DATA MINING PROJECT

Submitted by,
VIDYA V

PGPDSBA.O.2023.B
02.07.2023

CONTENTS

List of Figures

Case 1: 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.

A. **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.**

	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
vailable_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	0.31	0.35	0.28	0.2	0.41
CPM	0.0	0.0	0.0	0.0	0.0
CPC	0.0	0.0	0.0	0.0	0.0

Fig.1.1. Viewing 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
```

Fig.1.2. Data info

	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	96674.47	61538.33	33600.00	72000.00	72000.00	84000.00	216000.00
Available_Impressions	23066.0	2432043.67	4742887.76	1.00	33672.25	483771.00	2527711.75	27592861.00
Matched_Queries	23066.0	1295099.14	2512969.86	1.00	18282.50	258087.50	1180700.00	14702025.00
Impressions	23066.0	1241519.52	2429399.96	1.00	7990.50	225290.00	1112428.50	14194774.00
Clicks	23066.0	10678.52	17353.41	1.00	710.00	4425.00	12793.75	143049.00
Spend	23066.0	2706.63	4067.93	0.00	85.18	1425.12	3121.40	26931.87
Fee	23066.0	0.34	0.03	0.21	0.33	0.35	0.35	0.35
Revenue	23066.0	1924.25	3105.24	0.00	55.37	926.34	2091.34	21276.18
CTR	18330.0	0.07	0.08	0.00	0.00	0.08	0.13	1.00
CPM	18330.0	7.67	6.48	0.00	1.71	7.66	12.51	81.56
CPC	18330.0	0.35	0.34	0.00	0.09	0.16	0.57	7.26

Fig.1.3. Data Description

On a cursory viewing, the following are the observations about the dataset:

- It has 23066 rows and 19 columns.

- The columns include the length,width and size of ad, the number of times the ad is shown, the total number of clicks, the money spent on each variation of the ad, the revenue from the ads, Click through rate, cost per click, and cost per impression
- Null values present in CTR,CPC,CPM, which need to be treated
- High differences in standard deviations of various columns. Scaling needs to be done
- Spend field has a minimum value of 0? Is that possible?
- Probable outliers in Impressions, matched queries, Available_Impressions and Clicks field on the lower end and higher end
- Outliers probable in ad size, revenue on the upper end.
- There are no duplicate values present

B. Part 1 - Clustering: Treat missing values in CPC, CTR and CPM using the formula given.

```
<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                    23066 non-null  float64
 17  CPM                    23066 non-null  float64
 18  CPC                    23066 non-null  float64
dtypes: float64(6), int64(7), object(6)
memory usage: 3.3+ MB
```

Fig.1.4. Dataset info output after null value treatment

C. Part 1 - Clustering: Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an analyst your judgement may be different from another analyst)

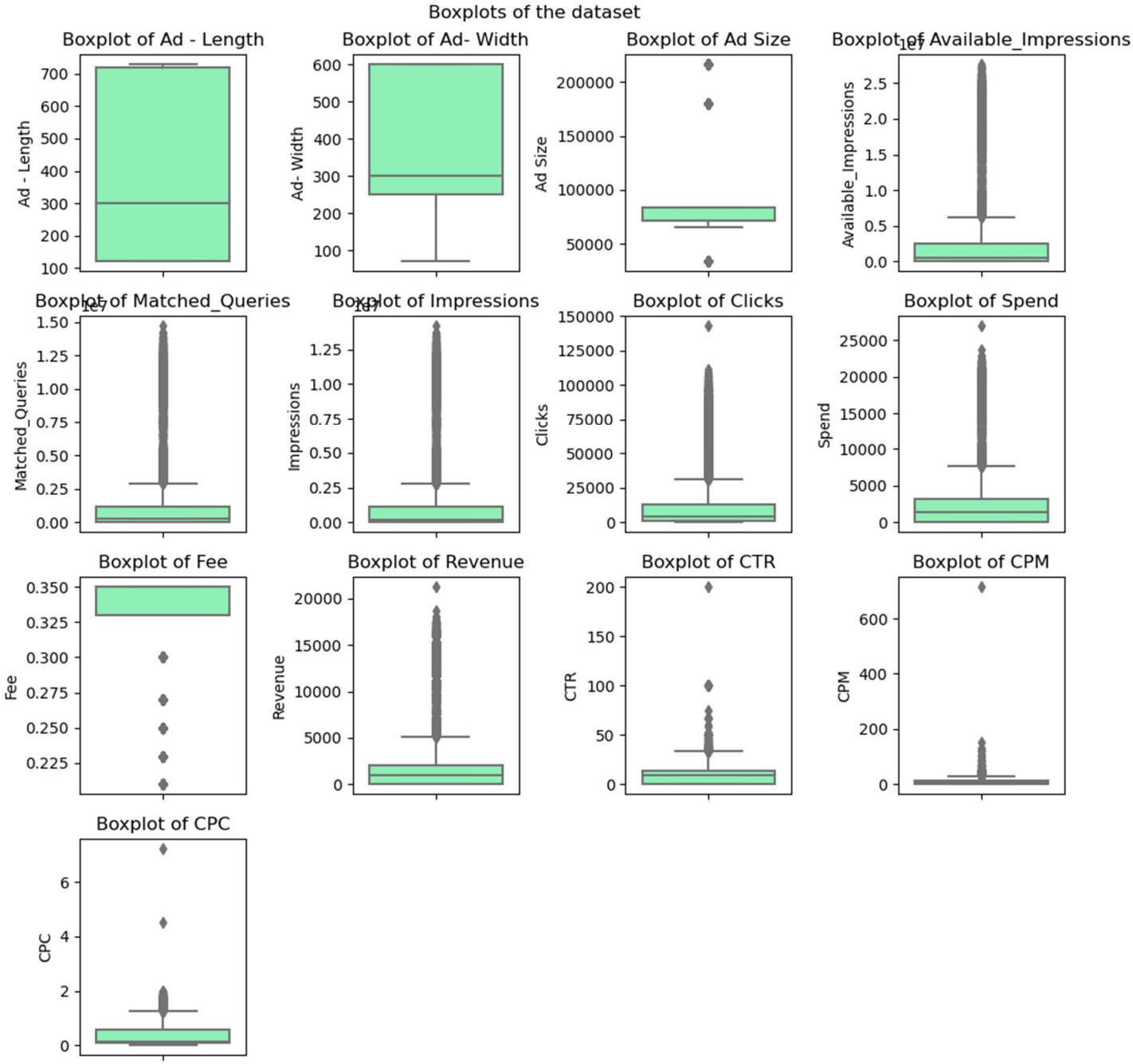


Fig.1.5. Boxplots of the fields in the dataset showing outliers

Observations:

- From the boxplots above, it can be seen that the fields Ad-Length and AD-Width do not have any outliers
- The columns Ad Size and Fee have a few outliers
- All other column have significant outliers
- Since the kmeans technique creates clusters based on distance, it is extremely outlier sensitive. Hence, the treatment of outliers is necessary
- Outlier treatment can be done on a given field by limiting the lower limit and upper limit as a function of the inter - quartile range and the first and the third quartiles of the data using the formula:
    - Lower limit= Q1-(1.5 * IQR)

- Upper Limit= Q3+(1.5 * IQR)

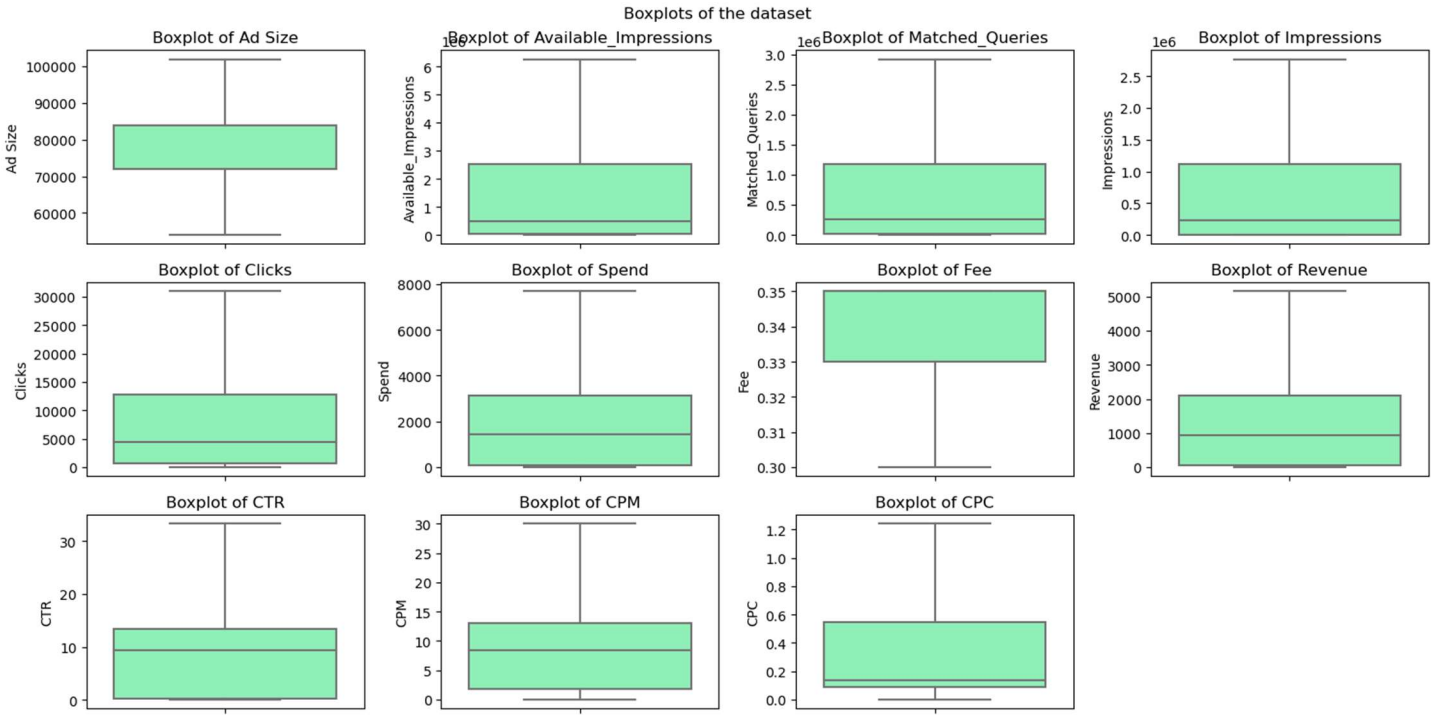


Fig. 1.6. Boxplots after outlier treatment

D. Part 1 - Clustering: Perform z-score scaling and discuss how it affects the speed of the algorithm.

	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	33.27
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	1.24

Fig.1.7. Data Before Scaling

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	-0.0	1.0	-1.13	-1.13	-0.36	1.43	1.47
Ad- Width	23066.0	0.0	1.0	-1.32	-0.43	-0.19	1.29	1.29
Ad Size	23066.0	-0.0	1.0	-1.47	-0.30	-0.30	0.48	1.65
Available_Impressions	23066.0	-0.0	1.0	-0.76	-0.74	-0.53	0.43	2.19
Matched_Queries	23066.0	0.0	1.0	-0.78	-0.76	-0.53	0.37	2.07
Impressions	23066.0	-0.0	1.0	-0.77	-0.76	-0.54	0.37	2.06
Clicks	23066.0	0.0	1.0	-0.87	-0.79	-0.41	0.47	2.36
Spend	23066.0	0.0	1.0	-0.89	-0.86	-0.31	0.39	2.27
Fee	23066.0	-0.0	1.0	-2.22	-0.57	0.54	0.54	0.54
Revenue	23066.0	0.0	1.0	-0.88	-0.85	-0.32	0.39	2.24
CTR	23066.0	0.0	1.0	-1.00	-0.96	0.14	0.64	3.03
CPM	23066.0	0.0	1.0	-1.19	-0.94	0.02	0.70	3.16
CPC	23066.0	0.0	1.0	-1.04	-0.76	-0.60	0.69	2.87

Fig.1.8. Data after scaling

Observations:

- After scaling, data has now been centered around 0, with standard deviation 1 for all the columns
- This will ensure that each of the columns carry the same weightage while performing clustering with a distance-based algorithm, and ensures that the variables with a higher value do not dominate those with lower values
- In addition, scaling aids the rapid convergence of clusters, thereby improving the speed and efficiency of the algorithm

E. Part 1 - Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.

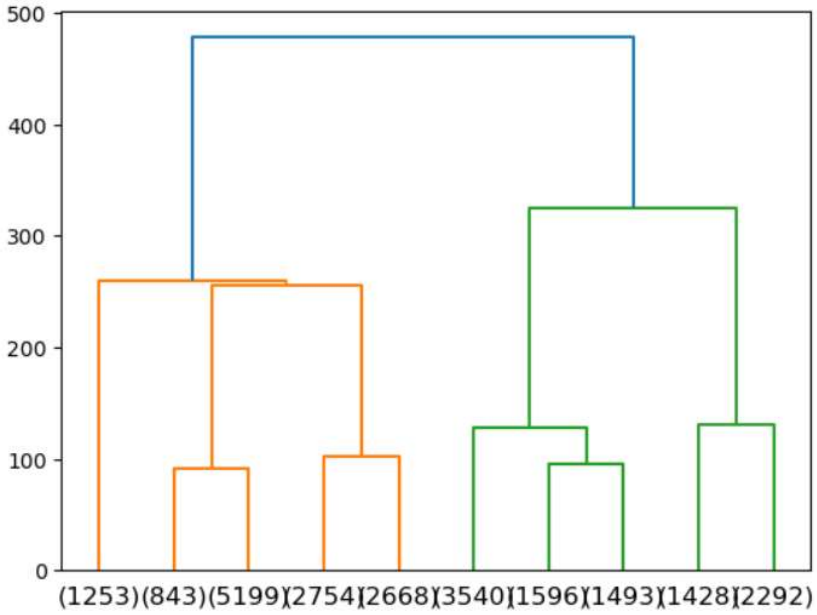


Fig.1.9. Dendrogram showing last 10 clusters

After hierarchical clustering, the given dataset was grouped into 5 clusters. The number of rows in each cluster is as follows:

```
1    1253
2    6042
3    5422
4    6629
5    3720
Name: hcluster, dtype: int64
```

Fig. 1.10. Number of rows in each cluster after hierarchical clustering

The below image shows the mean Spend, Clicks and Revenue across different clusters.

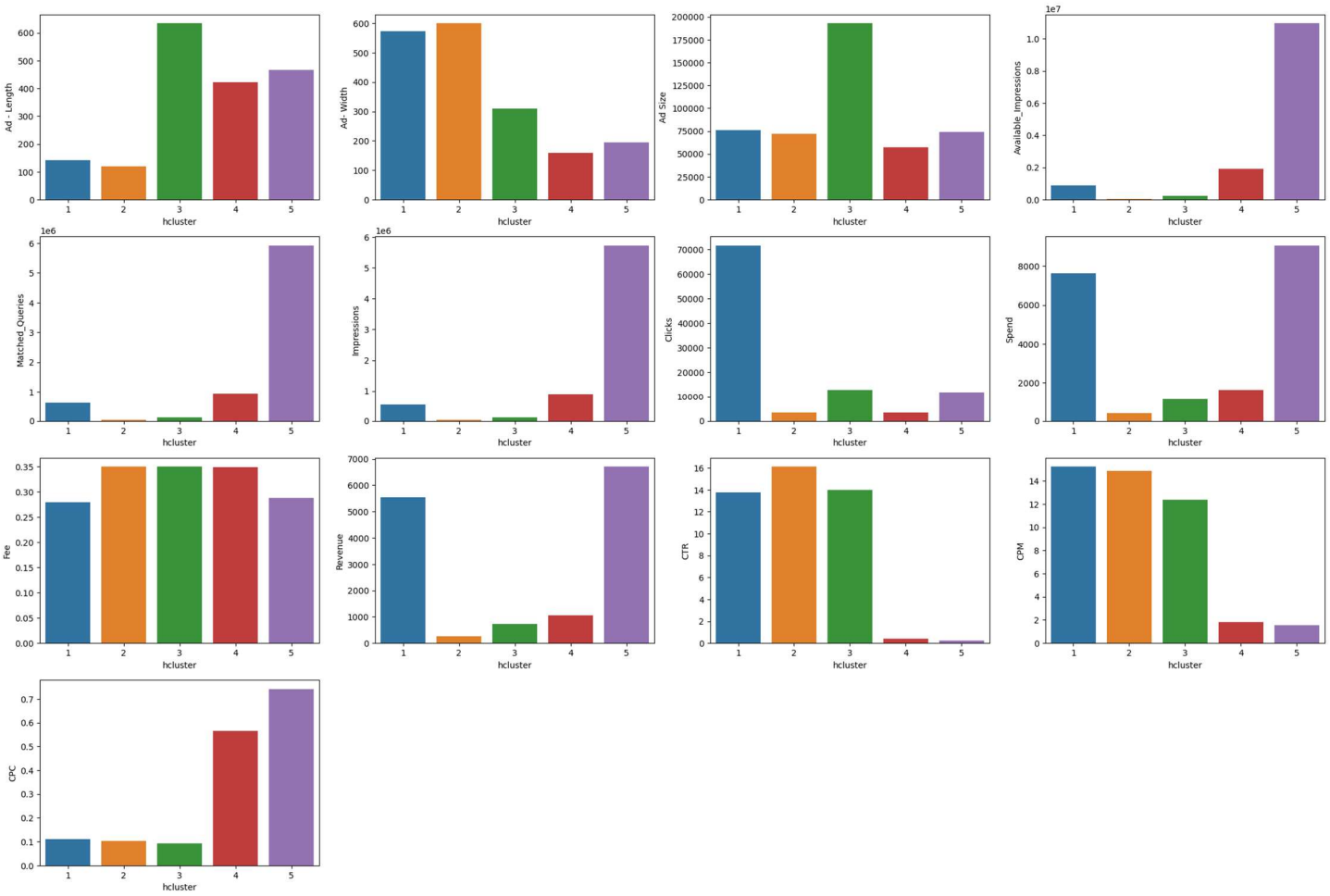


Fig.1.11. Mean of various columns across clusters after clustering

Fig.1.12. Sum totals of various columns across clusters

Observations:

- Cluster 1 contains medium sized wide ads. It has the highest mean number of clicks, though a moderate CTR. It also has a high CPM.
- Cluster 2 contains medium sized, very less frequently appearing ads. These have high CPM, and CTR, but the revenue is low.
- Cluster 3 contains big sized ads. These do not appear very frequently, but have a second best click through rate. The revenue from this is pretty low, but the fee is a little high. The available impressions in this cluster is low.
- Cluster 4 contains the smallest size ads. These have the second best mean available impressions, and the second-best matched queries. The mean spend for this cluster is low, and it has a low CTR, but high CPC.
- Cluster 5 contains moderate sized ads, with high available impressions and impressions. It has the highest matched_queries, which probably indicates that the content in it is in current trend. However, the CTR is pretty low. It has the highest Spend and CPC, but also the highest revenue. This cluster can be focused for an increase in revenue, by improving the CTR. This may be done by increasing the ad size.

F. Part 1 - Clustering: Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.

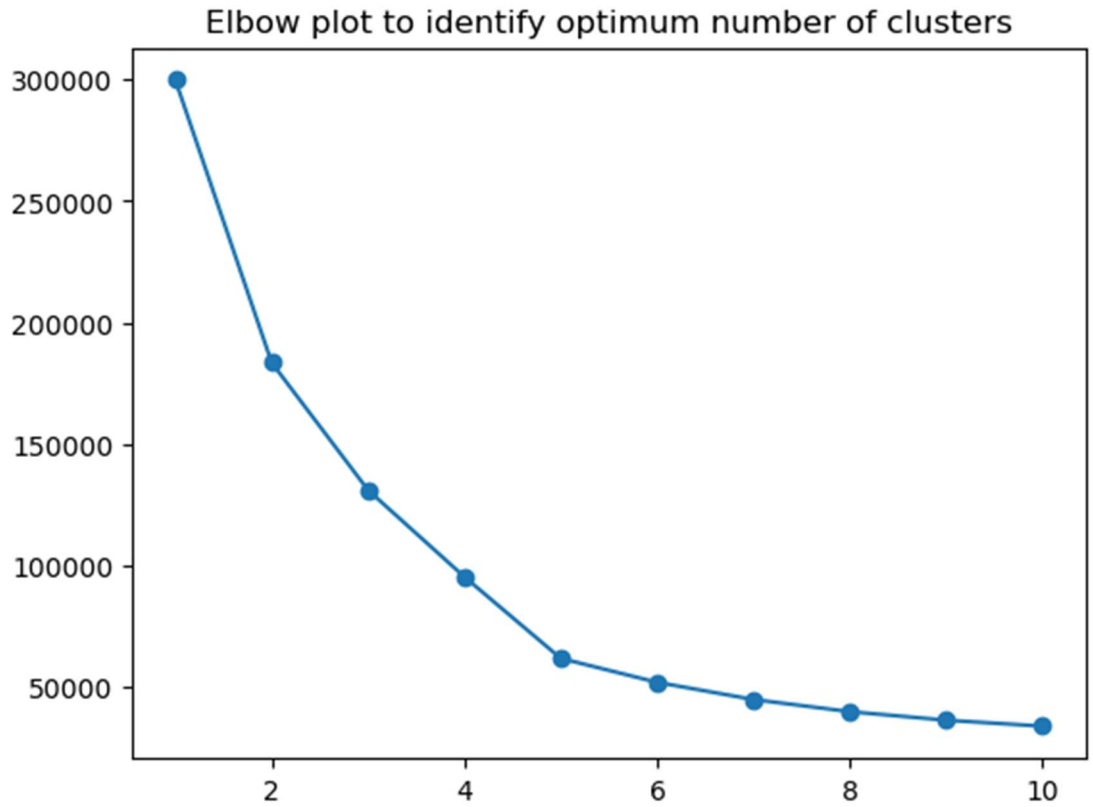


Fig.1.13. Elbow plot for k-means algorithm

From the above plot, it can be inferred that the optimum number of clusters is 5.

G. Part 1 - Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters.
Output:

```
Silhoutte scores for 2 to 10 clusters : [0.38563243892114524, 0.3824397416745965, 0.4452509118782309, 0.5240238007237571, 0.522
0833406497346, 0.516542969590082, 0.4797128672086253, 0.4367410164049805, 0.4448161952779515]
```

Observations:

- From the above array of silhouette scores, we find that the same is the highest when the number of clusters is 5.
- This is also in tune with the optimum number of clusters determined from the elbow plot.
- Hence, we can proceed with n_clusters=5

H. Part 1 - 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]

After performing k-means clustering, the following are the number of rows in each cluster:

