ML-1 GRADED PROJECT

NUPUR SARKAR

PGP-DSBA PGPDSBA.O.DEC23.A

Table of Contents

1 Clustering

1.1 Problem 1- Define the problem and perform Exploratory Data Analysis

1.1.1 Problem definition

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.

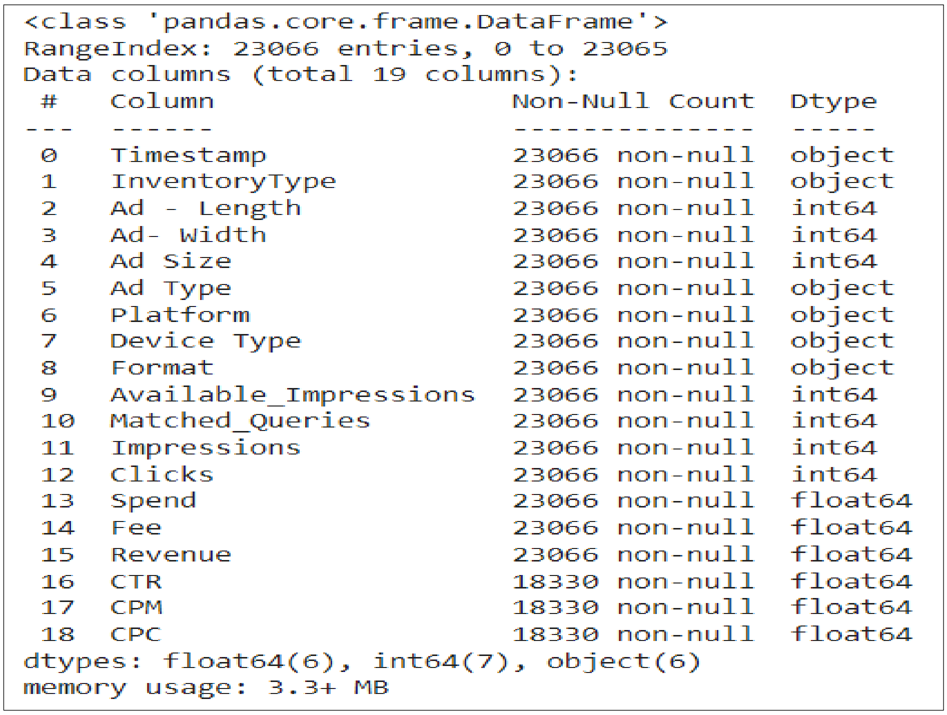
The Data Dictionary

Column Name	**Column Description**
Timestamp	The Timestamp of the particular Advertisement.
Inventory Type	The Inventory Type of the particular Advertisement. Format 1 to 7. This is a Categorical Variable.
Ad - Length	The Length Dimension of the particular Advertisement.
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 particular 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

