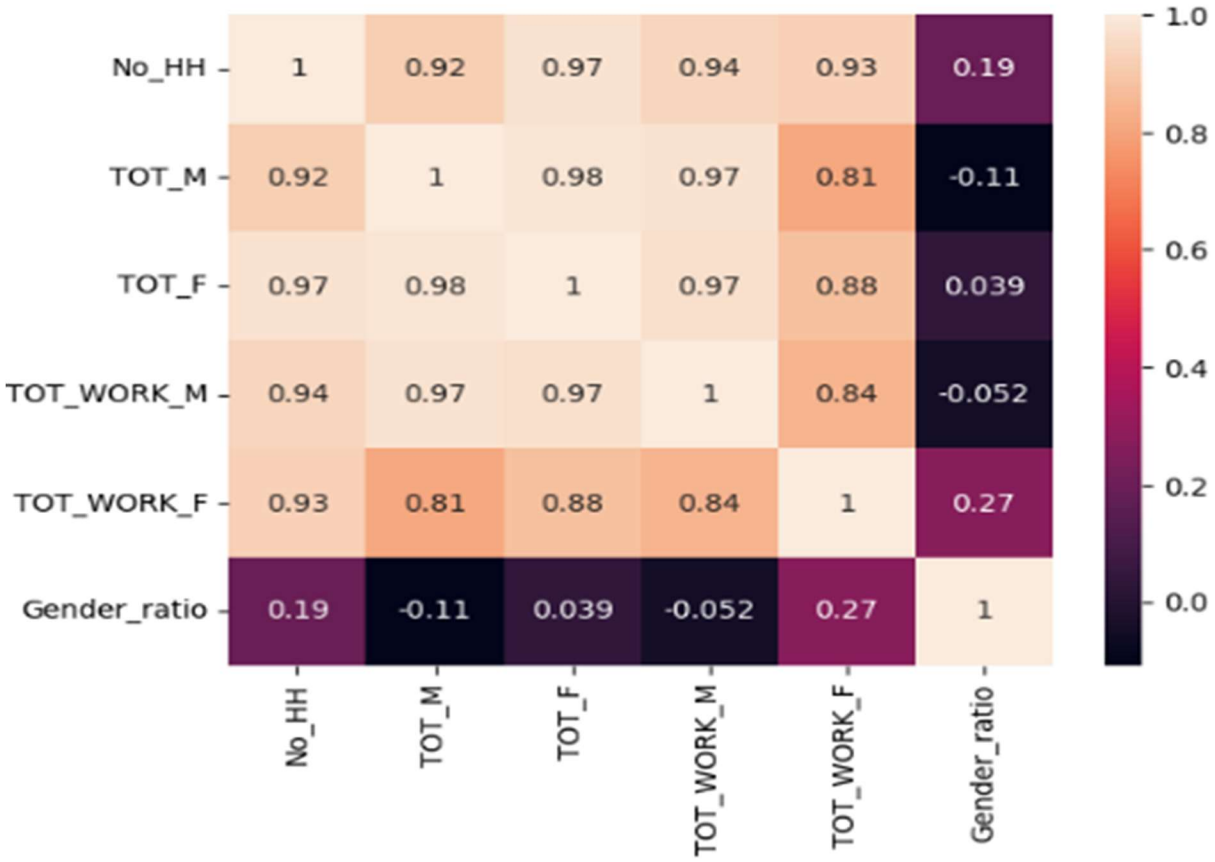


Fig. 2.2 L
Heat Map /Correlation Map for Numeric Variables

From above Heat map we can see that there is strong correlation between most of the variables which is very import feature in order to perform Principal Component Analysis (PCA)