```
0    4675
1    4054
2    6525
3    6275
4    1537
Name: Cluster_km, dtype: int64
```

Fig.1.14. Cluster distribution

The above is output is represented as bar plot below, with distinction of device types.

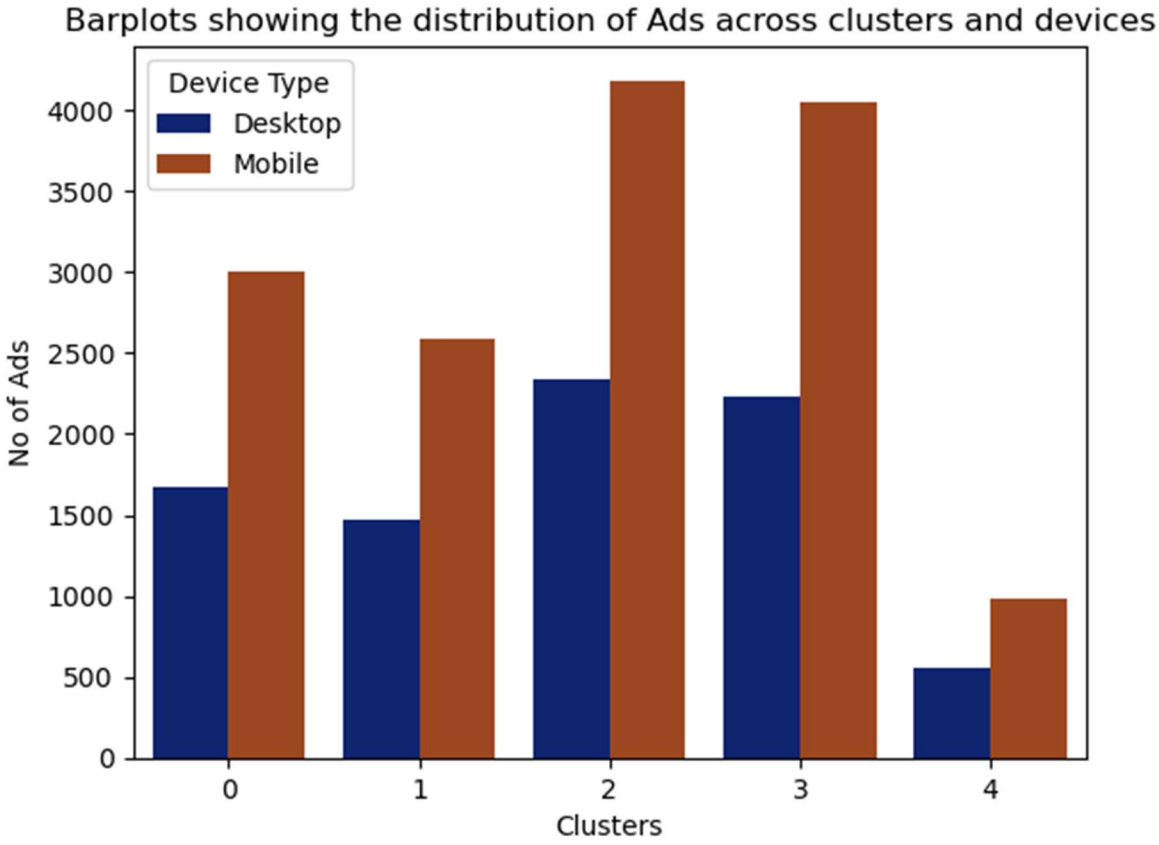


Fig.1.15 Barplot showing the distribution of Ads across clusters and Devices.

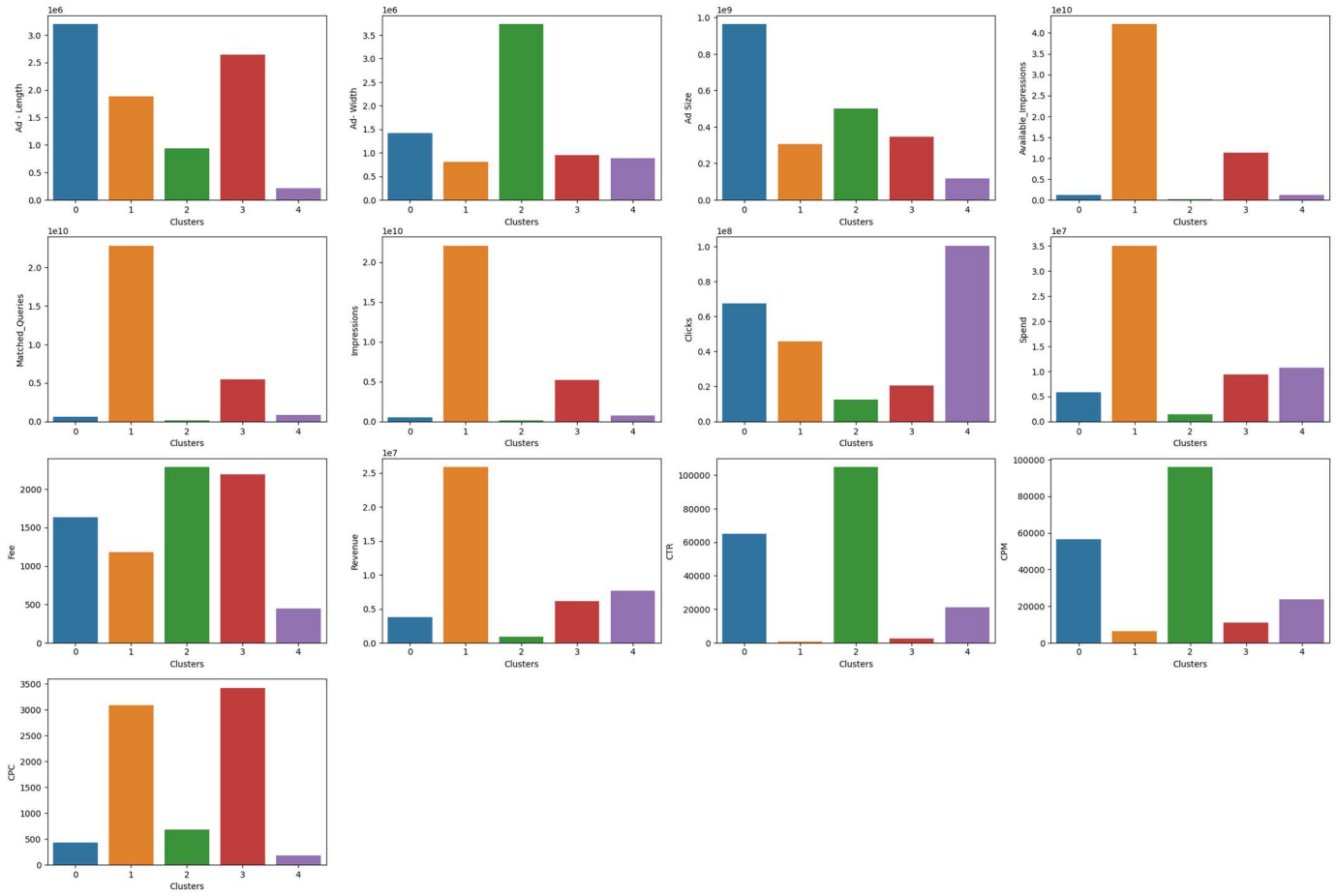


Fig.1.16. Sums of factors across clusters generated by k-means algorithm

Fig.1.17. Means of factors across clusters generated by k-means algorithm.

I. Part 1 - Clustering: Conclude the project by providing summary of your learnings.

Summary:

- Cluster 0: Contains the biggest ads. They have very low available impressions, and impressions, and have a moderate CPM. The Spend and revenue are low, though it has a decent CTR
- Cluster 1: Contains moderately sized ads that have a high mean matched_queries and available_impressions. The spend revenue and CPC are all high. It has the lowest CTR. This cluster can be the focus cluster since by improving the CTR, it holds the potential to generate high revenue.
- Cluster 2: Contains the highest number of ads. These are medium sized wide ads with low values across fields, except for CTR and CPC.
- Cluster 3: Contains the smallest ads, with second highest mean available_impressions and matched_queries. This has very low clicks and CTR
- Cluster 4: Contains the lowest number of ads, but has the highest mean number of clicks and high CTR. The mean CPM is highest for this cluster. The Revenue is very high for this cluster.

Case 2: PCA- India Census Data

A. Part 2 - PCA: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.

