MACHINE LEARNING 1 – CODED PROJECT

- DONE BY

R.SUKANYA

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Criteria	Points
Part 1: Clustering: Define the problem and perform Exploratory Data Analysis - Problem definition - Check shape, Data types, statistical summary - Univariate analysis - Bivariate analysis - Key meaningful observations on individual variables and the relationship between variables	6.5
Part 1: Clustering: Data Preprocessing - Missing value check and treatment - Outlier Treatment - z-score scaling Note: Treat missing values in CPC, CTR and CPM using the formula given.	2.5
Part 1: Clustering: Hierarchical Clustering - Construct a dendrogram using Ward linkage and Euclidean distance - Identify the optimum number of Clusters	4
Part 1: Clustering: K-means Clustering - Apply K-means Clustering - Plot the Elbow curve - Check Silhouette Scores - Figure out the appropriate number of clusters - Cluster Profiling	13
Part 1: Clustering: Actionable Insights & Recommendations - Extract meaningful insights (atleast 3) from the clusters to identify the most effective types of ads, target audiences, or marketing strategies that can be inferred from each segment. - Based on the clustering analysis and key insights, provide actionable recommendations (atleast 3) to Ads24x7 on how to optimize their digital marketing efforts, allocate budgets efficiently, and tailor ad content to specific audience segments.	6
Part 2: PCA: Define the problem and perform Exploratory Data Analysis - Problem Definition - Check shape, Data types, statistical summary - Perform an EDA on the data to extract useful insights Note: 1. 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 2. Example questions to answer from EDA - (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio?	6.5

Criteria	Points
Part 2: PCA: Data Preprocessing - Check for and treat (if needed) missing values - Check for and treat (if needed) data irregularities - Scale the Data using the z-score method - Visualize the data before and after scaling and comment on the impact on outliers	2.5
Part 2; PCA: PCA - Create the covariance matrix - Get eigen values and eigen vectors - Identify the optimum number of PCs - Show Scree plot - Compare PCs with Actual Columns and identify which is explaining most variance - Write inferences about all the PCs in terms of actual variables - Write linear equation for first PC Note: For the scope of this project, take at least 90% explained variance.	13
Quality of Business Report	6
Points	60

Part 1 - Problem Statement1:

Clustering:

Digital Ads Data:

The ads24x7 is a Digital Marketing company which has now got seed funding of $10 Million. They are expanding their wings in Marketing Analytics. They collected data from their Marketing Intelligence team and now wants you (their newly appointed data analyst) to segment type of ads based on the features provided. Use Clustering procedure to segment ads into homogeneous groups.

The following three features are commonly used in digital marketing:

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

CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column in the dataset and the Number of Clicks refers to the 'Clicks' Column in the dataset.

CTR = Total Measured Clicks / Total Measured Ad Impressions x 100. Note that the Total Measured Clicks refers to the 'Clicks' Column in the dataset and the Total Measured Ad Impressions refers to the 'Impressions' Column in the dataset.

Part 1: Clustering: Define the problem and perform Exploratory Data Analysis

1.1 Importing Libraries

Import the standard libraries of pandas, numpy, matplotlib and seaborn. Also, import clustering algorithm and silhouette score.

1.2 Loading the Dataset and Reading the Dataset

The dataset Clustering_Clean_Ads data is read and loaded into the Jupyter notebook. The dataset is read using the head() command.

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

Table 1 – Reading the dataset

1.3 Checking shape

Use the shape() function to get the number of rows and columns in the dataset. There are 23066 rows and 19 columns in the dataset.

1.4 Checking datatype

There are a total of 19 columns, out of which, 13 columns are numerical datatype while 6 columns are object datatype.

1.5 Exploratory Data Analysis

1.5.1 Statistical Summary

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	3.851631e+02	2.336514e+02	120.0000	120.000000	300.00000	7.200000e+02	728.00
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.0000	250.000000	300.00000	6.000000e+02	600.00
Ad Size	23066.0	9.667447e+04	6.153833e+04	33600.0000	72000.000000	72000.00000	8.400000e+04	216000.00
Available_Impressions	23066.0	2.432044e+06	4.742888e+06	1.0000	33672.250000	483771.00000	2.527712e+06	27592861.00
Matched_Queries	23066.0	1.295099e+06	2.512970e+06	1.0000	18282.500000	258087.50000	1.180700e+06	14702025.00
Impressions	23066.0	1.241520e+06	2.429400e+06	1.0000	7990.500000	225290.00000	1.112428e+06	14194774.00
Clicks	23066.0	1.067852e+04	1.735341e+04	1.0000	710.000000	4425.00000	1.279375e+04	143049.00
Spend	23066.0	2.706626e+03	4.067927e+03	0.0000	85.180000	1425.12500	3.121400e+03	26931.87
Fee	23066.0	3.351231e-01	3.196322e-02	0.2100	0.330000	0.35000	3.500000e-01	0.35
Revenue	23066.0	1.924252e+03	3.105238e+03	0.0000	55.365375	926.33500	2.091338e+03	21276.18
CTR	18330.0	7.366054e-02	7.515992e-02	0.0001	0.002600	0.08255	1.300000e-01	1.00
CPM	18330.0	7.672045e+00	6.481391e+00	0.0000	1.710000	7.66000	1.251000e+01	81.56
CPC	18330.0	3.510606e-01	3.433338e-01	0.0000	0.090000	0.16000	5.700000e-01	7.26

Table 2 – Statistical Summary

Observations:

- The average value and the median value of Available_Impressions is the highest in the dataset.
- Available impressions has maximum skewness followed by Matched queries and impressions.
- The available impressions has a very high range from 1 to 27592861.
- The matched queries, impressions and clicks have a very high range.
- The average ad size is 61538.33.
- The average value of fee payable is very less.
- The average revenue per ad is 1924.252. The minimum revenue is 0 while the maximum is 21276.
- The clicks range from 1 to 143049.
- The standard deviation is very highest for available impressions, matched queries and impressions.
- The variability is very less for CTR, CPM and CPC as well as fee.

1.5.2 Univariate Analysis:

a. Numerical Data

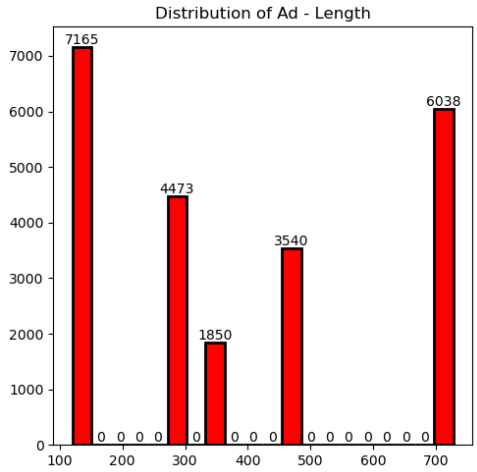
For univariate analysis, we make two datasets, one containing only the numerical fields while the other containing only the categorical fields with object datatype.

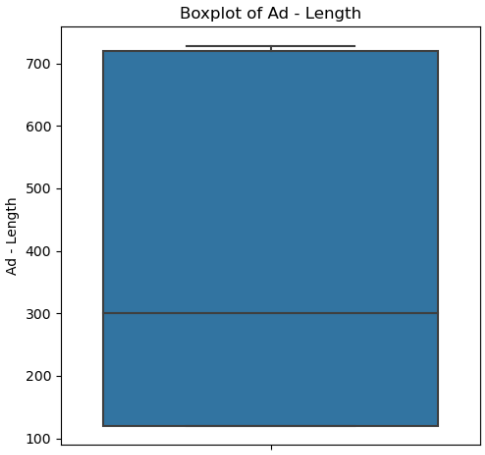
Numerical datatype dataset

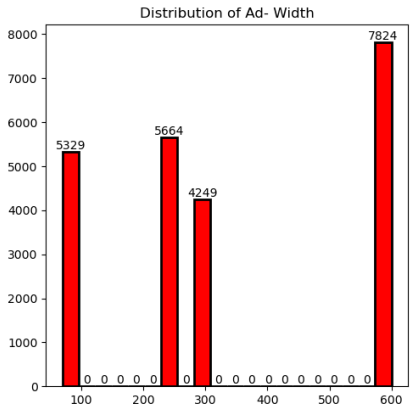
	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	300	250	75000	1806	325	323	1	0.0	0.35	0.0	0.0031	0.0	0.0
1	300	250	75000	1780	285	285	1	0.0	0.35	0.0	0.0035	0.0	0.0
2	300	250	75000	2727	356	355	1	0.0	0.35	0.0	0.0028	0.0	0.0
3	300	250	75000	2430	497	495	1	0.0	0.35	0.0	0.0020	0.0	0.0
4	300	250	75000	1218	242	242	1	0.0	0.35	0.0	0.0041	0.0	0.0

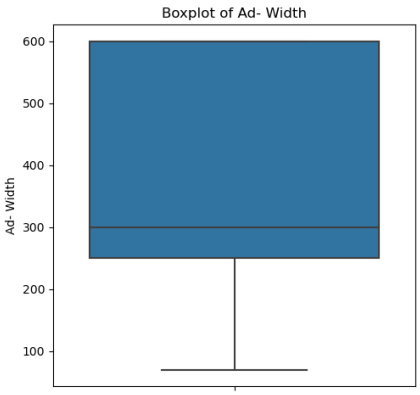
Table 3 – Numerical Variables

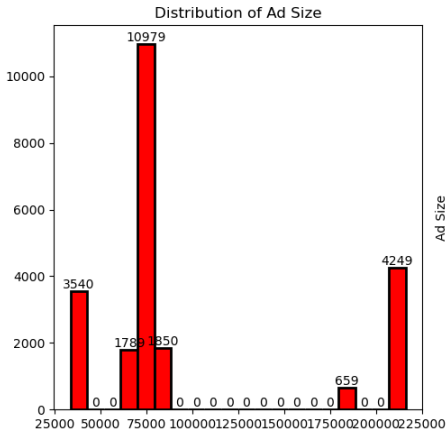
Histogram and Boxplot for all the numerical fields: A histogram and boxplot is plotted for all the variables in the dataset to understand the presence of skewness, distribution of data and presence of outliers.

