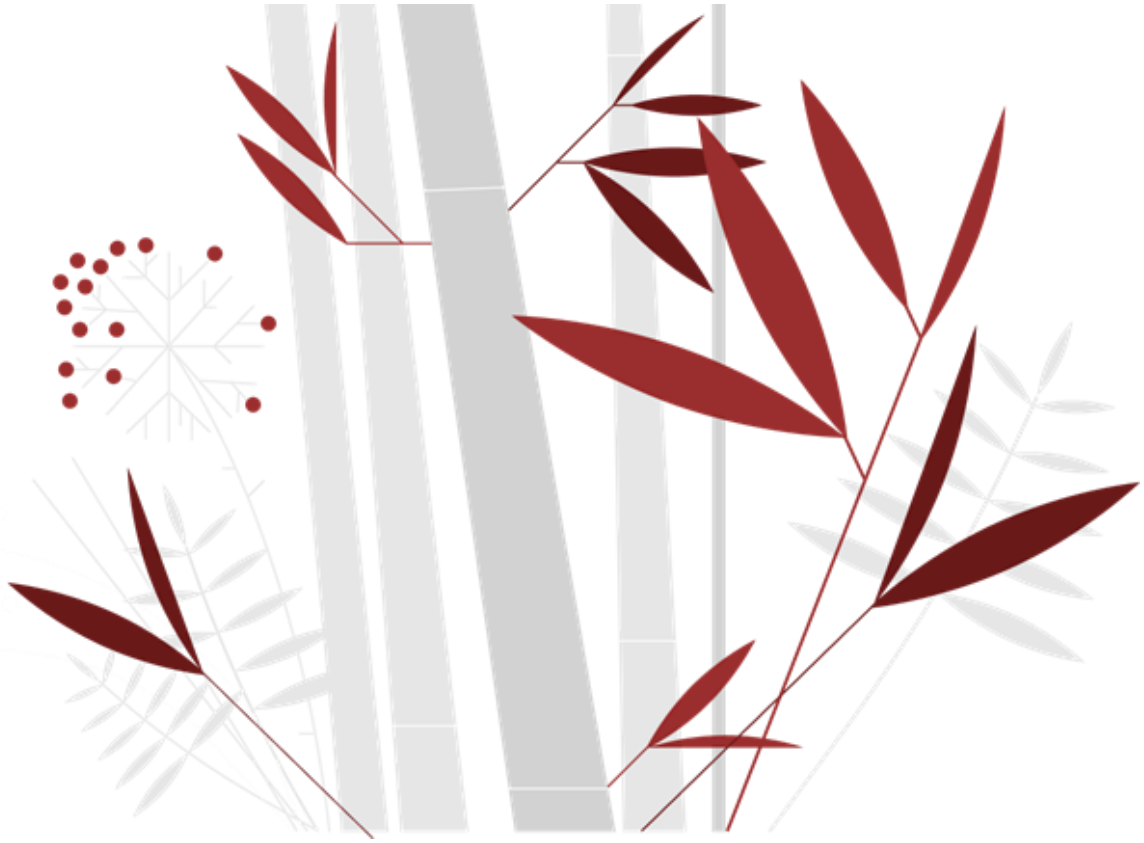
MACHINE LEARNING-1 (ML-1) PROJECT-CODED

BY

Harsh Patel

5th May 2024



Problem-1: ADS 24x7 Company

An Ads24x7 Digital Marketing company has $10 millions seed funding and is expanding its wings into market Marketing Analytics. They collected data from their Marketing Intelligence team and now wants us (their newly appointed data analyst) to segment type of ads based on the features provided. [Note: Use Clustering procedure to segment ads into homogeneous groups]

Data Variables:

Timestamp	The Timestamp of the particular Advertisement.
InventoryType	The Inventory Type of the particular Advertisement. Format 1 to 7. This is a Categorical Variable.
Ad - Length	The Length Dimension of the particular Adverstisement.
Ad- Width	The Width Dimension of the particular Advertisement.
Ad Size	The Overall Size of the particular Advertisement. Length*Width.
Ad Type	The type of the particular Advertisement. This is a Categorical Variable.
Platform	The platform in which the particular Advertisement is displayed. Web, Video or App. This is a Categorical Variable.
Device Type	The type of the device which supports the partciular Advertisement. This is a Categorical Variable.
Format	The Format in which the Advertisement is displayed. This is a Categorical Variable.
Available_Impressions	How often the particular Advertisement is shown. An impression is counted each time an Advertisement is shown on a search result page or other site on a Network.
Matched_Queries	Matched search queries data is pulled from Advertising Platform and consists of the exact searches typed into the search Engine that generated clicks for the particular Advertisement.
Impressions	The impression count of the particular Advertisement out of the total available impressions.
Clicks	It is a marketing metric that counts the number of times users have clicked on the particular advertisement to reach an online property.
Spend	It is the amount of money spent on specific ad variations within a specific campaign or ad set. This metric helps regulate ad performance.
Fee	The percentage of the Advertising Fees payable by Franchise Entities.
Revenue	It is the income that has been earned from the particular advertisement.
CTR	CTR stands for "Click through rate". CTR is the number of clicks that your ad receives divided by the number of times your ad is shown. Formula used here is CTR = Total Measured Clicks / Total Measured Ad Impressions x 100. Note that the Total Measured Clicks refers to the 'Clicks' Column and the Total Measured Ad Impressions refers to the 'Impressions' Column.
CPM	CPM stands for "cost per 1000 impressions." Formula used here is CPM = (Total Campaign Spend / Number of Impressions) * 1,000. Note that the Total Campaign Spend refers to the 'Spend' Column and the Number of Impressions refers to the 'Impressions' Column.
CPC	CPC stands for "Cost-per-click". Cost-per-click (CPC) bidding means that you pay for each click on your ads. The Formula used here is CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column and the Number of Clicks refers to the 'Clicks' Column.

Table-1 Data Variables

1.1 Data Pre-Processing and EDA:

First, we will look at first and last five rows using function head and tail respectively, of the dataset from excel file called Clustering clean ads data that we loaded using read excel function. In fig-1 and fig-2 shows below shows the dataset.

	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	R
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0	0.35	
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.0	0.35	
2	2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0	0.35	
3	2020-9-3-20	Format1	300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.0	0.35	
4	2020-9-4-15	Format1	300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.0	0.35	

fig-1 first five rows

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

fig-2 last five rows

Now, we use shape function the dataset and we get that there are 23066 row and 19 columns. Then, we use info function and found out the data type of each column and used value counts functions on the categorical variables as shown in the below table. we will check for the duplicated rows are present or not using duplicate function and found out that in the dataset there are zero same or duplicated rows present.

```
RangeIndex: 23066 entries, 0 to 23065
Data columns (total 19 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Timestamp              23066 non-null  object
 1   InventoryType          23066 non-null  object
 2   Ad - Length            23066 non-null  int64
 3   Ad- Width              23066 non-null  int64
 4   Ad Size                23066 non-null  int64
 5   Ad Type                23066 non-null  object
 6   Platform               23066 non-null  object
 7   Device Type            23066 non-null  object
 8   Format                 23066 non-null  object
 9   Available_Impressions  23066 non-null  int64
 10  Matched_Queries        23066 non-null  int64
 11  Impressions            23066 non-null  int64
 12  Clicks                 23066 non-null  int64
 13  Spend                  23066 non-null  float64
 14  Fee                    23066 non-null  float64
 15  Revenue                23066 non-null  float64
 16  CTR                    18330 non-null  float64
 17  CPM                    18330 non-null  float64
 18  CPC                    18330 non-null  float64
dtypes: float64(6), int64(7), object(6)
```

Table-2 Information table

We also checked the value counts for the categorical variables and can be seen in the below table.

```
InventoryType
Format4    7165
Format5    4249
Format1    3814
Format3    3540
Format6    1850
Format2    1789
Format7     659
Name: count, dtype: int64
Ad Type
Inter224    1658
Inter217    1655
Inter223    1654
Inter219    1650
Inter221    1650
Inter222    1649
Inter229    1648
Inter227    1647
Inter218    1645
inter230    1644
Inter220    1644
Inter225    1643
Inter226    1640
Inter228    1639
Name: count, dtype: int64
Platform
Video    9873
Web      8251
App      4942
Name: count, dtype: int64

Device Type
Mobile     14806
Desktop     8260
Name: count, dtype: int64
Format
Video      11552
Display    11514
Name: count, dtype: int64
```

Table-3 Value counts of Categorical variables.

We found that there are some missing values in the last three columns namely CTR, CPC and CPM. Now, we will use isnull function and found that there are 4736 numbers of empty variables present in the CTR, CPM, CPC variables equally.

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

Table-4 Missing values present before Treatment

We have got the formulas for these three variables as shown below:

CPM = (Total Campaign Spend / Number of Impressions) * 1,000.

CPC = Total Cost (spend) / Number of Clicks.

CTR = Total Measured Clicks / Total Measured Ad Impressions * 100.

Using these three formulae we replaced all the missing data or null values from all three variables and again check for the any null values using info and isnull function and found that we have successfully treated the empty values as shown in below tables.

```
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
```

Tabel-5 No missing values after treatment

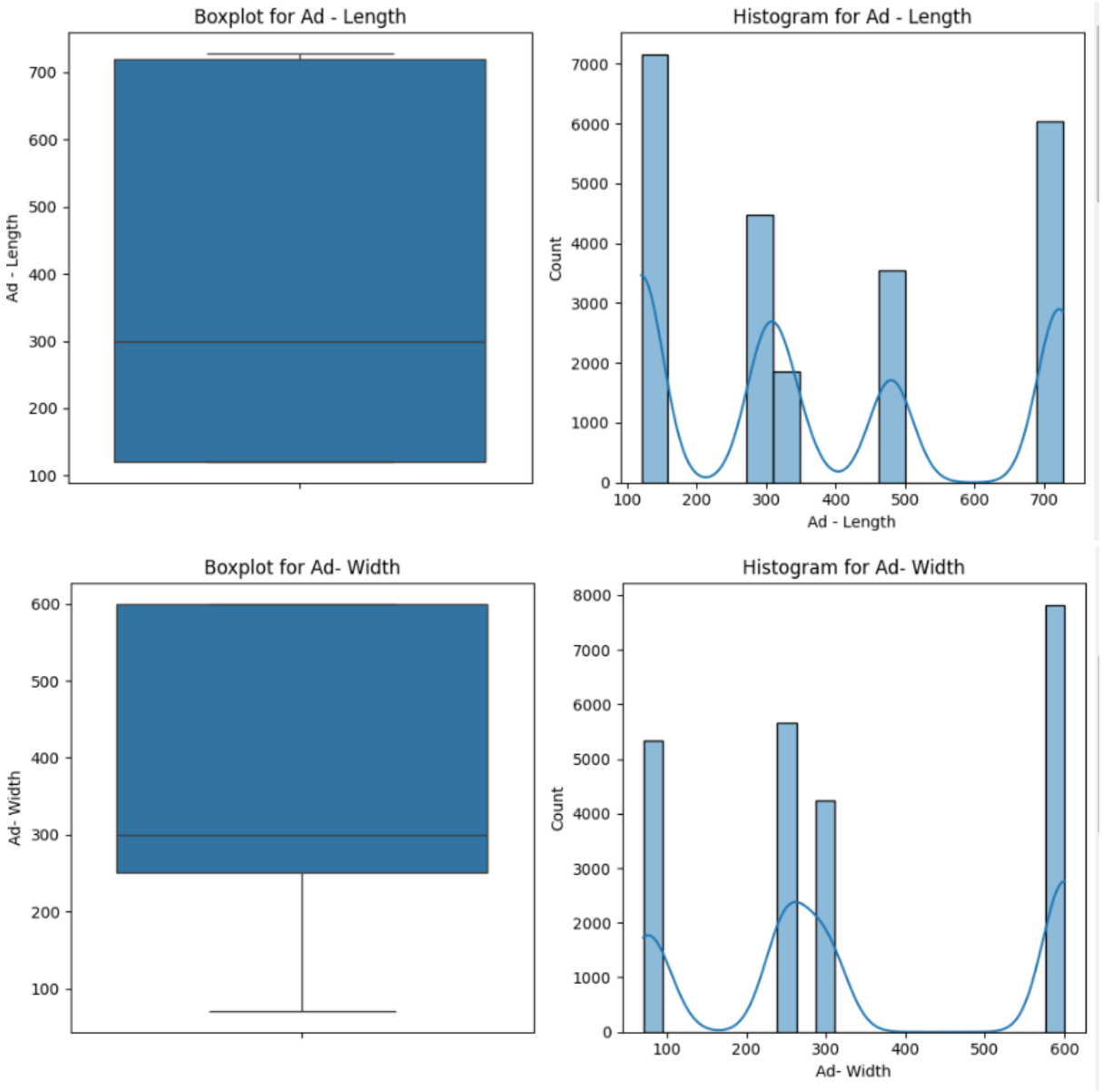
Now, we will use describe function on this dataset and obtain the min, max, std, count, mean, q1, q2, q3 values for all the numerical columns as shown in the below table. We can observe that for some the numerical variables the min, max values are too larger i.e. they are far apart from each other indicating presence of outliers in them as well as there is variation between variables so we will need to scale them.

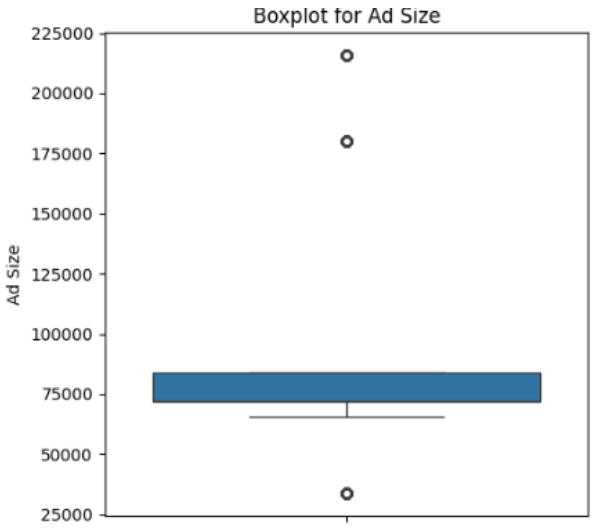
	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	3.851631e+02	2.336514e+02	120.000000	120.000000	300.000000	7.200000e+02	7.280000e+02
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.000000	250.000000	300.000000	6.000000e+02	6.000000e+02
Ad Size	23066.0	9.667447e+04	6.153833e+04	33600.000000	72000.000000	72000.000000	8.400000e+04	2.160000e+05
Available_Impressions	23066.0	2.432044e+06	4.742888e+06	1.000000	33672.250000	483771.000000	2.527712e+06	2.759286e+07
Matched_Queries	23066.0	1.295099e+06	2.512970e+06	1.000000	18282.500000	258087.500000	1.180700e+06	1.470202e+07
Impressions	23066.0	1.241520e+06	2.429400e+06	1.000000	7990.500000	225290.000000	1.112428e+06	1.419477e+07
Clicks	23066.0	1.067852e+04	1.735341e+04	1.000000	710.000000	4425.000000	1.279375e+04	1.430490e+05
Spend	23066.0	2.706626e+03	4.067927e+03	0.000000	85.180000	1425.125000	3.121400e+03	2.693187e+04
Fee	23066.0	3.351231e-01	3.196322e-02	0.210000	0.330000	0.350000	3.500000e-01	3.500000e-01
Revenue	23066.0	1.924252e+03	3.105238e+03	0.000000	55.365375	926.335000	2.091338e+03	2.127618e+04
CTR	23066.0	8.409941e+00	9.262048e+00	0.010874	0.265107	9.391248	1.347057e+01	2.000000e+02
CPM	23066.0	8.396849e+00	9.057760e+00	0.000000	1.749084	8.371566	1.304202e+01	7.150000e+02
CPC	23066.0	3.366776e-01	3.412527e-01	0.000000	0.089736	0.139347	5.462421e-01	7.264000e+00

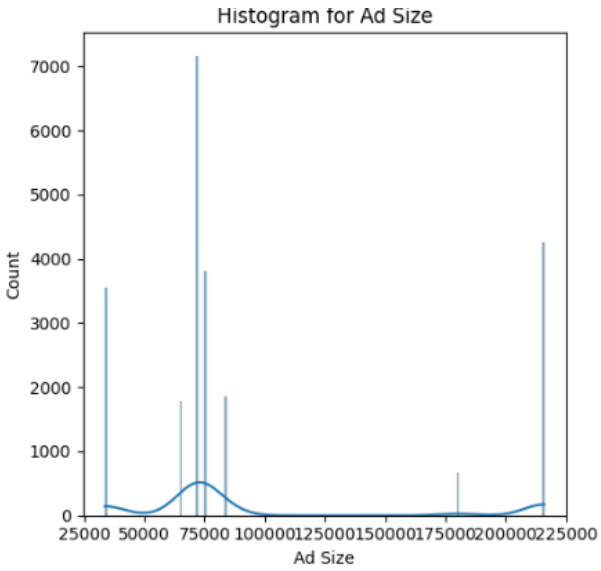
Table-6 Description of Numerical variables

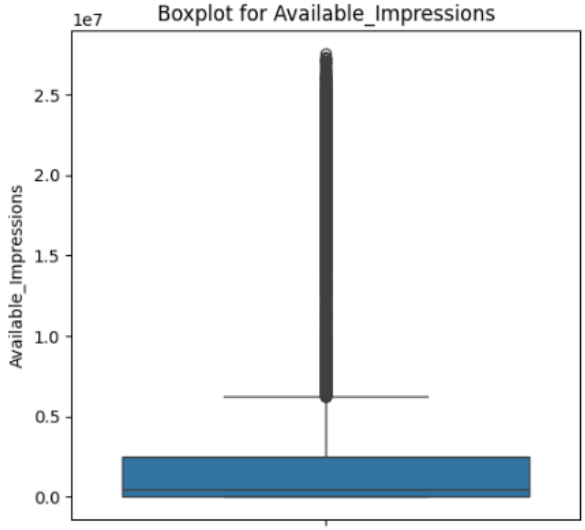
For K-Means clustering we will need to treat the outlier's presence as well scale the data for accurate cluster creation.

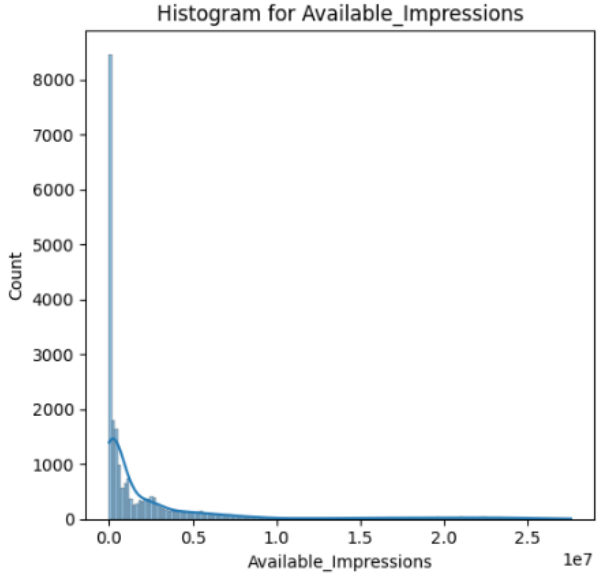
Now, will divide the dataset into dataset for numerical variables and categorical variables, and concatenate them after we have performed EDA on numerical variables as per requirements for clustering. Now, as shown below figure we can see the that we have performed Univariate analysis on the numerical dataset that is unscaled and has presence of outliers.