```
<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
 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
```

Fig.2.1. Census data- info

	count	mean	std	min	25%	50%	75%	max
State Code	640.0	17.11	9.43	1.0	9.00	18.0	24.00	35.0
Dist.Code	640.0	320.50	184.90	1.0	160.75	320.5	480.25	640.0
No_HH	640.0	51222.87	48135.41	350.0	19484.00	35837.0	68892.00	310450.0
TOT_M	640.0	79940.58	73384.51	391.0	30228.00	58339.0	107918.50	485417.0
TOT_F	640.0	122372.08	113600.72	698.0	46517.75	87724.5	164251.75	750392.0
M_06	640.0	12309.10	11500.91	56.0	4733.75	9159.0	16520.25	96223.0
F_06	640.0	11942.30	11326.29	56.0	4672.25	8663.0	15902.25	95129.0
M_SC	640.0	13820.95	14426.37	0.0	3466.25	9591.5	19429.75	103307.0
F_SC	640.0	20778.39	21727.89	0.0	5603.25	13709.0	29180.00	156429.0
M_ST	640.0	6191.81	9912.67	0.0	293.75	2333.5	7658.00	96785.0
F_ST	640.0	10155.64	15875.70	0.0	429.50	3834.5	12480.25	130119.0
M_LIT	640.0	57967.98	55910.28	286.0	21298.00	42693.5	77989.50	403261.0
F_LIT	640.0	66359.57	75037.86	371.0	20932.00	43796.5	84799.75	571140.0
M_ILL	640.0	21972.60	19825.61	105.0	8590.00	15767.5	29512.50	105961.0
F_ILL	640.0	56012.52	47116.69	327.0	22367.00	42386.0	78471.00	254160.0
TOT_WORK_M	640.0	37992.41	36419.54	100.0	13753.50	27936.5	50226.75	269422.0
TOT_WORK_F	640.0	41295.76	37192.36	357.0	16097.75	30588.5	53234.25	257848.0
MAINWORK_M	640.0	30204.45	31480.92	65.0	9787.00	21250.5	40119.00	247911.0
MAINWORK_F	640.0	28198.85	29998.26	240.0	9502.25	18484.0	35063.25	226166.0
MAIN_CL_M	640.0	5424.34	4739.16	0.0	2023.50	4160.5	7695.00	29113.0
MAIN_CL_F	640.0	5486.04	5326.36	0.0	1920.25	3908.5	7286.25	36193.0
MAIN_AL_M	640.0	5849.11	6399.51	0.0	1070.25	3936.5	8067.25	40843.0
MAIN_AL_F	640.0	8926.00	12864.29	0.0	1408.75	3933.5	10617.50	87945.0
MAIN_HH_M	640.0	883.89	1278.64	0.0	187.50	498.5	1099.25	16429.0
MAIN_HH_F	640.0	1380.77	3179.41	0.0	248.75	540.5	1435.75	45979.0
MAIN_OT_M	640.0	18047.10	26068.48	36.0	3997.50	9598.0	21249.50	240855.0
MAIN_OT_F	640.0	12406.04	18972.20	153.0	3142.50	6380.5	14368.25	209355.0

MARGWORK_M	640.0	7787.96	7410.79	35.0	2937.50	5627.0	9800.25	47553.0
MARGWORK_F	640.0	13096.91	10996.47	117.0	5424.50	10175.0	18879.25	66915.0
MARG_CL_M	640.0	1040.74	1311.55	0.0	311.75	606.5	1281.00	13201.0
MARG_CL_F	640.0	2307.68	3564.63	0.0	630.25	1226.0	2659.25	44324.0
MARG_AL_M	640.0	3304.33	3781.56	0.0	873.50	2062.0	4300.75	23719.0
MARG_AL_F	640.0	6463.28	6773.88	0.0	1402.50	4020.5	9089.25	45301.0
MARG_HH_M	640.0	316.74	462.66	0.0	71.75	166.0	356.50	4298.0
MARG_HH_F	640.0	786.63	1198.72	0.0	171.75	429.0	962.50	15448.0
MARG_OT_M	640.0	3126.15	3609.39	7.0	935.50	2036.0	3985.25	24728.0
MARG_OT_F	640.0	3539.32	4115.19	19.0	1071.75	2349.5	4400.50	36377.0
MARGWORK_3_6_M	640.0	41948.17	39045.32	291.0	16208.25	30315.0	57218.75	300937.0
MARGWORK_3_6_F	640.0	81076.32	82970.41	341.0	26619.50	56793.0	107924.00	676450.0
MARG_CL_3_6_M	640.0	6394.99	6019.81	27.0	2372.00	4630.0	8167.00	39106.0
MARG_CL_3_6_F	640.0	10339.86	8467.47	85.0	4351.50	8295.0	15102.00	50065.0
MARG_AL_3_6_M	640.0	789.85	905.64	0.0	235.50	480.5	986.00	7426.0
MARG_AL_3_6_F	640.0	1749.58	2496.54	0.0	497.25	985.5	2059.00	27171.0
MARG_HH_3_6_M	640.0	2743.64	3059.59	0.0	718.75	1714.5	3702.25	19343.0
MARG_HH_3_6_F	640.0	5169.85	5335.64	0.0	1113.75	3294.0	7502.25	36253.0
MARG_OT_3_6_M	640.0	245.36	358.73	0.0	58.00	129.5	276.00	3535.0
MARG_OT_3_6_F	640.0	585.88	900.03	0.0	127.75	320.5	719.25	12094.0
MARGWORK_0_3_M	640.0	2616.14	3036.96	7.0	755.00	1681.5	3320.25	20648.0
MARGWORK_0_3_F	640.0	2834.55	3327.84	14.0	833.50	1834.5	3610.50	25844.0
MARG_CL_0_3_M	640.0	1392.97	1489.71	4.0	489.50	949.0	1714.00	9875.0
MARG_CL_0_3_F	640.0	2757.05	2788.78	30.0	957.25	1928.0	3599.75	21611.0
MARG_AL_0_3_M	640.0	250.89	453.34	0.0	47.00	114.5	270.75	5775.0
MARG_AL_0_3_F	640.0	558.10	1117.64	0.0	109.00	247.5	568.75	17153.0
MARG_HH_0_3_M	640.0	560.69	762.58	0.0	136.50	308.0	642.00	6116.0
MARG_HH_0_3_F	640.0	1293.43	1585.38	0.0	298.00	717.0	1710.75	13714.0
MARG_OT_0_3_M	640.0	71.38	107.90	0.0	14.00	35.0	79.00	895.0
MARG_OT_0_3_F	640.0	200.74	309.74	0.0	43.00	113.0	240.00	3354.0
NON_WORK_M	640.0	510.01	610.60	0.0	161.00	326.0	604.50	6456.0
NON_WORK_F	640.0	704.78	910.21	5.0	220.50	464.5	853.50	10533.0

Fig.2.2. Data description

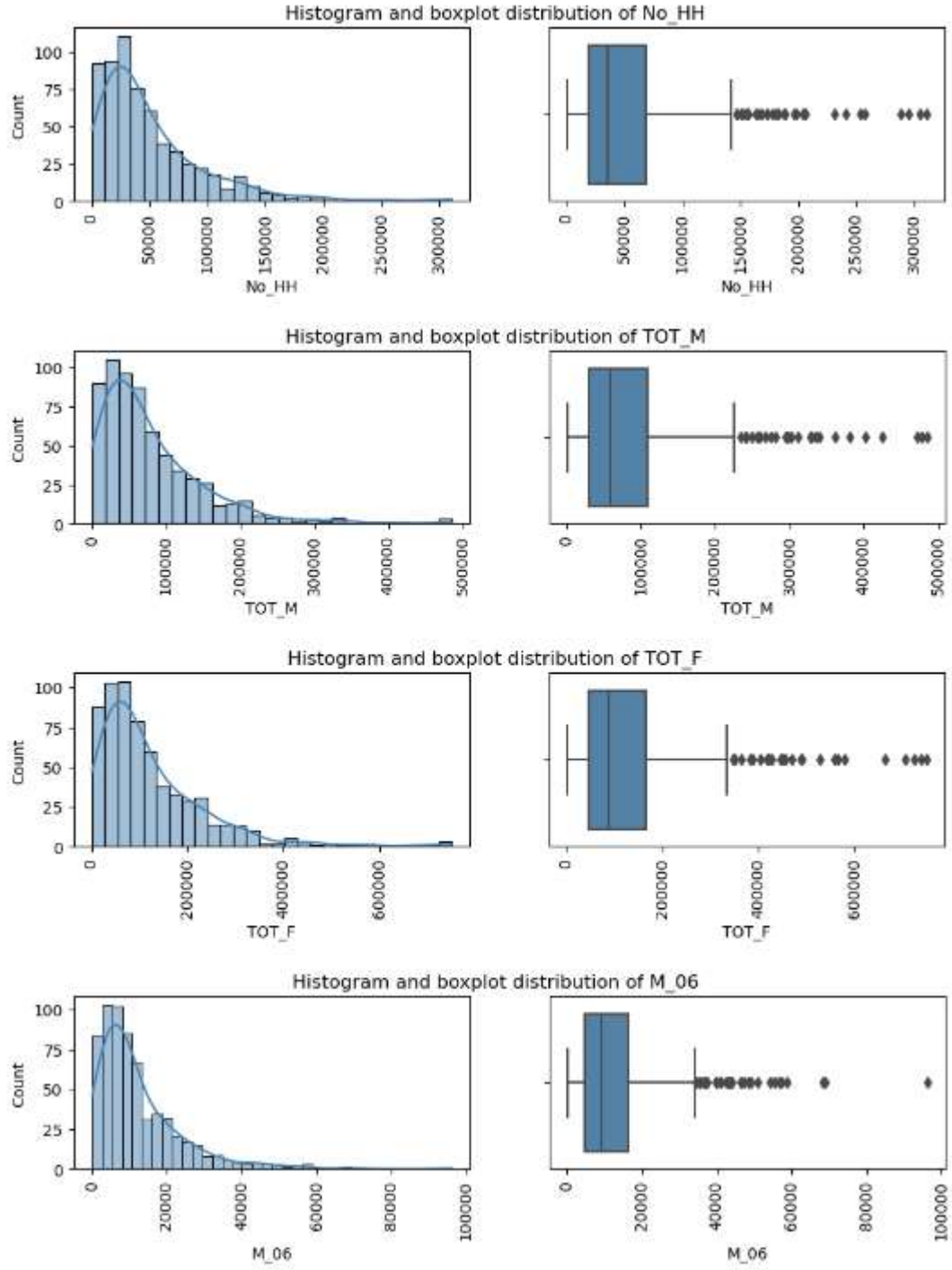
Observations:

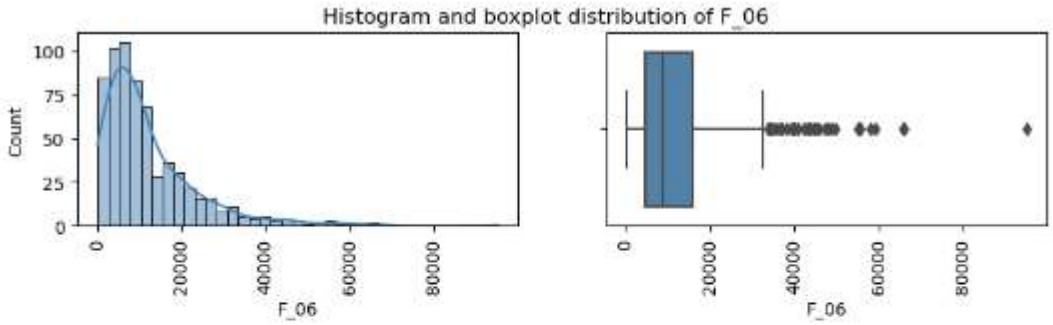
- The given dataset has 640 data points, and 61 fields containing elaborate details of various areas, including gender, population, literacy, occupation etc.
- There are no null values present in the given dataset
- The data magnitudes varies across fields. Scaling is required
- Few fields have a minimum value of 0, which seem plausible

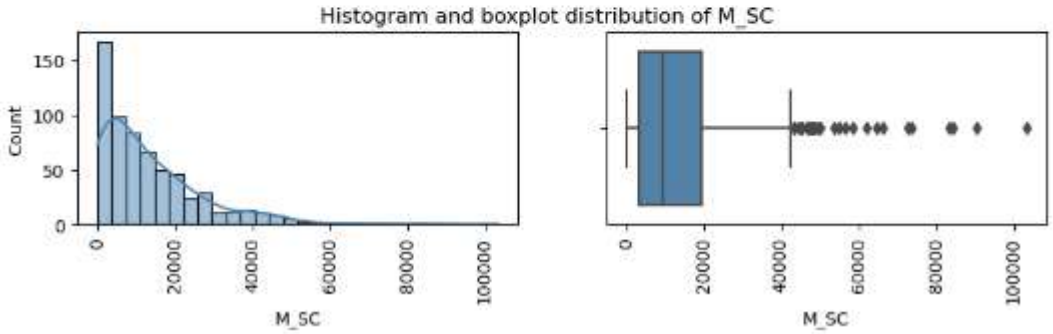
- Most of the distributions are not normal
- Few occupation related fields may have outliers
- There are no duplicates and bad data.

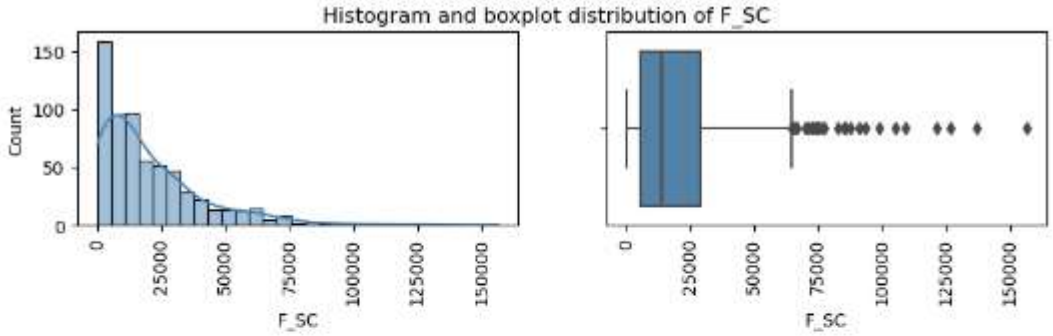
