|  |  |
| --- | --- |
|  |  |
| DATA MINING PROJECT |  |
|  |  |
|  | Submitted by,VIDYA V |
|  | PGPDSBA.O.2023.B 02.07.2023 |

**CONTENTS**

|  |  |
| --- | --- |
| Case 1: Digital Ads Data | 5 |
| 1. 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. | 5 |
| 1. Part 1 - Clustering: Treat missing values in CPC, CTR and CPM using the formula given. | 7 |
| 1. 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).e business. | 7 |
| 1. Part 1 - Clustering: Perform z-score scaling and discuss how it affects the speed of the algorithm. | 9 |
| 1. Part 1 - Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance. | 10 |
| 1. Part 1 - Clustering: Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm. | 13 |
| 1. Part 1 - Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters. | 14 |
| 1. 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]. | 14 |
| 1. Part 1 - Clustering: Conclude the project by providing summary of your learnings. | 17 |
| Case 2: PCA- India Census Data | 18 |
| 1. Part 2 - PCA: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc. | 18 |
| 1. 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 | 21 |
| 1. Part 2 - PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary? | 44 |
| 1. 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. | 44 |
| 1. Part 2 - PCA: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector. | 58 |
| 1. Part 2 - PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot. | 59 |
| 1. 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. | 60 |
| 1. Part 2 - PCA: Write linear equation for first PC. | 61 |

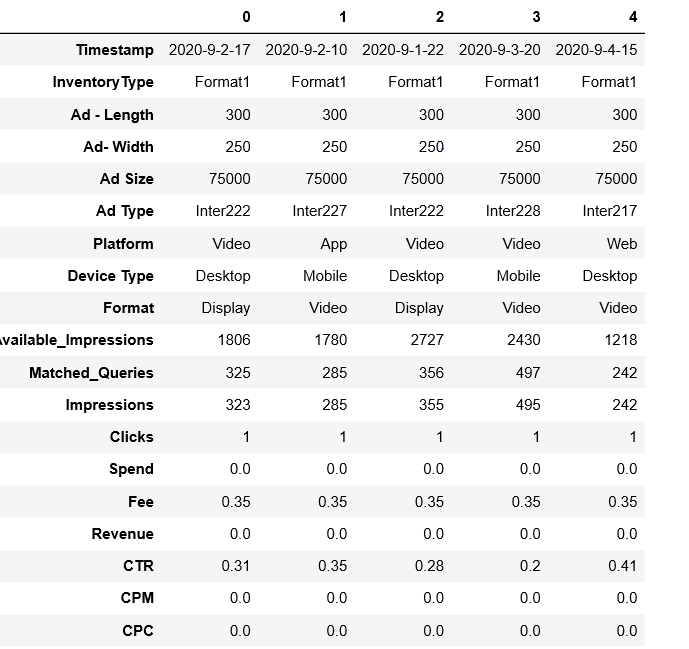
**List of Figures**

|  |  |
| --- | --- |
| **Name** | **Page No.** |
| **Fig.1.1. Viewing the dataset** | 5 |
| **Fig.1.2. Data info** | 6 |
| **Fig.1.3. Data Description** | 6 |
| **Fig.1.4. Dataset info output after null value treatment** | 7 |
| **Fig.1.5. Boxplots of the fields in the dataset showing outliers** | 8 |
| **Fig. 1.6. Boxplots after outlier treatment** | 9 |
| **Fig. 1.7. Data before scaling** | 9 |
| **Fig. 1.8. Data after scaling** | 10 |
| **Fig. 1.9. Dendrogram showing the last 10 clusters** | 11 |
| **Fig. 1.10. Number of rows in each cluster after hierarchical clustering** | 11 |
| **Fig.1.11. Mean of various columns across clusters after clustering** | 12 |
| **Fig.1.12. Sum totals of various columns across clusters after clustering** | 13 |
| **Fig.1.13. Elbow plot for k-means algorithm** | 14 |
| **Fig.1.14. Cluster distribution** | 15 |
| **Fig.1.15 Barplot showing the distribution of Ads across clusters and Devices.** | 15 |
| **Fig.1.16. Sums of factors across clusters generated by k-means algorithm** | 16 |
| **Fig.1.17. Means of factors across clusters generated by k-means algorithm** | 17 |
| **Fig.2.1. Census data- info** | 18 |
| **Fig.2.2. Data description** | 19-20 |
| **Fig.2.3 Univariate Analysis- Numerical fields** | 22-34 |
| **Fig.2.4. Univariate Analysis- Categorical field** | 35 |
| **Fig.2.5. Heatmap of numerical fields** | 36 |
| **Fig.2.6. Statewise mean literacy ratio** | 37 |
| **Fig.2.7. Statewise mean Illiteracy ratio** | 38 |
| **Fig.2.8. Statewise female illiteracy ratio** | 39 |
| **Fig.2.9. Comaprison of Female literacy and total literacy ratio** | 39 |
| **Fig.2.10. Statewise mean working ratio** | 40 |
| **Fig.2.11. Statewise Occupation chart** | 41 |
| **Fig.2.12. Statewise mean Agri\_ralated workers** | 42 |
| **Fig.2.13. Mean minority ratio** | 43 |
| **Fig.2.14. Data description after scaling** | 45-46 |
| **Fig.2.15. Boxplot distributions before and after scaling** | 47-58 |
| **Fig.2.16. Scree plot- line plot** | 59 |
| **Fig.2.17. Scree plot- Bar and Step plot** | 60 |
| **Fig.2.18. Comparison- PCs and actual columns** | 60 |
| **Fig.2.19. First 5 rows of dataset after PCA dimensionality reduction** | 61 |
| **Fig.2.20. Correlation heatmap after PCA** | 61 |

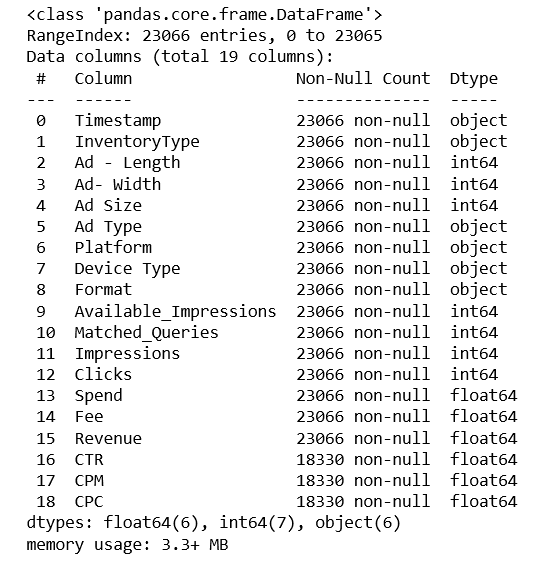
# Case 1: DIGITAL ADS DATA

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.

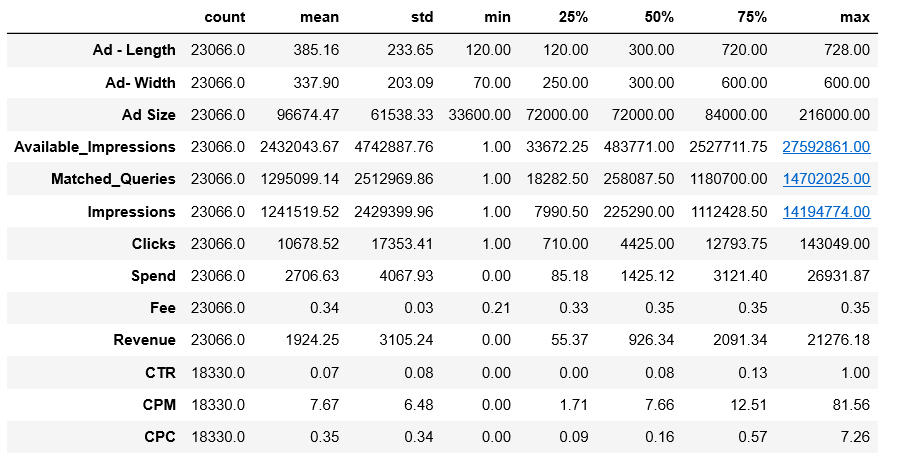
1. **Part 1 - Clustering: Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.**



**Fig.1.1. Viewing the dataset**



**Fig.1.2. Data info**

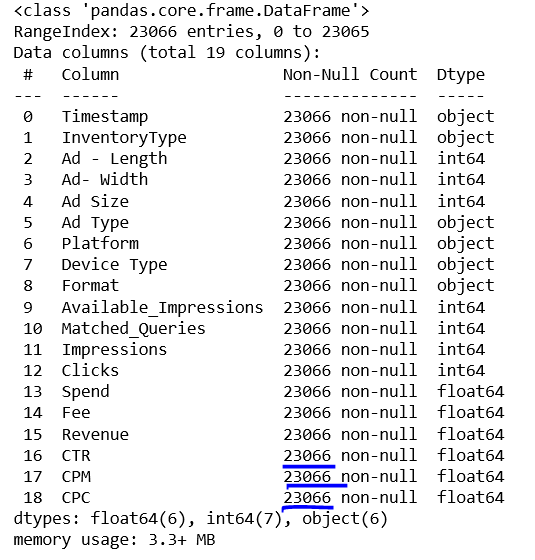


**Fig.1.3. Data Description**

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

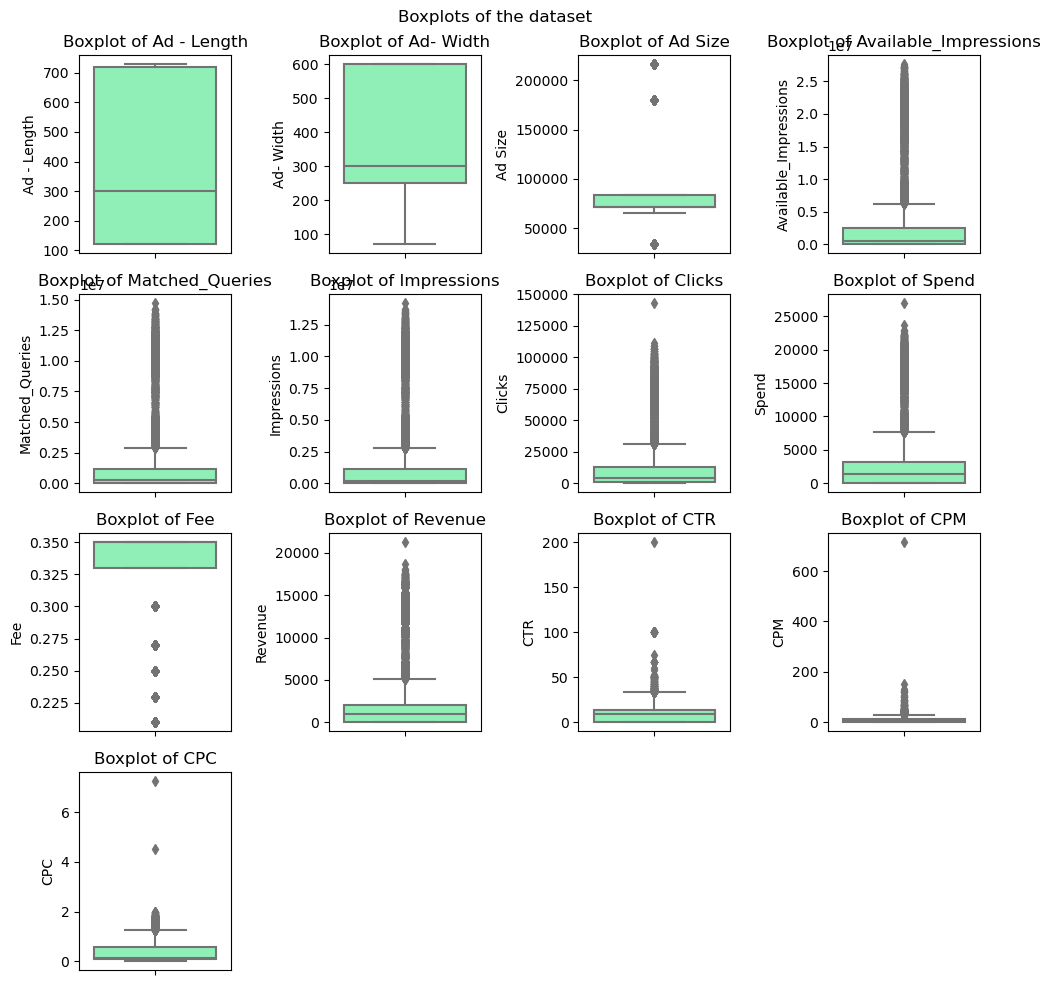
* It has 23066 rows and 19 columns.
* The columns include the length,width and size of ad, the number of times the ad is shown, the total number of clicks, the money spent on each variation of the ad, the revenue from the ads, Click through rate, cost per click, and cost per impression
* Null values 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