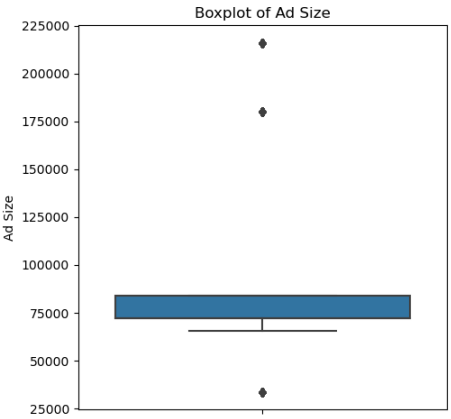

Distribution of Ad - Length
7165
4473
1850
3540
6038

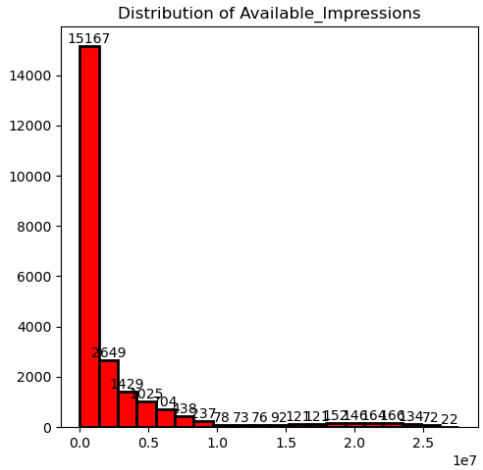

Boxplot of Ad - Length
Ad - Length

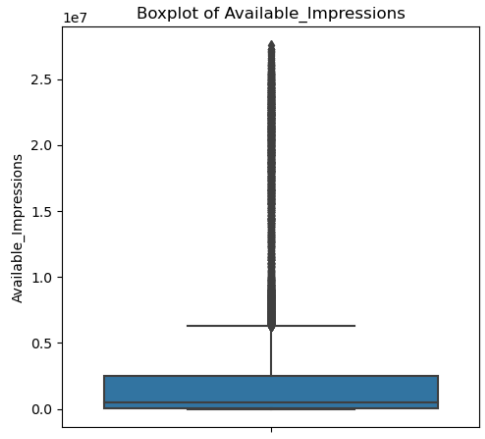

Distribution of Ad- Width
5329
5664
4249
7824

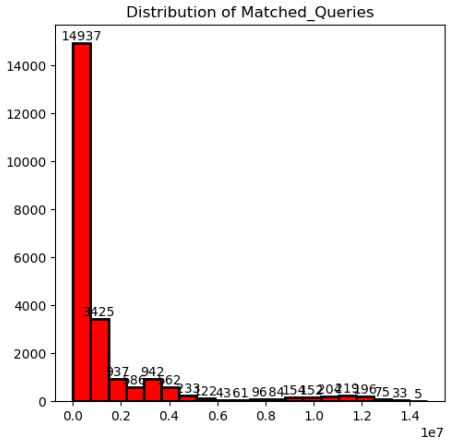

Boxplot of Ad- Width
Ad- Width

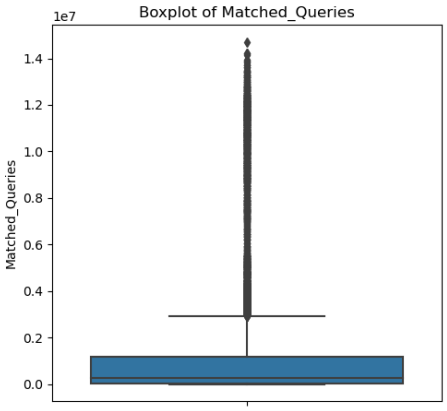

Distribution of Ad Size
10979
3540
1789
1850
659
4249

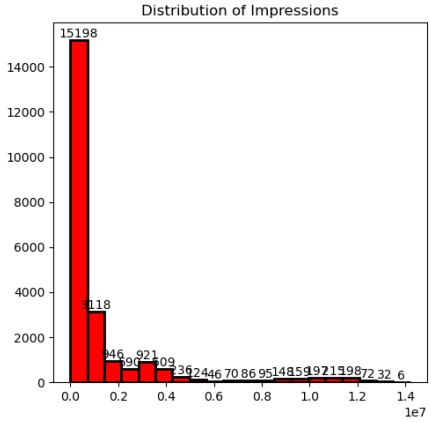

Boxplot of Ad Size
Ad Size

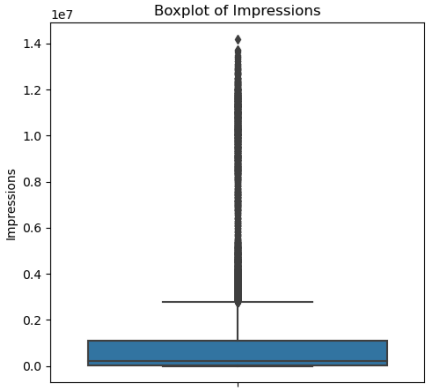

Distribution of Available_Impressions
15167
2649
1429
1025
704
338
237
1e7

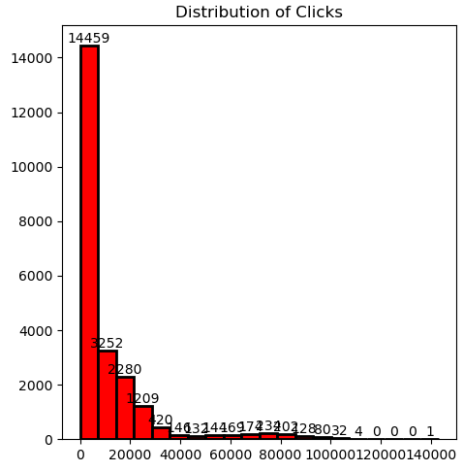

Boxplot of Available_Impressions
1e7
Available_Impressions

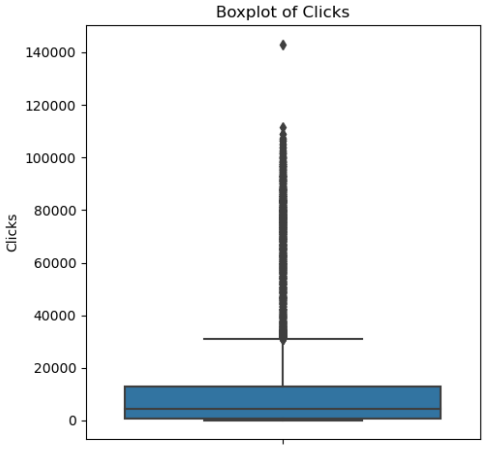

Distribution of Matched_Queries
14937
3425
1e7

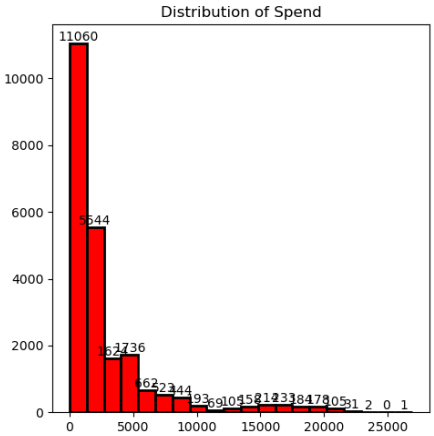

Boxplot of Matched_Queries
1e7
Matched_Queries

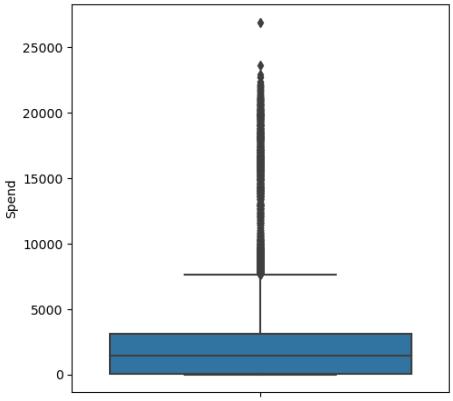

Distribution of Impressions
15198
3118
1e7

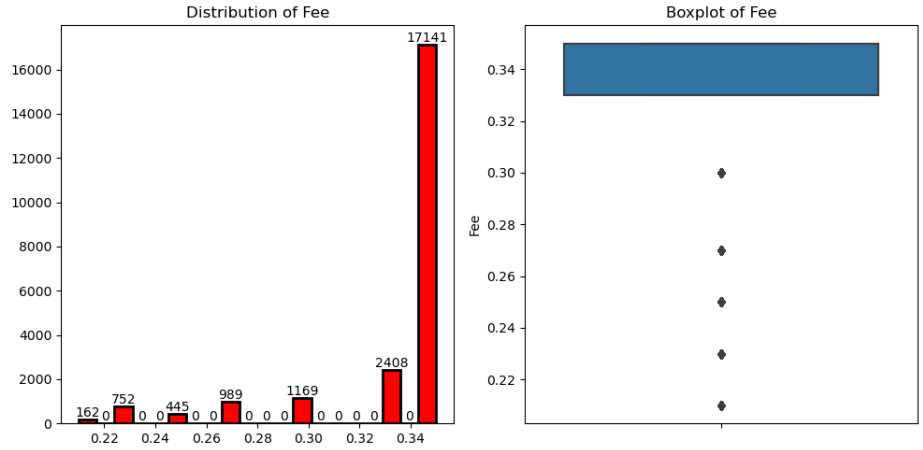

Boxplot of Impressions
1e7
Impressions

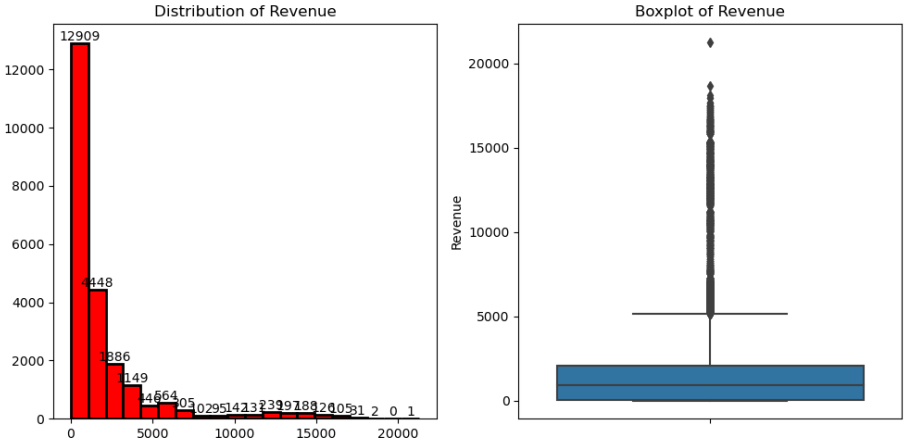

Distribution of Clicks
14459
3252
2280
1209

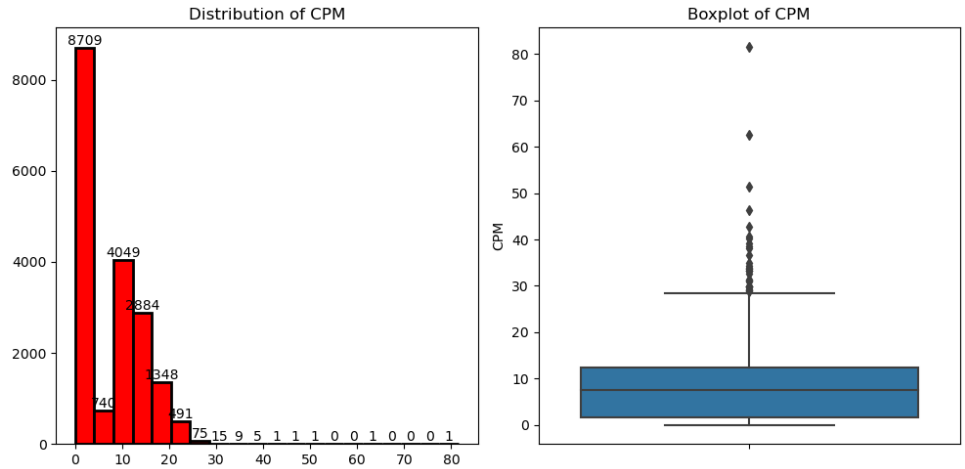

Boxplot of Clicks
Clicks

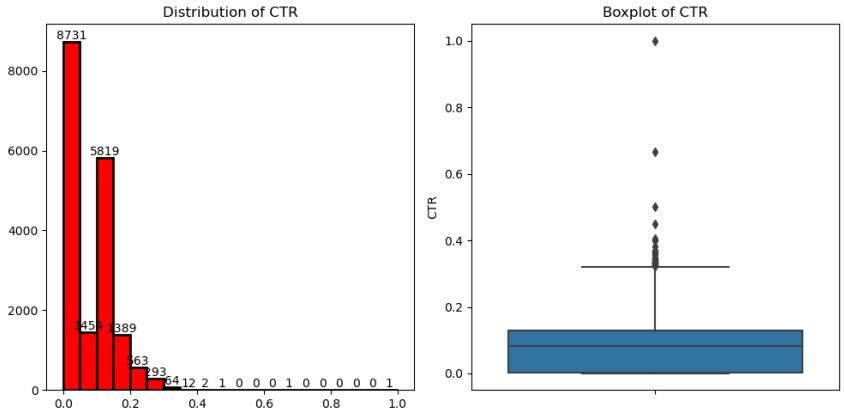

Distribution of Spend
11060
5544
1736


Boxplot of Spend
Spend


Distribution of Fee
17141
2408
1169
989
752
445
162
Boxplot of Fee
Fee


Distribution of Revenue
12909
4448
1386
1149
Boxplot of Revenue
Revenue


Distribution of CPM
8709
4049
2884
1348
740
491
Boxplot of CPM
CPM


Distribution of CTR
8731
5819
1451
1389
563
293
Boxplot of CTR
CTR

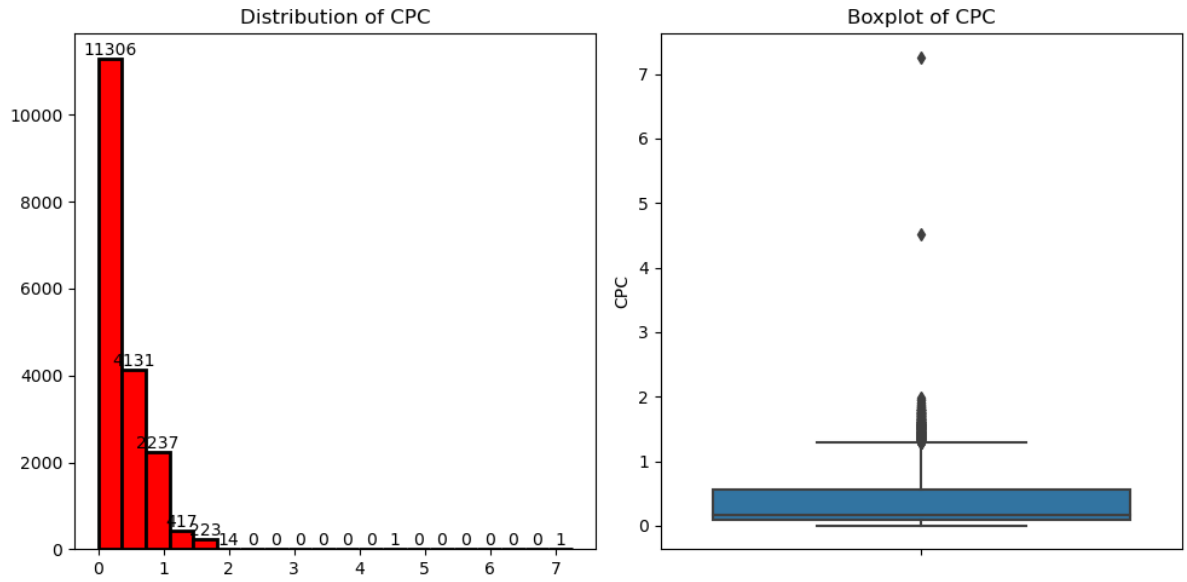


Figure 1 – Numerical Variables

Observations:

* There are 13 numerical fields in the dataset.

* Ad-length and Ad-width do not seem to have any outliers, but their product which makes the Ad-size has outliers.

* Available impressions seem to be highly right skewed with mean greater than the median.

* There are a huge number of outliers in Available impressions as well as matched queries fields.

* The impressions and clicks fields are also right-skewed with large number of outliers.

* Spend and Revenue fields are also right-skewed with many outliers.

* Fee field is the only field which is left-skewed.

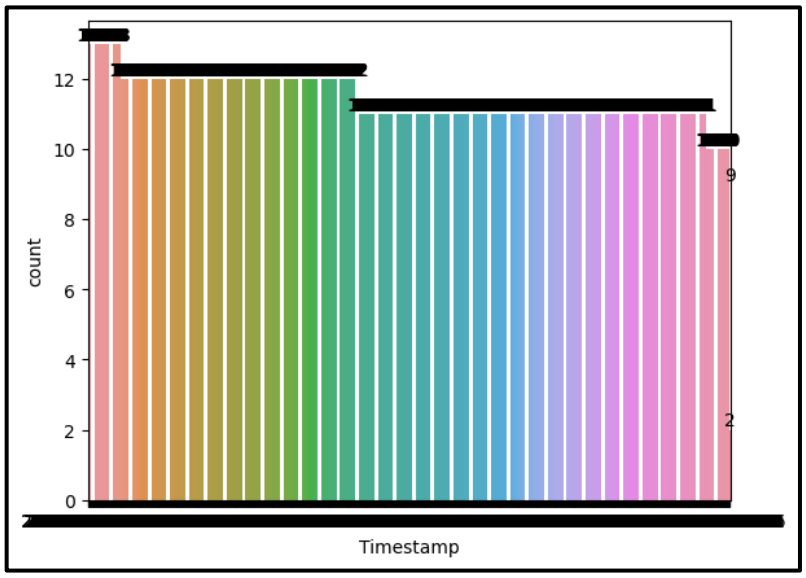
* CTR, CPM and CPC are the three fields which exhibit right-skewness with many outliers.

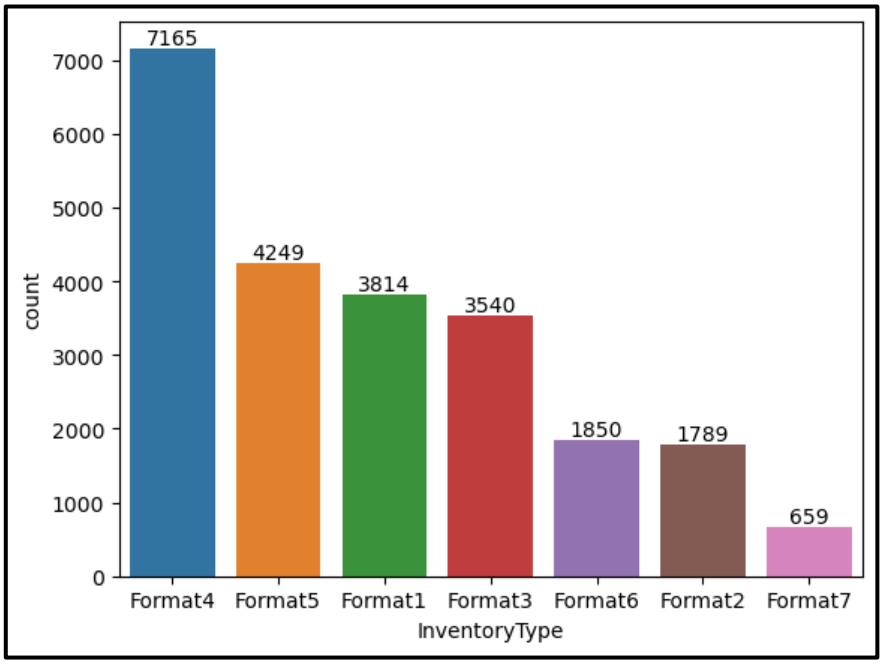
* The outliers need to be treated.

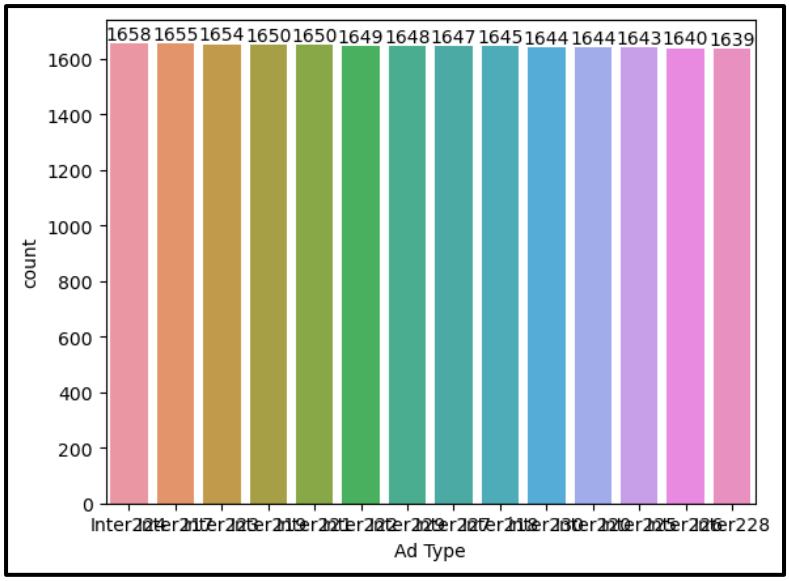
b) Categorical data: For the 6 categorical variables, histogram is plotted to understand the skewness and distribution of the data.

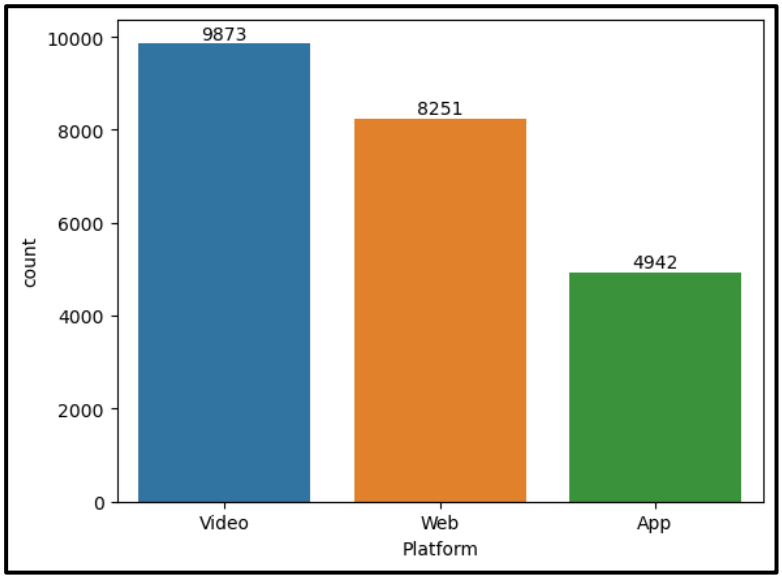
	Timestamp	InventoryType	Ad Type	Platform	Device Type	Format
0	2020-9-2-17	Format1	Inter222	Video	Desktop	Display
1	2020-9-2-10	Format1	Inter227	App	Mobile	Video
2	2020-9-1-22	Format1	Inter222	Video	Desktop	Display
3	2020-9-3-20	Format1	Inter228	Video	Mobile	Video
4	2020-9-4-15	Format1	Inter217	Web	Desktop	Video

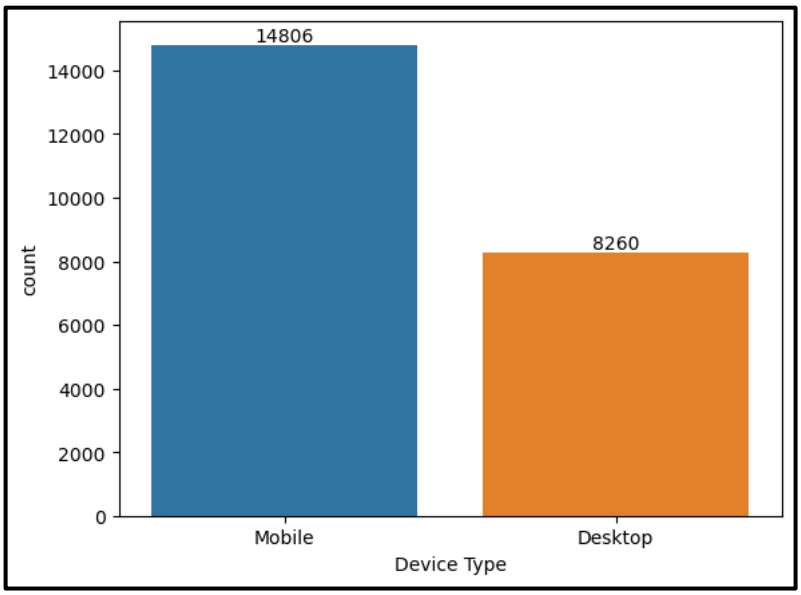
Table 4 – Categorical Variables






1658
1655
1654
1650
1650
1649
1648
1647
1645
1644
1644
1643
1640
1639
count
Ad Type


9873
8251
4942
count
Video
Web
App
Platform


14806
8260
count
Mobile
Desktop
Device Type

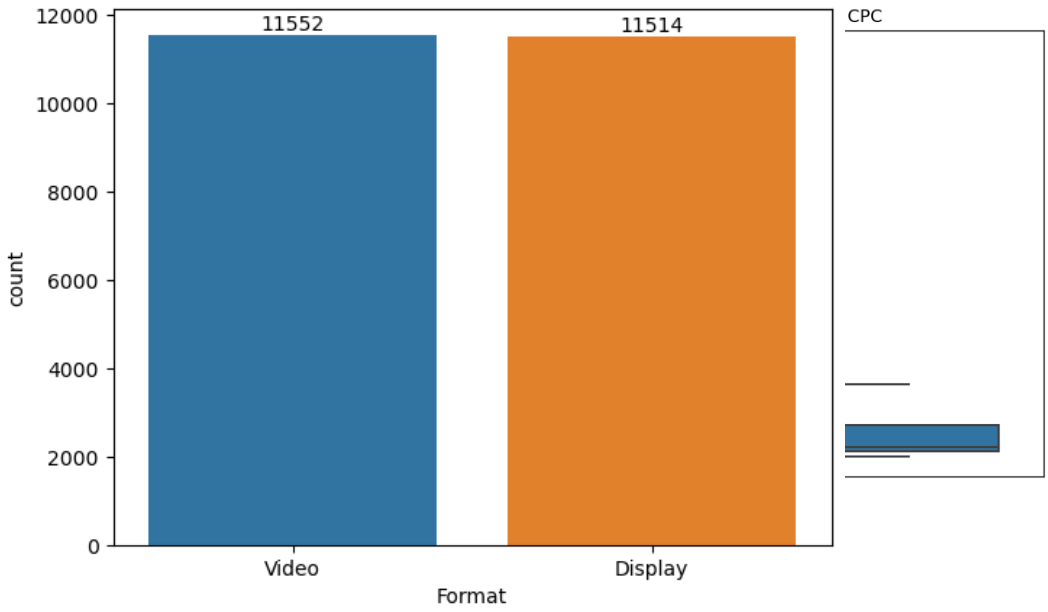


Figure 2 – Categorical Variables

Observations:

* Format 4 has maximum count of inventory type.

* Ad types are almost the same for all types.

* Video type has a much bigger count (9873) as compared to Web(8251) and App(4942).

* Mobile type is used much more than desktop device type.

* Video and Display formats are almost the same count.

1.5.3 Bivariate Analysis – Correlation between variables:

Use the pairplot and the heatmap to find the correlation between the numerical variables. If the correlation is positive and close to 1, it means that there is a high correlation between the corresponding variables. If the correlation is highly negative, that means that if one variable increases, the other decreases.