B. Part 2 - PCA: Perform detailed Exploratory analysis by creating certain questions like (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio? (Example Questions). Pick 5 variables out of the given 24 variables below for EDA: No_HH, 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

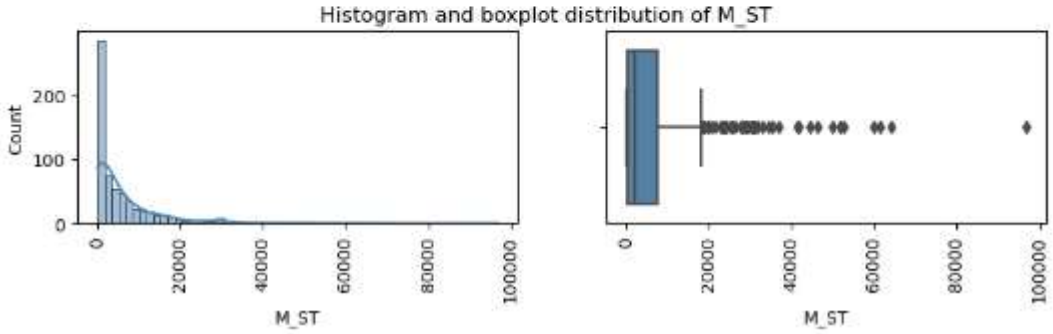
Univariate Analysis:

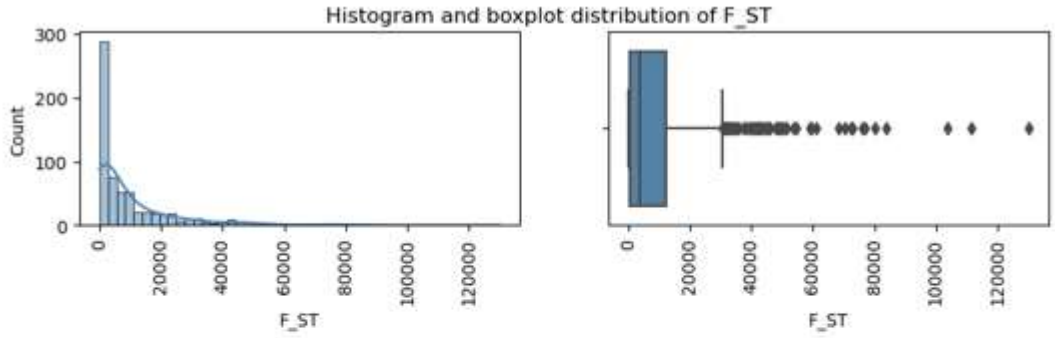


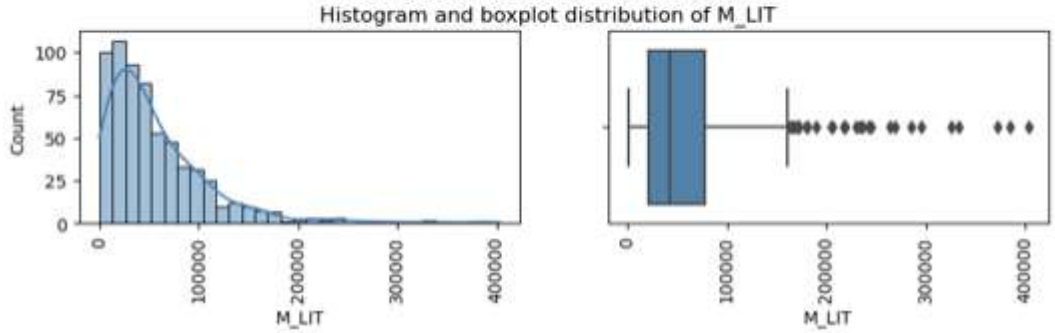

Histogram and boxplot distribution of F_06
Count
F_06
F_06

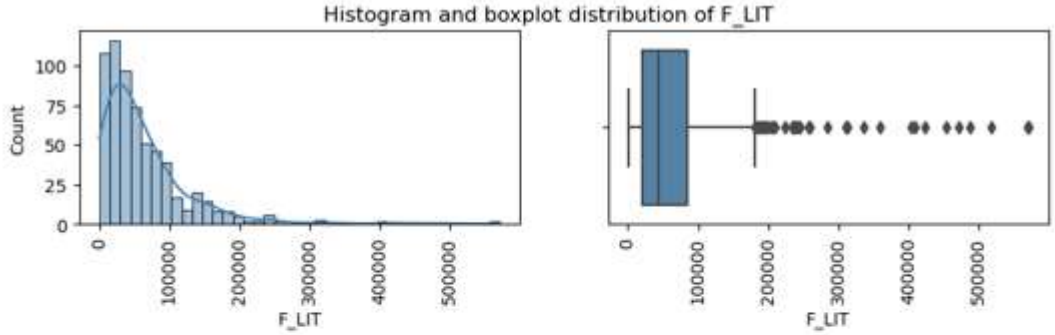

Histogram and boxplot distribution of M_SC
Count
M_SC
M_SC

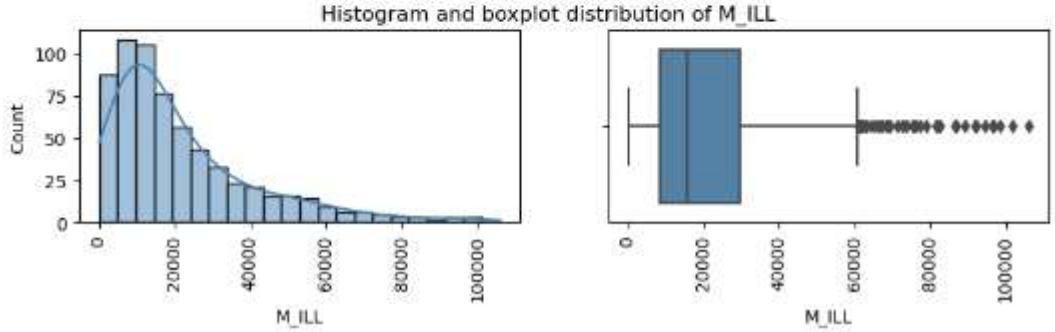

Histogram and boxplot distribution of F_SC
Count
F_SC
F_SC

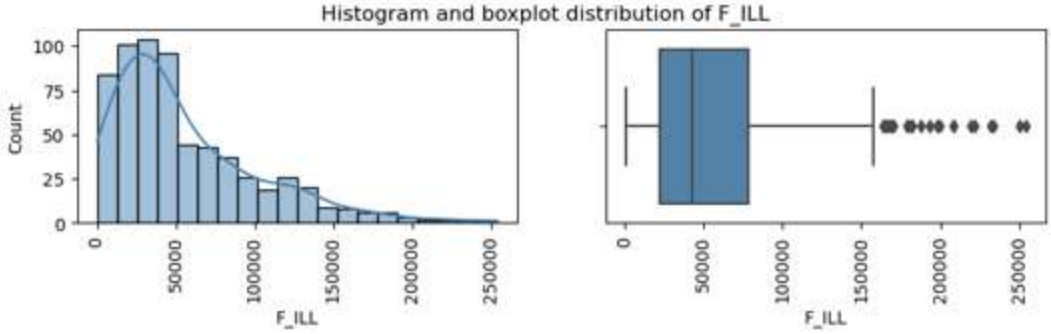

Histogram and boxplot distribution of M_ST
Count
M_ST
M_ST

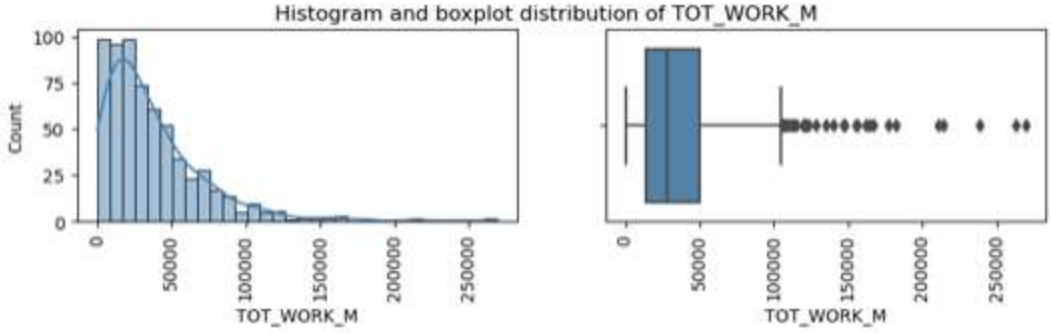

Histogram and boxplot distribution of F_ST
Count
F_ST

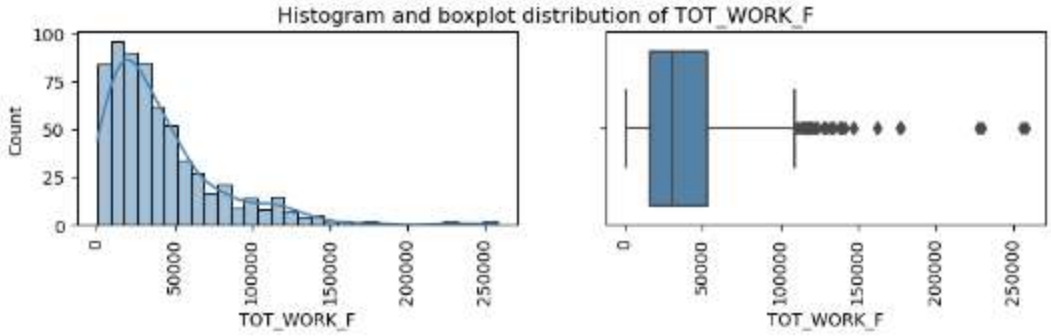

Histogram and boxplot distribution of M_LIT
Count
M_LIT

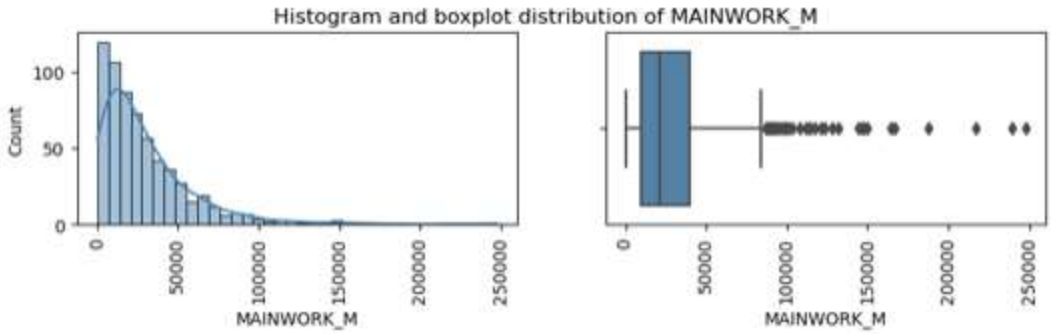

Histogram and boxplot distribution of F_LIT
Count
F_LIT

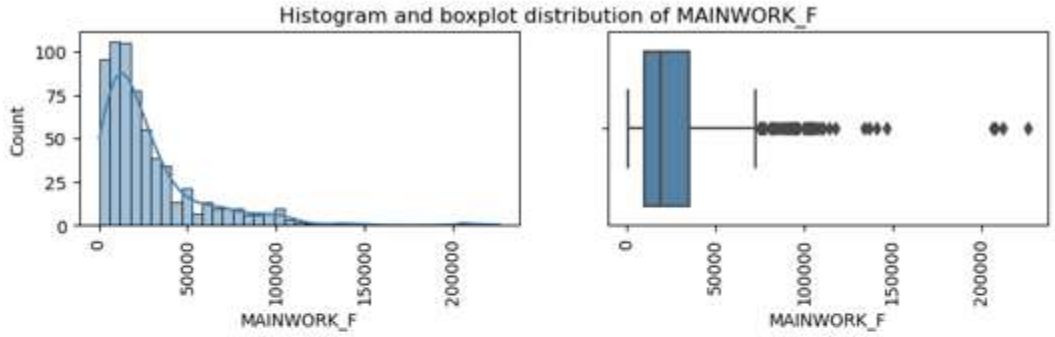

Histogram and boxplot distribution of M_ILL
Count
M_ILL

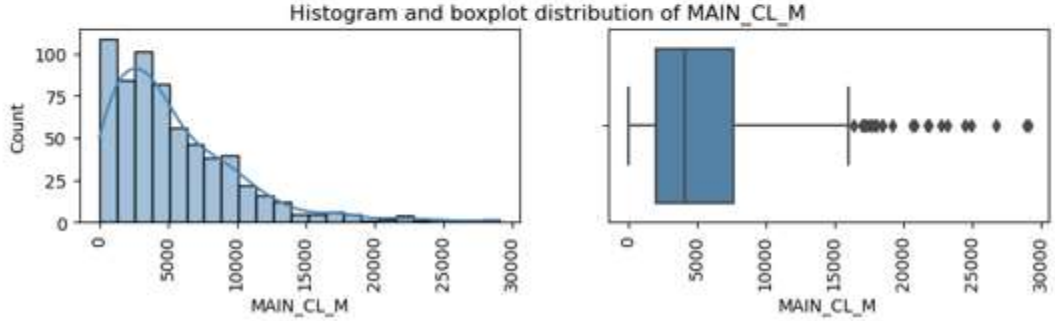

Histogram and boxplot distribution of F_ILL
Count
F_ILL

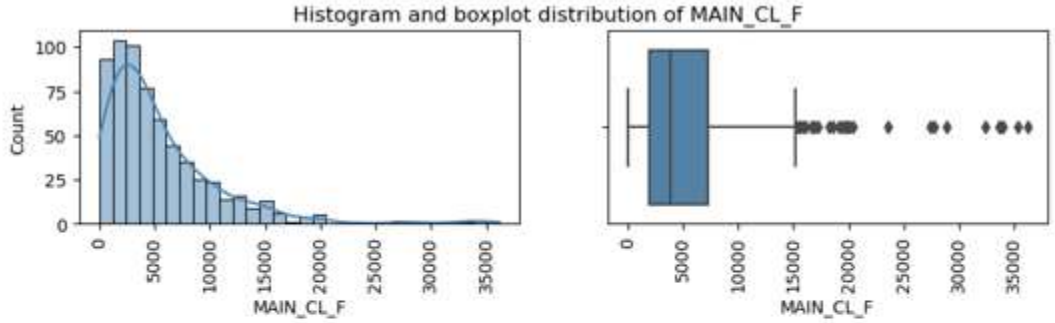

Histogram and boxplot distribution of TOT_WORK_M
Count
TOT_WORK_M

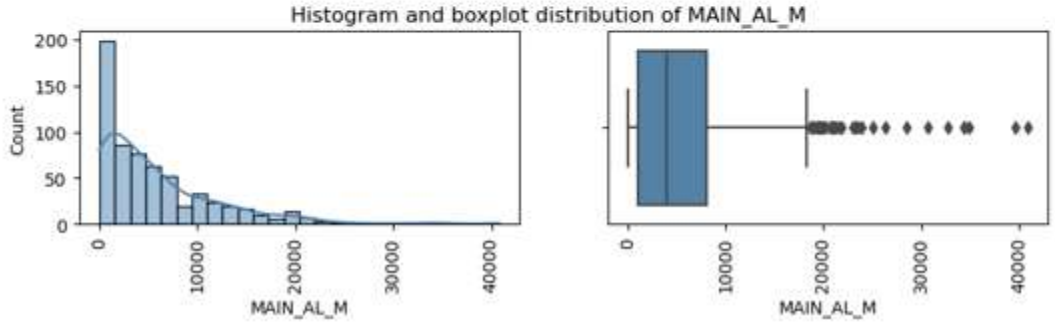

Histogram and boxplot distribution of TOT_WORK_F
Count
TOT_WORK_F

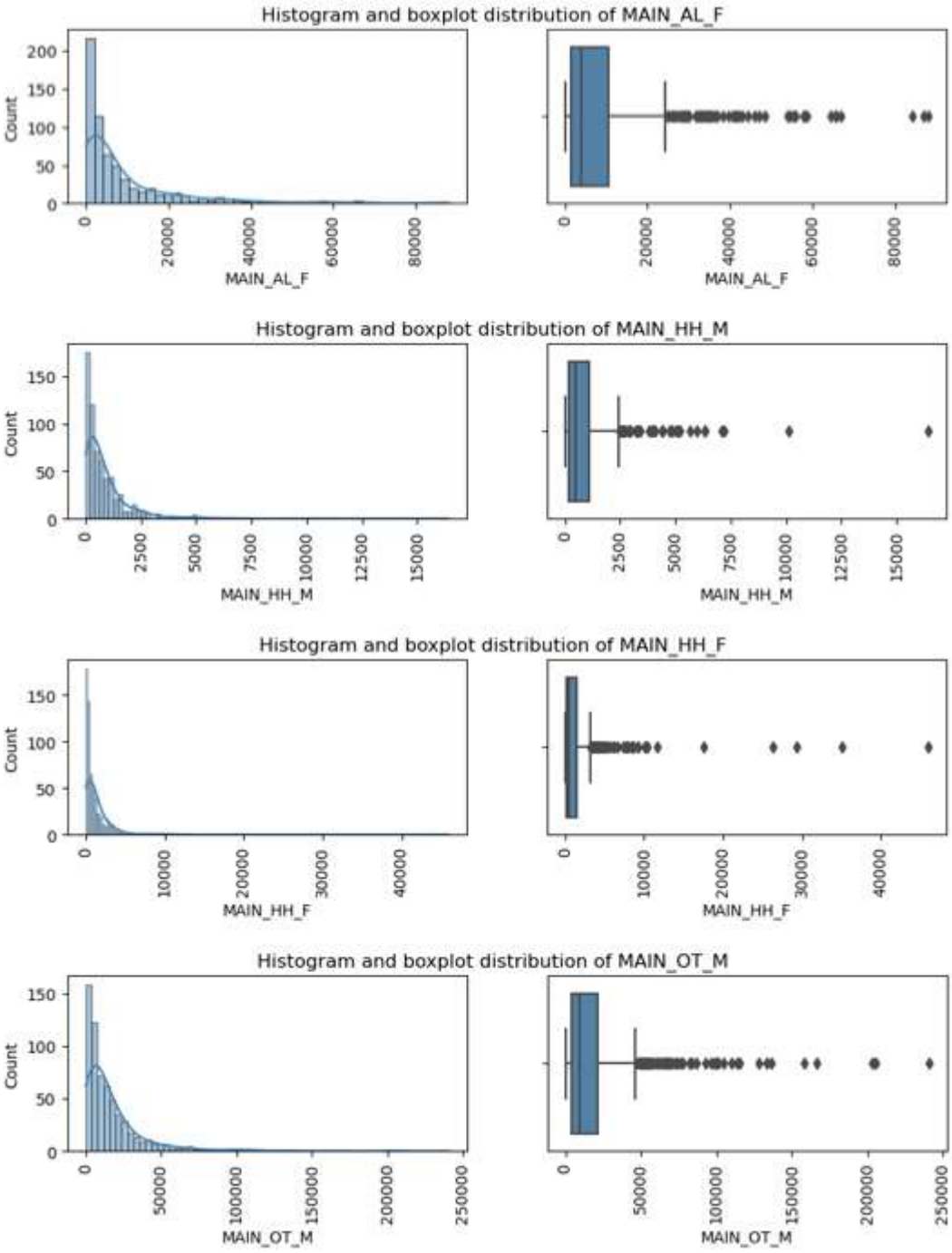

Histogram and boxplot distribution of MAINWORK_M
Count
MAINWORK_M

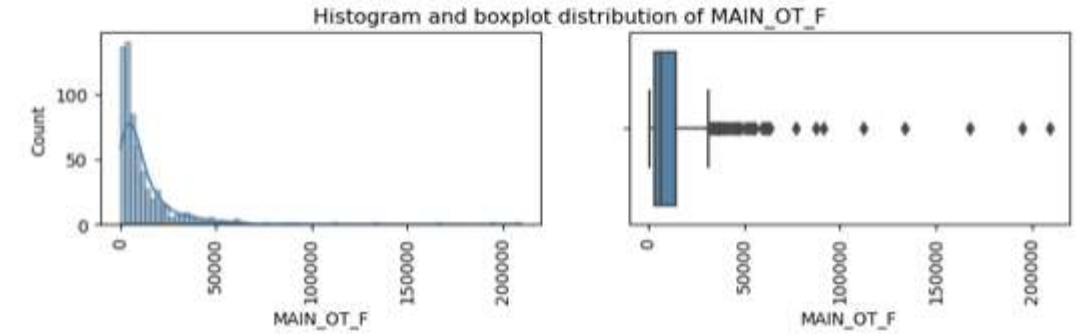

Histogram and boxplot distribution of MAINWORK_F
Count
MAINWORK_F
0
50000
100000
150000
200000

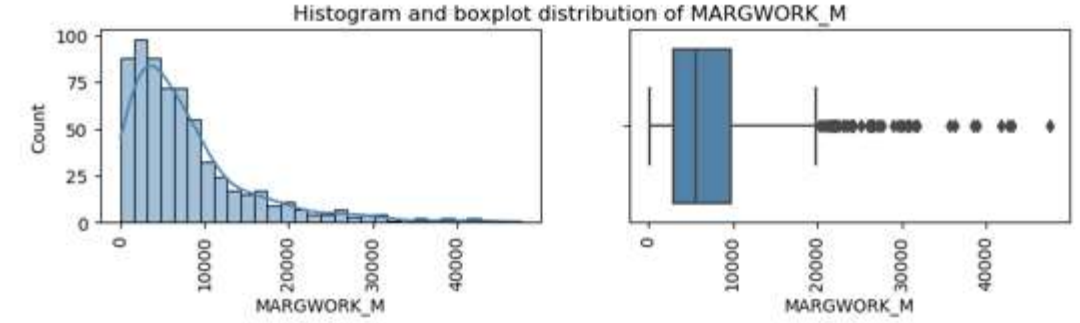

Histogram and boxplot distribution of MAIN_CL_M
Count
MAIN_CL_M
0
5000
10000
15000
20000
25000
30000

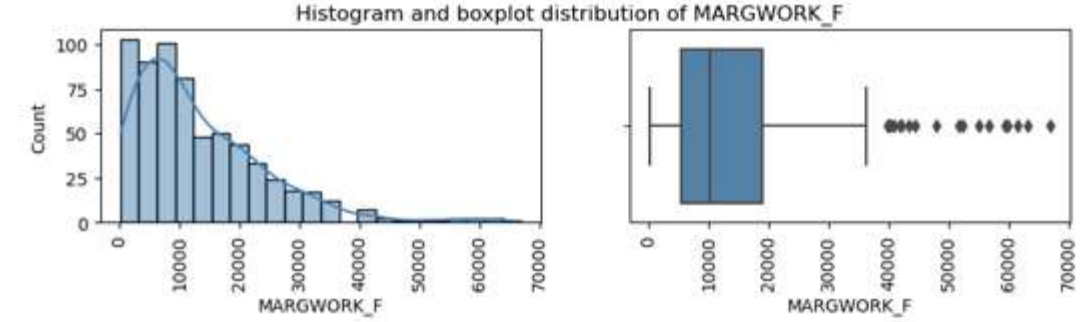

Histogram and boxplot distribution of MAIN_CL_F
Count
MAIN_CL_F
0
5000
10000
15000
20000
25000
30000
35000

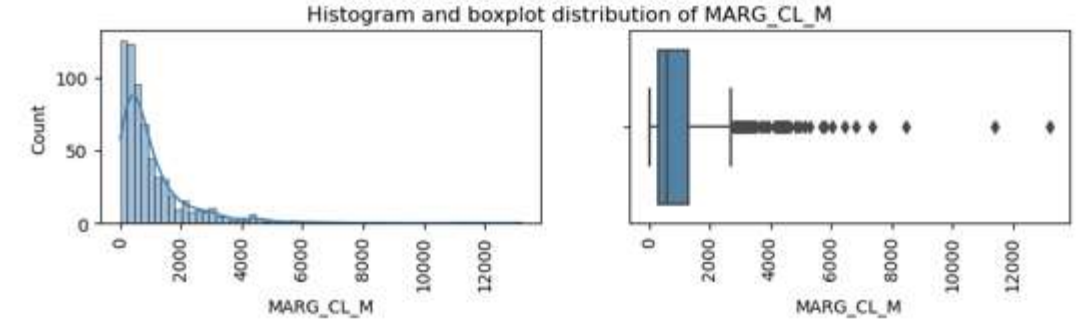

Histogram and boxplot distribution of MAIN_AL_M
Count
MAIN_AL_M
0
10000
20000
30000
40000

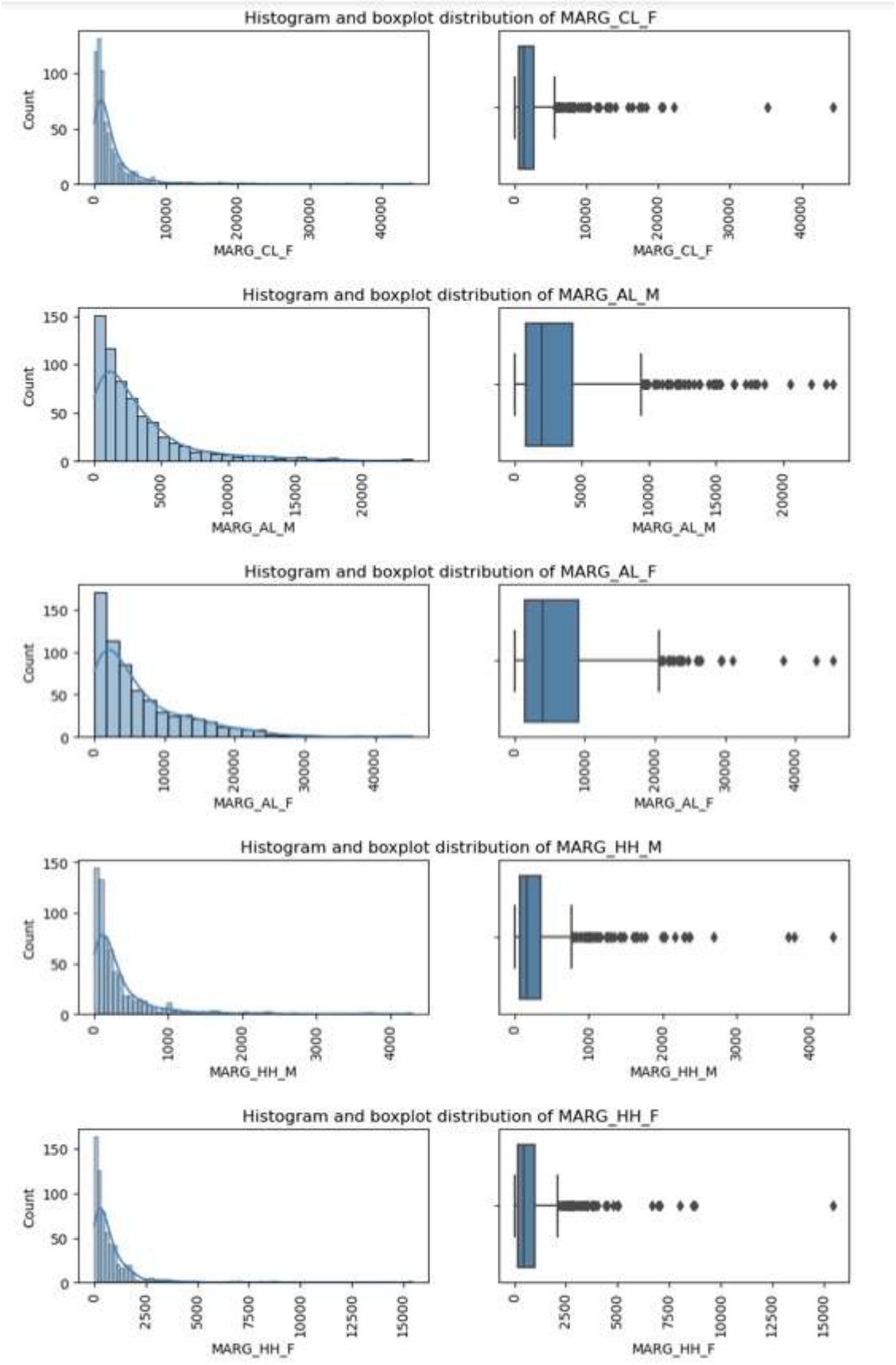

Histogram and boxplot distribution of MAIN_AL_F
Count
MAIN_AL_F
Histogram and boxplot distribution of MAIN_HH_M
Count
MAIN_HH_M
Histogram and boxplot distribution of MAIN_HH_F
Count
MAIN_HH_F
Histogram and boxplot distribution of MAIN_OT_M
Count
MAIN_OT_M

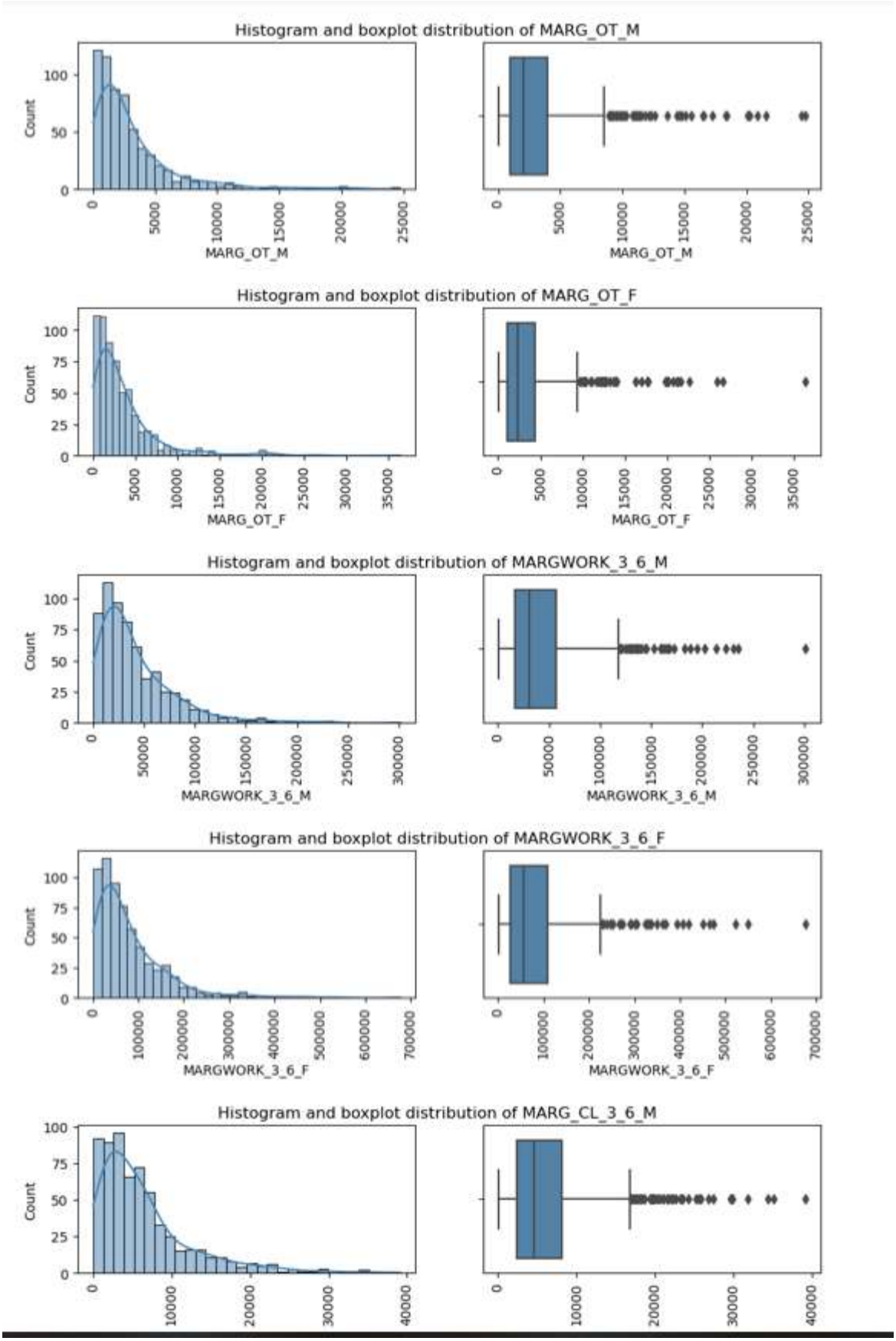

Histogram and boxplot distribution of MAIN_OT_F
Count
MAIN_OT_F
MAIN_OT_F

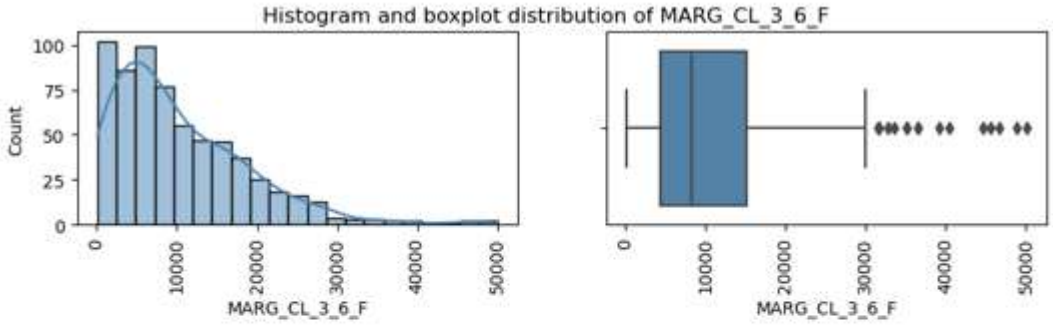

Histogram and boxplot distribution of MARGWORK_M
Count
MARGWORK_M
MARGWORK_M

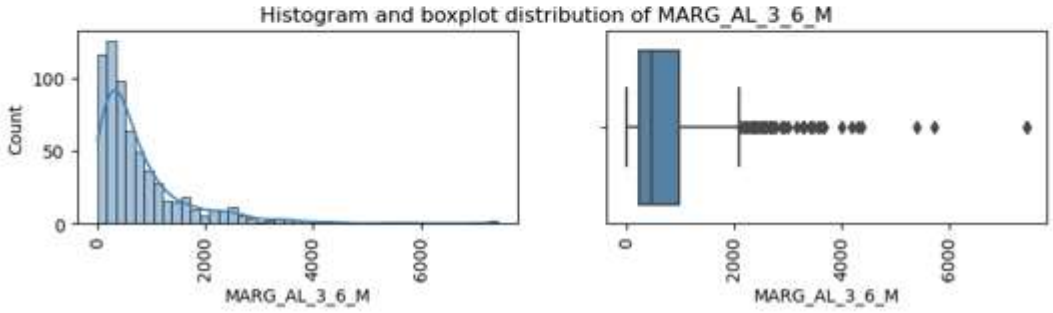

Histogram and boxplot distribution of MARGWORK_F
Count
MARGWORK_F
MARGWORK_F

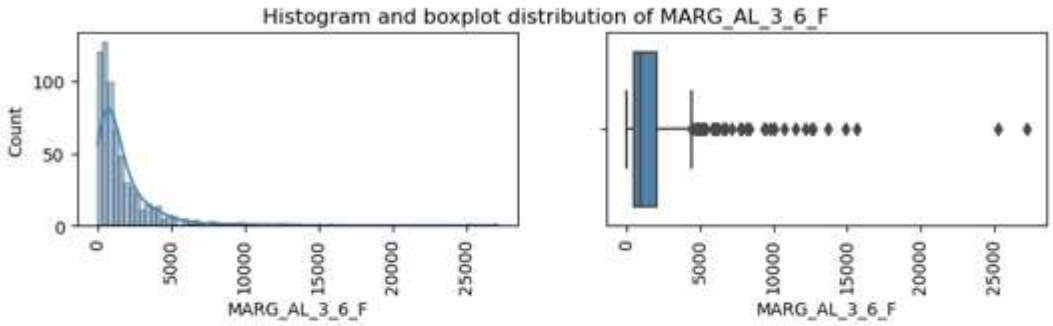

Histogram and boxplot distribution of MARG_CL_M
Count
MARG_CL_M
MARG_CL_M

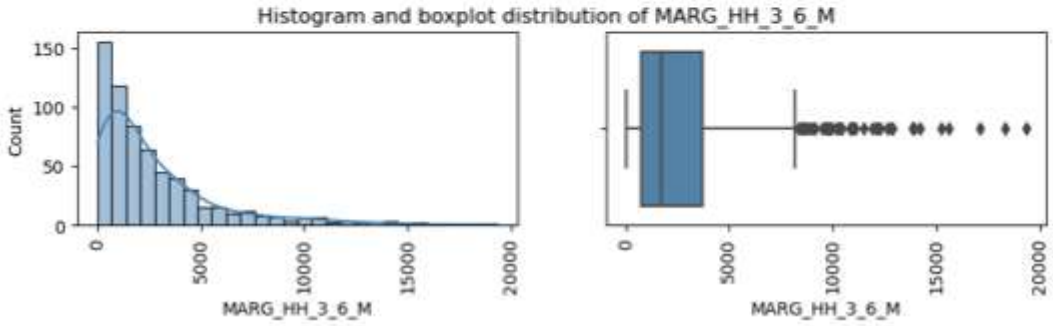

Histogram and boxplot distribution of MARG_CL_F
Count
MARG_CL_F
Histogram and boxplot distribution of MARG_AL_M
Count
MARG_AL_M
Histogram and boxplot distribution of MARG_AL_F
Count
MARG_AL_F
Histogram and boxplot distribution of MARG_HH_M
Count
MARG_HH_M
Histogram and boxplot distribution of MARG_HH_F
Count
MARG_HH_F

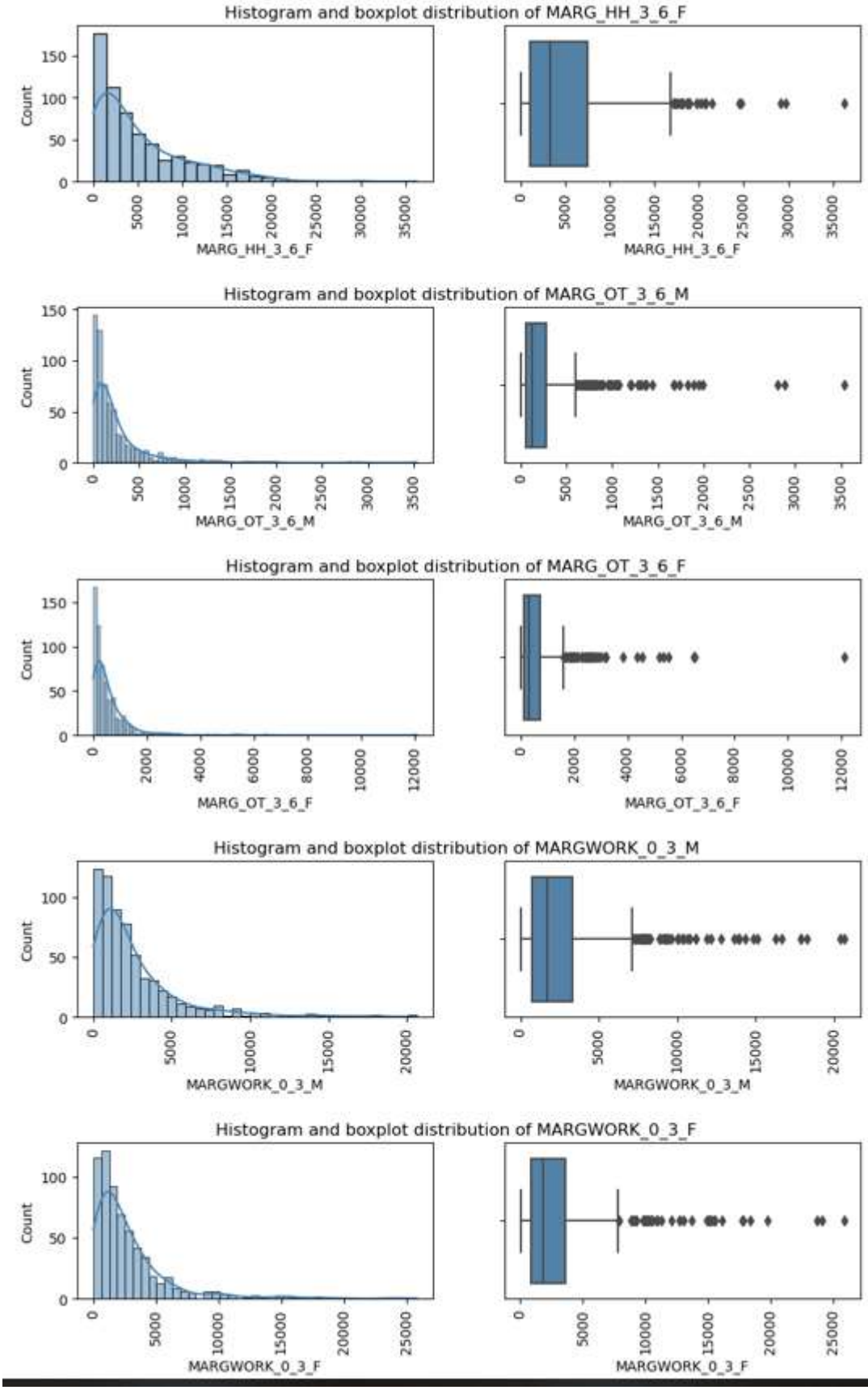

Histogram and boxplot distribution of MARG_OT_M
Count
MARG_OT_M
Histogram and boxplot distribution of MARG_OT_F
Count
MARG_OT_F
Histogram and boxplot distribution of MARGWORK_3_6_M
Count
MARGWORK_3_6_M
Histogram and boxplot distribution of MARGWORK_3_6_F
Count
MARGWORK_3_6_F
Histogram and boxplot distribution of MARG_CL_3_6_M
Count

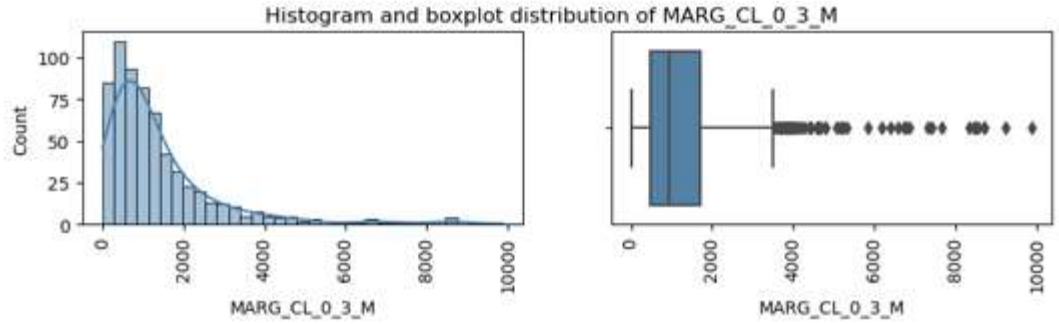

Histogram and boxplot distribution of MARG_CL_3_6_F
Count
MARG_CL_3_6_F
MARG_CL_3_6_F

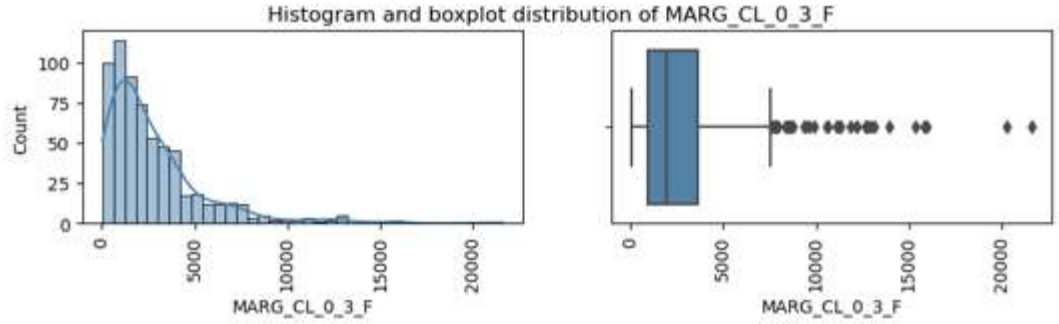

Histogram and boxplot distribution of MARG_AL_3_6_M
Count
MARG_AL_3_6_M
MARG_AL_3_6_M

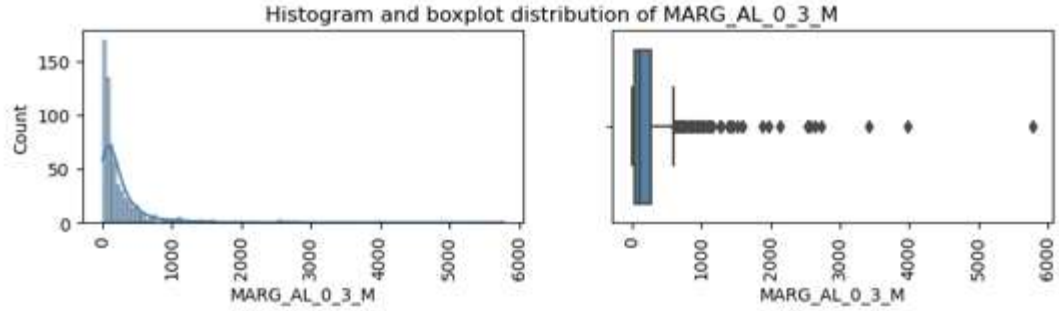

Histogram and boxplot distribution of MARG_AL_3_6_F
Count
MARG_AL_3_6_F
MARG_AL_3_6_F

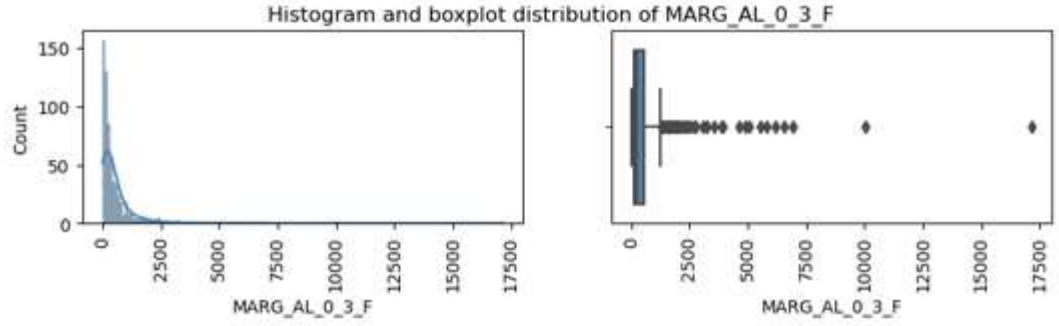

Histogram and boxplot distribution of MARG_HH_3_6_M
Count
MARG_HH_3_6_M
MARG_HH_3_6_M

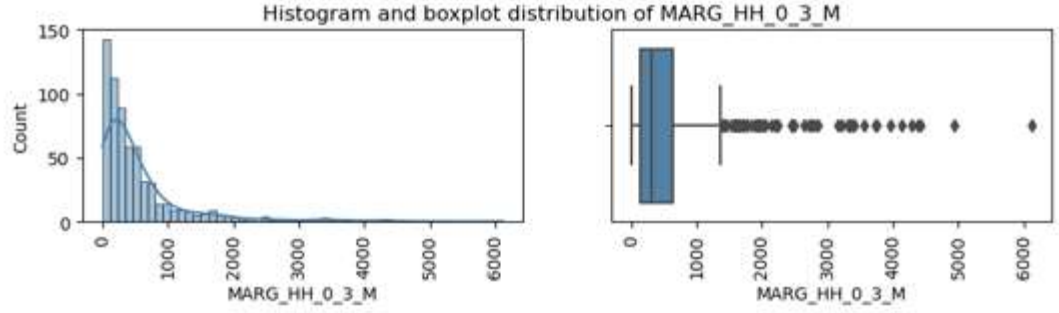

Histogram and boxplot distribution of MARG_HH_3_6_F
Count
MARG_HH_3_6_F
Histogram and boxplot distribution of MARG_OT_3_6_M
Count
MARG_OT_3_6_M
Histogram and boxplot distribution of MARG_OT_3_6_F
Count
MARG_OT_3_6_F
Histogram and boxplot distribution of MARGWORK_0_3_M
Count
MARGWORK_0_3_M
Histogram and boxplot distribution of MARGWORK_0_3_F
Count
MARGWORK_0_3_F


Histogram and boxplot distribution of MARG_CL_0_3_M
Count
MARG_CL_0_3_M


Histogram and boxplot distribution of MARG_CL_0_3_F
Count
MARG_CL_0_3_F


Histogram and boxplot distribution of MARG_AL_0_3_M
Count
MARG_AL_0_3_M


Histogram and boxplot distribution of MARG_AL_0_3_F
Count
MARG_AL_0_3_F


Histogram and boxplot distribution of MARG_HH_0_3_M