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



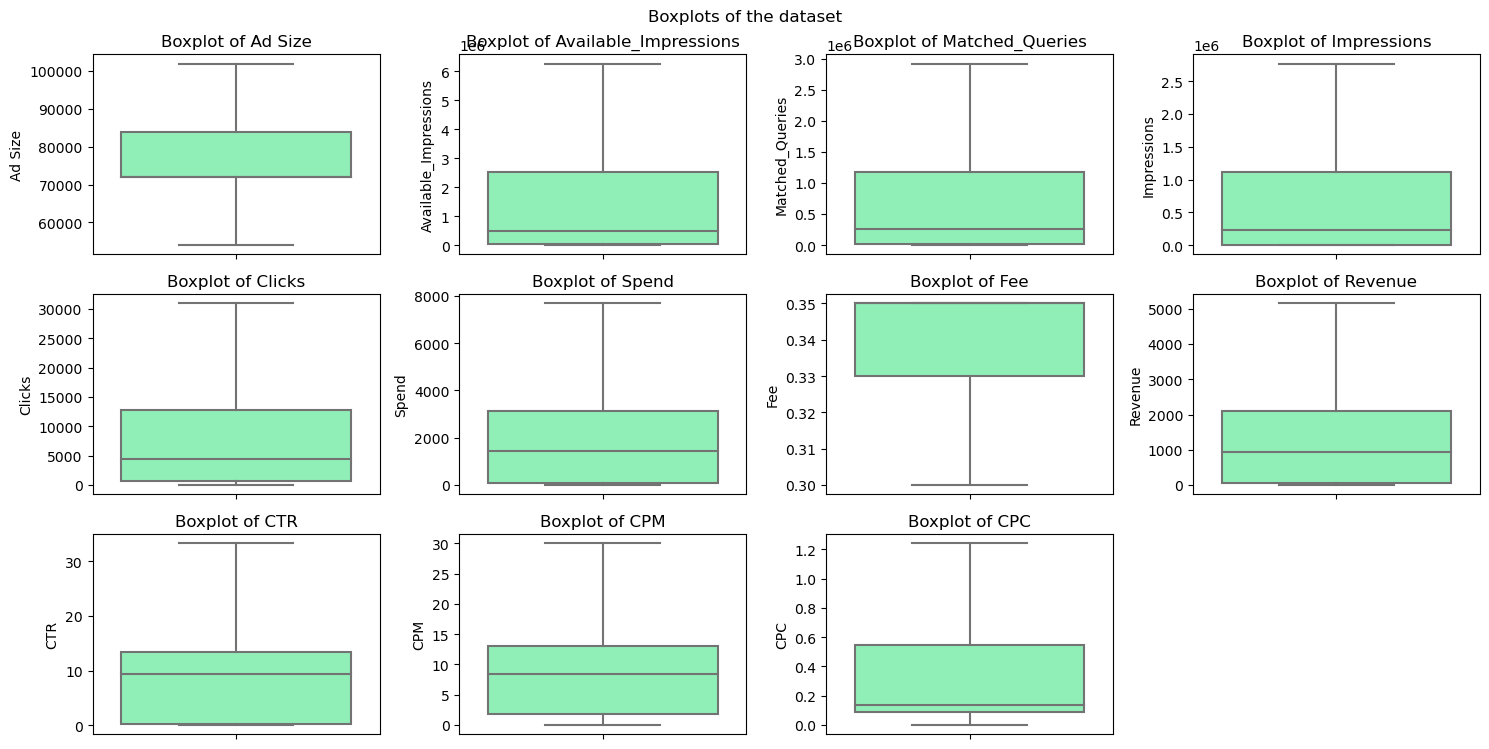
**Fig.1.4. Dataset info output after null value treatment**

1. **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)

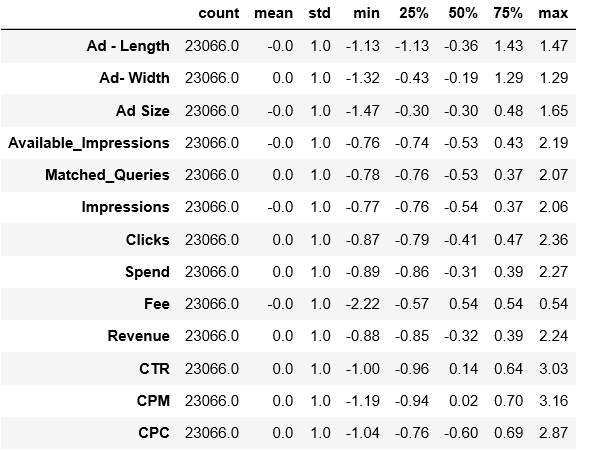


**Fig. 1.6. Boxplots after outlier treatment**

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



**Fig.1.7. Data Before Scaling**

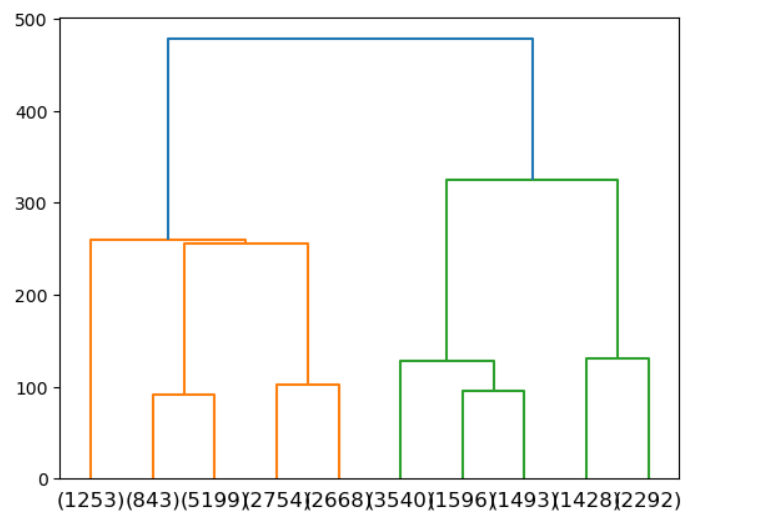


**Fig.1.8. Data after scaling**

**Observations:**

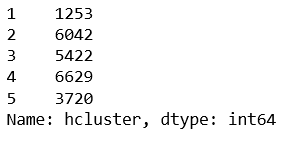
* After scaling, data has now been centered around 0, with standard deviation 1 for all the columns
* This will ensure that each of the columns carry the same weightage while performing clustering with a distance-based algorithm, and ensures that the variables with a higher value do not dominate those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

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



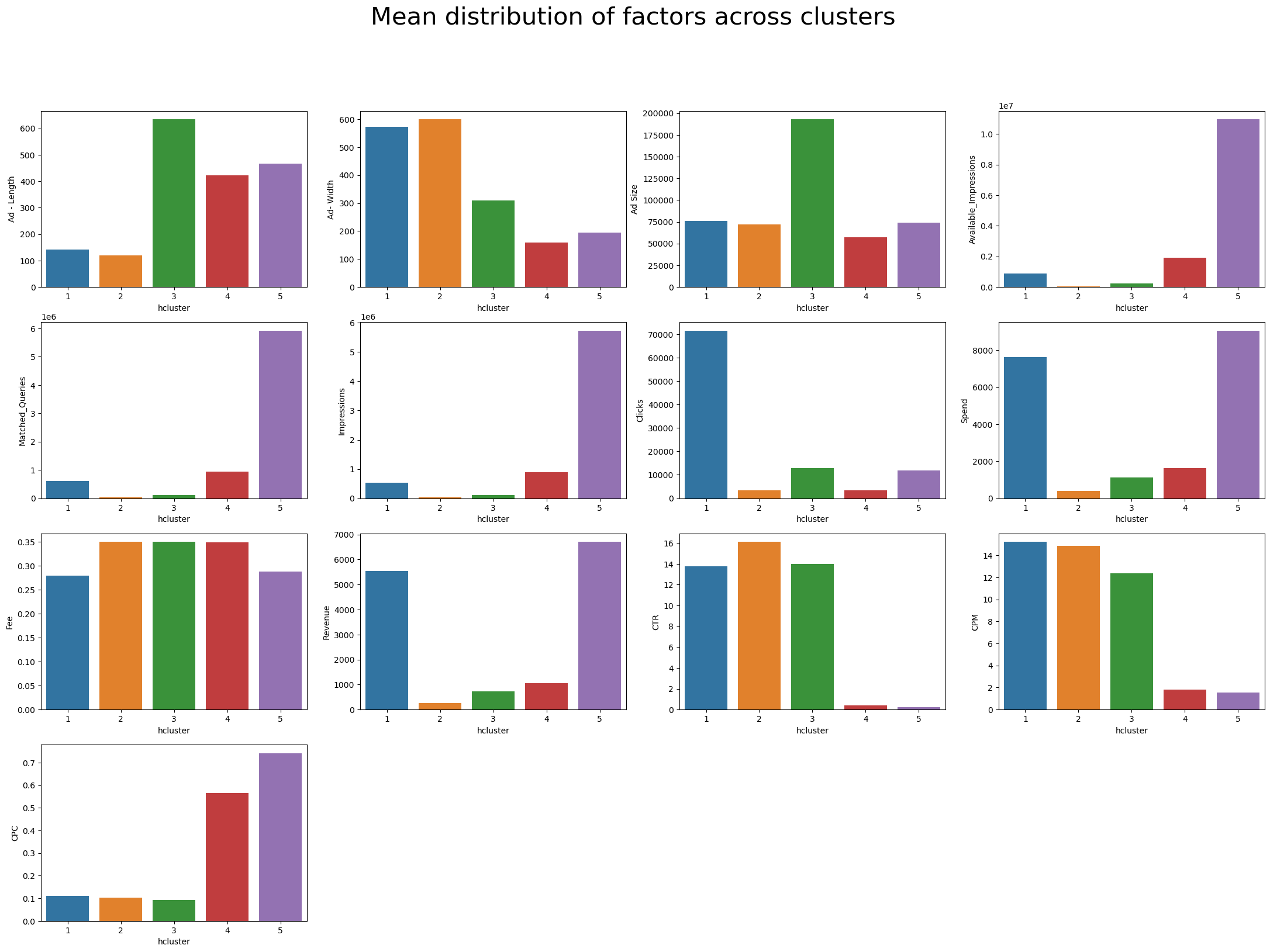
**Fig.1.9. Dendrogram showing last 10 clusters**

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

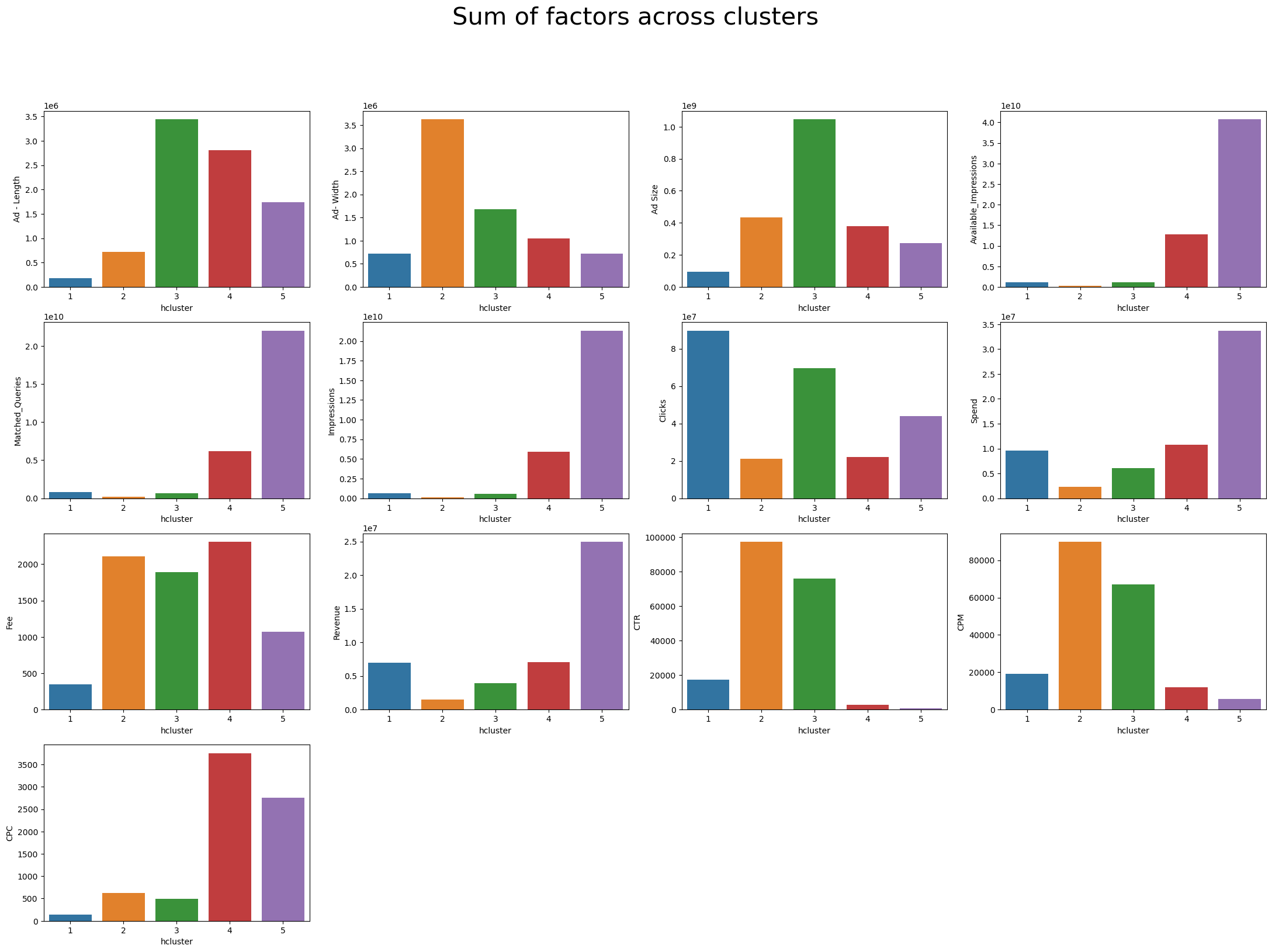


**Fig. 1.10. Number of rows in each cluster after hierarchical clustering**

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



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

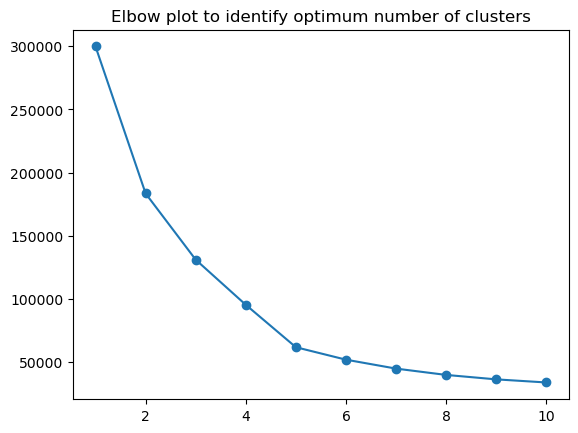


**Fig.1.12. Sum totals of various columns across clusters**

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.

1. **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.

