Machine Learning 1

Coded Project

Anirudh Sardiwal

25 February 2024

Table of Contents

1.0 Problem 1 - Digital Marketing - Clustering **3**

1.1 Aim **3**

1.2 EDA **3**

1.2.1 Basics 3

1.2.2 Univariate Analysis 4

1.2.3 Bivariate Analysis 6

1.3 Data Preprocessing **8**

1.4 K-Means Clustering **10**

1.5 Actionable Insights and Recommendations **12**

1.5.1 Clusters 12

1.5.2 Insights 12

1.5.3 Recommendations 13

2.0 Problem 2 - Census - PCA **14**

2.1 Basics **14**

2.2 EDA **14**

2.3 PCA **19**

1.0 Problem 1 - Digital Marketing - Clustering

1.1 Aim

Segment type of ads based on features provided. Use of Clustering to segment ads into homogenous groups.

1.2 EDA

1.2.1 Basics

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

Fig. 1 - Partial data head

- The data contains 23066 rows and 19 columns.
- 'Inventory Type' column has values from Format1 to Format7
- 'Ad Type' - Inter217 to Inter229
- 'Ad Size' - 75000, 84000, 216000, 180000, 72000, 33600, 65520
- Platform - Video, App, Web
- Device Type - Desktop, Mobile
- Format - Display, Video
- There are no duplicates

1.2.2 Univariate Analysis

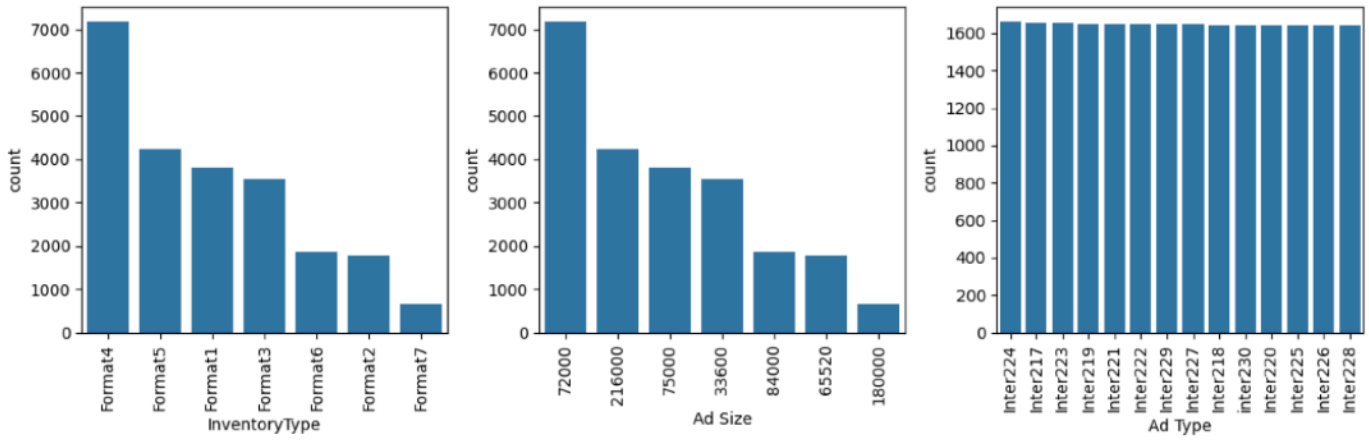

Fig. 2 - Countplots of Inventory Type, Ad Size, and Ad Type

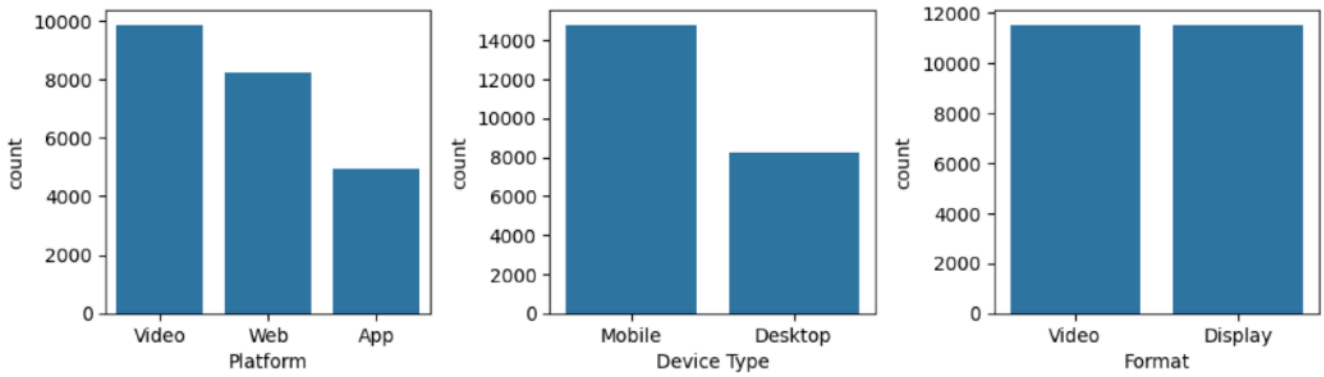

Fig. 2 - Countplots of Platform, Device Type, and Format

- 72000 is the most used Ad Size, followed by 216000
- Video platform is preferred on Mobile with equal number of Video and Display ads