Count
MARG_HH_0_3_M

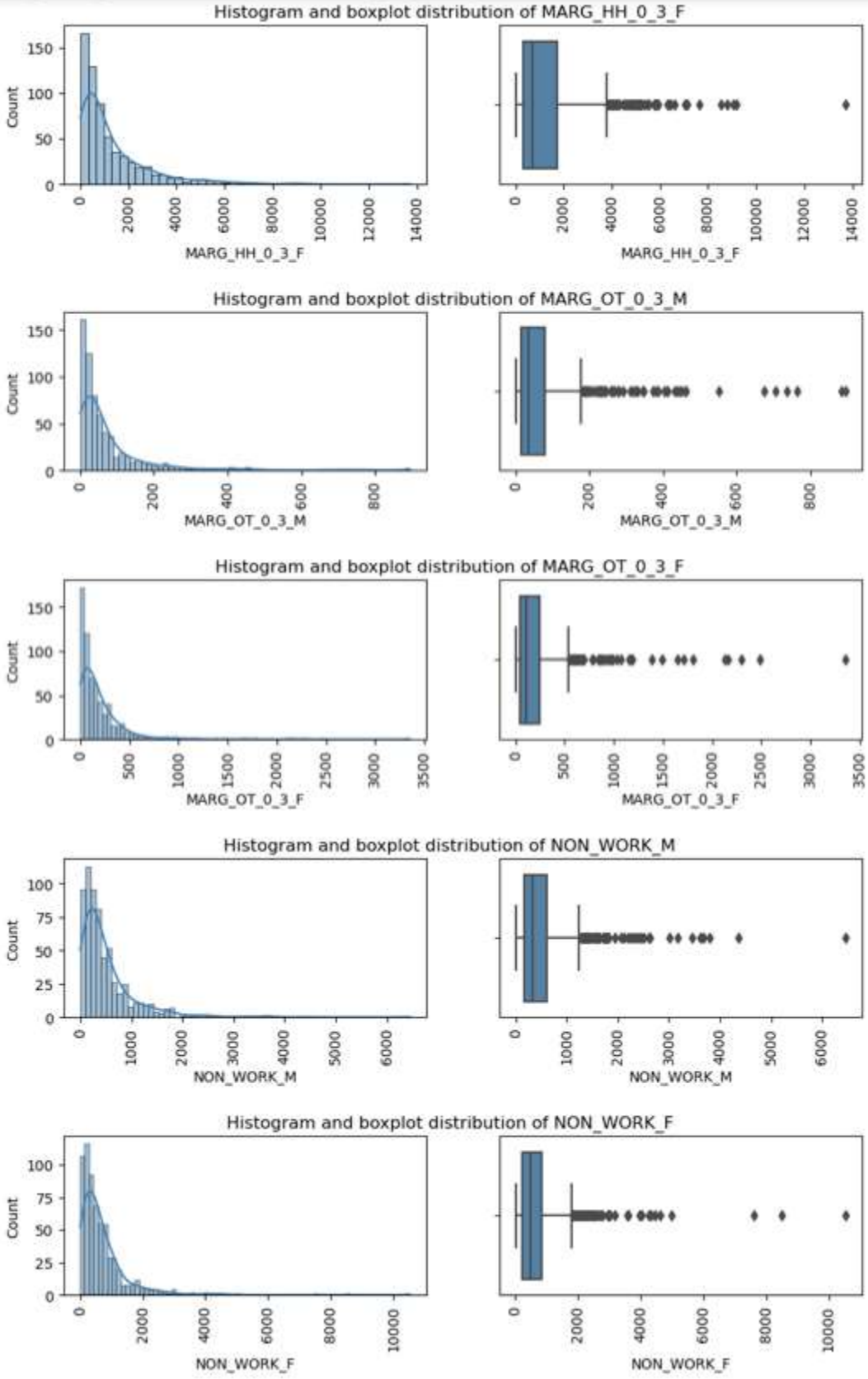


Fig.2.3 Univariate Analysis- Numerical fields

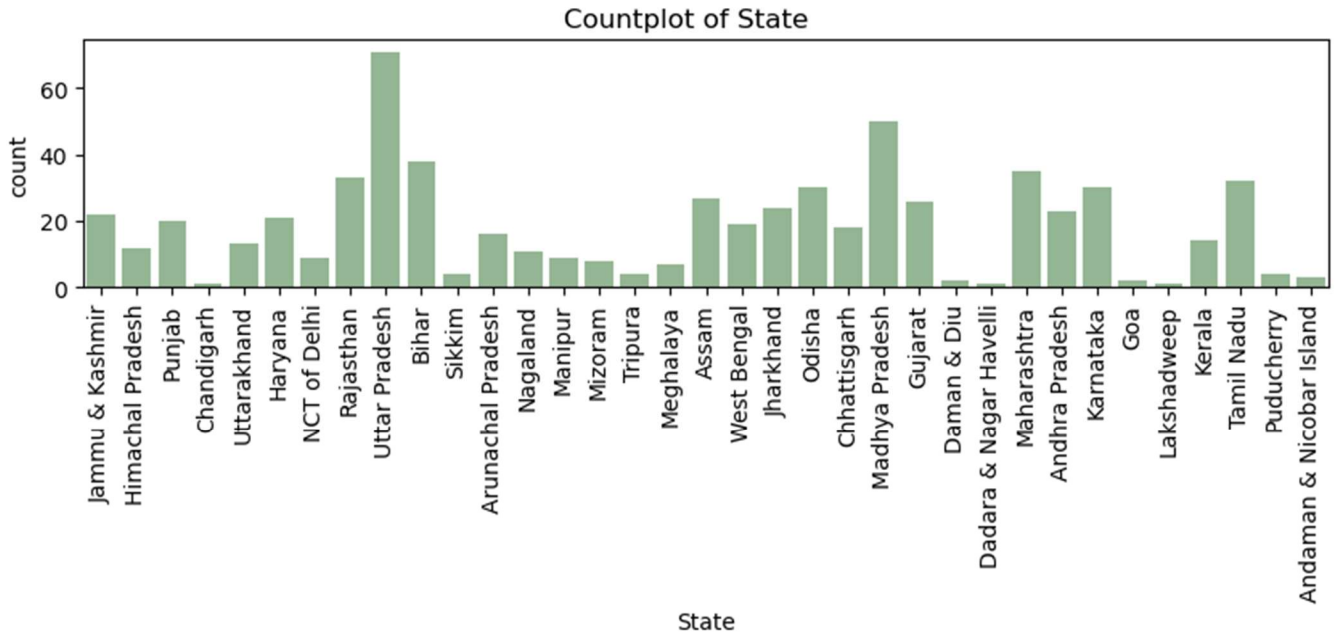


Fig.2.4. Univariate Analysis- Categorical field

Univariate Analysis Observations:

- All the fields have outliers
- Data varies in magnitude across fields
- Uttar Pradesh is the state with highest number of records
- Union Territories have the lowest number of records

Bivariate Analysis:

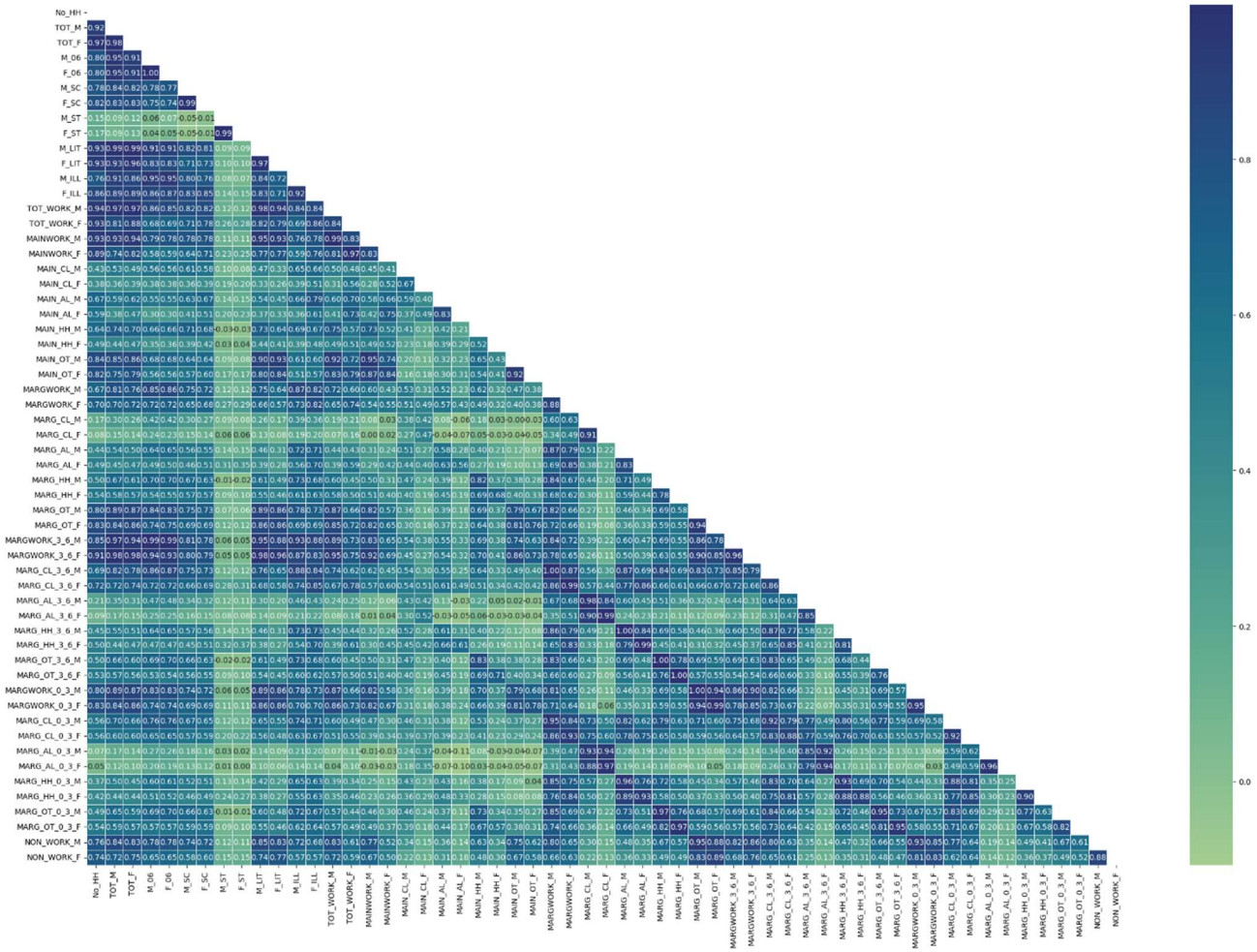


Fig.2.5. Heatmap of numerical fields

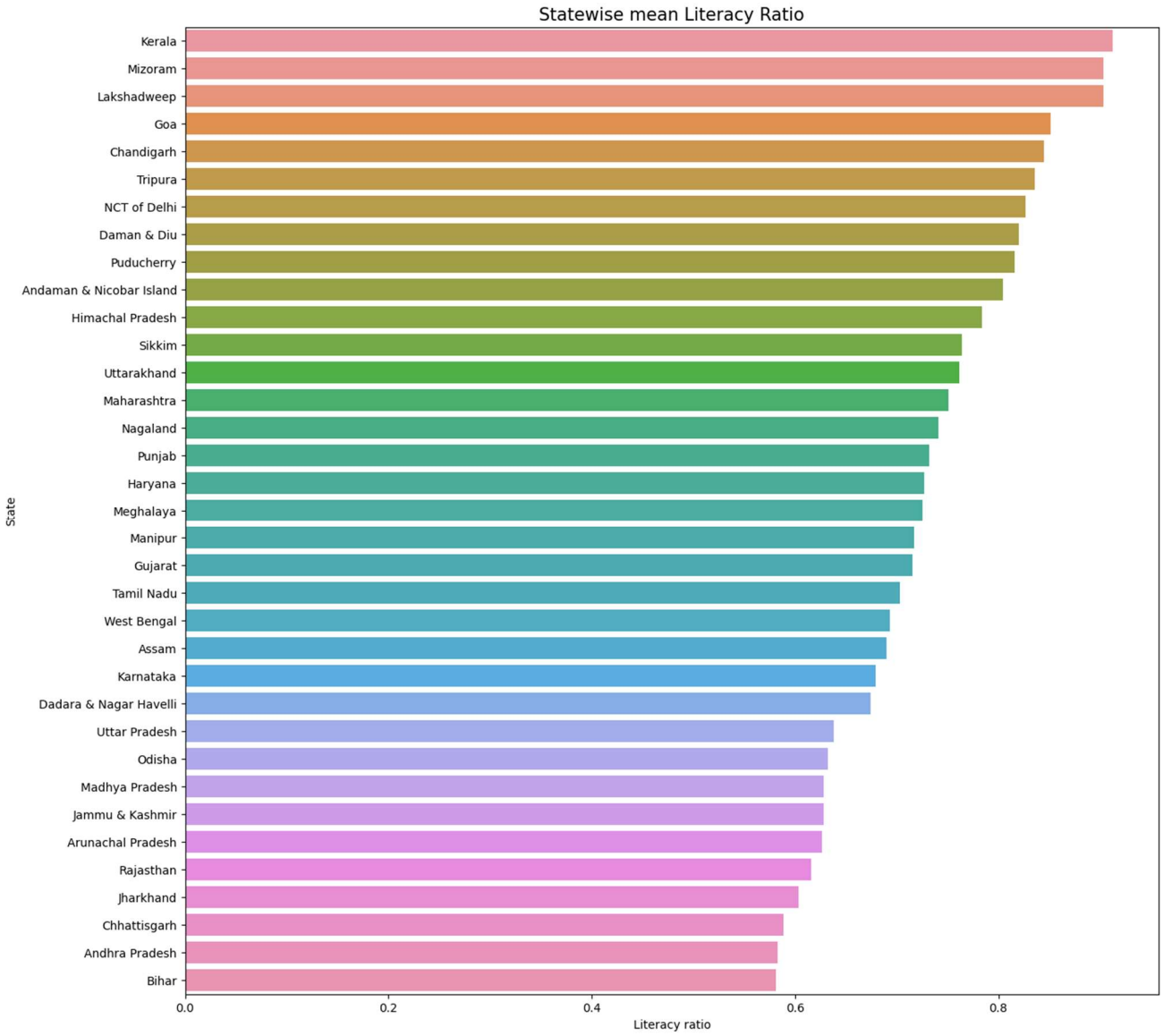


Fig.2.6. Statewise mean literacy ratio

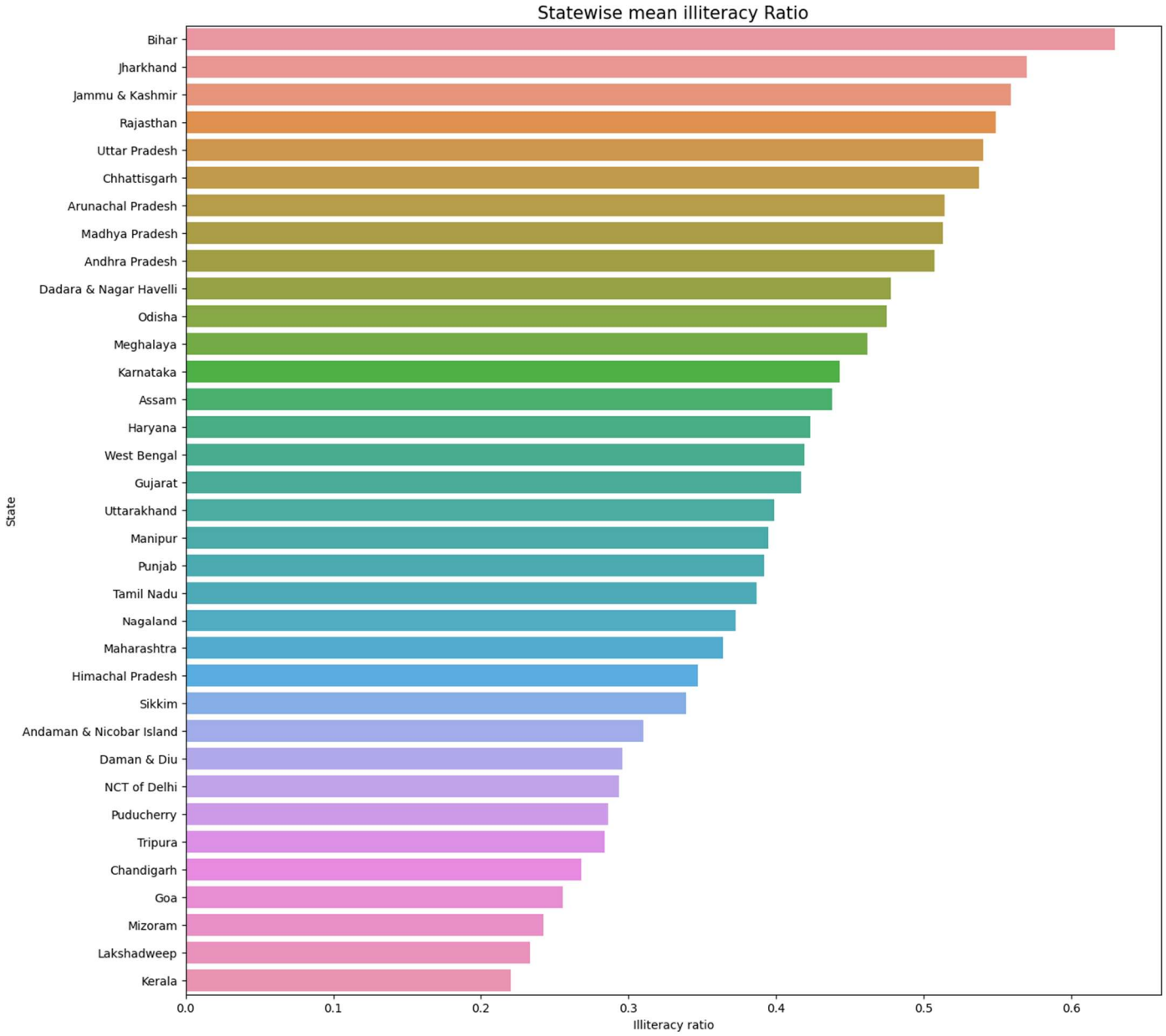


Fig.2.7. Statewise mean Illiteracy ratio

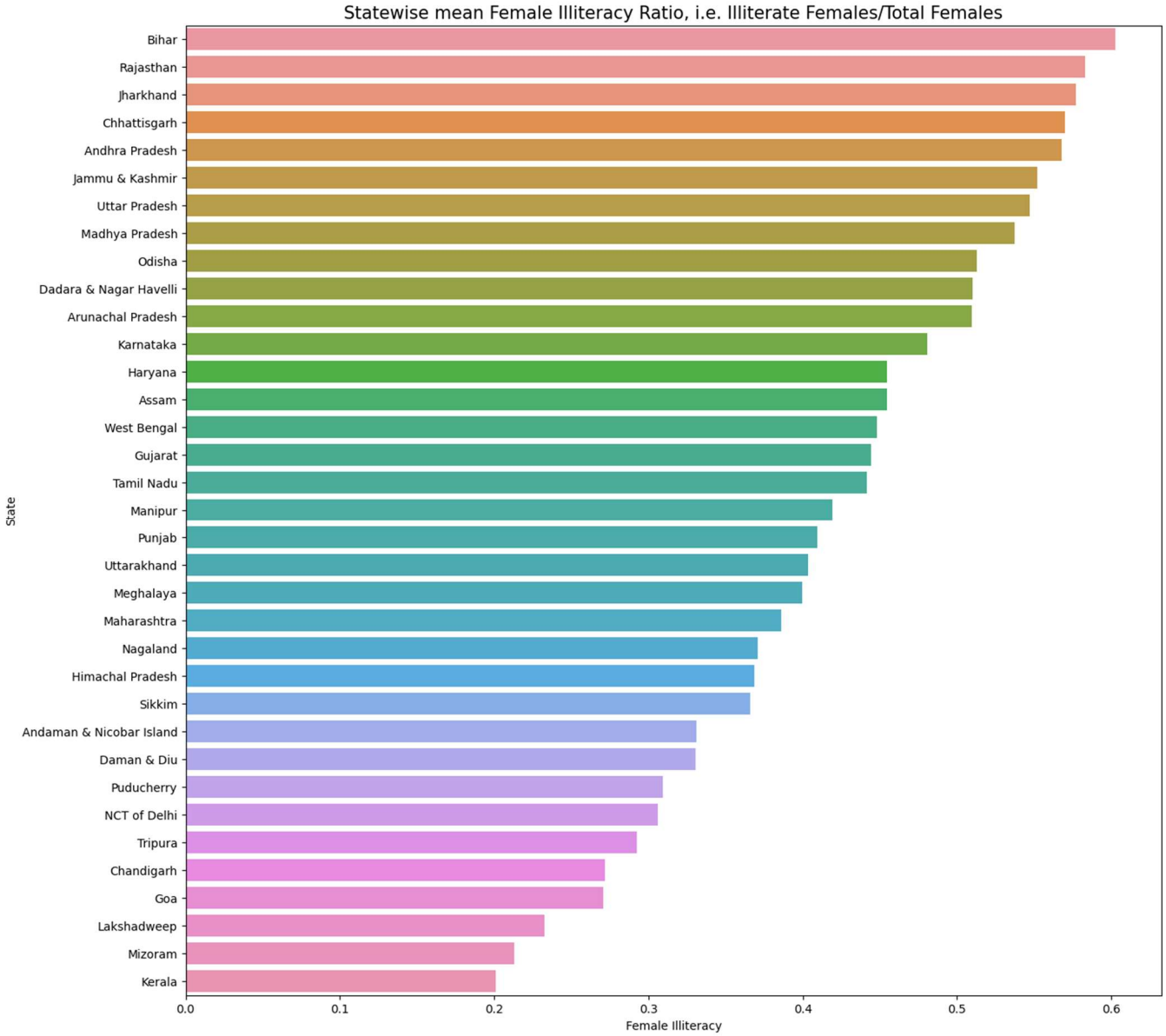


Fig.2.8. Statewise female illiteracy ratio

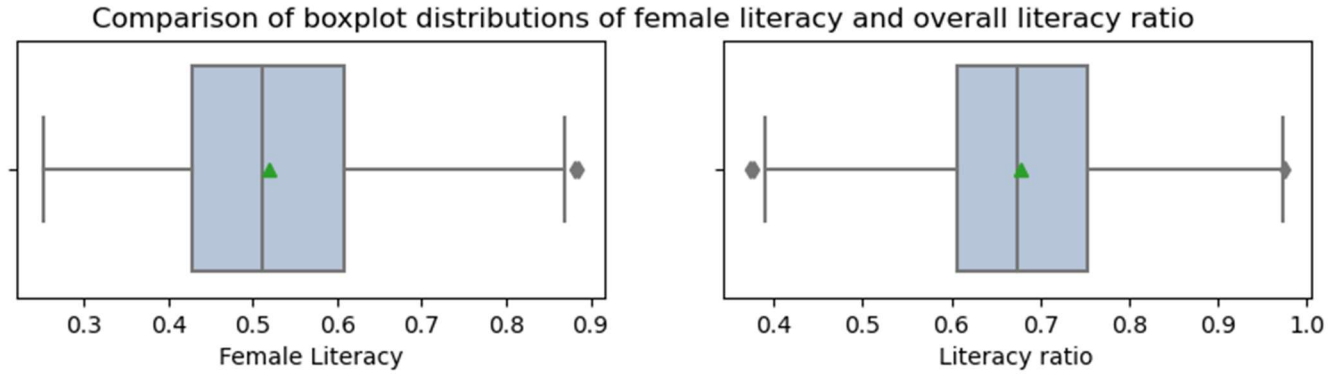


Fig.2.9. Comaprison of Female literacy and total literacy ratio

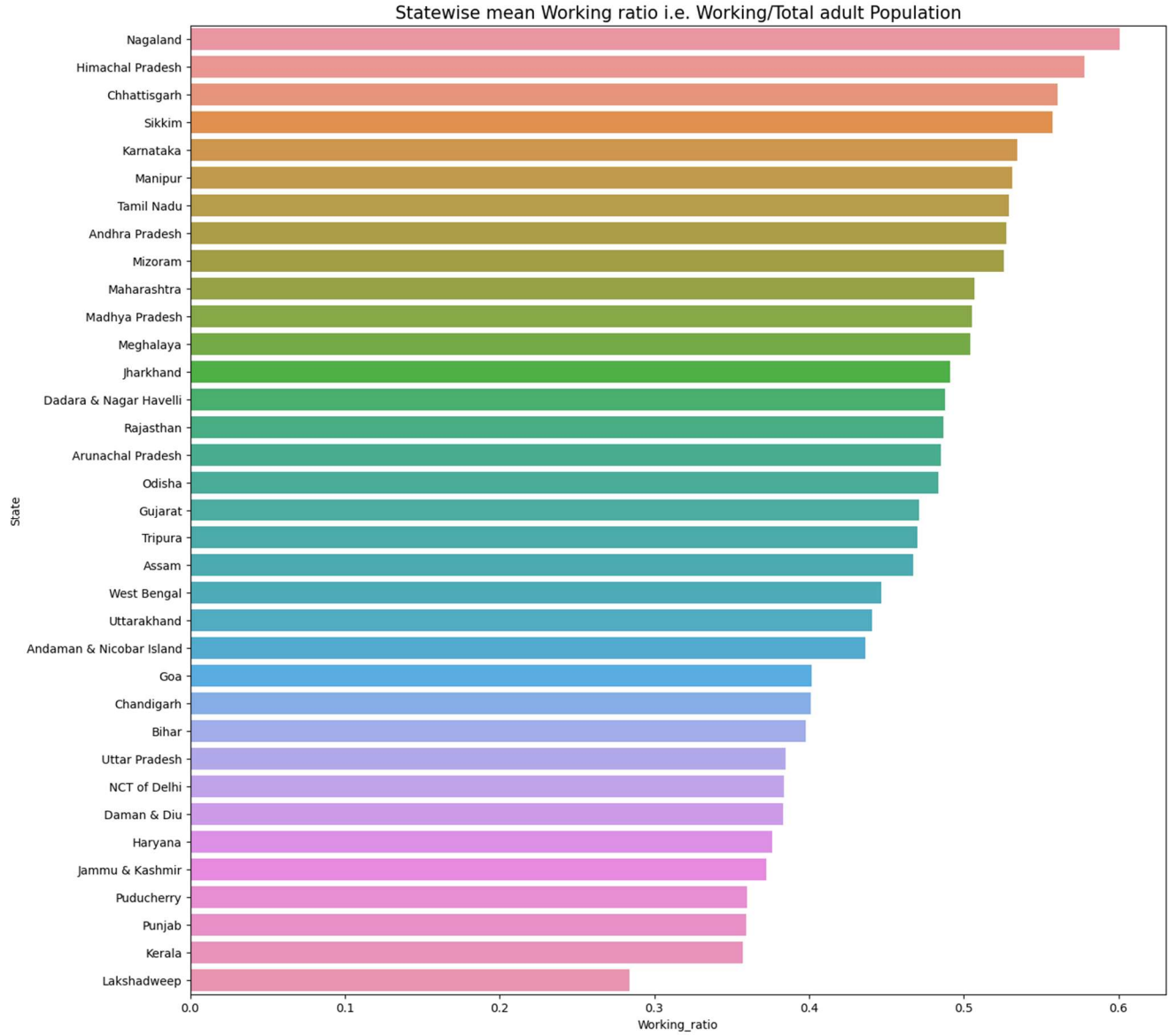


Fig.2.10. Statewise mean working ratio

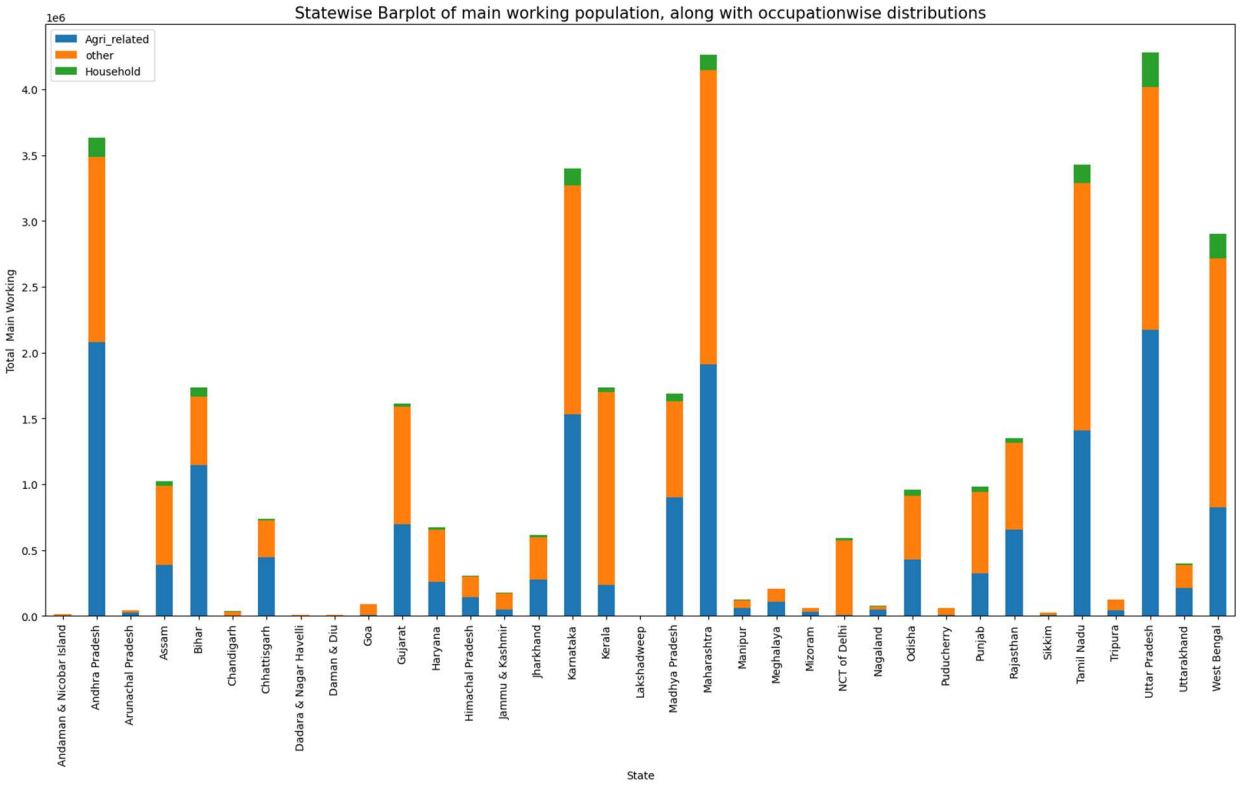
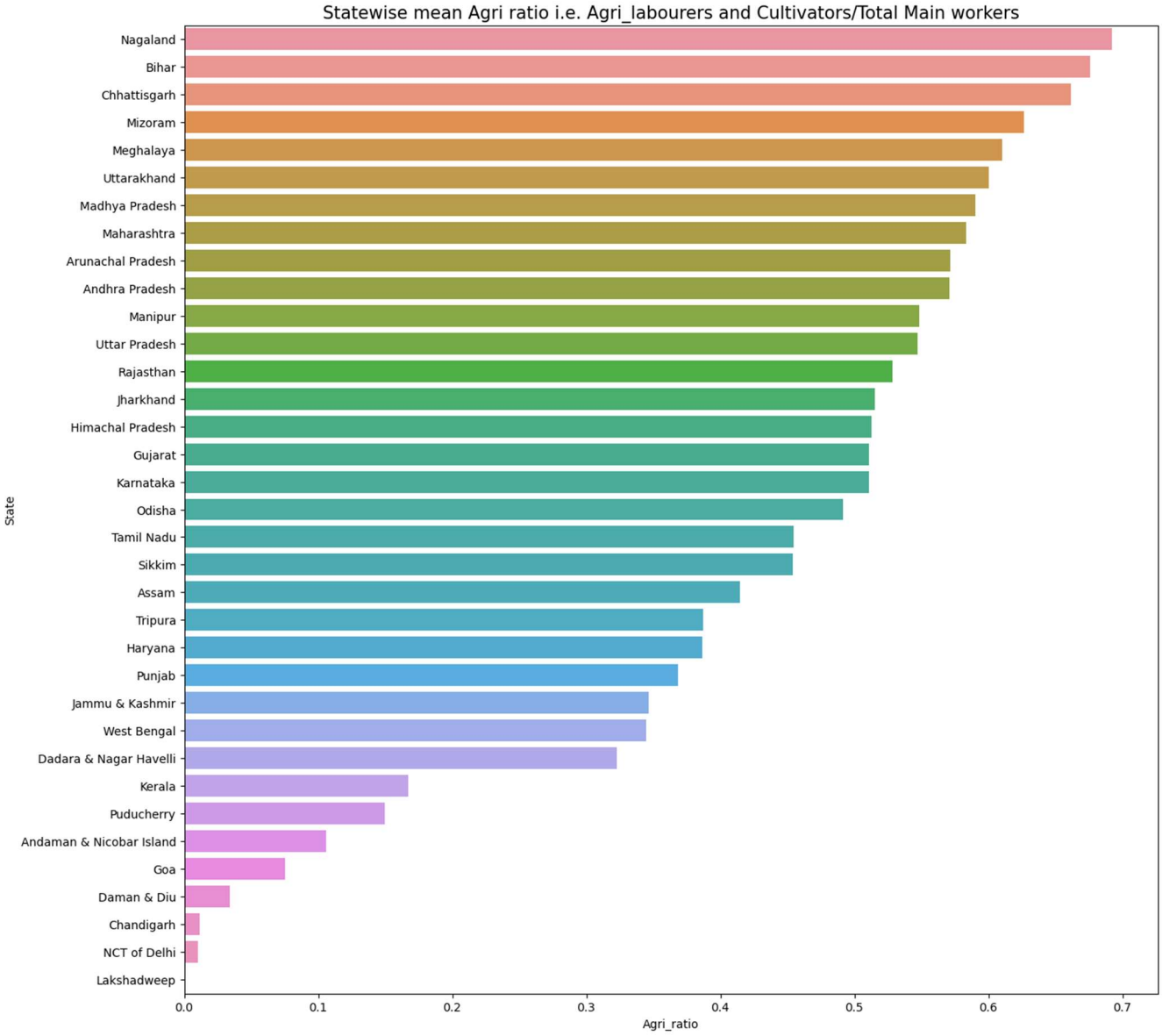


Fig.2.11. Statewise Occupation chart



2.12. Statewise mean Agri_ralated workers

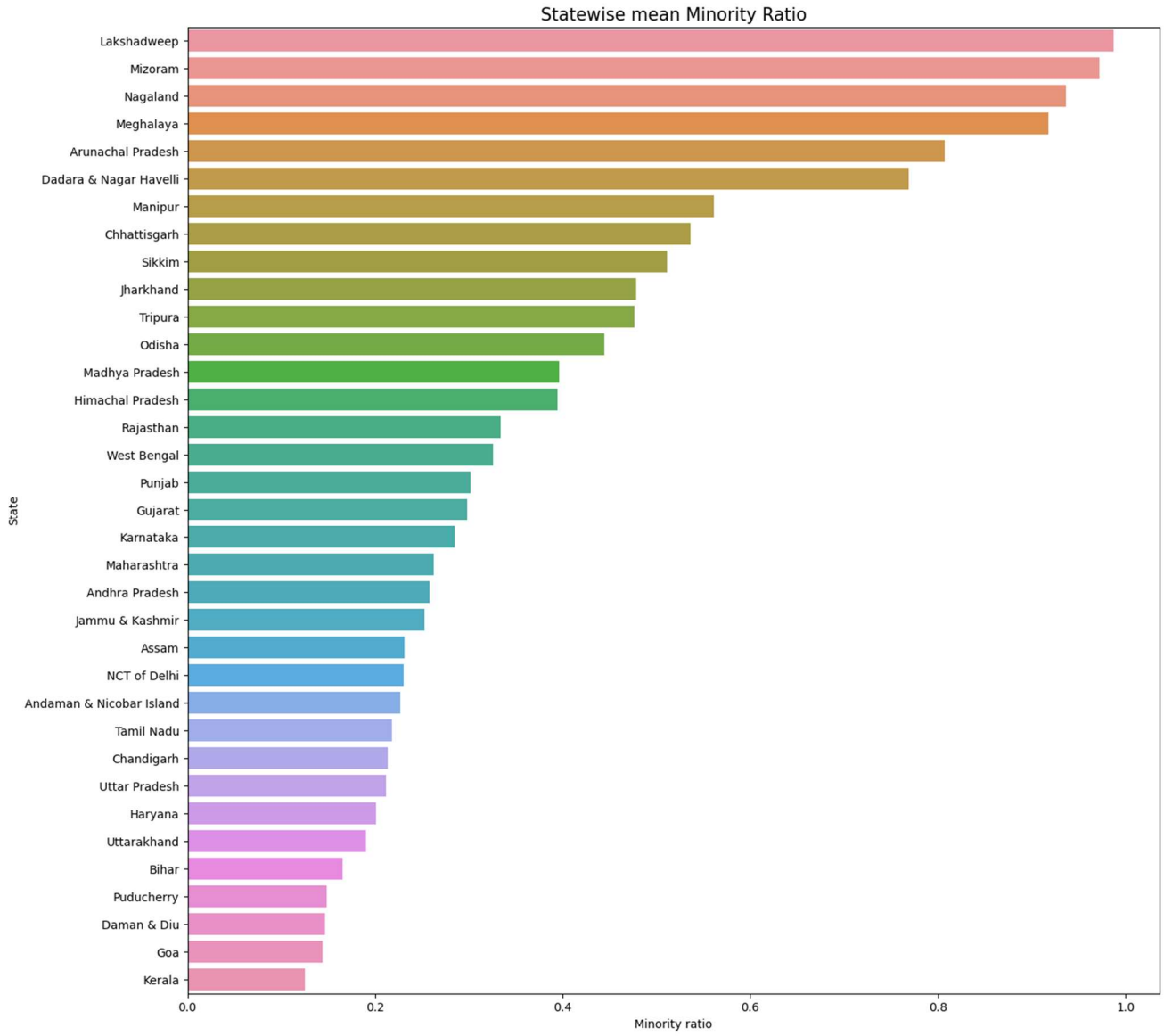


Fig.2.13. Mean minority ratio

Observations

- A lot of variables are highly correlated, as indicated by the dark blue areas of the above heatmap
- Difficult to discern details from heatmap
- For example, margwork_hh_m and marg_ot_3-6m, marg_ot_m, margword_0-3_m etc
- Kerala has the highest and Bihar has the lowest literacy ratio
- Bihar has the highest illiteracy rate among females, close to 60%
- Kerala has the lowest, close to 20%
- While the mean literacy ratio is close to 68%, the mean female literacy ratio is just above 50%
- This indicates that a large number of States still have high percentage of female illiteracy
- The number of districts having lower than average female literacy ratio is 345 , which is 53.91 % of the observed data
- Nagaland has the highest working ratio, close to 60%, and Lakshadweep has the lowest working ratio of 28%

- States like Andhra Pradesh, Bihar, Uttar Pradesh, Uttarkhand, Rajasthan and Madhya Pradesh, are agriculture intensive, with majority of the people working as either Cultivators or agricultural labourers
- Nagaland has the highest mean number of working people, and most of them are in the agriculture related jobs
- The states with lower literacy ratio are agriculture intensive
- The north-eastern states have a high minority ratio

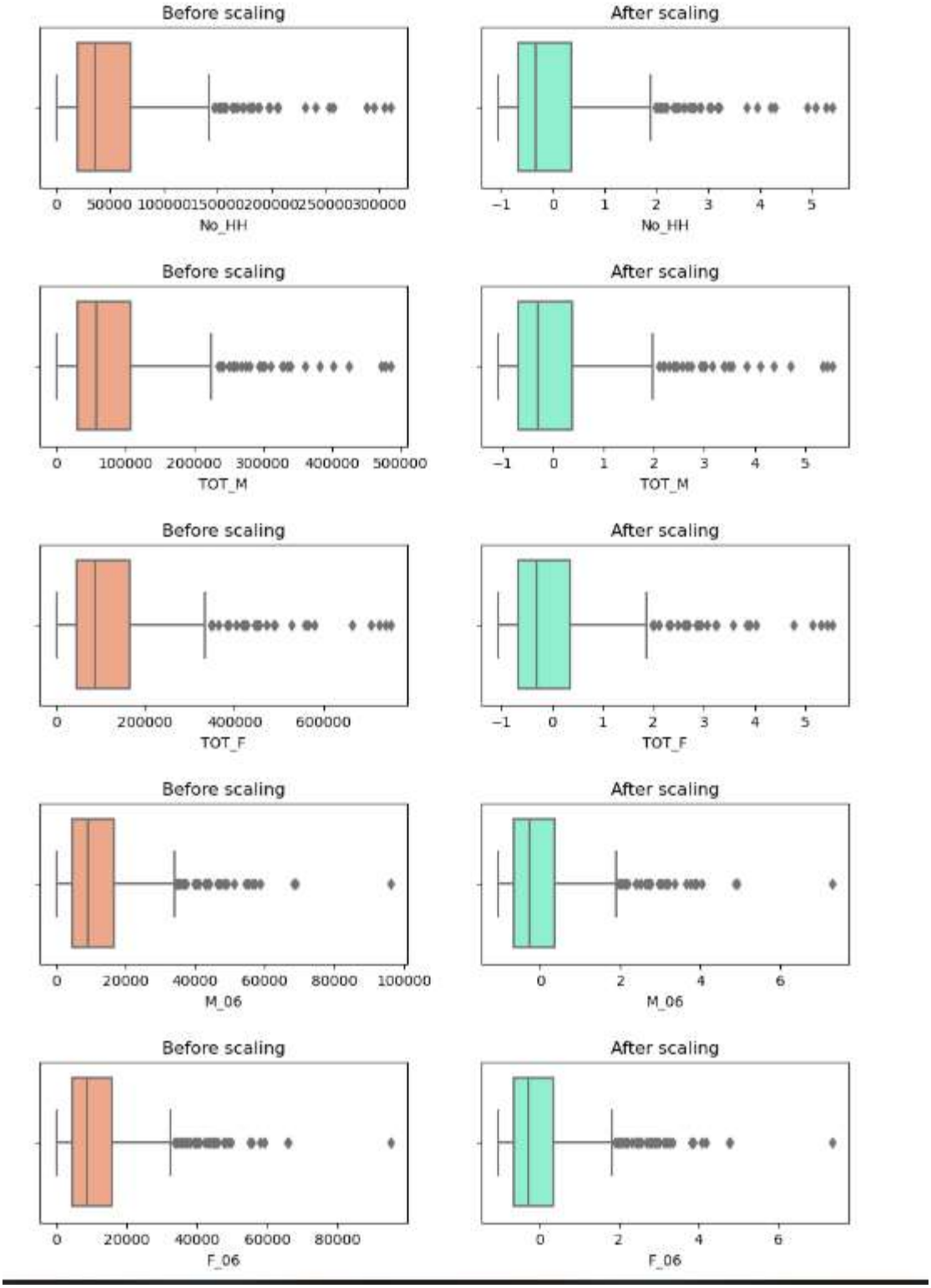
C. **Part 2 - PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?**
Generally,outliers have an impact on the results of the PCA, and hence need to be tackled. However, the outliers themselves may contain valuable data, that may provide useful insights. In a sensitive dataset like this census, eventhough PCA needs to be done, outliers may be left untreated in order to prevent loss of essential information.

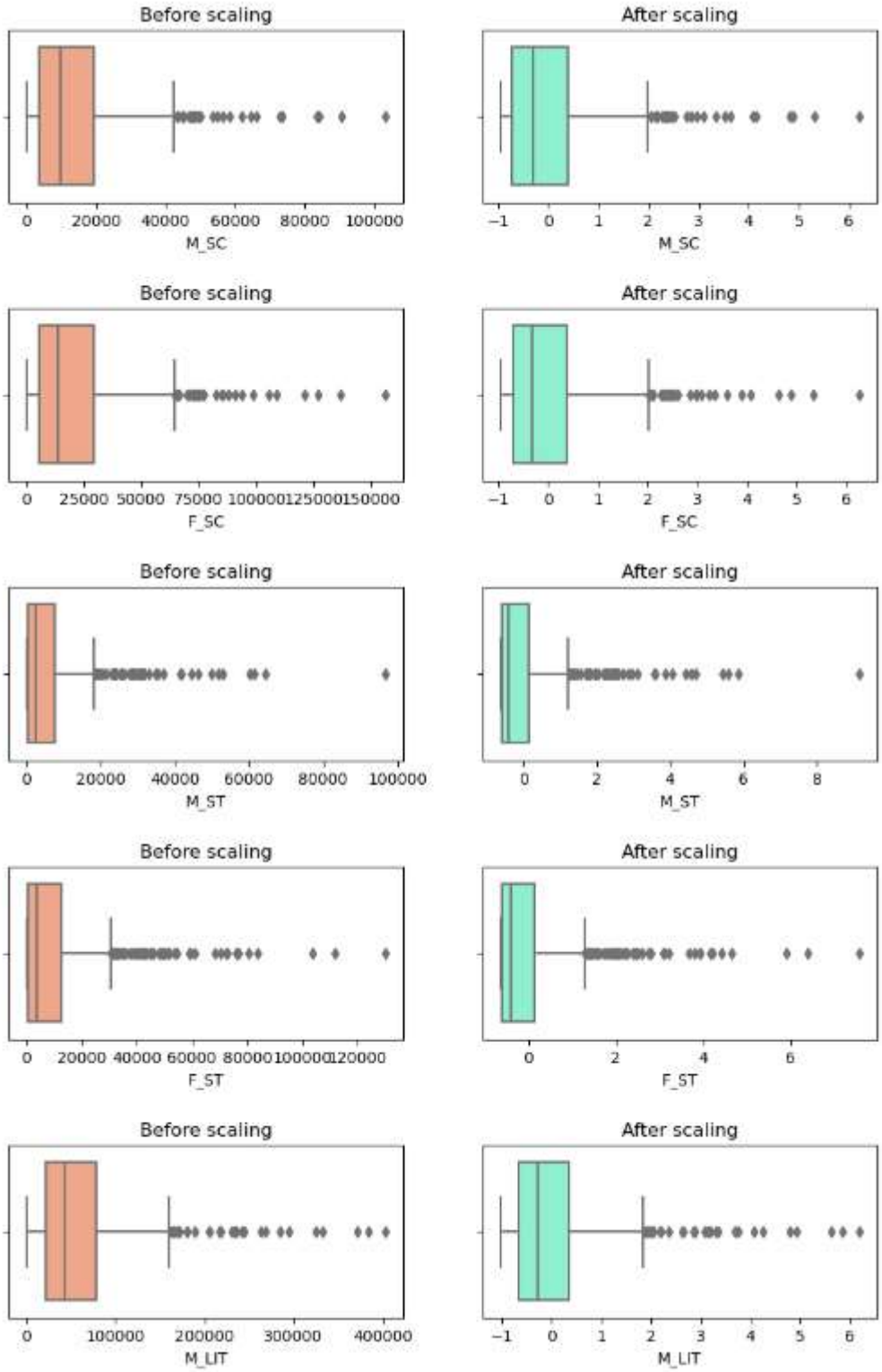
D. **Part 2 - PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.**

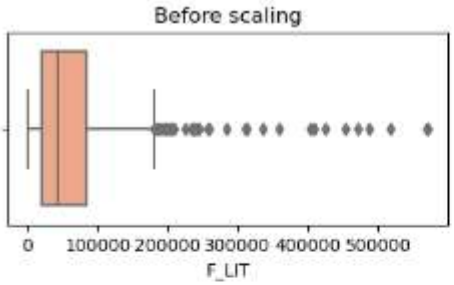
	count	mean	std	min	25%	50%	75%	max
No_HH	640.0	0.0	1.0	-1.06	-0.66	-0.32	0.37	5.39
TOT_M	640.0	-0.0	1.0	-1.08	-0.68	-0.29	0.38	5.53
TOT_F	640.0	-0.0	1.0	-1.07	-0.67	-0.31	0.37	5.53
M_06	640.0	-0.0	1.0	-1.07	-0.66	-0.27	0.37	7.30
F_06	640.0	0.0	1.0	-1.05	-0.64	-0.29	0.35	7.35
M_SC	640.0	-0.0	1.0	-0.96	-0.72	-0.29	0.39	6.21
F_SC	640.0	0.0	1.0	-0.96	-0.70	-0.33	0.39	6.25
M_ST	640.0	-0.0	1.0	-0.63	-0.60	-0.39	0.15	9.15
F_ST	640.0	-0.0	1.0	-0.64	-0.61	-0.40	0.15	7.56
M_LIT	640.0	0.0	1.0	-1.03	-0.66	-0.27	0.36	6.18
F_LIT	640.0	-0.0	1.0	-0.88	-0.61	-0.30	0.25	6.73
M_ILL	640.0	0.0	1.0	-1.10	-0.68	-0.31	0.38	4.24
F_ILL	640.0	-0.0	1.0	-1.18	-0.71	-0.29	0.48	4.21
TOT_WORK_M	640.0	-0.0	1.0	-1.04	-0.67	-0.28	0.34	6.36
TOT_WORK_F	640.0	-0.0	1.0	-1.10	-0.68	-0.29	0.32	5.83
MAINWORK_M	640.0	-0.0	1.0	-0.96	-0.65	-0.28	0.32	6.92
MAINWORK_F	640.0	0.0	1.0	-0.93	-0.62	-0.32	0.23	6.60
MAIN_CL_M	640.0	-0.0	1.0	-1.15	-0.72	-0.27	0.48	5.00
MAIN_CL_F	640.0	-0.0	1.0	-1.03	-0.67	-0.30	0.34	5.77
MAIN_AL_M	640.0	0.0	1.0	-0.91	-0.75	-0.30	0.35	5.47
MAIN_AL_F	640.0	0.0	1.0	-0.69	-0.58	-0.39	0.13	6.15
MAIN_HH_M	640.0	0.0	1.0	-0.69	-0.55	-0.30	0.17	12.17
MAIN_HH_F	640.0	-0.0	1.0	-0.43	-0.36	-0.26	0.02	14.04
MAIN_OT_M	640.0	-0.0	1.0	-0.69	-0.54	-0.32	0.12	8.55
MAIN_OT_F	640.0	0.0	1.0	-0.65	-0.49	-0.32	0.10	10.39
MARGWORK_M	640.0	0.0	1.0	-1.05	-0.66	-0.29	0.27	5.37
MARGWORK_F	640.0	-0.0	1.0	-1.18	-0.70	-0.27	0.53	4.90
MARG_CL_M	640.0	-0.0	1.0	-0.79	-0.56	-0.33	0.18	9.28
MARG_CL_F	640.0	-0.0	1.0	-0.65	-0.47	-0.30	0.10	11.80
MARG_AL_M	640.0	0.0	1.0	-0.87	-0.64	-0.33	0.26	5.40
MARG_AL_F	640.0	0.0	1.0	-0.95	-0.75	-0.36	0.39	5.74
MARG_HH_M	640.0	-0.0	1.0	-0.69	-0.53	-0.33	0.09	8.61

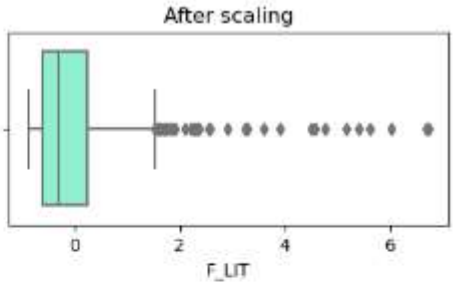
MARG_HH_F	640.0	0.0	1.0	-0.66	-0.51	-0.30	0.15	12.24
MARG_OT_M	640.0	0.0	1.0	-0.86	-0.61	-0.30	0.24	5.99