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

**Output:**



**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

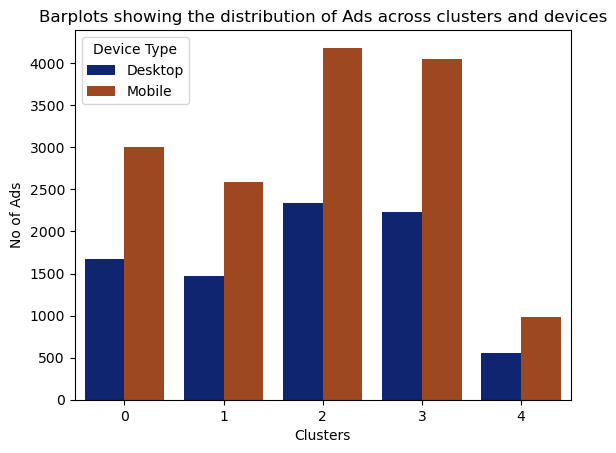
1. **Part 1 - Clustering: Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots]**

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

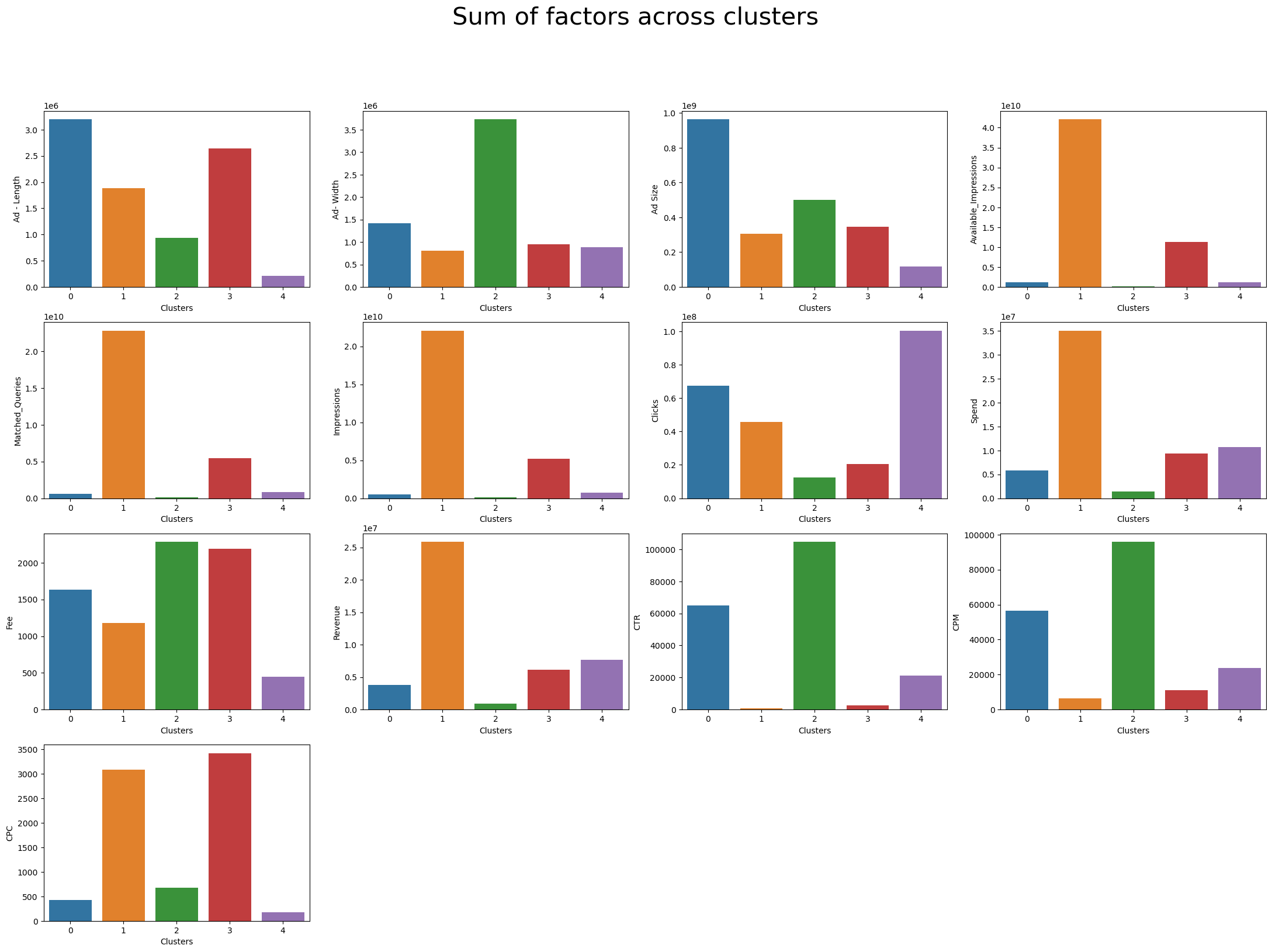


**Fig.1.14. Cluster distribution**

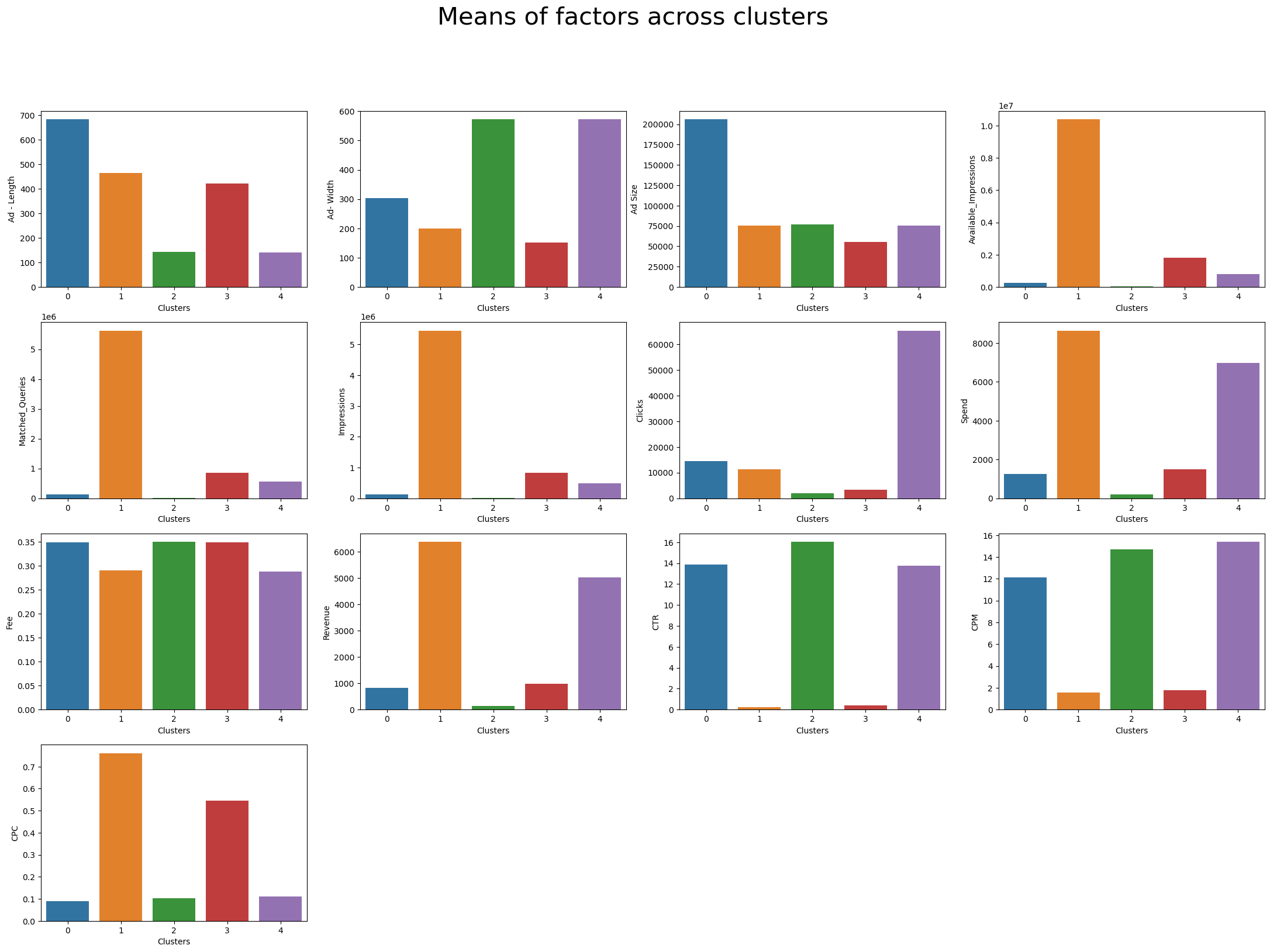
The above is output is represented as bar plot below, with distinction of device types.

****

**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.**

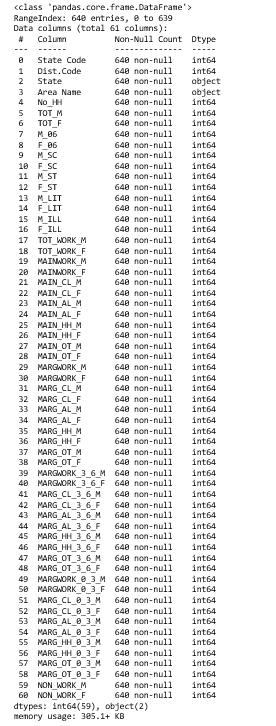
1. **Part 1 - Clustering: Conclude the project by providing summary of your learnings.**

**Summary:**

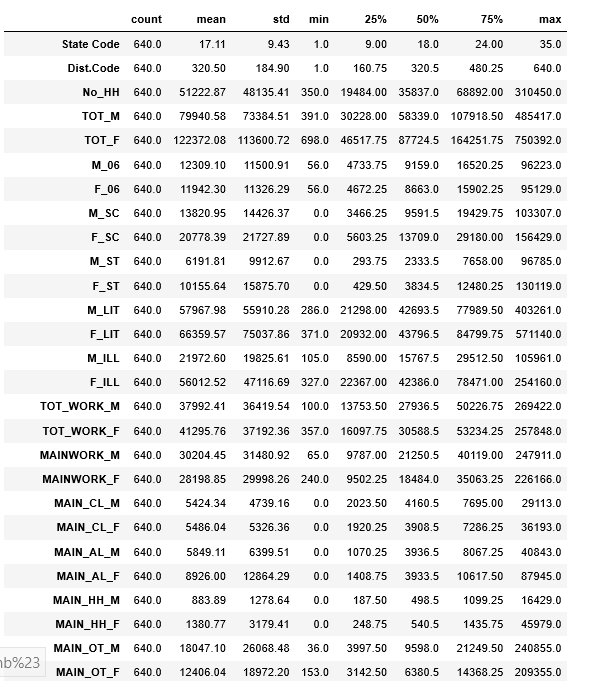
* Cluster 0: Contains the biggest ads. They have very low available impressions, and impressions, and have a moderate CPM. The Spend and revenue are low, though it has a decent CTR
* Cluster 1: Contains moderately sized ads that have a high mean matched\_queries and available\_impressions. The spend 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.

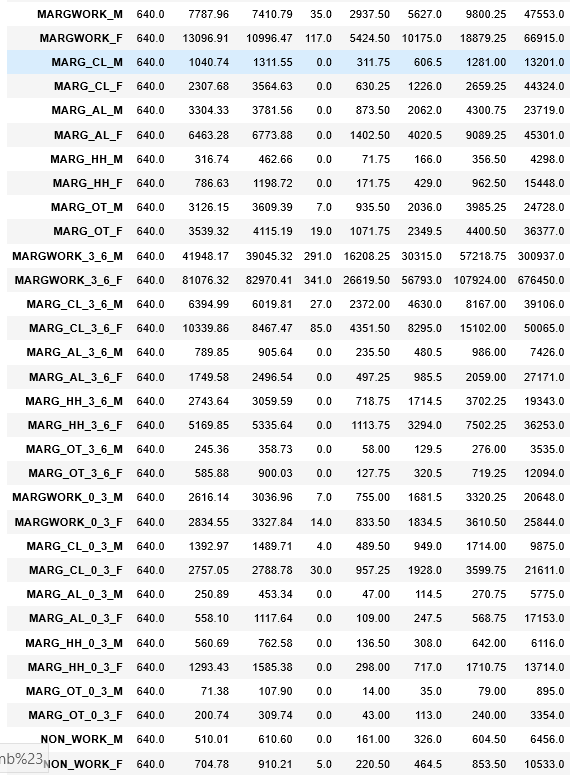
# Case 2: PCA- India Census Data

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



**Fig.2.1. Census data- info**



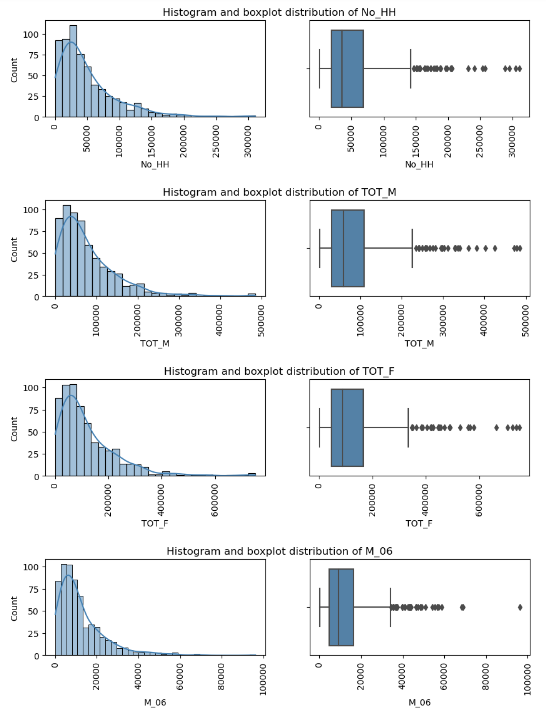


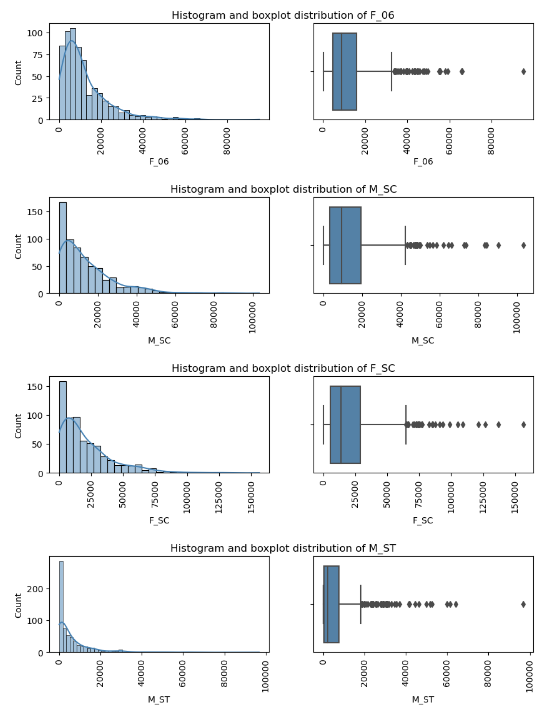
**Fig.2.2. Data description**

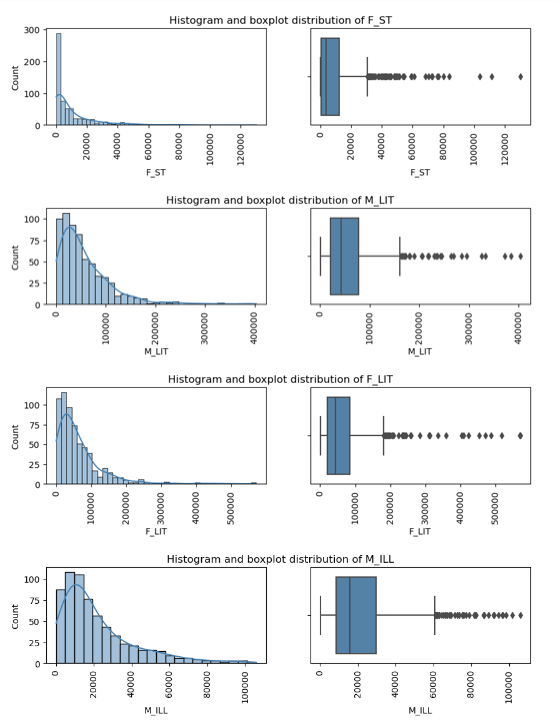
**Observations**:

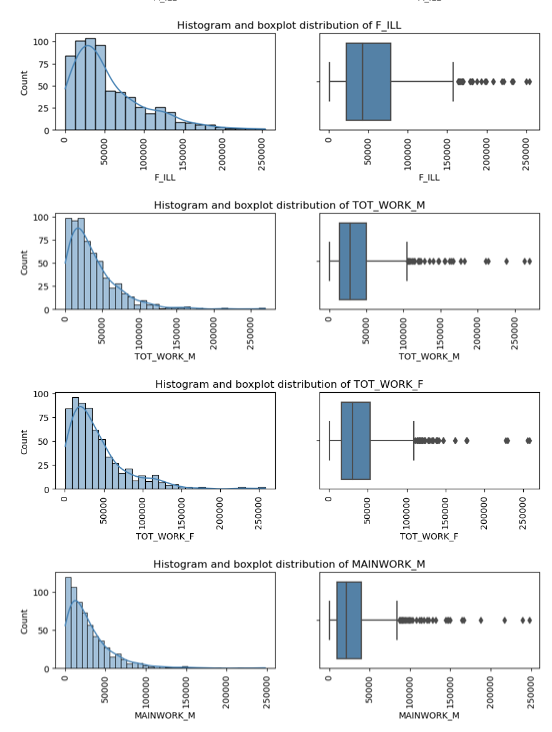
* The given dataset has 640 data points, and 61 fields containing elaborate details of various areas, including gender, population, literacy, occupation etc.
* There are no null values present in the given dataset
* The data magnitudes varies across fields. Scaling is required
* Few fields have a minimum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

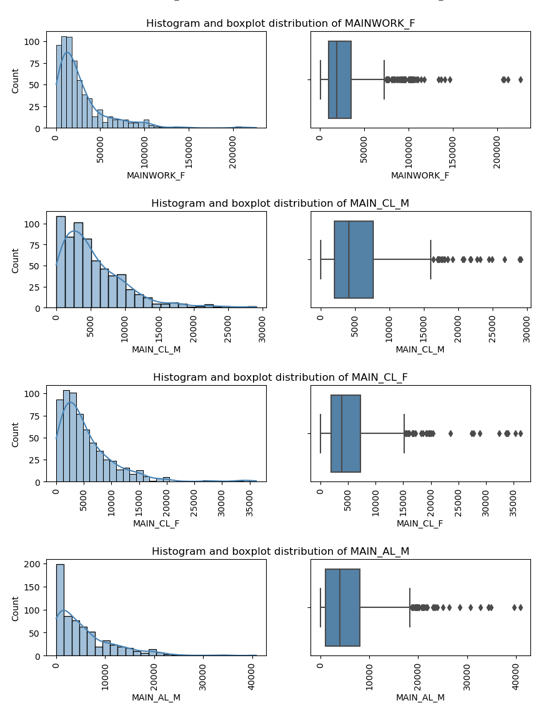
1. **Part 2 - PCA: Perform detailed Exploratory analysis by creating certain questions like (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio? (Example Questions). Pick 5 variables out of the given 24 variables below for EDA: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