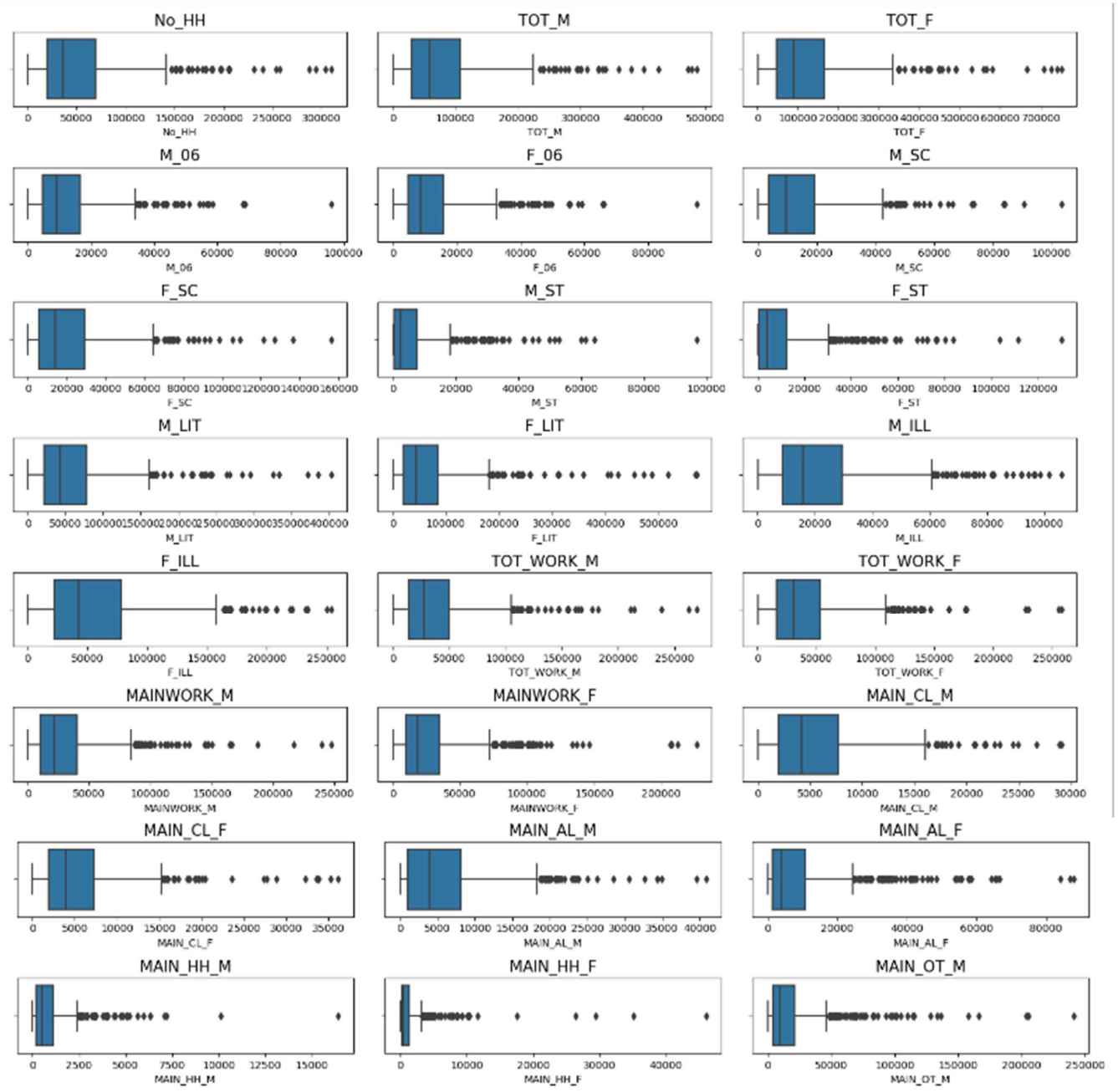
Part 2.3

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

As there is no such error in the data, so we prefer not to fix the outliers. Also it is considered that outliers may contains some important Information.