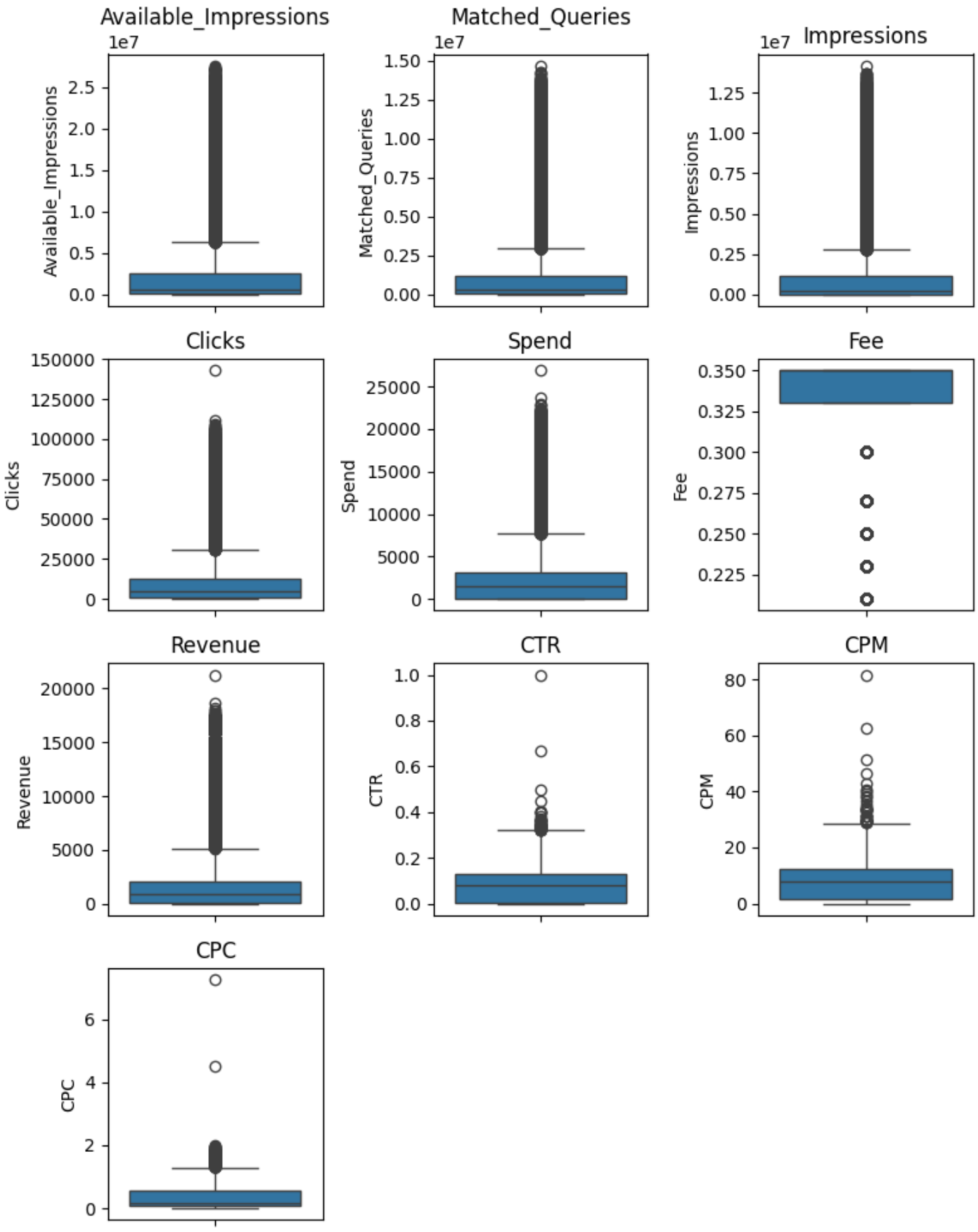

Fig. 4 - Boxplots of various features

1.2.3 Bivariate Analysis

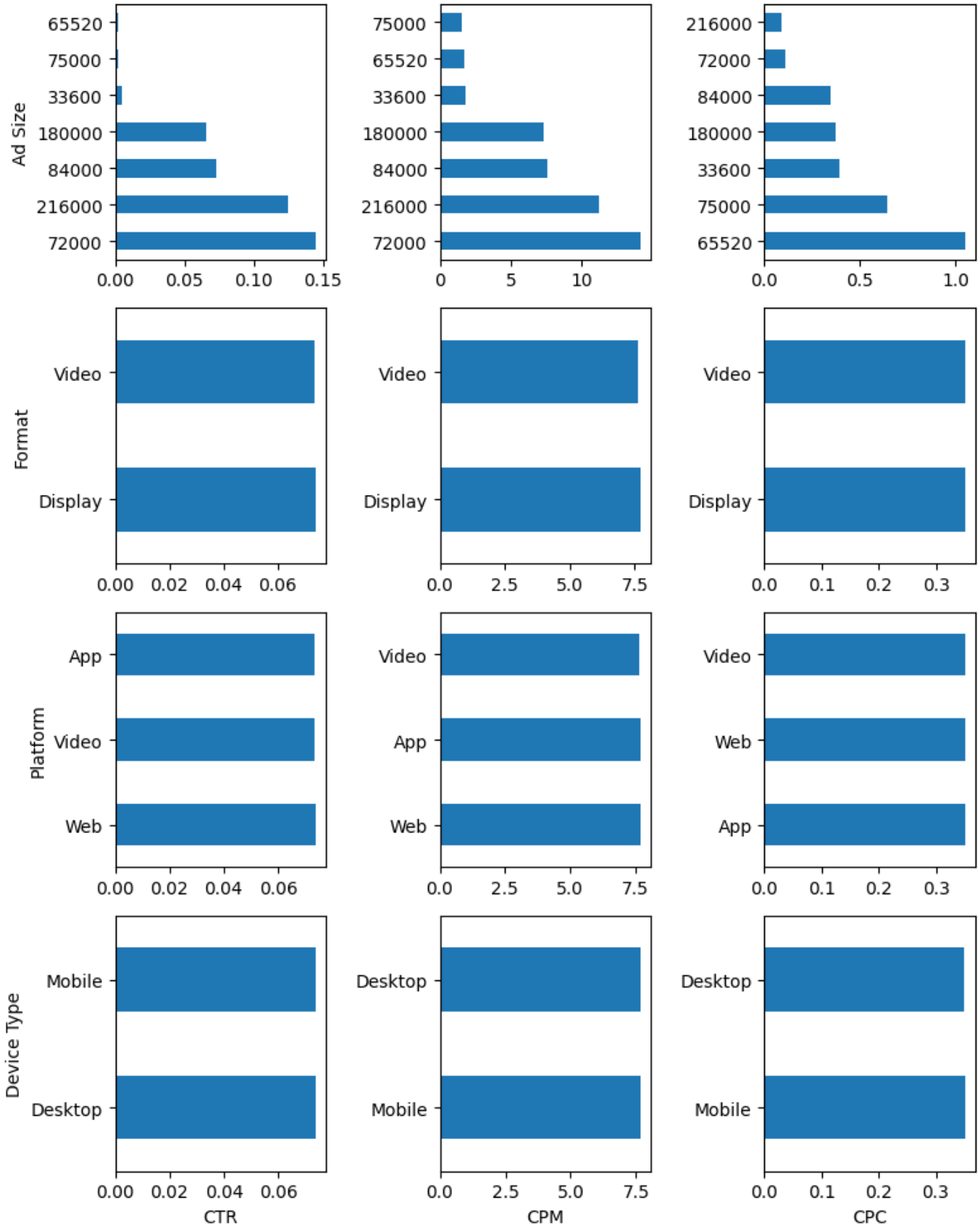

Fig. 5 - Device Type, Platform, Format, and Ad Size by Mean CTR, CPM, and CPC

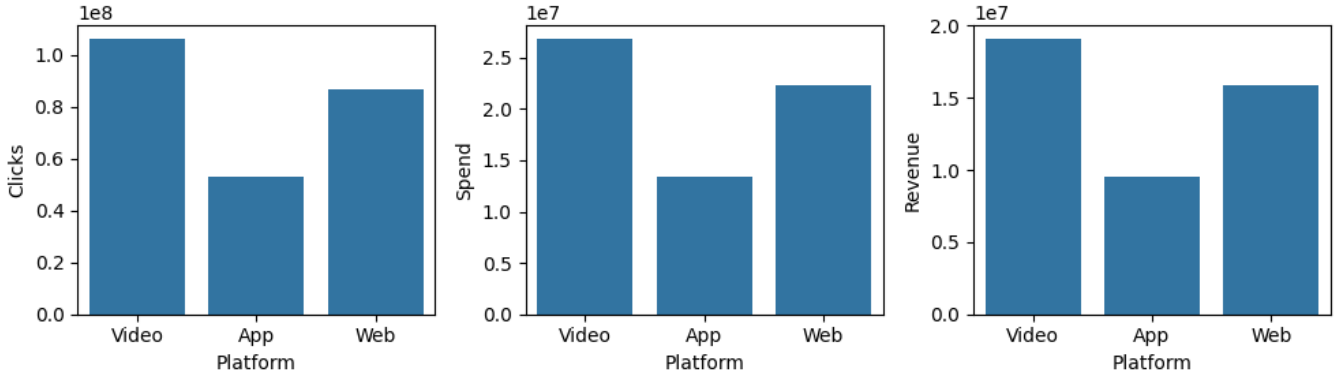

Fig.. 6 - Clicks, Spend, and Revenue by Platform

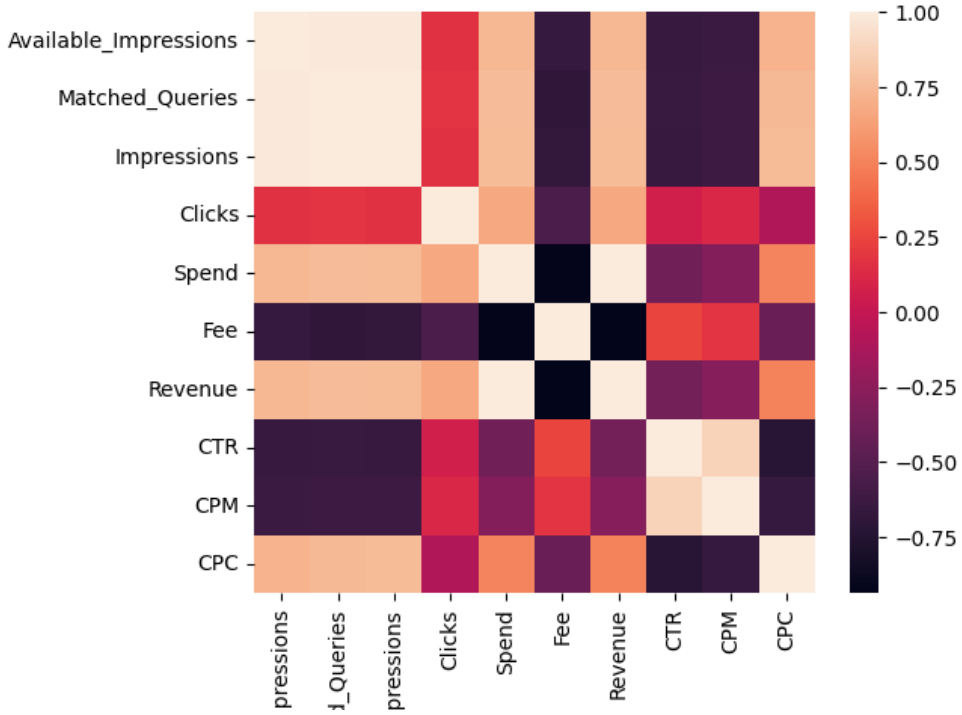

Fig. 7 - Heatmap of correlations of various features

- CTR and CPM have an negative correlation with CPC
- Revenue and Spend are positively correlated, that means the more is spent, the more revenue is generated for the company
- Ad sizes 72000 and 216000 get the maximum CTR
- 75000 and 65520 get the least CPM
- 216000 and 72000 have the least CPC
- Video platform is getting the maximum no. of Clicks, Spend, and Revenue
- *There seems to be something wrong with Device Type, Platform, and Format with CPC, CTR, and CPM values.*