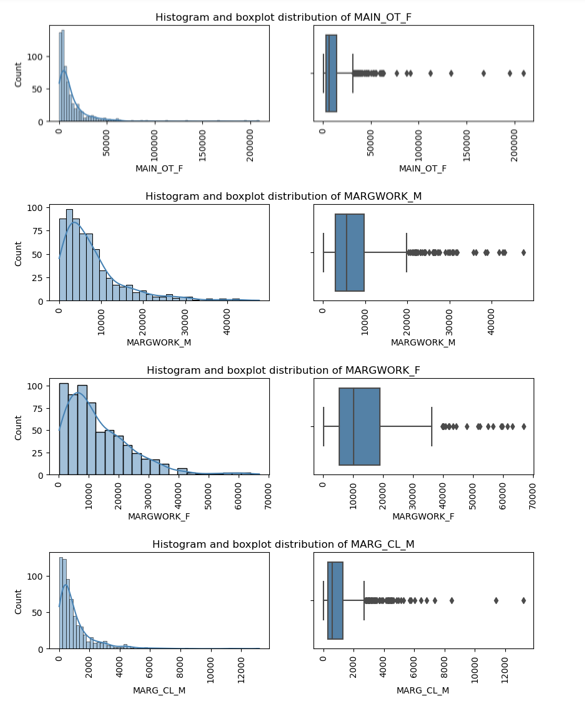


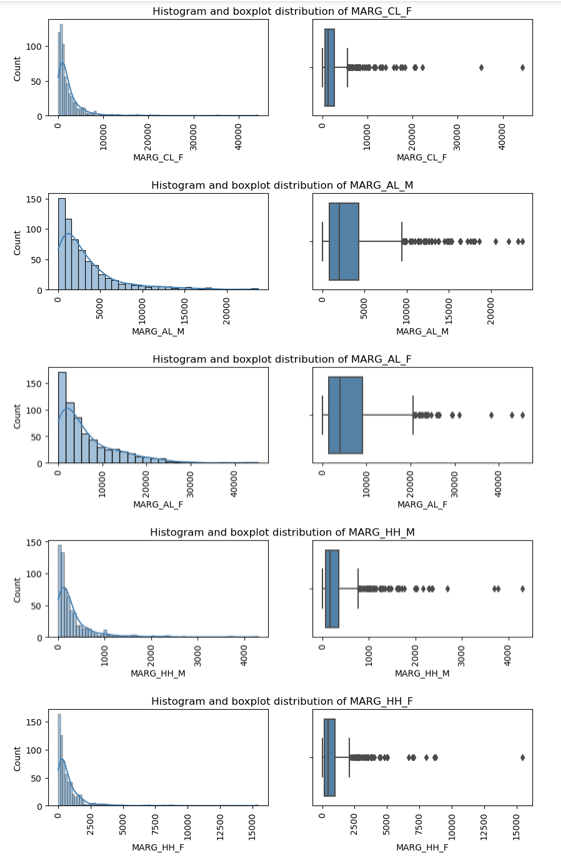


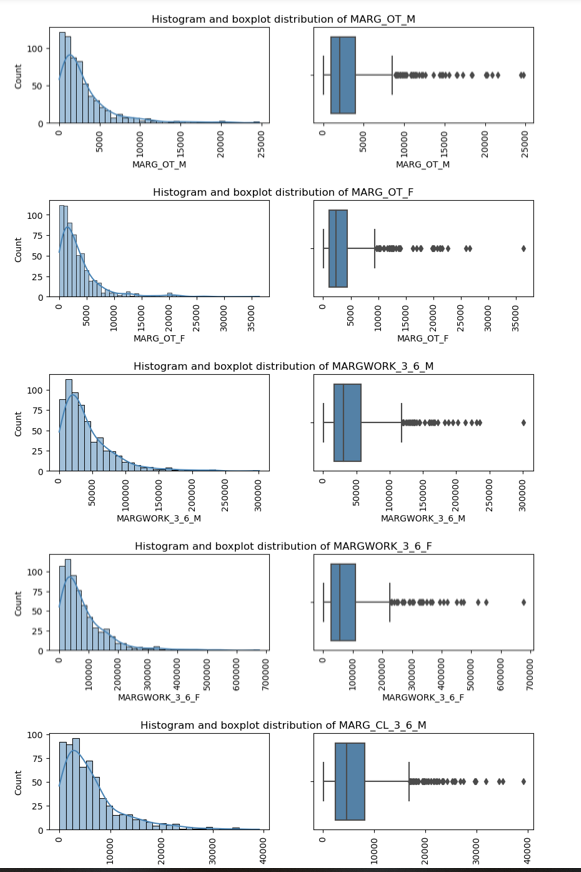




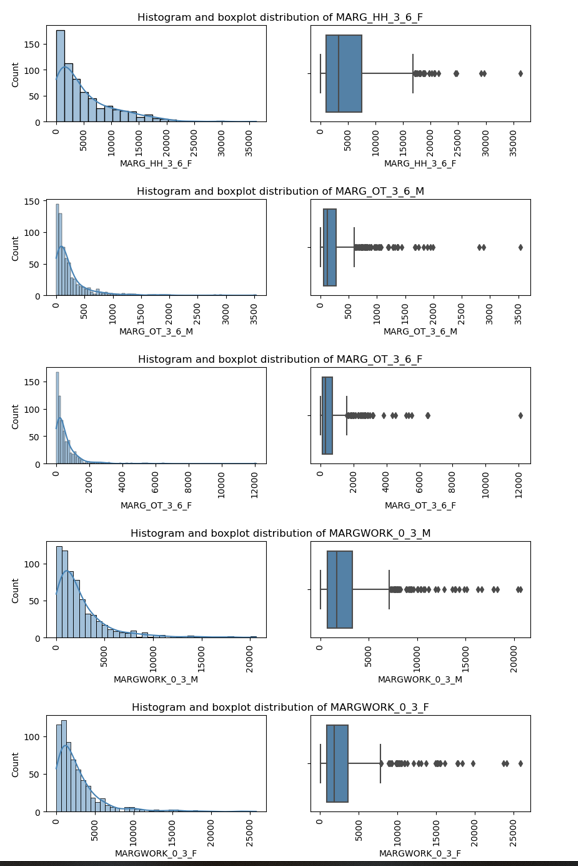


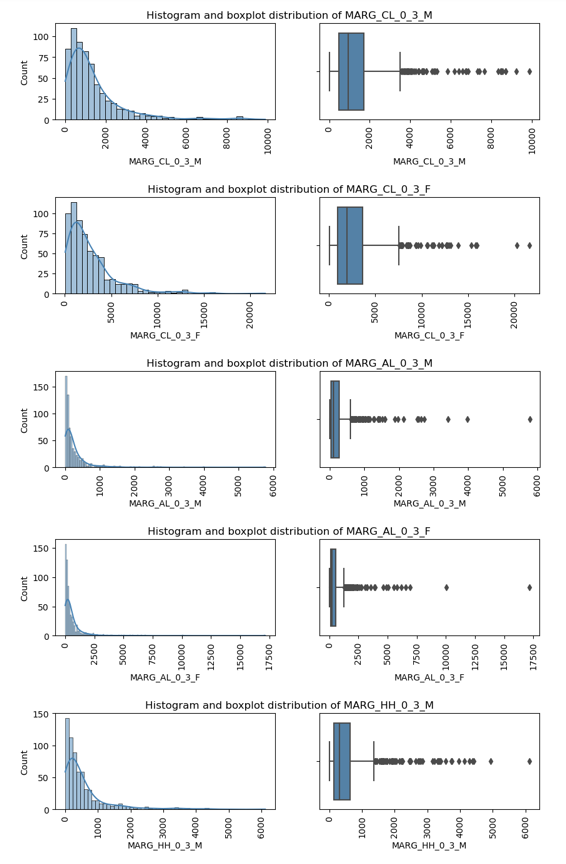


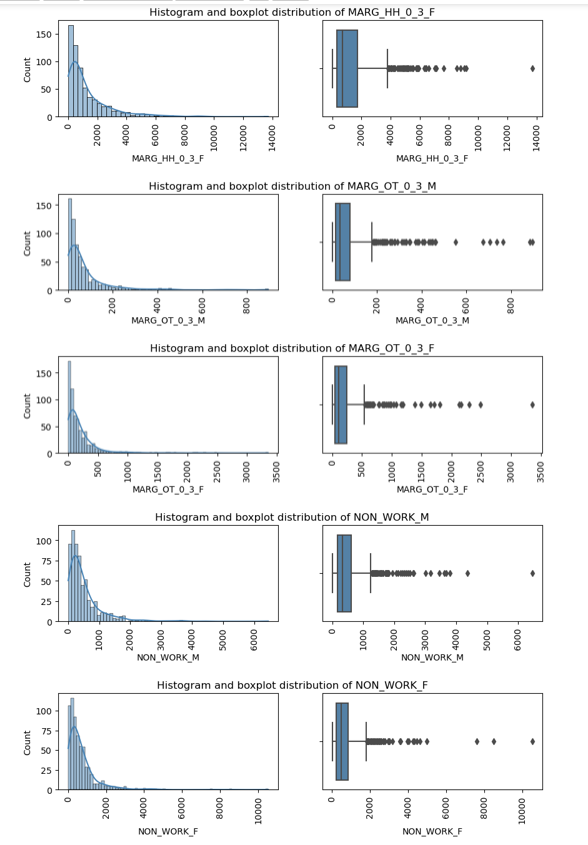




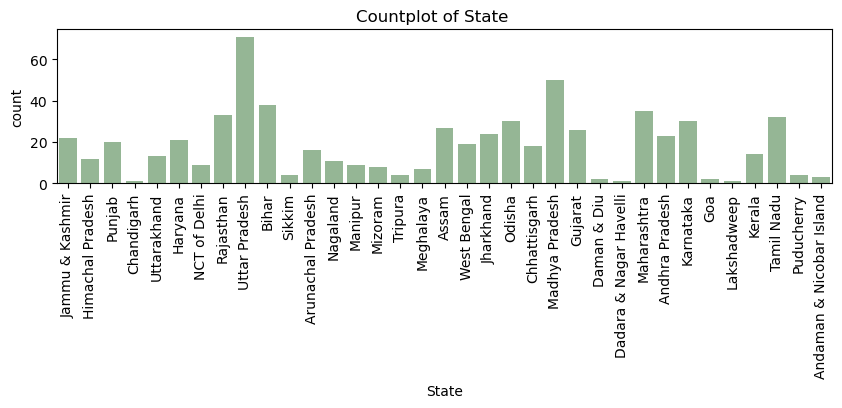








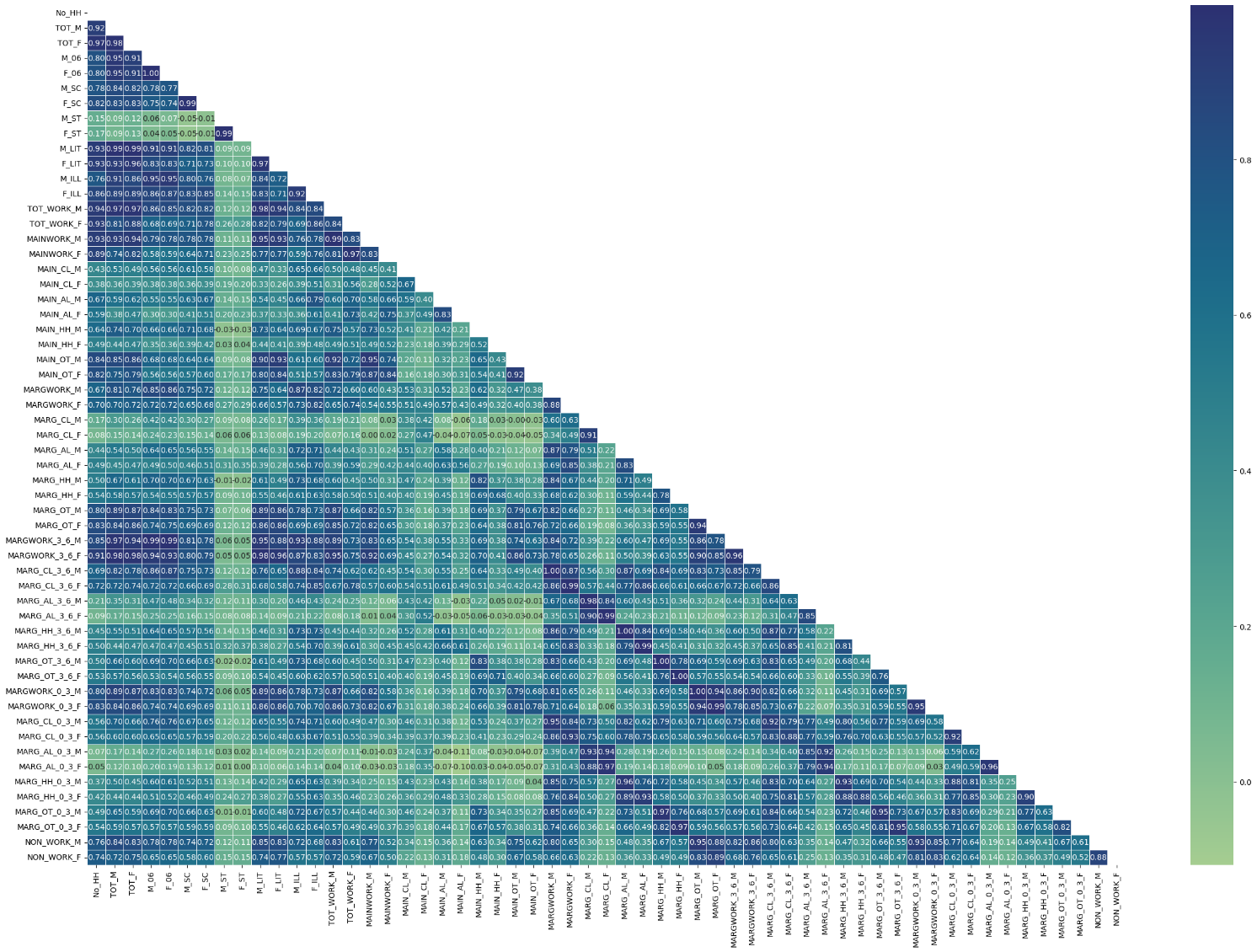
**Fig.2.3 Univariate Analysis- Numerical fields**

****

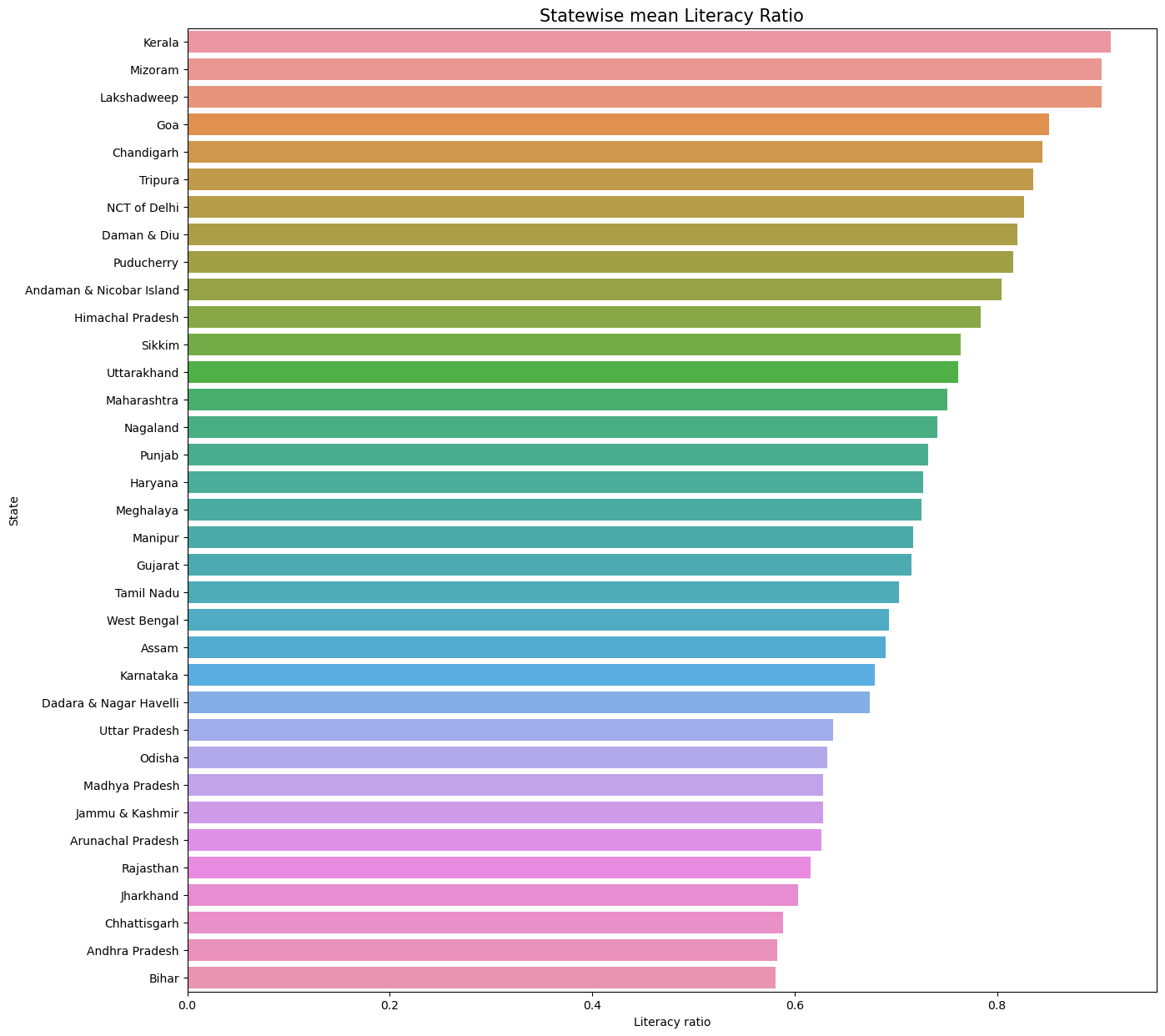
**Fig.2.4. Univariate Analysis- Categorical field**

**Univariate Analysis Observations:**

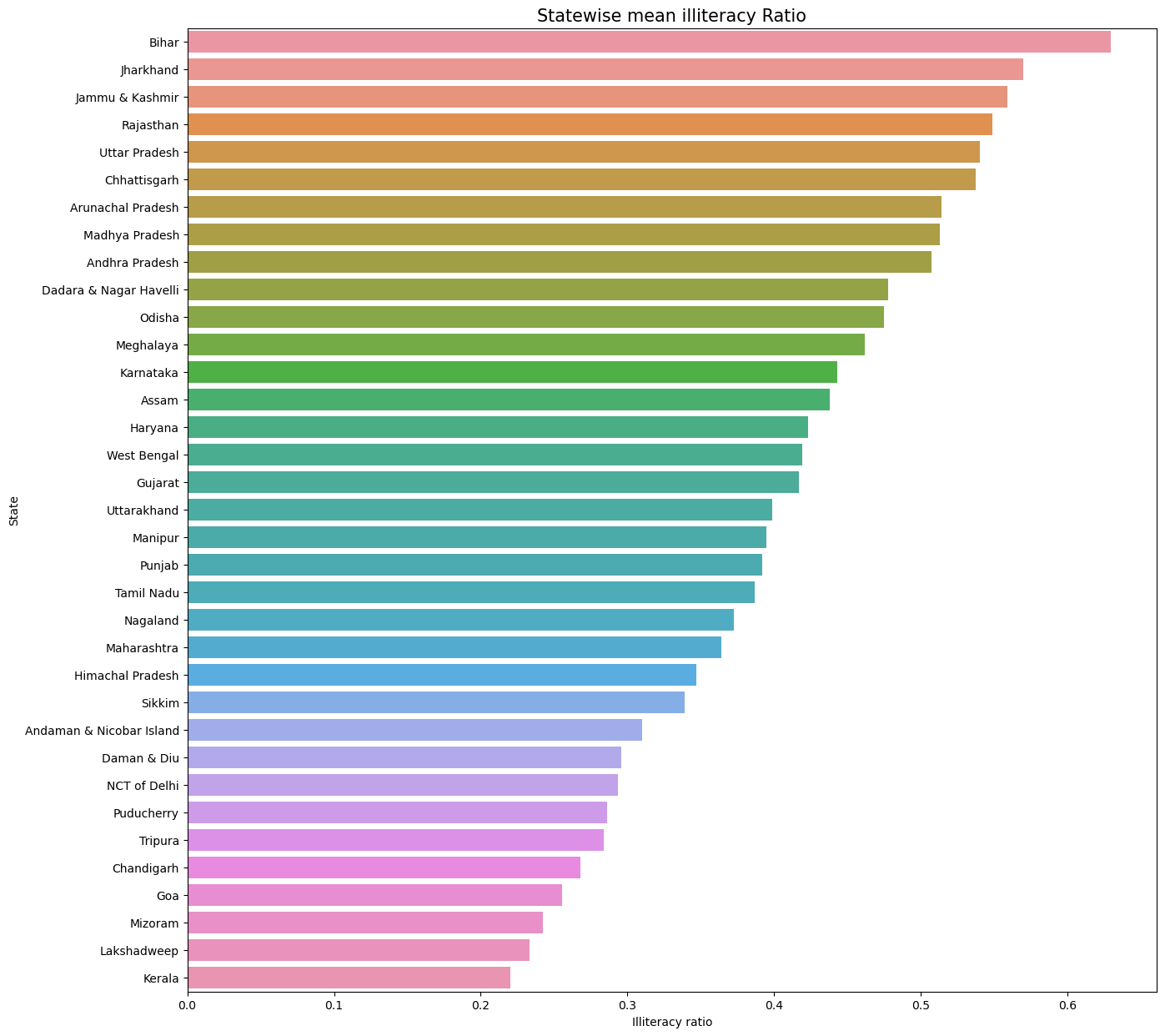
* All the fields have outliers
* Data varies in magnitude across fields
* Uttar Pradesh is the state with highest number of records
* Union Territories have the lowest number of records

**Bivariate Analysis:**

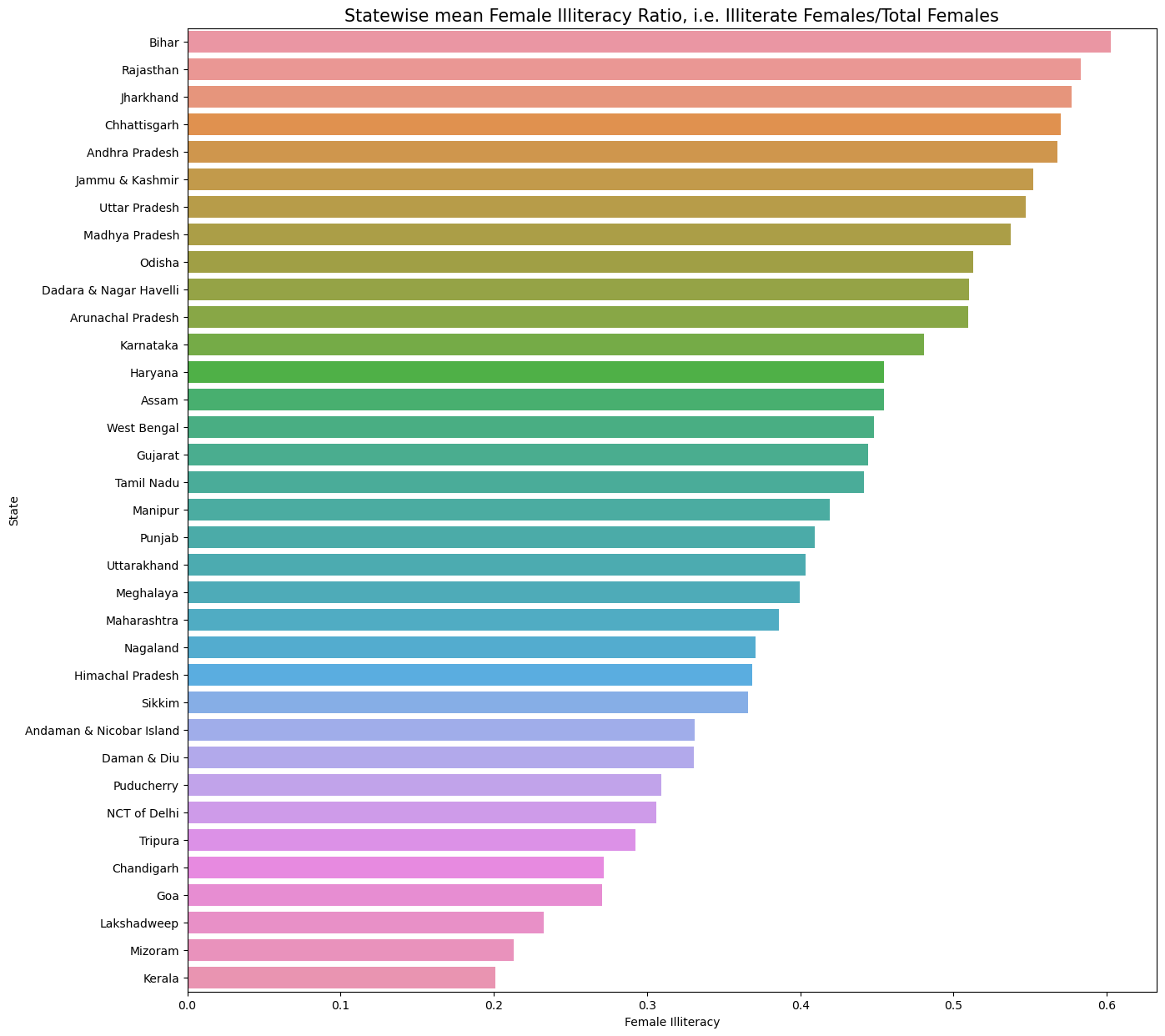
**Fig.2.5. Heatmap of numerical fields**