MARG_OT_F	640.0	-0.0	1.0	-0.86	-0.60	-0.29	0.21	7.99
MARGWORK_3_6_M	640.0	0.0	1.0	-1.07	-0.66	-0.30	0.39	6.64
MARGWORK_3_6_F	640.0	-0.0	1.0	-0.97	-0.66	-0.29	0.32	7.18
MARG_CL_3_6_M	640.0	-0.0	1.0	-1.06	-0.67	-0.29	0.29	5.44
MARG_CL_3_6_F	640.0	-0.0	1.0	-1.21	-0.71	-0.24	0.56	4.70
MARG_AL_3_6_M	640.0	0.0	1.0	-0.87	-0.61	-0.34	0.22	7.33
MARG_AL_3_6_F	640.0	0.0	1.0	-0.70	-0.50	-0.31	0.12	10.19
MARG_HH_3_6_M	640.0	-0.0	1.0	-0.90	-0.66	-0.34	0.31	5.43
MARG_HH_3_6_F	640.0	-0.0	1.0	-0.97	-0.76	-0.35	0.44	5.83
MARG_OT_3_6_M	640.0	0.0	1.0	-0.68	-0.52	-0.32	0.09	9.18
MARG_OT_3_6_F	640.0	0.0	1.0	-0.65	-0.51	-0.30	0.15	12.80
MARGWORK_0_3_M	640.0	0.0	1.0	-0.86	-0.61	-0.31	0.23	5.94
MARGWORK_0_3_F	640.0	0.0	1.0	-0.85	-0.60	-0.30	0.23	6.92
MARG_CL_0_3_M	640.0	-0.0	1.0	-0.93	-0.61	-0.30	0.22	5.70
MARG_CL_0_3_F	640.0	-0.0	1.0	-0.98	-0.65	-0.30	0.30	6.77
MARG_AL_0_3_M	640.0	0.0	1.0	-0.55	-0.45	-0.30	0.04	12.19
MARG_AL_0_3_F	640.0	-0.0	1.0	-0.50	-0.40	-0.28	0.01	14.86
MARG_HH_0_3_M	640.0	0.0	1.0	-0.74	-0.56	-0.33	0.11	7.29
MARG_HH_0_3_F	640.0	-0.0	1.0	-0.82	-0.63	-0.36	0.26	7.84
MARG_OT_0_3_M	640.0	-0.0	1.0	-0.66	-0.53	-0.34	0.07	7.64
MARG_OT_0_3_F	640.0	-0.0	1.0	-0.65	-0.51	-0.28	0.13	10.19
NON_WORK_M	640.0	-0.0	1.0	-0.84	-0.57	-0.30	0.15	9.75
NON_WORK_F	640.0	-0.0	1.0	-0.77	-0.53	-0.26	0.16	10.81

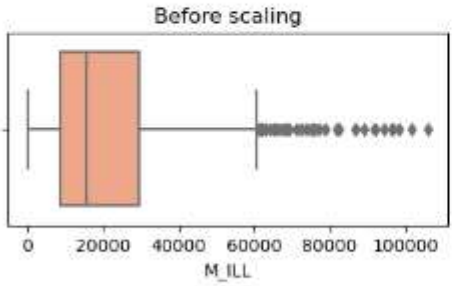
Fig.2.14. Data description after scaling

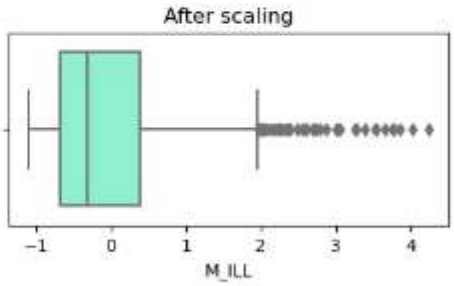

Before scaling
0 50000 100000150000200000250000300000
No_HH
After scaling
−1 0 1 2 3 4 5
No_HH
Before scaling
0 100000 200000 300000 400000 500000
TOT_M
After scaling
−1 0 1 2 3 4 5
TOT_M
Before scaling
0 200000 400000 600000
TOT_F
After scaling
−1 0 1 2 3 4 5
TOT_F
Before scaling
0 20000 40000 60000 80000 100000
M_06
After scaling
0 2 4 6
M_06
Before scaling
0 20000 40000 60000 80000
F_06
After scaling
0 2 4 6
F_06

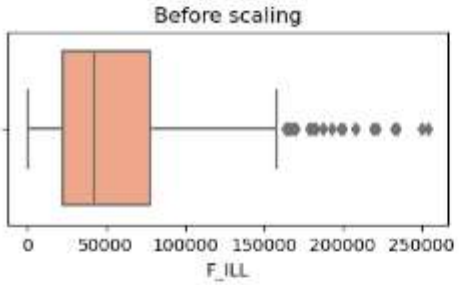

Before scaling
0
20000
40000
60000
80000
100000
M_SC
After scaling
−1
0
1
2
3
4
5
6
M_SC
Before scaling
0
25000
50000
75000
100000
125000
150000
F_SC
After scaling
−1
0
1
2
3
4
5
6
F_SC
Before scaling
0
20000
40000
60000
80000
100000
M_ST
After scaling
0
2
4
6
8
M_ST
Before scaling
0
20000
40000
60000
80000
100000
120000
F_ST
After scaling
0
2
4
6
F_ST
Before scaling
0
100000
200000
300000
400000
M_LIT
After scaling
−1
0
1
2
3
4
5
6
M_LIT

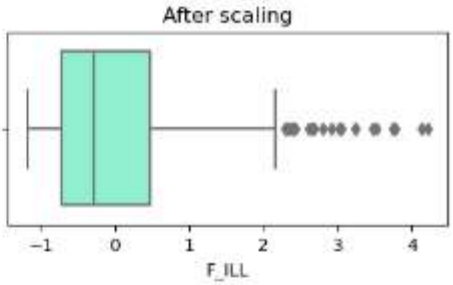

Before scaling
0
100000
200000
300000
400000
500000
F_LIT

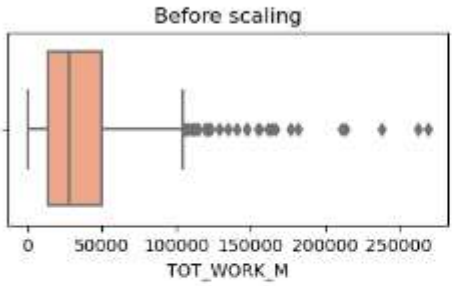

After scaling
0
2
4
6
F_LIT

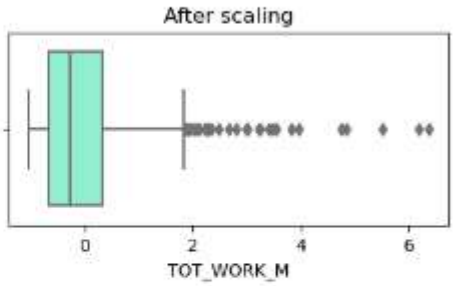

Before scaling
0
20000
40000
60000
80000
100000
M_ILL

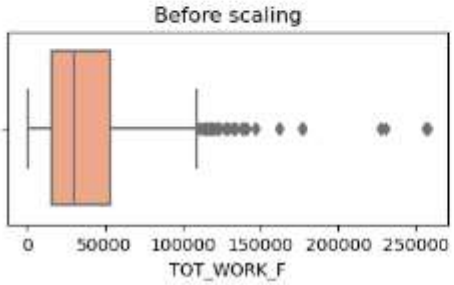

After scaling
−1
0
1
2
3
4
M_ILL

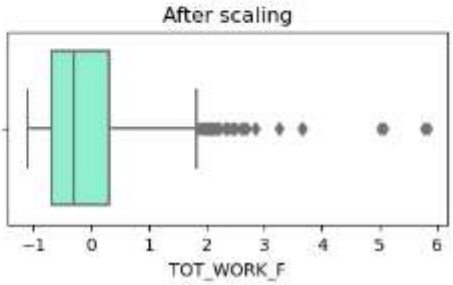

Before scaling
0
50000
100000
150000
200000
250000
F_ILL

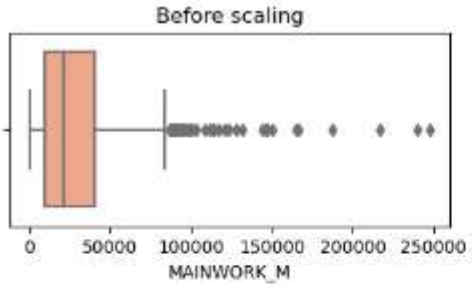

After scaling
−1
0
1
2
3
4
F_ILL

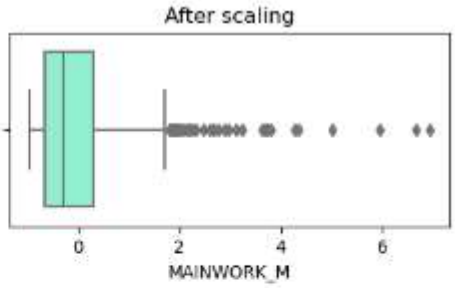

Before scaling
0
50000
100000
150000
200000
250000
TOT_WORK_M

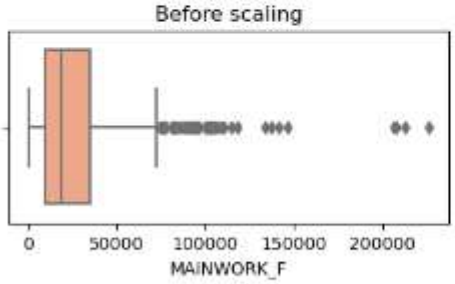

After scaling
0
2
4
6
TOT_WORK_M

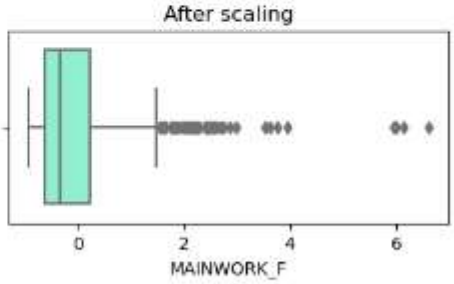

Before scaling
0
50000
100000
150000
200000
250000
TOT_WORK_F

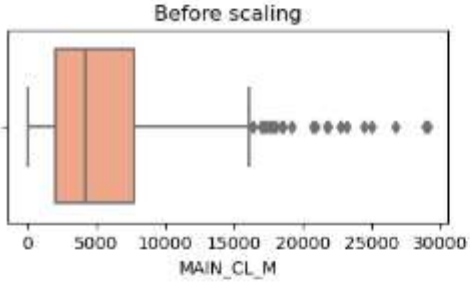

After scaling
−1
0
1
2
3
4
5
6
TOT_WORK_F

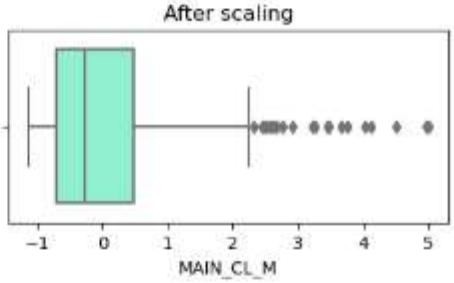

Before scaling
0
50000
100000
150000
200000
250000
MAINWORK_M

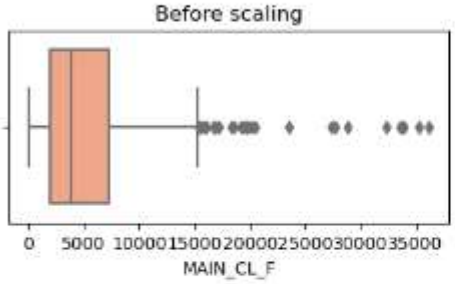

After scaling
0
2
4
6
MAINWORK_M

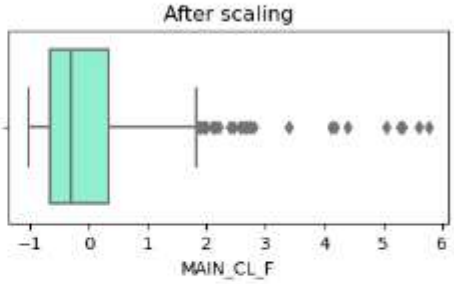

Before scaling
0
50000
100000
150000
200000
MAINWORK_F

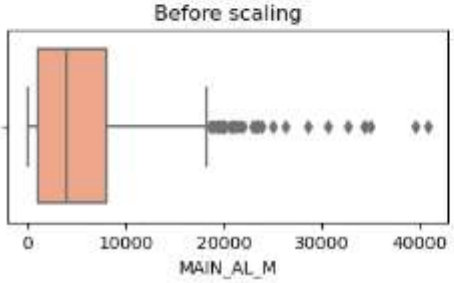

After scaling
0
2
4
6
MAINWORK_F

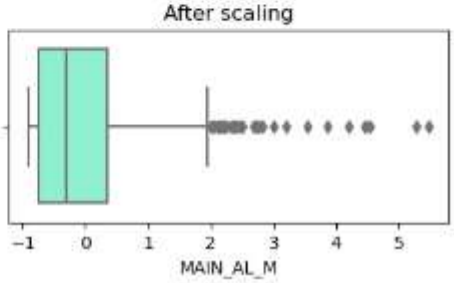

Before scaling
0
5000
10000
15000
20000
25000
30000
MAIN_CL_M

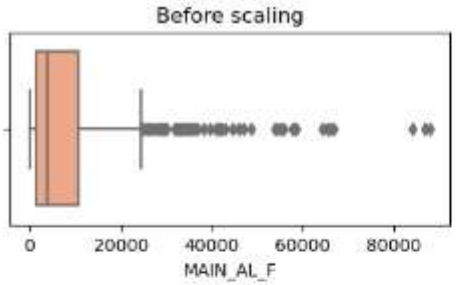

After scaling
−1
0
1
2
3
4
5
MAIN_CL_M

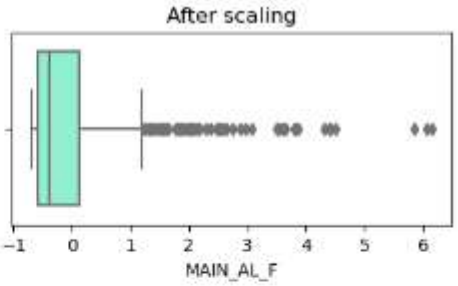

Before scaling
0
5000
10000
15000
20000
25000
30000
35000
MAIN_CL_F

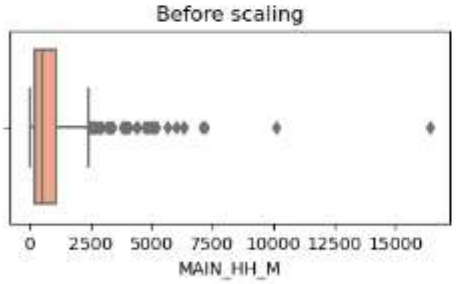

After scaling
−1
0
1
2
3
4
5
6
MAIN_CL_F

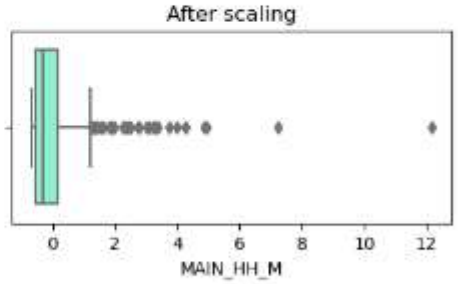

Before scaling
0
10000
20000
30000
40000
MAIN_AL_M

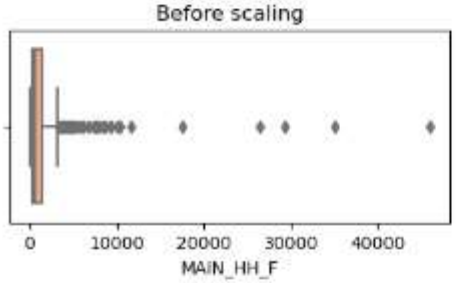

After scaling
−1
0
1
2
3
4
5
MAIN_AL_M

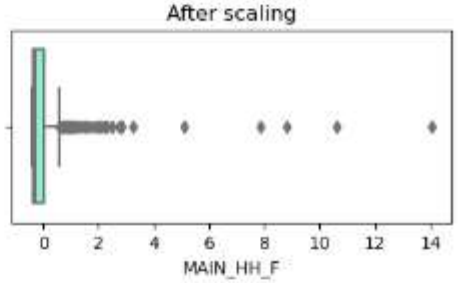

Before scaling
0
20000
40000
60000
80000
MAIN_AL_F

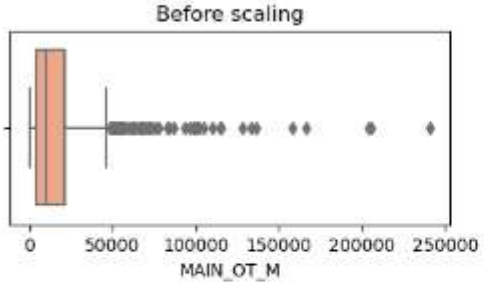

After scaling
−1
0
1
2
3
4
5
6
MAIN_AL_F

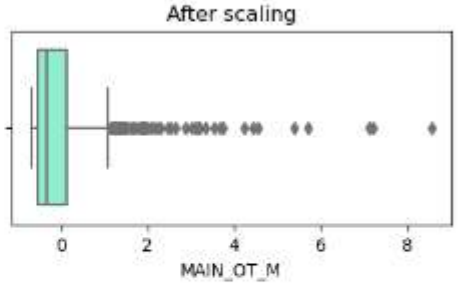

Before scaling
0
2500
5000
7500
10000
12500
15000
MAIN_HH_M

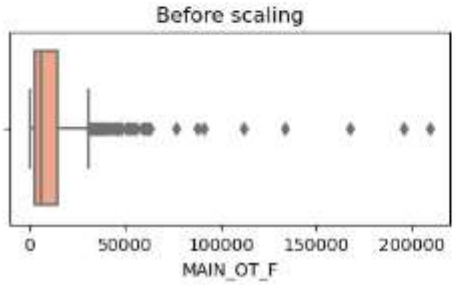

After scaling
0
2
4
6
8
10
12
MAIN_HH_M

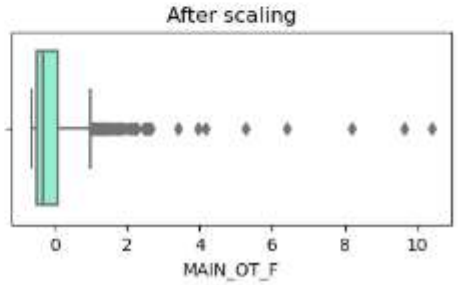

Before scaling
0
10000
20000
30000
40000
MAIN_HH_F

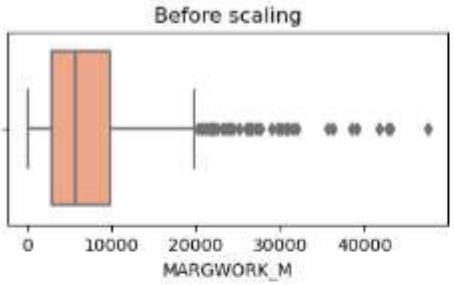

After scaling
0
2
4
6
8
10
12
14
MAIN_HH_F

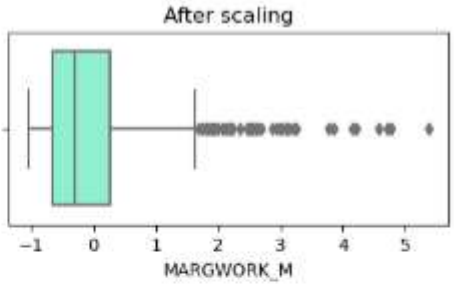

Before scaling
0
50000
100000
150000
200000
250000
MAIN_OT_M

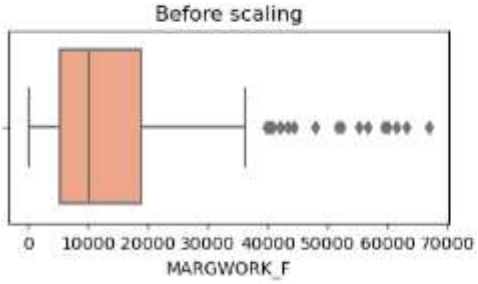

After scaling
0
2
4
6
8
MAIN_OT_M

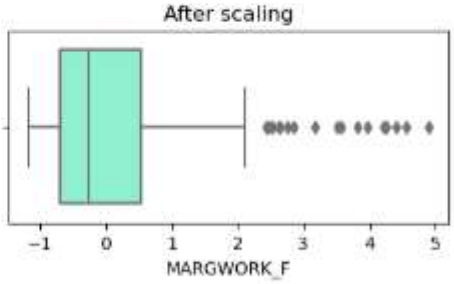

Before scaling
0
50000
100000
150000
200000