PCA2.4

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



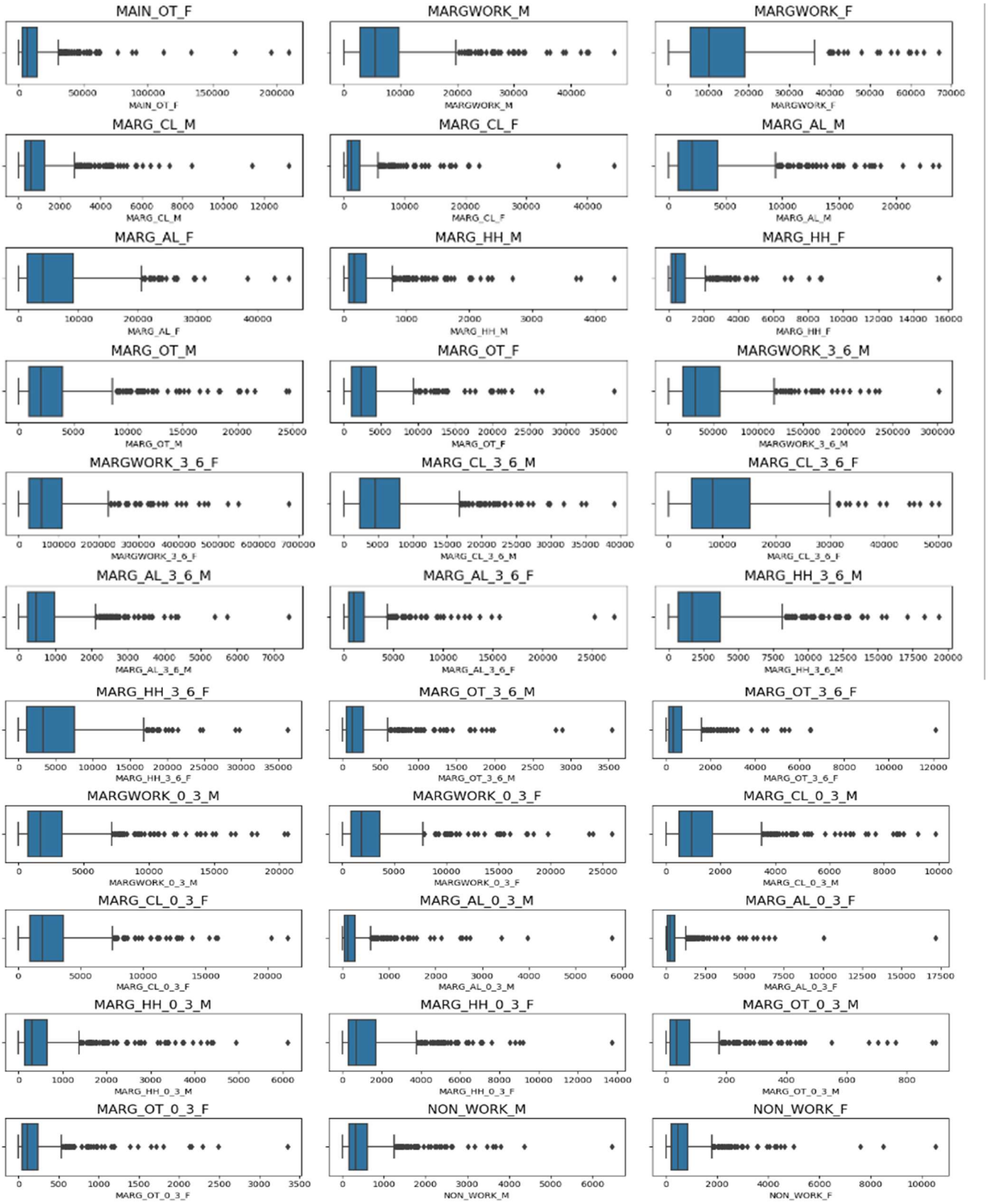


Fig. 2.4A

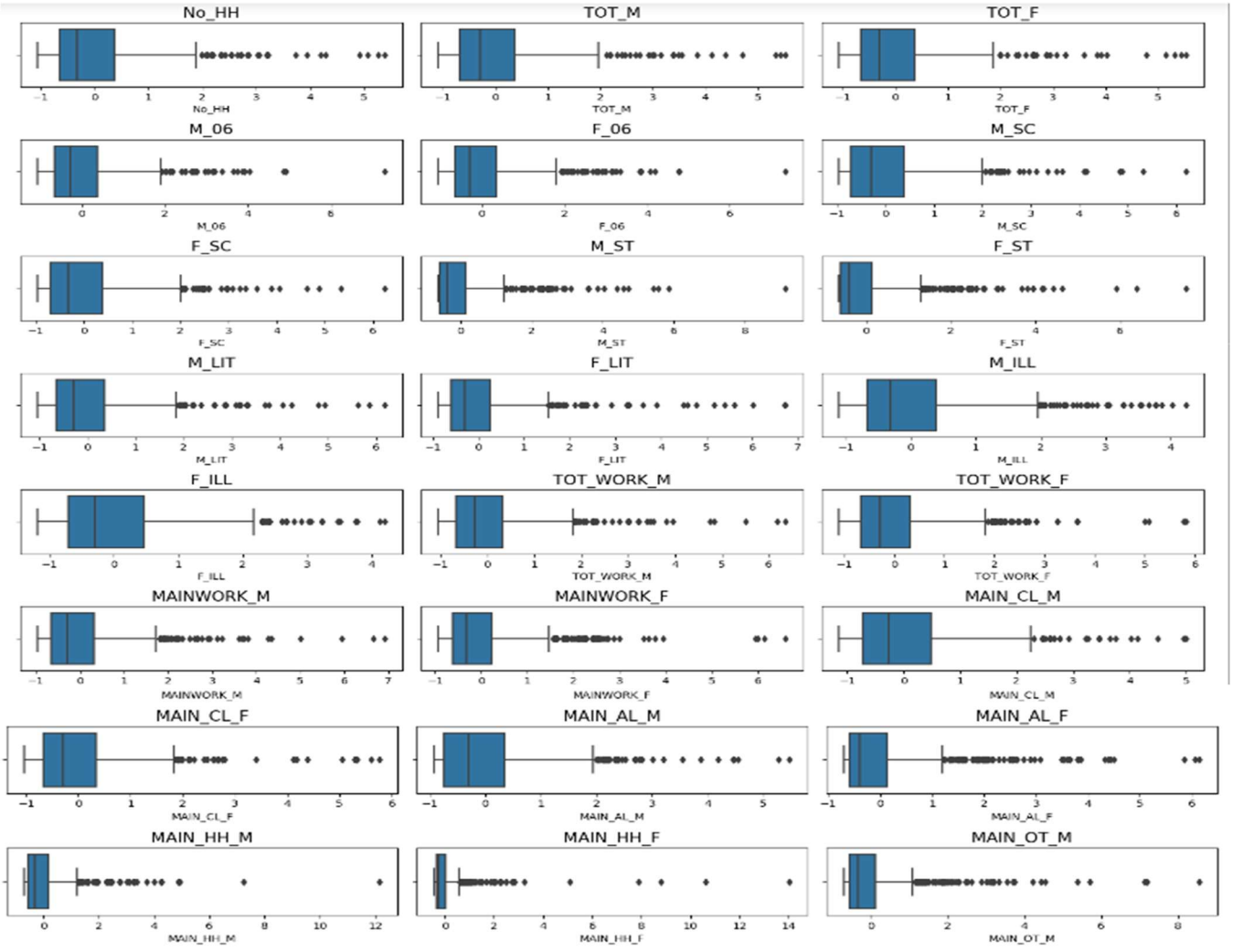
Boxplot of Numeric Variables (Before Scaling)

For scaling here we have imported ‘sklearn’ library
From ‘sklearn.preprocessing’ we had imported ‘StandardScaler’ function

StandardScaler () scales down each variables/columns such that each variables/columns Mean is 0 and Standard Deviation 1

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARGWORK_0_3_M	MARGWORK_0_3_F
0	-0.904738	-0.771236	-0.815563	-0.561012	-0.507738	-0.958575	-0.957049	-0.423306	-0.476423	-0.798097	...	-0.569151	-0.612451
1	-0.935695	-0.823100	-0.874534	-0.681096	-0.725367	-0.958297	-0.956772	-0.582014	-0.607607	-0.849434	...	-0.682181	-0.710490
2	-0.972412	-1.000919	-0.981466	-0.976956	-0.965262	-0.958575	-0.956772	-0.038951	-0.027273	-0.956457	...	-0.747099	-0.739059
3	-1.037530	-1.052224	-1.041001	-1.022118	-0.995393	-0.958783	-0.957049	-0.355965	-0.390060	-1.004643	...	-0.800484	-0.808528
4	-0.822676	-0.809381	-0.813933	-0.622359	-0.649908	-0.957395	-0.955529	0.149238	0.043330	-0.800568	...	-0.668341	-0.522533
...	...	...	...	...	...	...	...	...	...	...	...	...	...
635	-0.995677	-0.978990	-0.974268	-0.971387	-0.948916	-0.957326	-0.955667	-0.625124	-0.640197	-0.913820	...	-0.759622	-0.770936
636	-0.844340	-0.921822	-0.886965	-0.936754	-0.919757	-0.803806	-0.765670	-0.625124	-0.640197	-0.853390	...	-0.723702	-0.738157
637	-1.038465	-1.069066	-1.054885	-1.051356	-1.035331	-0.958783	-0.957049	-0.522953	-0.529880	-1.016367	...	-0.794223	-0.782665
638	-0.986758	-1.019276	-1.007472	-1.008195	-0.996541	-0.958783	-0.957049	-0.622297	-0.637046	-0.962328	...	-0.754019	-0.798604