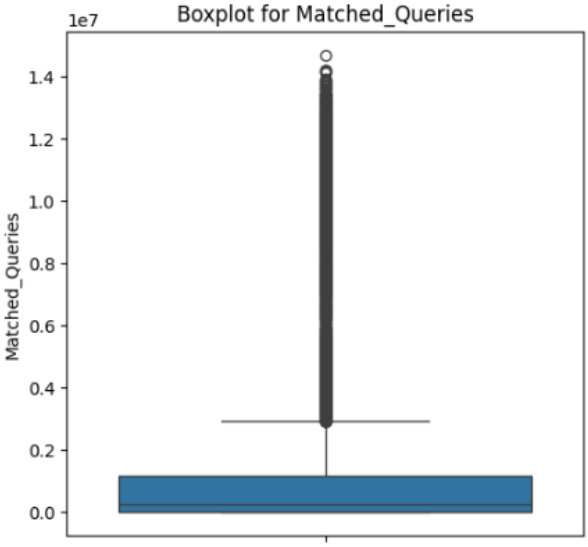


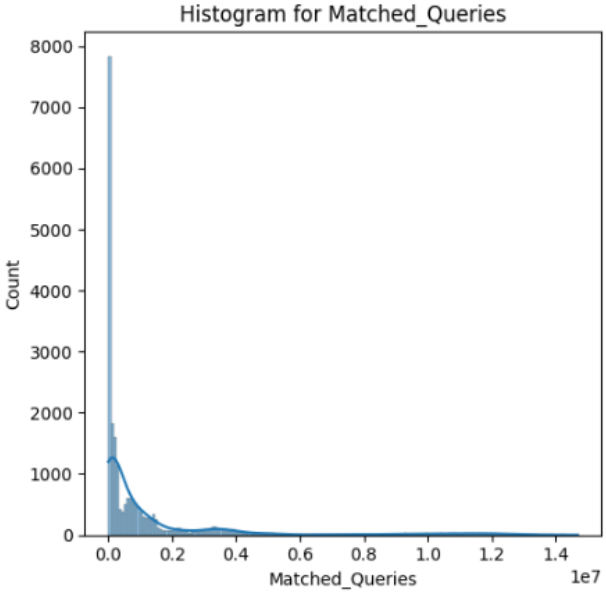

Boxplot for Ad Size
Ad Size
225000
200000
175000
150000
125000
100000
75000
50000
25000

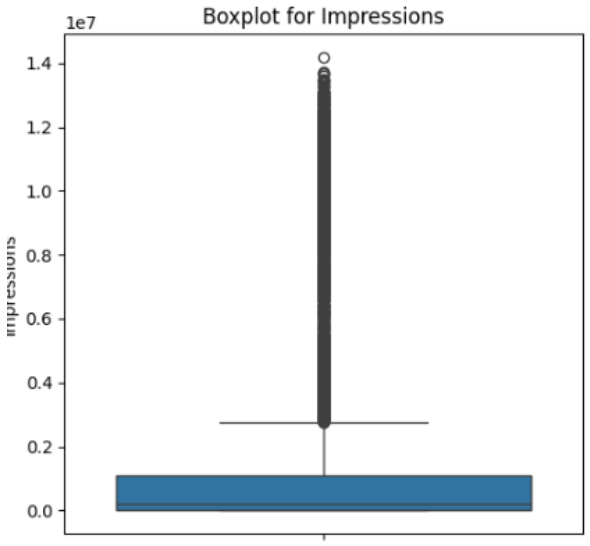

Histogram for Ad Size
Count
7000
6000
5000
4000
3000
2000
1000
0
25000 50000 75000 100000125000150000175000200000225000
Ad Size

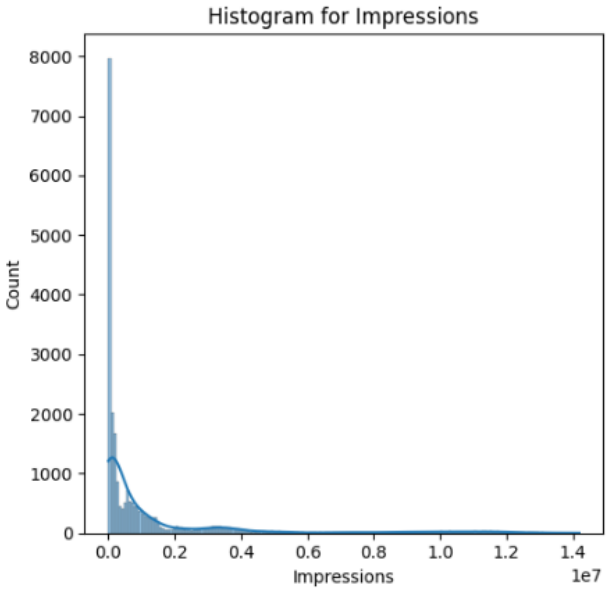

Boxplot for Available_Impressions
1e7
Available_Impressions
2.5
2.0
1.5
1.0
0.5
0.0

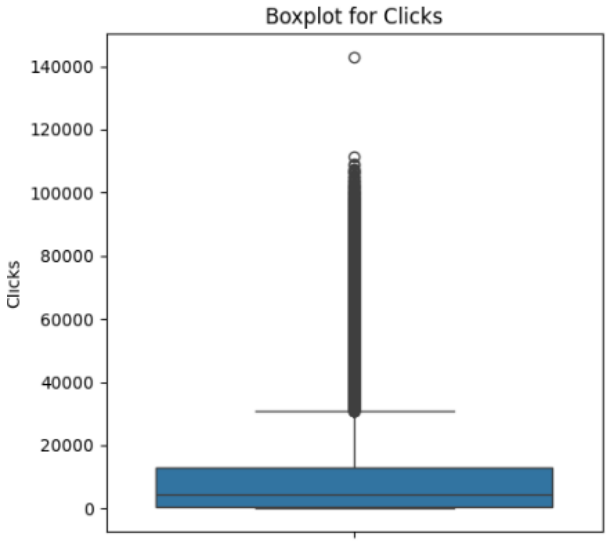

Histogram for Available_Impressions
Count
8000
7000
6000
5000
4000
3000
2000
1000
0
0.0 0.5 1.0 1.5 2.0 2.5
Available_Impressions
1e7

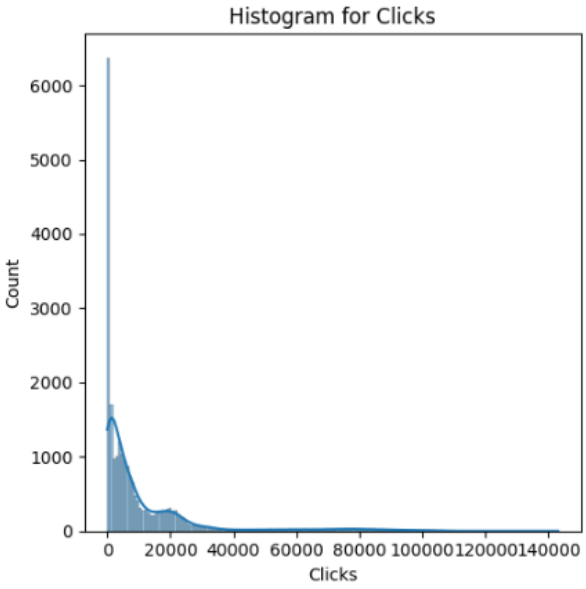

Boxplot for Matched_Queries
1e7
Matched_Queries
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0

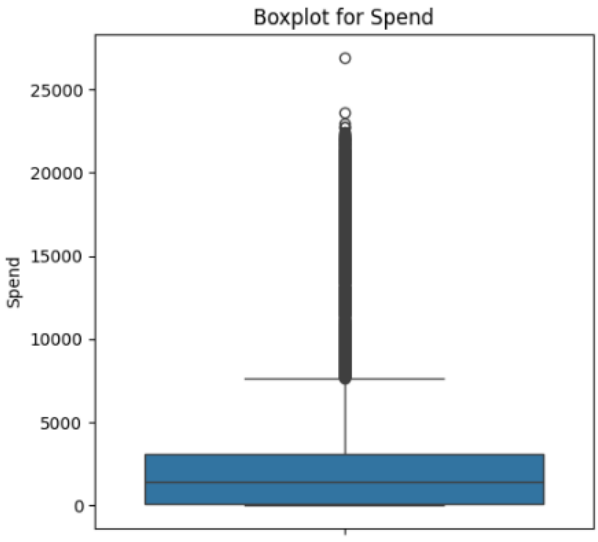

Histogram for Matched_Queries
Count
8000
7000
6000
5000
4000
3000
2000
1000
0
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4
Matched_Queries
1e7

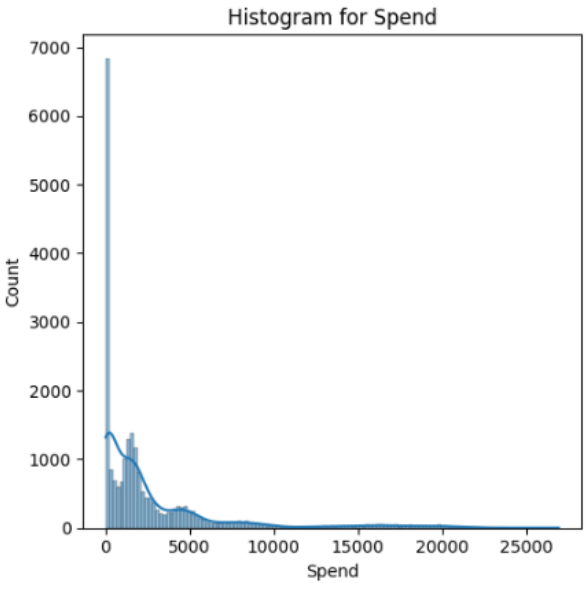

Boxplot for Impressions
1e7
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0

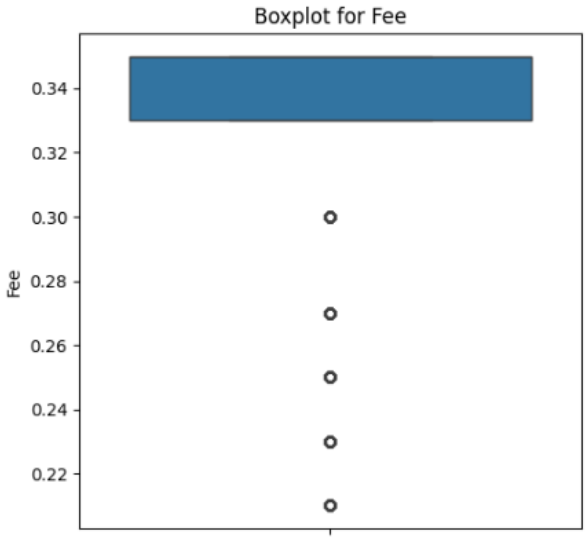

Histogram for Impressions
Count
8000
7000
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Impressions
1e7

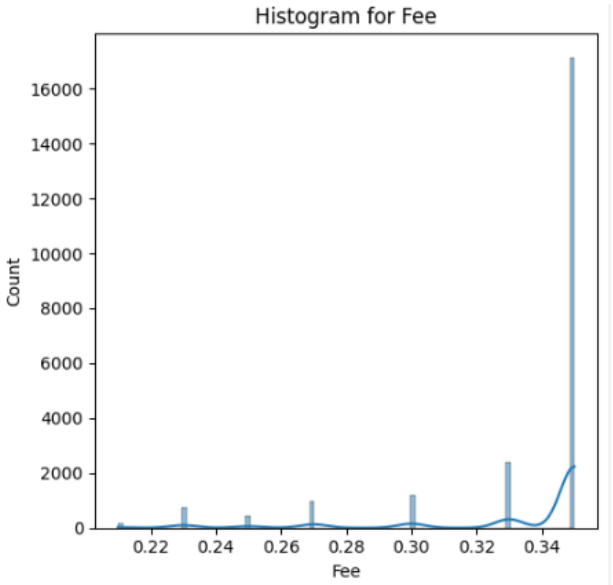

Boxplot for Clicks
Clicks
140000
120000
100000
80000
60000
40000
20000
0

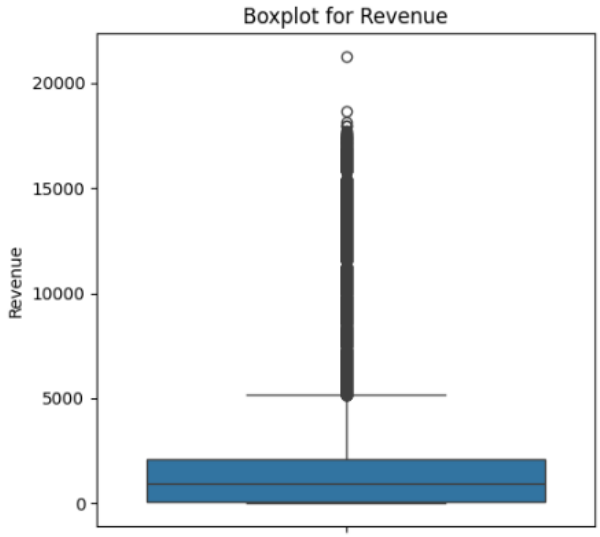

Histogram for Clicks
Count
6000
5000
4000
3000
2000
1000
0
0
20000
40000
60000
80000
100000
120000
140000
Clicks

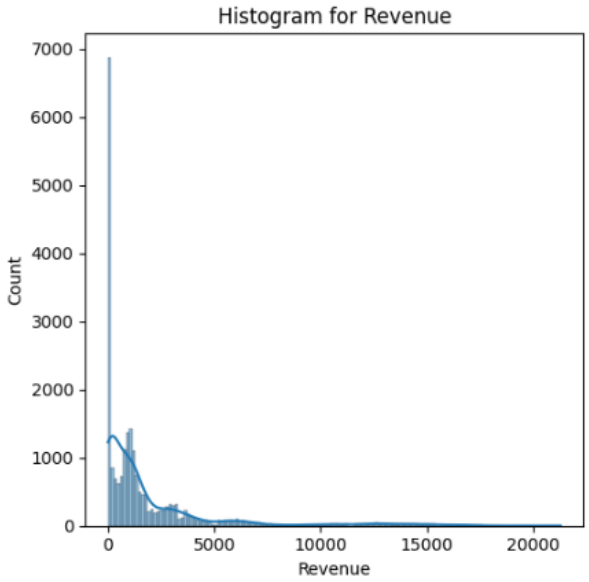

Boxplot for Spend
Spend
25000
20000
15000
10000
5000
0

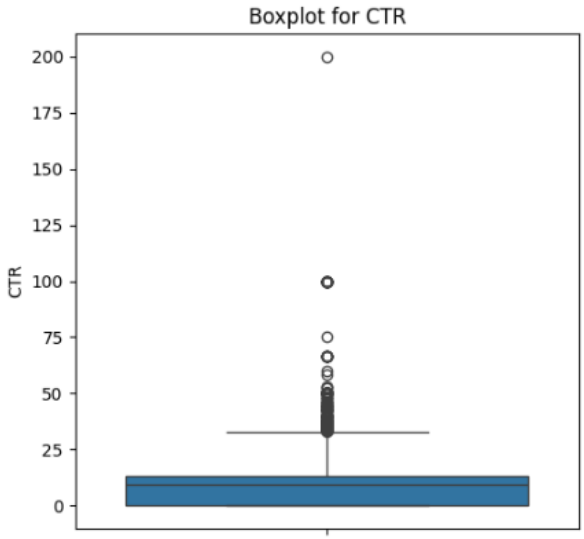

Histogram for Spend
Count
7000
6000
5000
4000
3000
2000
1000
0
0
5000
10000
15000
20000
25000
Spend

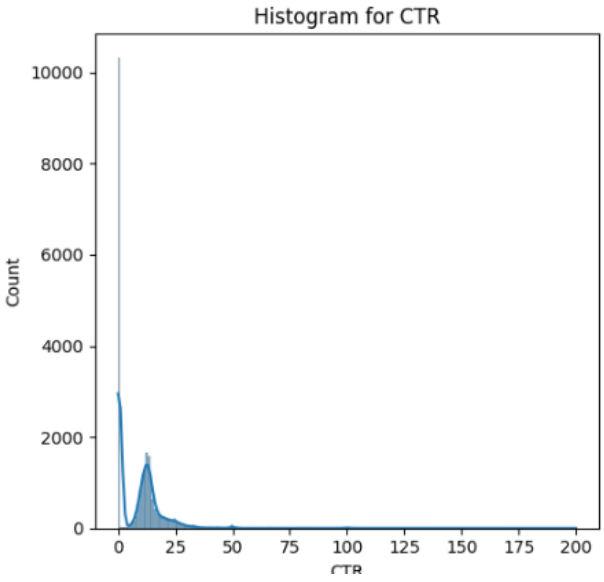

Boxplot for Fee
Fee
0.34
0.32
0.30
0.28
0.26
0.24
0.22


Histogram for Fee
Count
16000
14000
12000
10000
8000
6000
4000
2000
0
0.22
0.24
0.26
0.28
0.30
0.32
0.34
Fee


Boxplot for Revenue
Revenue
20000
15000
10000
5000
0


Histogram for Revenue
Count
7000
6000
5000
4000
3000
2000
1000
0
0
5000
10000
15000
20000
Revenue


Boxplot for CTR
CTR
200
175
150
125
100
75
50
25
0


Histogram for CTR
Count
10000
8000
6000
4000
2000
0
0
25
50
75
100
125
150
175
200

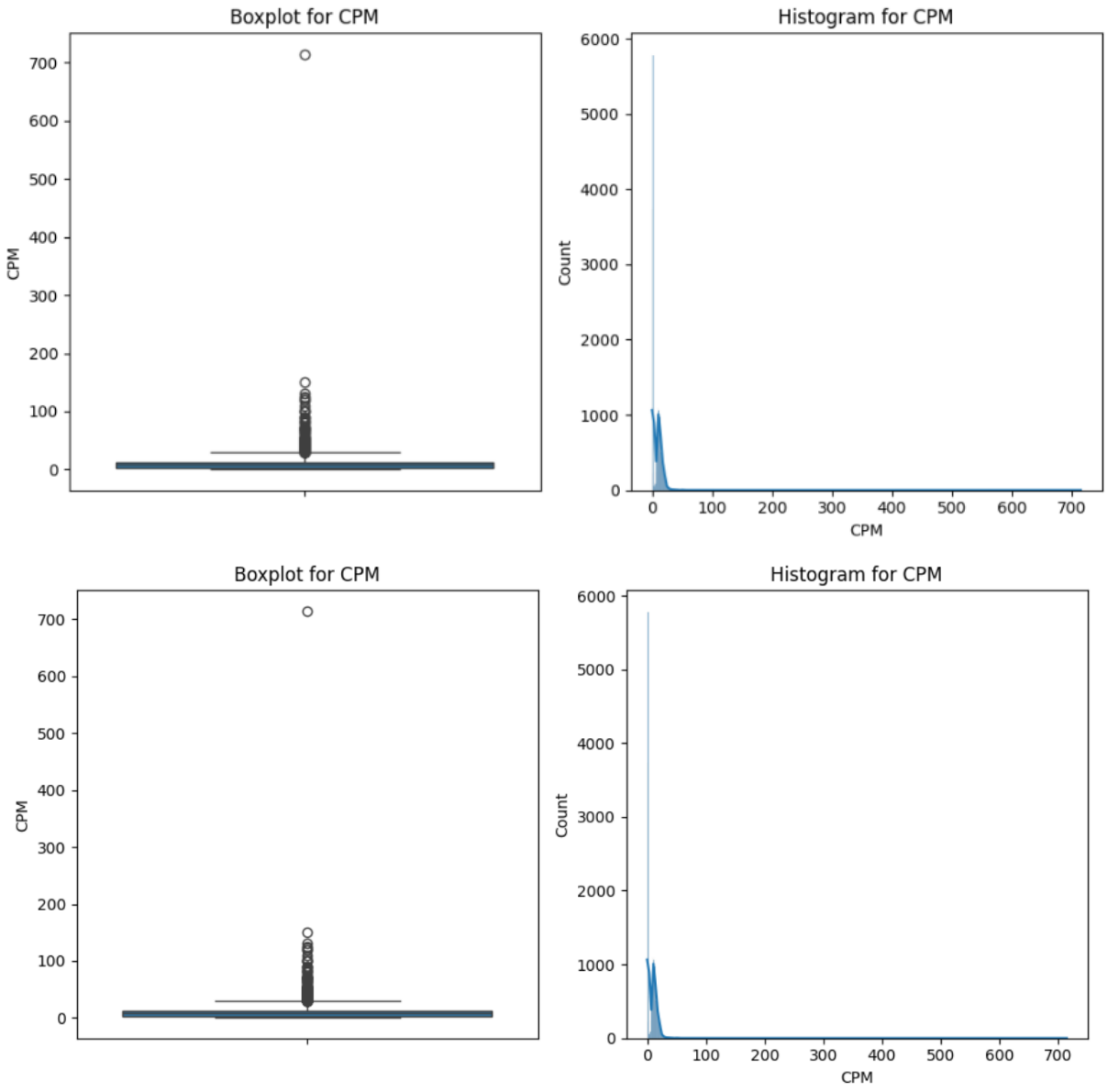
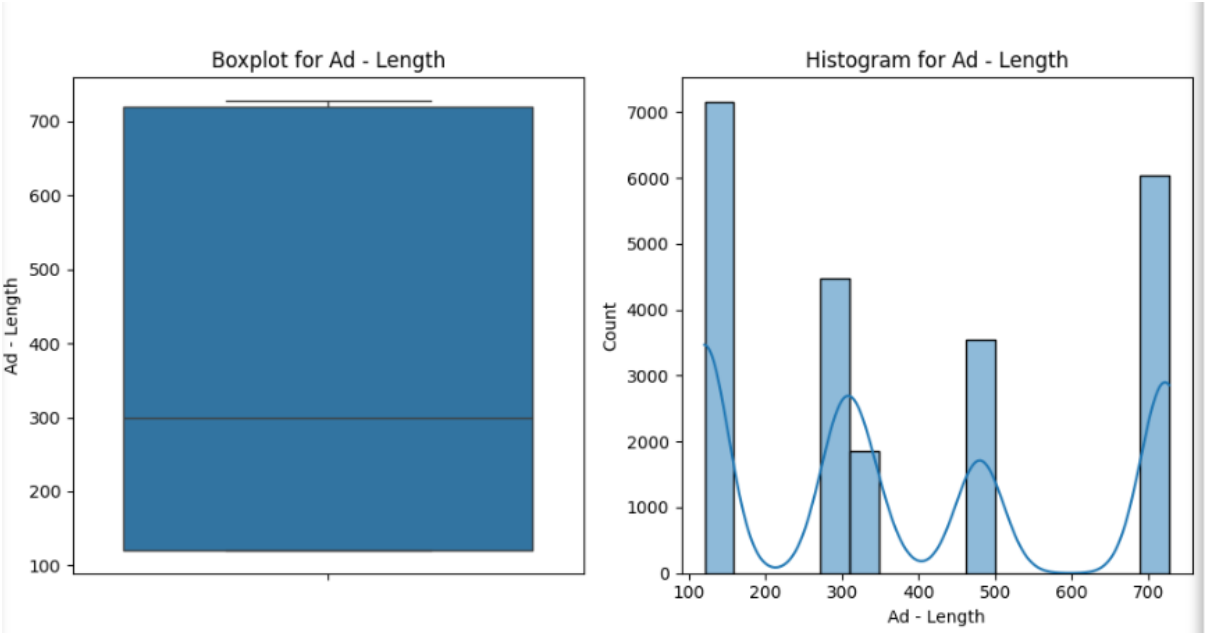
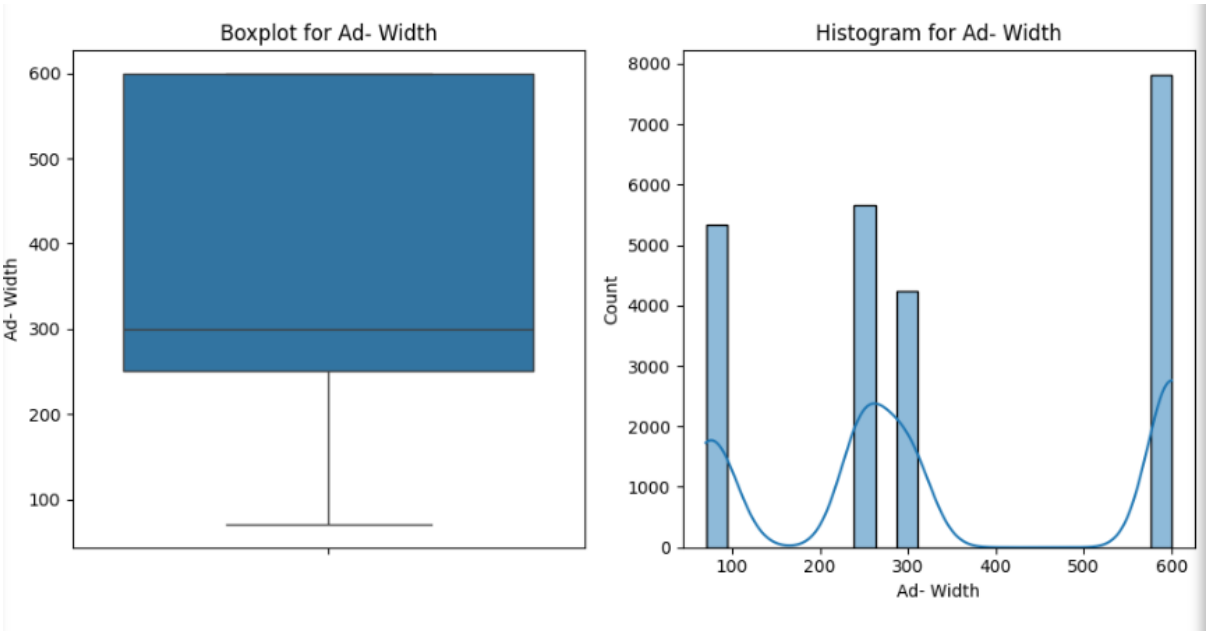