1.1.2 Check shape, Data types, statistical summary



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23066 entries, 0 to 23065
Data columns (total 19 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   Timestamp              23066 non-null  object 
 1   InventoryType          23066 non-null  object 
 2   Ad - Length            23066 non-null  int64  
 3   Ad- Width              23066 non-null  int64  
 4   Ad Size                23066 non-null  int64  
 5   Ad Type                23066 non-null  object 
 6   Platform               23066 non-null  object 
 7   Device Type            23066 non-null  object 
 8   Format                 23066 non-null  object 
 9   Available_Impressions  23066 non-null  int64  
 10  Matched_Queries        23066 non-null  int64  
 11  Impressions            23066 non-null  int64  
 12  Clicks                 23066 non-null  int64  
 13  Spend                  23066 non-null  float64
 14  Fee                    23066 non-null  float64
 15  Revenue                23066 non-null  float64
 16  CTR                    18330 non-null  float64
 17  CPM                    18330 non-null  float64
 18  CPC                    18330 non-null  float64
dtypes: float64(6), int64(7), object(6)
memory usage: 3.3+ MB
```

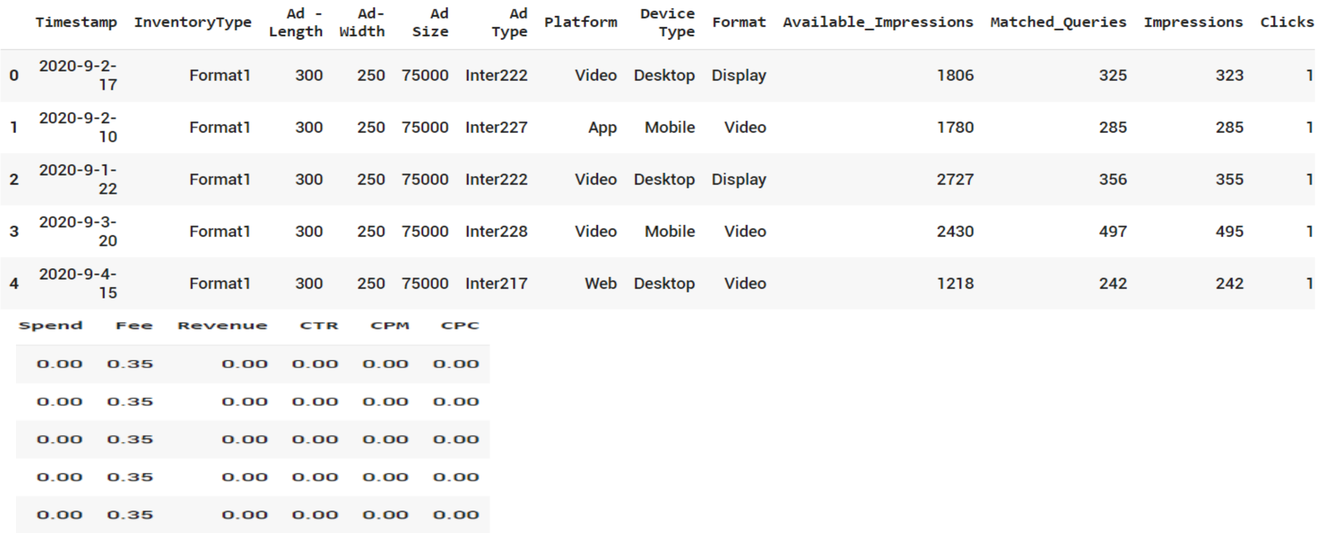
Figure 1

Observation 👀

-DATA SHAPE AND DATA TYPES

- Total 23066 entries, 19 columns(13 columns are numeric and 6 are non-numeric Type)
- 3 columns have some null values, lets fill them in next step with proper formula

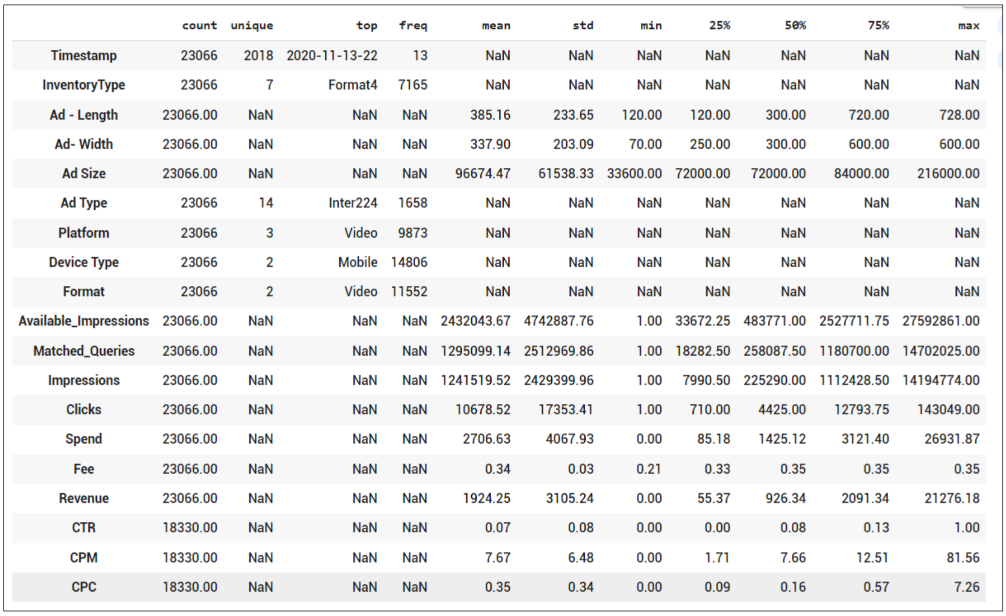
First Few records of the Dataset:



	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.00	0.35	0.00	0.00	0.00	0.00
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.00	0.35	0.00	0.00	0.00	0.00
2	2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.00	0.35	0.00	0.00	0.00	0.00
3	2020-9-3-20	Format1	300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.00	0.35	0.00	0.00	0.00	0.00
4	2020-9-4-15	Format1	300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.00	0.35	0.00	0.00	0.00	0.00

Figure 2

STATISTICAL SUMMARY



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Timestamp	23066	2018	2020-11-13-22	13	NaN	NaN	NaN	NaN	NaN	NaN	NaN
InventoryType	23066	7	Format4	7165	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Ad - Length	23066.00	NaN	NaN	NaN	385.16	233.65	120.00	120.00	300.00	720.00	728.00
Ad- Width	23066.00	NaN	NaN	NaN	337.90	203.09	70.00	250.00	300.00	600.00	600.00
Ad Size	23066.00	NaN	NaN	NaN	96674.47	61538.33	33600.00	72000.00	72000.00	84000.00	216000.00
Ad Type	23066	14	Inter224	1658	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Platform	23066	3	Video	9873	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Device Type	23066	2	Mobile	14806	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Format	23066	2	Video	11552	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Available_Impressions	23066.00	NaN	NaN	NaN	2432043.67	4742887.76	1.00	33672.25	483771.00	2527711.75	27592861.00
Matched_Queries	23066.00	NaN	NaN	NaN	1295099.14	2512969.86	1.00	18282.50	258087.50	1180700.00	14702025.00
Impressions	23066.00	NaN	NaN	NaN	1241519.52	2429399.96	1.00	7990.50	225290.00	1112428.50	14194774.00
Clicks	23066.00	NaN	NaN	NaN	10678.52	17353.41	1.00	710.00	4425.00	12793.75	143049.00
Spend	23066.00	NaN	NaN	NaN	2706.63	4067.93	0.00	85.18	1425.12	3121.40	26931.87
Fee	23066.00	NaN	NaN	NaN	0.34	0.03	0.21	0.33	0.35	0.35	0.35
Revenue	23066.00	NaN	NaN	NaN	1924.25	3105.24	0.00	55.37	926.34	2091.34	21276.18
CTR	18330.00	NaN	NaN	NaN	0.07	0.08	0.00	0.00	0.08	0.13	1.00
CPM	18330.00	NaN	NaN	NaN	7.67	6.48	0.00	1.71	7.66	12.51	81.56
CPC	18330.00	NaN	NaN	NaN	0.35	0.34	0.00	0.09	0.16	0.57	7.26

Figure 3

1. Timestamp: The recorded date and time. No outliers. Range: 2018 to 2020-11-13-22.
2. Inventory Type: Type of advertising format. No outliers. Range: 7 unique types.
3. Ad - Length: The length of advertisements. Outliers may exist beyond 720 characters. Range: 120 to 728 characters.
4. Ad - Width: The width of advertisements. No outliers. Range: 70 to 600 units.
5. Ad Size: The total size of advertisements. Outliers may exist beyond 84000 units. Range: 33600 to 216000 units.
6. Ad Type: Type of ad. No outliers. Range: 14 unique types.
7. Platform: The platform where ads are displayed. No outliers. Range: 3 unique types.
8. Device Type: Type of device where ads are viewed. No outliers. Range: 2 unique types.
9. Format: The format of ads. No outliers. Range: 2 unique types.
10. Available Impressions: The potential number of views for ads. Outliers may exist beyond 2527711.75. Range: 1 to 27592861 impressions.
11. Matched Queries: The number of user queries matched with ads. Outliers may exist beyond 1180700. Range: 1 to 14702025 queries.
12. Impressions: The actual number of times ads are viewed. Outliers may exist beyond 1112428.5. Range: 1 to 14194774 impressions.
13. Clicks: The number of times users clicked on ads. Outliers may exist beyond 12793.75. Range: 1 to 143049 clicks.
14. Spend: The amount of money spent on advertising. Outliers may exist beyond 3121.40.Range:0 to $26931.87.
15. Fee: The fee associated with advertising. No outliers. Range: 0.21to0.35.
16. Revenue: The earnings generated from ads. Outliers may exist beyond 2091.34.Range:0 to $21276.18.
17. CTR (Click-Through Rate): The percentage of viewers who clicked on ads. Outliers may exist beyond 0.13. Range: 0 to 1.00.

18. CPM (Cost Per Mille): The cost advertisers pay for every thousand views. Outliers may exist beyond 12.51.Range:0 to $81.56.
19. CPC (Cost Per Click): The cost advertisers pay for each click. Outliers may exist beyond 0.57.Range:0 to $7.26

1.1.3 Univariate analysis

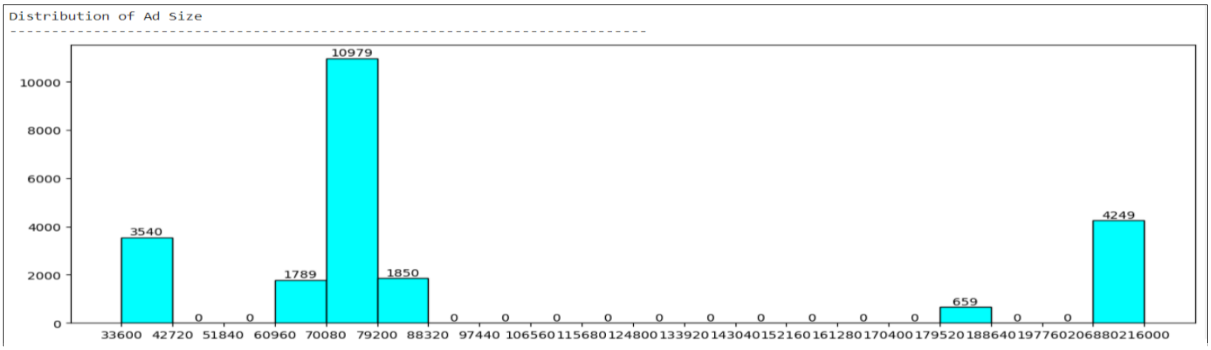


Figure 4

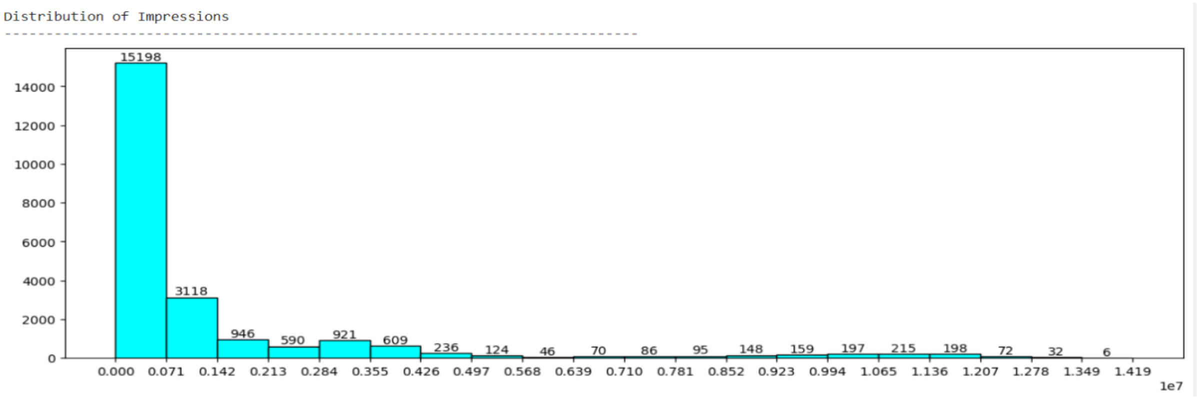


Figure 5

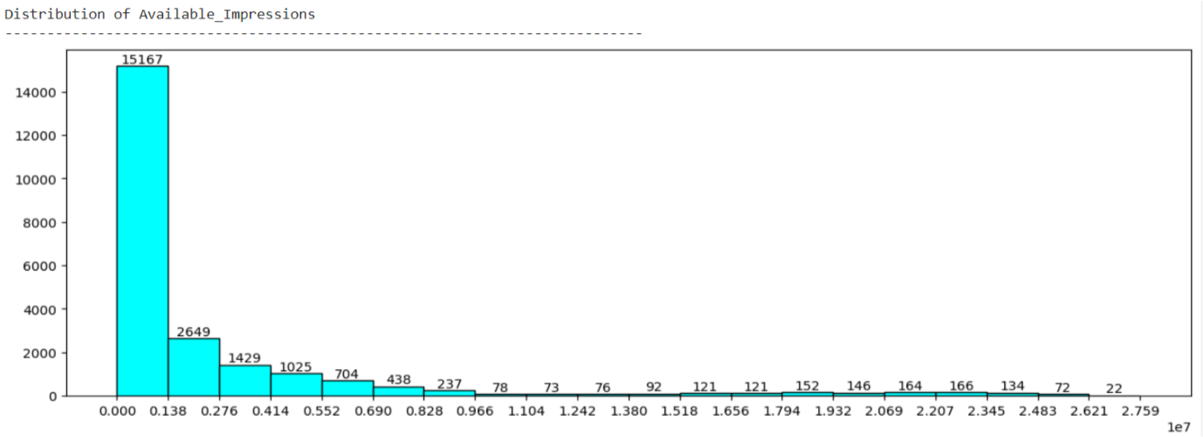


Figure 6

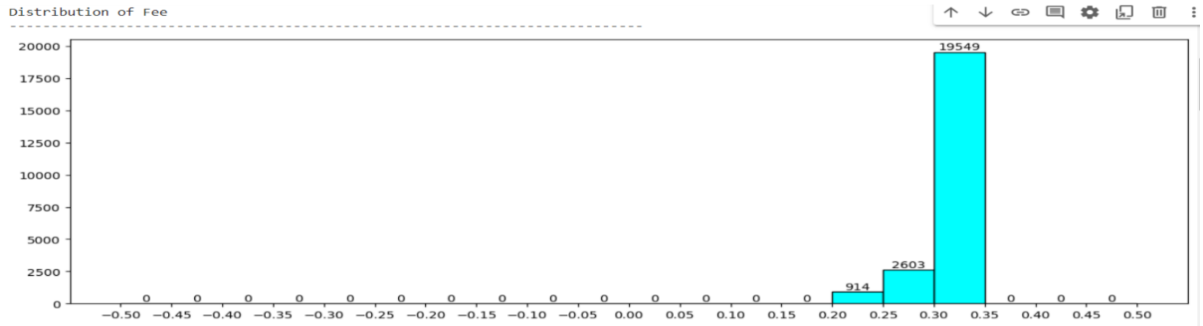


Figure 7

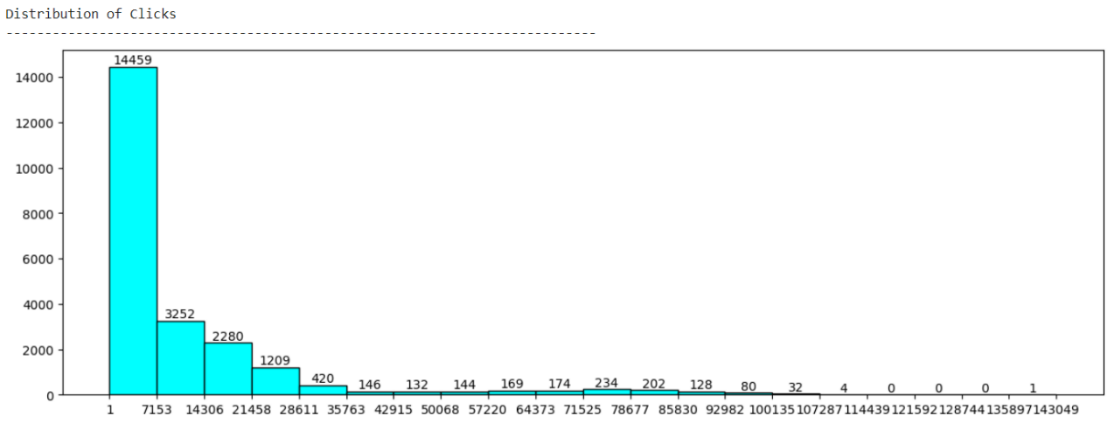


Figure 8

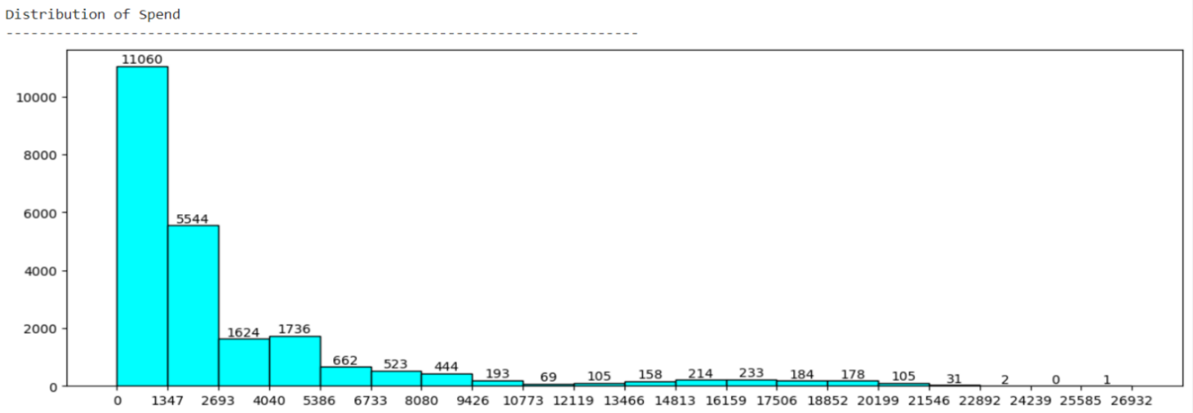


Figure 9

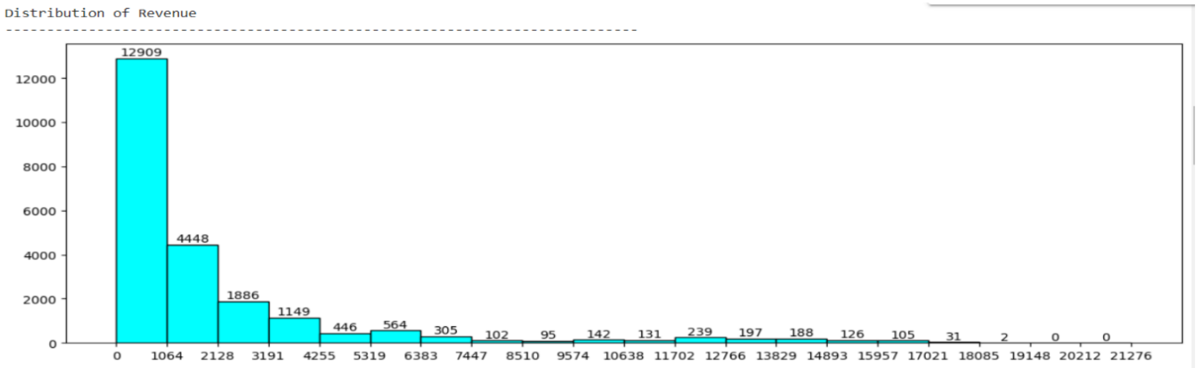


Figure 10

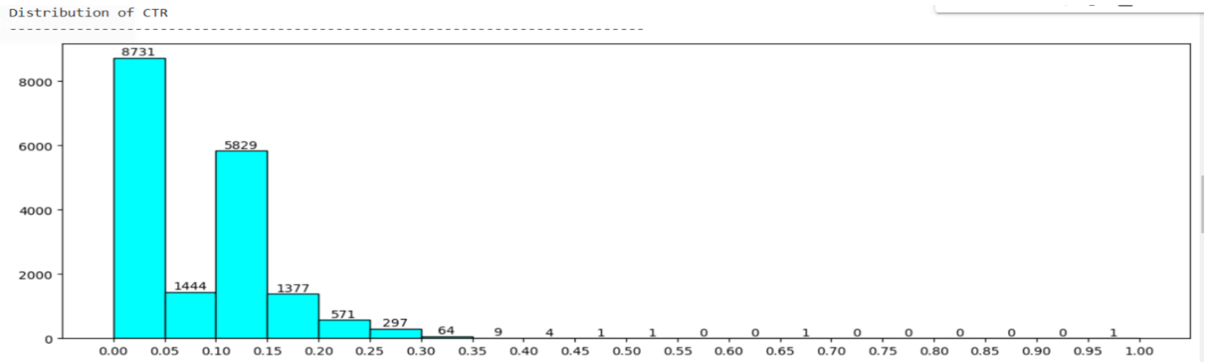


Figure 12

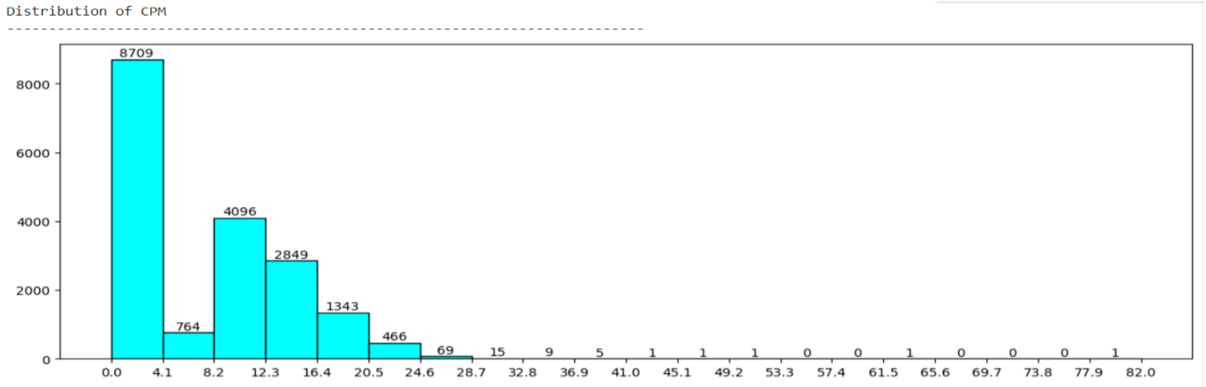


Figure 11

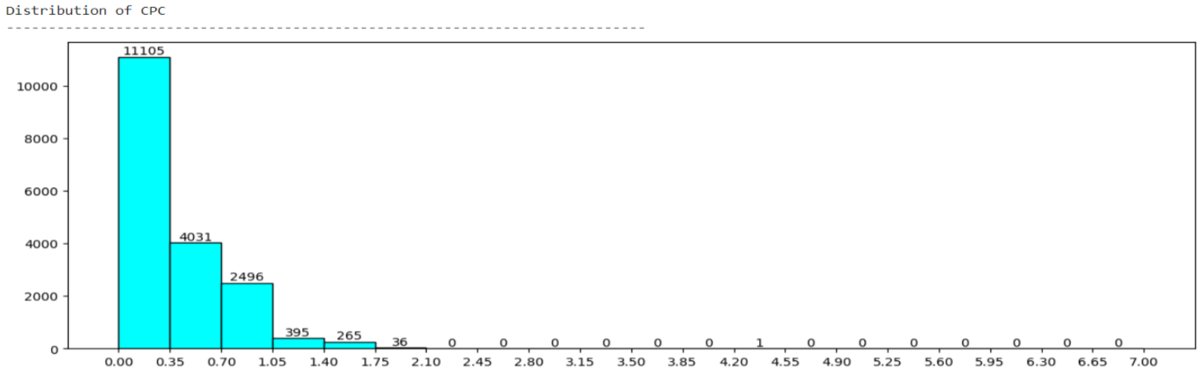