Table 2.4A
Scaled Dataset



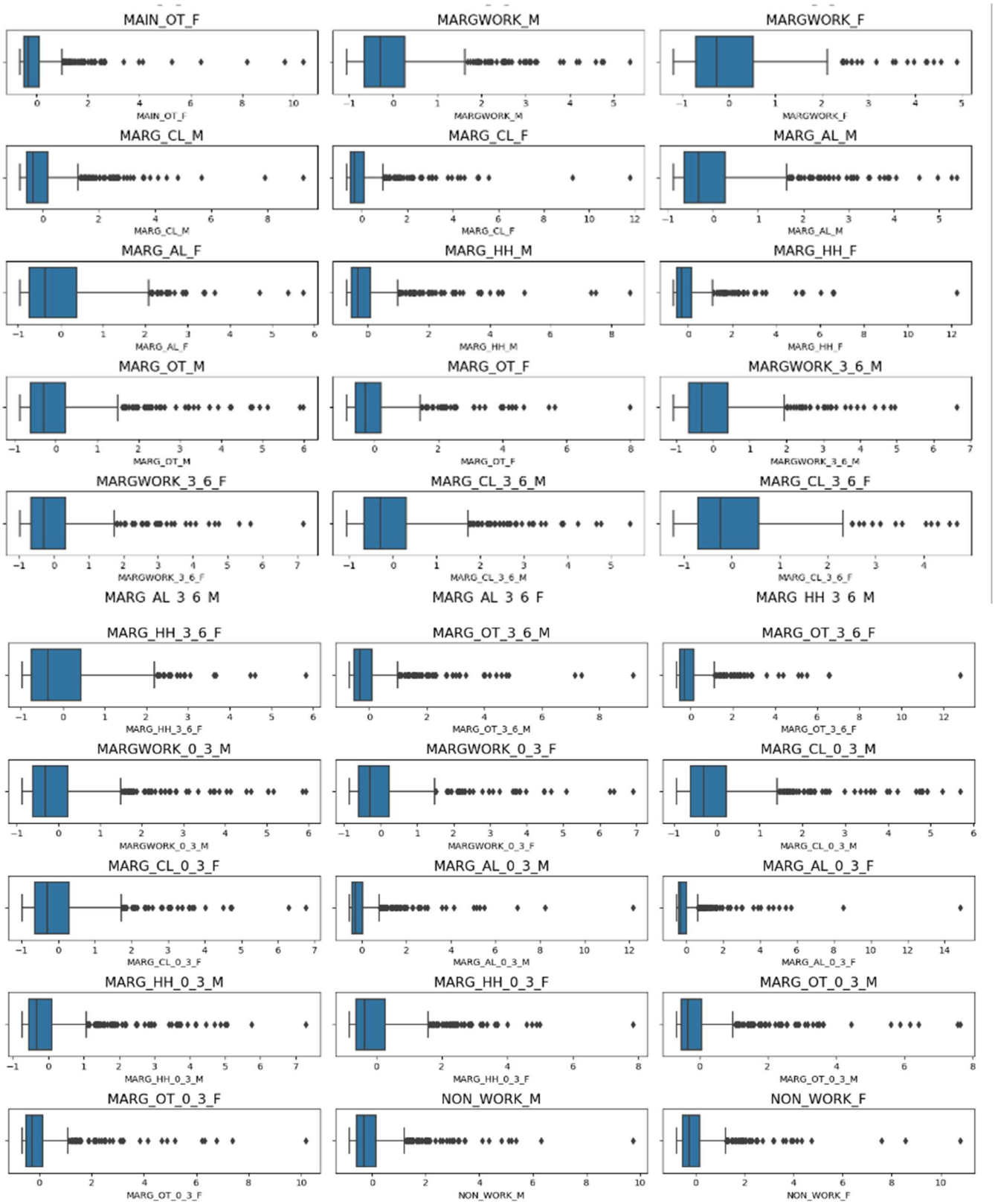


Fig. 2.4B
Boxplot of Numeric Variables (After Scaling)

	State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MARG_CL_0_3_M	MARG_CL_0_3_F
count	640.00	640.00	640.00	640.00	640.00	640.00	640.00	640.00	640.00	640.00	...	640.00	640.00
mean	17.11	320.50	51222.87	79940.58	122372.08	12309.10	11942.30	13820.95	20778.39	6191.81	...	1392.97	2757.05
std	9.43	184.90	48135.41	73384.51	113600.72	11500.91	11326.29	14426.37	21727.89	9912.67	...	1489.71	2788.78
min	1.00	1.00	350.00	391.00	698.00	56.00	56.00	0.00	0.00	0.00	...	4.00	30.00
25%	9.00	160.75	19484.00	30228.00	46517.75	4733.75	4672.25	3466.25	5603.25	293.75	...	489.50	957.25
50%	18.00	320.50	35837.00	58339.00	87724.50	9159.00	8663.00	9591.50	13709.00	2333.50	...	949.00	1928.00
75%	24.00	480.25	68892.00	107918.50	164251.75	16520.25	15902.25	19429.75	29180.00	7658.00	...	1714.00	3599.75
max	35.00	640.00	310450.00	485417.00	750392.00	96223.00	95129.00	103307.00	156429.00	96785.00	...	9875.00	21611.00

Table 2.4B
Describe () output Before Scaling

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARGWORK_0_3_M	MARGWORK_0_3_F
count	640.000	640.000	640.000	640.000	640.000	640.000	640.000	640.000	640.000	640.000	...	640.000	640.000
mean	0.000	-0.000	-0.000	-0.000	0.000	0.000	-0.000	-0.000	-0.000	-0.000	...	0.000	0.000
std	1.001	1.001	1.001	1.001	1.001	1.001	1.001	1.001	1.001	1.001	...	1.001	1.001
min	-1.058	-1.085	-1.072	-1.066	-1.050	-0.959	-0.957	-0.625	-0.640	-1.032	...	-0.860	-0.848
25%	-0.660	-0.678	-0.668	-0.659	-0.642	-0.718	-0.699	-0.595	-0.613	-0.656	...	-0.613	-0.602
50%	-0.320	-0.295	-0.305	-0.274	-0.290	-0.293	-0.326	-0.390	-0.398	-0.273	...	-0.308	-0.301
75%	0.367	0.382	0.369	0.366	0.350	0.389	0.387	0.148	0.147	0.358	...	0.232	0.233
max	5.390	5.530	5.533	7.302	7.350	6.208	6.248	9.146	7.562	6.181	...	5.942	6.920

Table 2.4C
Describe () output After Scaling

It can be seen from the boxplot & describe () output that before and after scaling there is not much impact on Outliers.