****

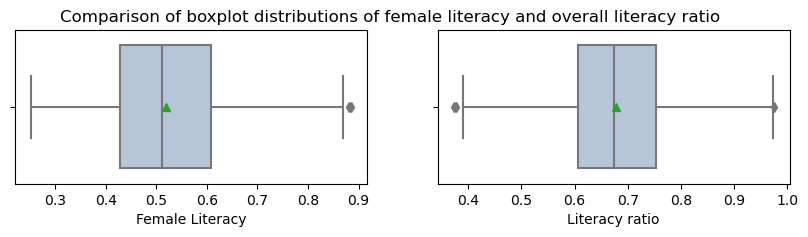
**Fig.2.6. Statewise mean literacy ratio**

****

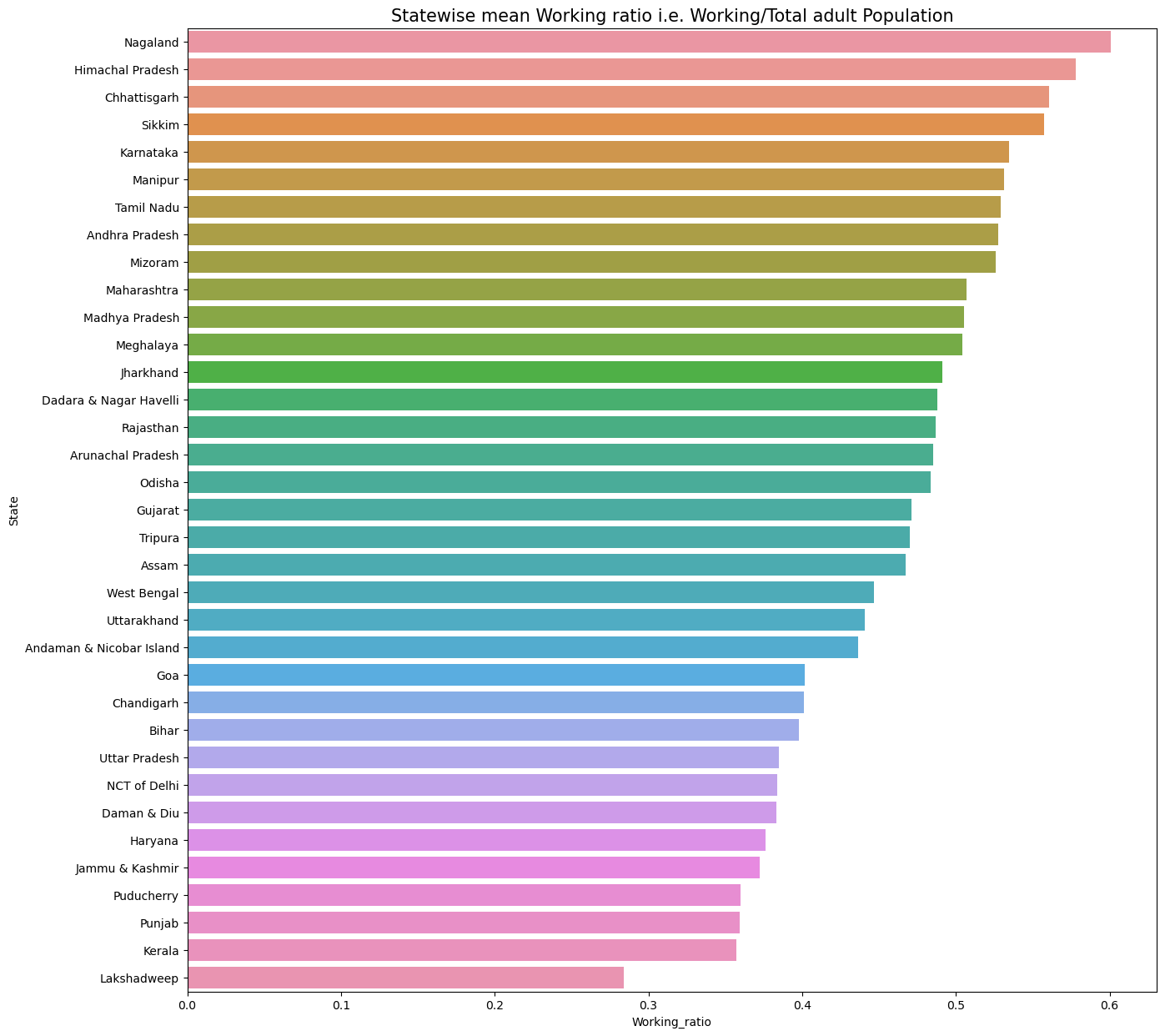
**Fig.2.7. Statewise mean Illiteracy ratio**

****

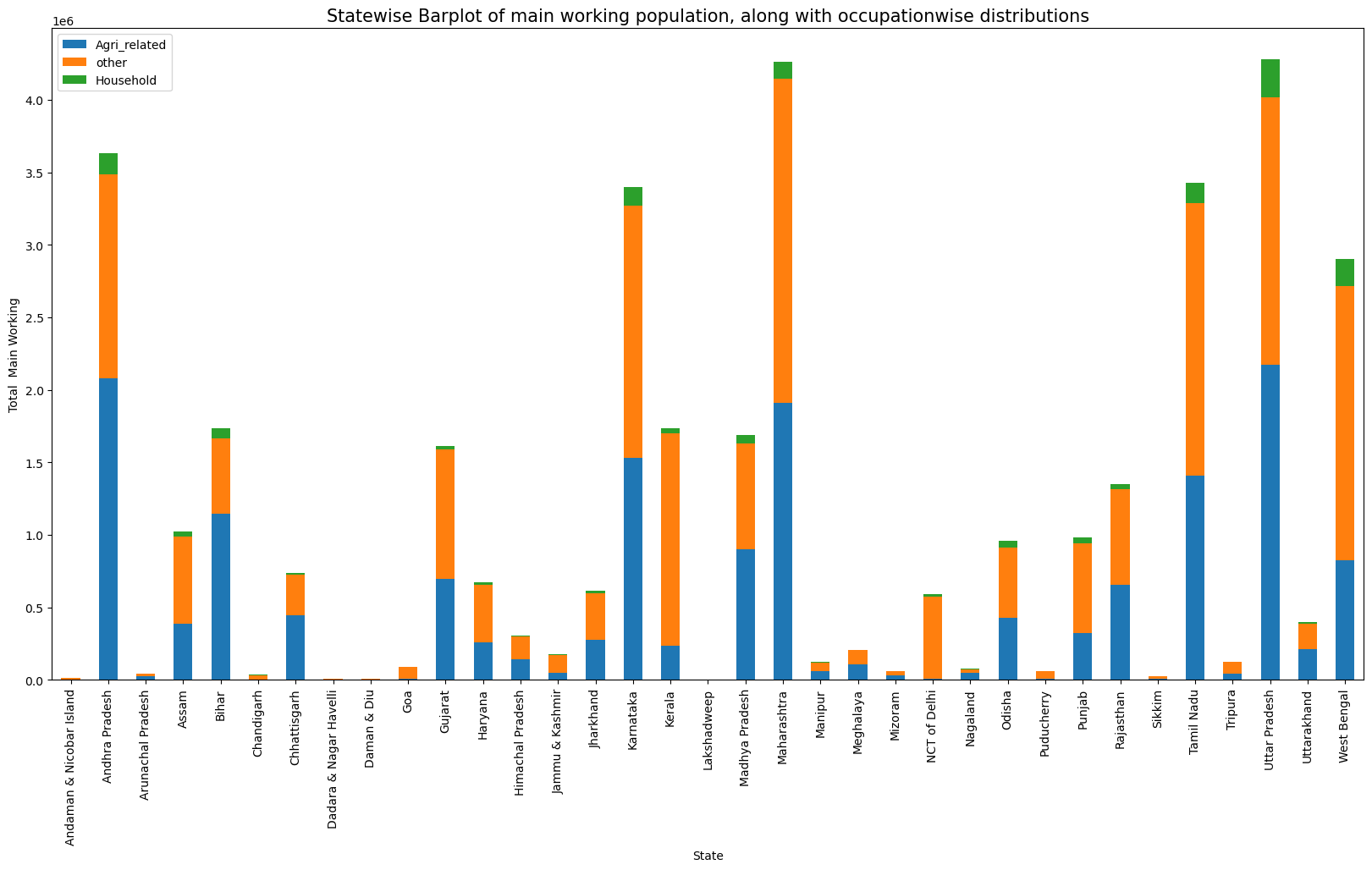
**Fig.2.8. Statewise female illiteracy ratio**

****

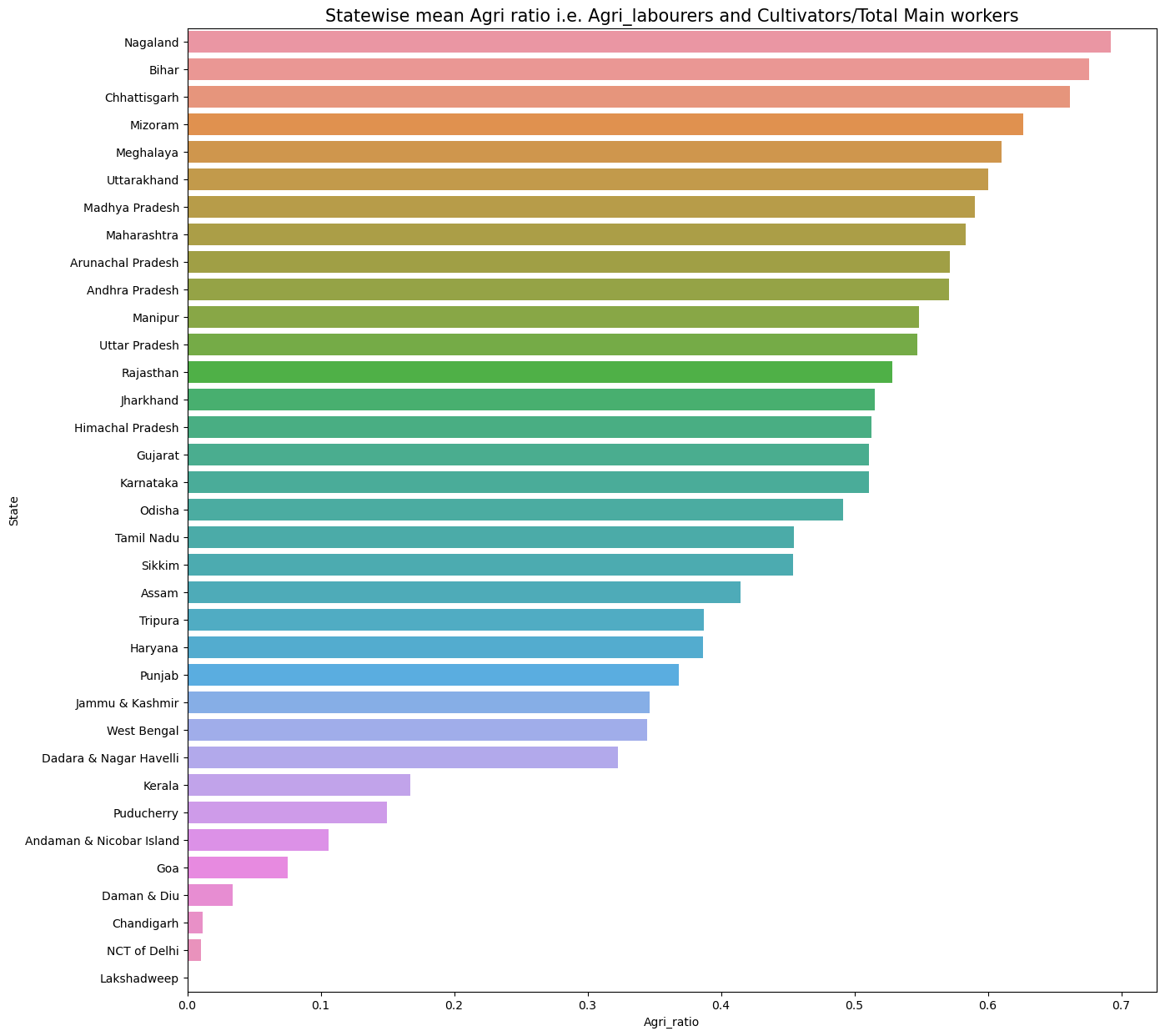
**Fig.2.9. Comaprison of Female literacy and total literacy ratio**