Figure 13

Observation 👀

Ad Length, Width, and Size: These histograms show the distribution of ad dimensions. We can observe the range and frequency of different ad sizes

Available Impressions, Impressions, and Clicks: These histograms display the availability of impressions, the actual number of impressions, and the number of clicks on the advertisement. We can see how often the ad was shown, how many times it was clicked, and its engagement level.

Spend, Revenue, and Fee: These histograms represent the spending, revenue, and fee associated with the advertisement. We can analyse the financial aspects of the advertising campaign.

CTR, CPM, and CPC: These histograms illustrate the click-through rate, cost per 1000 impression, and cost per click. They provide insights into the effectiveness and cost-efficiency of the advertisement.

In Detail:

1. Ad - Length:
   - The ad lengths are spread out with a mean of 385.16 and a standard deviation of 233.65.
   - The distribution appears to be right-skewed, as the mean is less than the median (300.00), indicating that there are longer ad lengths that pull the mean to the right.
2. Ad - Width:
   - The ad widths also show variability, with a mean of 337.90 and a standard deviation of 203.09.
   - Similar to ad length, the distribution seems right-skewed, as the mean (337.90) is less than the median (300.00), suggesting the presence of wider ad widths.

3. Ad Size:
    - Ad sizes exhibit significant variability, ranging from 33,600 to 216,000, with a mean of 96,674.47 and a standard deviation of 61,538.33.
    - The distribution of ad sizes appears to be moderately right-skewed, with a clustering around the median value of 72,000.
4. Available Impressions:
    - The available impressions vary widely, ranging from 1 to 27,592,861, with a mean of 2,432,043.67 and a standard deviation of 4,742,887.76.
    - The distribution of available impressions seems highly right-skewed, with a clustering of values towards the lower end of the range.
5. Matched Queries, Impressions, Clicks, Spend, Fee, Revenue:
    - These features also show variability in their spread, with differing means and standard deviations.
6. CTR, CPM, CPC:

- These features' distributions are not described in terms of percentiles or minimum/maximum values. Most of the Ads have CPC 35% , CPM 4.1 and CTR 5% OR 0.05

In the Python sheet Box plot is also given. But since its explanation would have been almost similar as above , So haven’t used here those pictures.

1.1.4 Bivariate analysis

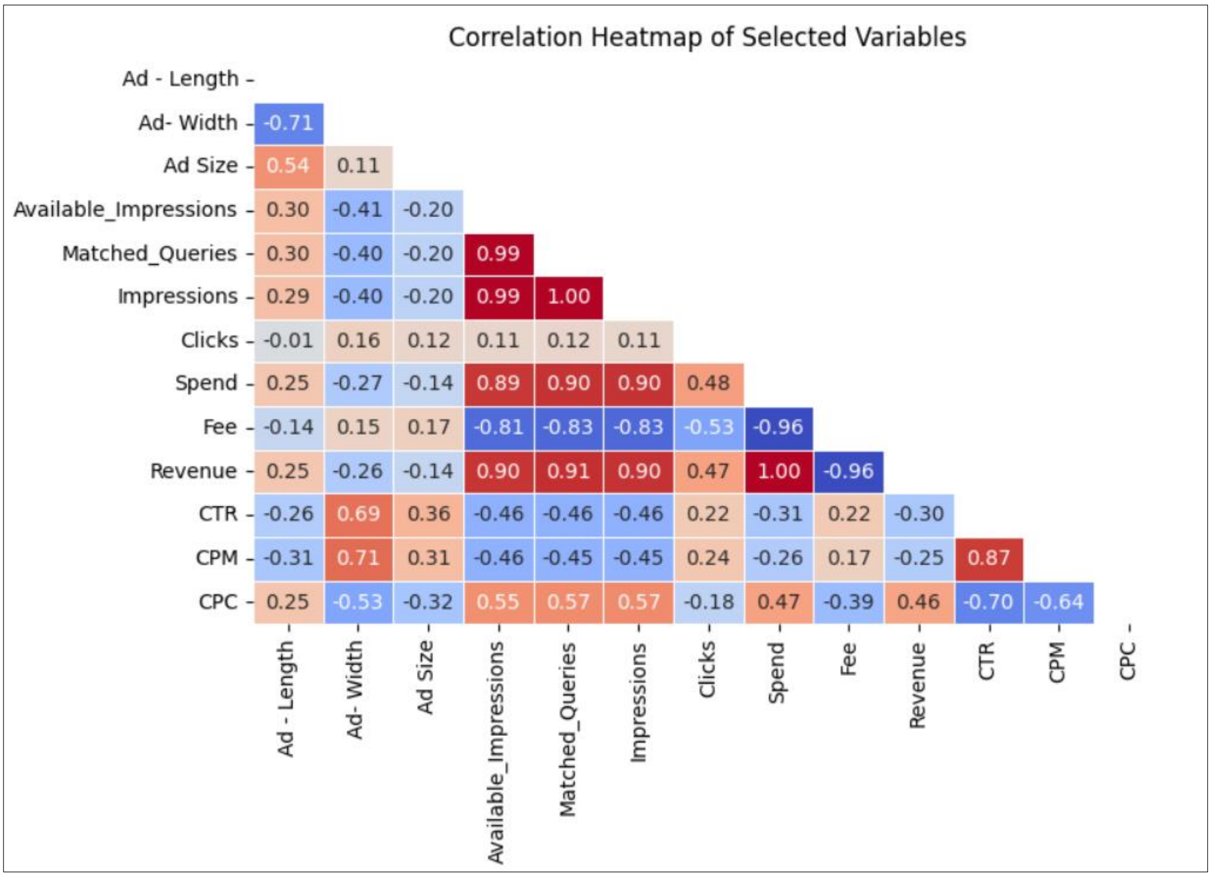


Figure 14

Insights 👀

Spend and Available Impressions, Matched Queries, Impressions: Spend is highly correlated with these variables, it suggests that as the number of Available Impressions, Matched Queries, or Impressions increases, the amount spent on advertising tends to increase as well. In other words, when there are more opportunities for the ad to be seen by users (impressions) or when the ad matches more search queries, companies tend to spend more money on advertising.

Revenue and Available Impressions, Matched Queries, Impressions, and Fee: Revenue is highly correlated with these variables, it means that as the number of Available Impressions, Matched Queries, or Impressions increases, the revenue earned from the advertising tends to increase. Additionally, if the fee is also correlated, it suggests that the amount of revenue earned is influenced by the fees associated with advertising. So, when there are more opportunities for the ad to be seen and clicked (impressions, matched queries), the revenue tends to rise, but it may also be affected by the fees paid.

CTR and Ad-Width: CTR is highly correlated with Ad-Width, it means that the click-through rate (CTR) tends to vary with changes in the width of the ad. In simple terms, wider ads might attract more clicks compared to narrower ads.

CPM and Ad-Width, CTR: CPM is highly correlated with Ad-Width and CTR, it suggests that the cost per thousand impressions (CPM) tends to be influenced by both the width of the ad and the click-through rate. This could mean that wider ads and ads with higher click-through rates may have higher costs per thousand impressions.

CPC and CTR, CPM: CPC is highly correlated with CTR and CPM, it means that the cost per click (CPC) tends to be influenced by both the click-through rate and the cost per thousand impressions. In simpler terms, the cost of each click may be higher for ads that have higher click-through rates or higher costs per thousand impressions. This suggests that advertisers may pay more for clicks on ads that are more likely to attract user engagement

1.1.5 Key meaningful observations on individual variables and the relationship between variables

Solution

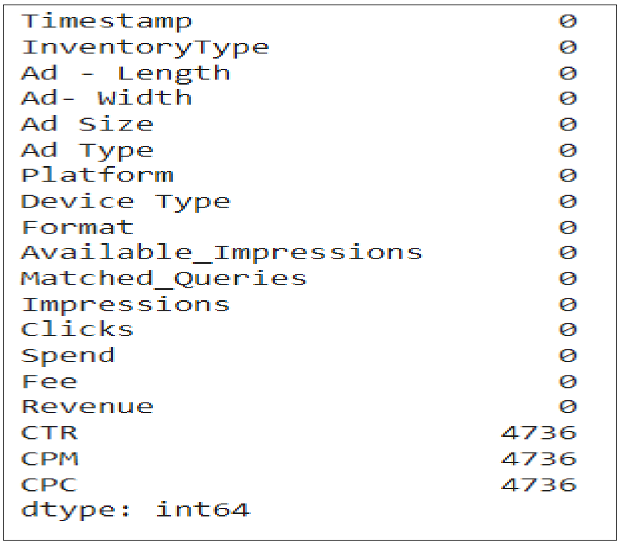
- The distribution of ad sizes appears to be moderately right-skewed, with a clustering around the median value of 72,000
- The distribution of available impressions seems highly right-skewed, with a clustering of values towards the lower end of the range.
- These features' distributions are not described in terms of percentiles or minimum/maximum values. Most of the Ads have CPC 35% , CPM 4.1 and CTR 5% OR 0.05
- When there are more opportunities for the ad to be seen by users (impressions) or when the ad matches more search queries, companies tend to spend more money on advertising
- The amount of revenue earned is influenced by the fees associated with advertising. So, when there are more opportunities for the ad to be seen and clicked (impressions, matched queries), the revenue tends to rise, but it may also be affected by the fees paid
- Wider ads might attract more clicks compared to narrower ads
- Wider ads and ads with higher click-through rates may have higher costs per thousand impressions
- Advertisers may pay more for clicks on ads that are more likely to attract user engagement

1.2 Problem 1 - Data Preprocessing

1.2.1 Missing value check and treatment

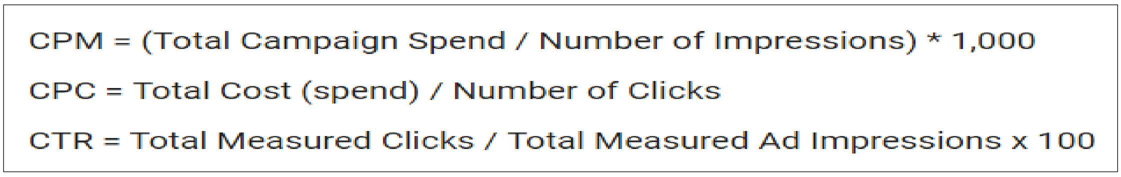
Observation 👀

Column CTR, CPM and CPC has 4736 missing values each. Filling those with the values fetched from the formula below:



```
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
```

Figure 15



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

CPC = Total Cost (spend) / Number of Clicks

CTR = Total Measured Clicks / Total Measured Ad Impressions x 100

Figure 16

1.2.2 Outlier Treatment

Solution

Let’s first check the outliers in each variable:Insights 👀

1. Ad Length & Width:

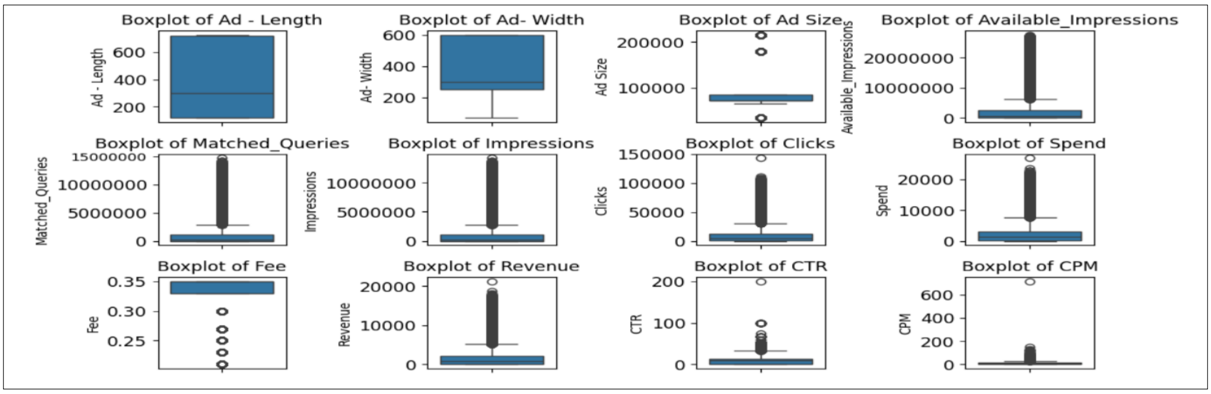


Figure 17

  - Both Ad length and width columns do not have outliers, with all values falling within the expected range.

2. Ad Size:
   - There are likely outliers with exceptionally large ad sizes, with a maximum of 216,000 square pixels. These outliers may represent ads with significantly larger dimensions compared to the majority.
3. Available Impressions:
   - Some ads may have exceptionally high numbers of available impressions, as indicated by the maximum value of 27,592,861. These outliers may represent ads with widespread reach or high visibility.
4. Matched Queries:
   - Similarly, there are likely outliers with a large number of matched queries, with a maximum of 14,702,025, indicating potential outliers with a significant amount of user engagement or interest.
5. Impressions:
   - Some ads may have exceptionally high numbers of impressions, as indicated by the maximum value of 14,194,774, suggesting outliers with widespread exposure.
6. Clicks:
   - There are likely outliers with a large number of clicks, with a maximum of 143,049, indicating potential outliers with high user engagement or effectiveness.
7. Spend:
   - Outliers may exist with unusually high spending on ads, as shown by the maximum spend of $26,931.87, potentially indicating campaigns with substantial investment or high-cost advertising strategies.
8. Fee:
   - The fee column shows little variability, with all values concentrated around 0.34%. However, there may be outliers with fees at either end of the range (e.g., 0.21% or 0.35%).
9. Revenue:
   - Some ads may generate exceptionally high revenue, as indicated by the maximum revenue of $21,276.18, suggesting potential outliers with highly profitable ad campaigns.
10. CTR (Click-through Rate):
    - Outliers may exist with unusually high or low click-through rates, as shown by the wide range from 0.01% to 200%. These outliers may represent ads with either exceptional effectiveness or poor performance.
11. CPM (Cost Per Mille):
    - There may be outliers with very high costs per thousand impressions, as indicated by the maximum value of $715.00, potentially representing ads with disproportionately high advertising costs.
12. CPC (Cost Per Click):
    - Outliers may exist with unusually high costs per click, as shown by the maximum value of $7.26, indicating potential outliers with costly advertising strategies.

Now, Let’s see the variables after removing outliers and get an insight on before and after treating outliers:

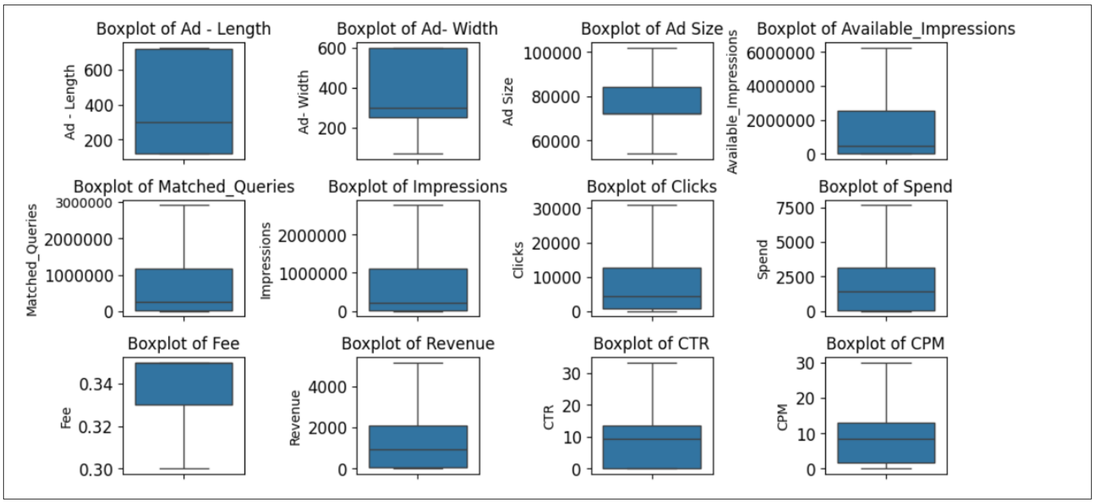


Figure 18

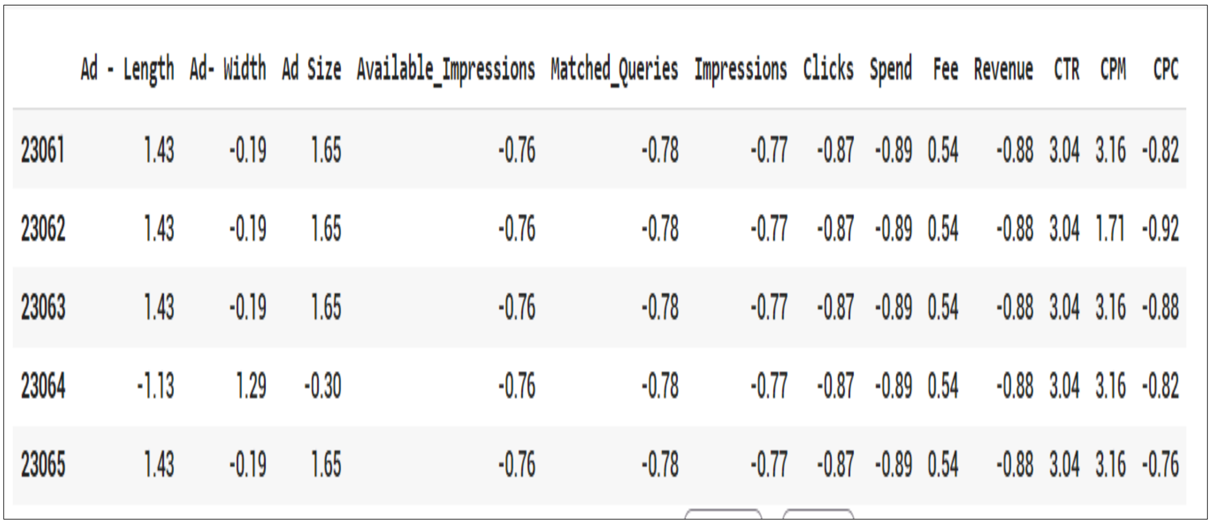
Insights 👀:

1. Ad Size:
   - The maximum ad size was 216,000 square pixels when outliers were not removed, while in the current table, the maximum ad size is reduced to 102,000 square pixels. This suggests a decrease in the occurrence of outliers with exceptionally large ad sizes.
2. Available Impressions:
   - The maximum number of available impressions was 27,592,861 when outliers were not removed, which is significantly higher than the maximum of 6,268,771 in the current table. This indicates a reduction in the occurrence of outliers with extremely high numbers of available impressions.
3. Matched Queries:
   - Similar to available impressions, the maximum number of matched queries was 14,702,025 when outliers were not removed, compared to 2,924,326 in the current table, indicating a decrease in outliers with exceptionally high numbers of matched queries.
4. Impressions:
   - The maximum number of impressions was 14,194,774 when outliers were not removed, which is higher than the maximum of 2,769,085.50 in the current table. This suggests a reduction in outliers with extremely high numbers of impressions.
5. Clicks:
   - The maximum number of clicks was 143,049 when outliers were not removed, compared to 30,919.38 in the current table, indicating a decrease in outliers with exceptionally high numbers of clicks.

6. Spend:
   - The maximum spending on ads was 26,931.87when outliers were not removed, whichishigherthanthemaximumof7,675.73 in the current table. This indicates a reduction in outliers with exceptionally high spending.
7. CTR (Click-through Rate):
   - The maximum CTR was 200.00% when outliers were not removed, compared to 33.28% in the current table, indicating a decrease in outliers with exceptionally high click-through rates.
8. CPM (Cost Per Mille):
   - Similarly, the maximum CPM was 715.00whenoutlierswerenotremoved, comparedto29.98 in the current table, suggesting a reduction in outliers with exceptionally high costs per thousand impressions.
9. CPC (Cost Per Click):
   - The maximum CPC was 7.26 when outliers were not removed, compared to 1.23 in the current table, indicating a decrease in outliers with exceptionally high costs per click.

1.2.3 Z-score scaling Note: Treat missing values in CPC, CTR and CPM using the formula given.

Z-score scaling is shown below and also missing values in CPC, CTR and CPM are also filled with the values generated form the formula. Let’s see the last few rows in the dataset after scaling



	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
23061	1.43	-0.19	1.65	-0.76	-0.78	-0.77	-0.87	-0.89	0.54	-0.88	3.04	3.16	-0.82
23062	1.43	-0.19	1.65	-0.76	-0.78	-0.77	-0.87	-0.89	0.54	-0.88	3.04	1.71	-0.92
23063	1.43	-0.19	1.65	-0.76	-0.78	-0.77	-0.87	-0.89	0.54	-0.88	3.04	3.16	-0.88
23064	-1.13	1.29	-0.30	-0.76	-0.78	-0.77	-0.87	-0.89	0.54	-0.88	3.04	3.16	-0.82
23065	1.43	-0.19	1.65	-0.76	-0.78	-0.77	-0.87	-0.89	0.54	-0.88	3.04	3.16	-0.76

Figure 19

1.3 Problem 1 - Hierarchical Clustering

1.3.1 Construct a dendrogram using Ward linkage and Euclidean distance

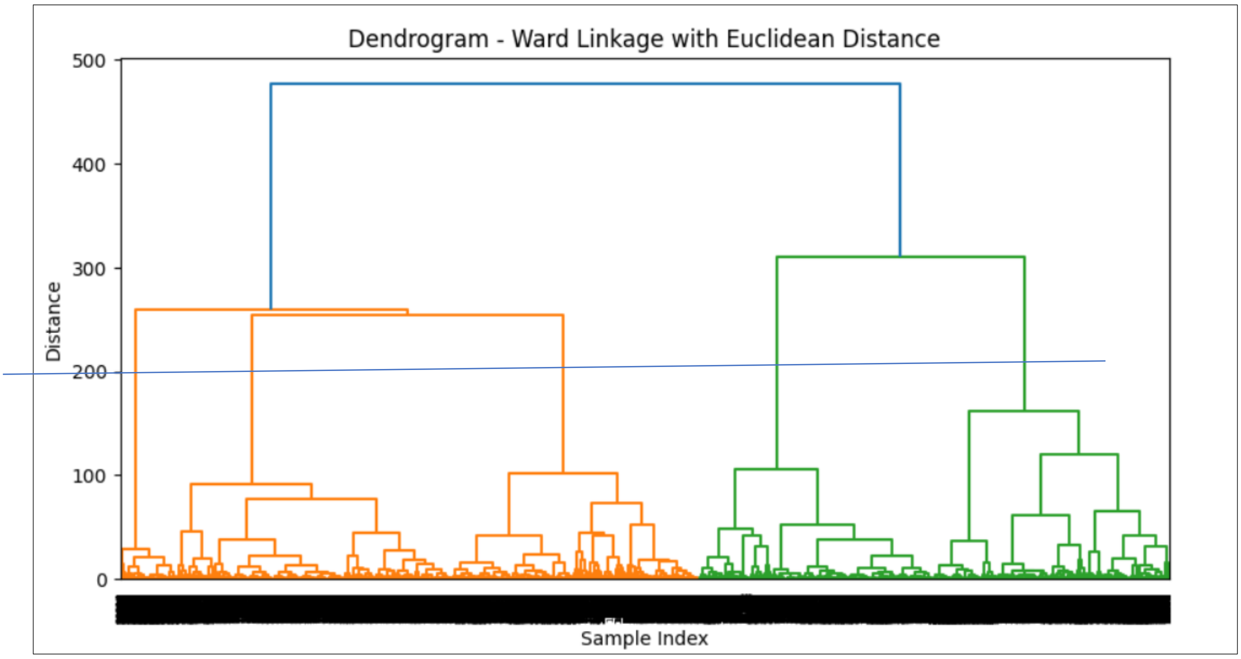


Figure 20

We have constructed Dendrogram here and the number of clusters can be seen as 5 through the blue horizontal line that's cutting the histograms at 5 points.

1.3.2 Identify the optimum number of Clusters

To figure out the optimum number of clusters, we can use methods like the Elbow Method or the Silhouette Score. 1.Elbow Method: The Elbow Method helps to find the optimal number of clusters by plotting the within-cluster sum of squares (WCSS) against the number of clusters. The "elbow" point in the plot indicates the optimal number of clusters

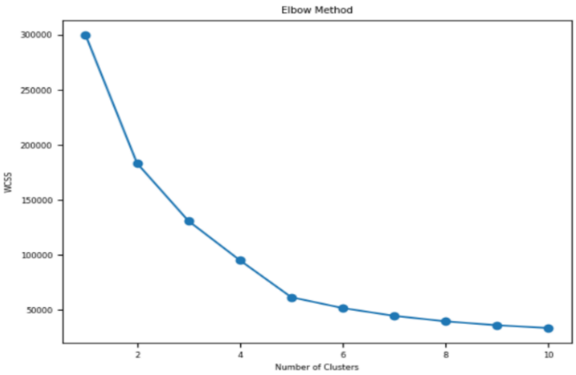


Figure 21

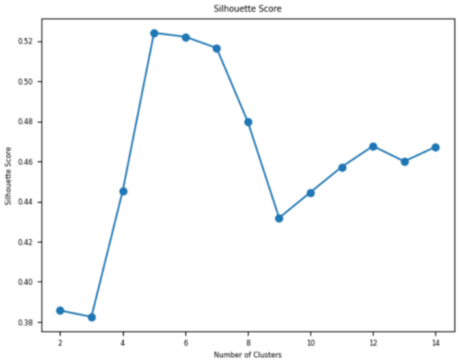
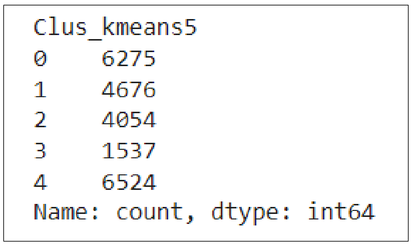


Figure 22

2.Silhouette Score: The Silhouette Score measures how similar an object is to its own cluster compared to other clusters. The score ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters

1.4 Problem 1 - K-means Clustering

1.4.1 Apply K-means Clustering

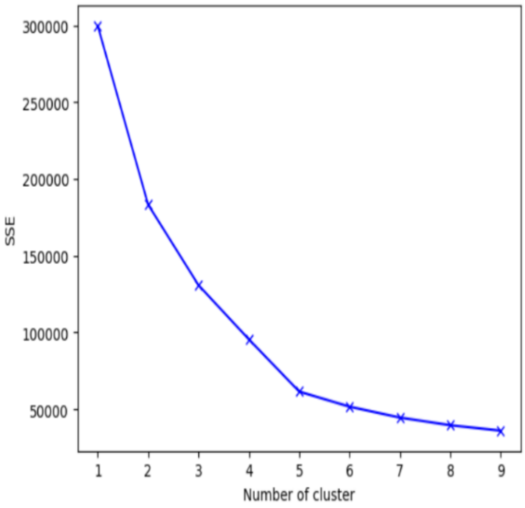