Fig-3 Boxplot Before Outlier Treatment

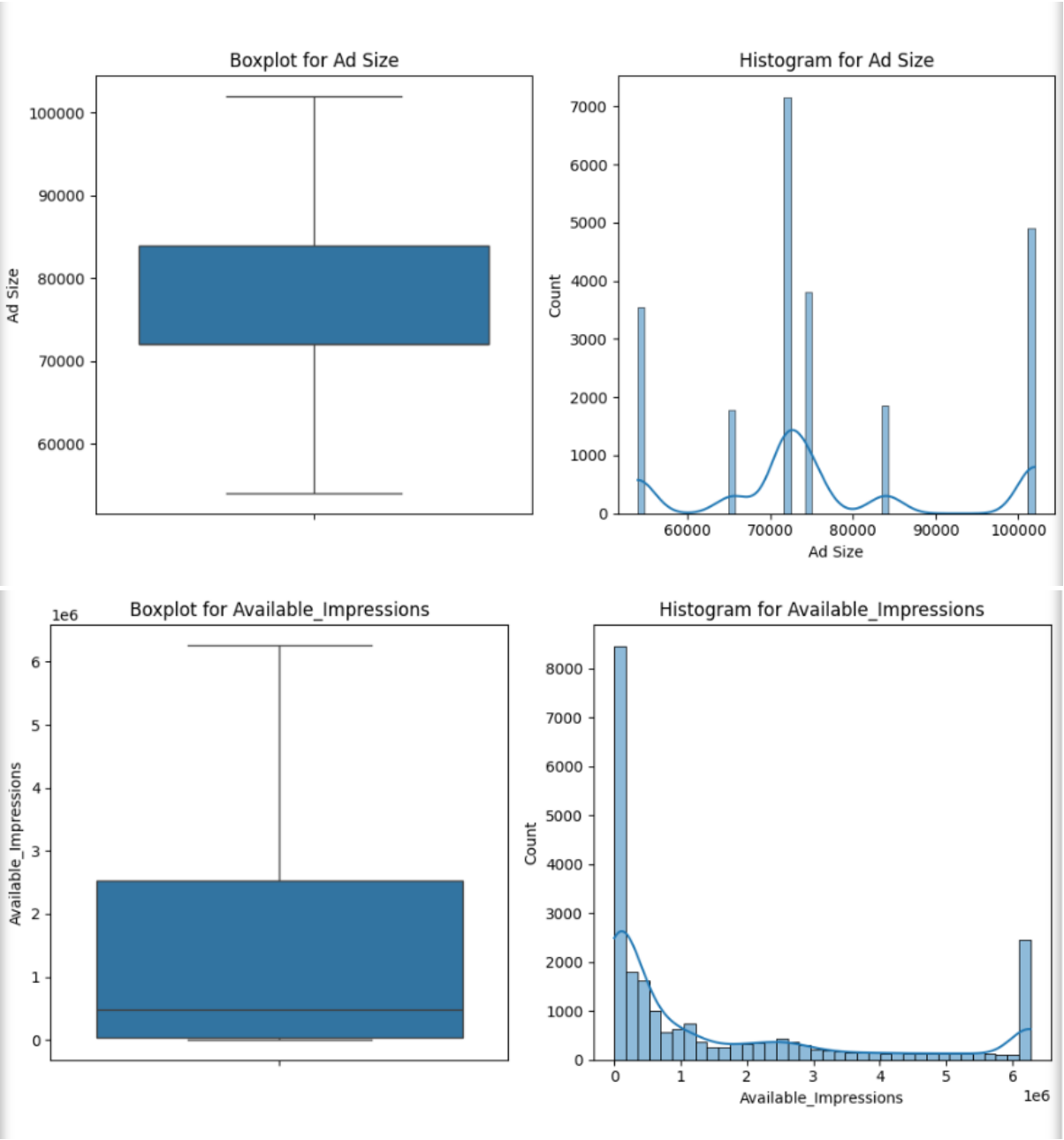
Now, we will perform outlier treatment using q1 and q3 to find IQR and using that to find out the upper and lower limit whiskers and finally bring all those outlier's points to these whiskers and obtain the below boxplots indicated that treatment has been done properly.

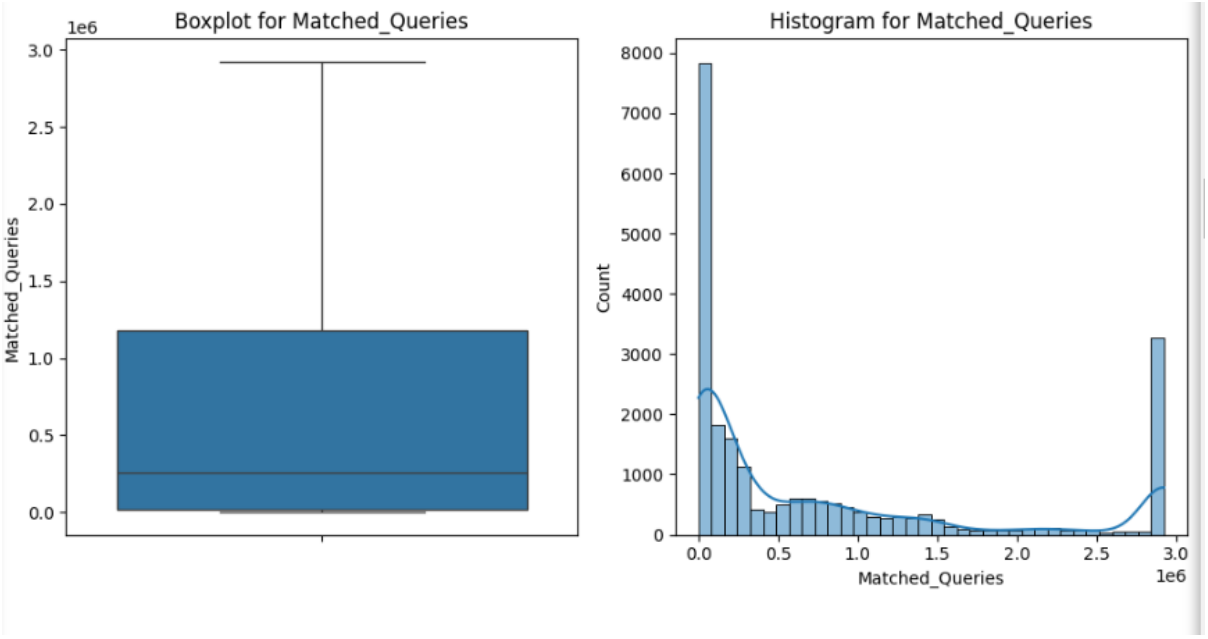

Boxplot for Ad - Length
Ad - Length
700
600
500
400
300
200
100

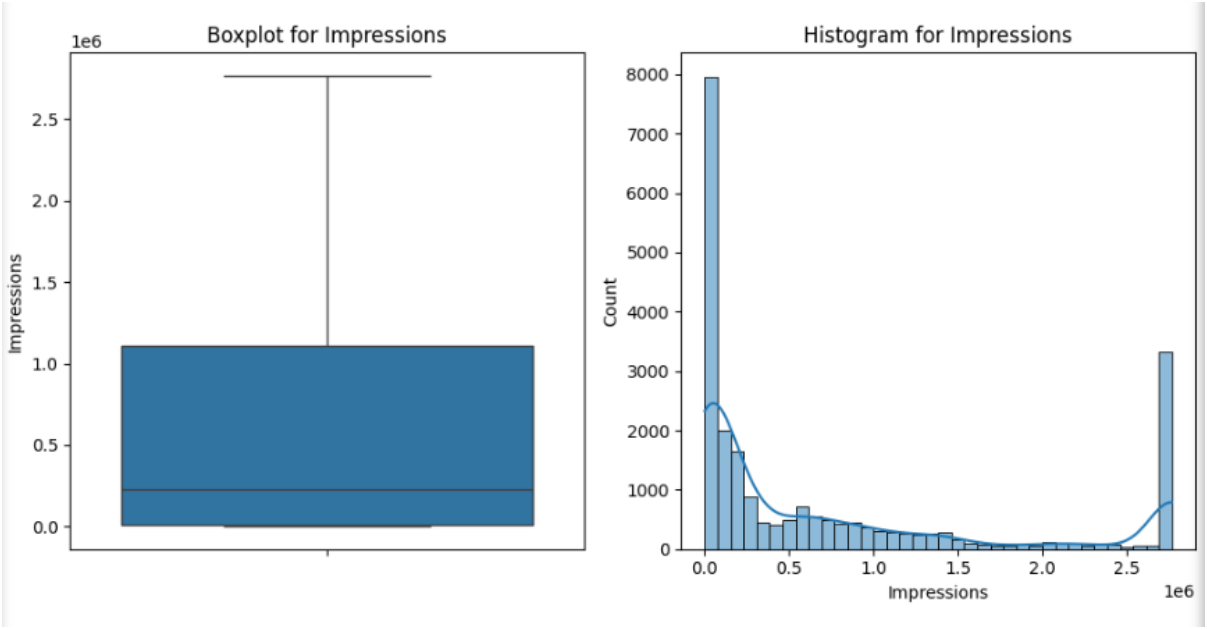
Histogram for Ad - Length
Count
7000
6000
5000
4000
3000
2000
1000
0
100
200
300
400
500
600
700
Ad - Length

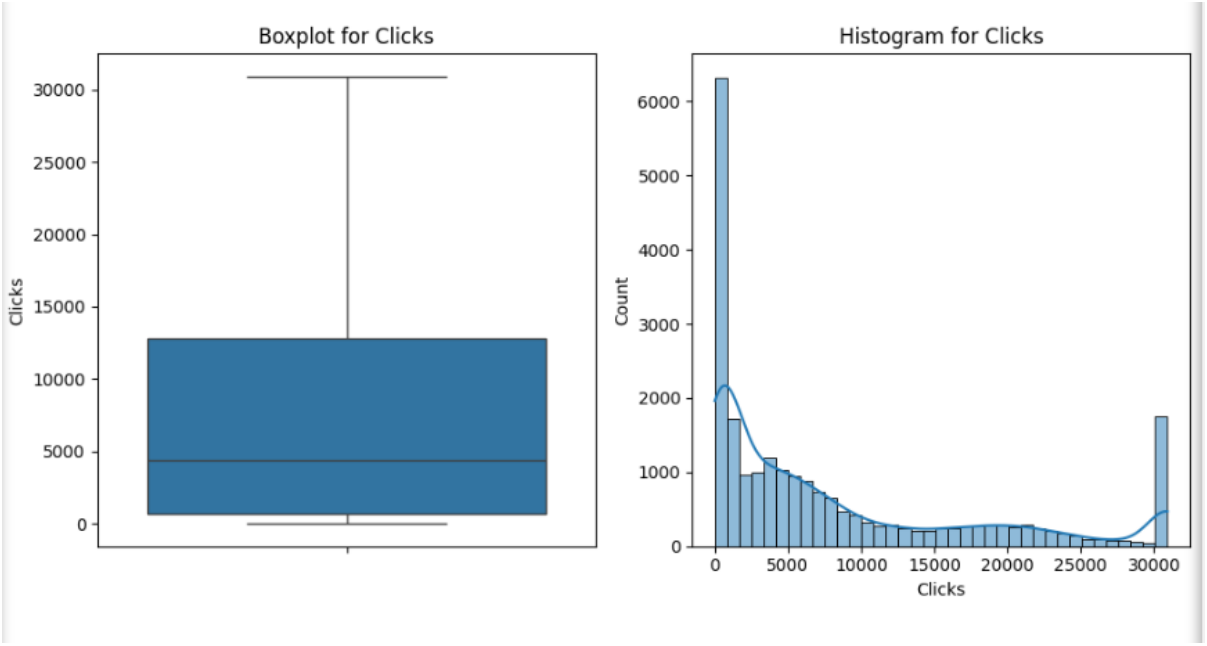

Boxplot for Ad- Width
Ad- Width
600
500
400
300
200
100

Histogram for Ad- Width
Count
8000
7000
6000
5000
4000
3000
2000
1000
0
100
200
300
400
500
600
Ad- Width

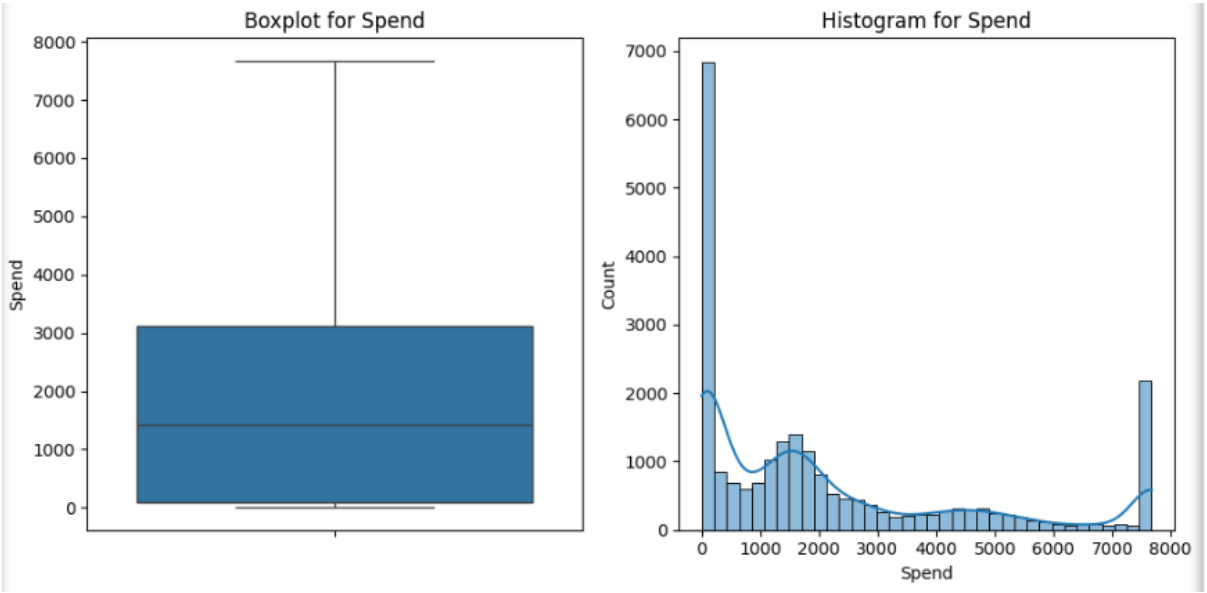

Boxplot for Ad Size
Ad Size
100000
90000
80000
70000
60000
Histogram for Ad Size
Count
Ad Size
7000
6000
5000
4000
3000
2000
1000
0
60000
70000
80000
90000
100000
Boxplot for Available_Impressions
1e6
Available_Impressions
6
5
4
3
2
1
0
Histogram for Available_Impressions
Count
Available_Impressions
1e6
8000
7000
6000
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6


Boxplot for Matched_Queries
1e6
Matched_Queries
Histogram for Matched_Queries
Count
Matched_Queries
1e6

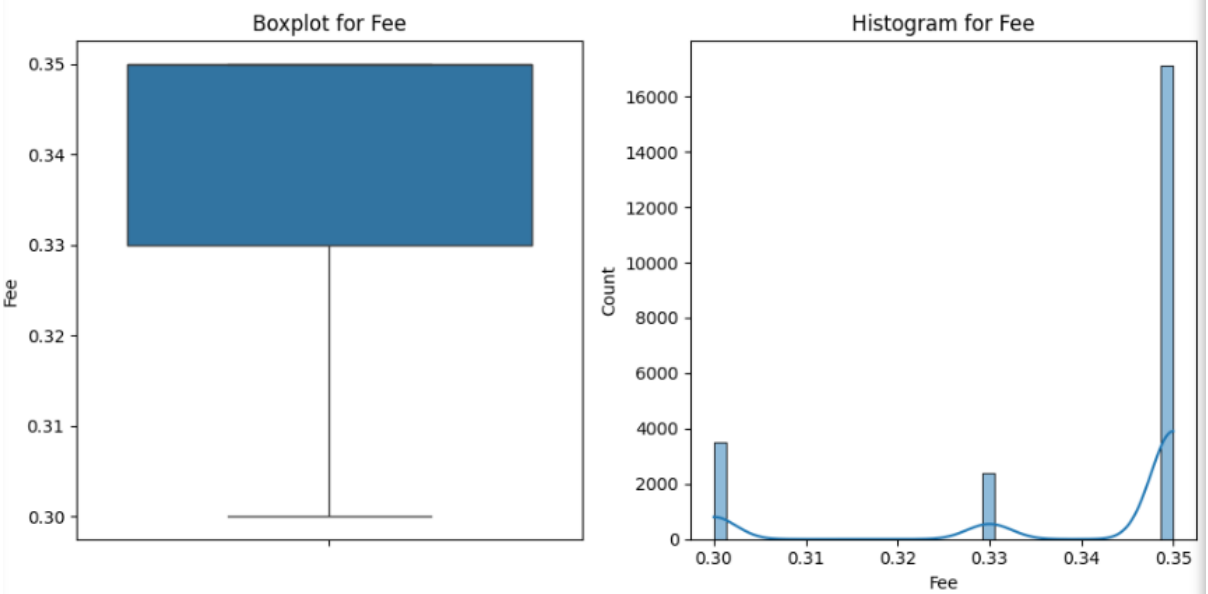

Boxplot for Impressions
1e6
Impressions
Histogram for Impressions
Count
Impressions
1e6