MAIN_OT_F

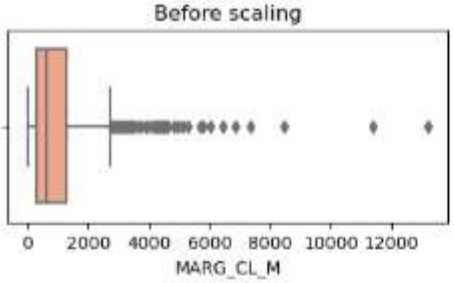

After scaling
0
2
4
6
8
10
MAIN_OT_F

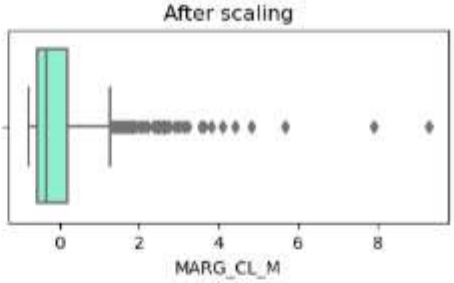

Before scaling
0
10000
20000
30000
40000
MARGWORK_M

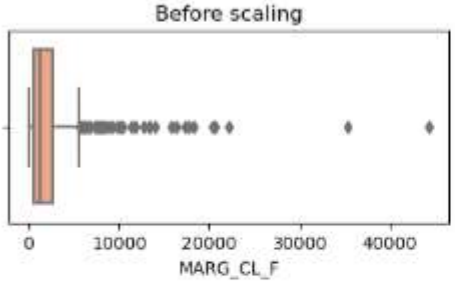

After scaling
−1
0
1
2
3
4
5
MARGWORK_M

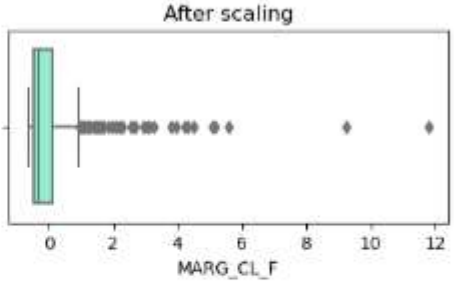

Before scaling
0
10000
20000
30000
40000
50000
60000
70000
MARGWORK_F

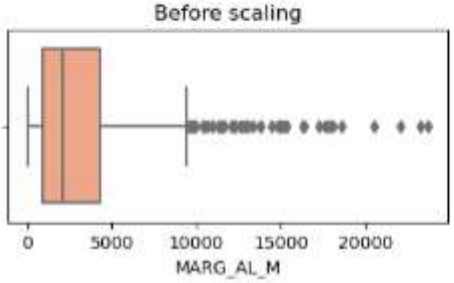

After scaling
−1
0
1
2
3
4
5
MARGWORK_F

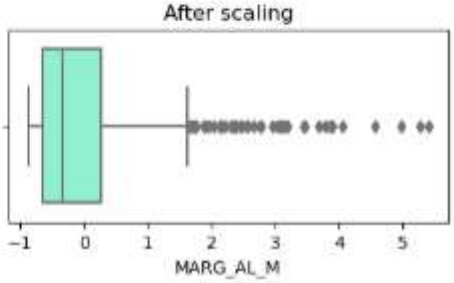

Before scaling
0
2000
4000
6000
8000
10000
12000
MARG_CL_M

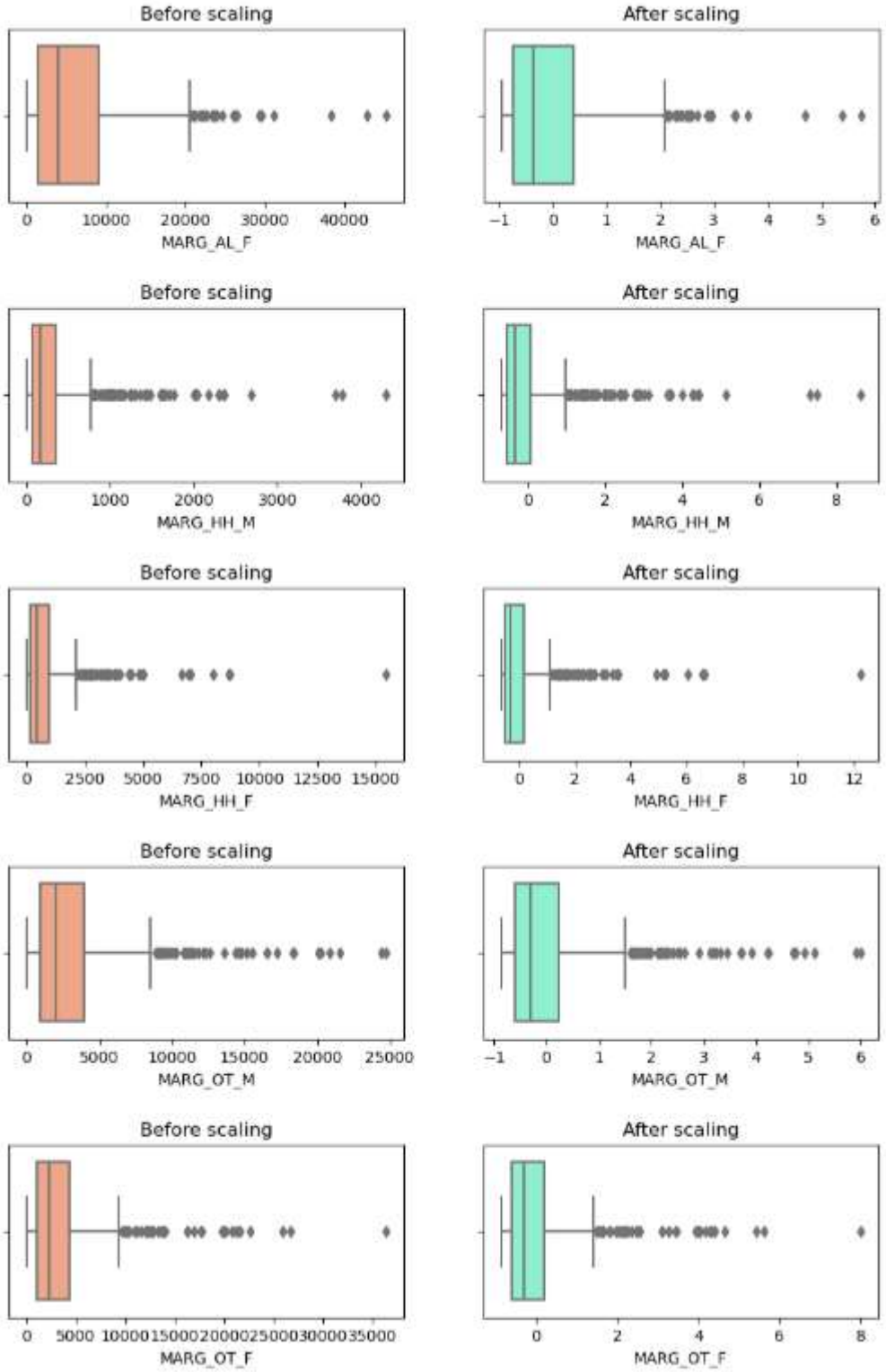

After scaling
0
2
4
6
8
MARG_CL_M

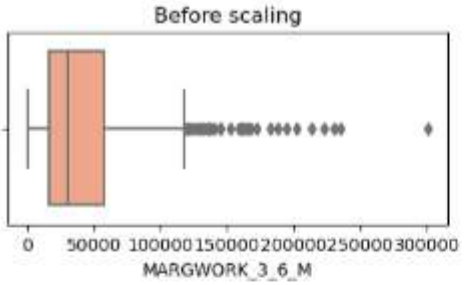

Before scaling
0
10000
20000
30000
40000
MARG_CL_F

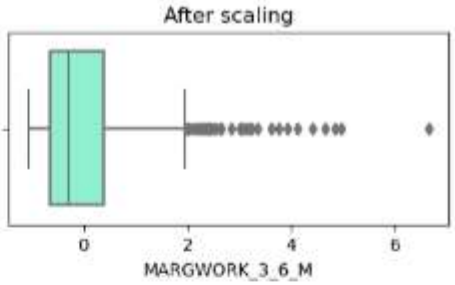

After scaling
0
2
4
6
8
10
12
MARG_CL_F

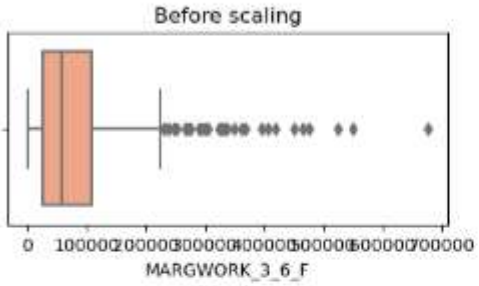

Before scaling
0
5000
10000
15000
20000
MARG_AL_M

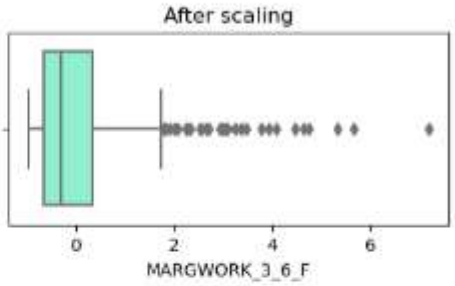

After scaling
−1
0
1
2
3
4
5
MARG_AL_M

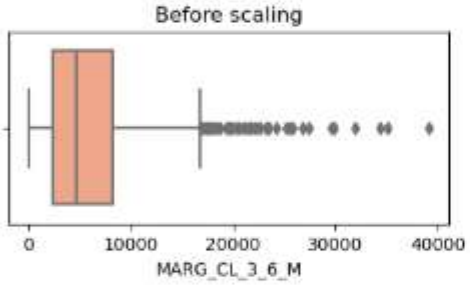

Before scaling
0
10000
20000
30000
40000
MARG_AL_F
After scaling
−1
0
1
2
3
4
5
6
MARG_AL_F
Before scaling
0
1000
2000
3000
4000
MARG_HH_M
After scaling
0
2
4
6
8
MARG_HH_M
Before scaling
0
2500
5000
7500
10000
12500
15000
MARG_HH_F
After scaling
0
2
4
6
8
10
12
MARG_HH_F
Before scaling
0
5000
10000
15000
20000
25000
MARG_OT_M
After scaling
−1
0
1
2
3
4
5
6
MARG_OT_M
Before scaling
0
5000
10000
15000
20000
25000
30000
35000
MARG_OT_F
After scaling
0
2
4
6
8
MARG_OT_F

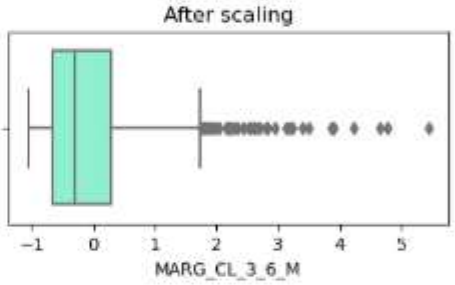

Before scaling
0
50000
100000
150000
200000
250000
300000
MARGWORK_3_6_M

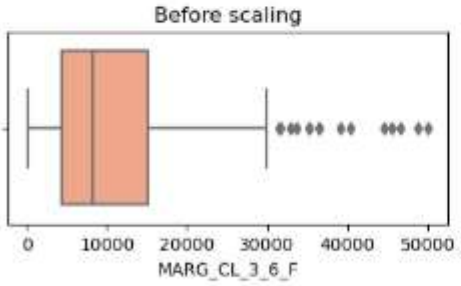

After scaling
0
2
4
6
MARGWORK_3_6_M

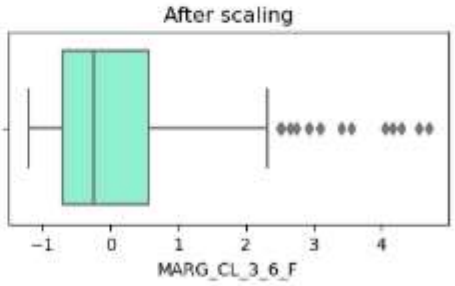

Before scaling
0
100000
200000
300000
400000
500000
600000
700000
MARGWORK_3_6_F

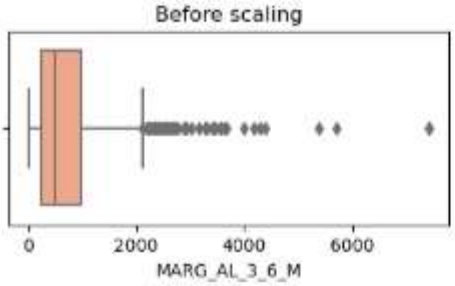

After scaling
0
2
4
6
MARGWORK_3_6_F

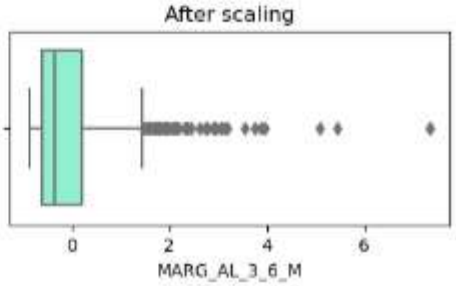

Before scaling
0
10000
20000
30000
40000
MARG_CL_3_6_M

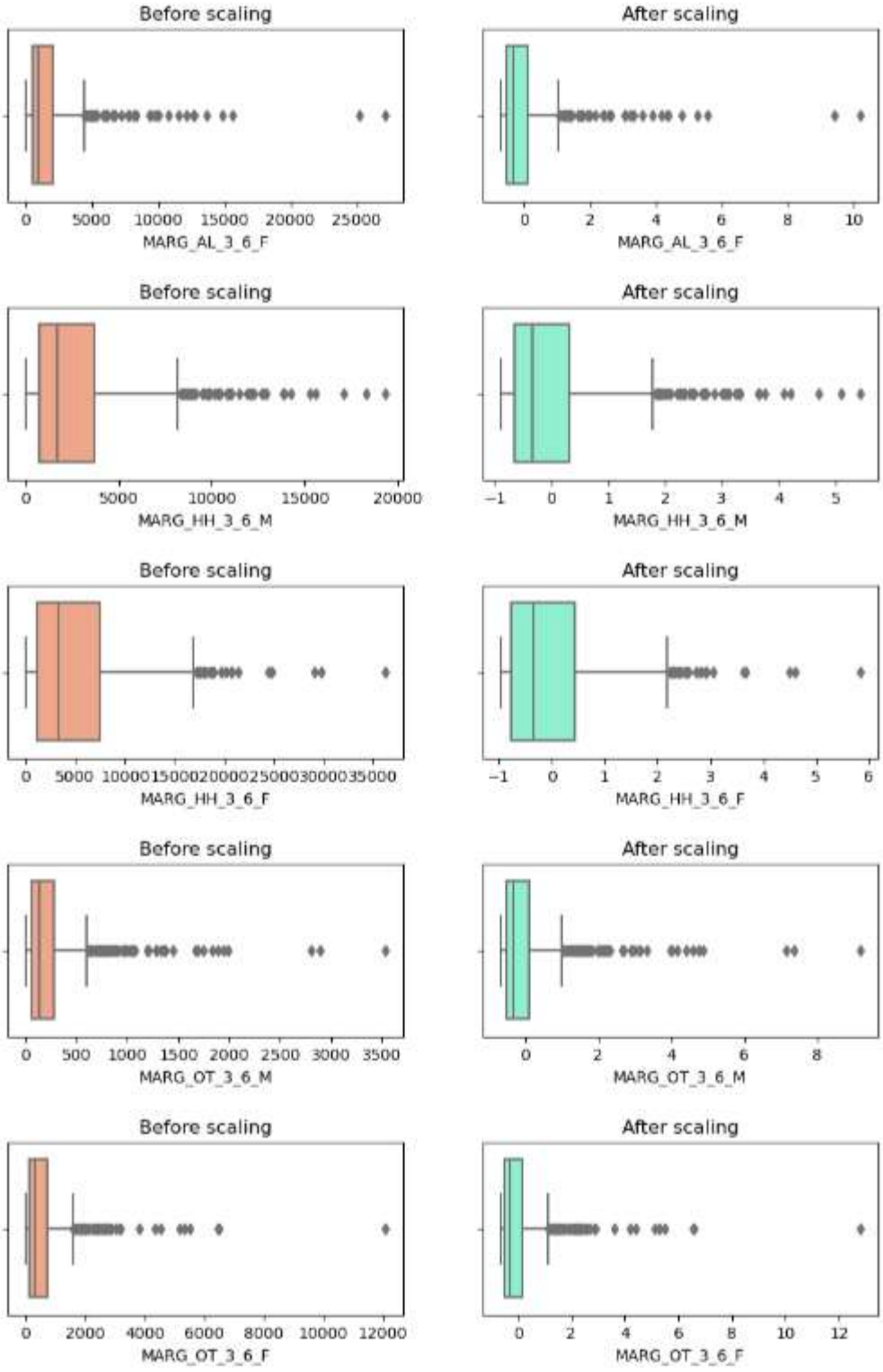

After scaling
−1
0
1
2
3
4
5
MARG_CL_3_6_M

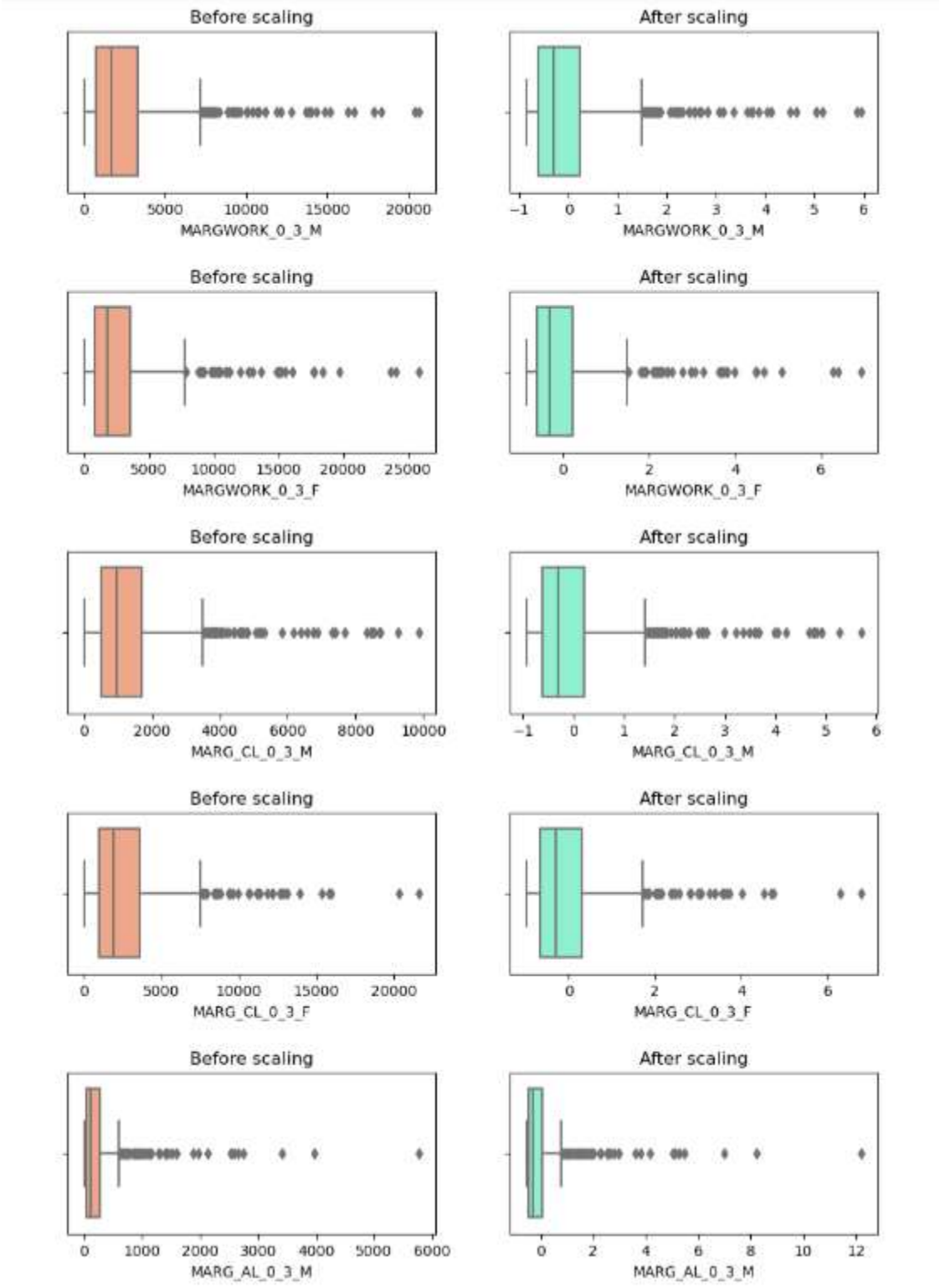

Before scaling
0
10000
20000
30000
40000
50000
MARG_CL_3_6_F

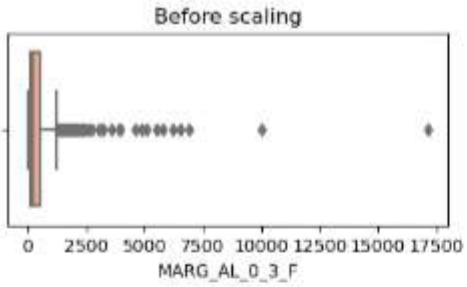

After scaling
−1
0
1
2
3
4
MARG_CL_3_6_F

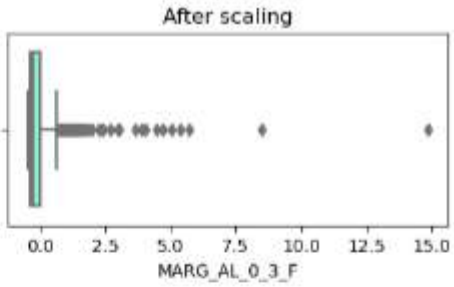

Before scaling
0
2000
4000
6000
MARG_AL_3_6_M

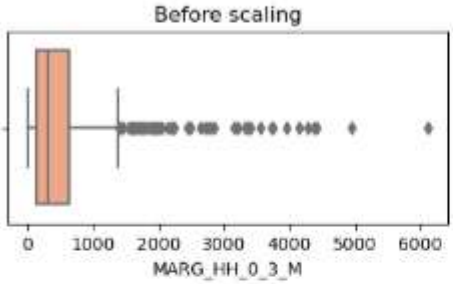

After scaling
0
2
4
6
MARG_AL_3_6_M

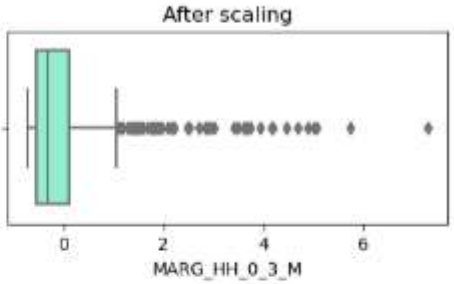

Before scaling
0 5000 10000 15000 20000 25000
MARG_AL_3_6_F
After scaling
0 2 4 6 8 10
MARG_AL_3_6_F
Before scaling
0 5000 10000 15000 20000
MARG_HH_3_6_M
After scaling
−1 0 1 2 3 4 5
MARG_HH_3_6_M
Before scaling
0 5000 10000150002000025000300003500
MARG_HH_3_6_F
After scaling
−1 0 1 2 3 4 5 6
MARG_HH_3_6_F
Before scaling
0 500 1000 1500 2000 2500 3000 3500
MARG_OT_3_6_M
After scaling
0 2 4 6 8
MARG_OT_3_6_M
Before scaling
0 2000 4000 6000 8000 10000 12000
MARG_OT_3_6_F
After scaling
0 2 4 6 8 10 12
MARG_OT_3_6_F

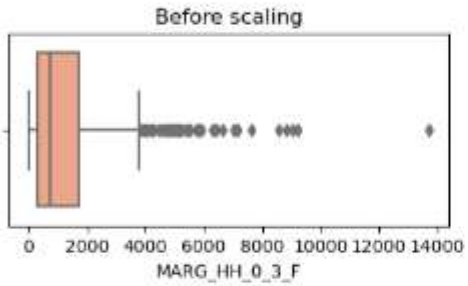

Before scaling
0
5000
10000
15000
20000
MARGWORK_0_3_M
After scaling
−1
0
1
2
3
4
5
6
MARGWORK_0_3_M
Before scaling
0
5000
10000
15000
20000
25000
MARGWORK_0_3_F
After scaling
0
2
4
6
MARGWORK_0_3_F
Before scaling
0
2000
4000
6000
8000
10000
MARG_CL_0_3_M
After scaling
−1
0
1
2
3
4
5
6
MARG_CL_0_3_M
Before scaling
0
5000
10000
15000
20000
MARG_CL_0_3_F
After scaling
0
2
4
6
MARG_CL_0_3_F
Before scaling
0
1000
2000
3000
4000
5000
6000
MARG_AL_0_3_M
After scaling
0
2
4
6
8
10
12
MARG_AL_0_3_M

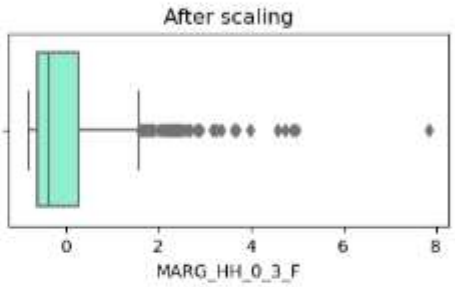

Before scaling
0 2500 5000 7500 10000 12500 15000 17500
MARG_AL_0_3_F

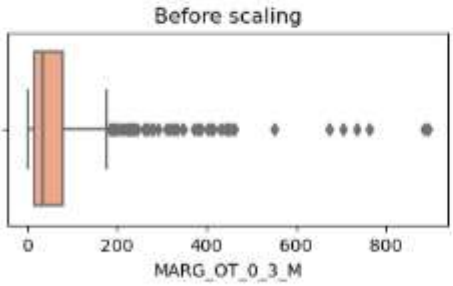

After scaling
0.0 2.5 5.0 7.5 10.0 12.5 15.0
MARG_AL_0_3_F

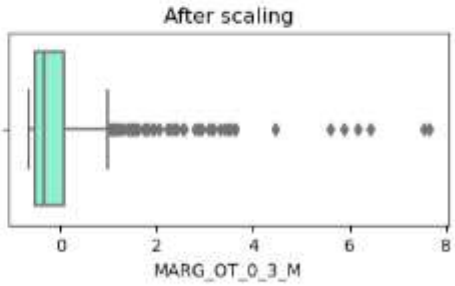

Before scaling
0 1000 2000 3000 4000 5000 6000
MARG_HH_0_3_M

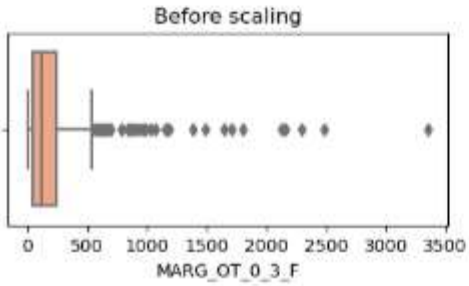

After scaling
0 2 4 6
MARG_HH_0_3_M

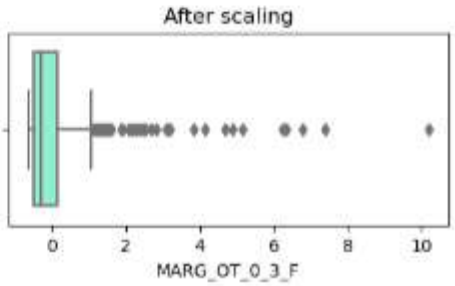

Before scaling
0 2000 4000 6000 8000 10000 12000 14000
MARG_HH_0_3_F


After scaling
0 2 4 6 8
MARG_HH_0_3_F


Before scaling
0 200 400 600 800
MARG_OT_0_3_M


After scaling
0 2 4 6 8
MARG_OT_0_3_M


Before scaling
0 500 1000 1500 2000 2500 3000 3500
MARG_OT_0_3_F


After scaling
0 2 4 6 8 10
MARG_OT_0_3_F

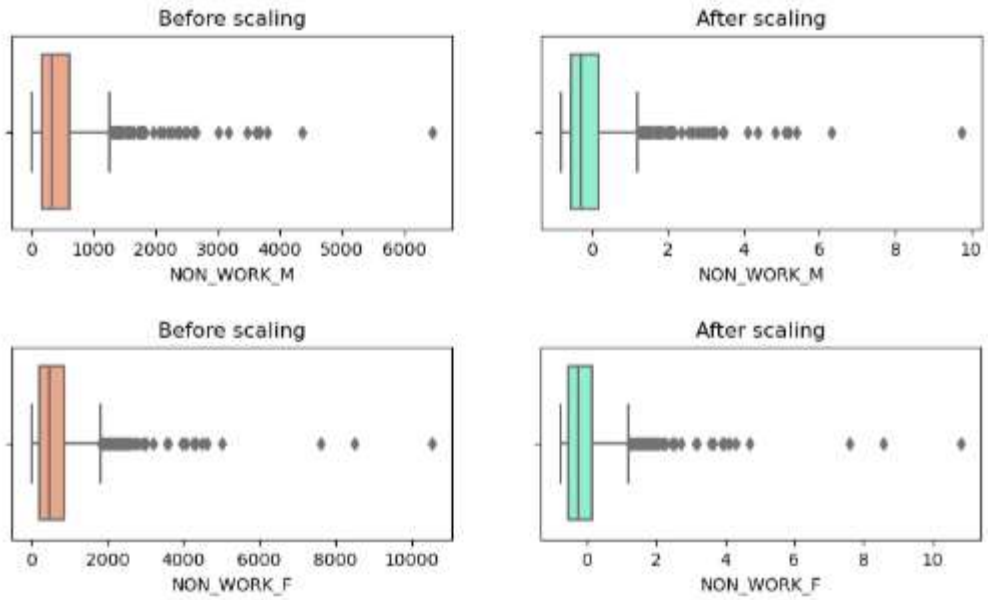


Fig.2.15. Boxplot distributions before and after scaling

As seen from the above boxplots, though the data is scaled, it does not impact the outliers in any field.

E. Part 2 - PCA: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.

- **Bartlett- Sphericity test:**
  - p=0.0
  - Since p value is less than 0.05, we reject the null hypothesis and conclude that there is a significant correlation between the factors.
- **kmo test:**
  - Output=0.81
  - Since the output of the kmo test is 0.8, the given dataset has a great suitability for factor analysis
- Co-variance matrix:

```
Covariance matrix:
 [[0.06173772 0.05024455 0.03317398 ... 0.03498547 0.03025035 0.02714776]
 [0.05024455 0.07966605 0.03858426 ... 0.0504162  0.03813376 0.03399651]
 [0.03317398 0.03858426 0.05855969 ... 0.04253603 0.03684299 0.03313212]
 ...
 [0.03498547 0.0504162  0.04253603 ... 0.04525044 0.03695346 0.03424301]
 [0.03025035 0.03813376 0.03684299 ... 0.03695346 0.03262291 0.03059199]
 [0.02714776 0.03399651 0.03313212 ... 0.03424301 0.03059199 0.03110606]]
```

  -
- Eigen-vectors:

```
The eigen vectors are:
 [[ 0.16  0.17  0.17 ...  0.13  0.15  0.13]
 [-0.13 -0.09 -0.1  ...  0.05 -0.07 -0.07]
 [-0.    0.06  0.04 ... -0.08  0.11  0.1 ]
 ...
 [ 0.    0.21  0.25 ... -0.07  0.   -0.07]
 [ 0.    0.29 -0.21 ...  0.04 -0.03  0.01]
 [-0.    0.19  0.03 ... -0.03 -0.14 -0.02]]
```

- Eigen-values:

```
The eigen values are:
 [3.181e+01 7.870e+00 4.150e+00 3.670e+00 2.210e+00 1.940e+00 1.180e+00
 7.500e-01 6.200e-01 5.300e-01 4.300e-01 3.500e-01 3.000e-01 2.800e-01
 1.900e-01 1.400e-01 1.100e-01 1.100e-01 1.000e-01 8.000e-02 6.000e-02
 4.000e-02 4.000e-02 3.000e-02 3.000e-02 2.000e-02 1.000e-02 1.000e-02
 1.000e-02 1.000e-02 1.000e-02 1.000e-02 0.000e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 0.000e+00]
```

F. Part 2 - PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.

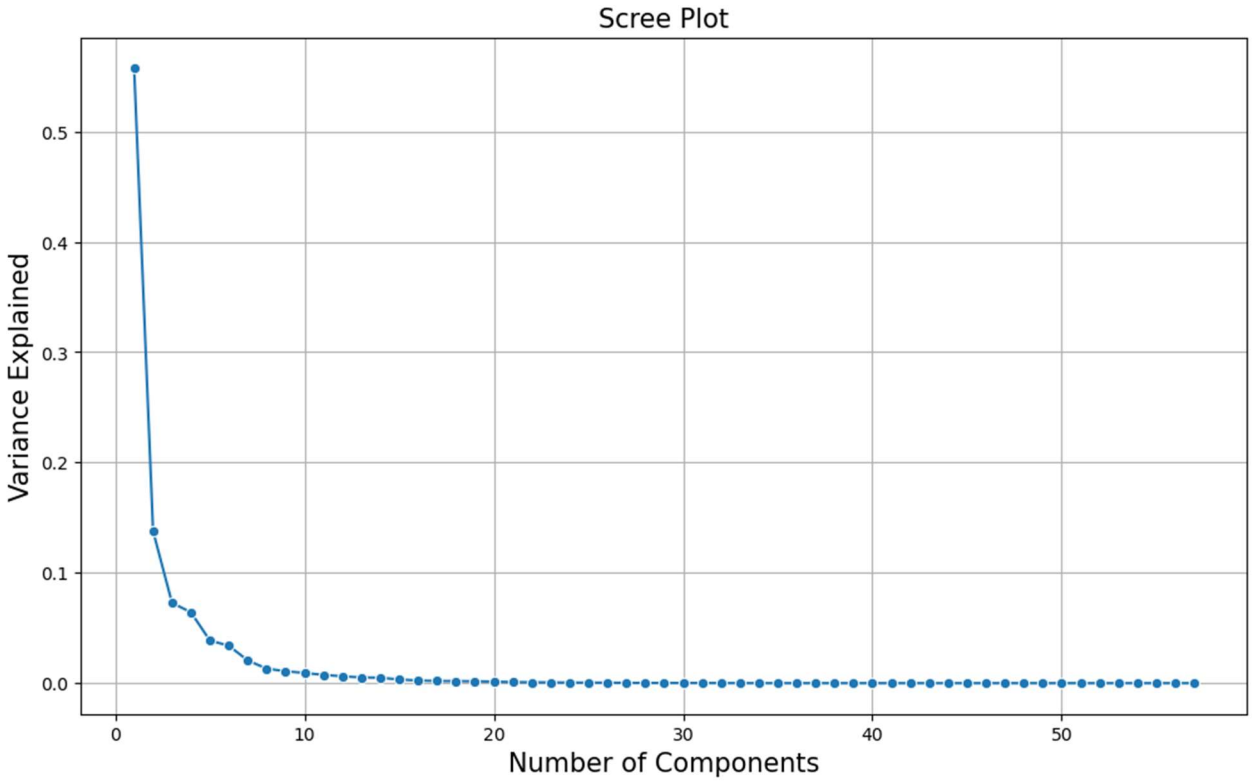


Fig.2.16. Scree plot- line plot

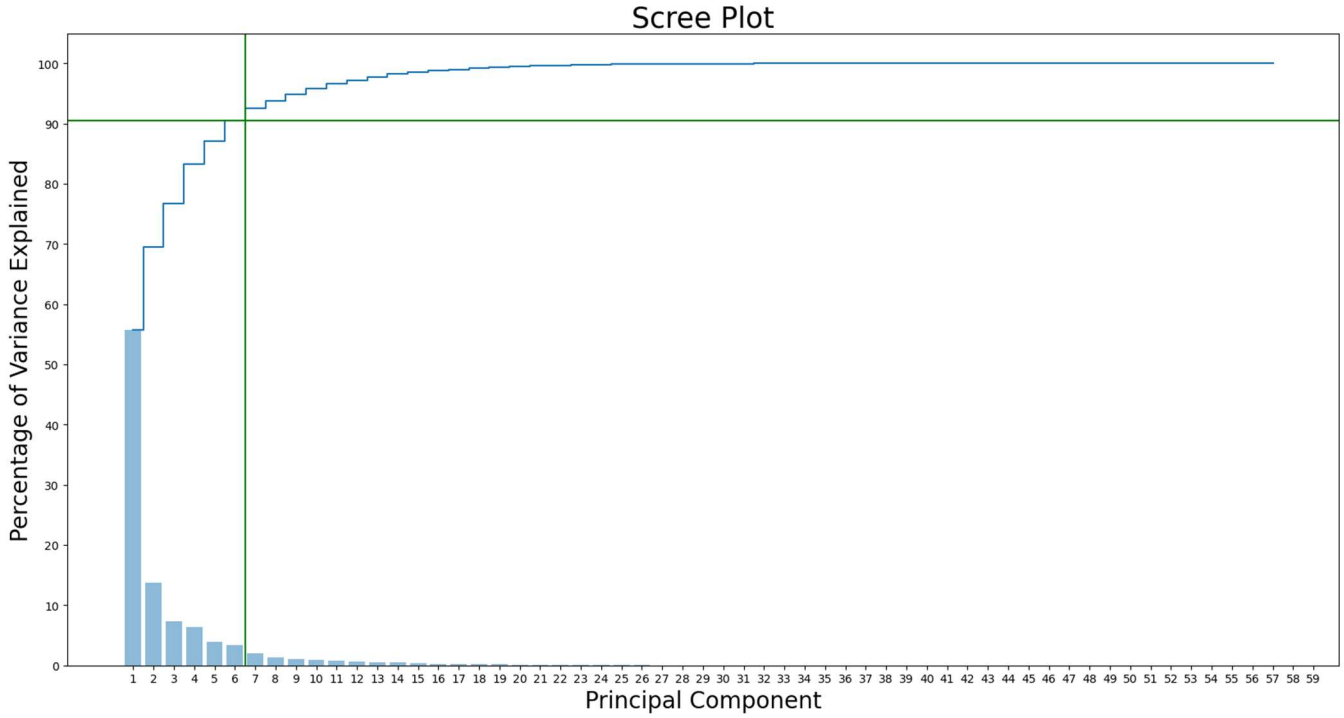


Fig.2.17. Scree plot- Bar and Step plot

Summary:

From the plots above, 90.47% of the variance is explained by the first 6 principal components. Hence, the optimum number of PCs is 6.

G. Part 2 - 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.

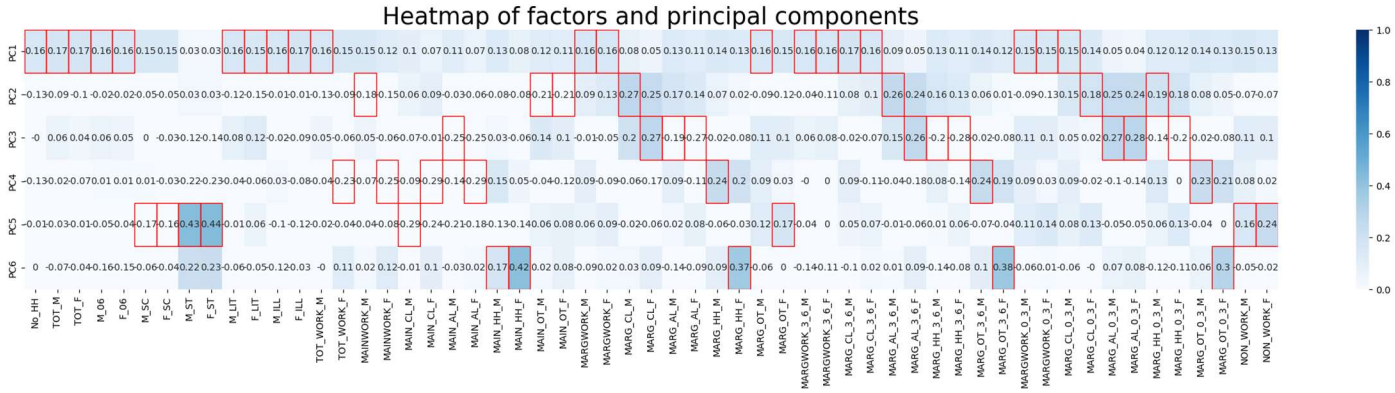


Fig.2.18. Comparison- PCs and actual columns

Observations and Inferences:

- PC1 explains close to 56% of the variance, as gathered from the PCA
- PC1- Mostly contains all aggregate population parameters- like No_HH,Total males and females, total children, Total literates and illiterates etc
- PC2- Contains the aggregate working related parameters- like mainwork males,main_ot males and females, marg_work cultivators and agriculturalists
- PC3- Mostly contains fields related to Agriculture cultivators and lobourers
- PC4- Mostly contains factors involving female workers across different segments

- PC5- Contains minority groups and non-working population
- PC6- Contains household and other workers

	State Code	Dist.Code	State	Area Name	Population	Working	Agri_related	Female_workers	Not_working	Household_other
0	1	1	Jammu & Kashmir	Kupwara	-4.62	0.14	0.33	1.54	0.35	-0.42
1	1	2	Jammu & Kashmir	Badgam	-4.77	-0.11	0.24	1.96	-0.15	0.42
2	1	3	Jammu & Kashmir	Leh(Ladakh)	-5.96	-0.29	0.37	0.62	0.48	0.28
3	1	4	Jammu & Kashmir	Kargil	-6.28	-0.50	0.21	1.07	0.30	0.05
4	1	5	Jammu & Kashmir	Punch	-4.48	0.89	1.08	0.54	0.80	0.34

Fig.2.19. First 5 rows of dataset after PCA dimensionality reduction

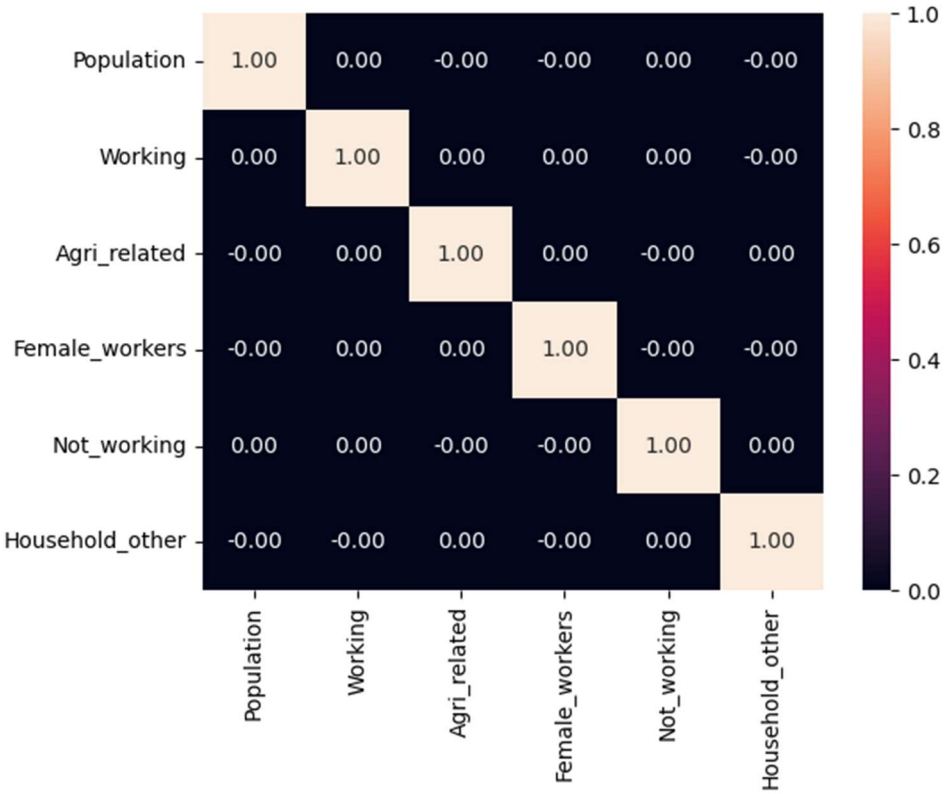


Fig.2.20. Correlation heatmap after PCA

H. Part 2 - PCA: Write linear equation for first PC.

Equation:

PC1= w1X1+w2X2+…+wnXn

That is,

PC1 = 0.16 * X1 + 0.17 * X2 + 0.17 * X3 + 0.16 * X4 + 0.16 * X5 + 0.15 * X6 + 0.15 * X7 + 0.03 * X8 + 0.03 * X9 + 0.16 * X10 + 0.15 * X11 + 0.16 * X12 + 0.17 * X13 + 0.16 * X14 + 0.15 * X15 + 0.15 * X16 + 0.12 * X17 + 0.1 * X18 + 0.07 * X19 + 0.11 * X20 + 0.07 * X21 + 0.13 * X22 + 0.08 * X23 + 0.12 * X24 + 0.11 * X25 + 0.16 * X26 + 0.16 * X27 + 0.08 * X28 + 0.05 * X29 + 0.13 * X30 + 0.11 * X31 + 0.14 * X32 + 0.13 * X33 + 0.16 * X34 + 0.15 * X35 + 0.16 * X36 + 0.16 * X37 + 0.17 * X38 + 0.16 * X39 + 0.09 * X40 + 0.05 * X41 + 0.13 * X42 + 0.11 * X43 + 0.14 * X44 + 0.12 * X45 + 0.15 * X46 + 0.15 * X47 + 0.15 * X48 + 0.14 * X49 + 0.05 * X50 + 0.04 * X51 + 0.12 * X52 + 0.12 * X53 + 0.14 * X54 + 0.13 * X55 + 0.15 * X56 + 0.13 * X57