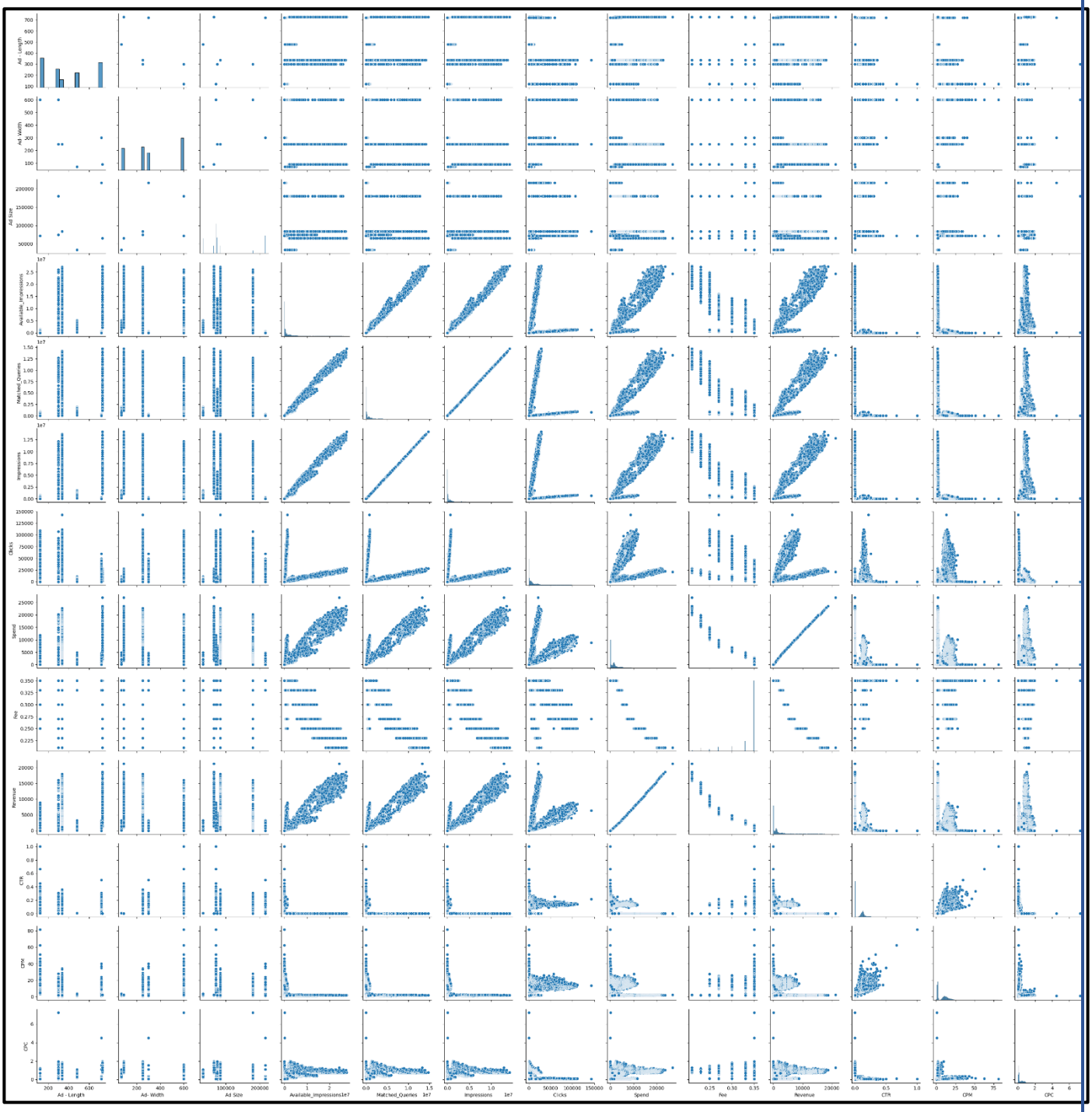


Figure 3 – Pairplot of all the numerical variables

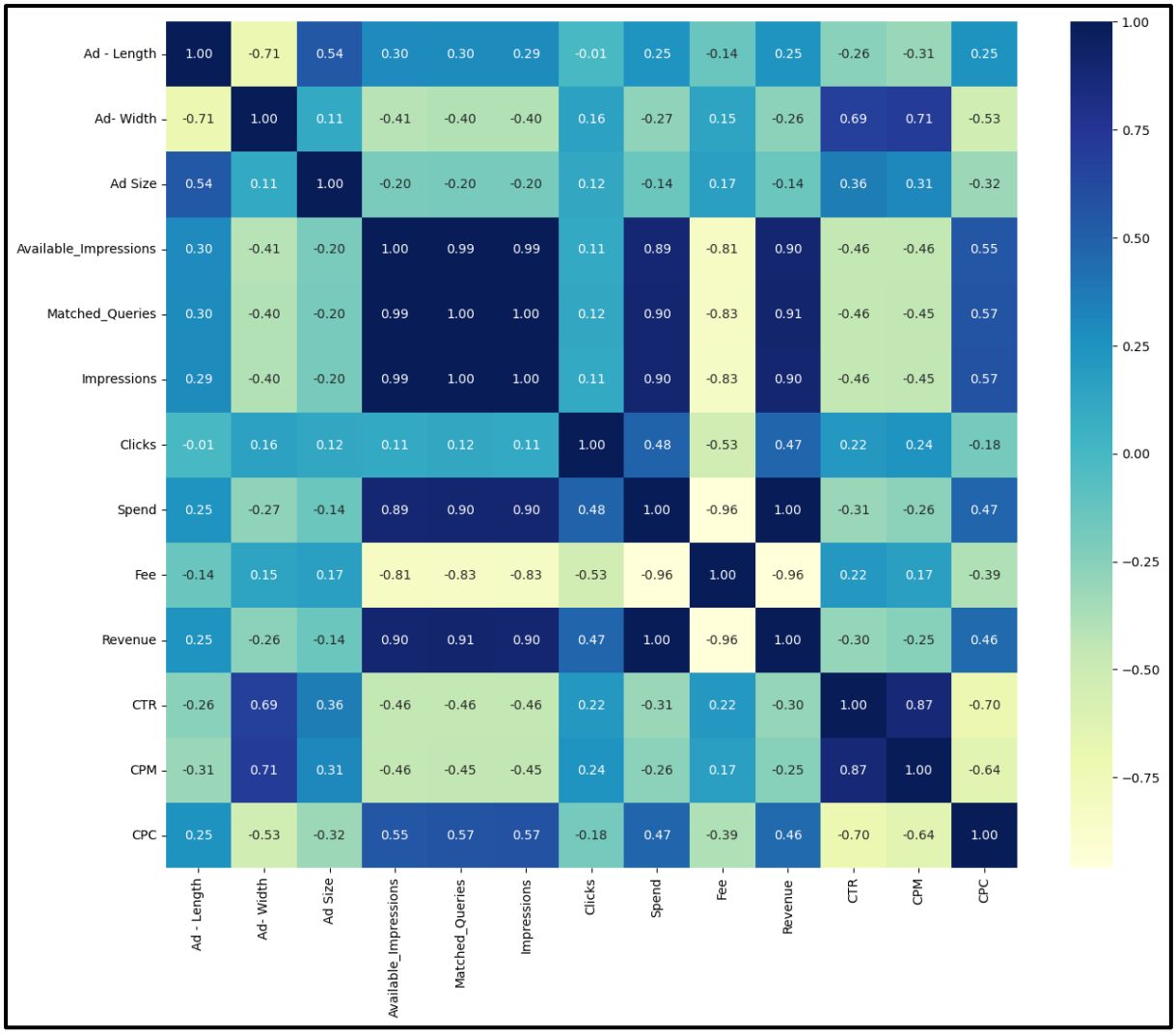


Figure 4 – Correlation between variables in heatmap

Observations:

* There is a very high correlation between:
  Available impressions and Matched queries
  Impressions and Matched queries
  Impressions and Available impressions
  Revenue and Spend.

* There is good correlation between:
  Matched Queries and Spend
  Impressions and Spend
  Matched queries and Revenue
  Impressions and Revenue
  Available Impressions and Revenue
  Available Impressions and Spend

* There is negative correlation between:
  CPC and CTR
  Fee and other fields like Revenue, Spend, Impressions, Matched queries

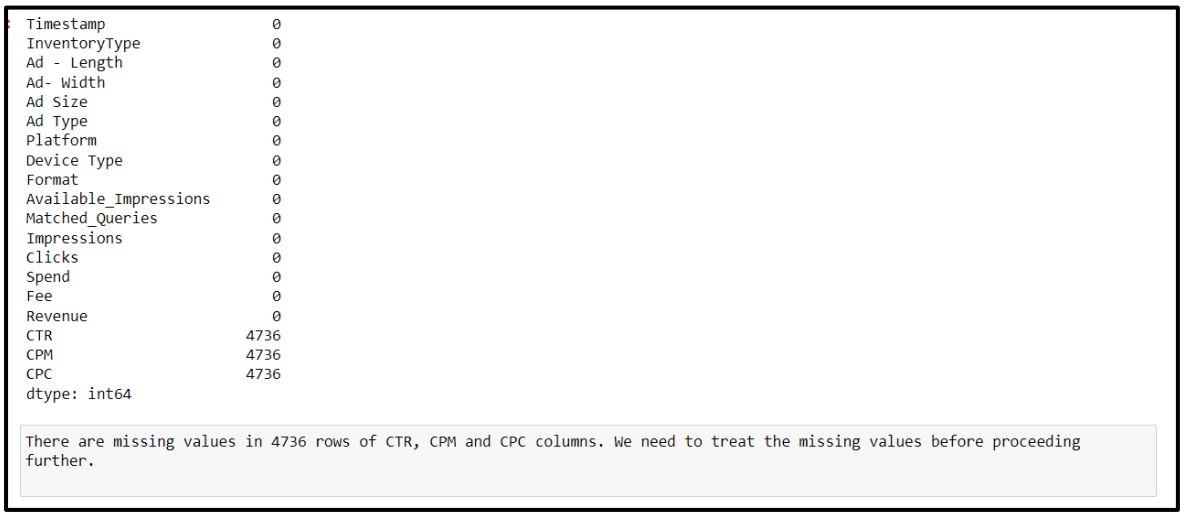
2.1 Part 1: Clustering: Data Preprocessing

2.1.1 Check for duplicate values:

We use the duplicated() function to find if the dataset has any duplicate values. We find that there are no duplicate values and we are good to go.

2.1.2 Check for missing values:

There are missing values in 4736 rows of CTR, CPM and CPC columns. We need to treat the missing values before proceeding further.



```
Timestamp                  0
InventoryType              0
Ad - Length                0
Ad- Width                  0
Ad Size                    0
Ad Type                    0
Platform                   0
Device Type                0
Format                     0
Available_Impressions      0
Matched_Queries            0
Impressions                0
Clicks                     0
Spend                      0
Fee                        0
Revenue                    0
CTR                     4736
CPM                     4736
CPC                     4736
dtype: int64
```

```
There are missing values in 4736 rows of CTR, CPM and CPC columns. We need to treat the missing values before proceeding
further.
```

Table 5 – Missing values

We impute the missing values in the three columns using the fillna() function and the formula given in the data dictionary.

CPM = (Spend / Impressions) * 1000
CPC = Spend/Clicks
CTR = (Clicks/Impressions)*100

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

Table 6 – Treating missing values

2.1.3 Dropping Object type columns:

Drop the object datatype columns which are Timestamp, Inventory Type, Ad Type, Platform, Device Type and Format. We also drop the Ad-Length and Ad-width columns as they are multiplied to get the Ad-size column.

	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	75000	1806	325	323	1	0.0	0.35	0.0	0.0031	0.0	0.0
1	75000	1780	285	285	1	0.0	0.35	0.0	0.0035	0.0	0.0
2	75000	2727	356	355	1	0.0	0.35	0.0	0.0028	0.0	0.0
3	75000	2430	497	495	1	0.0	0.35	0.0	0.0020	0.0	0.0
4	75000	1218	242	242	1	0.0	0.35	0.0	0.0041	0.0	0.0

Table 7 – Dropping columns

2.1.4 Checking for outliers:

Define a function which returns the Upper and Lower limit to detect outliers for each feature.

Call the function with the column names.

Cap & floor the values beyond the outlier boundaries.

Observations:

We see that all outliers have been removed and treated. We can proceed with the scaling of data.

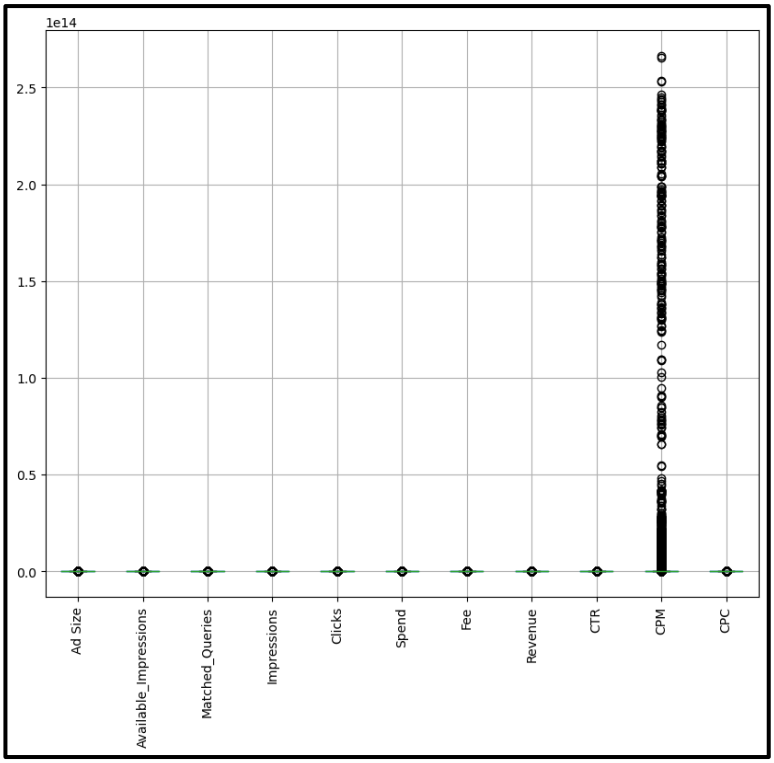


Figure 5 – Outliers in the dataset

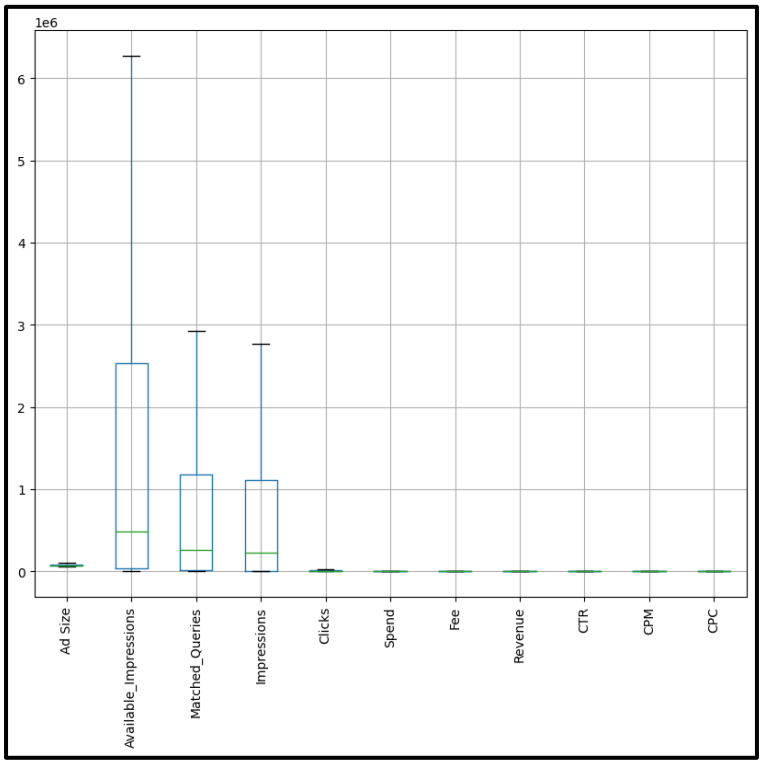


Figure 6 – Treating Outliers

2.1.5 Scaling and Standardizing the data:

First, import the z-score library from scipy.stats. Apply z_score on the dataset to get a scaled dataset.

	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	-0.102518	-0.755333	-0.778949	-0.768478	-0.867488	-0.893170	0.535724	-0.880093	-0.891201	-0.958326	-0.924681
1	-0.102518	-0.755345	-0.778988	-0.768516	-0.867488	-0.893170	0.535724	-0.880093	-0.888615	-0.958326	-0.924681
2	-0.102518	-0.754900	-0.778919	-0.768445	-0.867488	-0.893170	0.535724	-0.880093	-0.893142	-0.958326	-0.924681
3	-0.102518	-0.755040	-0.778781	-0.768302	-0.867488	-0.893170	0.535724	-0.880093	-0.898315	-0.958326	-0.924681
4	-0.102518	-0.755610	-0.779030	-0.768560	-0.867488	-0.893170	0.535724	-0.880093	-0.884734	-0.958326	-0.924681
...	...	...	...	...	...	...	...	...	...	...	...
23061	1.652896	-0.756182	-0.779265	-0.768806	-0.867488	-0.893141	0.535724	-0.880066	2.027108	1.838610	-0.822706
23062	1.652896	-0.756181	-0.779264	-0.768805	-0.867488	-0.893154	0.535724	-0.880078	2.027108	1.838610	-0.866409
23063	1.652896	-0.756182	-0.779265	-0.768806	-0.867488	-0.893150	0.535724	-0.880074	2.027108	1.838610	-0.851841
23064	-0.297564	-0.756179	-0.779265	-0.768806	-0.867488	-0.893141	0.535724	-0.880066	2.027108	1.838610	-0.822706
23065	1.652896	-0.756182	-0.779264	-0.768805	-0.867488	-0.893133	0.535724	-0.880058	2.027108	1.838610	-0.793570

Table 8 – Scaled Dataset

Inferences:

Scaling is necessary because it is a distance algorithm. In this problem, each variable has a different range. One variable might dominate another variable because it has a larger range than the other. To avoid this, we need to preprocess and scale your data so that each variable has a comparable impact on the clustering outcome.

3. Part1: Clustering: Hierarchical Clustering

3.1 Constructing a dendrogram using Ward linkage and Euclidean distance: Use the scaled data and apply the dendrogram function() on it.

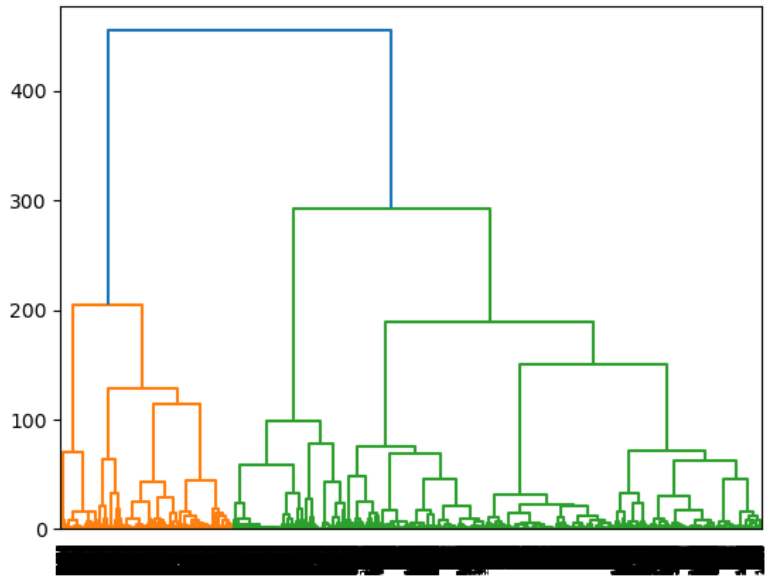


Figure 7 - Dendrogram

To see only the last 10 clusters, we apply the dendrogram function with the truncate mode.

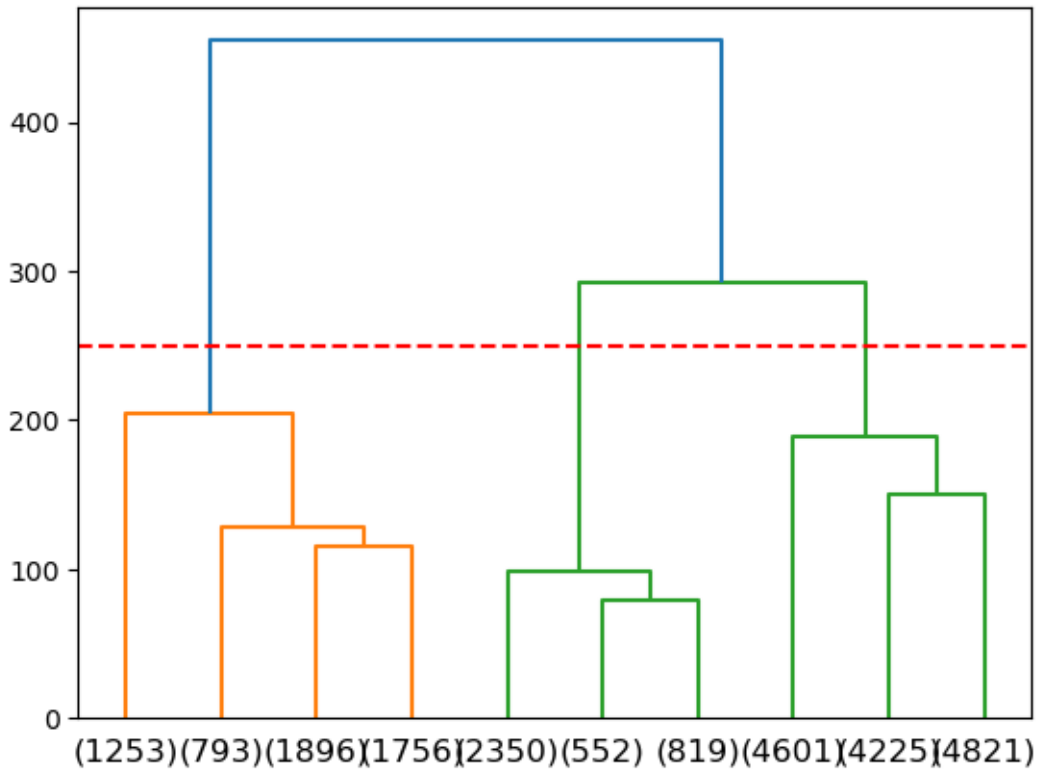


Figure 8

Inferences:

- Dendrogram helps in visualizing hierarchical clustering and assist in choosing optimum number of clusters. 23066 records are merged based on Euclidean distance and formed into one cluster.
- For better visualization, dendrogram is plot with last 10 mergers. X axis represents sample index and Y axis represent the minimum distance at which the merge happened.
- Drawing a horizontal line at the distance of 250 leads to the result of forming 3 clusters.
- It is inferred from dendrogram that Optimum number of clusters is 3.