Part 2.5

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

Steps to Perform Principal Component Analysis (PCA)

Step 1: - We need to perform bartlett_sphericity test . It tells us about the correlation between the variables. If the P-Value is less than 0.5 , there is a strong correlation and good to proceed further.
If it is less than 0.5 , we cannot perform PCA

Here P-Value is 0, which means there is a strong correlation between the variables

Step 2: - We need to perform calculate_kmo test. It tells us if we have significant sample size to conduct PCA. If kmo_model value is more than 0.5 , we can perform PCA. If kmo_model value is less than 0.5 we need to get more samples or we cannot perform PCA.

Here kmo_model values is 0.8, which means we can perform PCA on this dataset

Step 3: - To perform PCA we need to load 'sklearn' library.

From sklearn.decomposition we need to import PCA function.

Covariance Matrix

```
array([[ 3.18135647e+01,  3.20244613e-15, -6.22697859e-16, ...,
         1.64191706e-31, -1.48142893e-32, -4.44428679e-32],
       [ 3.20244613e-15,  7.86942415e+00,  4.44784185e-16, ...,
        -2.83940545e-32, -4.69119161e-32,  6.17262054e-32],
       [-6.22697859e-16,  4.44784185e-16,  4.15340812e+00, ...,
        -6.17262054e-32, -1.23452411e-32,  5.18500125e-32],
       ...,
       [ 1.64191706e-31, -2.83940545e-32, -6.17262054e-32, ...,
         2.46823875e-31,  7.53446493e-34,  1.20732223e-33],
       [-1.48142893e-32, -4.69119161e-32, -1.23452411e-32, ...,
         7.53446493e-34,  2.47218134e-31, -9.32080599e-34],
       [-4.44428679e-32,  6.17262054e-32,  5.18500125e-32, ...,
         1.20732223e-33, -9.32080599e-34,  2.46306247e-31]])
```

Fig. 2.5A
Covariance Matrix for Given Dataset

Eigen Vectors

```
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.        ,  0.2077636 ,  0.24647657, ..., -0.07217993,
         0.00399206, -0.06929081],
       [ 0.        ,  0.2887035 , -0.20596721, ...,  0.04019745,
        -0.03192722,  0.00778048],
       [-0.        ,  0.18790022,  0.02642675, ..., -0.02597314,
        -0.13972835, -0.02147533]])
```

Fig 2.5B

Eigen Vectors of Given Dataset

Eigen Values

```
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])
```

Fig. 2.5C

Eigen Values of given Dataset

<u>Explained Variance Ratio</u>

```
array([0.557, 0.138, 0.073, 0.064, 0.039, 0.034, 0.021, 0.013, 0.011,
       0.009, 0.008, 0.006, 0.005, 0.005, 0.003, 0.002, 0.002, 0.002,
       0.002, 0.001, 0.001, 0.001, 0.001, 0.001, 0.   , 0.   , 0.   ,
       0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ,
       0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ,
       0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ,
       0.   , 0.   , 0.   ])
```

Fig. 2.5D

Explained Variance Ratio (Rounding up to 3 Decimal)

Part 2.6

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

Cumulative Explained variance ratio

```
array([0.55503817, 0.6965685 , 0.77046291, 0.8365171 , 0.87362788,
       0.90857579, 0.92918864, 0.94239263, 0.95222687, 0.96118762,
       0.96898215, 0.97472283, 0.97994971, 0.98460163, 0.98770208,
       0.98976981, 0.99171383, 0.99329944, 0.99469207, 0.99576917,
       0.99652724, 0.99721774, 0.99778393, 0.99829324, 0.99871192,
       0.99904749, 0.99927834, 0.99946022, 0.99960979, 0.99975544,
       0.99989147, 0.99993876, 0.99996795, 0.99998706, 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ])
```

Fig 2.6A

Cumulative Explained Ratio

Since it is given in the project to take at least 90% of explained variance.

From the above fig. 2.6A Cumulative Explained Variance ratio, the optimum number of PCs can be 6.