```
Clus_kmeans5
0    6275
1    4676
2    4054
3    1537
4    6524
Name: count, dtype: int64
```

Figure 23

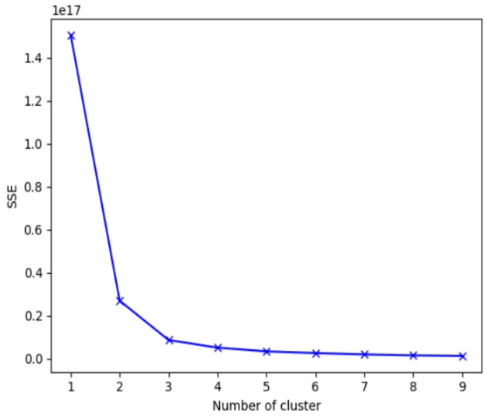
k-means clustering done and results of it is 5 clusters are formed with the bifurcation of observation given in each in above figure.

1.4.2 Plot the Elbow curve - Check Silhouette Scores



```
Elapsed time for scaled dataset: 7.76 seconds
```

Figure 24



```
Elapsed time for not scaled Dataset: 16.54 seconds
```

Figure 25

Insights 👀

Both the curves give different number of optimal clusters as we can see. The unscaled data gives 3 clusters and the scaled data gives 5 as optimal cluster. Also, the time consumed in unscaled data for clustering is taking more than the scaled data. Hence, in K-means scaling of data is very important.

1.4.3 Figure out the appropriate number of clusters

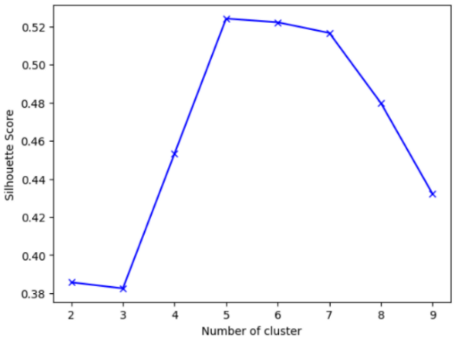
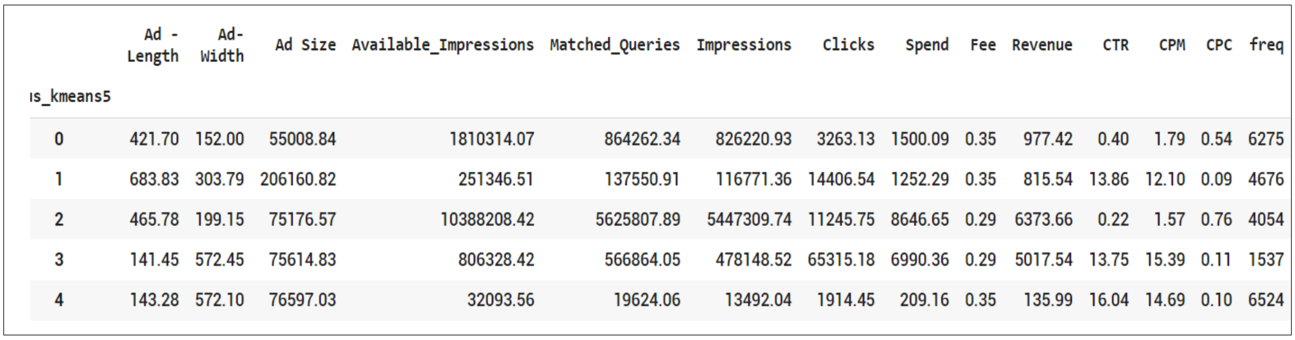


Figure 26

As we can see from the graph above, Silhouette score is highest for no of clusters as 5 and also inertia gets stagnant after cluster 5. It is 0.52 which is near to 1 which state that clusters are properly separated from each other. Hence the optimal number of clusters must be 5

1.4.4 Cluster Profiling



	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	freq
ıs_kmeans5														
0	421.70	152.00	55008.84	1810314.07	864262.34	826220.93	3263.13	1500.09	0.35	977.42	0.40	1.79	0.54	6275
1	683.83	303.79	206160.82	251346.51	137550.91	116771.36	14406.54	1252.29	0.35	815.54	13.86	12.10	0.09	4676
2	465.78	199.15	75176.57	10388208.42	5625807.89	5447309.74	11245.75	8646.65	0.29	6373.66	0.22	1.57	0.76	4054
3	141.45	572.45	75614.83	806328.42	566864.05	478148.52	65315.18	6990.36	0.29	5017.54	13.75	15.39	0.11	1537
4	143.28	572.10	76597.03	32093.56	19624.06	13492.04	1914.45	209.16	0.35	135.99	16.04	14.69	0.10	6524

Figure 27

Above we can see profiling of Ads in clusters 0 to 4 in 5 clusters

1.5 Problem 1 - Actionable Insights & Recommendations

1.5.1 Extract meaningful insights (at least 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 provided, we have five distinct clusters (Clus_kmeans5) with different characteristics. Here are some meaningful insights and actionable recommendations based on each cluster:

Cluster 0:

Insights: This cluster has ads with moderate Ad Length and Ad Width but relatively low Ad Size. They have a moderate number of Available Impressions and Impressions. Clicks, Spend, and Revenue is also moderate.

Cluster 1:

Insights: This cluster has ads with relatively high Ad Length and Ad Width, resulting in a significantly higher Ad Size compared to other clusters. They also have the highest Clicks, Spend, and Revenue among all clusters.

Cluster 2:

Insights: This cluster represents ads with moderate Ad Length and Ad Width but significantly higher Ad Size. They have the highest Available Impressions and Impressions among all clusters, but Clicks, Spend, and Revenue is relatively lower.

Cluster 3:

Insights: This cluster has ads with the lowest Ad Length and moderate Ad Width, resulting in relatively lower Ad Size. However, they have the highest Clicks among all clusters, indicating high engagement.

Cluster 4:

Insights: This cluster consists of ads with the lowest Ad Length and Ad Width, resulting in the smallest Ad Size. They also have the lowest Available Impressions and Impressions, but Clicks, Spend, and Revenue are relatively higher compared to some other clusters.

1.5.2 Based on the clustering analysis and key insights, provide actionable recommendations (at least 3) to Ads24x7 on how to optimize their digital marketing efforts, allocate budgets efficiently, and tailor ad content to specific audience segments.

Cluster 0

Recommendations: Optimize ads in this cluster to increase Ad Size without compromising on Ad Length and Width to attract more attention. Focus on targeting platforms or devices where these ads perform well to maximize impressions and clicks. Experiment with different ad formats and content to increase engagement and revenue from this cluster.

Cluster 1

Recommendations: Invest more budget in advertising formats similar to those in this cluster as they have shown to be highly effective in driving clicks and revenue. Focus on targeting specific audience segments or platforms where these ads perform exceptionally well to maximize ROI. Continuously monitor and analyse the performance of ads in this cluster to identify any changes in audience behaviour or preferences.

Cluster2

Recommendations: Explore ways to increase engagement and conversion rates for ads in this cluster to capitalize on the large number of impressions. Consider adjusting targeting strategies to reach more relevant audiences who are likely to convert. Experiment with different ad placements and creatives to improve click-through rates and overall performance.

Cluster 3

Recommendations: Focus on optimizing ad content to maintain high engagement levels while also increasing Ad Size to potentially drive higher revenue. Allocate additional budget to platforms or devices where ads in this cluster perform exceptionally well to further increase clicks and conversions. Implement retargeting campaigns to capitalize on the high engagement levels and drive repeat conversions from users who have previously interacted with ads in this cluster.

Cluster 4

Recommendations: Consider optimizing ad targeting to reach a more relevant audience and increase impressions while maintaining or improving click-through rates. Experiment with different ad formats or placements to increase visibility and reach a larger audience without compromising on engagement. Focus on optimizing conversion funnels to maximize revenue from the clicks generated by ads in this cluster. Overall, the recommendations aim to optimize digital marketing efforts by tailoring ad content, targeting strategies, and budget allocation to specific audience segments identified through clustering analysis, ultimately improving ROI and maximizing revenue for Ads24x7.

2 PCA

2.1 Problem 2 - Define the problem and perform Exploratory Data Analysis

2.1.1 Problem Definition

Solution

Context 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

Objective Perform detailed EDA and identify Optimum Principal Components that explains the most variance in data

Data Description: 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.

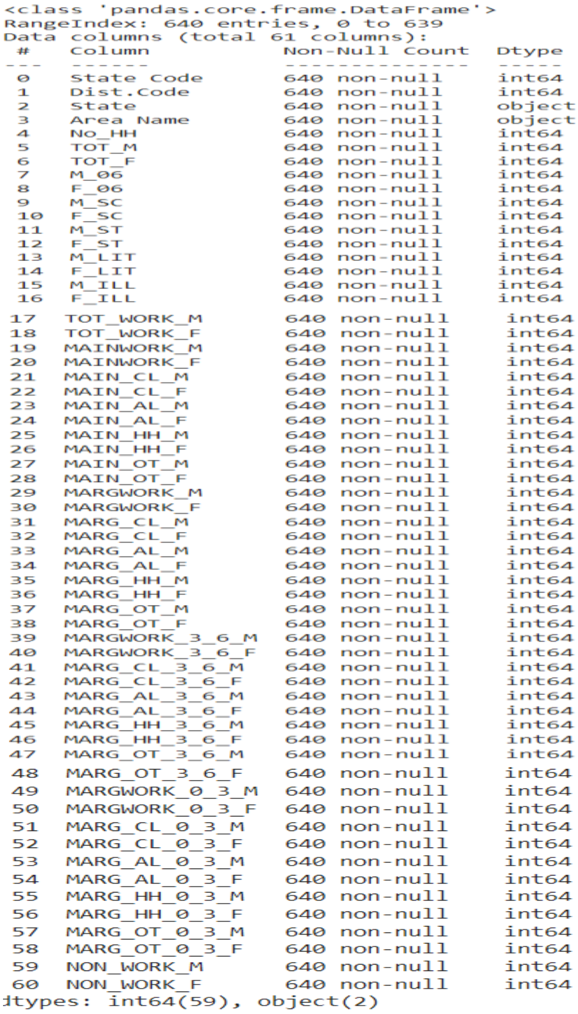
Data Dictionary

Name	Description
State	State Code
District	District Code
Name	Name
TRU1	Area Name
No_HH	No of Household
TOT_M	Total population Male
TOT_F	Total population Female
M_06	Population in the age group 0-6 Male
F_06	Population in the age group 0-6 Female
M_SC	Scheduled Castes population Male
F_SC	Scheduled Castes population Female
M_ST	Scheduled Tribes population Male
F_ST	Scheduled Tribes population Female
M_LIT	Literates population Male
F_LIT	Literates population Female
M_ILL	Illiterate Male
F_ILL	Illiterate Female
TOT_WORK_M	Total Worker Population Male
TOT_WORK_F	Total Worker Population Female
MAINWORK_M	Main Working Population Male
MAINWORK_F	Main Working Population Female
MAIN_CL_M	Main Cultivator Population Male
MAIN_CL_F	Main Cultivator Population Female

MAIN_AL_M	Main Agricultural Labourers Population Male
MAIN_AL_F	Main Agricultural Labourers Population Female
MAIN_HH_M	Main Household Industries Population Male
MAIN_HH_F	Main Household Industries Population Female
MAIN_OT_M	Main Other Workers Population Male
MAIN_OT_F	Main Other Workers Population Female
MARGWORK_M	Marginal Worker Population Male
MARGWORK_F	Marginal Worker Population Female
MARG_CL_M	Marginal Cultivator Population Male
MARG_CL_F	Marginal Cultivator Population Female
MARG_AL_M	Marginal Agriculture Labourers Population Male
MARG_AL_F	Marginal Agriculture Labourers Population Female
MARG_HH_M	Marginal Household Industries Population Male
MARG_HH_F	Marginal Household Industries Population Female
MARG_OT_M	Marginal Other Workers Population Male
MARG_OT_F	Marginal Other Workers Population Female
MARGWORK_3_6_M	Marginal Worker Population 3-6 Male
MARGWORK_3_6_F	Marginal Worker Population 3-6 Female
MARG_CL_3_6_M	Marginal Cultivator Population 3-6 Male
MARG_CL_3_6_F	Marginal Cultivator Population 3-6 Female
MARG_AL_3_6_M	Marginal Agriculture Labourers Population 3-6 Male
MARG_AL_3_6_F	Marginal Agriculture Labourers Population 3-6 Female
MARG_HH_3_6_M	Marginal Household Industries Population 3-6 Male
MARG_HH_3_6_F	Marginal Household Industries Population 3-6 Female
MARG_OT_3_6_M	Marginal Other Workers Population Person 3-6 Male
MARG_OT_3_6_F	Marginal Other Workers Population Person 3-6 Female
MARGWORK_0_3_M	Marginal Worker Population 0-3 Male
MARGWORK_0_3_F	Marginal Worker Population 0-3 Female
MARG_CL_0_3_M	Marginal Cultivator Population 0-3 Male
MARG_CL_0_3_F	Marginal Cultivator Population 0-3 Female
MARG_AL_0_3_M	Marginal Agriculture Labourers Population 0-3 Male
MARG_AL_0_3_F	Marginal Agriculture Labourers Population 0-3 Female
MARG_HH_0_3_M	Marginal Household Industries Population 0-3 Male
MARG_HH_0_3_F	Marginal Household Industries Population 0-3 Female
MARG_OT_0_3_M	Marginal Other Workers Population 0-3 Male
MARG_OT_0_3_F	Marginal Other Workers Population 0-3 Female
NON_WORK_M	Non Working Population Male
NON_WORK_F	Non Working Population Female

Table 2

2.1.2 Check shape, Data types, statistical summary