3.2 Forming clusters

Create the clusters using both the methods:

1) Method 1 using maxclust criterion in fcluster() function. This will create the clusters for all the rows. Append this column of clusters onto the main dataset.

2) Method 2 using distance criterion in fcluster() function.

We see that there are 3 clusters which have counts like this:

1 5698
2 3721
3 13647

Inferences:

The linkage Method used for forming clusters is Ward's linkage and the distance is calculated using Euclidean distance and thereby at the distance of 250, the clusters obtained are 3 clusters.

4. Part1: Clustering: K-Means Clustering

4.1 Applying K-means clustering:

- Create K Means cluster and store the result in the object k_means.
- Fit K means on the scaled_df.
- The within sum of squares (wss) inertia value of clusters is obtained.
- The point beyond which the WSS value drop is not significant, that point is taken as optimum point for determining number of clusters.

Output:

((

```
The WSS value for  1 cluster is 253725.99999999985
The WSS value for  2 cluster is 149080.88857441364
The WSS value for  3 cluster is 105411.429057287
The WSS value for  4 cluster is 82828.27967059704
The WSS value for  5 cluster is 64258.158574817266
The WSS value for  6 cluster is 52740.712814016246
The WSS value for  7 cluster is 45767.1097879499
The WSS value for  8 cluster is 40675.08201330931
The WSS value for  9 cluster is 36189.84949512461
The WSS value for  10 cluster is 31861.547294957563
```

))

Inferences:

1. WSS keeps reducing as the number of clusters, k keeps increasing.
2. Larger the drop, better it is for model. Lesser the drop, additional cluster is not useful.

4.2 Plotting the Elbow Curve:

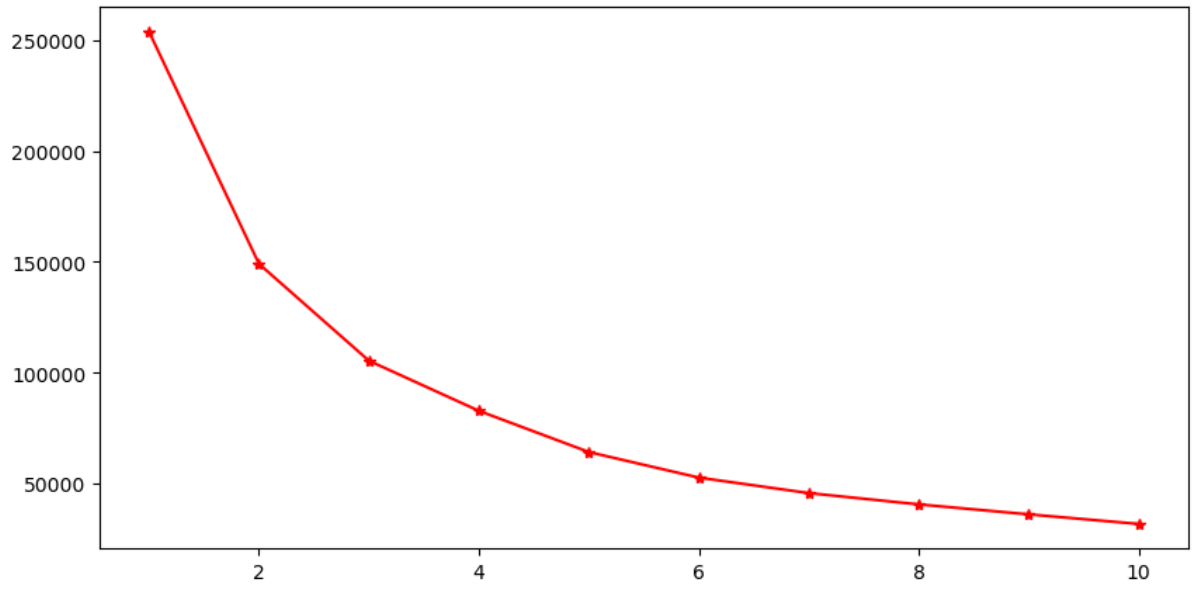


Figure 9 – Elbow Curve

Inferences:

We can see from the plot that there is a consistent dip from 3 to 5 and there doesn't seem to be a clear 'elbow' here. We may chose any from 23 to 5 as our number of clusters.

So, let's look at another method. Let's create a plot with Silhouette scores to see how it varies with k.

4.3 Checking the silhouette scores:

The silhouette_score() function is applied to the scaled data and the cluster labels obtained in the previous step.

Output:

((

```
For n_clusters=2, the silhouette score is 0.4523334291487271
For n_clusters=3, the silhouette score is 0.44203236146517316
```

```
For n_clusters=4, the silhouette score is 0.4699974809745648
For n_clusters=5, the silhouette score is 0.4243306876109383
For n_clusters=6, the silhouette score is 0.44834128335513374
For n_clusters=7, the silhouette score is 0.44552630662744597
For n_clusters=8, the silhouette score is 0.4630388125178437
For n_clusters=9, the silhouette score is 0.4714466522525009
For n_clusters=10, the silhouette score is 0.46397944565123433))
```

Inferences:

We observe that after number of clusters = 4, there is a sharp decline in the silhouette score. Hence, the optimum number of clusters should be 4.

4.4 Elbow Plot for Silhouette scores

Plot the elbow plot for the silhouette scores vs the number of clusters and note down the clusters where the silhouette score is maximum.

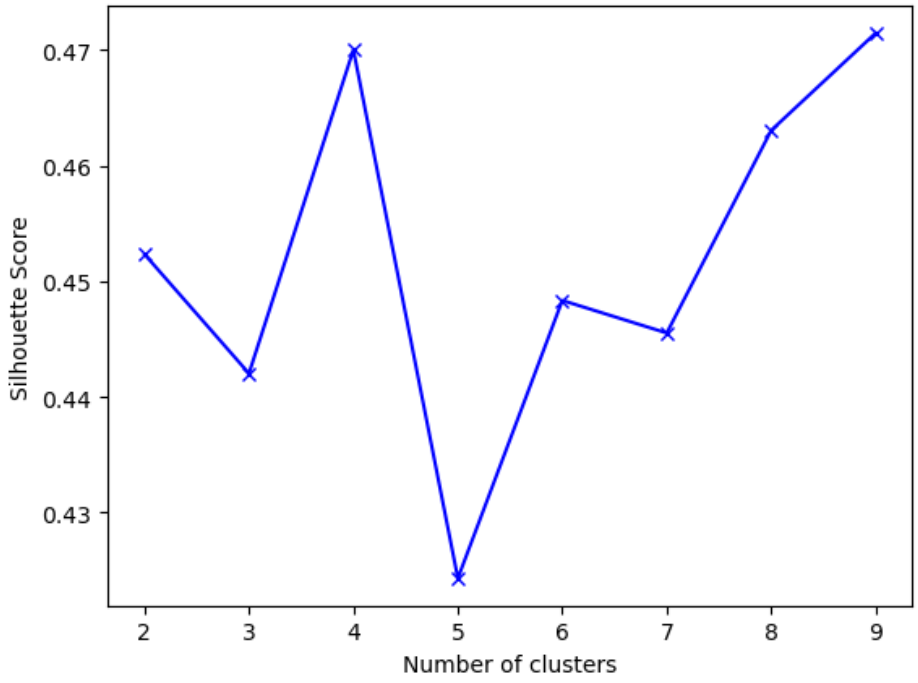


Figure 10 – Elbow Plot for Silhouette Scores

Inferences:

The silhouette score plot clearly shows that the silhouette score is the highest at 4 clusters. It can be observed that the maximum silhouette score is obtained for K=4.

Hence, the optimum number of clusters are 4.

4.5 Cluster Profiling

We append the clus_kmeans column to the original dataset. So, every row in the dataset is associated with a particular cluster.

The number of rows in each cluster is:

Cluster 1 - 4208

Cluster 2 - 13380

Cluster 3 - 1567

Cluster 4 - 3911

We group the dataset based on clus_kmeans i.e.clusters and then find the mean of all the numerical fields.

Clus_kmeans	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenu
0	460.598859	200.903042	75238.288973	10135433.938688	5488822.149477	5315528.403517	11000.566302	8442.988771	0.292115	6215.3322
1	388.857399	339.788490	101393.641256	755604.165845	364769.498804	341824.975336	6025.063154	982.914146	0.349876	639.2831
2	167.509892	558.774729	81802.169751	803443.535418	562998.441608	475164.563497	64985.560306	6910.902642	0.289075	4957.9789
3	378.566096	390.319611	109552.452058	531483.872155	259002.500895	243146.166965	4493.125543	747.163304	0.349376	487.8356

Table 9 – Dataset grouped based on clusters

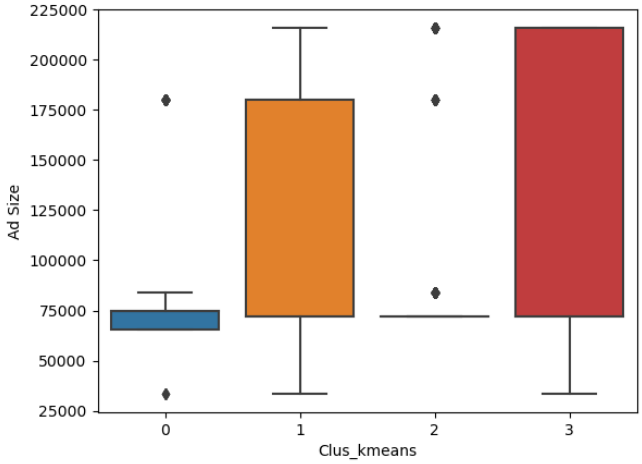
Inferences:

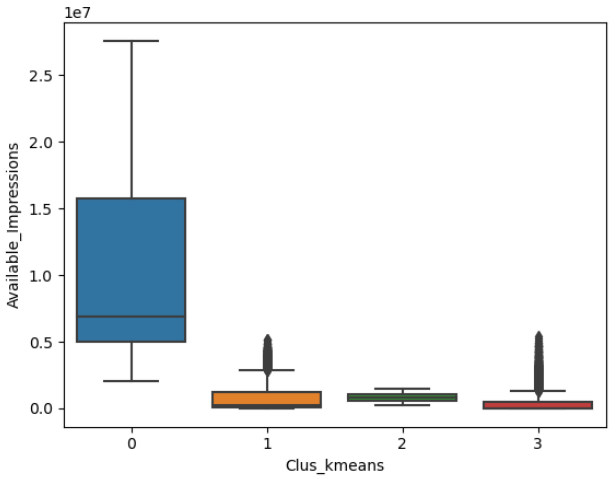
1. The impressions is the highest for Cluster 1.
2. The available impressions is the maximum for Cluster 1.
3. The number of clicks is maximum for Cluster 3
4. The number of matched queries is the highest for Cluster 1.
5. CTR is the highest for Cluster 4.
6. CPC is the highest for Cluster 3.

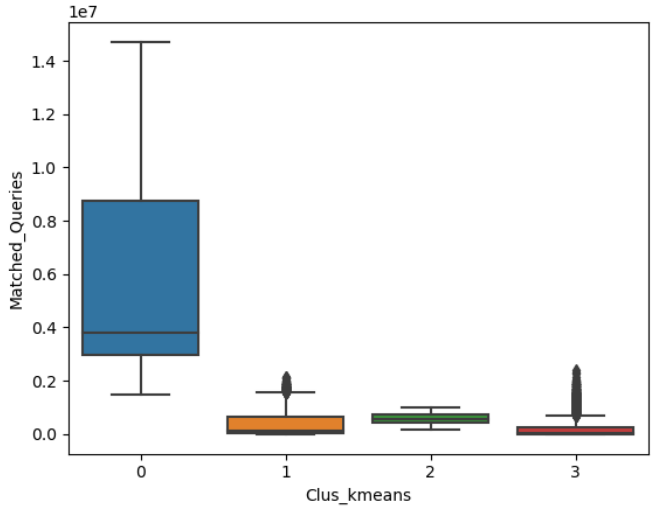
   Revenue / Spend for Cluster 1 = 73%
   Revenue /Spend for Cluster 2 = 65%
   Revenue / Spend for Cluster 3 = 71.7%
   Revenue / Spend for Cluster 4 = 65.2%

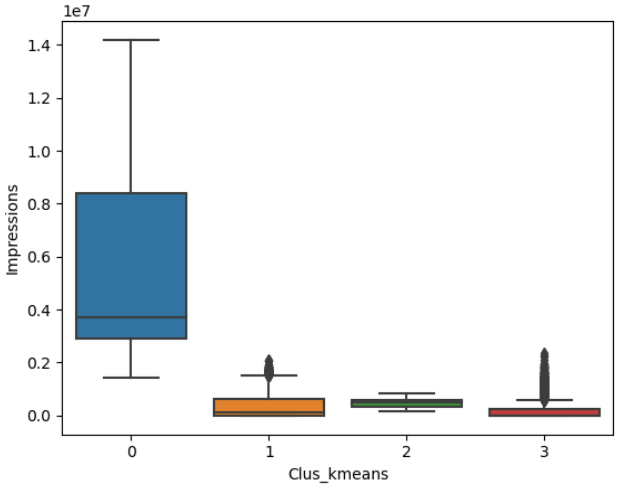
7. We observe that Revenue per every spend is maximum for Cluster 1 followed by Cluster 3.

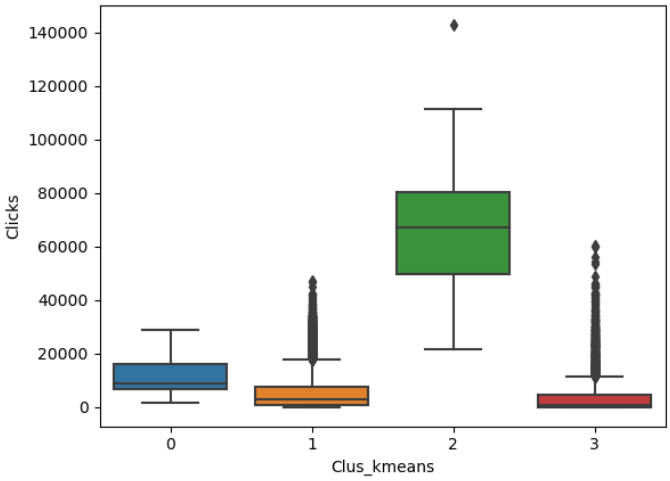
Visualization

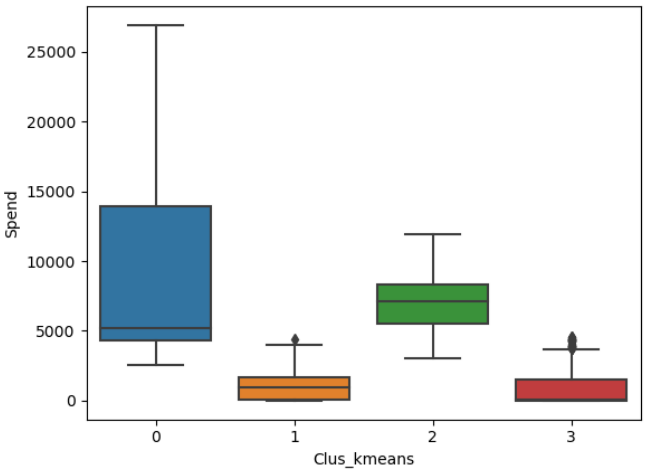












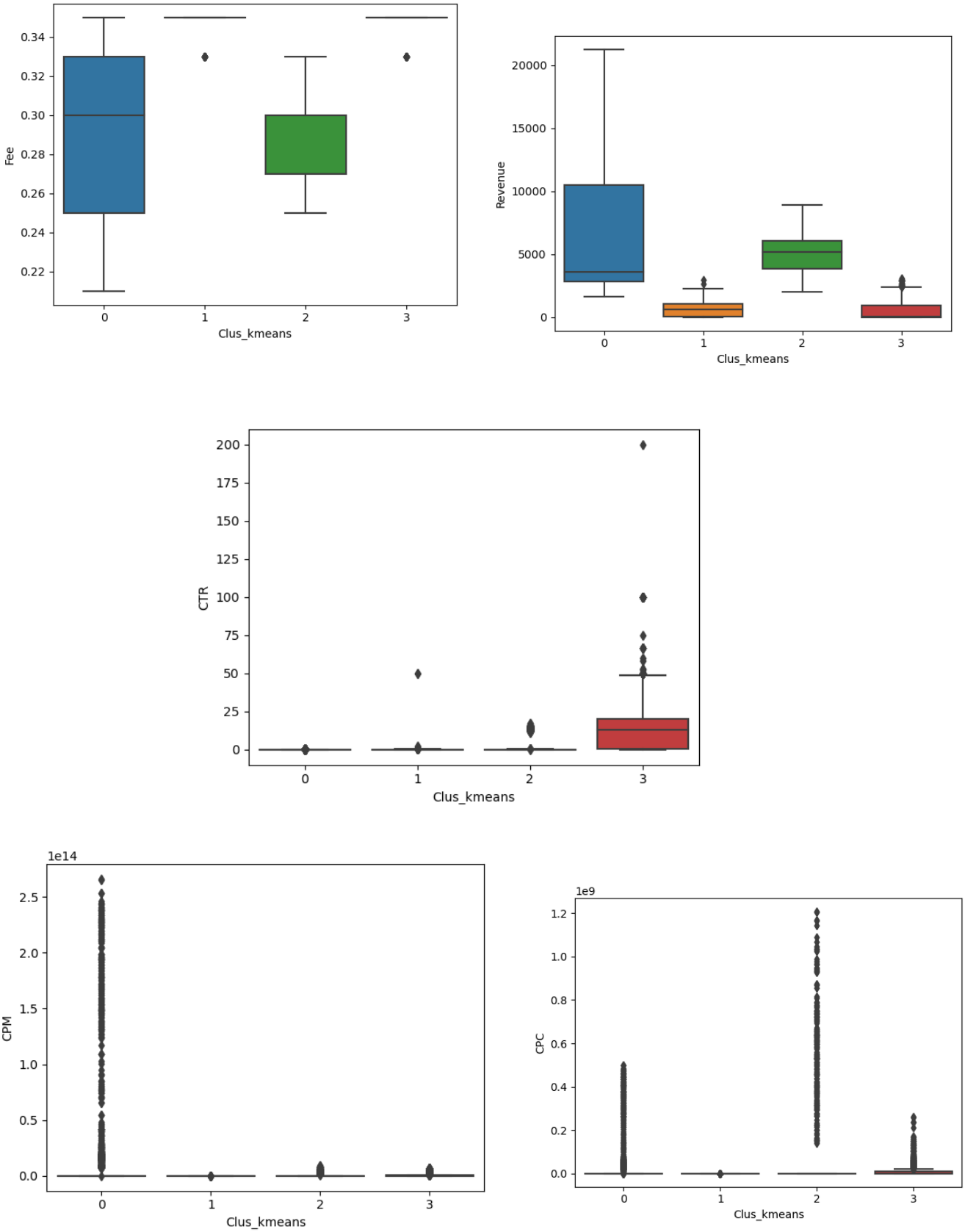


Figure 11 – Boxplots for variation w.r.t clusters

5. Part1: Clustering: Actionable Insights and Business Recommendations:

1. Cluster 1 is the best in terms of the maximum impressions that they are gathering. Cluster 1 advertisements are shown the maximum. From a business perspective, cluster 1 should be used for maximum advertisement impressions.