1.3 Data Preprocessing

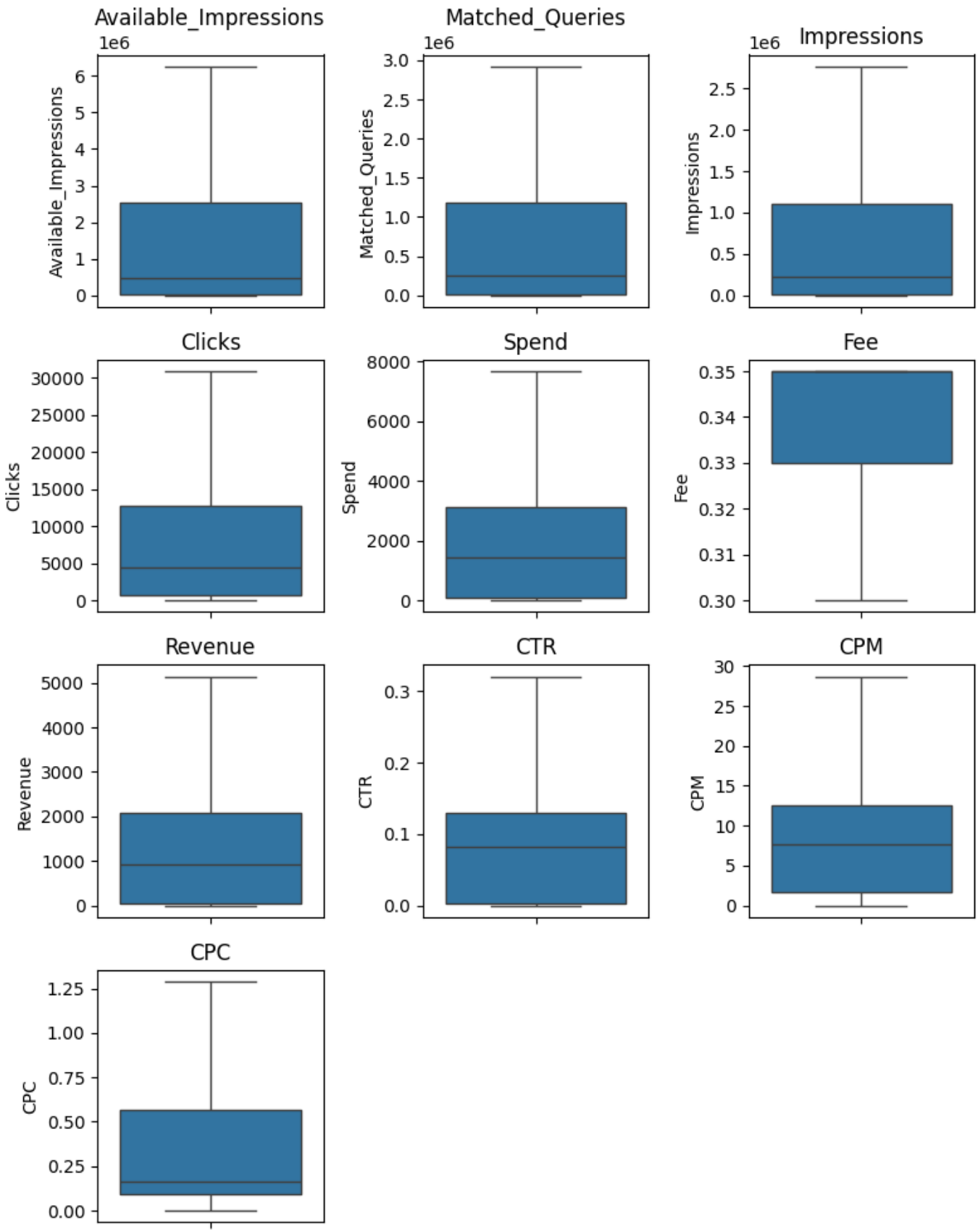

Fig. 8 - Numerical features after bringing outliers to Lower and Upper limits

	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	1806.0	325.0	323.0	1.0	0.0	0.35	0.0	0.0	0.0	0.0
1	1780.0	285.0	285.0	1.0	0.0	0.35	0.0	0.0	0.0	0.0
2	2727.0	356.0	355.0	1.0	0.0	0.35	0.0	0.0	0.0	0.0
3	2430.0	497.0	495.0	1.0	0.0	0.35	0.0	0.0	0.0	0.0
4	1218.0	242.0	242.0	1.0	0.0	0.35	0.0	0.0	0.0	0.0

Fig. 9 - Selected numerical features for Clustering

	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	-0.755333	-0.778949	-0.768478	-0.867488	-0.89317	0.535724	-0.880093	-0.984667	-1.194635	-1.041011
1	-0.755345	-0.778988	-0.768516	-0.867488	-0.89317	0.535724	-0.880093	-0.984667	-1.194635	-1.041011
2	-0.754900	-0.778919	-0.768445	-0.867488	-0.89317	0.535724	-0.880093	-0.984667	-1.194635	-1.041011
3	-0.755040	-0.778781	-0.768302	-0.867488	-0.89317	0.535724	-0.880093	-0.984667	-1.194635	-1.041011
4	-0.755610	-0.779030	-0.768560	-0.867488	-0.89317	0.535724	-0.880093	-0.984667	-1.194635	-1.041011

Fig. 10 - Selected features scaled by Z-score method

1.4 K-Means Clustering

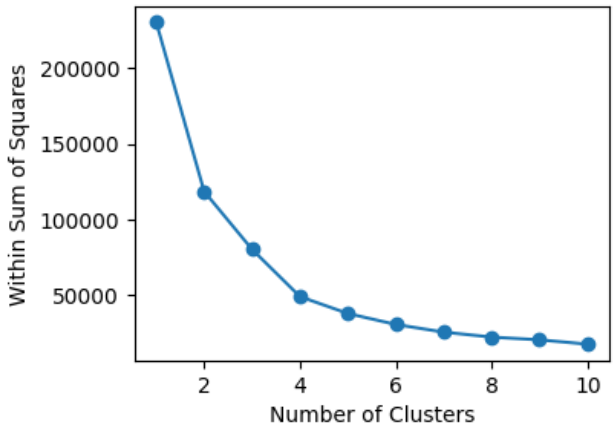

Fig. 10 - WSS Analysis

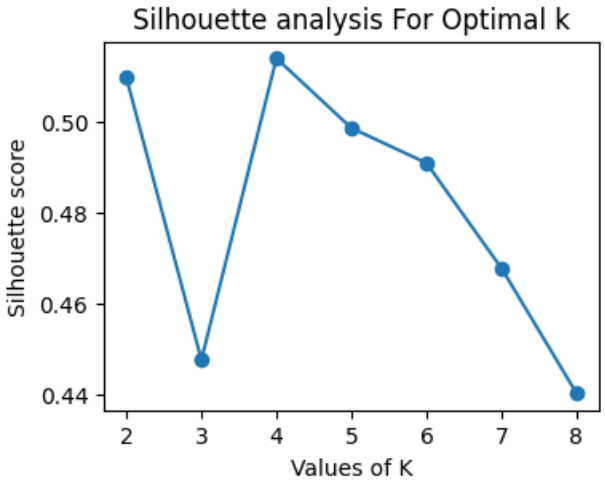

Fig. 11 - Silhoutte score analysis

Clus_kmeans4	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	freq
0	790535.85	553268.22	466992.19	30638.79	6367.58	0.31	4347.17	0.14	1.52	0.11	1630
1	5712214.21	2813243.41	2677457.92	11300.23	5759.93	0.31	3892.92	0.00	0.16	0.75	4020
2	1814910.99	867540.62	829613.22	3262.56	1505.11	0.35	980.96	0.00	0.18	0.53	6336
3	119308.83	66276.29	54314.68	6819.99	622.01	0.35	404.40	0.15	1.33	0.10	11080