Boxplot for Clicks
Clicks
30000
25000
20000
15000
10000
5000
0

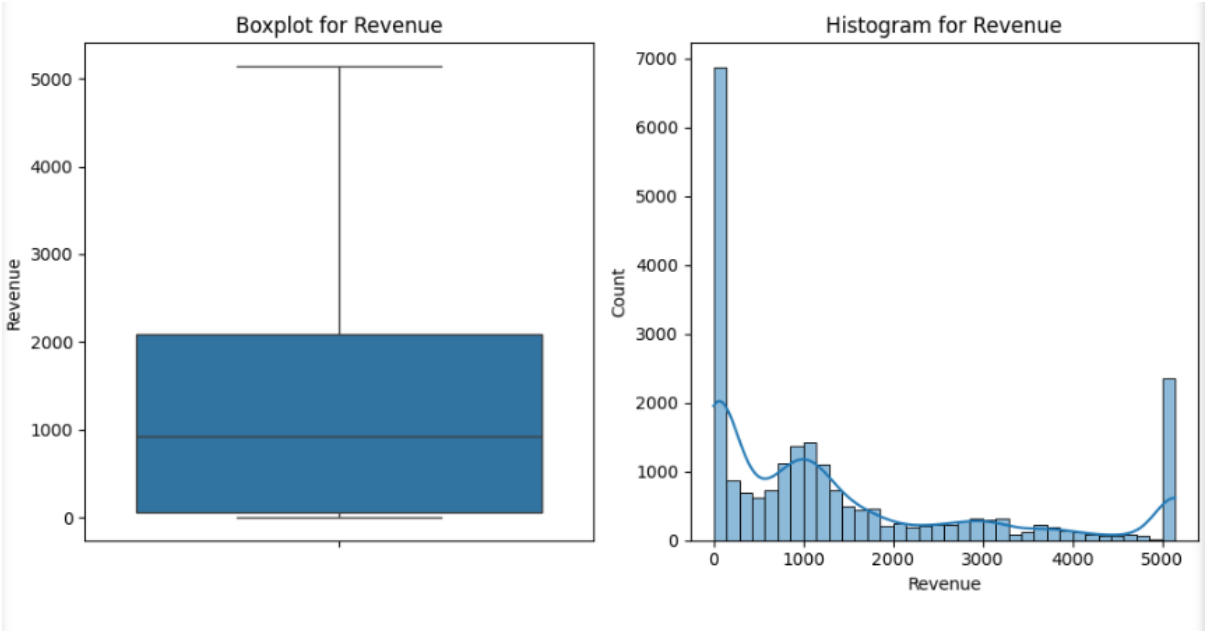
Histogram for Clicks
Count
6000
5000
4000
3000
2000
1000
0
0
5000
10000
15000
20000
25000
30000
Clicks

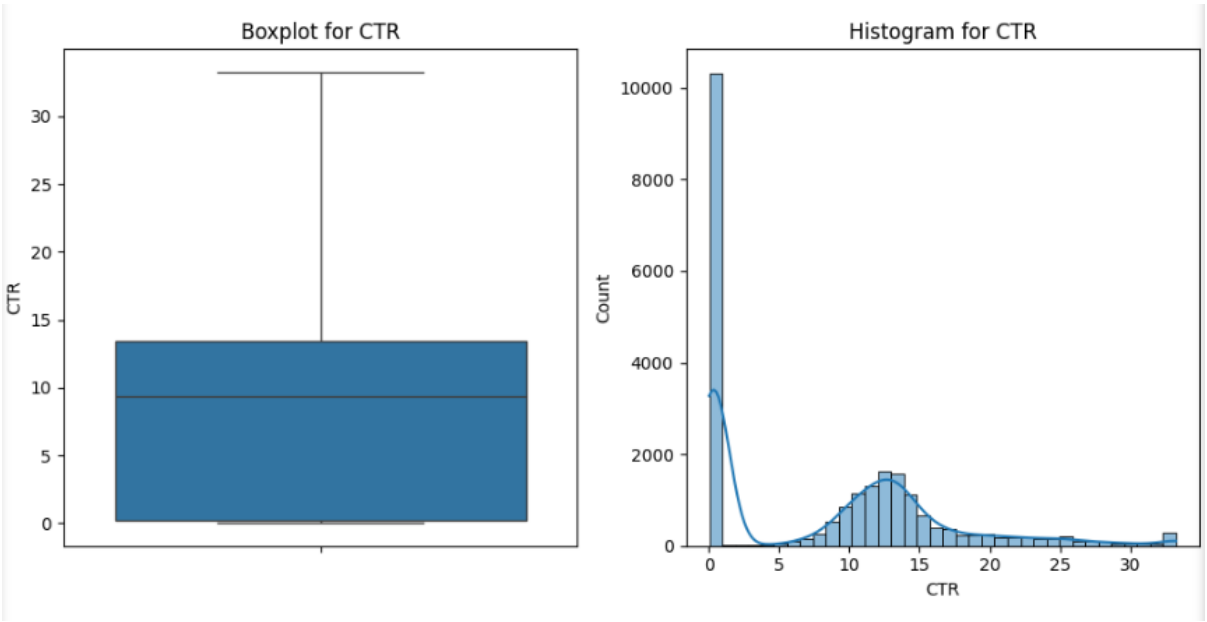

Boxplot for Spend
Spend
8000
7000
6000
5000
4000
3000
2000
1000
0

Histogram for Spend
Count
7000
6000
5000
4000
3000
2000
1000
0
0
1000
2000
3000
4000
5000
6000
7000
8000
Spend


Boxplot for Fee
Fee
0.35
0.34
0.33
0.32
0.31
0.30

Histogram for Fee
Count
16000
14000
12000
10000
8000
6000
4000
2000
0
0.30
0.31
0.32
0.33
0.34
0.35
Fee


Boxplot for Revenue
Revenue
0
1000
2000
3000
4000
5000
Histogram for Revenue
Count
Revenue
0
1000
2000
3000
4000
5000
6000
7000


Boxplot for CTR
CTR
0
5
10
15
20
25
30
Histogram for CTR
Count
CTR
0
2000
4000
6000
8000
10000
0
5
10
15
20
25
30

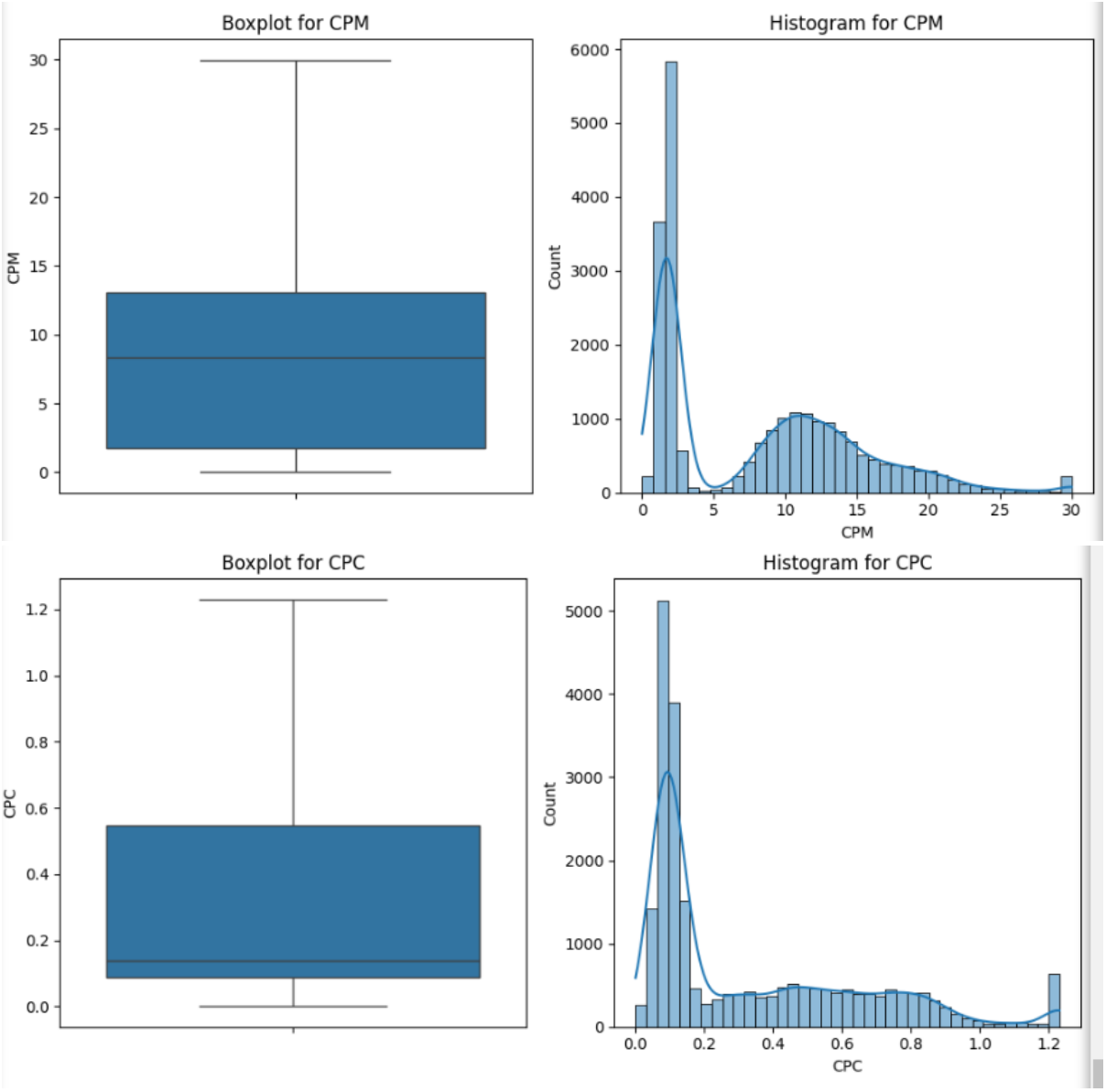


Fig-4 Boxplot After Outlier Treatment

Now, we will perform Bivariate Analysis on the numerical data to see and find any correlations between the variables and also find any hidden pattern the below figure shows the pair plot, heatmap.

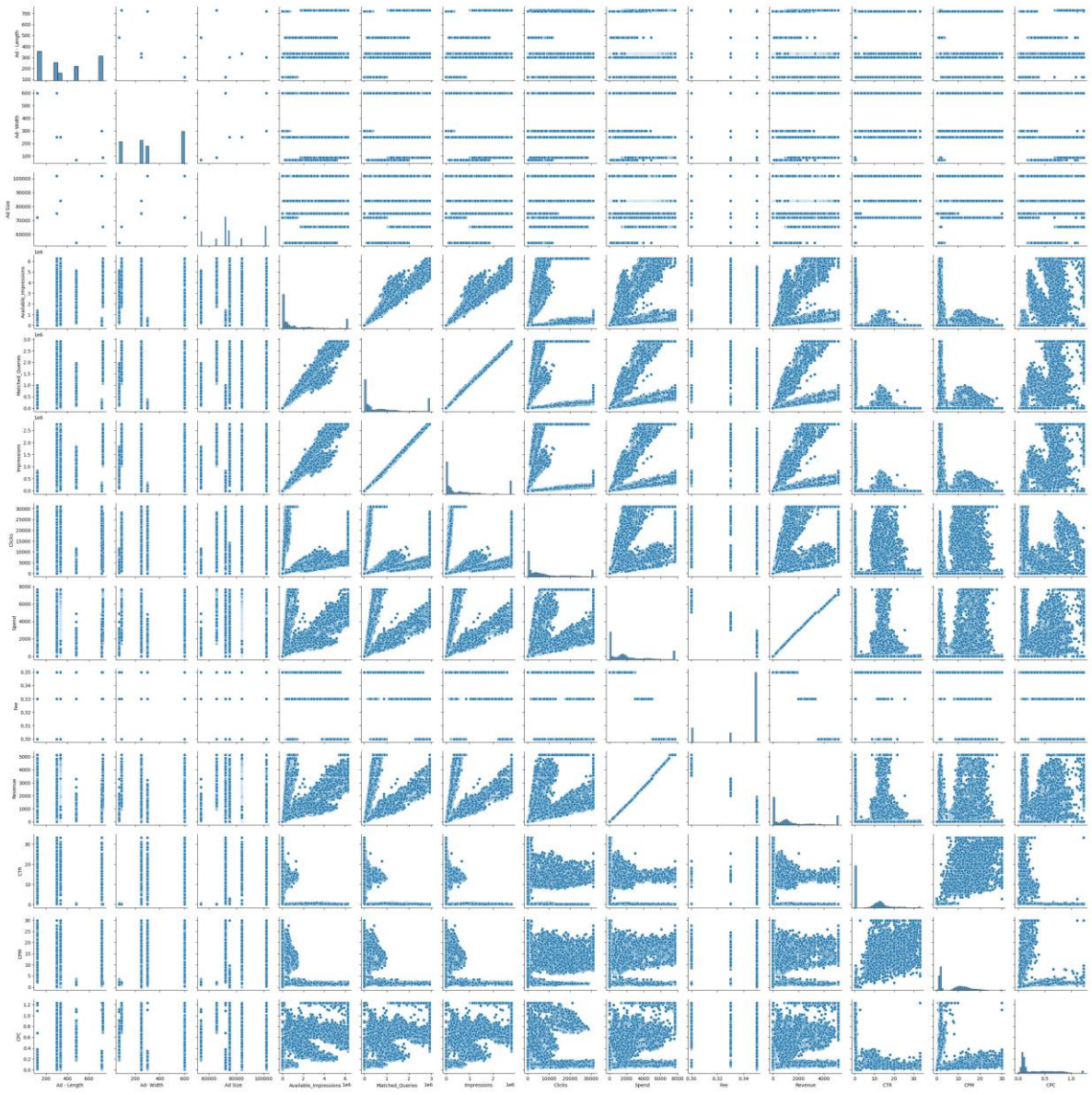


Fig-5 Pair plot for Numerical variable

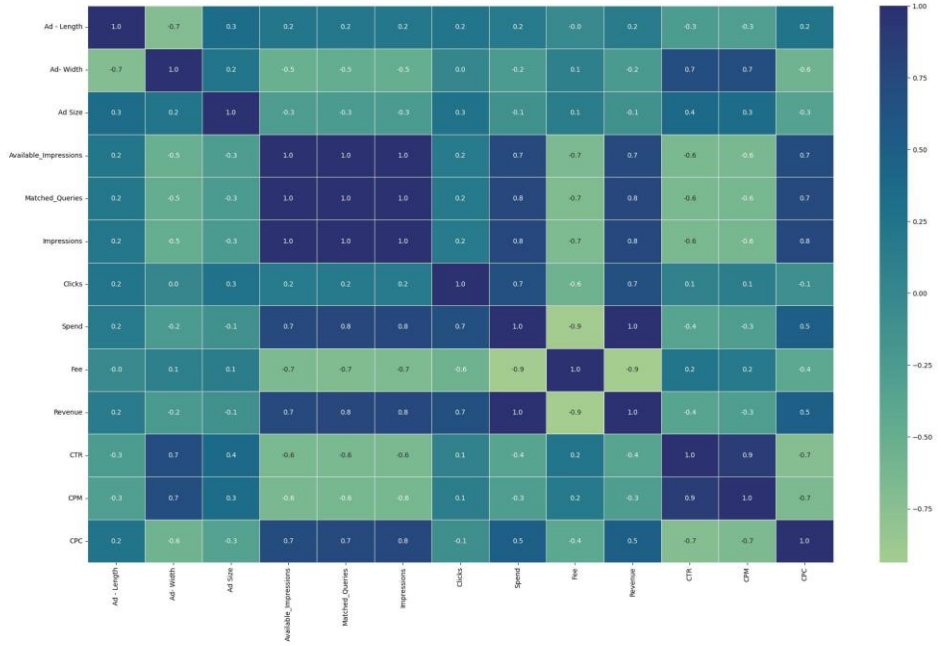


Fig-6 Heatmap for Numerical variable

From the Univariate Analysis and Bivariate Analysis on the unscaled numerical dataset we can observe the following things: -

1) We can see that the data is mostly Right skewed.

2) We can also see that there is a strong correlation between variables.

Now, the final step is to scale the numerical dataset using z-score scaling and we obtain the table after scaling as shown below,

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	-0.364496	-0.432797	-0.102518	-0.755333	-0.778949	-0.768478	-0.867488	-0.89317	0.535724	-0.880093	-0.958836	-1.194498	-1.042561
1	-0.364496	-0.432797	-0.102518	-0.755345	-0.778988	-0.768516	-0.867488	-0.89317	0.535724	-0.880093	-0.953835	-1.194498	-1.042561
2	-0.364496	-0.432797	-0.102518	-0.754900	-0.778919	-0.768445	-0.867488	-0.89317	0.535724	-0.880093	-0.962218	-1.194498	-1.042561
3	-0.364496	-0.432797	-0.102518	-0.755040	-0.778781	-0.768302	-0.867488	-0.89317	0.535724	-0.880093	-0.971871	-1.194498	-1.042561
4	-0.364496	-0.432797	-0.102518	-0.755610	-0.779030	-0.768560	-0.867488	-0.89317	0.535724	-0.880093	-0.946281	-1.194498	-1.042561

Table-7 Scaled data for Clustering

Effects of scaling:

The influence of scaling on an algorithm's efficiency is significant. Analysing the dataset's statistical summary reveals a broad spectrum of values, ranging from minimal to substantial. These values are not uniform in scale and carry varying degrees of significance, which underscores the necessity of scaling.

Essentially, scaling standardizes the range of independent variables or features of data, ensuring that each feature contributes proportionately to the final prediction. This homogenization is crucial because it can impact the performance of many machine learning algorithms, particularly those that use distance calculations like k-nearest neighbours (k-NN) or gradient descent-based algorithms like linear regression. Without scaling, features with larger ranges could unduly influence the model, leading to biased results.

1.2 Hierarchical and K-Means Clustering:

For, Hierarchical Clustering we will plot a Dendrogram using wards linkage method and Euclidean distance and truncating dendrogram up to p value to 10 and obtain the following plot as shown below.

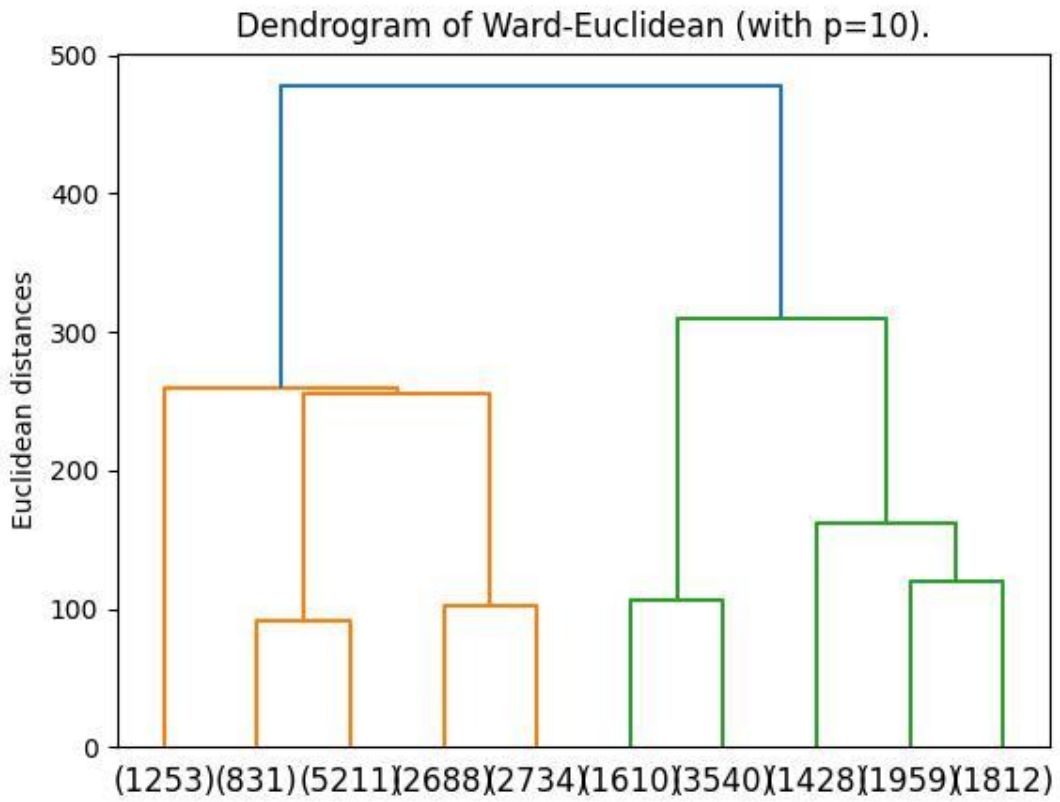


fig-7 Dendrogram plot

Now, will find out the WSS (within sum of squares) values for 10 clusters and plot them in elbow plot to find out the optimum number of K for K-Means algorithm. We can see the plot in below.