2. Cluster 3 gathers the highest number of clicks. This shows that the users have clicked on cluster 3 to reach an online property or for online shopping. From a business point of view, Cluster 3 can be used to further increase the revenue as the customers are clicking on this maximum number of times.

3. CTR should be maximum to generate maximum revenue. We see that CTR is the highest for cluster 4. That means, the number of times the ad receives clicks each time the ad is shown is maximum for cluster 4. From a business perspective, cluster 4 ads have been generating the maximum number of clicks by the customers. They can boost up these advertisements to get a maximum reach among the customers.

4. Revenue/Spend is maximum for Cluster 1 with 73% followed by Cluster 3. Thus, Cluster 3 should be used to increase the target audience as their spending is more. This will help to generate more revenue.

5. Revenue seems to be maximum for Cluster 1 as compared to all the clusters, as seen from the boxplot.

Part 2 - Problem 2

PCA

PCA FH (FT): Primary census abstract for female headed households excluding institutional households (India & States/UTs - District Level), Scheduled tribes - 2011 PCA for Female Headed Household Excluding Institutional Household. The Indian Census has the reputation of being one of the best in the world. The first Census in India was conducted in the year 1872. This was conducted at different points of time in different parts of the country. In 1881 a Census was taken for the entire country simultaneously. Since then, Census has been conducted every ten years, without a break. Thus, the Census of India 2011 was the fifteenth in this unbroken series since 1872, the seventh after independence and the second census of the third millennium and twenty first century. The census has been uninterruptedly continued despite of several adversities like wars, epidemics, natural calamities, political unrest, etc. The Census of India is conducted under the provisions of the Census Act 1948 and the Census Rules, 1990. The Primary Census Abstract which is important publication of 2011 Census gives basic information on Area, Total Number of Households, Total Population, Scheduled Castes, Scheduled Tribes Population, Population in the age group 0-6, Literates, Main Workers and Marginal Workers classified by the four broad industrial categories, namely, (i) Cultivators, (ii) Agricultural Laborers, (iii) Household Industry Workers, and (iv) Other Workers and also Non-Workers. The characteristics of the Total Population include Scheduled Castes, Scheduled Tribes, Institutional and Houseless Population and are presented by sex and rural-urban residence. Census 2011 covered 35 States/Union Territories, 640 districts, 5,924 sub-districts, 7,935 Towns and 6,40,867 Villages. The data collected has so many variables thus making it difficult to find useful details without using Data Science Techniques. You are tasked to perform detailed EDA and identify Optimum Principal Components that explains the most variance in data. Use Sklearn only.

Note: The 24 variables given in the Rubric is just for performing EDA. You will have to consider the entire dataset, including all the variables for performing PCA. Data file - PCA India Data Census.xlsx

2.1 Part2: PCA: Statistical Summary

2.1.1 Importing necessary libraries

The standard libraries like pandas, numpy, matplotlib and seaborn are imported. Additional libraries like PCA and TSNE are imported from sklearn.

2.1.2 Loading and Reading the data

The data PCA_India_Data_Census is first loaded and then read using the head() function.

	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180	
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123	
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	...	114	188	44	

Table 10 – Reading the first 3 rows of PCA_India_Data_Census

2.1.3 Checking the shape

Use the shape() function to determine the number of rows and columns in the dataset.

```
There are 640 rows and 61 columns.
```

2.1.4 Checking the datatype of columns

The datatype of columns is found out using the info() function.

Output:
((

```
#   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
))
```

We also did an analysis of the number of rows with 0 value in the Male and Female SC and ST columns.

Output:

((

```
% of records with 0 value in M_SC field: 6.88%
% of records with 0 value in F_SC field: 7.03%
% of records with 0 value in M_ST field: 8.75%
% of records with 0 value in F_ST field: 8.91%
```

))

Inferences:

- There are 640 observations and 61 columns.
- 59 variables are integer data type while 2 are object data type.
- All the variables except State and Area Name are of numeric data type.
- As the percentage is too low of the records having 0 value in the four fields, namely, M_SC, F_SC, M_ST, F_ST, hence, we would not delete these fields.

2.1.5 EDA of any five fields

Out of the given 24 fields, the five fields that I have selected to perform EDA are M_LIT, F_LIT, TOT_WORK_M, TOT_WORK_F and F_ILL

Use the describe() function to get the statistical summary of these columns.

	count	mean	std	min	25%	50%	75%	max
M_LIT	640.0	57967.979688	55910.282466	286.0	21298.00	42693.5	77989.50	403261.0
F_LIT	640.0	66359.565625	75037.860207	371.0	20932.00	43796.5	84799.75	571140.0
TOT_WORK_M	640.0	37992.407813	36419.537491	100.0	13753.50	27936.5	50226.75	269422.0
TOT_WORK_F	640.0	41295.760938	37192.360943	357.0	16097.75	30588.5	53234.25	257848.0
F_ILL	640.0	56012.518750	47116.693769	327.0	22367.00	42386.0	78471.00	254160.0

Table 11 – Summary of 5 columns of dataset

Inferences:

- The literate males range from a minimum of 286 to a maximum of 403261.

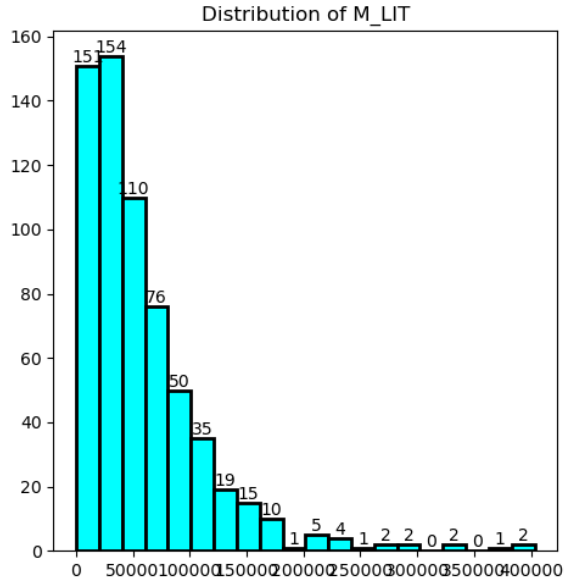
- The literate females range from a minimum of 371 to a maximum of 571140.
- The maximum number of literate females is much more than the literate males.
- The mean literacy rate of females is much more than that compared to the males.
- The range of total working male population is very high from 100 to 269422.
- The range of total working female population is also quite high, from 357 to 257848.
- The mean working population of females is higher than the mean working male population.
- 78471 constitutes 75% of the female population who are illiterate.
- The standard deviation of literate females is the highest.
- The average female illiterate population is around 56000.

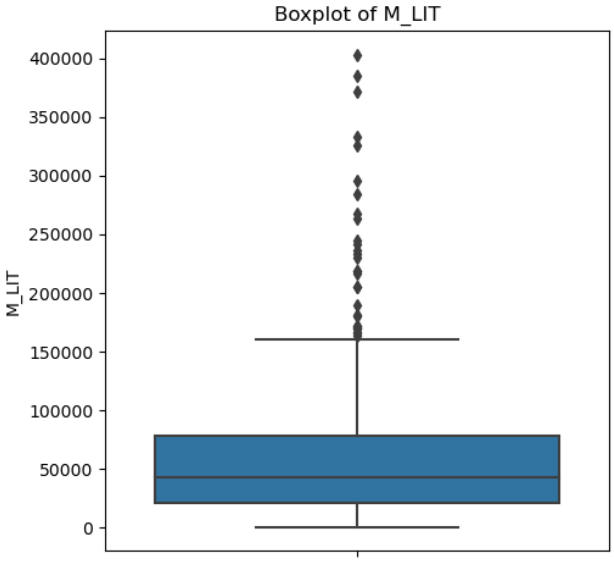
2.1.6 EDA – Data Visualization

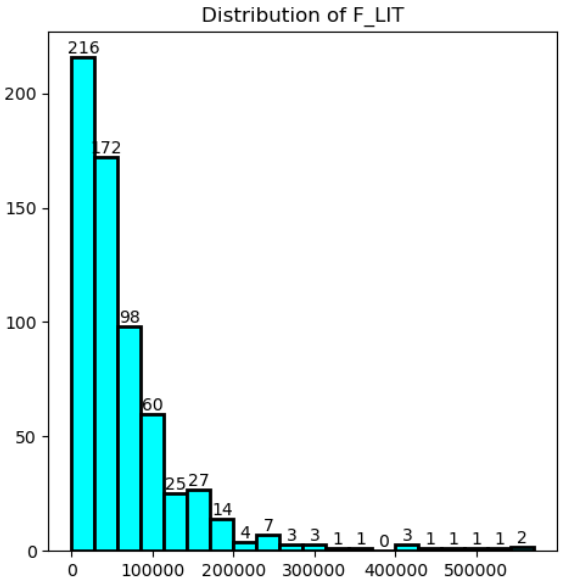
2.1.6.1 Univariate Analysis of the 5 numeric variables

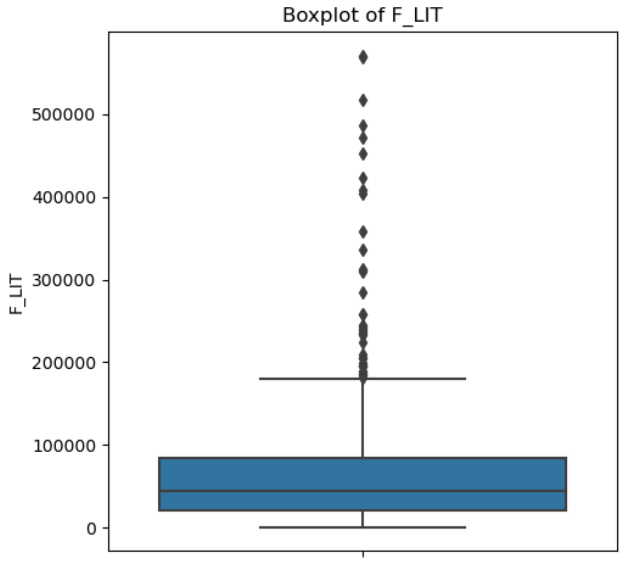
We define a function 'univariateAnalysis_numeric' to display information as part of univariate analysis of numeric variables. The function will accept column name and number of bins as arguments.

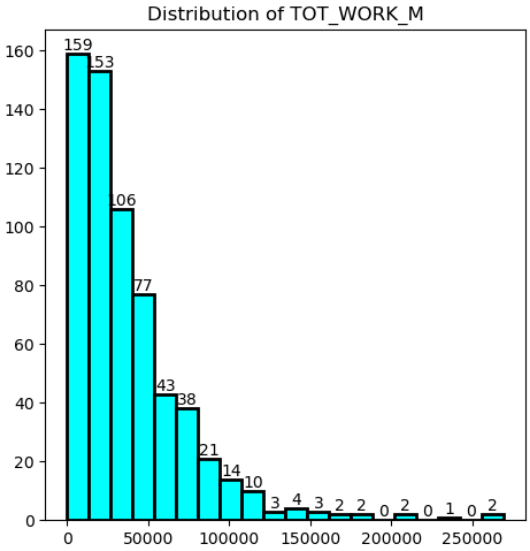
The function will display the statistical description of the numeric variable, histogram or distplot to view the distribution and the box plot to view 5 point summary and outliers if any.

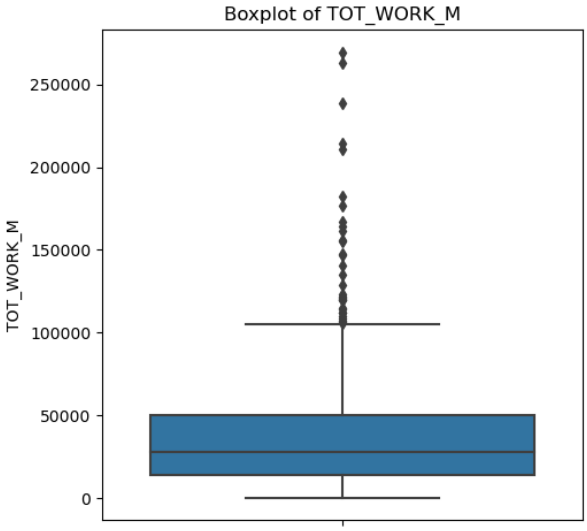


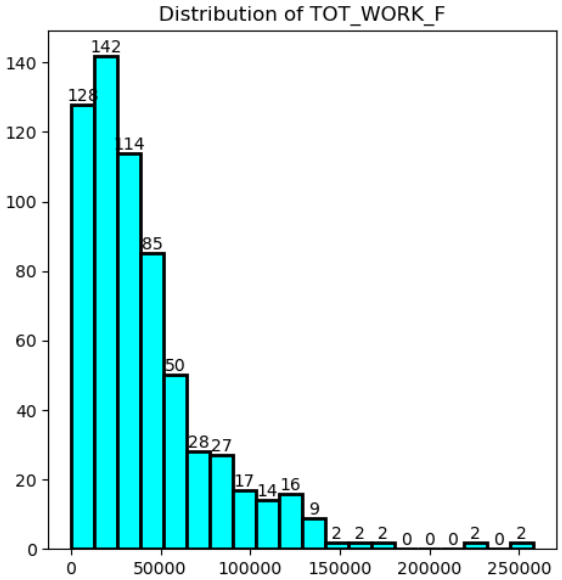


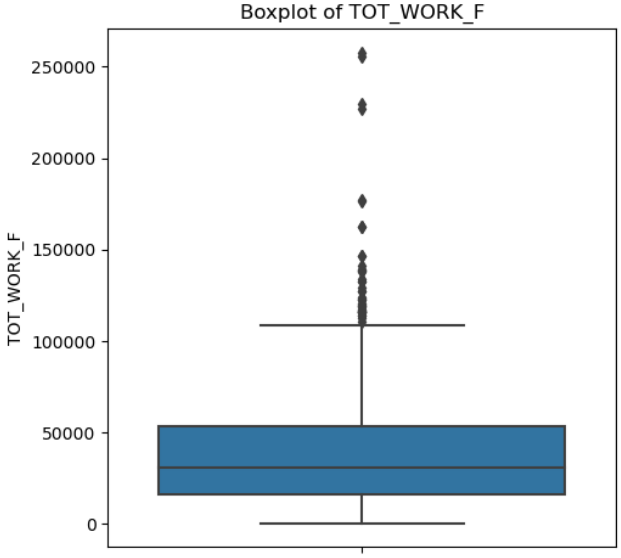

Distribution of F_LIT
216
172
98
60
25
27
14
4
7
3
3
1
1
0
3
1
1
1
1
2
0
100000
200000
300000
400000
500000


Boxplot of F_LIT
F_LIT
500000
400000
300000
200000
100000
0


Distribution of TOT_WORK_M
159
153
106
77
43
38
21
14
10
3
4
3
2
2
0
2
0
1
0
2
0
50000
100000
150000
200000
250000


Boxplot of TOT_WORK_M
TOT_WORK_M
250000
200000
150000
100000
50000
0


Distribution of TOT_WORK_F
142
128
114
85
50
28
27
17
14
16
9
2
2
2
0
0
0
2
0
2
0
50000
100000
150000
200000
250000


Boxplot of TOT_WORK_F
TOT_WORK_F
250000
200000
150000
100000
50000
0

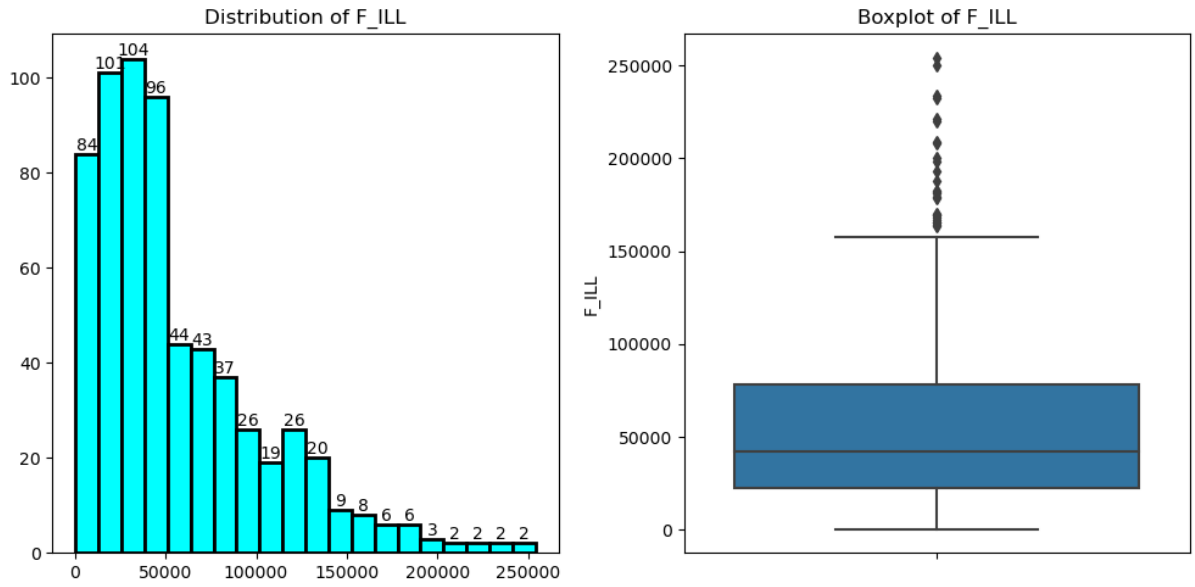


Figure 12 – Univariate Analysis

Inferences:

- All the 5 variables namely, M_LIT, F_LIT, TOT_WORK_M, TOT_WORK_F and F_ILL are right-skewed.
- All the five variables have outliers which need to be treated.

2.1.6.2 Bivariate Analysis

The variation of each of the 5 variables with each other is observed using a pairplot.

Inferences:

There is a positive correlation between all these variables:

a) M_LIT and F_ILL
b) M_LIT and TOT_WORK_F
c) M_LIT and TOT_WORK_M
d) M_LIT and F_LIT
e) F_LIT and F_ILL
f) F_LIT and TOT_WORK_F
g) F_LIT and TOT_WORK_M
h) TOT_WORK_M and F_ILL
i) TOT_WORK_M and TOT_WORK_F
j) TOT_WORK_F and F_ILL

As one variable increases, the other variable also increases, indicating positive correlation.