Fig. 12 - Means of numerical columns aggregated by Cluster

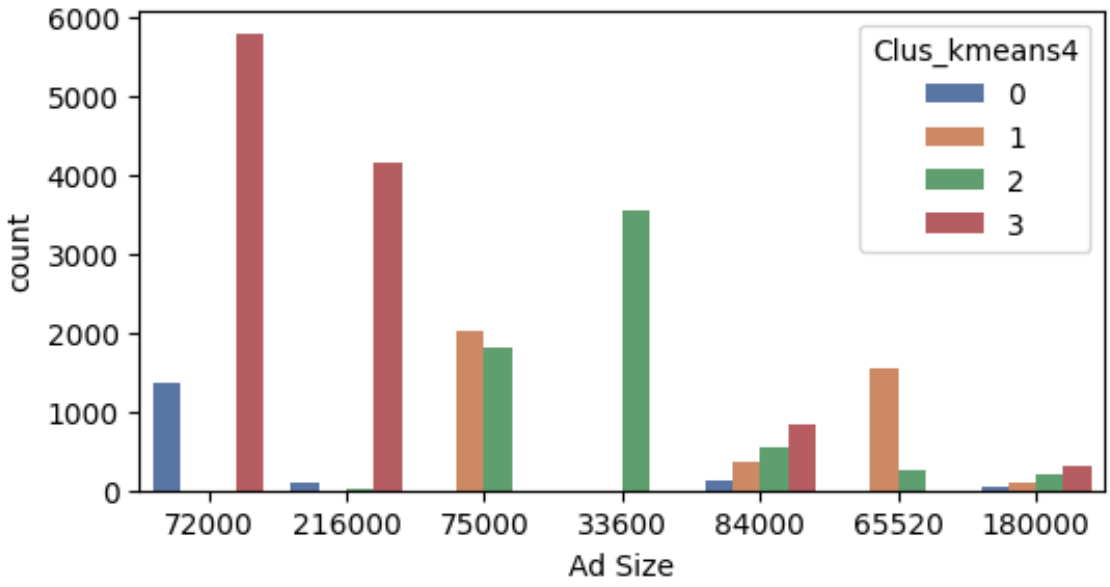

Fig. 13 - Count of Ad Sizes by Cluster

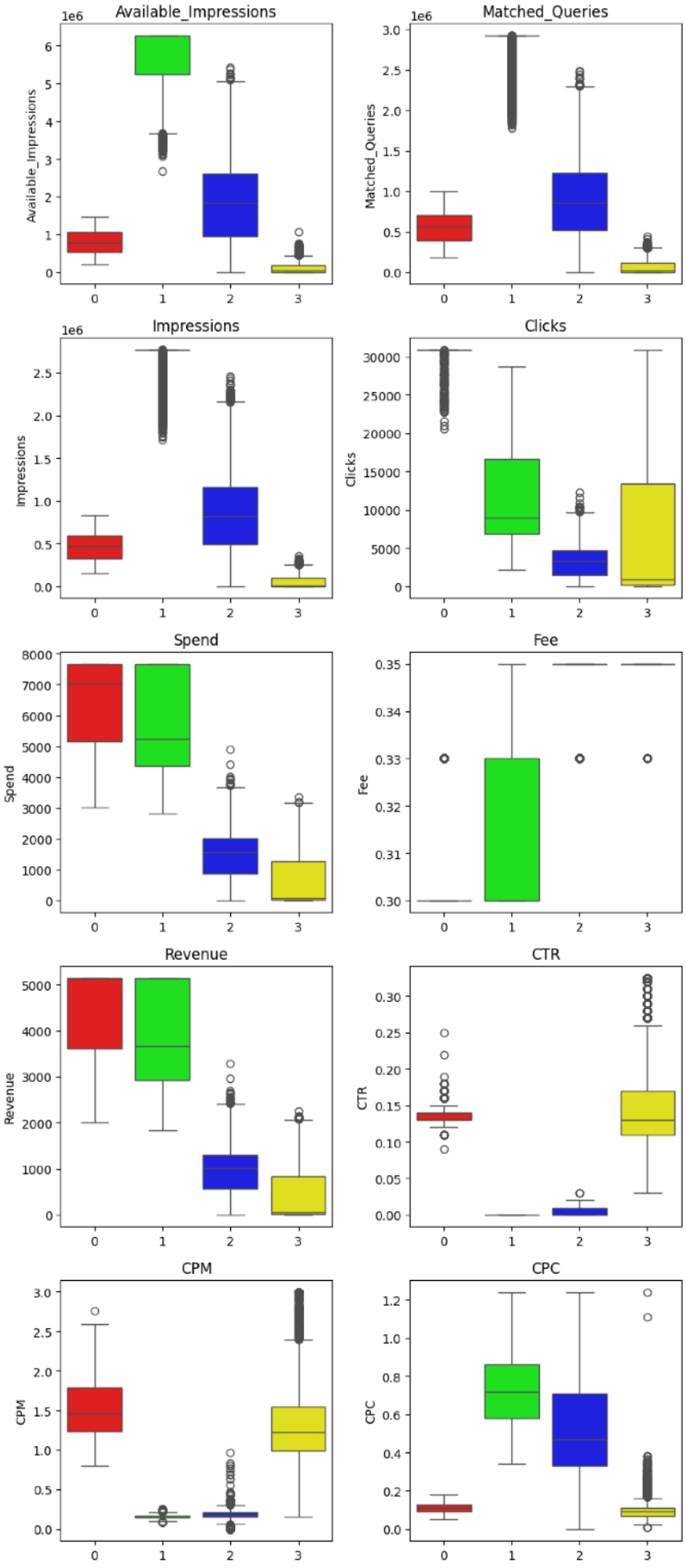

Fig. 14 - Boxplots of various features by Cluster

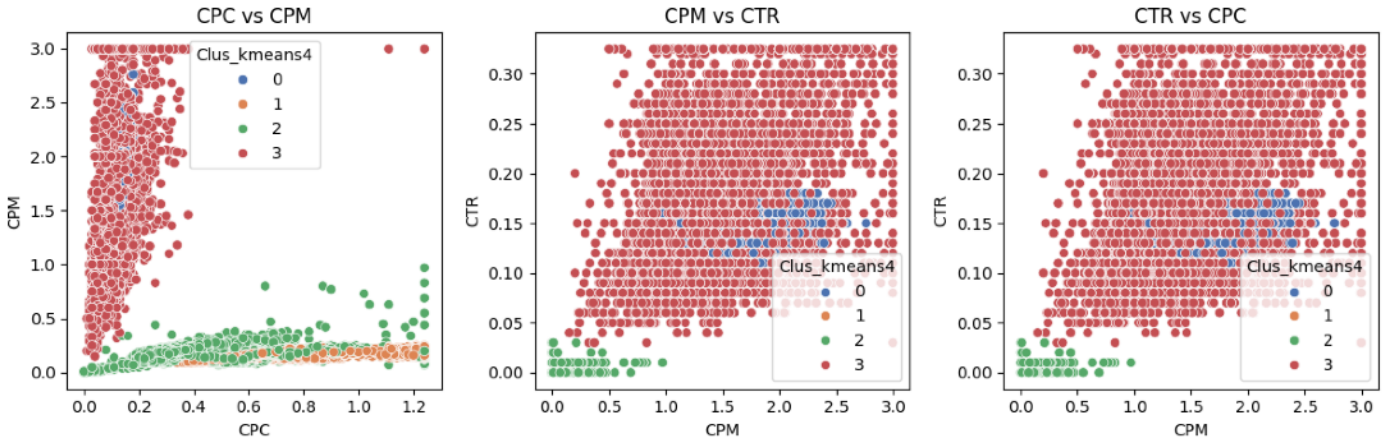

Fig. 15 - Scatterplots of CPC, CPM, and CTR by cluster

1.5 Actionable Insights and Recommendations

1.5.1 Clusters

Cluster 0 - highest Revenue, highest Spend, high CTR, highest CPM, low CPC, lowest frequency, most used Ad Size - 72000

Cluster 1 - second highest Spend, second highest Revenue, lowest CTR, lowest CPM, highest CPC, most used Ad Size - 75000, 65520

Cluster 2 - most used Ad Size - 33600, low Spend, low Revenue, lowest CTR, low CPM, high CPC, high frequency

Cluster 3 - most used Ad Size - 72000, lowest Spend, lowest Revenue, high CTR, high CPM, low CPC, highest frequency

1.5.2 Insights

- Cluster 0 is producing most Revenue but with highest CPM. This means that people are clicking a lot but the landing page experience isn't good. 72000 most used. Has lowest frequency.
- Cluster 1 has high Revenue and high Spend, but lowest CTR and highest CPC. This means ads are not relevant to the audience. The frequency is on the lower side. 75000, 65520 most used
- Cluster 2 is similar to Cluster 1. Ads are not relevant to audience. These ads are used with high frequency. 33600 most used.
- Cluster 3 has the highest frequency but has lowest Revenue but also lowest spend. Metrics are similar to Cluster 0. 72000 most used.

1.5.3 Recommendations

- The landing pages of Cluster 0 need to be improved, that will bring down the CPM and spends. If these ads can increase in number they can bring in good Revenue.
- Ads of Cluster 1 need to be made relevant for the audience. This will increase CTR thus bringing down the CPC and Spend. Sizes 75000 and 65520 can then be increased in number for increased Revenue.
- Similarly, ads for Cluster 2 also need to be made relevant for the audience thus increasing Revenue, and relevance to audience needs to increase thus increasing CTR. Frequency is already high. With these changes Revenue should increase substantially.
- For Cluster 3 only CPM needs to improve. That means the landing pages need to be made relevant to the audience. This should increase the Revenue and since the frequency is already high it should bring in good results.