So we will consider 6 Principal Components for the given Dataset

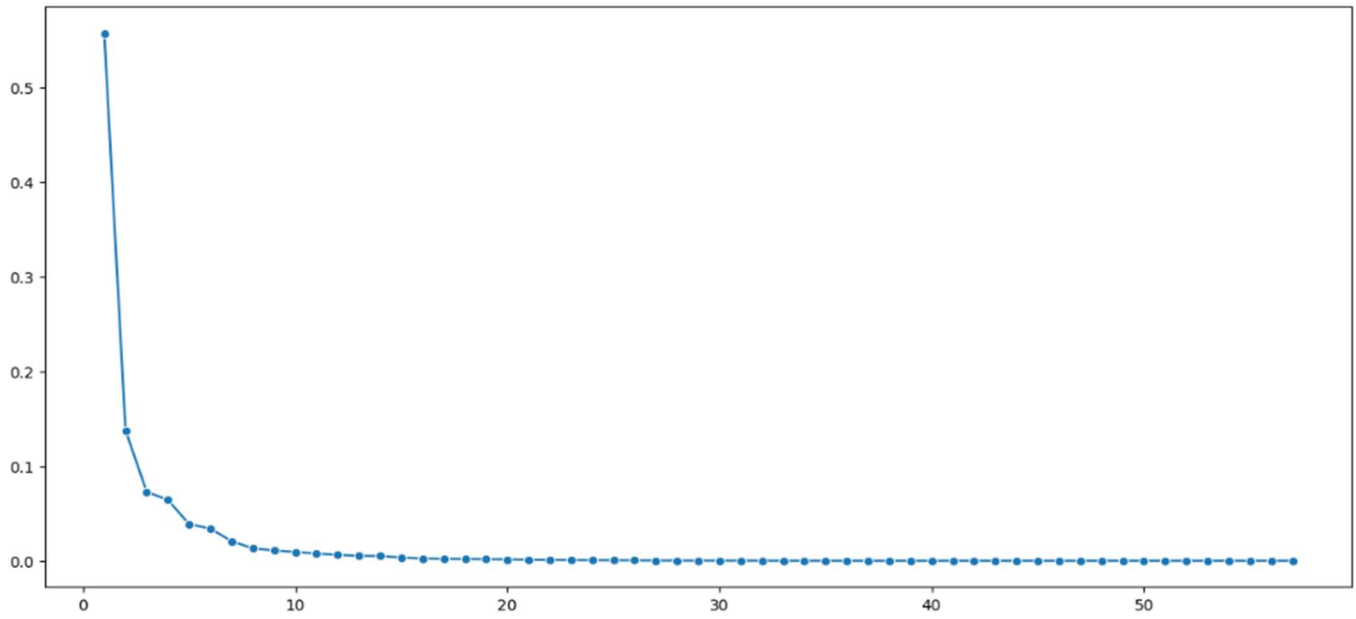


Fig 2.6B

Scree Plot of the given Dataset

From the scree plot also we see that after n-components =6 there is no significant drop.

Hence we will consider 6 PCs for the given dataset.

Part 2.7

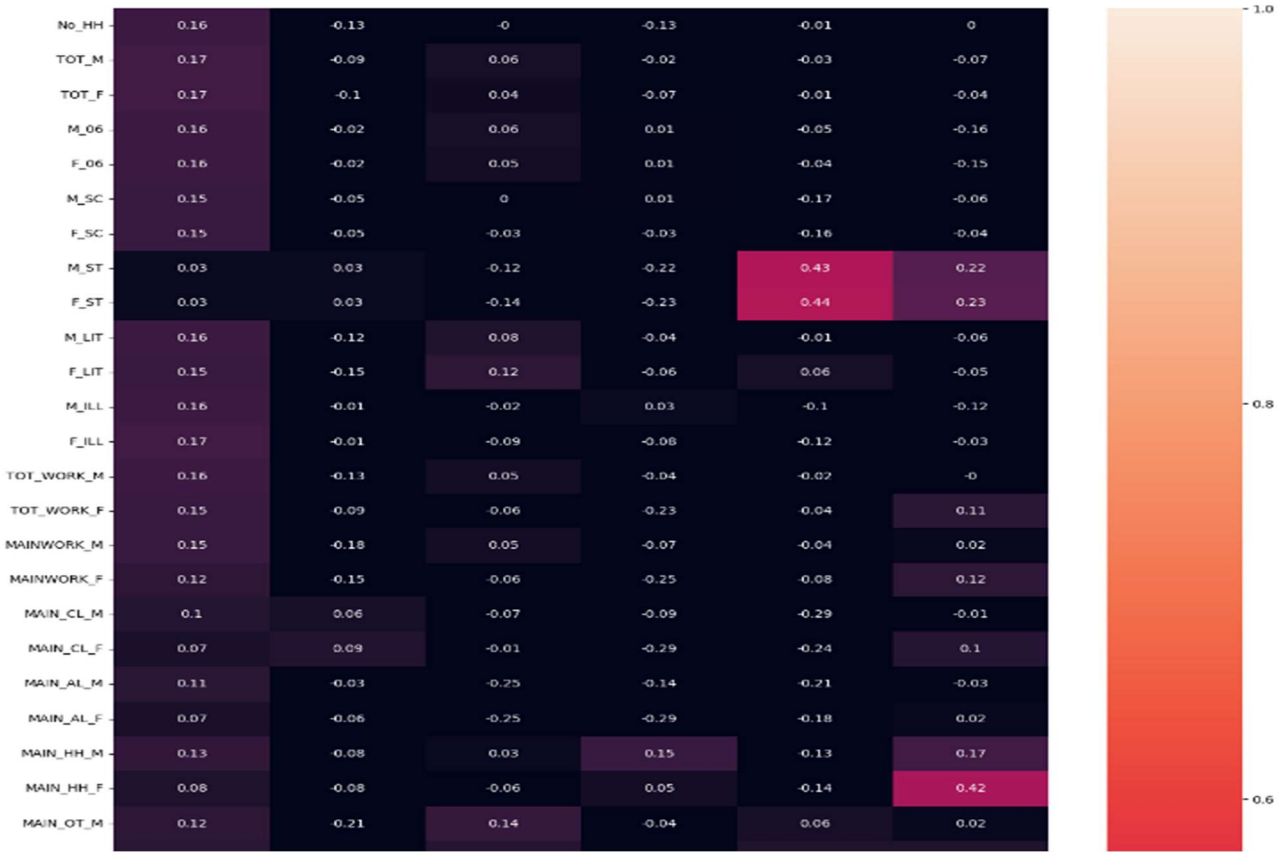
PCA: Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.