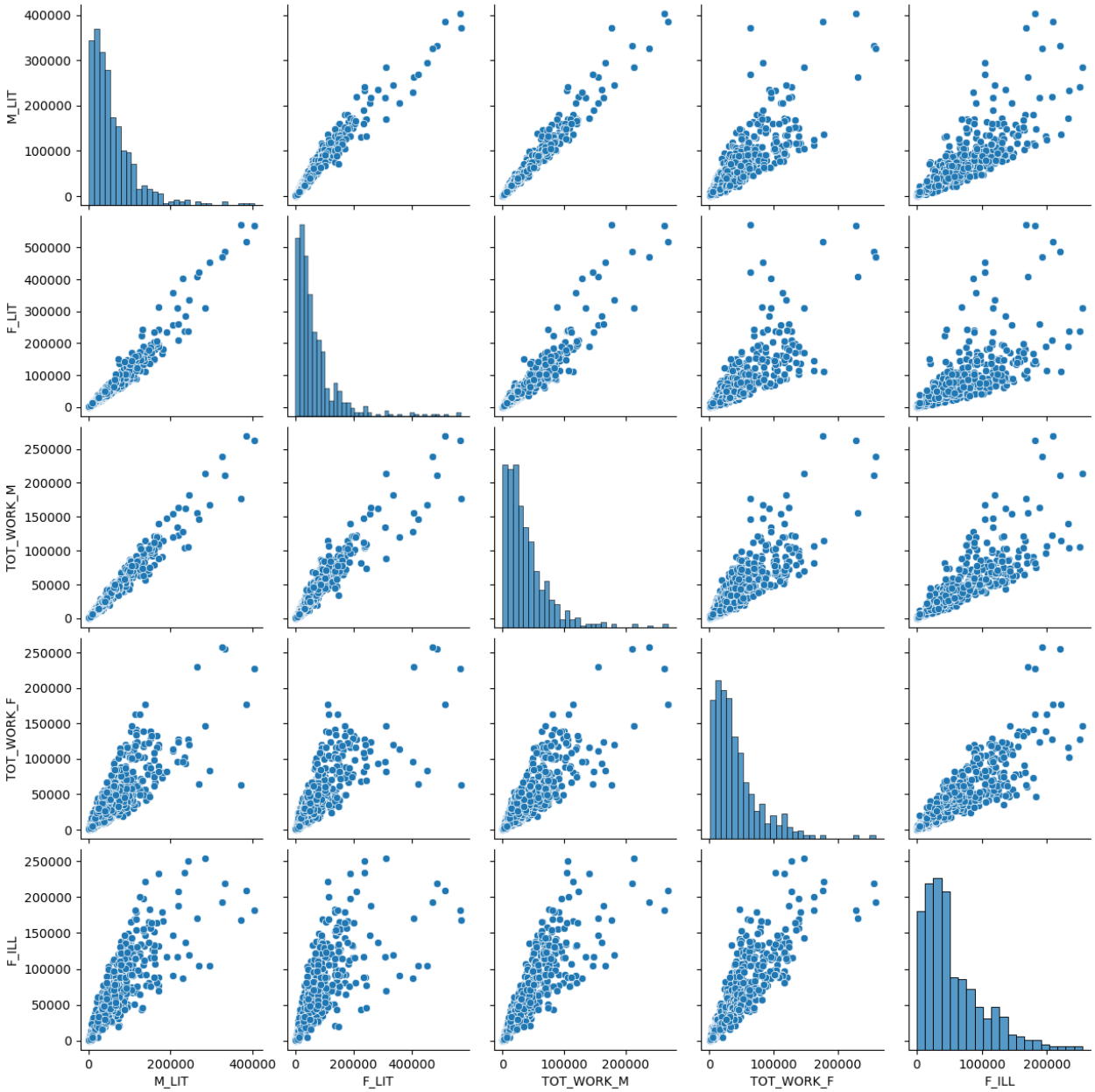


Figure 13 – Pairplot for correlation b/w numerical variables

2.1.6.3 Multivariate Analysis

Inferences:

From the heatmap, it is clearly seen that the correlation between these variables is very high:

a) M_LIT and TOT_WORK_M
b) F_LIT and TOT_WORK_M
c) F_LIT and M_LIT

* The correlation between F_LIT and F_ILL is the least.

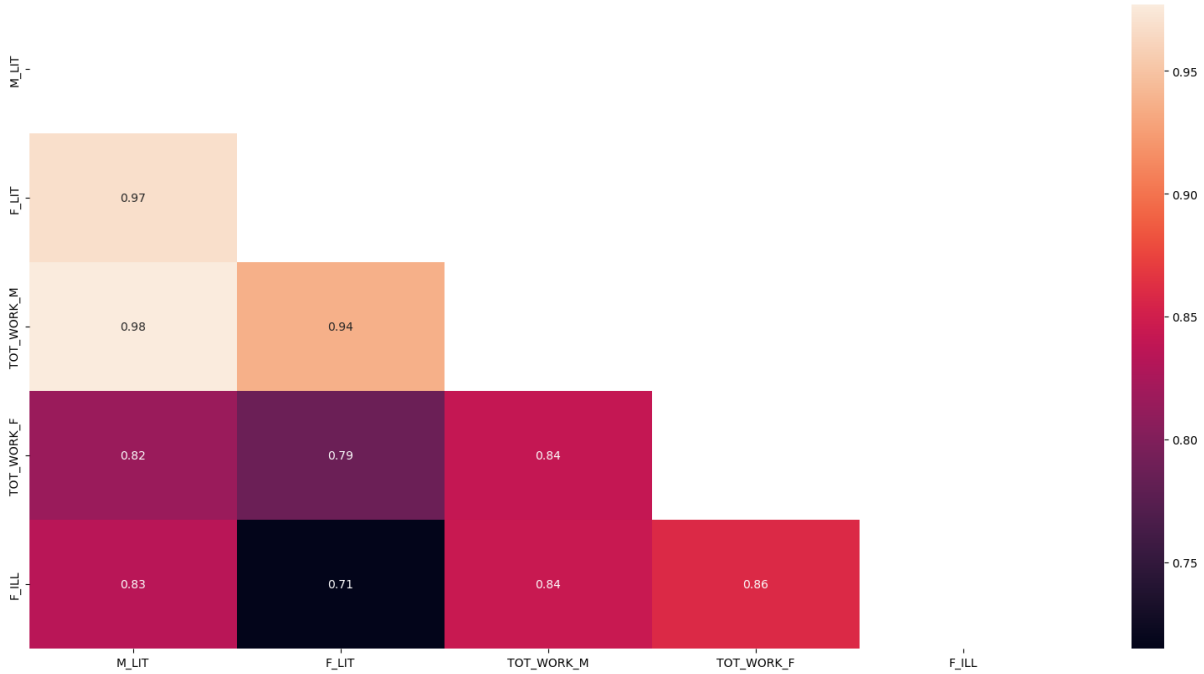


Figure 14 – Heatmap to show correlation

2.1.7 Group by

a) Which state has the highest and the lowest gender ratio?

Ans: Use the groupby() function with respect to State and mean of total males and mean of total females. The gender ratio is a ratio of the number of males to the number of females.

Output:

((

```
State
Lakshadweep      0.868061
Haryana          0.779129
NCT of Delhi     0.775077
```

))

The gender ratio, that is, the number of males to the number of females is highest in Lakshadweep followed by Haryana. In Lakshadweep, the gender ratio is 0.86.

b) Which district has the highest and the lowest gender ratio?

Ans: Use the groupby() function for Area Name and mean of total number of males and mean of total number of females.

Output:

((

```
Area Name
Lakshadweep       0.868061
Badgam            0.847762
Mahamaya Nagar    0.847313
```

))

Lakshadweep district has the highest gender ratio of 0.86.

Output:

((

```
Area Name
Virudhunagar    0.449352
Koraput         0.440769
Krishna         0.437972
```

))

Krishna district has the lowest gender ratio.

2.2 PCA – Data Preprocessing

2.2.1 Check for duplicate rows

Use the duplicated function() to find the number of duplicate rows.

```
No. of duplicate rows is 0
```

As the number of duplicated rows is 0, we are good to proceed.

2.2.2 Check for missing values

Use the isnull() function to check for missing values.

```
State Code        0
Dist.Code         0
State             0
Area Name         0
```

No_HH	0
TOT_M	0
TOT_F	0
M_06	0
F_06	0
M_SC	0
F_SC	0
M_ST	0
F_ST	0
M_LIT	0
F_LIT	0
M_ILL	0
F_ILL	0
TOT_WORK_M	0
TOT_WORK_F	0
MAINWORK_M	0
MAINWORK_F	0
MAIN_CL_M	0
MAIN_CL_F	0
MAIN_AL_M	0
MAIN_AL_F	0
MAIN_HH_M	0
MAIN_HH_F	0
MAIN_OT_M	0
MAIN_OT_F	0
MARGWORK_M	0
MARGWORK_F	0
MARG_CL_M	0
MARG_CL_F	0
MARG_AL_M	0
MARG_AL_F	0
MARG_HH_M	0
MARG_HH_F	0
MARG_OT_M	0
MARG_OT_F	0
MARGWORK_3_6_M	0
MARGWORK_3_6_F	0
MARG_CL_3_6_M	0
MARG_CL_3_6_F	0
MARG_AL_3_6_M	0
MARG_AL_3_6_F	0
MARG_HH_3_6_M	0
MARG_HH_3_6_F	0
MARG_OT_3_6_M	0
MARG_OT_3_6_F	0
MARGWORK_0_3_M	0
MARGWORK_0_3_F	0
MARG_CL_0_3_M	0
MARG_CL_0_3_F	0
MARG_AL_0_3_M	0
MARG_AL_0_3_F	0
MARG_HH_0_3_M	0
MARG_HH_0_3_F	0
MARG_OT_0_3_M	0
MARG_OT_0_3_F	0
NON_WORK_M	0
NON_WORK_F	0

There are no missing values in the dataset. We are good to proceed further.

2.2.3 Dropping the object type datatype columns

Drop the object columns State, Area Name, State Code and Dist.Code columns as they are not useful for PCA.

After dropping the 4 columns namely, State, Area Name, State Code and District Code, the number of columns left in the dataset is 57 columns.

2.2.4 Checking for Outliers/Visualization of data before scaling

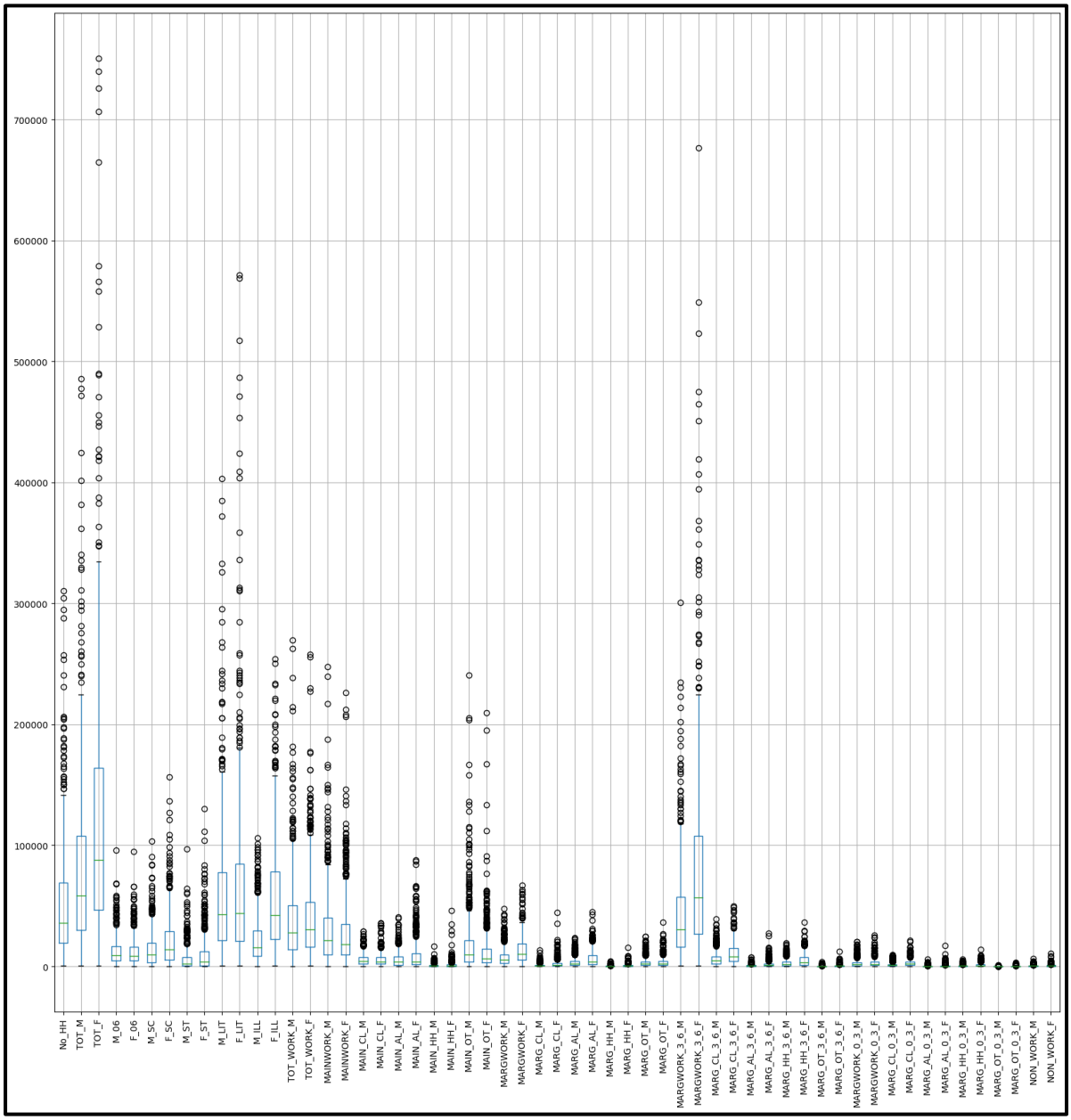


Figure 15 – Checking for outliers

Inferences:

All the fields have outliers. Most of them are right-skewed.

The outlier treatment needs to be done before proceeding further.

2.2.5 Treatment of outliers in the dataset

Define a function which returns the Upper and Lower limit to detect outliers for each feature.
Find the Q1 and Q3 which are 25% and 75% of the columns. The Inter-Quartile Range is the difference between Q1 and Q3.
Cap & floor the values beyond the outlier boundaries.

2.2.6 Visualization after treatment of Outliers

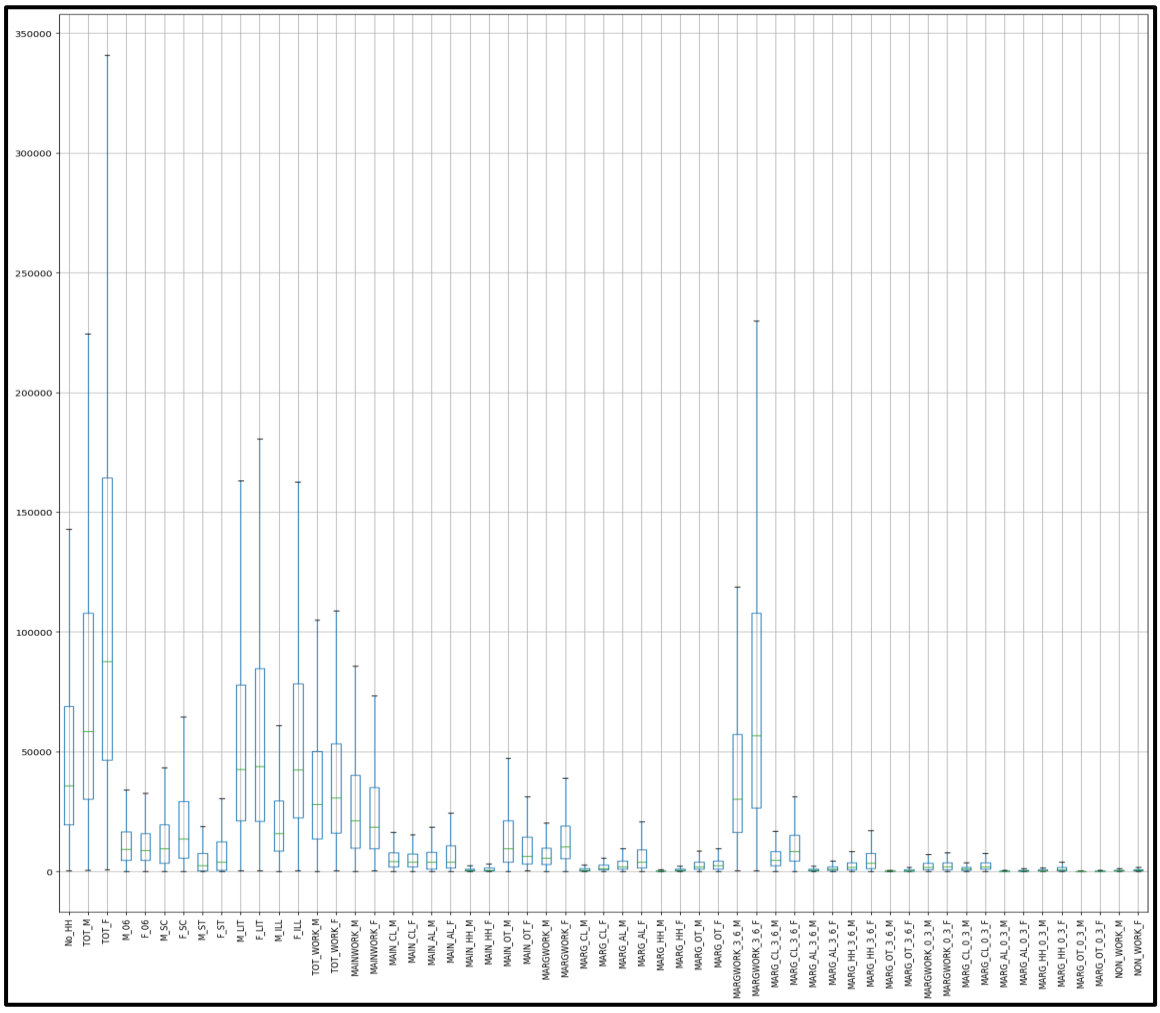


Figure 16 – Treatment of outliers

After treating the outliers, we see that there are no outliers for any of the fields.

2.2.7 Scaling the data or Standardizing the data

Apply the z-score to the dataset to standardize the data. Standardising the data is very important to bring the entire data to the same scale.

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M
0	-1.038986	-0.874837	-0.937027	-0.624685	-0.561282	-1.080201	-1.079963	-0.510440	-0.574198	-0.939617	...	-0.093587	-0.860882	-0.041800
1	-1.076896	-0.938023	-1.009723	-0.773932	-0.835657	-1.079873	-1.079635	-0.771833	-0.782092	-1.005083	...	-0.719169	-0.877096	-0.347040
2	-1.121858	-1.154665	-1.141539	-1.141642	-1.138104	-1.080201	-1.079635	0.122588	0.137599	-1.141561	...	-1.130551	-1.128423	-0.770091
3	-1.201599	-1.217171	-1.214930	-1.197772	-1.176091	-1.080447	-1.079963	-0.399531	-0.437333	-1.203009	...	-1.050477	-1.100286	-0.679055
4	-0.938495	-0.921309	-0.935018	-0.700931	-0.740523	-1.078807	-1.078160	0.432534	0.249489	-0.942767	...	-0.369844	-0.298617	1.484398
5	-0.819062	-0.843235	-0.854167	-0.592997	-0.661764	-0.868183	-0.861395	0.453486	0.276262	-0.869084	...	0.565025	0.468239	2.241473
6	-0.916701	-0.884722	-0.929479	-0.833847	-0.895033	-0.641074	-0.702606	-0.678549	-0.702771	-0.863058	...	-0.742190	-0.950539	-0.148902
7	-0.995526	-0.881948	-0.929284	-0.799455	-0.794885	-1.080447	-1.079963	-0.685699	-0.715858	-0.909692	...	-0.394867	-0.799362	-0.106061
8	-1.138102	-1.091977	-1.118799	-1.080645	-1.053330	-1.080447	-1.079963	-0.552010	-0.603169	-1.112206	...	-0.729178	-0.788393	-0.427366
9	-0.850887	-0.615211	-0.692876	-0.580668	-0.616091	-1.079545	-1.079799	-0.809578	-0.810065	-0.655293	...	-0.936371	-1.012059	-0.952164
10	-1.161932	-1.143135	-1.153235	-1.107683	-1.105464	-1.080447	-1.079963	-0.603890	-0.641531	-1.148021	...	-0.919355	-0.975815	-0.791511

Table 12 – Scaled data

Visualisation after scaling the data:

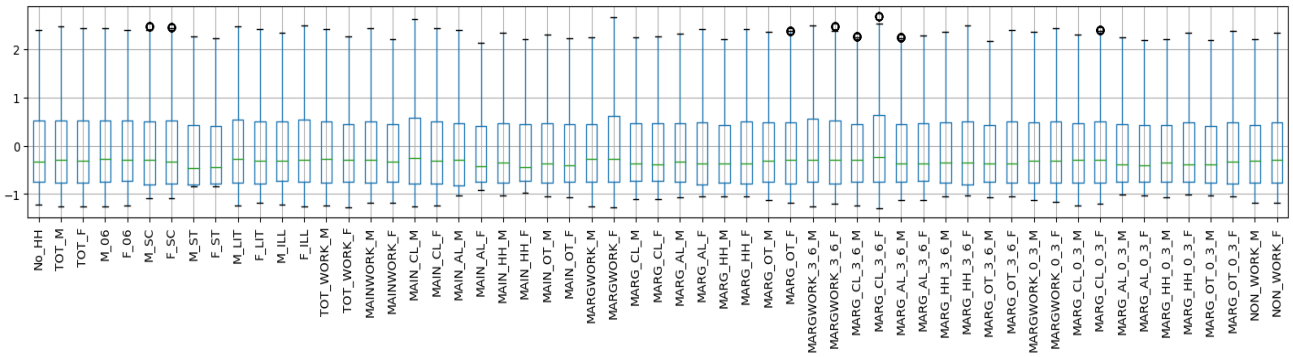


Figure 17 – Visualization after scaling the data

2.2.8 Significance of correlation test - Bartlett Sphericity Test

Bartletts Test of Sphericity to Statisticaly confirm presence of Multicolinearity
Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.