2.0 Problem 2 - Census - PCA

2.1 Basics

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

Fig. 2.1 - Census data head

The data has 640 Rows and 61 Columns

There are no null or duplicated values.

2.2 EDA

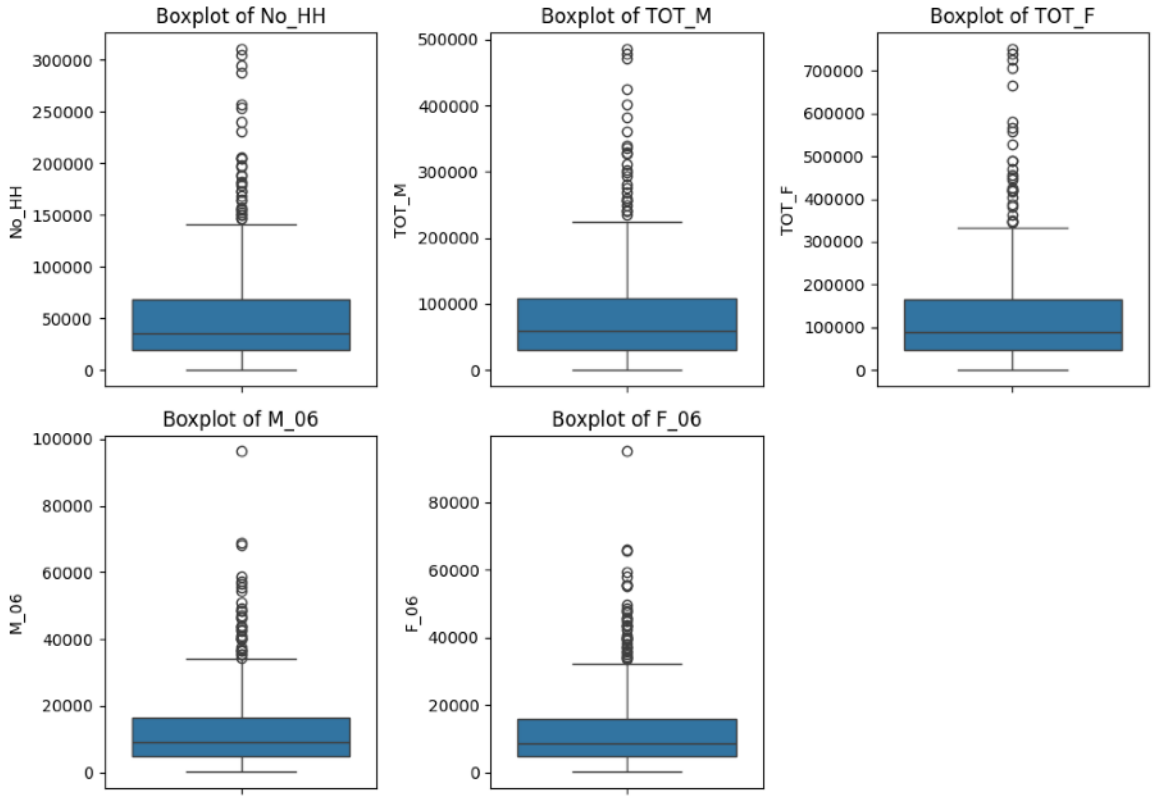

Fig. 2.2 - Boxplots of selected features

The 5 selected features are:

1. No_HH : Number of Households
2. TOT_M : Total Male Population
3. TOT_F : Total Female Population
4. M_06 : Male Population below the age of 6
5. F_06 : Female Population below the age of 6