	PC1	PC2	PC3	PC4	PC5	PC6
No_HH	0.156000	-0.126000	-0.003000	-0.125000	-0.007000	0.004000
TOT_M	0.167000	-0.090000	0.057000	-0.020000	-0.033000	-0.073000
TOT_F	0.166000	-0.105000	0.039000	-0.071000	-0.013000	-0.044000
M_06	0.162000	-0.022000	0.058000	0.012000	-0.050000	-0.158000
F_06	0.163000	-0.020000	0.050000	0.015000	-0.044000	-0.154000
M_SC	0.151000	-0.045000	0.003000	0.012000	-0.173000	-0.064000
F_SC	0.152000	-0.052000	-0.025000	-0.030000	-0.160000	-0.041000
M_ST	0.027000	0.028000	-0.124000	-0.222000	0.433000	0.223000
F_ST	0.028000	0.030000	-0.140000	-0.230000	0.439000	0.226000
M_LIT	0.162000	-0.115000	0.082000	-0.035000	-0.009000	-0.055000
F_LIT	0.147000	-0.153000	0.117000	-0.060000	0.058000	-0.048000
M_ILL	0.162000	-0.007000	-0.022000	0.025000	-0.097000	-0.115000
F_ILL	0.165000	-0.009000	-0.093000	-0.076000	-0.120000	-0.029000
TOT_WORK_M	0.160000	-0.134000	0.045000	-0.040000	-0.020000	-0.002000
TOT_WORK_F	0.146000	-0.085000	-0.059000	-0.225000	-0.040000	0.105000
MAINWORK_M	0.146000	-0.176000	0.054000	-0.068000	-0.037000	0.019000
MAINWORK_F	0.124000	-0.151000	-0.056000	-0.247000	-0.083000	0.124000
MAIN_CL_M	0.103000	0.062000	-0.067000	-0.090000	-0.286000	-0.006000
MAIN_CL_F	0.075000	0.086000	-0.009000	-0.289000	-0.242000	0.103000
MAIN_AL_M	0.113000	-0.031000	-0.248000	-0.136000	-0.206000	-0.031000
MAIN_AL_F	0.074000	-0.059000	-0.252000	-0.290000	-0.178000	0.019000
MAIN_HH_M	0.132000	-0.076000	0.027000	0.152000	-0.134000	0.174000
MAIN_HH_F	0.083000	-0.082000	-0.061000	0.049000	-0.139000	0.422000
MAIN_OT_M	0.124000	-0.213000	0.137000	-0.040000	0.065000	0.023000
MAIN_OT_F	0.111000	-0.210000	0.096000	-0.120000	0.081000	0.083000
MARGWORK_M	0.165000	0.093000	-0.009000	0.093000	0.060000	-0.091000
MAINWORK_F	0.124000	-0.151000	-0.056000	-0.247000	-0.083000	0.124000
MAIN_CL_M	0.103000	0.062000	-0.067000	-0.090000	-0.286000	-0.006000
MAIN_CL_F	0.075000	0.086000	-0.009000	-0.289000	-0.242000	0.103000
MAIN_AL_M	0.113000	-0.031000	-0.248000	-0.136000	-0.206000	-0.031000
MAIN_AL_F	0.074000	-0.059000	-0.252000	-0.290000	-0.178000	0.019000
MAIN_HH_M	0.132000	-0.076000	0.027000	0.152000	-0.134000	0.174000
MAIN_HH_F	0.083000	-0.082000	-0.061000	0.049000	-0.139000	0.422000
MAIN_OT_M	0.124000	-0.213000	0.137000	-0.040000	0.065000	0.023000
MAIN_OT_F	0.111000	-0.210000	0.096000	-0.120000	0.081000	0.083000
MARGWORK_M	0.165000	0.093000	-0.009000	0.093000	0.060000	-0.091000
MARGWORK_F	0.155000	0.125000	-0.049000	-0.089000	0.089000	0.018000
MARG_CL_M	0.082000	0.269000	0.199000	-0.063000	-0.022000	0.032000
MARG_CL_F	0.049000	0.247000	0.269000	-0.168000	-0.059000	0.092000
MARG_AL_M	0.129000	0.166000	-0.190000	0.092000	0.019000	-0.142000
MARG_AL_F	0.114000	0.141000	-0.268000	-0.106000	0.081000	-0.085000
MARG_HH_M	0.141000	0.068000	-0.021000	0.238000	-0.060000	0.090000
MARG_HH_F	0.128000	0.024000	-0.083000	0.196000	-0.034000	0.365000
MARG_OT_M	0.155000	-0.089000	0.112000	0.087000	0.119000	-0.061000
MARG_OT_F	0.147000	-0.118000	0.100000	0.027000	0.167000	0.002000
MARGWORK_3_6_M	0.165000	-0.044000	0.064000	-0.000000	-0.044000	-0.136000
MARGWORK_3_6_F	0.161000	-0.106000	0.080000	0.004000	0.001000	-0.107000
MARG_CL_3_6_M	0.166000	0.077000	-0.024000	0.093000	0.054000	-0.097000
MARG_CL_3_6_F	0.156000	0.103000	-0.072000	-0.108000	0.073000	0.024000
MARG_AL_3_6_M	0.093000	0.264000	0.154000	-0.038000	-0.008000	0.013000
MARG_AL_3_6_F	0.052000	0.244000	0.256000	-0.180000	-0.061000	0.094000
MARG_HH_3_6_M	0.129000	0.159000	-0.200000	0.080000	0.008000	-0.144000
MARG_HH_3_6_F	0.111000	0.125000	-0.280000	-0.136000	0.064000	-0.077000
MARG_OT_3_6_M	0.140000	0.062000	-0.021000	0.238000	-0.066000	0.097000

MARG_OT_3_6_F	0.125000	0.015000	-0.083000	0.191000	-0.045000	0.385000
MARGWORK_0_3_M	0.154000	-0.093000	0.110000	0.086000	0.109000	-0.062000
MARGWORK_0_3_F	0.146000	-0.126000	0.096000	0.027000	0.141000	0.009000
MARG_CL_0_3_M	0.150000	0.151000	0.055000	0.087000	0.081000	-0.061000
MARG_CL_0_3_F	0.140000	0.181000	0.024000	-0.022000	0.130000	-0.002000
MARG_AL_0_3_M	0.053000	0.251000	0.268000	-0.105000	-0.049000	0.085000
MARG_AL_0_3_F	0.042000	0.241000	0.285000	-0.136000	-0.052000	0.084000
MARG_HH_0_3_M	0.122000	0.185000	-0.139000	0.133000	0.062000	-0.124000
MARG_HH_0_3_F	0.116000	0.181000	-0.202000	0.004000	0.128000	-0.106000
MARG_OT_0_3_M	0.140000	0.085000	-0.023000	0.230000	-0.036000	0.061000
MARG_OT_0_3_F	0.132000	0.051000	-0.079000	0.206000	0.000000	0.296000
NON_WORK_M	0.150000	-0.065000	0.112000	0.085000	0.163000	-0.052000
NON_WORK_F	0.131000	-0.074000	0.103000	0.021000	0.238000	-0.025000