****

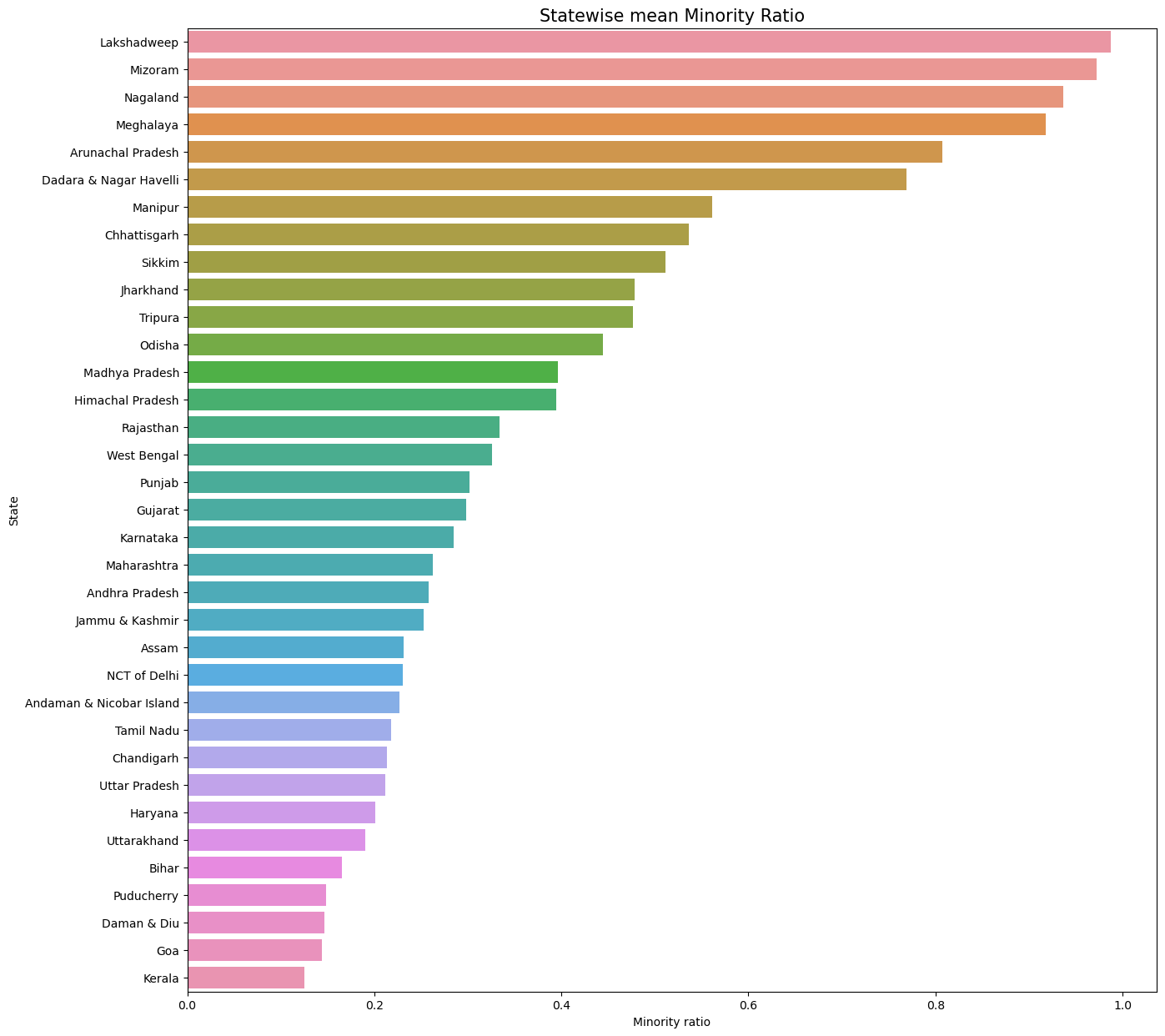
**Fig.2.10. Statewise mean working ratio**

****

**Fig.2.11. Statewise Occupation chart**

****

**2.12. Statewise mean Agri\_ralated workers**

****

**Fig.2.13. Mean minority ratio**

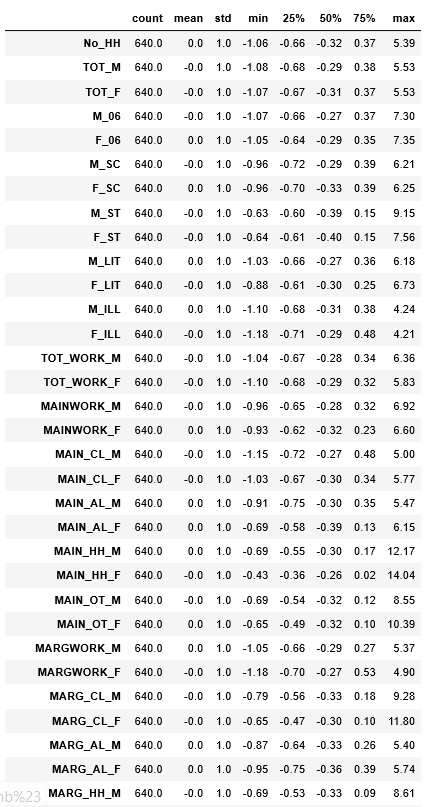
**Observations**

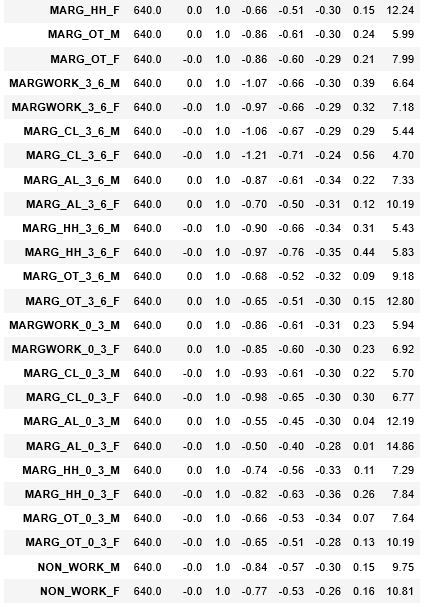
* A lot of variables are highly correlated, as indicated by the dark blue areas of the above heatmap
* Difficult to discern details from heatmap
* For example, margwork\_hh\_m and marg\_ot\_3-6m, marg\_ot\_m, margword\_0-3\_m etc
* Kerala has the highest and Bihar has the lowest literacy ratio
* Bihar 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

1. **Part 2 - PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?**

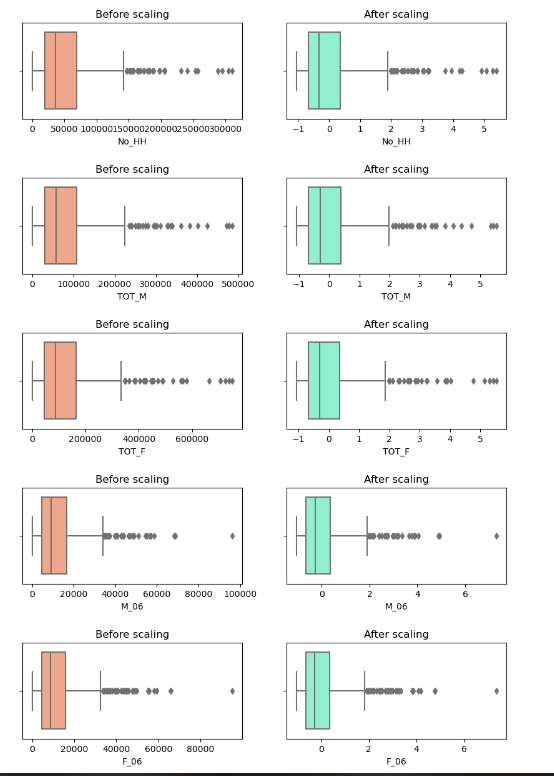
Generally,outliers have an impact on the results of the PCA, and hence need to be tackled. However, the outliers themselves may contain valuable data, that may provide useful insights. In a sensitive dataset like this census, eventhough PCA needs to be done, outliers may be left untreated in order to prevent loss of essential information.

1. **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.**

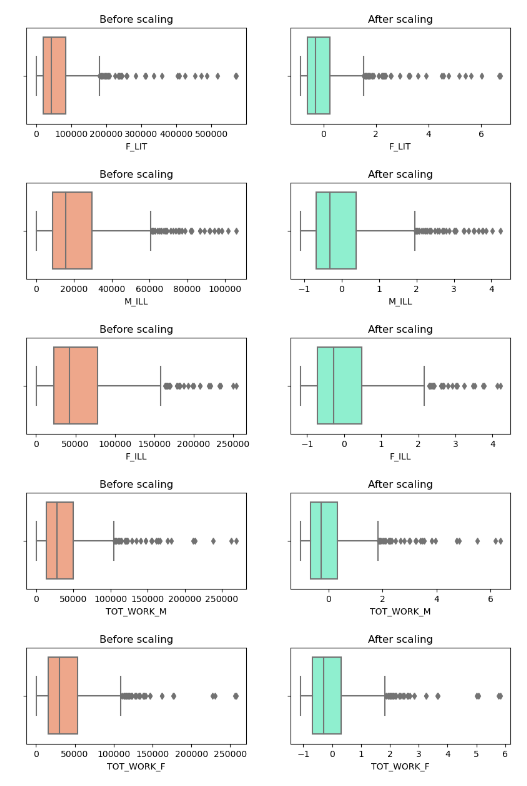


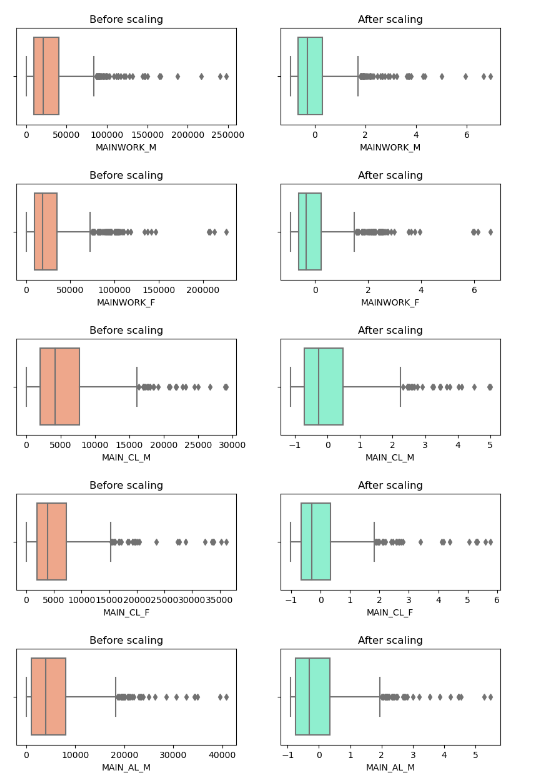


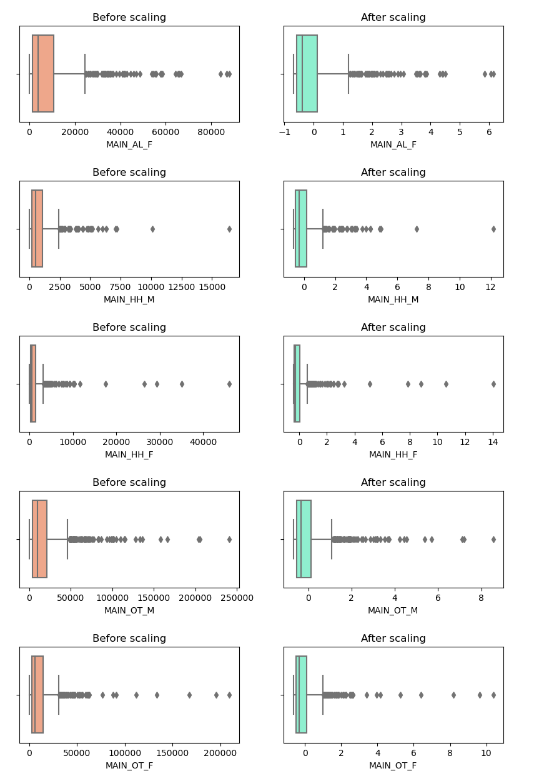
**Fig.2.14. Data description after scaling**