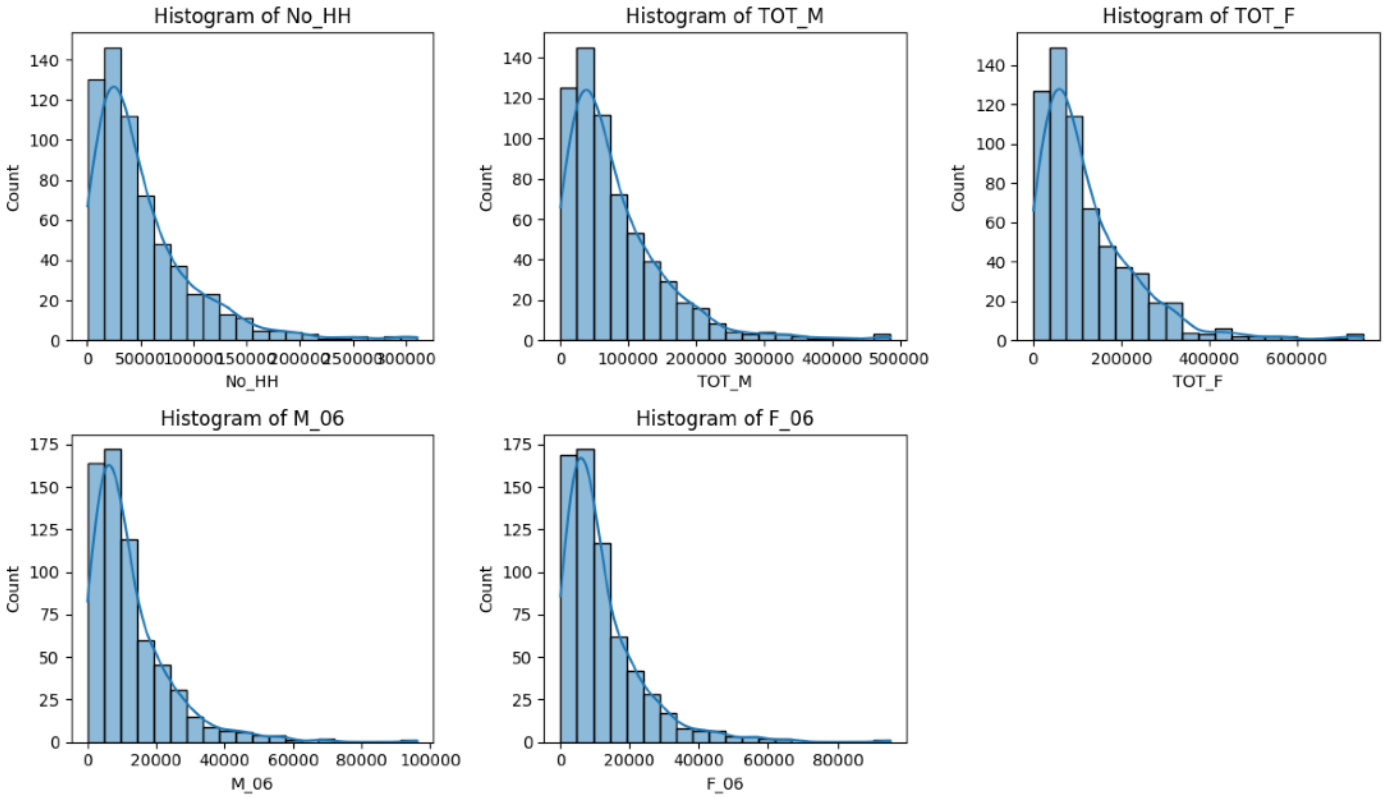

Fig. 2.3 - Histograms of selected features

	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587

Fig. 2.4 - Head of selected data

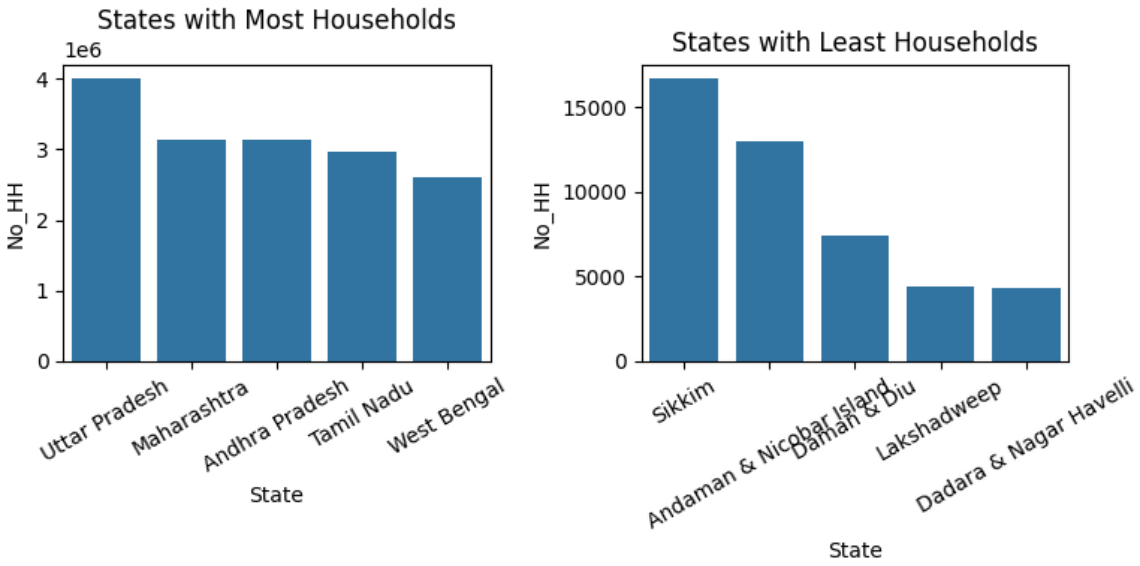

Fig. 2.4 - States with most and least households

There seems to be a problem with adult MF ratio. It cannot be as low as 0.8 and below!