H0: All variables in the data are uncorrelated.
Ha: At least one pair of variables in the data are correlated.
If the null hypothesis cannot be rejected, then PCA is not advisable.

If the p-value is small, then we can reject the null hypothesis and agree that there is atleast one pair of variables in the data which are correlated hence PCA is recommended.

Inferences:

p_value = 0

As the p_value < 0.05, we can say that the alternate hypothesis is true.

So, there are significant correlations between the fields.

2.2.9 Kmo Test to check the adequacy of sample size

KMO Test to check enough number of observations there to perform PCA
The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.

Generally, if MSA is less than 0.5, PCA is not recommended, since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

Inferences:

Kmo_model = `0.9361896166652829`

As the kmo_model test gives a value of 0.93 which is greater than 0.7, we can say that the sample size is adequate.

2.3 Part2: PCA

2.3.1 Generating the Covariance and Correlation Matrix

Correlation Matrix: Apply the corr() function to the scaled dataset.

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	\
No_HH	1.00	0.91	0.97	0.81	0.81	0.81	0.86	0.12	0.12	
TOT_M	0.91	1.00	0.98	0.96	0.96	0.88	0.86	0.02	0.01	
TOT_F	0.97	0.98	1.00	0.91	0.91	0.86	0.88	0.08	0.07	
M_06	0.81	0.96	0.91	1.00	1.00	0.83	0.80	-0.01	-0.02	
F_06	0.81	0.96	0.91	1.00	1.00	0.82	0.79	0.01	-0.01	
M_SC	0.81	0.88	0.86	0.83	0.82	1.00	0.98	-0.10	-0.10	
F_SC	0.86	0.86	0.88	0.80	0.79	0.98	1.00	-0.05	-0.05	
M_ST	0.12	0.02	0.08	-0.01	0.01	-0.10	-0.05	1.00	0.99	
F_ST	0.12	0.01	0.07	-0.02	-0.01	-0.10	-0.05	0.99	1.00	
M_LIT	0.93	0.99	0.98	0.92	0.91	0.87	0.86	0.03	0.02	
F_LIT	0.94	0.94	0.96	0.84	0.83	0.80	0.82	0.05	0.04	
M_ILL	0.78	0.93	0.88	0.97	0.97	0.82	0.78	0.02	0.01	
F_ILL	0.89	0.92	0.93	0.90	0.90	0.84	0.86	0.11	0.11	
TOT_WORK_M	0.94	0.98	0.97	0.90	0.89	0.87	0.86	0.06	0.05	
TOT_WORK_F	0.95	0.82	0.90	0.73	0.73	0.73	0.80	0.25	0.26	
MAINWORK_M	0.93	0.93	0.94	0.83	0.82	0.84	0.84	0.05	0.04	
MAINWORK_F	0.92	0.77	0.86	0.65	0.65	0.69	0.76	0.22	0.22	
MAIN_CL_M	0.52	0.63	0.59	0.65	0.65	0.64	0.62	0.07	0.06	

Table 13 - Correlation Matrix

Covariance Matrix: Apply the cov() function to the scaled dataset.

```
[[1.   0.91 0.97 ... 0.65 0.77 0.8 ]
 [0.91 1.   0.98 ... 0.73 0.87 0.79]
 [0.97 0.98 1.   ... 0.71 0.84 0.81]
 ...
 [0.65 0.73 0.71 ... 1.   0.76 0.72]
 [0.77 0.87 0.84 ... 0.76 1.   0.9 ]
 [0.8  0.79 0.81 ... 0.72 0.9  1.  ]]
```

Table 14 – Covariance Matrix

Inferences:

We see that once the data is scaled, the correlation matrix and the covariance matrix are the same.

2.3.2 Fit and Transform PCA Model

Apply the PCA function to all the 57 fields in the dataset and fit transform the same. The output is as follows:

```
array([[-5.53,  0.43, -1.47, ...,  0.01,  0.  ,  0.  ],
       [-5.49, -0.11, -2.02, ..., -0.  ,  0.01, -0.01],
       [-7.47, -0.22, -0.25, ..., -0.  , -0.  ,  0.  ],
       ...,
       [-7.89, -1.  , -0.91, ..., -0.  ,  0.  , -0.  ],
       [-7.86, -1.  , -0.85, ..., -0.  , -0.  ,  0.  ],
       [-7.42, -1.41, -0.87, ..., -0.  , -0.  ,  0.  ]])
```

Table 15 – PCA Model

Inferences:

The fit_transform() method will first fit the PCA model to the data. This involves computing the mean and covariance of the data. The model is then used to transform the data into a new space, where the principal components are aligned with the directions of greatest variance in the data.

2.3.3 Calculating the Eigen Vectors

The Eigen vectors represent the new set of axes of the Principal component space and also the Eigen values carry the information of the amount of variance that each eigenvector has.

Eigen vectors represent the direction of a line. Eigen values are numbers that indicate how spread out a data set is on the line.

In PCA, eigenvalues are coefficients applied to eigenvectors that give the vectors their length or magnitude.

```
array([[ 0.15,  0.16,  0.16, ...,  0.14,  0.15,  0.14],
       [-0.12, -0.08, -0.09, ...,  0.04, -0.05, -0.04],
       [ 0.1 , -0.04,  0.03, ..., -0.1 , -0.13, -0.03],
       ...,
       [ 0.  , -0.01,  0.02, ..., -0.01,  0.06, -0.01],
       [ 0.  ,  0.05,  0.  , ...,  0.01, -0.08, -0.  ],
       [-0.  , -0.  ,  0.01, ...,  0.  ,  0.01,  0.  ]])
```

Table 16 – Eigen Vectors

2.3.4 Calculating the Eigen Values

```
array([35.6,  7.6,  3.8,  2.8,  1.9,  1.2,  1. ,  0.5,  0.4,  0.3,  0.3,
       0.2,  0.2,  0.2,  0.1,  0.1,  0.1,  0.1,  0.1,  0.1,  0.1,  0.1,
       0. ,  0. ,  0. ,  0. ,  0. ,  0. ,  0. ,  0. ,  0. ,  0. ,  0. ,
       0. ,  0. ,  0. ,  0. ,  0. ,  0. ,  0. ,  0. ,  0. ,  0. ,  0. ,
       0. ,  0. ,  0. ,  0. ,  0. ,  0. ,  0. ,  0. ,  0. ,  0. ,  0. ,
       0. ,  0. ])
```

Table 17 – Eigen Values

By default, Eigen Values are always generated in descending order.

2.3.5 Calculate the Explained variance for each PC

Obtain percentage of variability explained by each PC.

Explained variance = $\frac{\text{Eigen value of each PC}}{Sum\ of\ eigen\ values\ of\ all\ PCs}$

```
array([0.624, 0.134, 0.066, 0.049, 0.033, 0.02 , 0.017, 0.008, 0.007,
       0.006, 0.005, 0.004, 0.003, 0.003, 0.002, 0.002, 0.002, 0.002,
       0.002, 0.001, 0.001, 0.001, 0.001, 0.001, 0.001, 0.001, 0.001,
       0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ,
       0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ,
       0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ,
       0.   , 0.   , 0.   ])
```

Table 18 – Variance for each PC

Obtain percentage of variability explained by each PCs

```
[62.4 13.4  6.6  4.9  3.3  2.   1.7  0.8  0.7  0.6  0.5  0.4  0.3  0.3
  0.2  0.2  0.2  0.2  0.2  0.1  0.1  0.1  0.1  0.1  0.1  0.1  0.1  0.
  0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.
  0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.
  0. ]
```

Table 19 – Percentage of variance of each PC

2.3.6 Calculating the cut-off for selecting the optimum number of PCs

```
Cumulative Variance Explained in Percentage: [62.4 75.8 82.4 87.3 90.6 92.6 94.3 95.1 95.8 96.4 96.9 97.3 97.6 97.9
 98.1 98.3 98.5 98.7 98.9 99.  99.1 99.2 99.3 99.4 99.5 99.6 99.7 99.7
 99.7 99.7 99.7 99.7 99.7 99.7 99.7 99.7 99.7 99.7 99.7 99.7 99.7 99.7
 99.7 99.7 99.7 99.7 99.7 99.7 99.7 99.7 99.7 99.7 99.7 99.7 99.7 99.7
 99.7]
```

Table 20 – Cumsum for selecting PCs

Inferences:

- 5 PCs contribute 90% of variance.
- 90% of variability is explained 5 PCs.
- We observe that the cumulative variance reaches 90% at PC5. Hence, we select 5 Principal Components as the optimum number.

2.3.7 Creating a scree plot to identify the optimum number of PCs

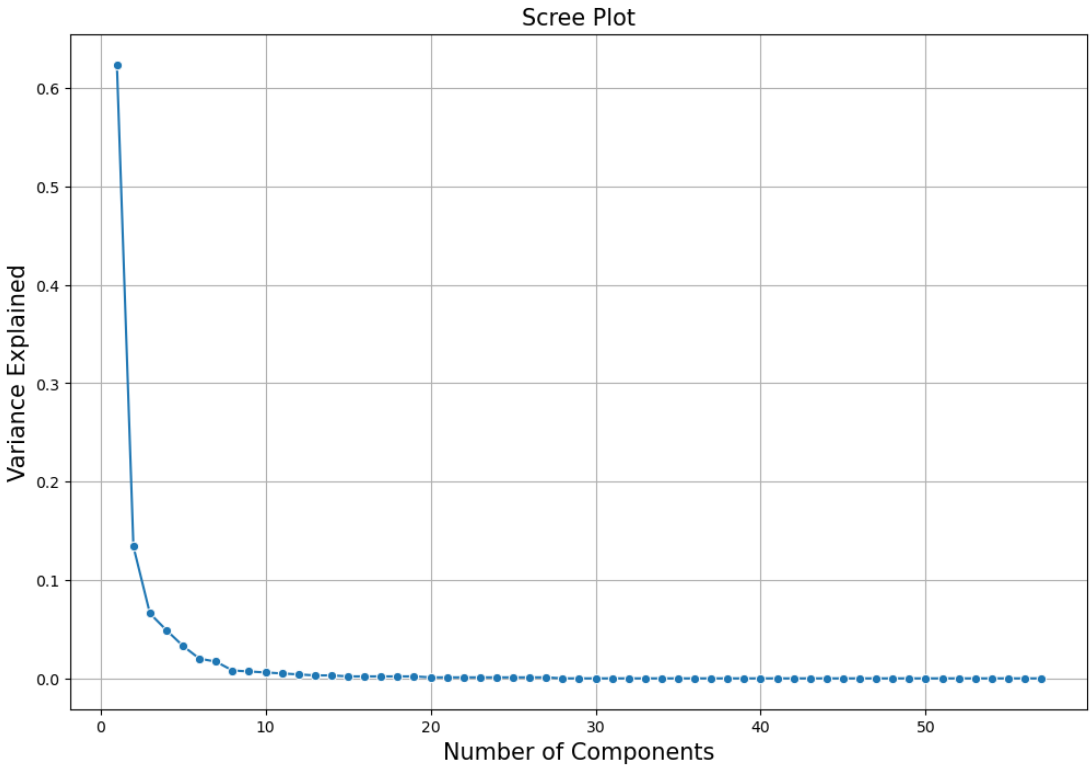


Figure 18 – Scree Plot

2.3.8 Creating a bar plot for identifying the optimum number of PCs

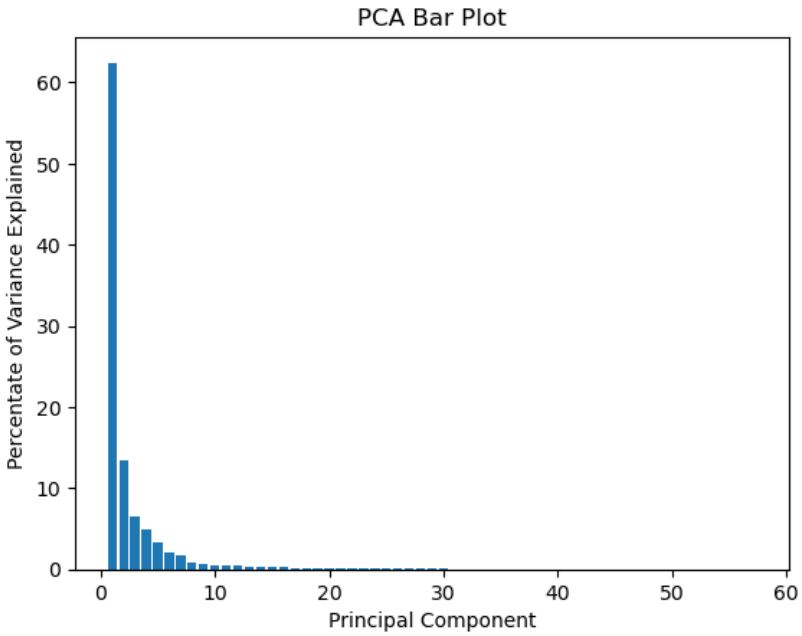


Figure 19 – Bar Plot of PCs vs % Variance

2.3.9 Building PCA model with 5 components

Apply PCA for the number of decided components to get the loadings and component output. We are generating only 5 PCA dimensions (dimensionality reduction from 57 to 5)

```
array([[-5.52816148e+00,  4.30377559e-01, -1.47382695e+00,
        -1.27804898e+00,  3.76357641e-01],
       [-5.49201646e+00, -1.06110331e-01, -2.01564100e+00,
        -1.75016759e+00, -6.85689225e-03],
       [-7.47464297e+00, -2.17193764e-01, -2.47428211e-01,
         6.07916548e-03,  5.56282180e-01],
       ...,
       [-7.88626804e+00, -1.00353656e+00, -9.09284569e-01,
        -1.23800927e+00,  1.46031242e-01],
       [-7.86425952e+00, -9.99337996e-01, -8.51569237e-01,
        -7.82561039e-01, -8.16813905e-02],
       [-7.41622568e+00, -1.41214300e+00, -8.65921210e-01,
        -6.80528005e-01,  9.68605787e-02]])
```

Table 21 – PCA model with 5 components

We have 5 sets of Eigen Vectors because we gave 5 PCs. Each Eigen Vector will have 57 coefficients in it.

2.3.10 Linear Equation

The Linear equation of 1st component:

0.149 * No_HH + 0.159 * TOT_M + 0.158 * TOT_F + 0.156 * M_06 + 0.157 * F_06 + 0.14 3 * M_SC + 0.144 * F_SC + 0.019 * M_ST + 0.018 * F_ST + 0.155 * M_LIT + 0.145 * F_L IT + 0.155 * M_ILL + 0.158 * F_ILL + 0.154 * TOT_WORK_M + 0.143 * TOT_WORK_F + 0.142 * MAINWORK_M + 0.126 * MAINWORK_F + 0.112 * MAIN_CL_M + 0.083 * MAIN_CL_F + 0.119 * MAIN_AL_M + 0.09 * MAIN_AL_F + 0.142 * MAIN_HH_M + 0. 134 * MAIN_HH_F + 0.123 * MAIN_OT_M + 0.117 * MAIN_OT_F + 0.157 * MARGWO RK_M + 0.149 * MARGWORK_F + 0.088 * MARG_CL_M + 0.065 * MARG_CL_F + 0.1 27 * MARG_AL_M + 0.116 * MARG_AL_F + 0.145 * MARG_HH_M + 0.142 * MARG_ HH_F + 0.151 * MARG_OT_M + 0.148 * MARG_OT_F + 0.158 * MARGWORK_3_6_M + 0.156 * MARGWORK_3_6_F + 0.158 * MARG_CL_3_6_M + 0.15 * MARG_CL_3_6_F + 0.095 * MARG_AL_3_6_M + 0.067 * MARG_AL_3_6_F + 0.128 * MARG_HH_3_6_M + 0.114 * MARG_HH_3_6_F + 0.145 * MARG_OT_3_6_M + 0.141 * MARG_OT_3_6_F + 0.151 * MARGWORK_0_3_M + 0.148 * MARGWORK_0_3_F + 0.143 * MARG_CL_0_ 3_M + 0.134 * MARG_CL_0_3_F + 0.063 * MARG_AL_0_3_M + 0.057 * MARG_AL_0_ 3_F + 0.119 * MARG_HH_0_3_M + 0.113 * MARG_HH_0_3_F + 0.142 * MARG_OT_0_ 3_M + 0.141 * MARG_OT_0_3_F + 0.148 * NON_WORK_M + 0.142 * NON_WORK_F +

2.3.11 Extracting factor loadings

	0	1	2	3	4
No_HH	0.149222	-0.115487	0.101528	0.076814	-0.012090
TOT_M	0.159169	-0.080239	-0.038662	0.052976	-0.042344
TOT_F	0.158209	-0.093718	0.028959	0.070022	-0.022927
M_06	0.156340	-0.020341	-0.074419	0.028520	-0.080339
F_06	0.156814	-0.014310	-0.068223	0.016398	-0.078326
M_SC	0.143350	-0.079667	-0.037619	0.010210	-0.167893
F_SC	0.143537	-0.087098	0.021350	0.016244	-0.158092
M_ST	0.018849	0.069101	0.323827	0.091143	0.418412
F_ST	0.017878	0.067316	0.338705	0.079554	0.415965
M_LIT	0.155152	-0.105986	-0.032107	0.089187	-0.014033
F_LIT	0.145450	-0.133234	-0.005133	0.125412	0.029084
M_ILL	0.154551	-0.009460	-0.047054	-0.034665	-0.104073
F_ILL	0.158283	-0.021793	0.079345	-0.010578	-0.110332
TOT_WORK_M	0.154076	-0.120912	-0.001116	0.069046	-0.023104
TOT_WORK_F	0.142530	-0.076003	0.194130	0.111057	-0.018931
MAINWORK_M	0.141932	-0.166700	0.019821	0.100188	-0.043225
MAINWORK_F	0.125732	-0.142250	0.209976	0.133013	-0.054674
MAIN_CL_M	0.111692	0.042552	0.033131	0.078851	-0.303376
MAIN_CL_F	0.083035	0.095893	0.188822	0.265022	-0.257925
MAIN_AL_M	0.119291	-0.053342	0.225831	-0.121379	-0.253131
MAIN_AL_F	0.090089	-0.072467	0.356566	-0.020989	-0.199220

Table 22 – Factor Loadings

2.3.12 Creating a dataframe with the coefficients of all PCs

	PC1	PC2	PC3	PC4	PC5
No_HH	0.149222	-0.115487	0.101528	0.076814	-0.012090
TOT_M	0.159169	-0.080239	-0.038662	0.052976	-0.042344
TOT_F	0.158209	-0.093718	0.028959	0.070022	-0.022927
M_06	0.156340	-0.020341	-0.074419	0.028520	-0.080339
F_06	0.156814	-0.014310	-0.068223	0.016398	-0.078326
M_SC	0.143350	-0.079667	-0.037619	0.010210	-0.167893
F_SC	0.143537	-0.087098	0.021350	0.016244	-0.158092
M_ST	0.018849	0.069101	0.323827	0.091143	0.418412
F_ST	0.017878	0.067316	0.338705	0.079554	0.415965
M_LIT	0.155152	-0.105986	-0.032107	0.089187	-0.014033
F_LIT	0.145450	-0.133234	-0.005133	0.125412	0.029084
M_ILL	0.154551	-0.009460	-0.047054	-0.034665	-0.104073
F_ILL	0.158283	-0.021793	0.079345	-0.010578	-0.110332
TOT_WORK_M	0.154076	-0.120912	-0.001116	0.069046	-0.023104
TOT_WORK_F	0.142530	-0.076003	0.194130	0.111057	-0.018931
MAINWORK_M	0.141932	-0.166700	0.019821	0.100188	-0.043225
MAINWORK_F	0.125732	-0.142250	0.209976	0.133013	-0.054674
MAIN_CL_M	0.111692	0.042552	0.033131	0.078851	-0.303376
MAIN_CL_F	0.083035	0.095893	0.188822	0.265022	-0.257925
MAIN_AL_M	0.119291	-0.053342	0.225831	-0.121379	-0.253131
MAIN_AL_F	0.090089	-0.072467	0.356566	-0.020989	-0.199220
MAIN_HH_M	0.141850	-0.101835	-0.102202	-0.021969	-0.060812

Table 23 – Dataframe with coefficients of PCs

Inferences:

- Let's identify which features have maximum loading across the components.
- We will first plot the component loading on a heatmap.
- For each feature, we find the maximum loading value across the components and mark the same with help of rectangular box.
- Features marked with rectangular red box are the one having maximum loading on the respective component. We consider these marked features to decide the context that the component represents.