Table 2.7A

Dataframe Comparing PCs with Actual Columns



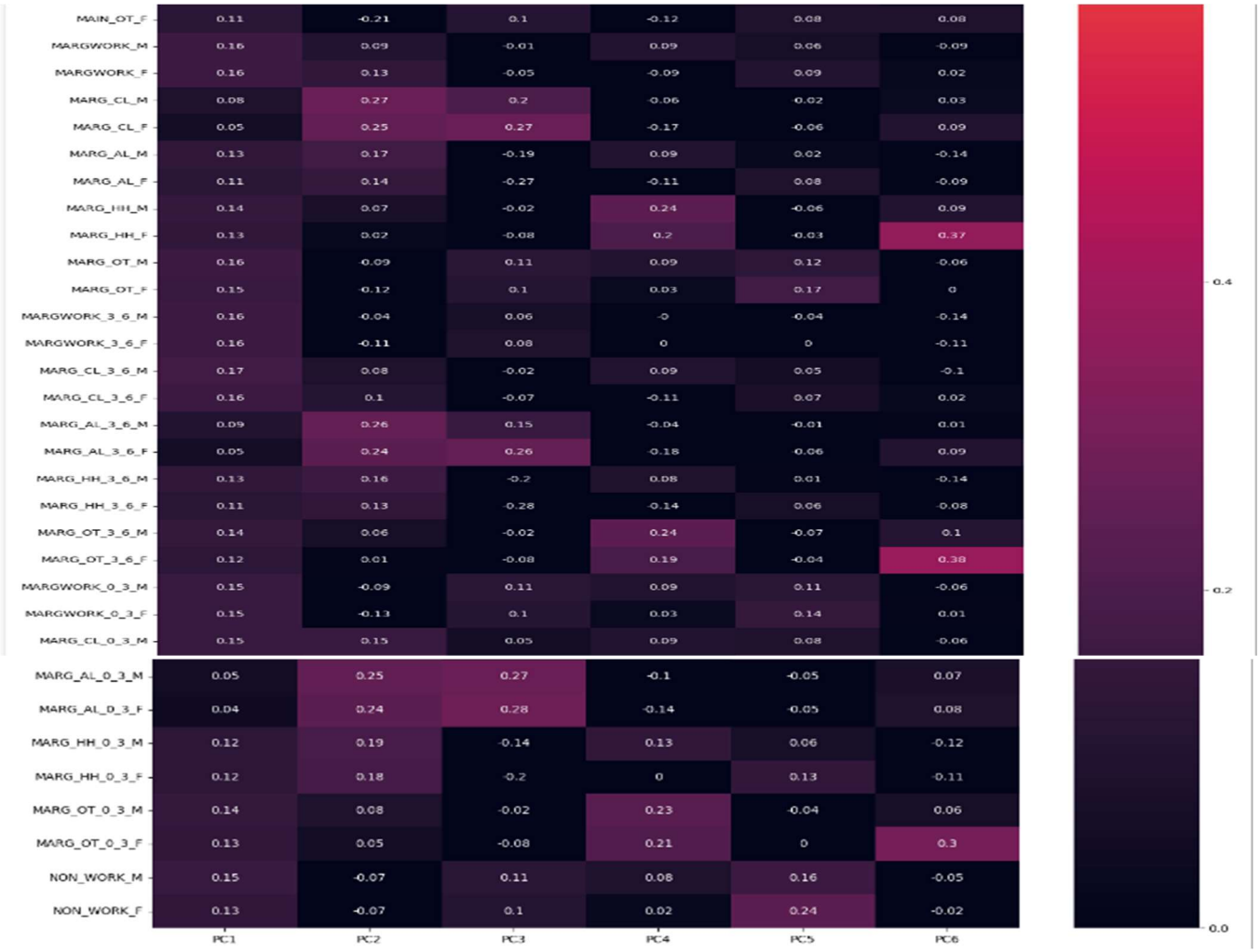


Fig 2.7A Heat-map to compare between PCs and Actual Columns

Observation from Dataframe & Heat-Map Comparing PCs and Actual Columns: -

- **PC1 has all the features of almost equal value.**
- **PC2 has high of MARG_CL_M /MARG_CL_F, also it has high of MARG_AL_3_6_M as well as MARG_AL_3_6_F, MARG_AL_0_6_M as well as F**
- **PC3 has high of MARG_CL_F, MARG_AL_3_6_F, MARG_AL_0_3_F**
- **PC4 has high of MARG_HH_M, MARG_OT_3_6_M, MARG_OT_0_3_M as well as MARG_OT_0_3_M**
- **PC5 has high of NON_WORK_F, M_ST as well as F_st**
- **PC6 has high of MAIN_HH_F, MARG_HH_F, MARG_OT_3_6_F and MARG_OT_0_3_F**

Part 2.8

PCA: Write linear equation for first PC.

Linear Equation for PC1 = a1x1 + a2x2 + a3x3+ a4x4 +…….+ a55x55 + a56x56 + a57+x57

OR

PC1 = 0.15602058* No_HH + 0.16711763*TOT_M + 0.16555318*TOT_F +.....+ MARG_OT_0_3_M*0.13219 + MARG_OT_0_3_M *224 + 0.15037558*NON_WORK_M + 0.1310662**NON_WORK_F