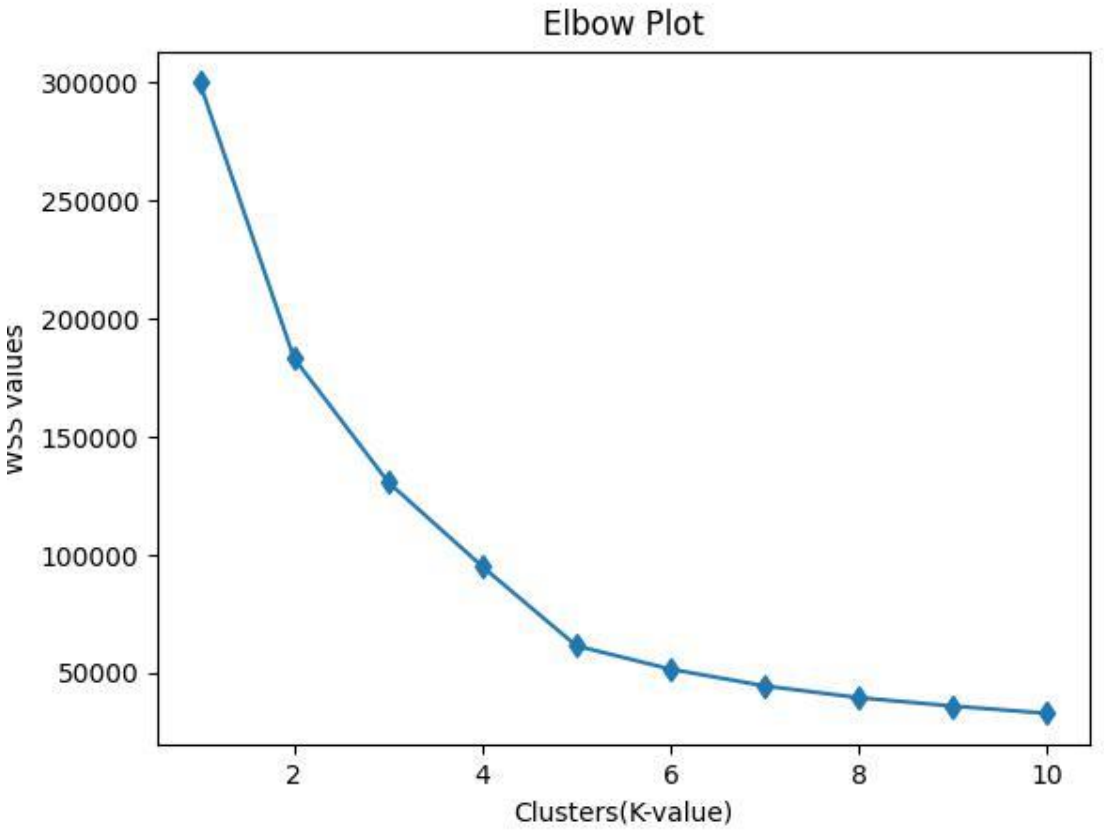


fig-8 Elbow plot

We can see that after clusters-5 the drop is significantly low so we can take the K=5 value but to make sure we will use silhouette score for 10 clusters and plot the elbow graph again.

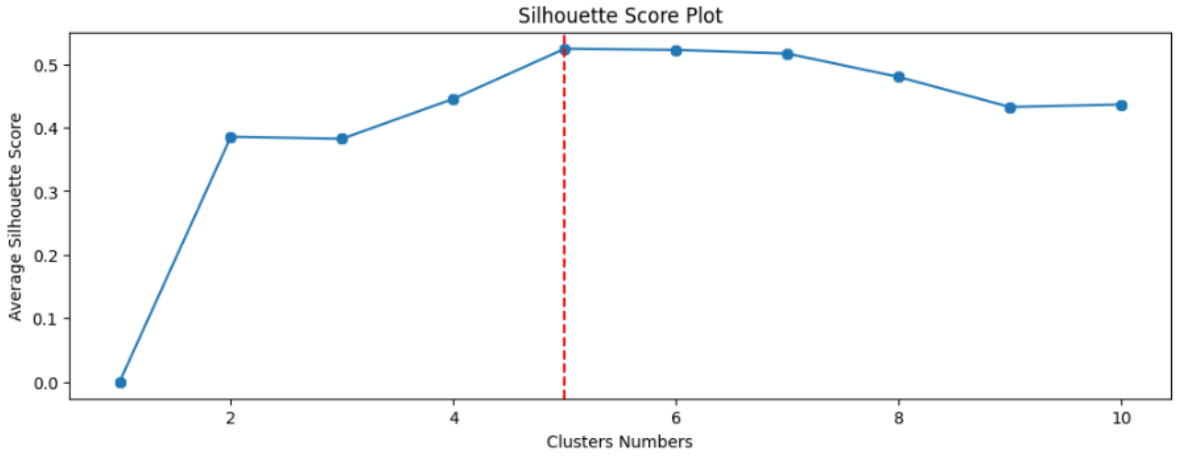


fig-9 Silhouette score plot

From, the above plot and Average Silhouette score for all 10 clusters we can say that no. of clusters or K value should be 5 for caring out the K-Means clustering.

Now, we will perform K-means clustering on the dataset and cluster the whole dataset into five clusters. After that create a new column called Kmeans clusters add it to the numerical dataset and concatenate the numerical and categorical variable into one whole dataset. We see that in the results in the below table.

Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	Device Type	Kmean_clusters
325.0	323.0	1.0	0.0	0.35	0.0	0.309598	0.0	0.0	Desktop	2
285.0	285.0	1.0	0.0	0.35	0.0	0.350877	0.0	0.0	Mobile	2
356.0	355.0	1.0	0.0	0.35	0.0	0.281690	0.0	0.0	Desktop	2
497.0	495.0	1.0	0.0	0.35	0.0	0.202020	0.0	0.0	Mobile	2
242.0	242.0	1.0	0.0	0.35	0.0	0.413223	0.0	0.0	Desktop	2

Table-8 Clustered Column Added in Dataset

1.3 Cluster profiling:

For this we will use the variables like clicks, spend, revenue, CPM, CPC, CTR based on k-mean clusters and device type variables for profiling the clusters. We use group by function and make the following bar plots.

First, we make Bar plot for Clicks variable obtain the plot as shown below. We can say that for both devices cluster-0 is highest and followed by cluster-4 so on.

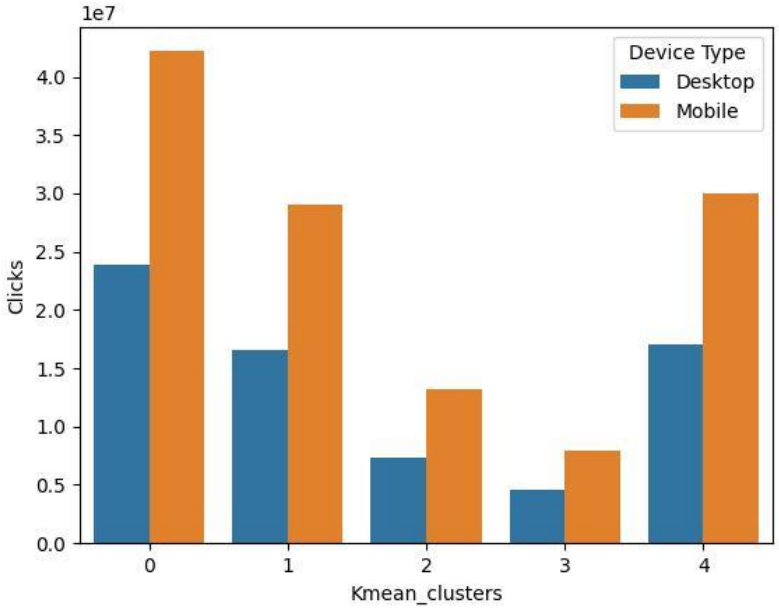


fig-10 Bar plot for Clicks

Now, we make Bar plot for Spend variable as shown below and can conclude that cluster-1 is the highest and cluster-3 has the lowest value.

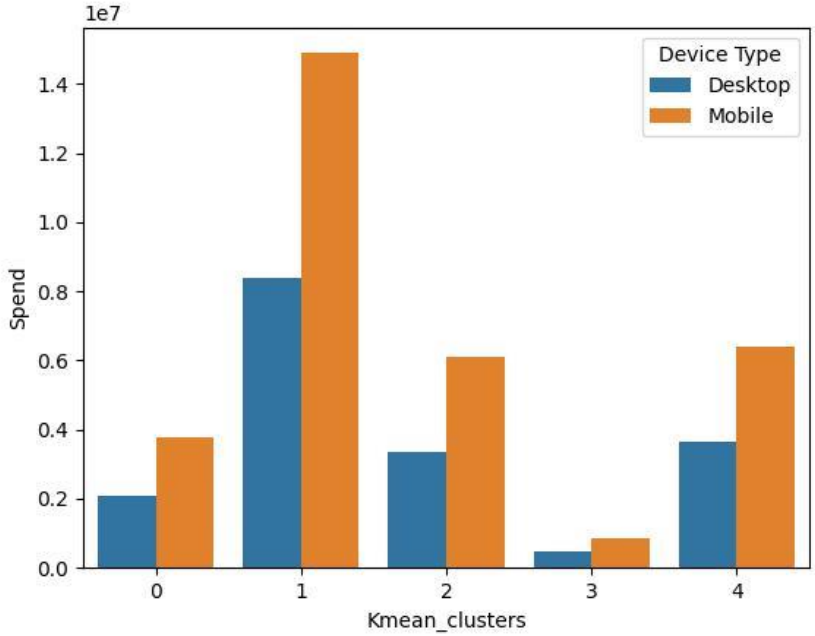


fig-11 Bar plot for Spend

Now we make Bar plot for Revenue variable as shown below and can conclude that cluster-1 has the highest value followed by cluster-4 then cluster-2 , cluster-0 and at last cluster-3.

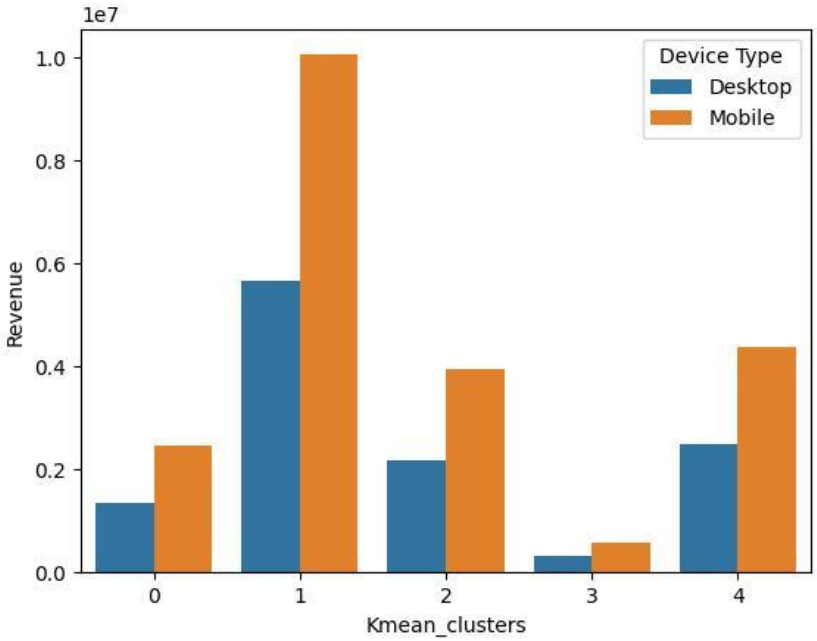


fig-12 Bar plot for Revenue

Finally, we create the Bar plot for CTR, CPM, CPC as shown below and conclude that for CPM is hight for cluster 3 and low for cluster 1. CPC is hight for cluster 1 and low or negligible for cluster 3 and 4. CTR is high for cluster 3 and low for cluster 1.

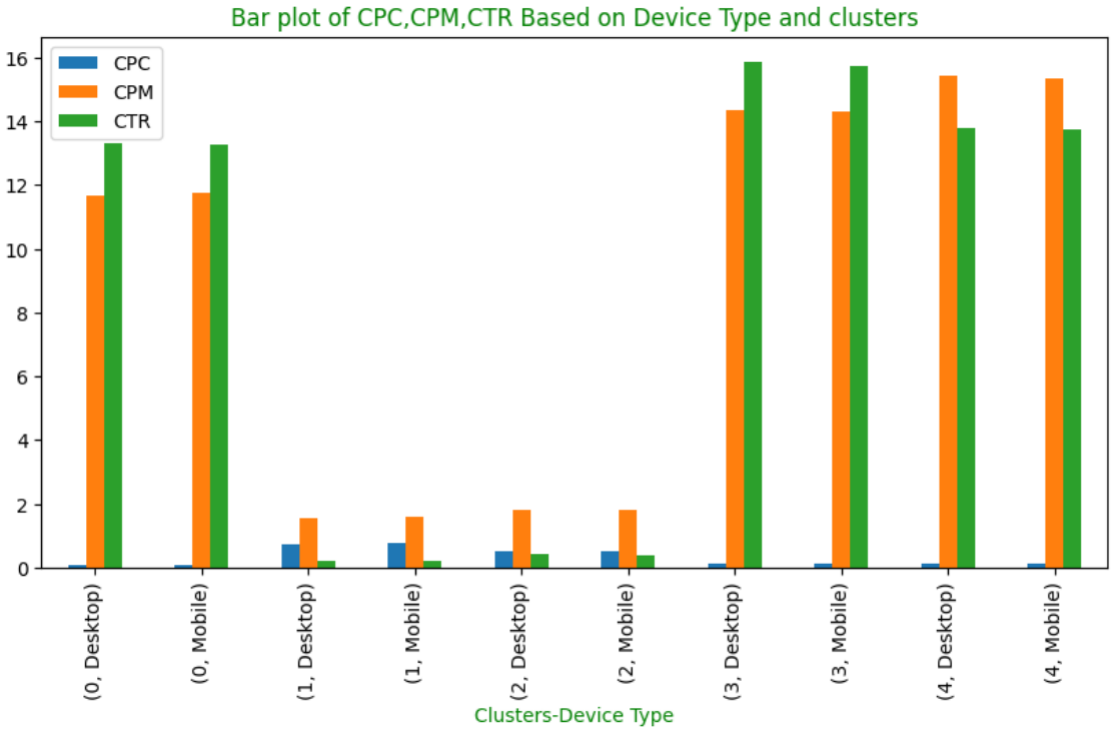


fig-13 Bar plot for CTR, CPM, CPC

1.4 CONCLUSION:

1) Ad Length vs. Ad Width:

- Ad Length (Ad Length > Ad Width):
- CTR (Click-Through Rate): Approximately 0.21%.
- CPM (Cost Per Mille): Lowest at 1.54.
- CPC (Cost Per Click): Highest at 0.75.

Recommendation: Investigate the reasons behind the low CTR. Optimize creative content, targeting, and placement. Ensure relevance to the audience.

2) Ad Width (Ad Width > Ad Length):

- CTR: High at approximately 15.81%.
- CPM: Highest at 14.61.
- CPC: Lowest at 0.10.

Recommendation: Capitalize on the successful aspects (high CTR, low CPC, and high CPM). Continuously monitor and test variations for further optimization.

3) Revenue Impact:

The Ad Width > Ad Length cluster contributes almost 0.65% of the total revenue from ad spends.

Recommendation: Maintain a balanced approach by allocating resources to both ad types. Prioritize revenue while managing costs effectively.

They're learning outcome is that I learned about the Clustering types and the thought process is required when selecting any clustering type on dataset and got some refresher on some previous topics like EDA, scaling, etc.

Problem-2: India Census 2011

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.

2.1 Data- preprocessing and EDA:

First, we will look at first and last five rows using function head and tail respectively, of the dataset from excel file called PCA India data census that we loaded using read excel function. In fig-1 and fig-2 shows below shows the dataset.

	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_3_F	MARG_I
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180	237	
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123	229	
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	...	114	188	44	89	
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	...	194	247	61	128	
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	...	874	1928	465	1043	

fig-14 First 5 rows

	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MAR
635	34	636	Puducherry	Mahe	3333	8154	11781	1146	1203	21	...	32	47	0	0	
636	34	637	Puducherry	Karaikal	10612	12346	21691	1544	1533	2234	...	155	337	3	14	
637	35	638	Andaman & Nicobar Island	Nicobars	1275	1549	2630	227	225	0	...	104	134	9	4	
638	35	639	Andaman & Nicobar Island	North & Middle Andaman	3762	5200	8012	723	664	0	...	136	172	24	44	
639	35	640	Andaman & Nicobar Island	South Andaman	7975	11977	18049	1470	1358	0	...	173	122	6	2	

fig-15 Last 5 rows

We observed that there are many columns named 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 etc.