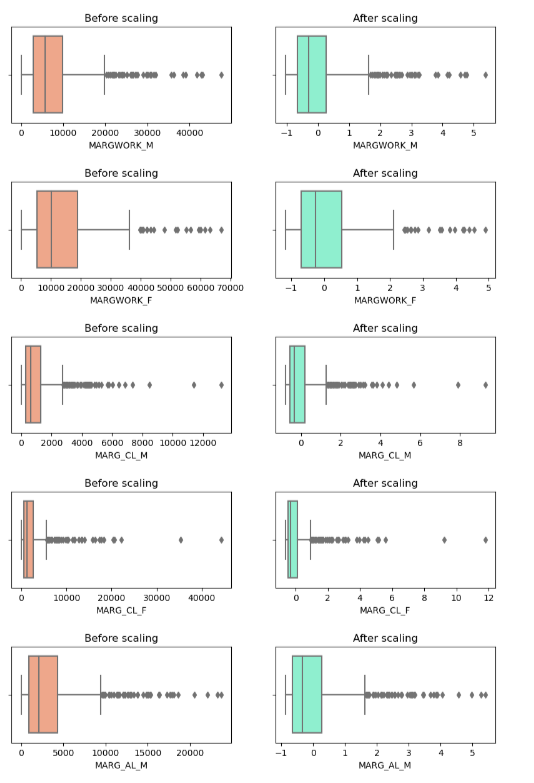


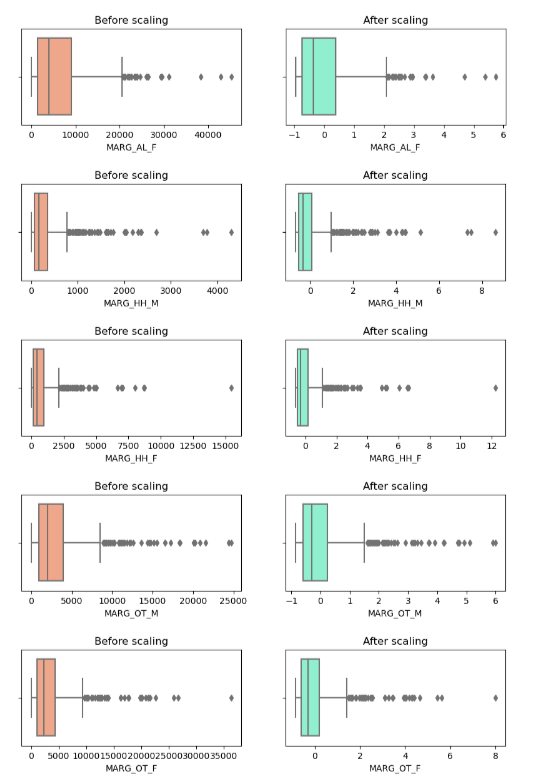


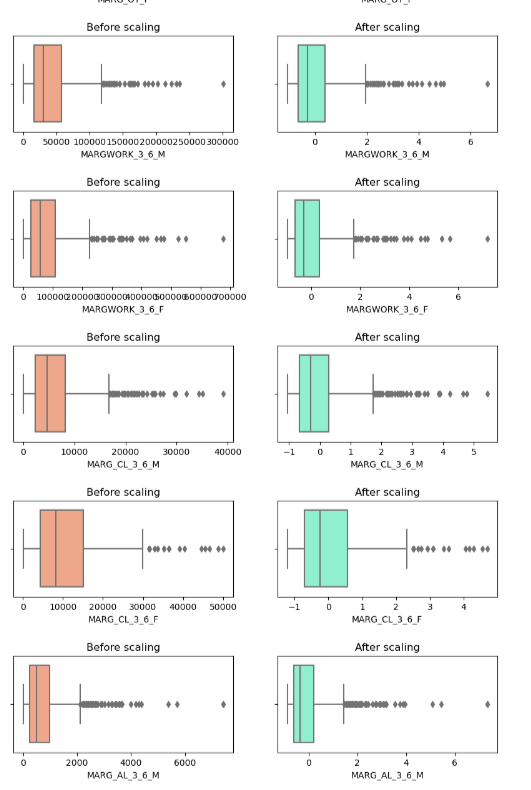


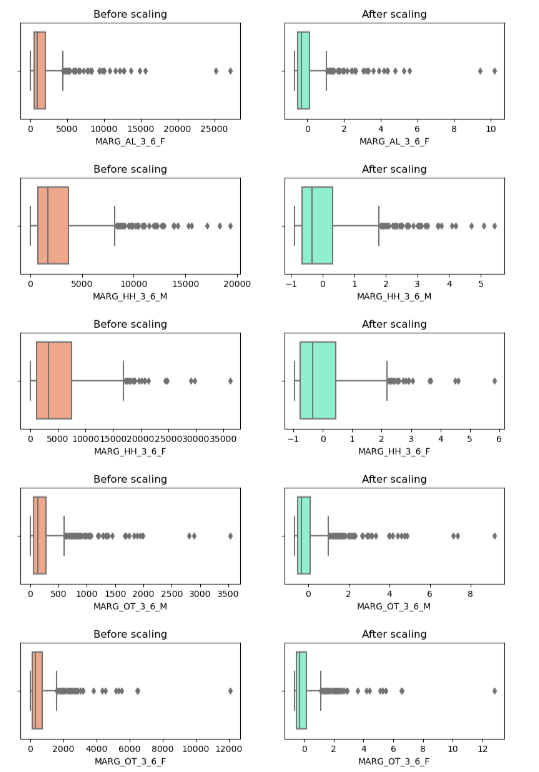


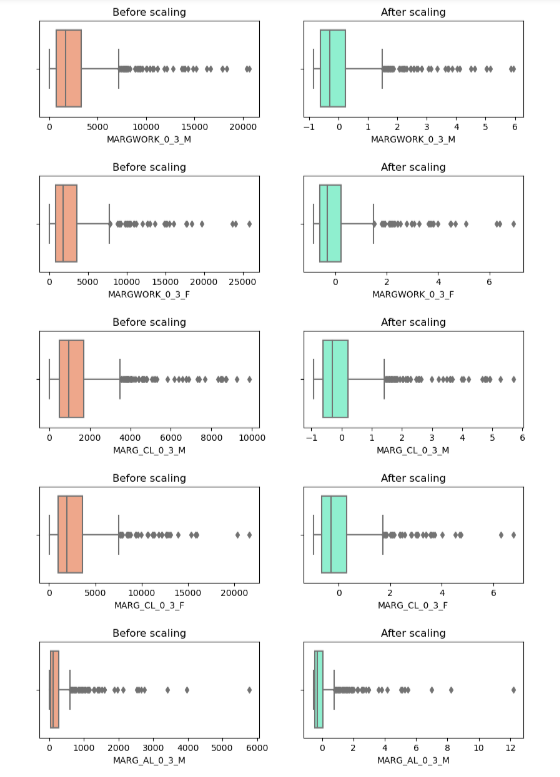


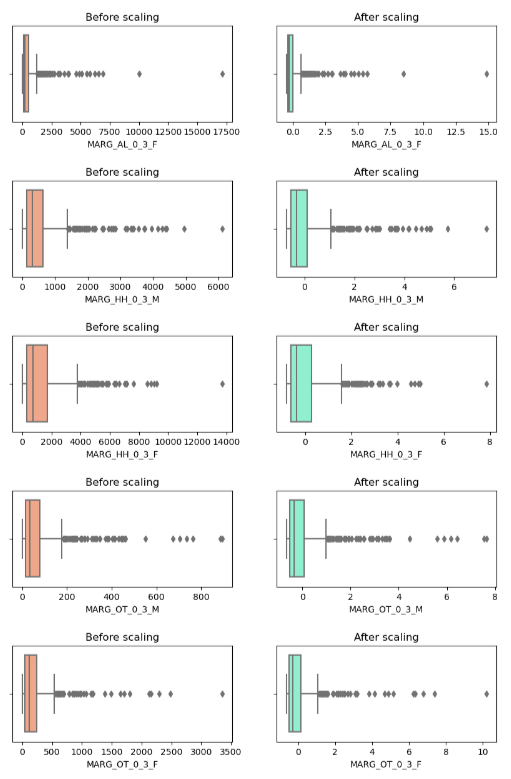


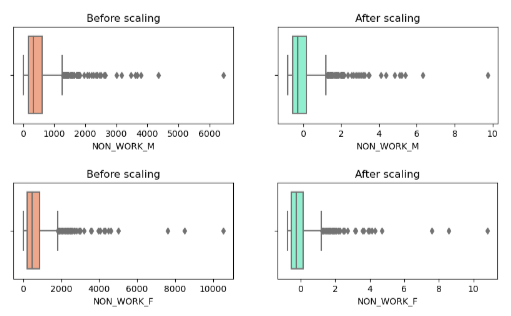








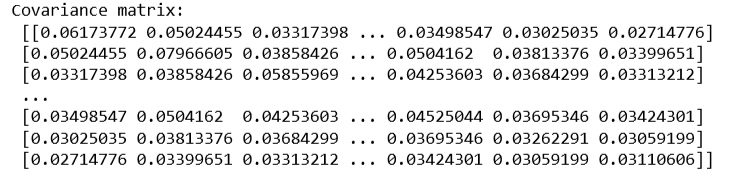
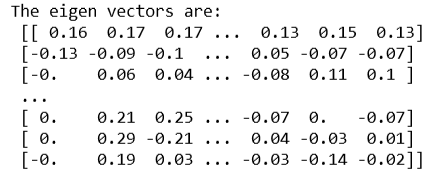




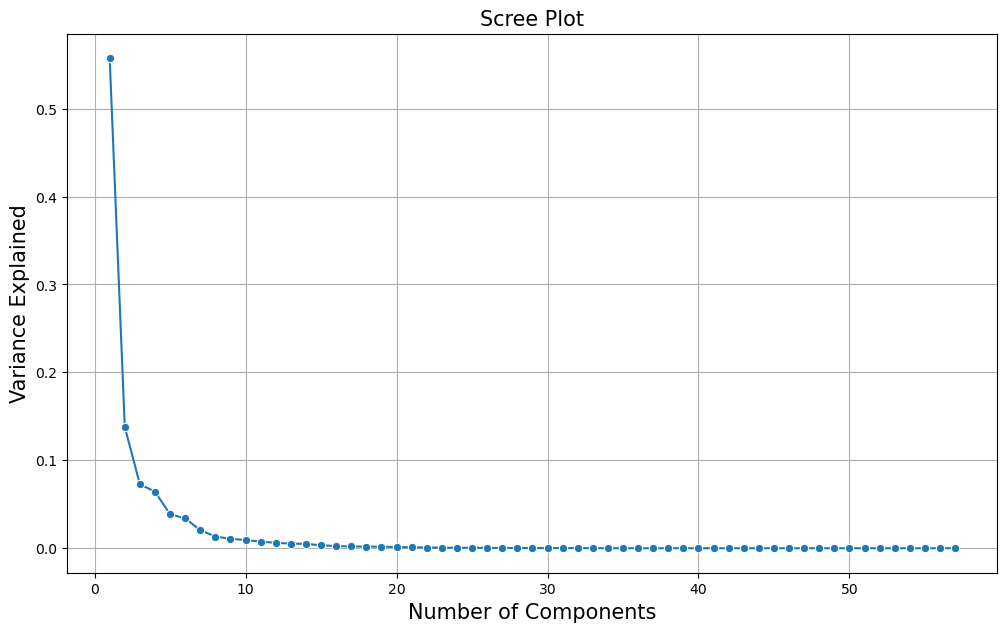
**Fig.2.15. Boxplot distributions before and after scaling**

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

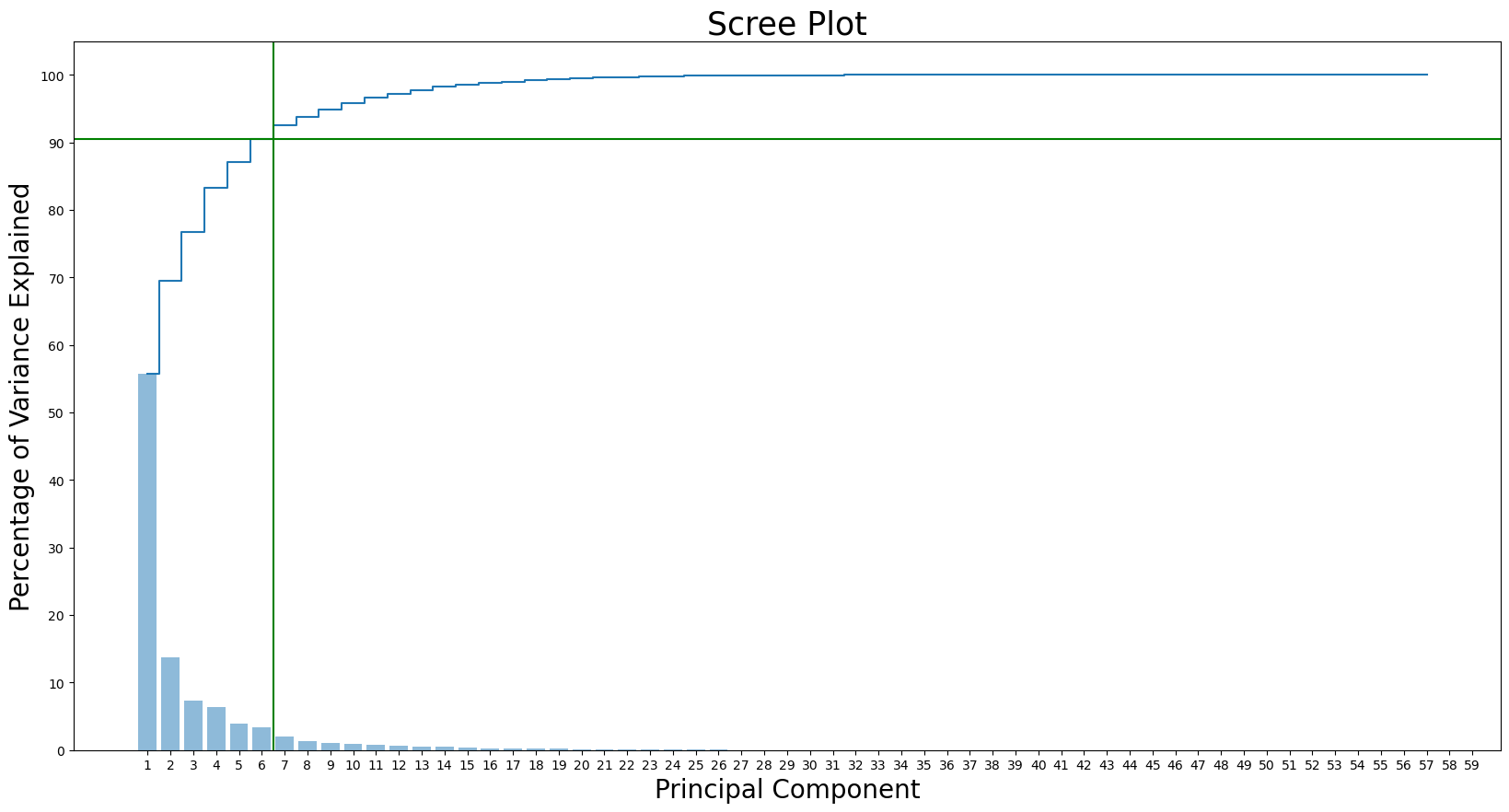
1. **Part 2 - PCA: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.**

* **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:
  + 
* Eigen-vectors:
  + 
* Eigen-values:
  + 

1. **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**

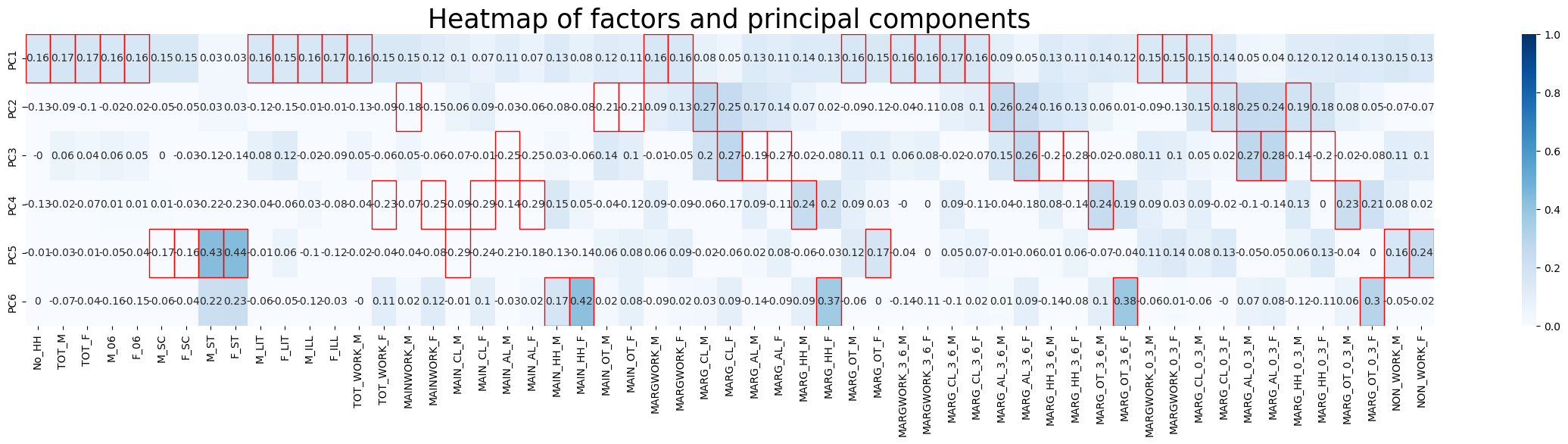
****

**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.

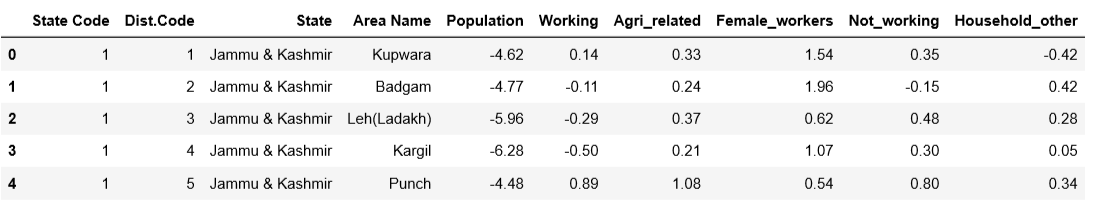
1. **Part 2 - PCA: Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.**

****

**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



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

****

**Fig.2.20. Correlation heatmap after PCA**

1. **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