```
<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)
```

Figure 28

Data Shape and Data Type

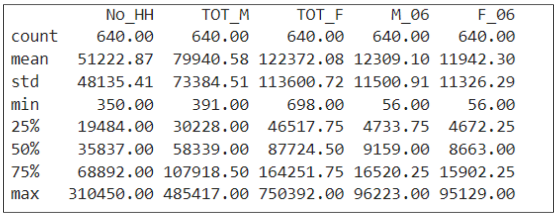
Observation 👀

- This dataset has 640 entries and 61 variables
- 59 variables are numeric and other 2 are non-numeric or categorial. Also, out of these 59 one variables 'Dist.Code' is shown numeric but it's a categorical Variable and no mathematical operations can be performed on it and hence its converted into 'Object' Type so that carrying analysis gets easy in future with this column
- This dataset has no missing or null values and the format is right

2.1.3 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?

Solution: 5 Variables chosen No_HH, TOT_M, TOT_F, M_06, F_06

Summary Statistics



```
          No_HH      TOT_M      TOT_F      M_06      F_06
count    640.00     640.00     640.00    640.00    640.00
mean   51222.87   79940.58  122372.08  12309.10  11942.30
std    48135.41   73384.51  113600.72  11500.91  11326.29
min      350.00     391.00     698.00     56.00     56.00
25%    19484.00   30228.00   46517.75   4733.75   4672.25
50%    35837.00   58339.00   87724.50   9159.00   8663.00
75%    68892.00  107918.50  164251.75  16520.25  15902.25
max   310450.00  485417.00  750392.00  96223.00  95129.00
```

Figure 29

Insights:

Number of Households (No_HH):

- Count: There are data for 640 different areas or communities.
- Mean: On average, each area has around 51,223 households.
- Standard Deviation: The typical difference from the mean is about 48,135 households. This suggests a wide variation in the number of households across different areas.
- Minimum: The smallest number of households observed in any area is 350.
- 25%: 25% of the areas have 19,484 households or fewer.
- Median (50%): Half of the areas have 35,837 households or fewer.
- 75%: 75% of the areas have 68,892 households or fewer.
- Maximum: The largest number of households observed in any area is 310,450.

Total Number of Males (TOT_M):

- Count: We have data for 640 different places.
- Mean: On average, each place has around 79,941 males.
- Standard Deviation: Males' numbers tend to vary a lot between places. On average, the difference from the average number is about 73,385 males.
- Minimum: The smallest number of males observed in any place is 391.
- 25%: In 25% of places, there are 30,228 males or fewer.
- Median (50%): Half of the places have 58,339 males or fewer.
- 75%: In 75% of places, there are 107,919 males or fewer.
- Maximum: The largest number of males observed in any place is 485,417.

Total Number of Females (TOT_F):

- Count: We have data for 640 different places.
- Mean: On average, each place has around 122,372 females.
- Standard Deviation: The number of females tends to vary quite a bit from one place to another. On average, the difference from the average number is about 113,601 females.
- Minimum: The smallest number of females observed in any place is 698.
- 25%: In 25% of places, there are 46,518 females or fewer.
- Median (50%): Half of the places have 87,725 females or fewer.
- 75%: In 75% of places, there are 164,252 females or fewer.
- Maximum: The largest number of females observed in any place is 750,392.

Number of Males Aged 6 and Older (M_06):

- Count: We have data for 640 different places.
- Mean: On average, each place has around 12,309 males aged 6 and older.
- Standard Deviation: The number of males aged 6 and older tends to vary somewhat from one place to another. On average, the difference from the average number is about 11,501 males.
- Minimum: The smallest number of males aged 6 and older observed in any place is 56.
- 25%: In 25% of places, there are 4,734 males aged 6 and older or fewer.
- Median (50%): Half of the places have 9,159 males aged 6 and older or fewer.
- 75%: In 75% of places, there are 16,520 males aged 6 and older or fewer.
- Maximum: The largest number of males aged 6 and older observed in any place is 96,223.

Number of Females Aged 6 and Older (F_06):

- Count: We have data for 640 different places.
- Mean: On average, each place has around 11,942 females aged 6 and older.
- Standard Deviation: The number of females aged 6 and older tends to vary somewhat from one place to another. On average, the difference from the average number is about 11,326 females.
- Minimum: The smallest number of females aged 6 and older observed in any place is 56.
- 25%: In 25% of places, there are 4,672 females aged 6 and older or fewer.
- Median (50%): Half of the places have 8,663 females aged 6 and older or fewer.
- 75%: In 75% of places, there are 15,902 females aged 6 and older or fewer.
- Maximum: The largest number of females aged 6 and older observed in any place is 95,129.

These numbers provide detailed insights into the distribution and characteristics of females and males, including their age demographics, across different areas.

Gender Disparities in Population distribution

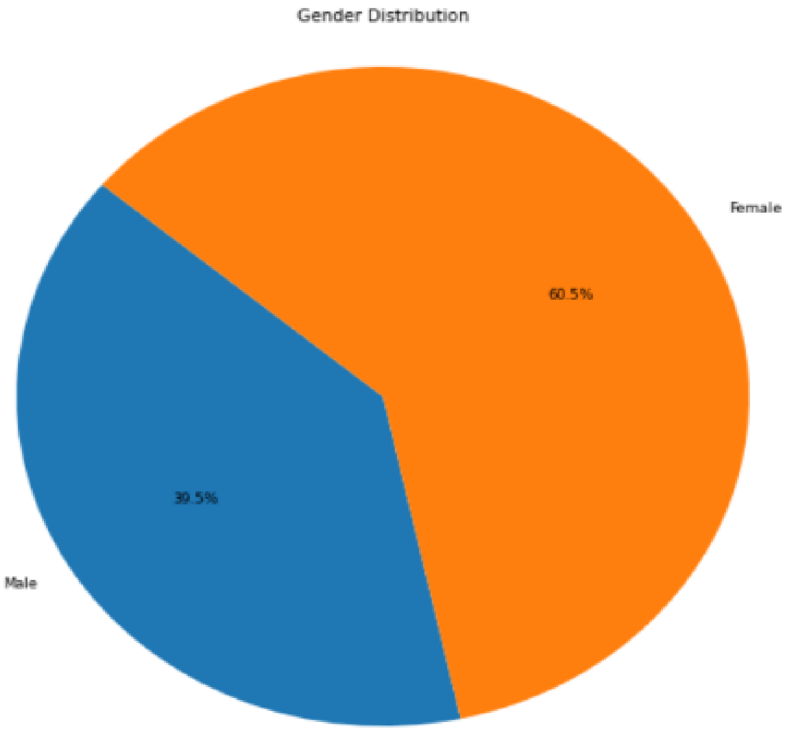


Figure 30

Age Group Analysis

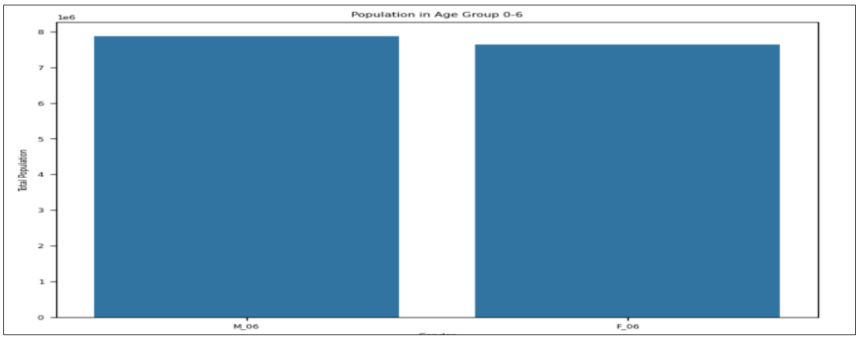


Figure 31

Bi-variate Analysis (Correlation Analysis)

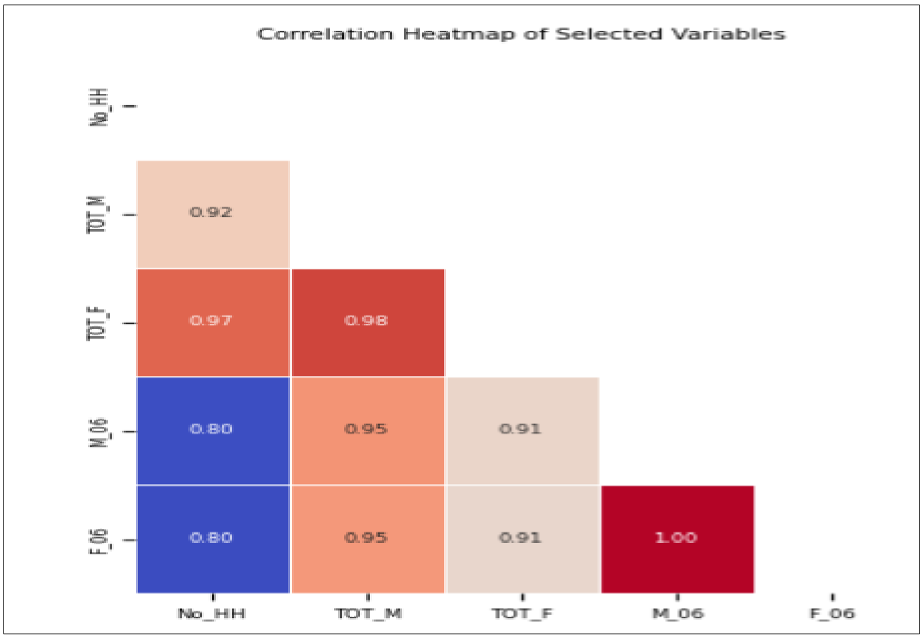


Figure 32

Insights:

- 60.5% of the Population is female and 39.5% is Male , it’s a female dominant data.
- There is not much difference in the population distribution between Male and female in the age group of 6
- if there's an increase in the total number of males and females in a community, it might lead to more households being formed as more people might decide to live independently or start families. Similarly, if there's a decrease in the number of males or females aged 6 and over, it could affect the number of households, perhaps because families with children move out or elderly individuals pass away. So, these factors tend to move together, suggesting a strong connection or correlation between them

Aggregation and Grouping

(i)Andhra Pradesh has the highest Gender Ratio and Lakshadweep has the lowest.

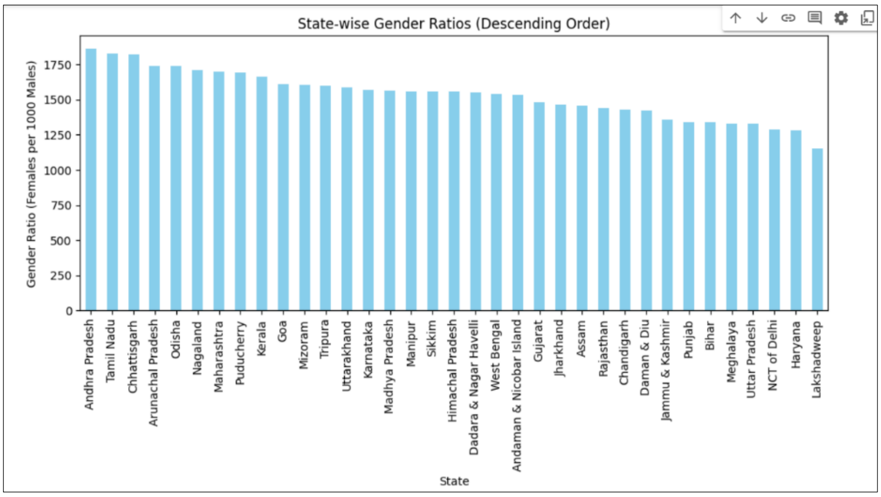
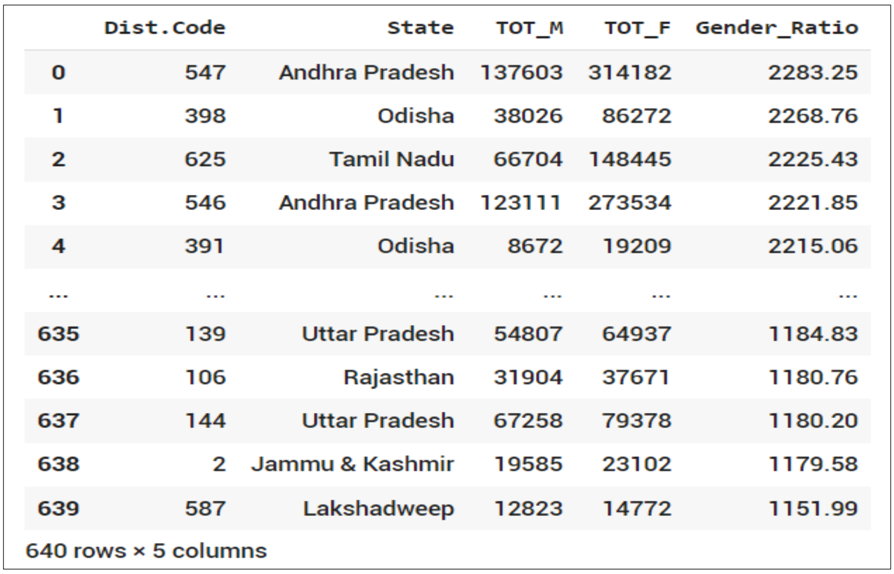


Figure 33

(ii) District 547 has the highest Gender ratio and 587 has the lowest



	Dist.Code	State	TOT_M	TOT_F	Gender_Ratio
0	547	Andhra Pradesh	137603	314182	2283.25
1	398	Odisha	38026	86272	2268.76
2	625	Tamil Nadu	66704	148445	2225.43
3	546	Andhra Pradesh	123111	273534	2221.85
4	391	Odisha	8672	19209	2215.06
...	...	...	...	...	...
635	139	Uttar Pradesh	54807	64937	1184.83
636	106	Rajasthan	31904	37671	1180.76
637	144	Uttar Pradesh	67258	79378	1180.20
638	2	Jammu & Kashmir	19585	23102	1179.58
639	587	Lakshadweep	12823	14772	1151.99

640 rows × 5 columns

Figure 34

2.2 Problem 2 - Data Preprocessing

2.2.1 Check for and treat (if needed) missing values

There are no missing values in this dataset

2.2.2 Check for and treat (if needed) data irregularities

- one variable 'Dist.Code' is shown numeric but it’s a categorical Variable and no mathematical operations can be performed on it and hence I have converted it into 'Object' Type so that carrying analysis gets easy in future with this column.This is for overall Dataset.
- No Data regularities as such in this pca dataset as it has only numeric variables and all of them are in right format

2.2.3 Scale the Data using the z-score method - Visualize the data before and after scaling and comment on the impact on outliers.

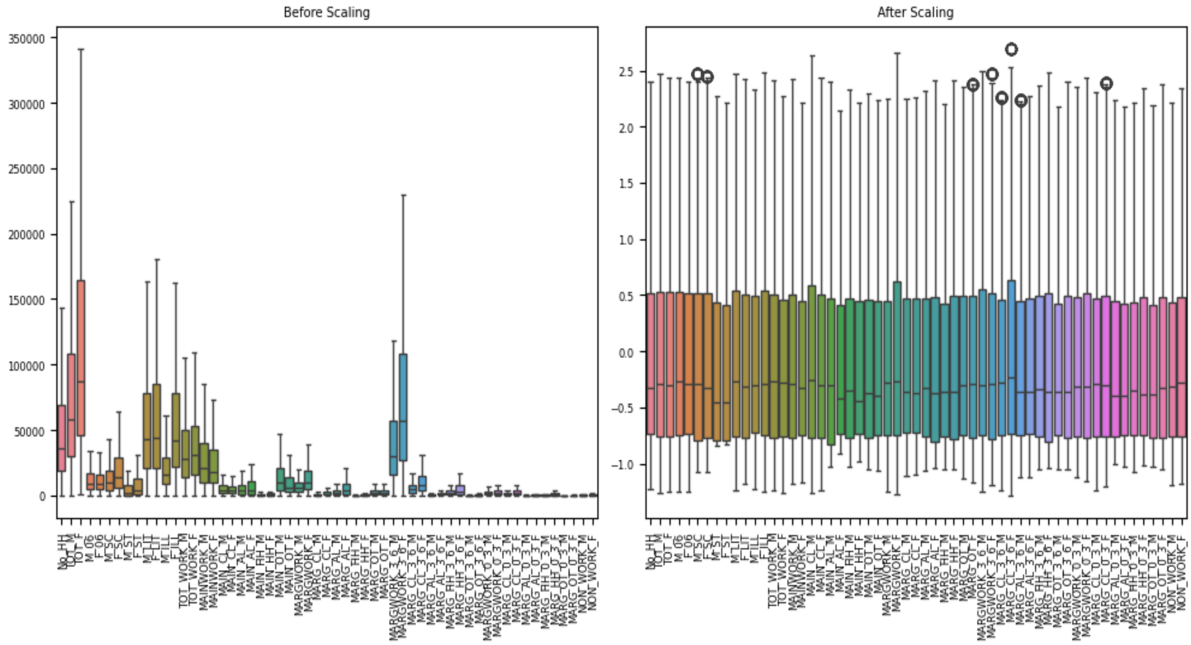


Figure 35

Impact of Scaling on Outliers:

Before scaling, outliers are present in the data.After scaling, outliers are still present, but their positions are altered due to scaling.To interpret the outliers and their potential effects, let's focus on variables where outliers were identified:

MAIN_CL_M and MAIN_CL_F: Outliers in these variables suggest extreme values in the proportion of individuals engaged in main work in male-headed and female-headed households. Possible effects: These outliers could indicate specific regions or demographics where the proportion of main workers significantly deviates from the norm, possibly due to unique economic conditions or cultural factors.

MAIN_AL_M and MAIN_AL_F: Outliers in these variables indicate extreme values in the proportion of individuals engaged in alternative work arrangements (excluding main work) in male-headed and female-headed households. Possible effects: These outliers might reflect areas with significant informal economies or unconventional employment patterns, potentially highlighting areas with irregular or seasonal employment opportunities.

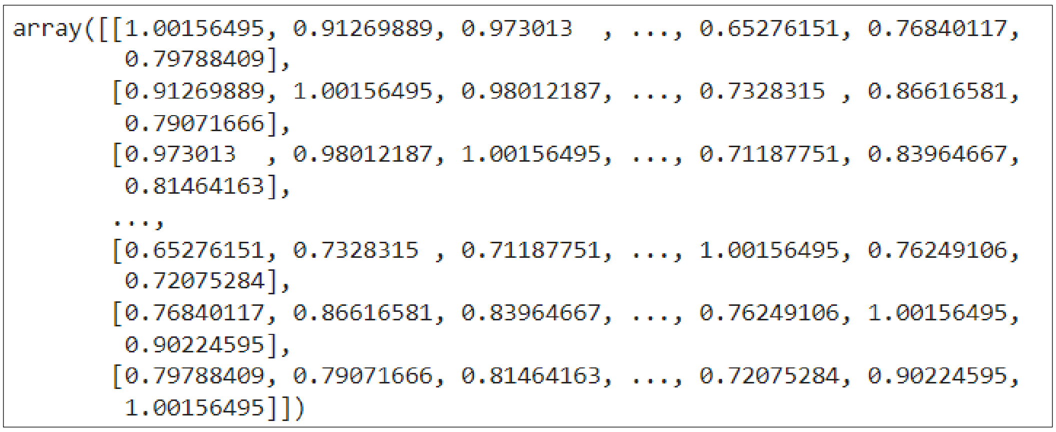
MAIN_HH_M and MAIN_HH_F: Outliers here suggest extreme values in the proportion of individuals engaged in household work (mainly managing the household) in male-headed and female-headed households. Possible effects: These outliers could represent regions where traditional gender roles strongly influence household responsibilities or areas with unique cultural or socioeconomic dynamics impacting household work distribution.

MAIN_OT_M and MAIN_OT_F: Outliers in these variables point to extreme values in the proportion of individuals engaged in other types of work (excluding main work) in male-headed and female-headed households. Possible effects: These outliers might indicate areas with high rates of secondary employment or diverse occupational structures, potentially reflecting regions with vibrant informal sectors or varied employment opportunities.

It's essential to note that while outliers can provide valuable insights into unique situations or phenomena, they can also skew statistical analyses. Further investigation, including qualitative research or domain-specific knowledge, would be necessary to understand the underlying reasons behind these outliers and their potential impacts accurately

2.3 Problem 2 - PCA

2.3.1 Create the covariance matrix