Now, we use shape function the dataset and we get that there are 640 row and 60 columns. Then, we use info function and found out the data type of each column as shown in the below table. we will check for the duplicated rows are present or not using duplicate function and found out that in the dataset there are zero same or duplicated rows present. We also used isnull function and find there are zero or no empty values present in the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 640 entries, 0 to 639
Data columns (total 61 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   State Code      640 non-null    int64
 1   Dist.Code       640 non-null    int64
 2   State           640 non-null    object
 3   Area Name       640 non-null    object
 4   No_HH           640 non-null    int64
 5   TOT_M           640 non-null    int64
 6   TOT_F           640 non-null    int64
 7   M_06            640 non-null    int64
 8   F_06            640 non-null    int64
 9   M_SC            640 non-null    int64
 10  F_SC            640 non-null    int64
 11  M_ST            640 non-null    int64
 12  F_ST            640 non-null    int64
 13  M_LIT           640 non-null    int64
 14  F_LIT           640 non-null    int64
 15  M_ILL           640 non-null    int64
 16  F_ILL           640 non-null    int64
 17  TOT_WORK_M      640 non-null    int64
 18  TOT_WORK_F      640 non-null    int64
 19  MAINWORK_M      640 non-null    int64
 20  MAINWORK_F      640 non-null    int64
 21  MAIN_CL_M       640 non-null    int64
 22  MAIN_CL_F       640 non-null    int64
 23  MAIN_AL_M       640 non-null    int64
 24  MAIN_AL_F       640 non-null    int64
 25  MAIN_HH_M       640 non-null    int64
 26  MAIN_HH_F       640 non-null    int64
 27  MAIN_OT_M       640 non-null    int64
 28  MAIN_OT_F       640 non-null    int64
 29  MARGWORK_M      640 non-null    int64
 30  MARGWORK_F      640 non-null    int64
 31  MARG_CL_M       640 non-null    int64
 32  MARG_CL_F       640 non-null    int64
 33  MARG_AL_M       640 non-null    int64
 34  MARG_AL_F       640 non-null    int64
 35  MARG_HH_M       640 non-null    int64
 36  MARG_HH_F       640 non-null    int64
 37  MARG_OT_M       640 non-null    int64
 38  MARG_OT_F       640 non-null    int64
 39  MARGWORK_3_6_M  640 non-null    int64
 40  MARGWORK_3_6_F  640 non-null    int64
 41  MARG_CL_3_6_M   640 non-null    int64
 42  MARG_CL_3_6_F   640 non-null    int64
 43  MARG_AL_3_6_M   640 non-null    int64
 44  MARG_AL_3_6_F   640 non-null    int64
 45  MARG_HH_3_6_M   640 non-null    int64
 46  MARG_HH_3_6_F   640 non-null    int64
 47  MARG_OT_3_6_M   640 non-null    int64
 48  MARG_OT_3_6_F   640 non-null    int64
 49  MARGWORK_0_3_M  640 non-null    int64
 50  MARGWORK_0_3_F  640 non-null    int64
 51  MARG_CL_0_3_M   640 non-null    int64
 52  MARG_CL_0_3_F   640 non-null    int64
 53  MARG_AL_0_3_M   640 non-null    int64
 54  MARG_AL_0_3_F   640 non-null    int64
 55  MARG_HH_0_3_M   640 non-null    int64
 56  MARG_HH_0_3_F   640 non-null    int64
 57  MARG_OT_0_3_M   640 non-null    int64
 58  MARG_OT_0_3_F   640 non-null    int64
 59  NON_WORK_M      640 non-null    int64
 60  NON_WORK_F      640 non-null    int64
dtypes: int64(59), object(2)
```

Table-9 Info function Table

Now, will create a new column called Gender Ratio and added in the dataset into dataset and using head function and after it is added we use the describe function again on the dataset to get the five summary important values for the newly created Gender Ratio. Now, we will use describe function on this dataset and obtain the min, max, std, count, mean, q1, q2, q3 values for all the numerical columns as shown in the below table. Graph is shown below for mean Gender ratio for each state.

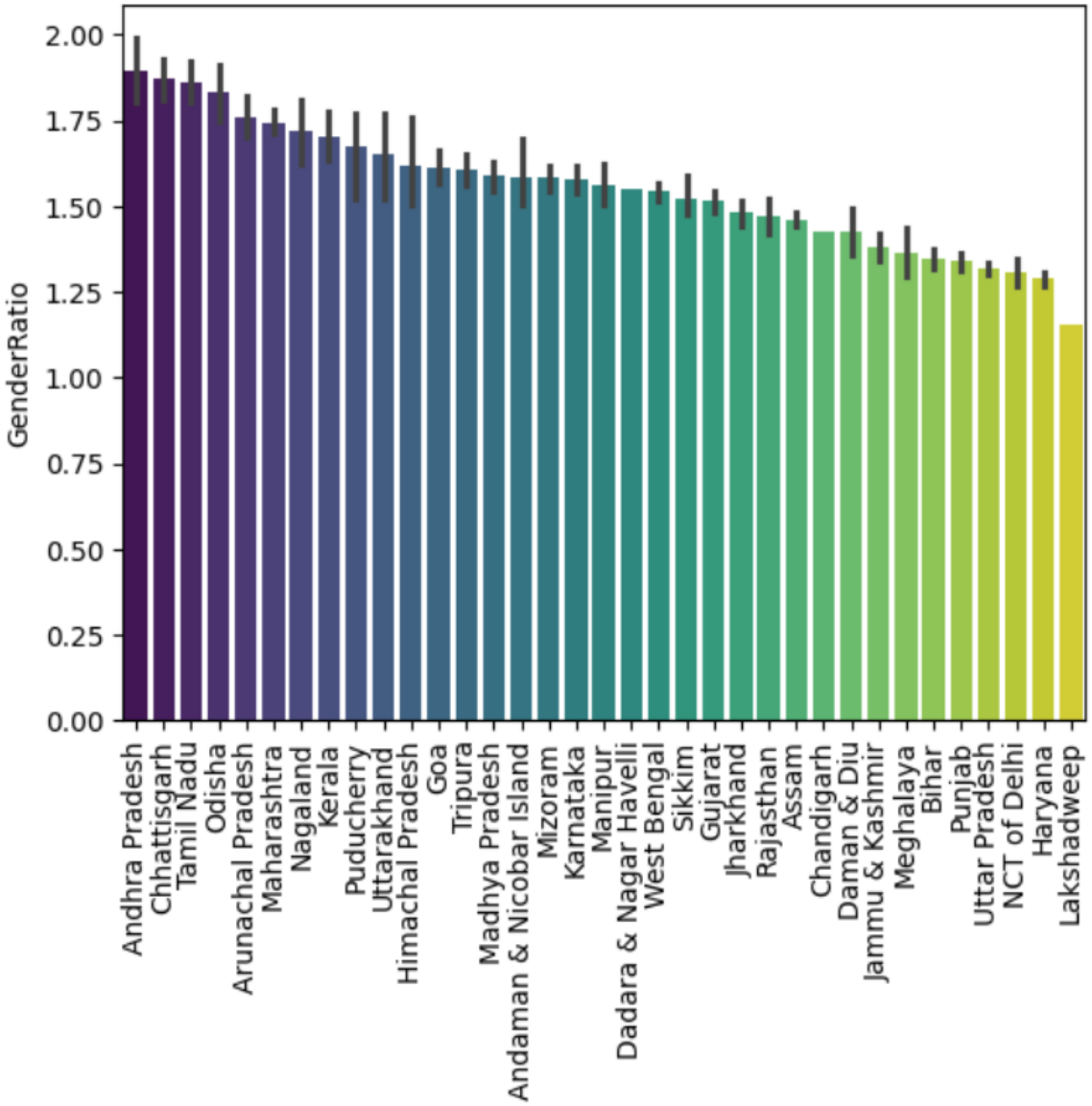


fig-16 Gender Ratio Bar Plot

	count	mean	std	min	25%	50%
State Code	640.0	17.114062	9.426486	1.000000	9.000000	18.000000
Dist.Code	640.0	320.500000	184.896367	1.000000	160.750000	320.500000
No_HH	640.0	51222.871875	48135.405475	350.000000	19484.000000	35837.000000
TOT_M	640.0	79940.576563	73384.511114	391.000000	30228.000000	58339.000000
TOT_F	640.0	122372.084375	113600.717282	698.000000	46517.750000	87724.500000
M_06	640.0	12309.098438	11500.906881	56.000000	4733.750000	9159.000000
F_06	640.0	11942.300000	11326.294567	56.000000	4672.250000	8663.000000
M_SC	640.0	13820.946875	14426.373130	0.000000	3466.250000	9591.500000
F_SC	640.0	20778.392188	21727.887713	0.000000	5603.250000	13709.000000
M_ST	640.0	6191.807813	9912.668948	0.000000	293.750000	2333.500000
F_ST	640.0	10155.640625	15875.701488	0.000000	429.500000	3834.500000
M_LIT	640.0	57967.979688	55910.282466	286.000000	21298.000000	42693.500000
F_LIT	640.0	66359.565625	75037.860207	371.000000	20932.000000	43796.500000
M_ILL	640.0	21972.596875	19825.605268	105.000000	8590.000000	15767.500000
F_ILL	640.0	56012.518750	47116.693769	327.000000	22367.000000	42386.000000

Table-10 First few rows of describe Table

we have performed EDA on numerical variables as per requirements but out of all variables we can only select five variables to perform EDA and thus, we need to ask and find the answers to our questions as given below:

Q1) Which State has the highest and the lowest gender ratio?

A1) The Uttar Pradesh has 93.41 highest while, Lakshadweep has lowest 1.15 gender ratio.

Now, as shown below figure we can see the that we see the Gender Ratio for the states as well as table shown below

State	
Uttar Pradesh	93.419861
Madhya Pradesh	79.387259
Maharashtra	61.028511
Tamil Nadu	59.555615
Odisha	54.909147
Bihar	51.108680
Rajasthan	48.517953
Karnataka	47.346656
Andhra Pradesh	43.587142
Assam	39.430086
Gujarat	39.351217
Jharkhand	35.555111
Chhattisgarh	33.651022
Jammu & Kashmir	30.381027
West Bengal	29.296690
Arunachal Pradesh	28.122140
Haryana	27.038507
Punjab	26.791864
Kerala	23.802752
Uttarakhand	21.452670
Himachal Pradesh	19.394036
Nagaland	18.893102
Manipur	14.074579
Mizoram	12.650335
NCT of Delhi	11.770049
Meghalaya	9.562205
Puducherry	6.693767
Tripura	6.433467
Sikkim	6.096870
Andaman & Nicobar Island	4.745611
Goa	3.227768
Daman & Diu	2.854771
Dadara & Nagar Havelli	1.551275
Chandigarh	1.428496
Lakshadweep	1.151993

Name: GenderRatio, dtype: float64

Table-11 State wise Gender ratio

Q2) Q2) What is highest and lowest value for gender ratio as per districts?

A2) From the below code we can see that the district with code 547 has highest 2.28 gender ratio value and code 587 has lowest of 1.15 value. It can be seen in the table given below.

Dist.Code	
547	2.283250
398	2.268763
625	2.225429
546	2.221849
391	2.215060
...	...
139	1.184830
106	1.180761
144	1.180202
2	1.179576
587	1.151993

Name: GenderRatio, Length: 640, dtype: float64

Table- 12 District code wise Gender ratio

Q3) Which states has highest and lowest literacy rate as per their gender?

A3) Andaman & Nicobar Island has 82.27% highest literacy rate while West Bengal has 74.92% lowest for male.

For female Andaman & Nicobar Island has 70.53% highest literacy rate while West Bengal has 57.83% lowest. It can also be seen in the table below.

FOR MALE LITERACY RATE:

State	
Andaman & Nicobar Island	0.827085
Andhra Pradesh	0.724712
Arunachal Pradesh	0.671484
Assam	0.711972
Bihar	0.598354
Chandigarh	0.803583
Chhattisgarh	0.733391
Dadara & Nagar Havelli	0.733171
Daman & Diu	0.827188
Goa	0.835282
Gujarat	0.760907
Haryana	0.749246
Himachal Pradesh	0.802359
Jammu & Kashmir	0.672121
Jharkhand	0.665078
Karnataka	0.749135
Kerala	0.811806
Lakshadweep	0.826718
Madhya Pradesh	0.713084
Maharashtra	0.788496
Manipur	0.757676
Meghalaya	0.610784
Mizoram	0.814862
NCT of Delhi	0.791835
Nagaland	0.759543
Odisha	0.737274
Puducherry	0.827295
Punjab	0.742014
Rajasthan	0.703164
Sikkim	0.796205
Tamil Nadu	0.808522
Tripura	0.815733

Uttar Pradesh	0.665239
Uttarakhand	0.755365
West Bengal	0.749542
dtype: float64	
FOR FEMALE LITERACY RATE:	
State	
Andaman & Nicobar Island	0.705343
Andhra Pradesh	0.439314
Arunachal Pradesh	0.514466
Assam	0.550760
Bihar	0.406581
Chandigarh	0.728288
Chhattisgarh	0.461043
Dadara & Nagar Havelli	0.490075
Daman & Diu	0.669304
Goa	0.730168
Gujarat	0.586118
Haryana	0.551532
Himachal Pradesh	0.654789
Jammu & Kashmir	0.502746
Jharkhand	0.435937
Karnataka	0.543499
Kerala	0.798583
Lakshadweep	0.767262
Madhya Pradesh	0.491609
Maharashtra	0.647051
Manipur	0.589823
Meghalaya	0.617477
Mizoram	0.831862
NCT of Delhi	0.690211
Nagaland	0.669607
Odisha	0.510190
Puducherry	0.657692
Punjab	0.611202
Rajasthan	0.442871
Sikkim	0.653018
Tamil Nadu	0.571286
Tripura	0.714791
Uttar Pradesh	0.463640
Uttarakhand	0.604570
West Bengal	0.578332
dtype: float64	

Table- 13 State wise Literacy rate

Q4) Which group has highest and lowest gender ratio as per Population in the age group 0-6 for Male and Female as per the states?

A4) Haryana has the lowest value of 858.48 while, Arunachal Pradesh has the highest value of 1085.05. It can be seen in the table below.

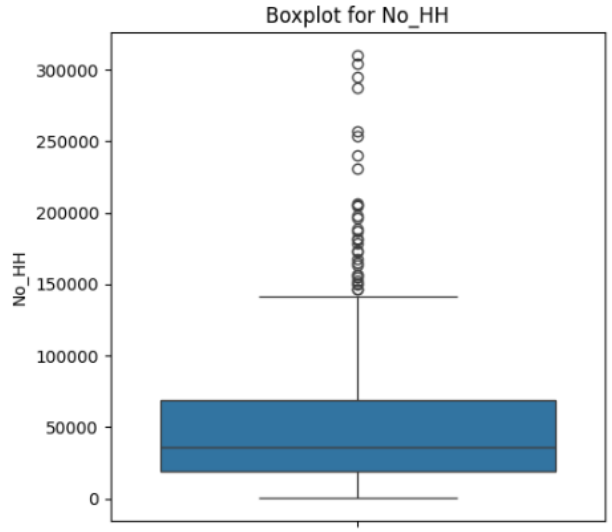
State	
Haryana	858.480597
Punjab	869.658139
Chandigarh	874.544128
NCT of Delhi	889.584269
Jammu & Kashmir	892.391196
Uttarakhand	900.009853
Himachal Pradesh	907.177823
Lakshadweep	923.211169
Andaman & Nicobar Island	928.512397
Rajasthan	934.733155
Gujarat	945.291581
Tripura	950.208787
Tamil Nadu	953.878768
Madhya Pradesh	960.657686
Daman & Diu	961.674528
Manipur	964.006983
Odisha	964.071416
Andhra Pradesh	970.265375
Nagaland	975.936968
Karnataka	977.466847
Puducherry	978.837209
Kerala	981.337471
Sikkim	984.455959
Uttar Pradesh	985.131016
Assam	990.065048
West Bengal	992.555412
Maharashtra	995.240084
Bihar	997.996966
Chhattisgarh	1001.604245
Goa	1002.356268
Meghalaya	1006.781787
Mizoram	1014.876751
Jharkhand	1015.004941
Dadara & Nagar Havelli	1041.811847
Arunachal Pradesh	1085.058618

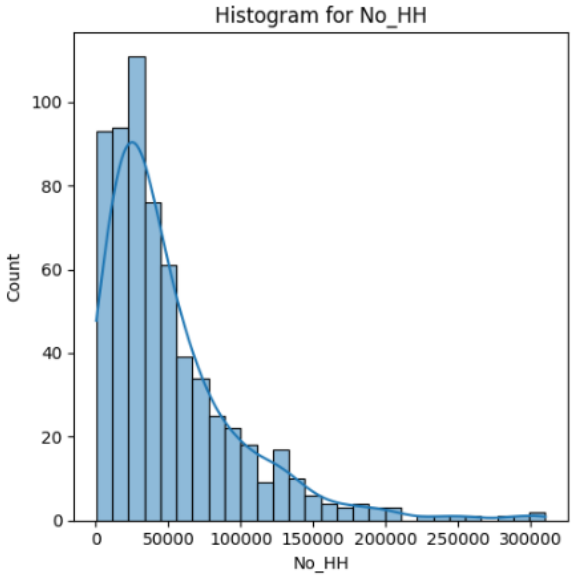
dtype: float64

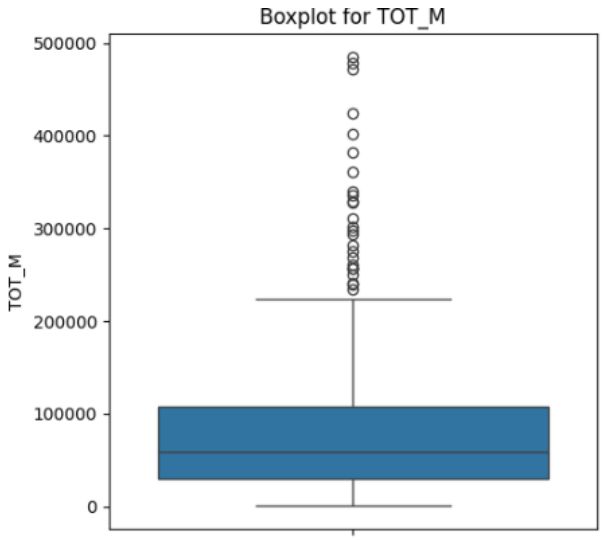
Table- 14 Group wise Gender ratio in states

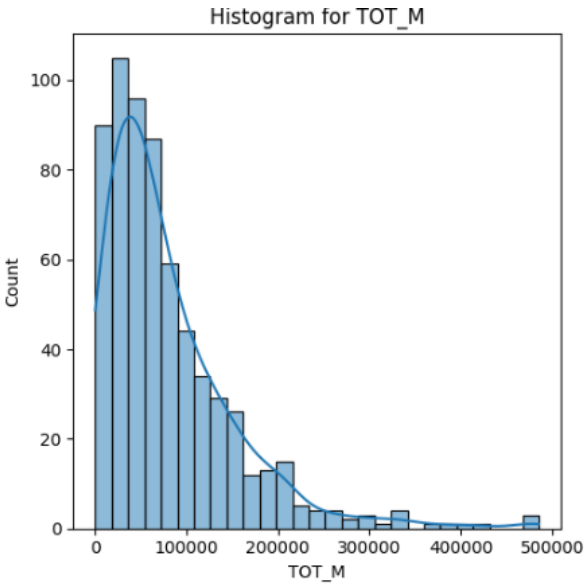
Now, from the above question we have selected five variables that we are going to perform EDA on, and they are No_HH, TOT_M, TOT_F, M_06, F_06 variables. We will create new dataset with these five selected variables and perform EDA on them. In the below two tables can be seen that has visual view for this new dataset and the described function table.

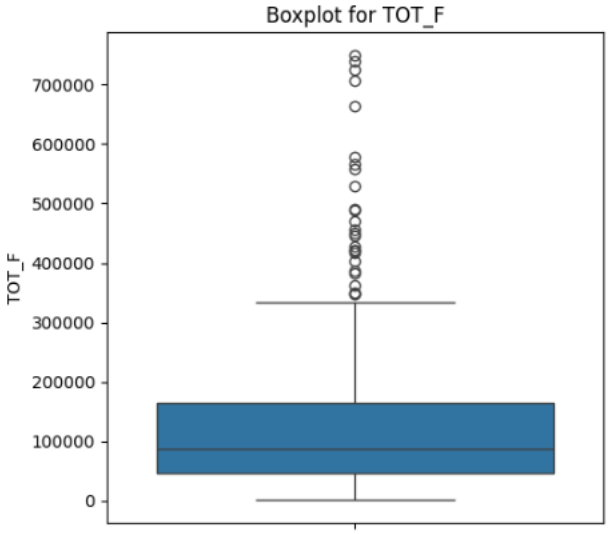
Univariate Analysis of selected five variables and the boxplot and histogram has been shown below.

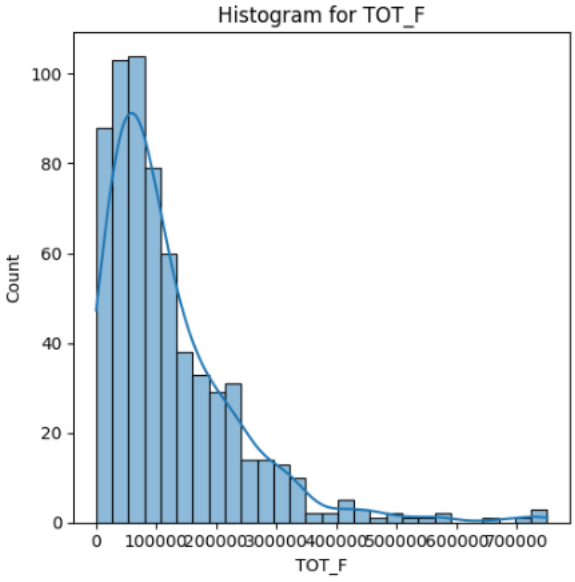

Boxplot for No_HH
No_HH
300000
250000
200000
150000
100000
50000
0


Histogram for No_HH
Count
100
80
60
40
20
0
0
50000
100000
150000
200000
250000
300000
No_HH


Boxplot for TOT_M
TOT_M
500000
400000
300000
200000
100000
0


Histogram for TOT_M
Count
100
80
60
40
20
0
0
100000
200000
300000
400000
500000
TOT_M


Boxplot for TOT_F
TOT_F
700000
600000
500000
400000
300000
200000
100000
0


Histogram for TOT_F
Count
100
80
60
40
20
0
0
100000
200000
300000
400000
500000
600000
700000
TOT_F

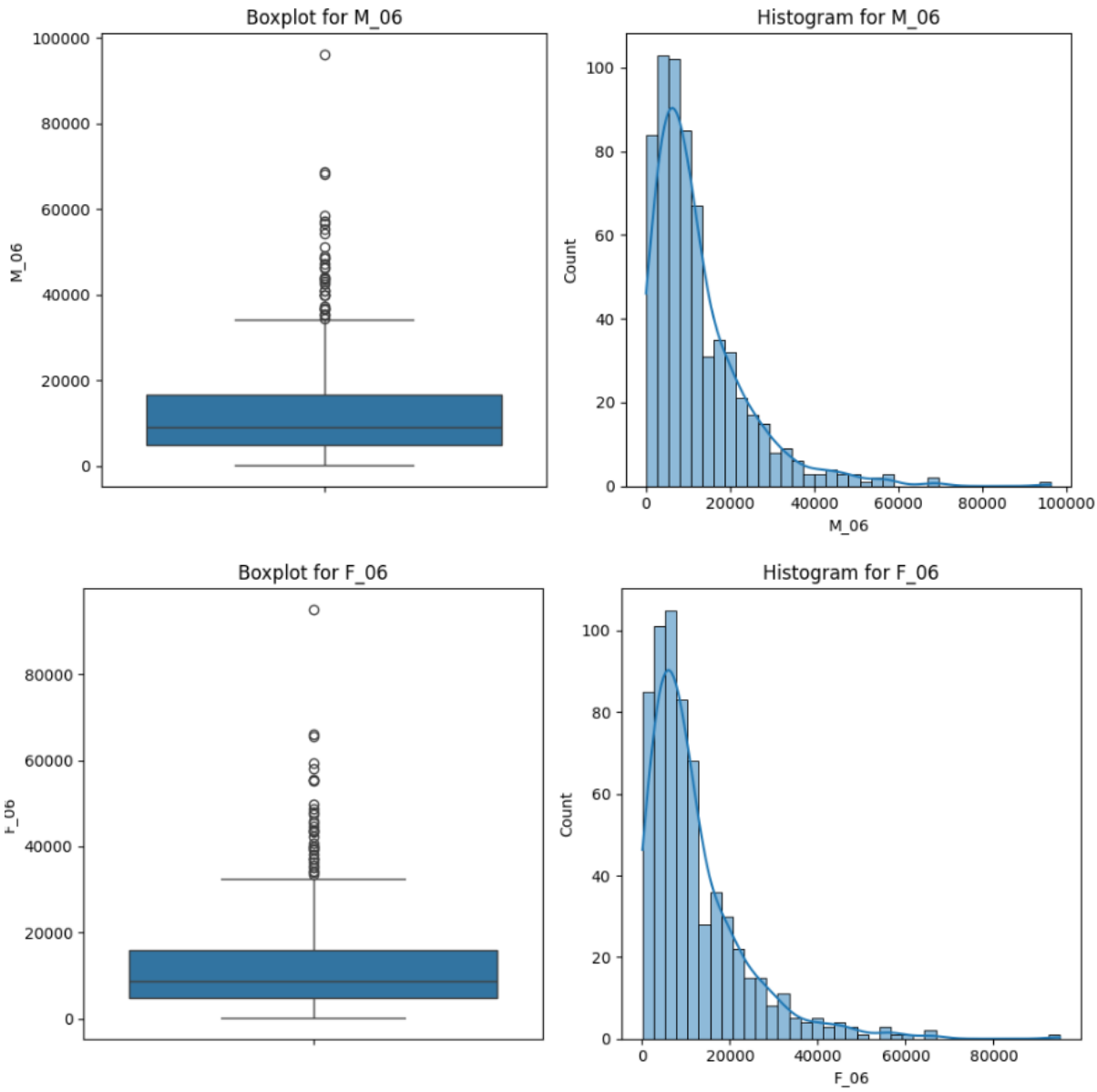


fig-17 Boxplot for 5 variables

From, the plot we can see that there is presence of outliers and, they are unscaled variables, we will need to do scaling and outliers treatment overall original dataset once the EDA is done these five variables. We can also see that they are right skewed variables and have huge variables within the dataset.

Now, we perform Bivariate Analysis and obtain the following pair plot and heatmap for the five-variable dataset

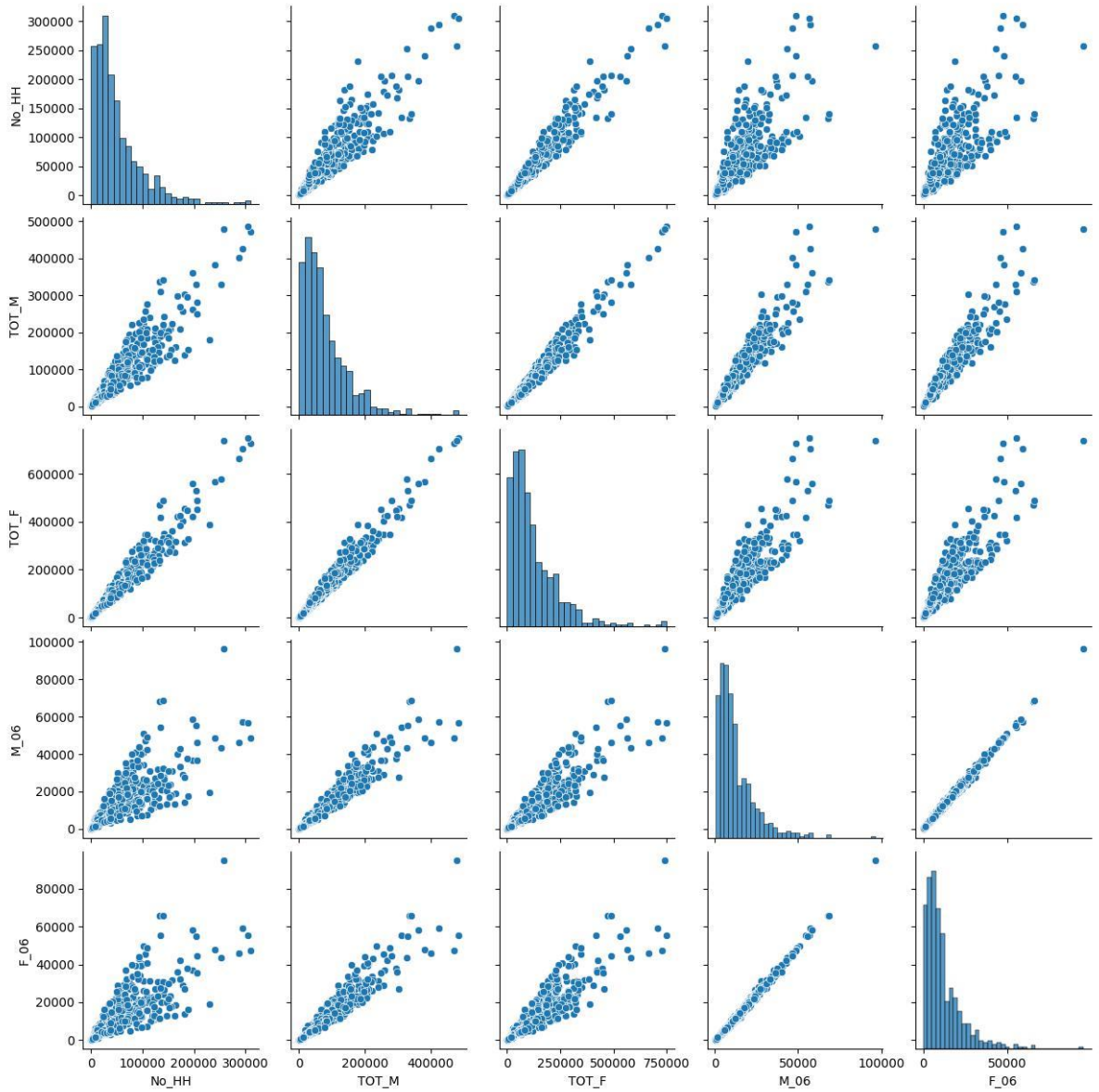


fig-18 Pair plot for 5 variables

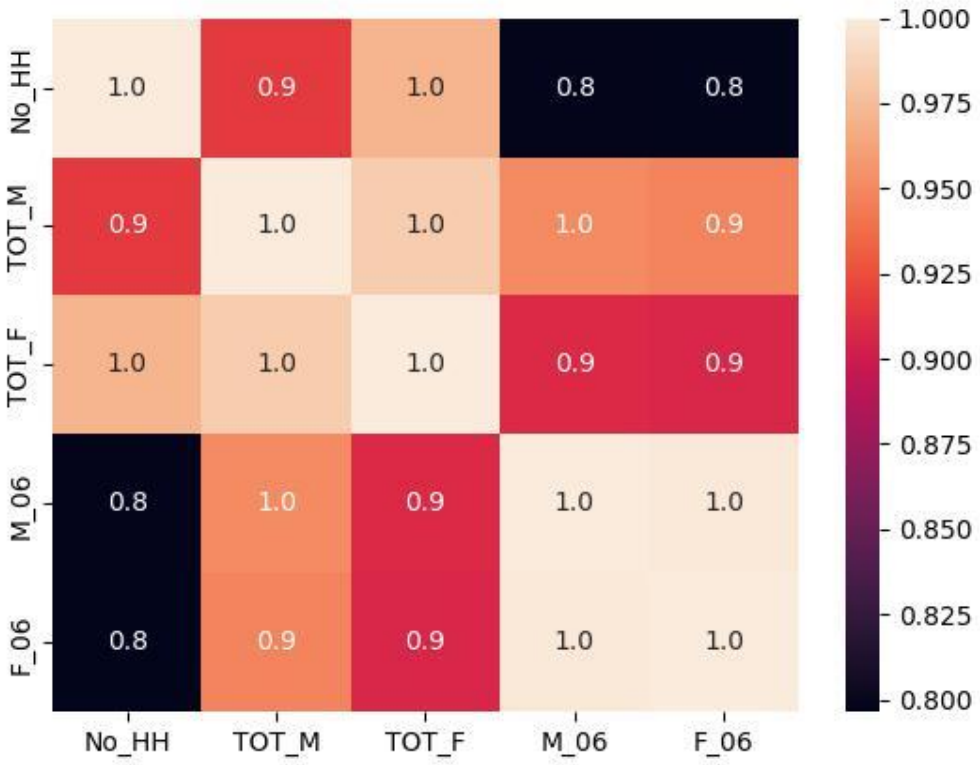


fig-19 Heatmap for 5 variables

We can observe that all these five variables have a significant correlation with each other and have a positive correlation as one variables value increases so does the other variable.

We will drop the categorical variables that are not needed for PCA like State, Area name, Gender Ratio, dist. Code and so on.
Now, to perform PCA we will have to scale the date and in the below graph you can see the unscaled boxplots as shown below.

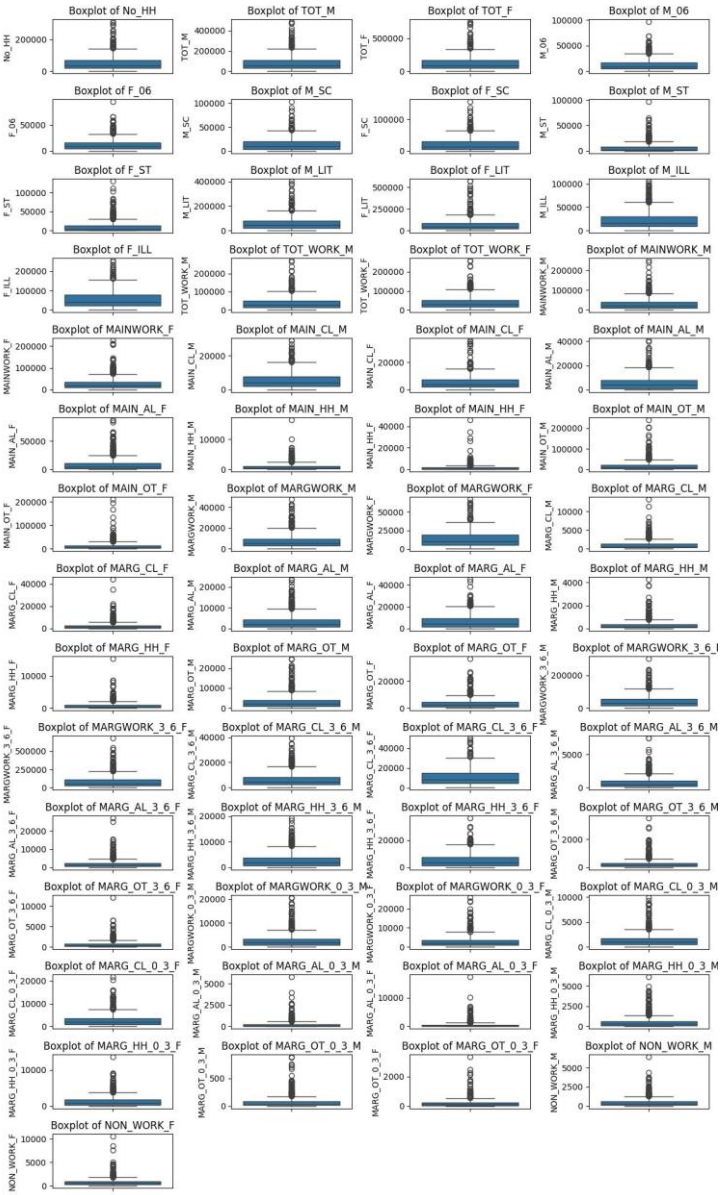


fig-20 Boxplot for unscaled variables

We will apply z-score method and obtained the scaled data shown in the figure.
Now, we will again apply the same code for plotting the boxplots and get the below figure and we can observe that scaling does not affect the outlier presence in the variables as scaling all it does is bring the variables to the same scale for accurate calculations that we need to perform.

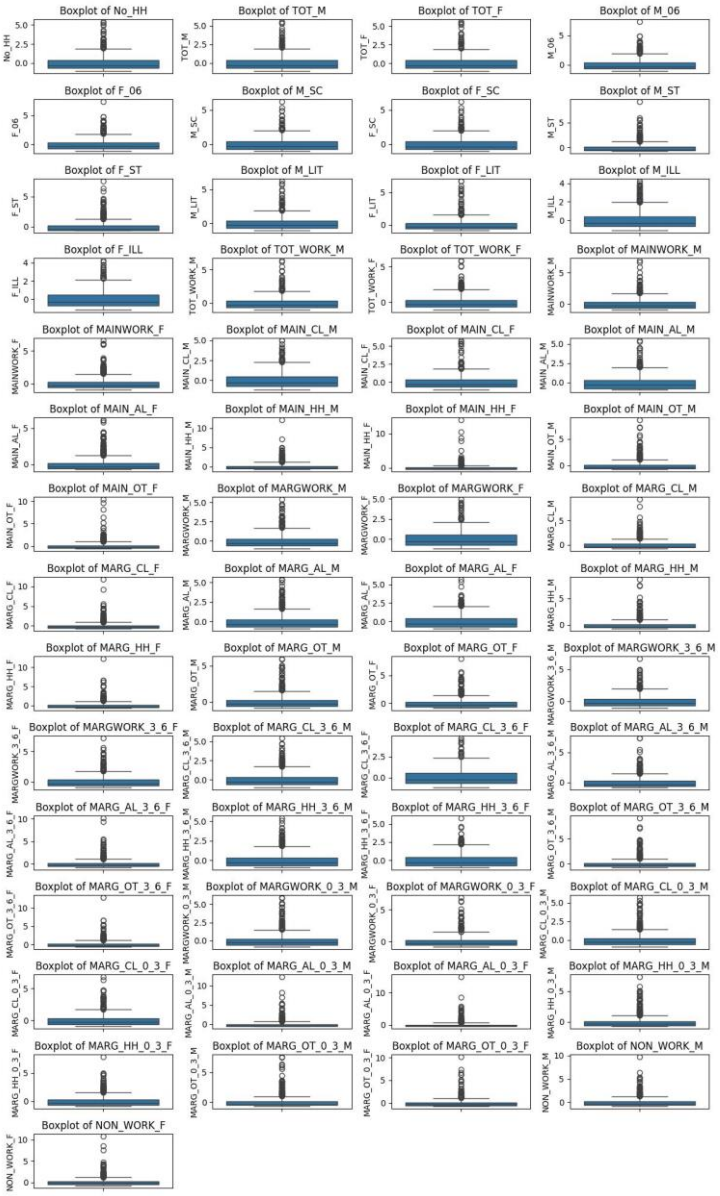


fig-21 Boxplot for scaled variables

We will perform the outlier treatment the same way we did in problem-1 and obtained the new boxplots for the scaled dataset.

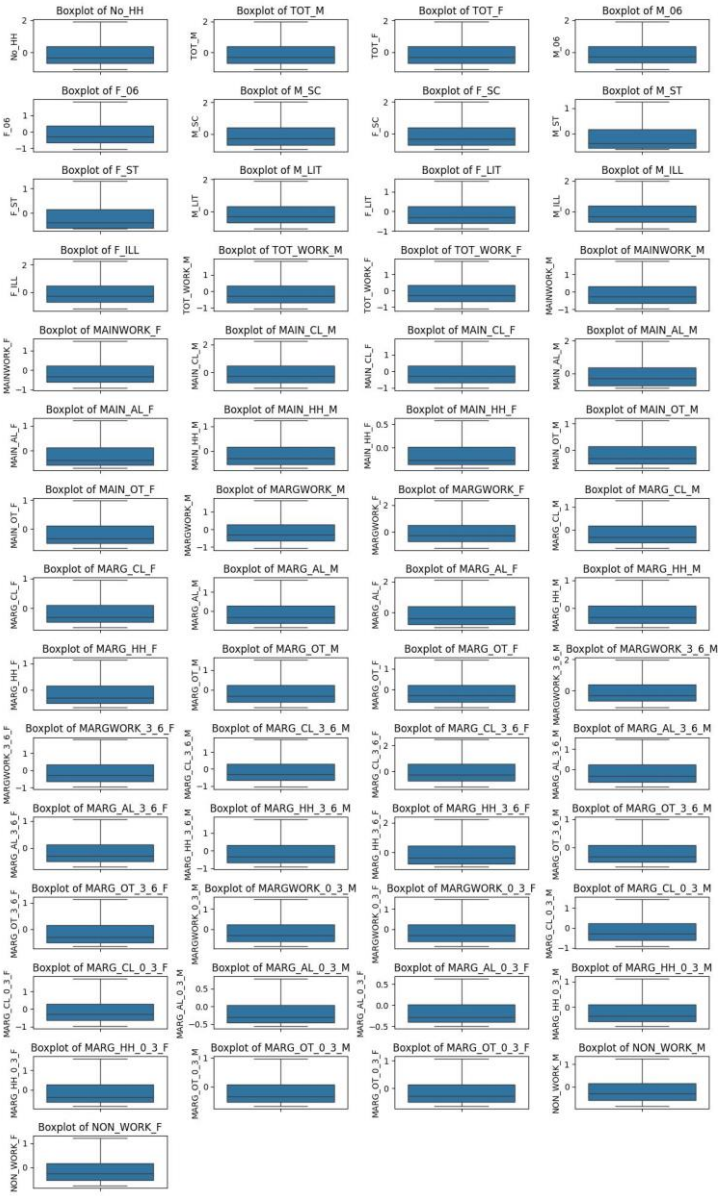


fig-22 Boxplot for scaled variables and outlier treated

We have created a heatmap to show the correlation between the variables for the scaled and treated dataset.

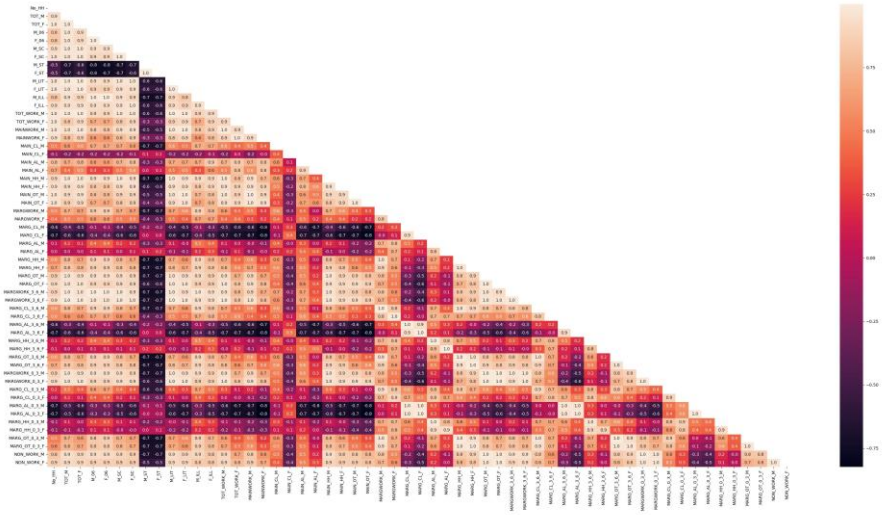


fig-23 heatmap for scaled dataset

Now, we will use the Bartlett sphericity calculation to see if there are any significant correlations or not and the value we obtained the p-value <0.05 and can say that yes there is a significant correlation.
We also performed calculation for Kmo model and obtained value 0.93 which is > 0.7, indicating we can perform PCA.

2.2 Perform PCA:

We will create a covariance matrix and can be seen in the below table.

	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	MARG_AL_0_3_F	M
No_HH	1.00	0.91	0.97	0.81	0.81	0.81	0.86	0.12	0.12	0.93	...	0.60	0.62	0.09	0.08	
TOT_M	0.91	1.00	0.98	0.97	0.96	0.88	0.86	0.02	0.01	0.99	...	0.74	0.64	0.23	0.19	
TOT_F	0.97	0.98	1.00	0.91	0.91	0.86	0.88	0.08	0.07	0.98	...	0.70	0.65	0.19	0.16	
M_06	0.81	0.97	0.91	1.00	1.00	0.83	0.80	-0.01	-0.02	0.92	...	0.80	0.68	0.35	0.31	
F_06	0.81	0.96	0.91	1.00	1.00	0.82	0.79	0.01	-0.01	0.92	...	0.81	0.69	0.36	0.31	
M_SC	0.81	0.88	0.86	0.83	0.82	1.00	0.98	-0.10	-0.10	0.87	...	0.65	0.55	0.19	0.15	
F_SC	0.86	0.86	0.88	0.80	0.79	0.98	1.00	-0.05	-0.05	0.86	...	0.62	0.57	0.15	0.13	
M_ST	0.12	0.02	0.08	-0.01	0.01	-0.10	-0.05	1.00	0.99	0.03	...	0.09	0.20	0.09	0.09	
F_ST	0.12	0.01	0.07	-0.02	-0.01	-0.10	-0.05	0.99	1.00	0.02	...	0.08	0.21	0.07	0.08	
M_LIT	0.93	0.99	0.98	0.92	0.92	0.87	0.86	0.03	0.02	1.00	...	0.69	0.60	0.19	0.15	
F_LIT	0.94	0.94	0.96	0.84	0.84	0.81	0.82	0.05	0.04	0.97	...	0.62	0.56	0.12	0.09	
M_ILL	0.78	0.93	0.88	0.97	0.97	0.82	0.78	0.02	0.01	0.87	...	0.78	0.66	0.32	0.27	
F_ILL	0.90	0.92	0.93	0.90	0.90	0.84	0.86	0.11	0.11	0.88	...	0.73	0.71	0.27	0.24	
TOT_WORK_M	0.94	0.98	0.97	0.90	0.89	0.87	0.87	0.06	0.05	0.98	...	0.66	0.57	0.13	0.09	
TOT_WORK_F	0.95	0.83	0.90	0.73	0.73	0.73	0.80	0.25	0.26	0.84	...	0.57	0.63	0.14	0.14	
MAINWORK_M	0.93	0.94	0.94	0.83	0.83	0.84	0.84	0.05	0.04	0.95	...	0.53	0.45	0.01	-0.01	
MAINWORK_F	0.92	0.77	0.86	0.65	0.65	0.69	0.76	0.22	0.22	0.81	...	0.40	0.45	-0.03	-0.03	

Table- 15 Covariance Matrix

Now we perform Eigen decomposition and obtain the following matrixes that has PCs, explained variance and explained variance ratio respectively as shown below.