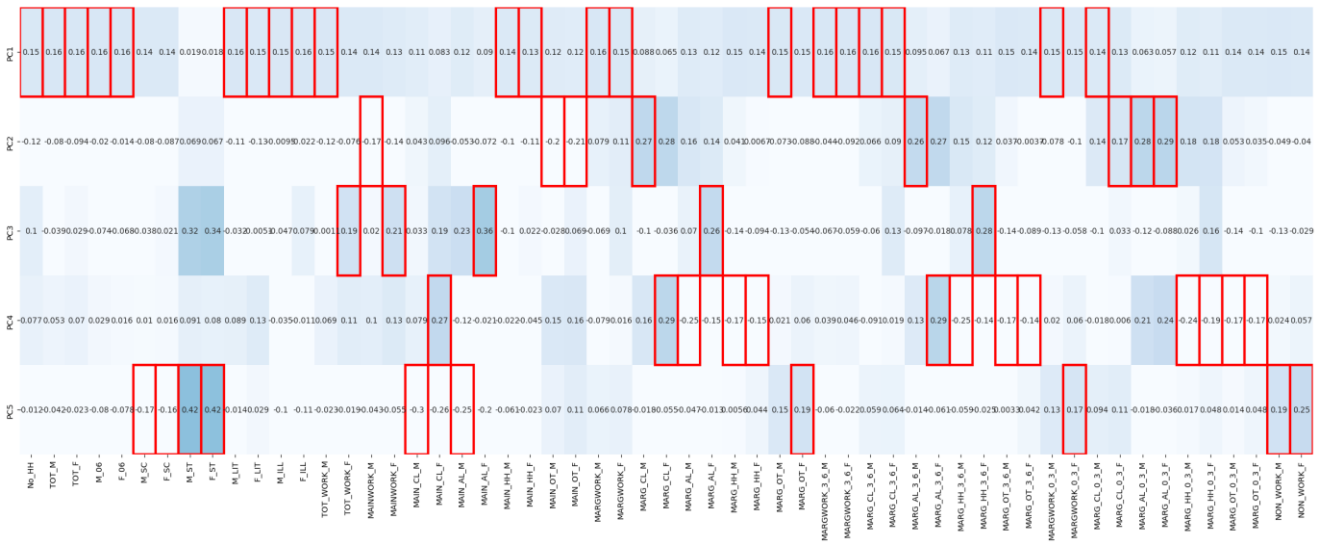


Figure 20 – Heatmap of PCs vs Variables

Inferences:

- The first Principal component, PC1, is a measure of Total Males, Total Females, Males and Females under 6 years, Literate males and Females, Illiterate Males and Females, Total Working males, Male and Female Marginal Workers, Marginal Workers Male and Female of 3-6 years, Marginal Cultivator Population of 3-6 years.
- The second principal component, PC2, is a measure of Marginal Agricultural Labourers Male and Female from 0-3 years.
- The third principal component, PC3, is a measure of Female Main Agricultural Labourers.
- The fourth principal component, PC4, is a measure of Female Marginal Cultivation Labourers and Female Marginal Agricultural labourers from 3-6 years.
- The fifth principal component, PC5, is a measure of Male and Female Scheduled Tribes.

Principal Component	Columns having maximum loading on the respective component	Principal Component Name
PC1	TOT_M, TOT_F, M_06, M_06	pc_total_males_and_females
PC2	MARG_AL_0_3_M, MARG_AL_0_3_F	pc_male_female_marginal_al_0_3
PC3	MAIN_AL_F	pc_female_main_al
PC4	MARG_CL_F, MARG_AL_3_6_F	pc_female_marginal_cl_al_3_6
PC5	M_ST, F_ST	pc_male_female_st

Table 24 – Principal Components

	State	Area Name	pc_total_males_and_females	pc_male_female_marginal_al_0_3	pc_female_main_al	pc_female_marginal_cl_al_3_6	pc_male_female_st
0	Jammu & Kashmir	Kupwara	-5.53	0.43	-1.47	-1.28	0.38
1	Jammu & Kashmir	Badgam	-5.49	-0.11	-2.02	-1.75	-0.01
2	Jammu & Kashmir	Leh(Ladakh)	-7.47	-0.22	-0.25	0.01	0.56
3	Jammu & Kashmir	Kargil	-7.92	-0.65	-0.66	-0.74	0.27
4	Jammu & Kashmir	Punch	-5.18	2.30	-1.16	1.06	1.08
5	Jammu & Kashmir	Rajouri	-3.65	4.60	-1.74	3.30	1.21
6	Jammu & Kashmir	Kathua	-6.18	-0.26	-1.23	-0.12	-0.17

Table 25 – PCs in terms of actual variables

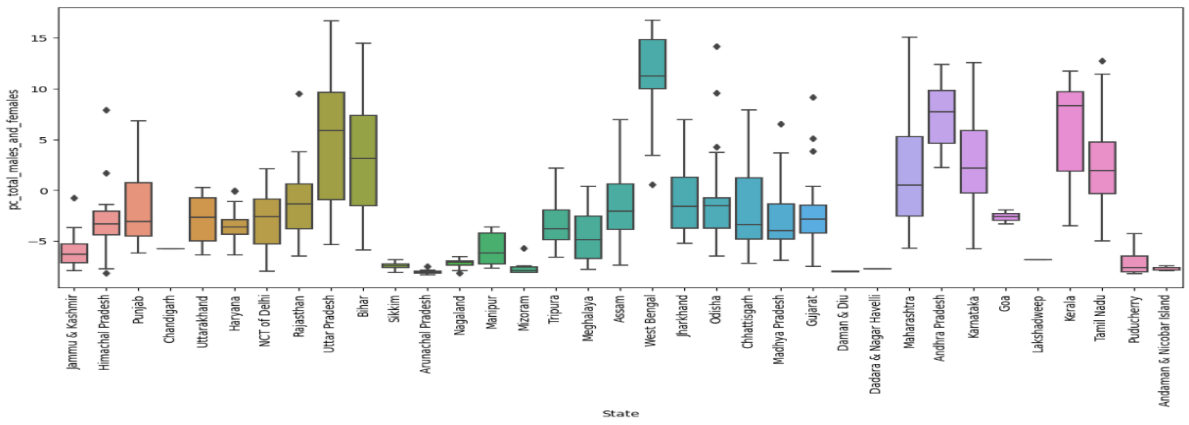


Figure 21 – Boxplot of State vs Total Males and Females

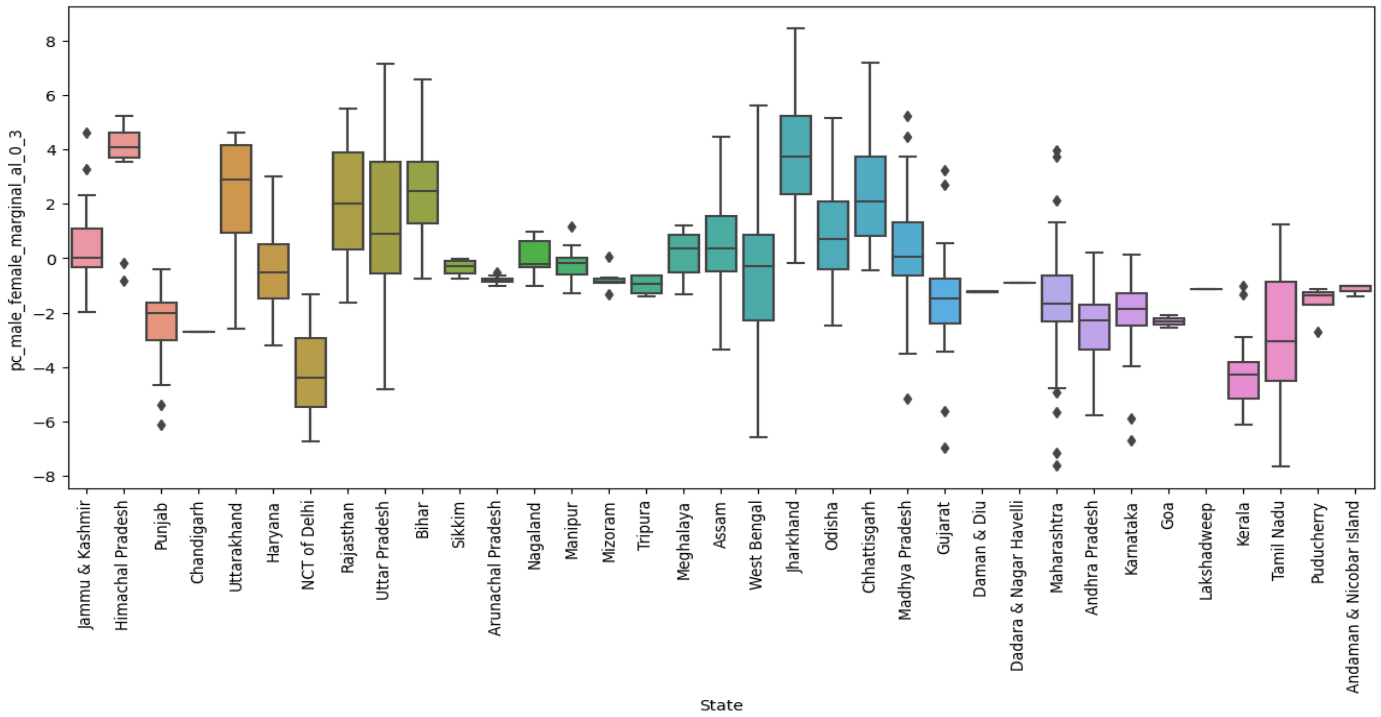


Figure 22- Boxplot of State vs Marginal Agricultural Labourers

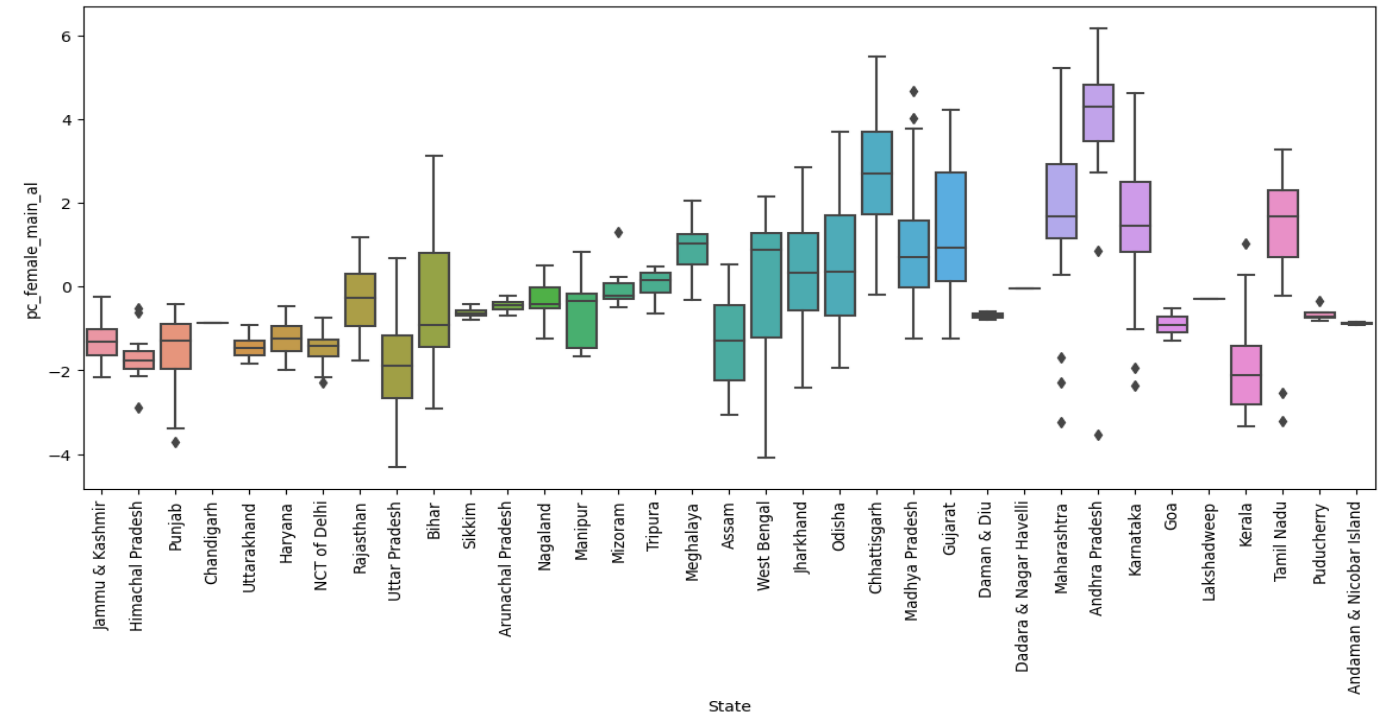


Figure 23 – Boxplot of State vs Main Female Agricultural Labourers

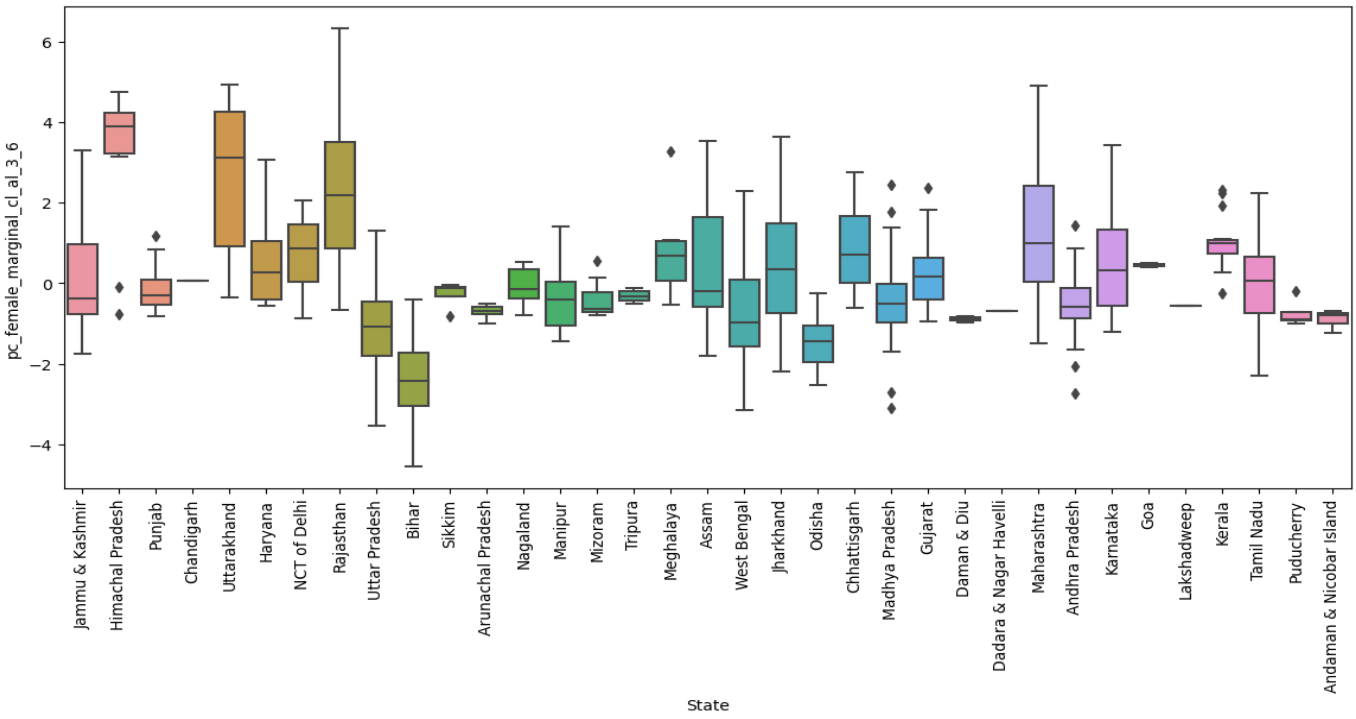


Figure 24 – Boxplot of State vs Female Marginal Cultivator and Agricultural Labourers from 3-6 years

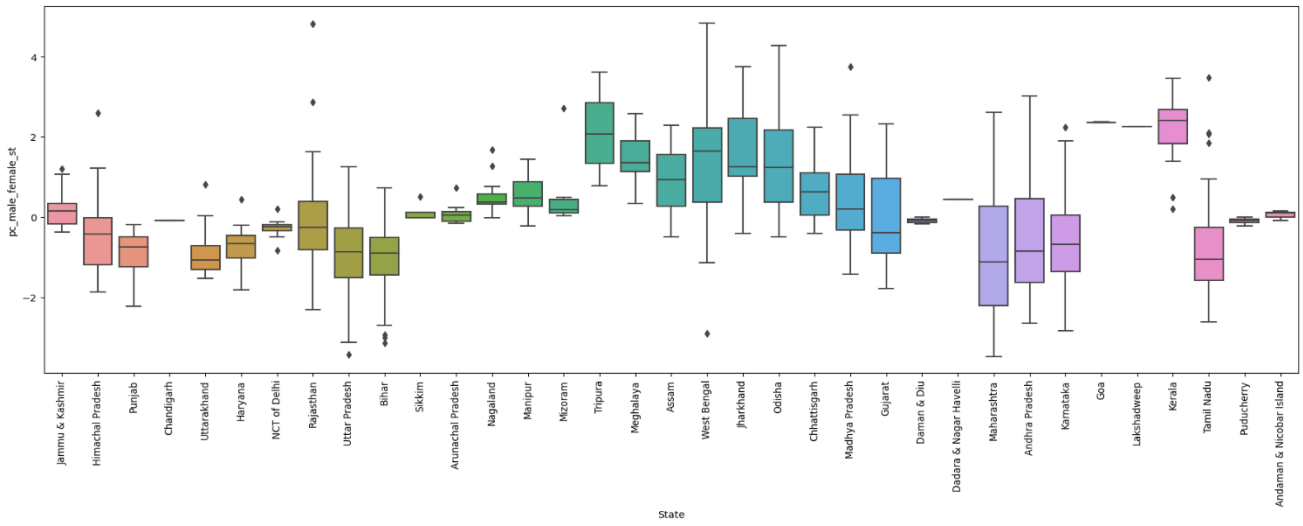


Figure 25 – Boxplot of State vs Male and Female ST

Inferences:

1) Total males and females is maximum in the state of West Bengal and minimum in Arunachal Pradesh.

2) Marginal Agricultural Labourers are the maximum in Jharkhand while the least in NCT of Delhi.
3) Main Female Agricultural Labourers are the least in Kerala and the highest in Andhra Pradesh.
4) Female Marginal Cultivator and Agricultural Labourers from 3-6 years is maximum in Uttarakhand and Himachal Pradesh and the least in Bihar.
5) Male and Female Scheduled Tribes are maximum in Tripura and the lowest in Maharashtra

****