```
array([[1.00156495, 0.91269889, 0.973013  , ..., 0.65276151, 0.76840117,
        0.79788409],
       [0.91269889, 1.00156495, 0.98012187, ..., 0.7328315 , 0.86616581,
        0.79071666],
       [0.973013  , 0.98012187, 1.00156495, ..., 0.71187751, 0.83964667,
        0.81464163],
       ...,
       [0.65276151, 0.7328315 , 0.71187751, ..., 1.00156495, 0.76249106,
        0.72075284],
       [0.76840117, 0.86616581, 0.83964667, ..., 0.76249106, 1.00156495,
        0.90224595],
       [0.79788409, 0.79071666, 0.81464163, ..., 0.72075284, 0.90224595,
        1.00156495]])
```

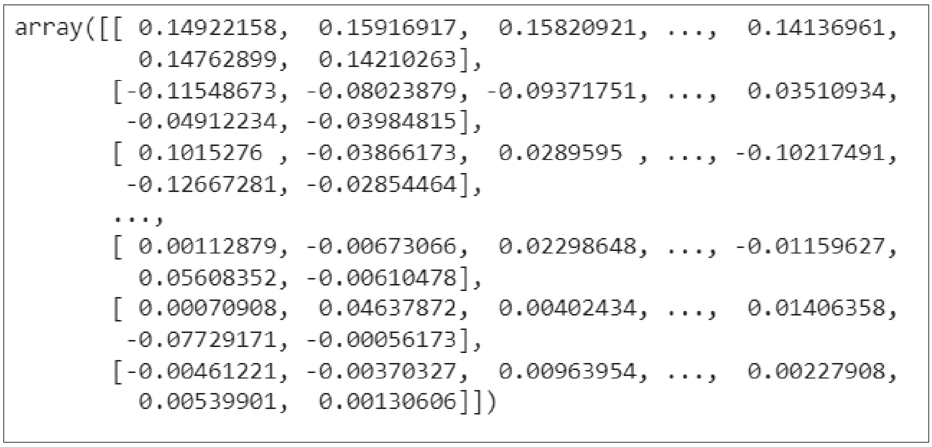
Figure 36

2.3.2 Get eigen values and eigen vectors

Solution

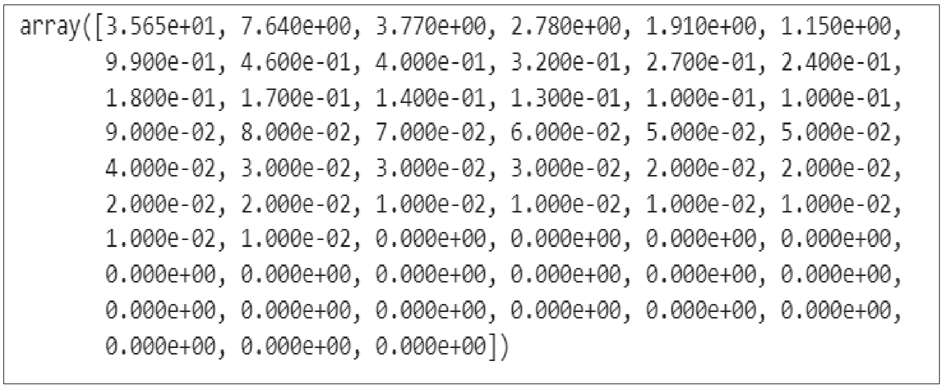
Eigen Vectors

Eigen Value



```
array([[ 0.14922158,  0.15916917,  0.15820921, ...,  0.14136961,
         0.14762899,  0.14210263],
       [-0.11548673, -0.08023879, -0.09371751, ...,  0.03510934,
        -0.04912234, -0.03984815],
       [ 0.1015276 , -0.03866173,  0.0289595 , ..., -0.10217491,
        -0.12667281, -0.02854464],
       ...,
       [ 0.00112879, -0.00673066,  0.02298648, ..., -0.01159627,
         0.05608352, -0.00610478],
       [ 0.00070908,  0.04637872,  0.00402434, ...,  0.01406358,
        -0.07729171, -0.00056173],
       [-0.00461221, -0.00370327,  0.00963954, ...,  0.00227908,
         0.00539901,  0.00130606]])
```

Figure 37

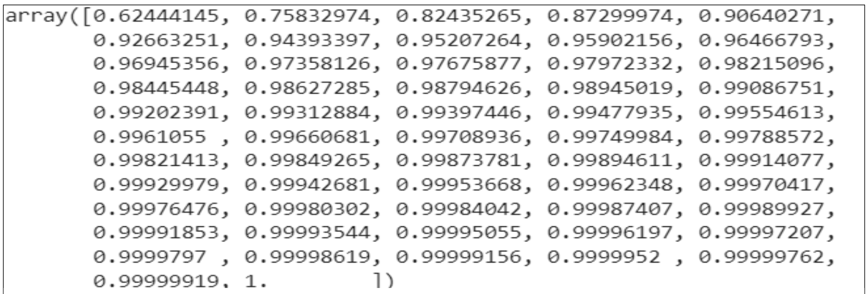


```
array([3.565e+01, 7.640e+00, 3.770e+00, 2.780e+00, 1.910e+00, 1.150e+00,
       9.900e-01, 4.600e-01, 4.000e-01, 3.200e-01, 2.700e-01, 2.400e-01,
       1.800e-01, 1.700e-01, 1.400e-01, 1.300e-01, 1.000e-01, 1.000e-01,
       9.000e-02, 8.000e-02, 7.000e-02, 6.000e-02, 5.000e-02, 5.000e-02,
       4.000e-02, 3.000e-02, 3.000e-02, 3.000e-02, 2.000e-02, 2.000e-02,
       2.000e-02, 2.000e-02, 1.000e-02, 1.000e-02, 1.000e-02, 1.000e-02,
       1.000e-02, 1.000e-02, 0.000e+00, 0.000e+00, 0.000e+00, 0.000e+00,
       0.000e+00, 0.000e+00, 0.000e+00, 0.000e+00, 0.000e+00, 0.000e+00,
       0.000e+00, 0.000e+00, 0.000e+00, 0.000e+00, 0.000e+00, 0.000e+00,
       0.000e+00, 0.000e+00, 0.000e+00])
```

Figure 38

2.3.3 Identify the optimum number of PCs - Show Scree plot

We can see below that more than 90% of the variance is explained by 5 Principal Components