```
array([1.99168024e+00, 6.26759062e-01, 1.98469001e-01, 8.86356128e-02,
       5.47773348e-02, 1.90622311e-02, 1.11218380e-02, 3.78064216e-03,
       2.31589770e-03, 1.43622281e-03, 1.28755450e-03, 8.10613528e-04,
       5.93416913e-04, 3.46163830e-04, 3.22891995e-04, 2.25606257e-04,
       1.81507581e-04, 1.67136036e-04, 1.20707249e-04, 8.59828662e-05,
       8.46576027e-05, 7.09429636e-05, 5.06190285e-05, 3.67036975e-05,
       2.88577826e-05, 2.40896688e-05, 2.22627435e-05, 1.94683585e-05,
       1.76911514e-05, 1.51129180e-05, 1.29034059e-05, 1.20862652e-05,
       9.65573613e-06, 8.65563950e-06, 7.19610801e-06, 6.46918361e-06,
       5.42618038e-06, 5.07864487e-06, 4.30983748e-06, 4.15724003e-06,
       3.89586823e-06, 3.10776784e-06, 2.62618093e-06, 2.20410719e-06,
       1.88599831e-06, 1.75518863e-06, 1.18508243e-06, 9.36588210e-07,
       8.73237390e-07, 6.97056021e-07, 4.16392415e-07, 3.48365467e-07,
       2.32633142e-07, 1.42975783e-07, 7.87370088e-08, 3.71358661e-08,
       2.96200216e-31])
```

fig-24 Array for PCs matrix