```
array([0.62444145, 0.75832974, 0.82435265, 0.87299974, 0.90640271,
       0.92663251, 0.94393397, 0.95207264, 0.95902156, 0.96466793,
       0.96945356, 0.97358126, 0.97675877, 0.97972332, 0.98215096,
       0.98445448, 0.98627285, 0.98794626, 0.98945019, 0.99086751,
       0.99202391, 0.99312884, 0.99397446, 0.99477935, 0.99554613,
       0.9961055 , 0.99660681, 0.99708936, 0.99749984, 0.99788572,
       0.99821413, 0.99849265, 0.99873781, 0.99894611, 0.99914077,
       0.99929979, 0.99942681, 0.99953668, 0.99962348, 0.99970417,
       0.99976476, 0.99980302, 0.99984042, 0.99987407, 0.99989927,
       0.99991853, 0.99993544, 0.99995055, 0.99996197, 0.99997207,
       0.9999797 , 0.99998619, 0.99999156, 0.9999952 , 0.99999762,
       0.99999919, 1.        ])
```

Figure 39

Solution: Scree Plot is shown below

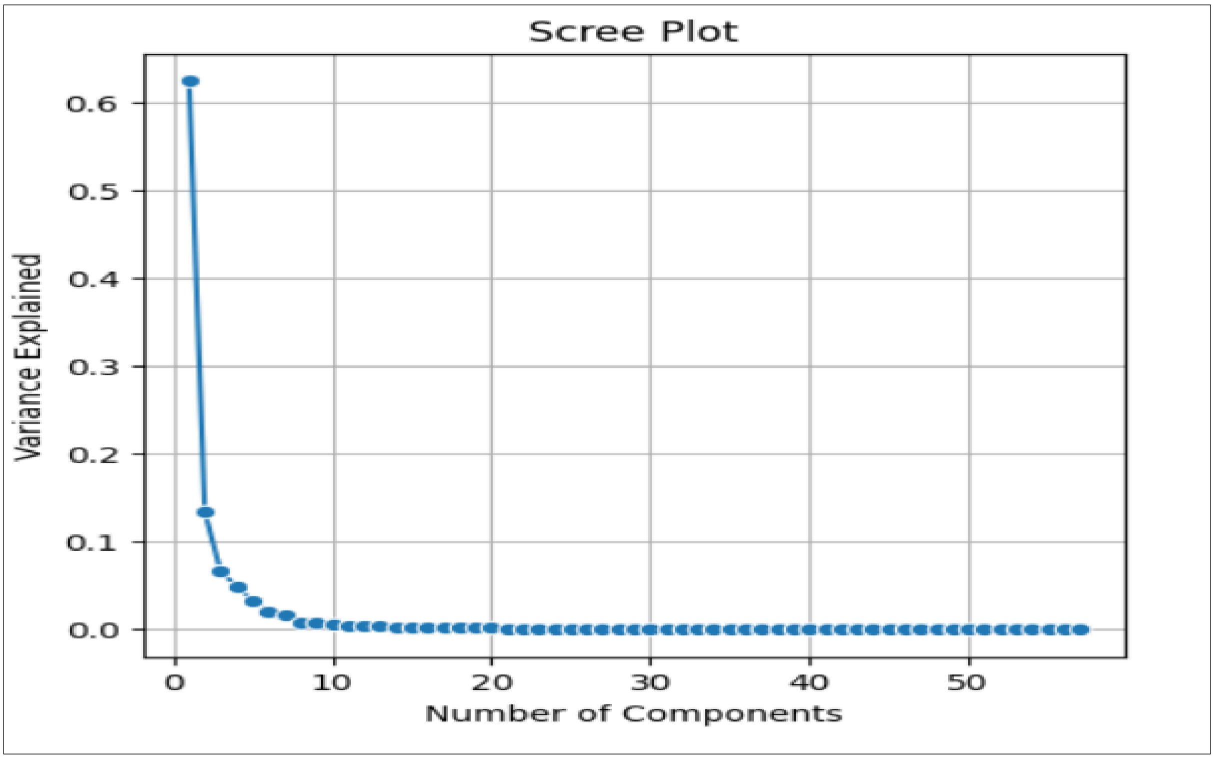


Figure 40

The optimum number of PCs through this graph comes out to be 5 as after that what we can see is the Variance gets constant.

2.3.4 Compare PCs with Actual Columns and identify which is explaining most variance

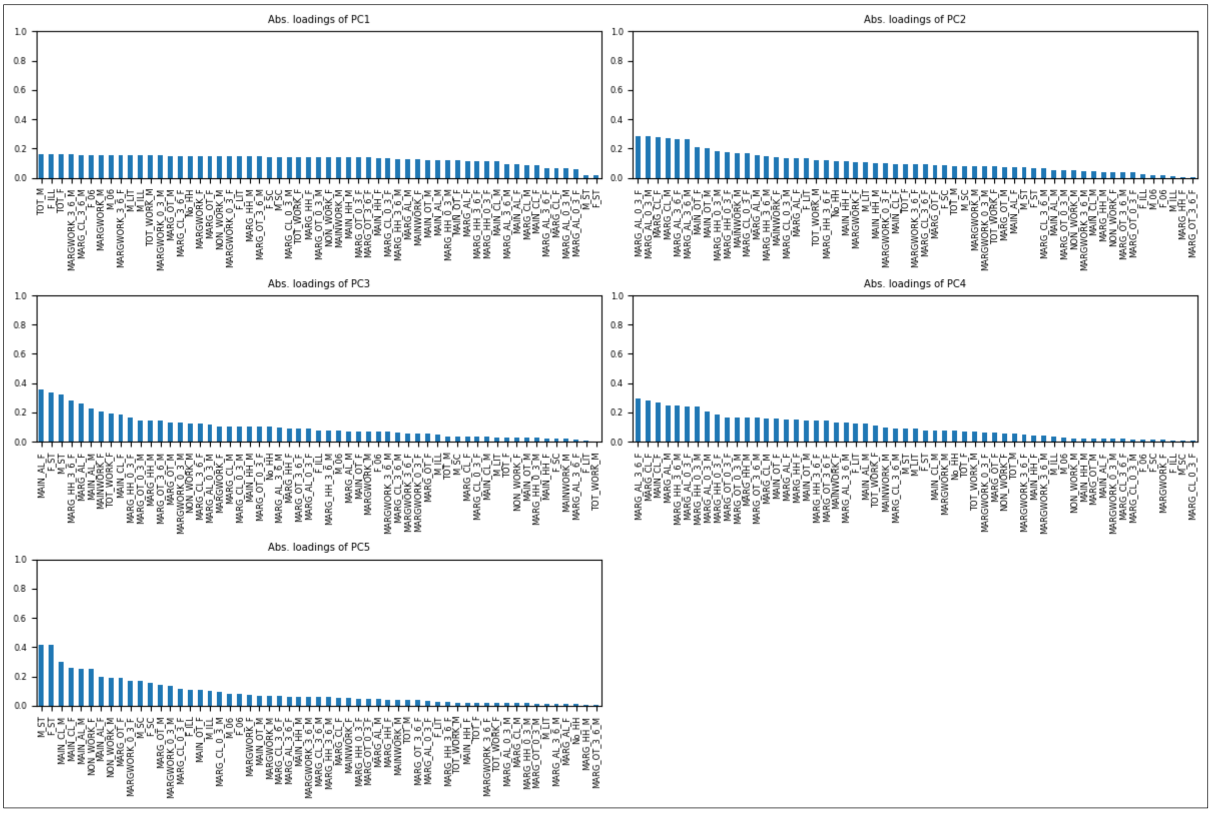


Figure 41

PC1 has the maximum explained variance, where TOT_M ,F_ILL,TOT_F has the most variance shown and M_ST and F_ST has the least variance shown.

PC2 has the maximum explained variance,where MARG_AL_0_3_M, MARG_AL_0_3_F,MARG_CL_M,MARG_CL_F has the most variance and MARG_HH_F,MARG_OT_3_6_F has the lowest variance.

PC3 has lesser variance as compared to PC1 and PC2 in which MAIN_AL_F ,F_ST,M_ST has the maximum variance and variables like F_LIT,TOT_WORK_M has the lowest variance.

PC4 and PC5 has the least variance, hence not that important to explain the variance of features in it.

2.3.5 Write inferences about all the PCs in terms of actual variables

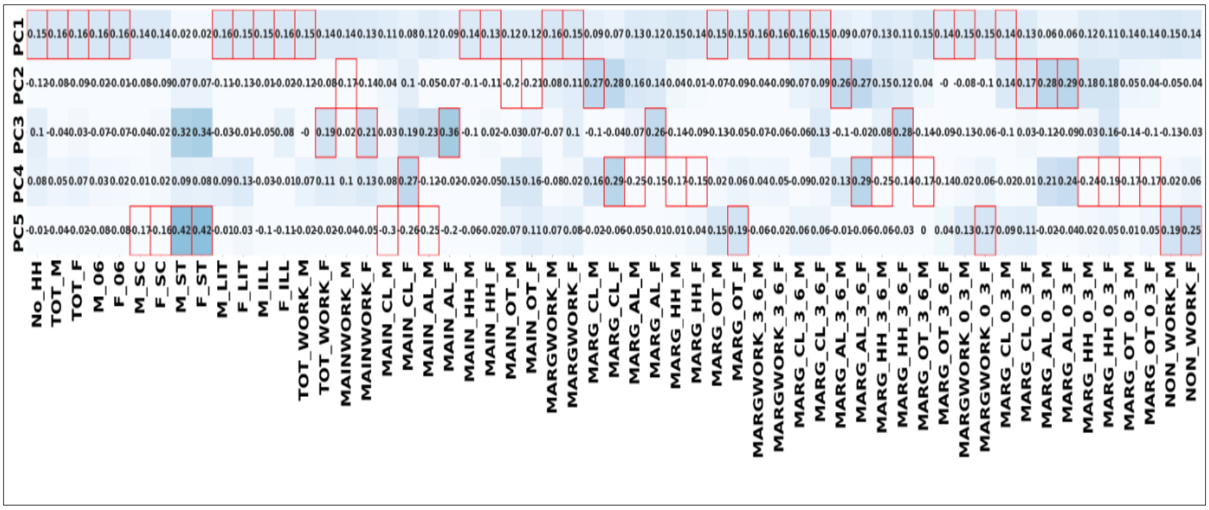


Figure 42

PC1: TOT_M , TOT_F,M_06,F_06 ,M_LITF_ILL,MARGWORK_M,MARGWORK_3_6M,MARGWORK_3_6_F has the highest positive loading, indicating that it contributes the most to PC1 , lets name it **Population Characteristics Profile** captures the composition and demographic features of the population, including gender distribution, age groups, literacy status, and engagement in marginal work.

PC2 : MAIN_OT_M,MAIN_OT_F,MARG_CL_M,MARG_AL_3_6_M, MARG_AL_0_3_M,MARG_AL_0_3_F has the highest positive loading, indicating that it contributes the most to PC2 , lets name it **Occupational Distribution and Marginal Employment Profile** captures the distribution of occupations, particularly in terms of main occupation status, and the prevalence of marginal employment within the population.

PC3: M_ST,F_ST,MAIN_AL_F,MARG_HH_3_6_F has the highest positive loading, indicating that it contributes the most to PC3, LETS NAME IT AS *Tribal Gender Dynamics and Household Composition Profile* captures the focus on gender dynamics within tribal communities along with insights into household characteristics

PC4: MAIN_CL_F,MARG_CL_F ,MARG_AL_F,MARG_AL_M,MARG_AL_3_6_F,MARG_HH_0_3M,MARG_HH_0_3F has the highest positive loading, indicating that it contributes the most to PC4 ,lets name it **Marginal Agricultural Employment and Household Structure Profile**. capture the focus on employment in agriculture, particularly among marginal workers, along with insights into household structure and composition within this context

PC5: M_SC,F_SC_,M_ST,F_ST,MAIN_CL_F,NON_WORK_F has the highest positive loading, indicating that it contributes the most to PC5, lets name it as **Scheduled Community Workforce Profile** capture the focus on employment status and dynamics within scheduled caste and scheduled tribe communities

2.3.6 Write linear equation for first PC Note: For the scope of this project, take at least 90% explained variance.

Solution

Equation of the first principal component:

Population Characteristics Profile(PC1) = 0.150*No_HH + 0.160*TOT_M + 0.160*TOT_F + 0.160*M_06 + 0.160*F_06 + 0.140*M_SC + 0.140*F_SC + 0.020*M_ST + 0.020*F_ST + 0.160*M_LIT + 0.150*F_LIT + 0.150*M_ILL + 0.160*F_ILL + 0.150*TOT_WORK_M + 0.140*TOT_WORK_F + 0.140*MAINWORK_M + 0.130*MAINWORK_F + 0.110*MAIN_CL_M + 0.080*MAIN_CL_F + 0.120*MAIN_AL_M + 0.090*MAIN_AL_F + 0.140*MAIN_HH_M + 0.130*MAIN_HH_F + 0.120*MAIN_OT_M + 0.120*MAIN_OT_F + 0.160*MARGWORK_M + 0.150*MARGWORK_F + 0.090*MARG_CL_M + 0.070*MARG_CL_F + 0.130*MARG_AL_M + 0.120*MARG_AL_F + 0.150*MARG_HH_M + 0.140*MARG_HH_F + 0.150*MARG_OT_M + 0.150*MARG_OT_F + 0.160*MARGWORK_3_6_M + 0.160*MARGWORK_3_6_F + 0.160*MARG_CL_3_6_M + 0.150*MARG_CL_3_6_F + 0.090*MARG_AL_3_6_M + 0.070*MARG_AL_3_6_F + 0.130*MARG_HH_3_6_M + 0.110*MARG_HH_3_6_F + 0.150*MARG_OT_3_6_M + 0.140*MARG_OT_3_6_F + 0.150*MARGWORK_0_3_M + 0.150*MARGWORK_0_3_F + 0.140*MARG_CL_0_3_M + 0.130*MARG_CL_0_3_F + 0.060*MARG_AL_0_3_M + 0.060*MARG_AL_0_3_F + 0.120*MARG_HH_0_3_M + 0.110*MARG_HH_0_3_F + 0.140*MARG_OT_0_3_M + 0.140*MARG_OT_0_3_F + 0.150*NON_WORK_M + 0.140*NON_WORK_F