```
array([[-0.17183796, -0.17774185, -0.17524095, ..., -0.11513521,
        -0.15750061, -0.13519652],
       [ 0.06234143, -0.01999535,  0.01589756, ..., -0.14880893,
        -0.06161616, -0.03089135],
       [ 0.05923685, -0.03781452,  0.00671237, ...,  0.01853063,
        -0.11138935, -0.0460998 ],
       ...,
       [ 0.0598666 , -0.21721558, -0.242402  , ..., -0.09471876,
         0.0231657 , -0.00193034],
       [ 0.02200652, -0.38453037, -0.04711229, ...,  0.0437575 ,
         0.11246383, -0.09926668],
       [-0.09599873,  0.32058118,  0.33321909, ...,  0.03786379,
        -0.01217672, -0.04261217]])
```

fig-25 Array for Explained Variance

```
array([6.63308280e-01, 2.08735553e-01, 6.60980259e-02, 2.95191641e-02,
       1.82430187e-02, 6.34847676e-03, 3.70401186e-03, 1.25910334e-03,
       7.71285516e-04, 4.78318992e-04, 4.28806565e-04, 2.69966360e-04,
       1.97631299e-04, 1.15286244e-04, 1.07535803e-04, 7.51358052e-05,
       6.04492020e-05, 5.56629095e-05, 4.02002874e-05, 2.86356946e-05,
       2.81943295e-05, 2.36268123e-05, 1.68581382e-05, 1.22237827e-05,
       9.61078277e-06, 8.02281234e-06, 7.41437397e-06, 6.48373327e-06,
       5.89185300e-06, 5.03319932e-06, 4.29734441e-06, 4.02520426e-06,
       3.21574197e-06, 2.88267024e-06, 2.39658854e-06, 2.15449397e-06,
       1.80713264e-06, 1.69138958e-06, 1.43534631e-06, 1.38452533e-06,
       1.29747818e-06, 1.03500959e-06, 8.74622101e-07, 7.34054855e-07,
       6.28112018e-07, 5.84547223e-07, 3.94679313e-07, 3.11920911e-07,
       2.90822583e-07, 2.32147220e-07, 1.38675140e-07, 1.16019477e-07,
       7.74760356e-08, 4.76165897e-08, 2.62225376e-08, 1.23677119e-08,
       9.86463850e-32])
```

fig-26 Array for Explained Variance Ratio

Now we extract the 57 PCs into dataset and we will plot a scree plot for explained variance ratio as shown below.

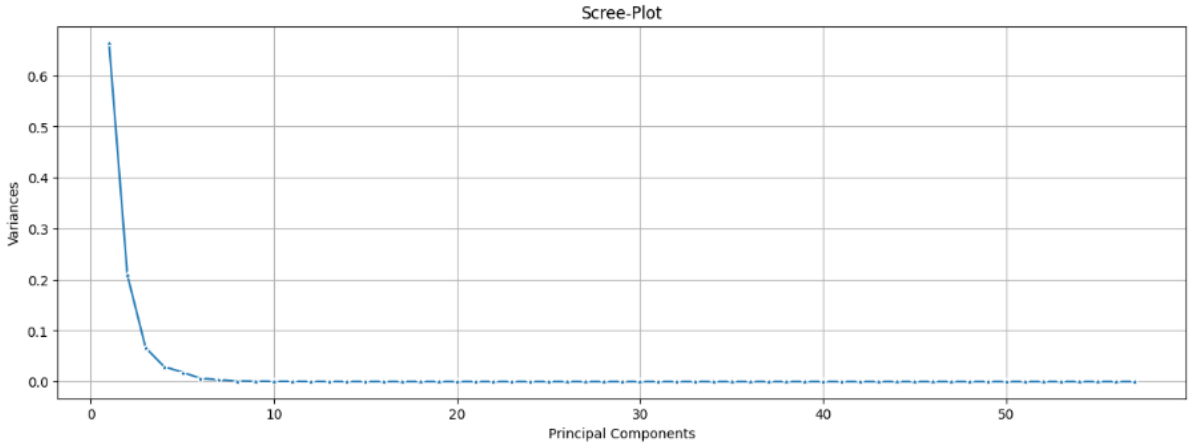


fig-27 Scree plot

But since we must cut down the PCs and take at least 90% of explained variance. We will calculate the cumulative summation of explained variance ratio and obtained the take below as well as plot a scree plot for this summation.

```
array([0.66330828, 0.87204383, 0.93814186, 0.96766102, 0.98590404,
       0.99225252, 0.99595653, 0.99721563, 0.99798692, 0.99846524,
       0.99889404, 0.99916401, 0.99936164, 0.99947693, 0.99958446,
       0.9996596 , 0.99972005, 0.99977571, 0.99981591, 0.99984455,
       0.99987274, 0.99989637, 0.99991323, 0.99992545, 0.99993506,
       0.99994308, 0.9999505 , 0.99995698, 0.99996287, 0.99996791,
       0.99997221, 0.99997623, 0.99997945, 0.99998233, 0.99998473,
       0.99998688, 0.99998869, 0.99999038, 0.99999181, 0.9999932 ,
       0.9999945 , 0.99999553, 0.99999641, 0.99999714, 0.99999777,
       0.99999835, 0.99999875, 0.99999906, 0.99999935, 0.99999958,
       0.99999972, 0.99999984, 0.99999991, 0.99999996, 0.99999999,
       1.        , 1.        ])
```

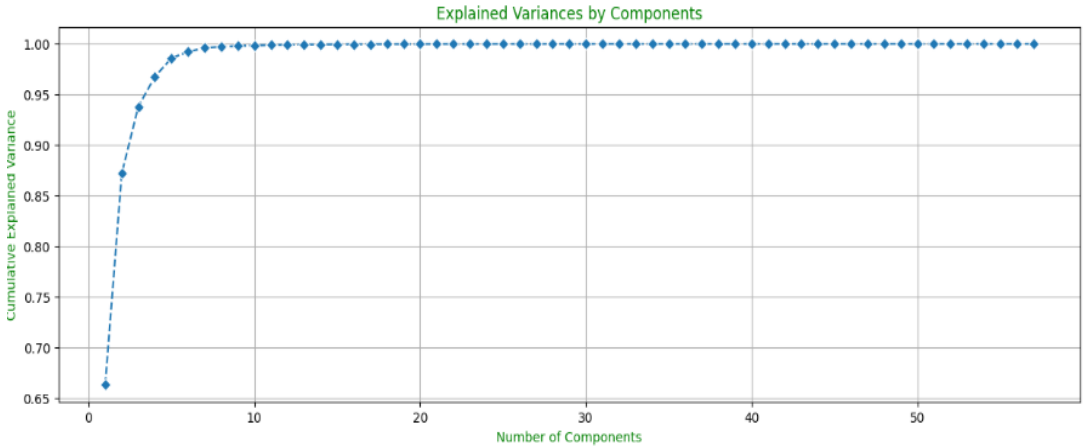


fig-28 Cumulative Explained variance ratio Scree plot

From the table and graph, we can conclude that we can cut the eigen vector from PC1 to PC6. And we create a new dataset with this six PCs. And we can see the Absolute loading of each PCs in the below plot.

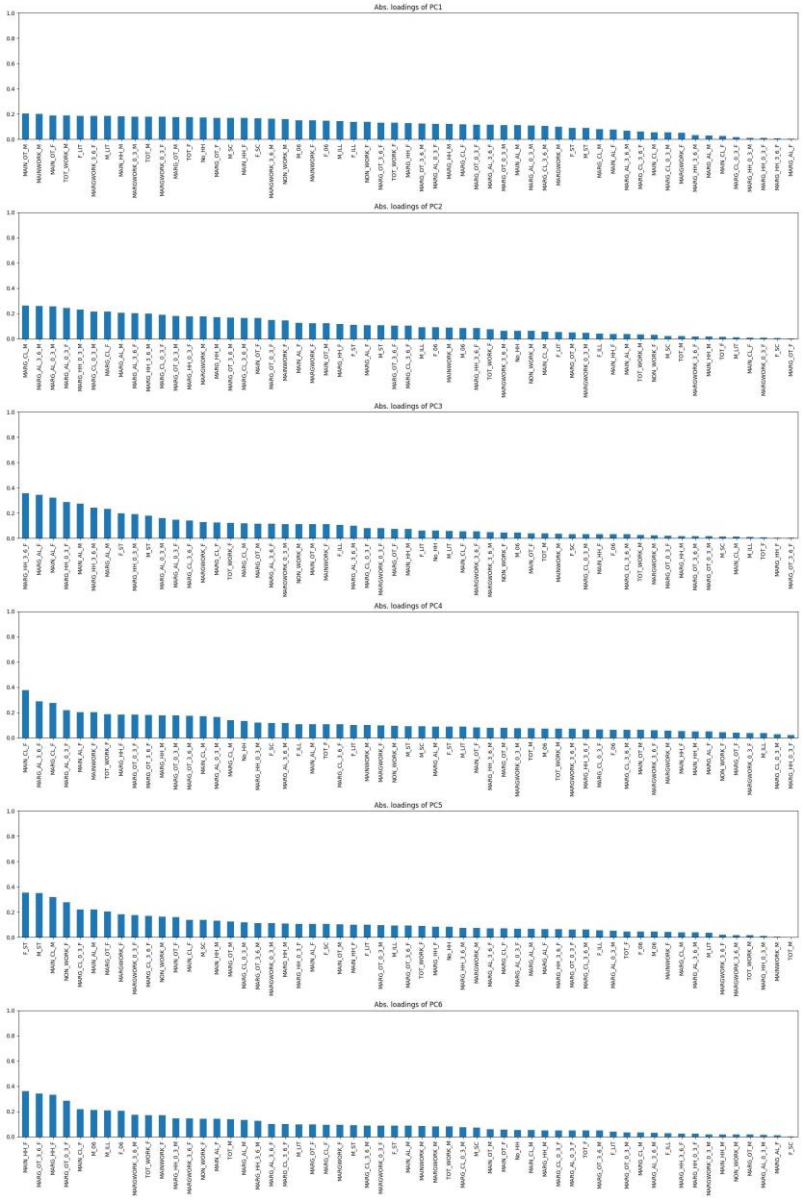


Fig-29 Absolute loading of PCs

The heatmap for the component loading is also shown below.

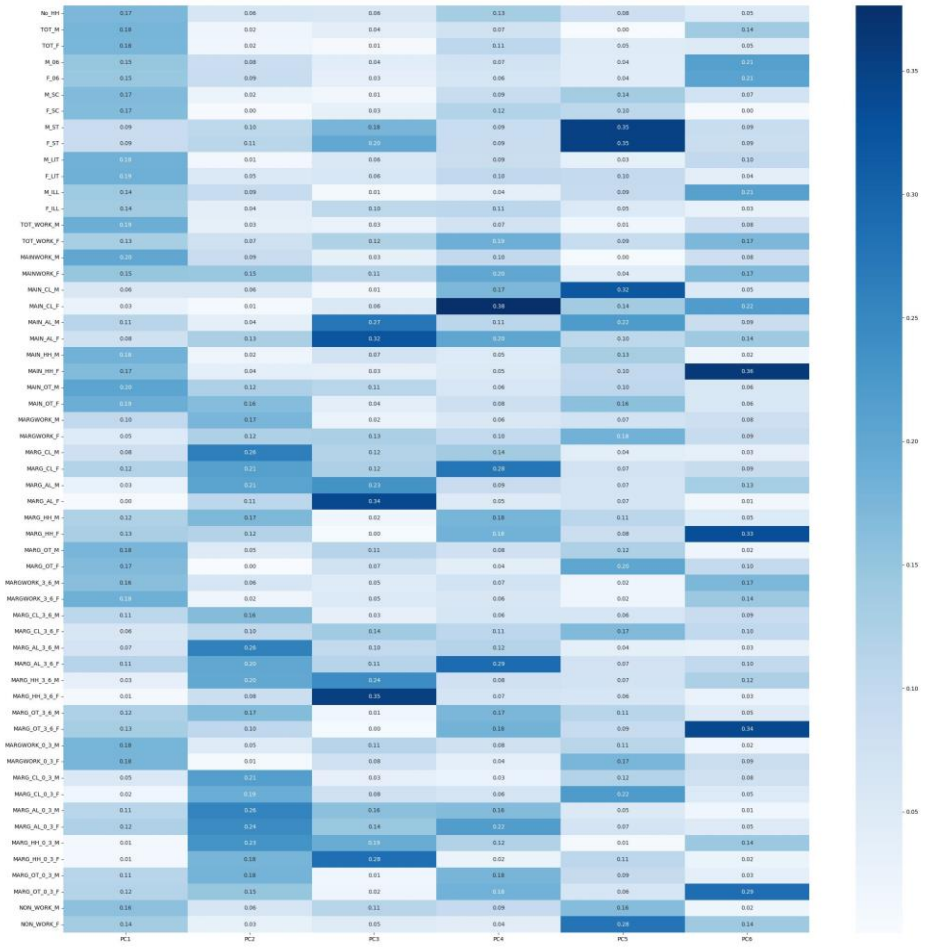


Fig-30 PCs heatmap

Now we perform a fit transform PCA on this six PCs and put them in a new dataset and obtain the following table.

	PC1	PC2	PC3	PC4	PC5	PC6
0	-1.141010	0.588138	0.164899	0.282802	-0.141400	0.057689
1	-1.348574	0.017570	-0.134072	0.175221	-0.011739	-0.135731
2	-1.341161	0.207062	0.027676	0.267599	-0.109592	-0.037613
3	-1.029050	-0.358927	-0.156145	0.164678	0.070711	-0.216209
4	-1.016880	-0.401646	-0.113016	0.151000	0.064642	-0.207636

Table- 16 Selected PCs Dataset

We will create a heatmap of correlations for this PCs and obtained the given below figure.

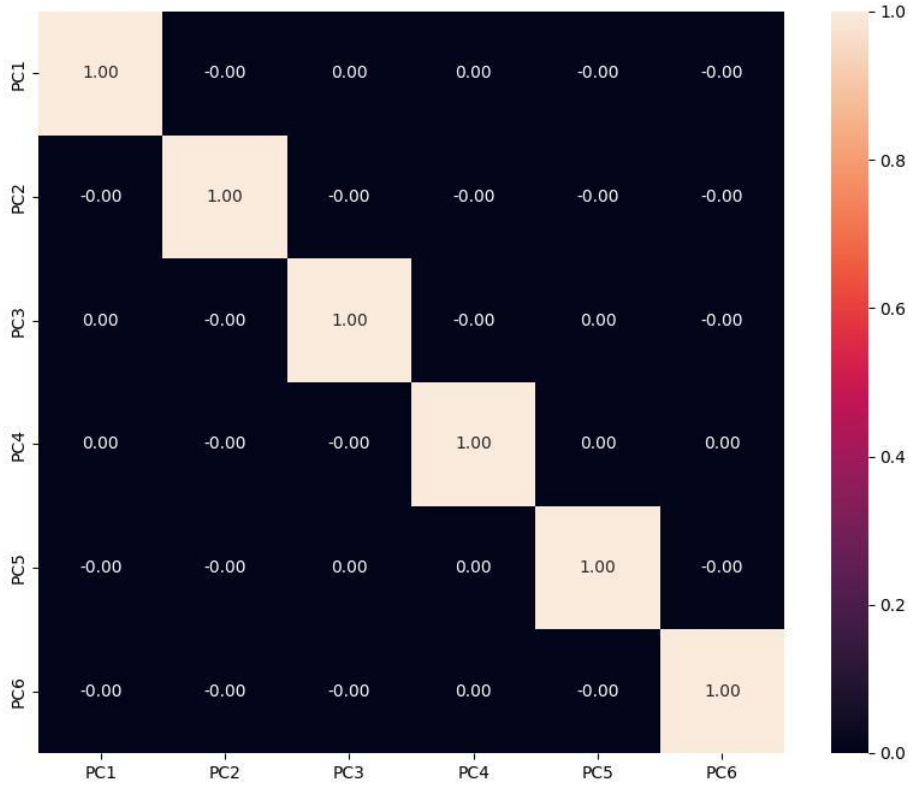


Fig-31 Selected PCs heatmap

Finally, we will write the Linear Equation for First Principal Components and we obtained the below equation shown in the figure.

The linear equation for the first PC with intercept is:
(-0.17) * No_HH + (-0.18) * TOT_M + (-0.18) * TOT_F + (-0.15) * M_06 + (-0.15) * F_06 + (-0.17) * M_SC + (-0.17) * F_SC + (0.09) * M_ST + (0.09) * F_ST + (-0.18) * M_LIT + (-0.19) * F_LIT + (-0.14) * M_ILL + (-0.14) * F_ILL + (-0.19) * TOT_WORK_M + (-0.13) * TOT_WORK_F + (-0.20) * MAINWORK_M + (-0.15) * MAINWORK_F + (-0.06) * MAIN_CL_M + (0.03) * MAIN_CL_F + (-0.11) * MAIN_AL_M + (-0.08) * MAIN_AL_F + (-0.18) * MAIN_HH_M + (-0.17) * MAIN_HH_F + (-0.20) * MAIN_OT_M + (-0.19) * MAIN_OT_F + (-0.10) * MARGWORK_M + (-0.05) * MARGWORK_F + (0.08) * MARG_CL_M + (0.12) * MARG_CL_F + (-0.03) * MARG_AL_M + (-0.00) * MARG_AL_F + (-0.12) * MARG_HH_M + (-0.13) * MARG_HH_F + (-0.18) * MARG_OT_M + (-0.17) * MARG_OT_F + (-0.16) * MARGWORK_3_6_M + (-0.18) * MARGWORK_3_6_F + (-0.11) * MARG_CL_3_6_M + (-0.06) * MARG_CL_3_6_F + (0.07) * MARG_AL_3_6_M + (0.11) * MARG_AL_3_6_F + (-0.03) * MARG_HH_3_6_M + (-0.01) * MARG_HH_3_6_F + (-0.12) * MARG_OT_3_6_M + (-0.13) * MARG_OT_3_6_F + (-0.18) * MARGWORK_0_3_M + (-0.18) * MARGWORK_0_3_F + (-0.05) * MARG_CL_0_3_M + (-0.02) * MARG_CL_0_3_F + (0.11) * MARG_AL_0_3_M + (0.12) * MARG_AL_0_3_F + (-0.01) * MARG_HH_0_3_M + (0.01) * MARG_HH_0_3_F + (-0.11) * MARG_OT_0_3_M + (-0.12) * MARG_OT_0_3_F + (-0.16) * NON_WORK_M + (-0.14) * NON_WORK_F + (3.63) = 0

Fig-32 PC-1 Linear equation