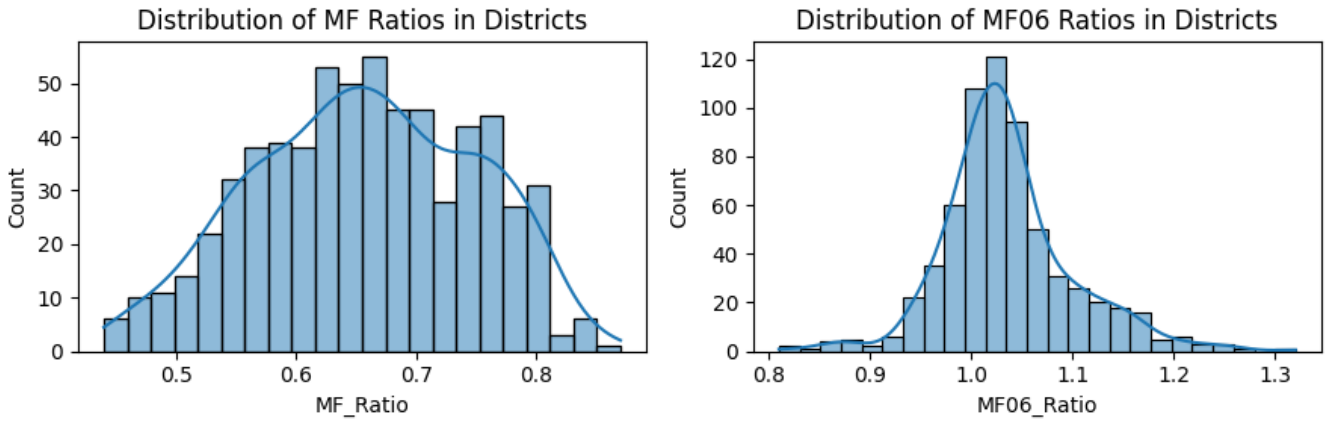

Fig. 2.5 - Distribution of Male to Female Ratios in Districts

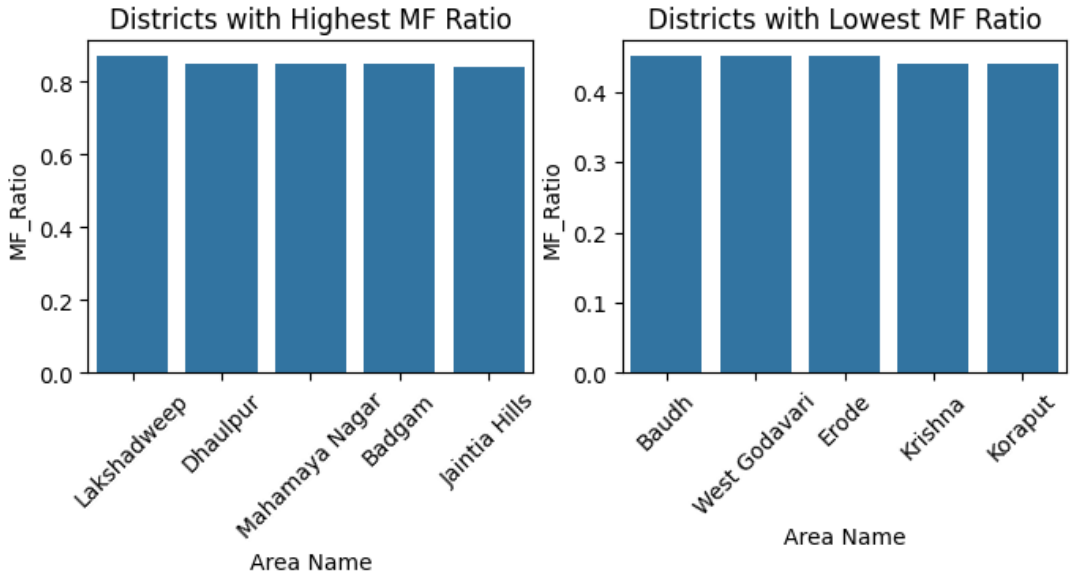

Fig. 2.6 - Districts with highest and lowest adult MF ratio

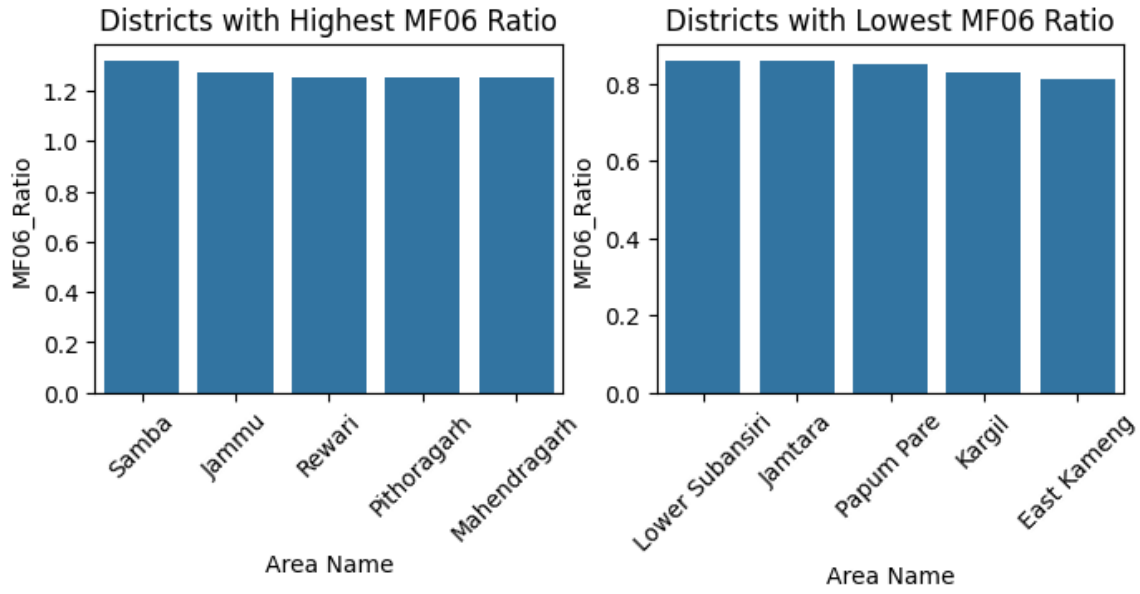

Fig. 2.7 - Districts with highest and lowest 0-6 MF ratio

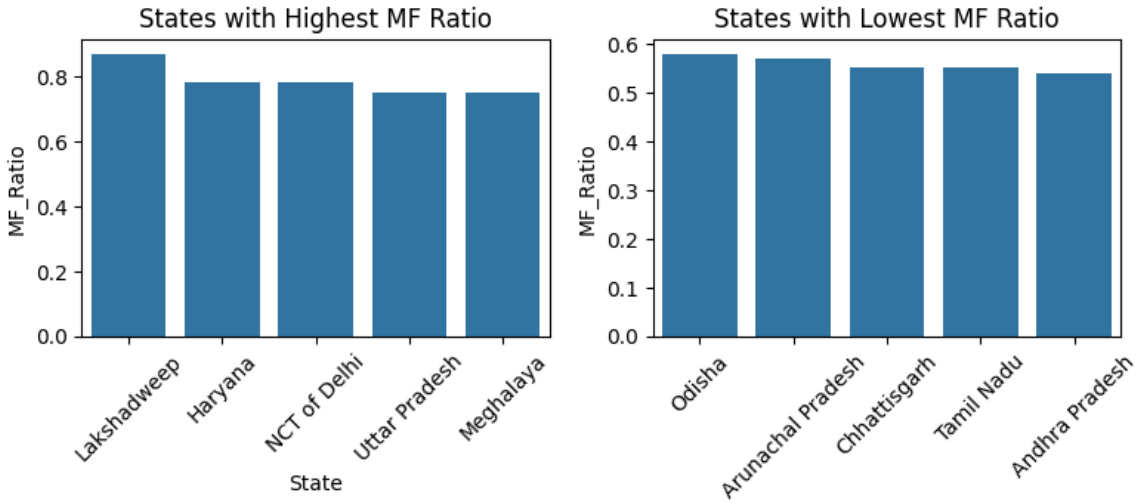

Fig. 2.8 - States with highest and lowest adult MF ratio

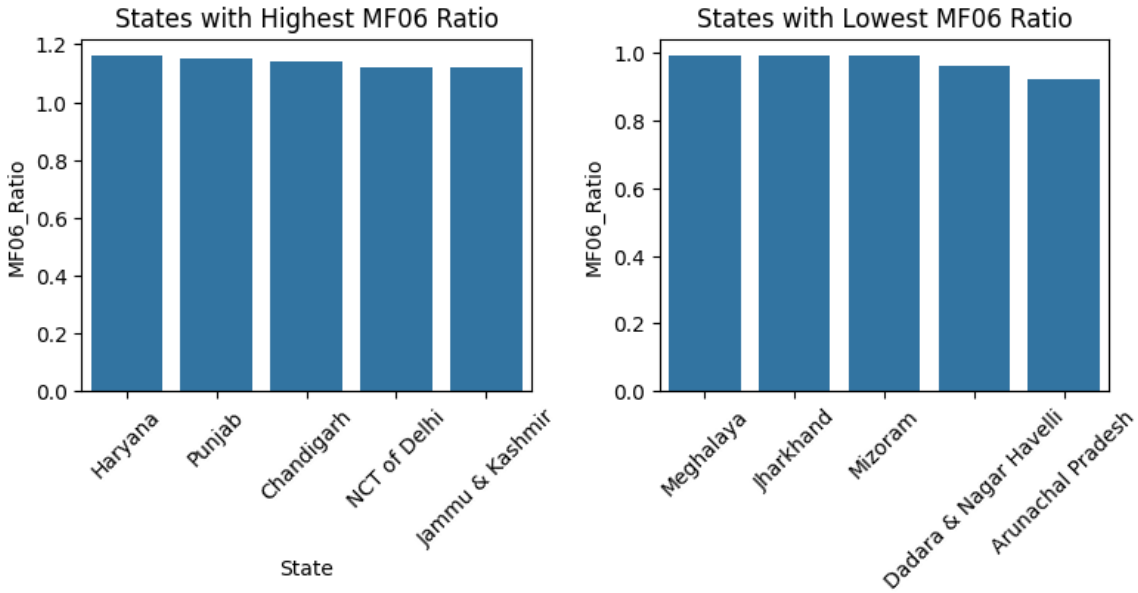

Fig. 2.9 - States with highest and lowest 0-6 MF ratio

	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
0	7707	23388	29796	5862	6196	3	0	1999	2598	13381	...	1150	749
1	6218	19585	23102	4482	3733	7	6	427	517	10513	...	525	715
2	4452	6546	10964	1082	1018	3	6	5806	9723	4534	...	114	188
3	1320	2784	4206	563	677	0	0	2666	3968	1842	...	194	247
4	11654	20591	29981	5157	4587	20	33	7670	10843	13243	...	874	1928

Fig. 2.10 - Head of numerical data sliced for PCA

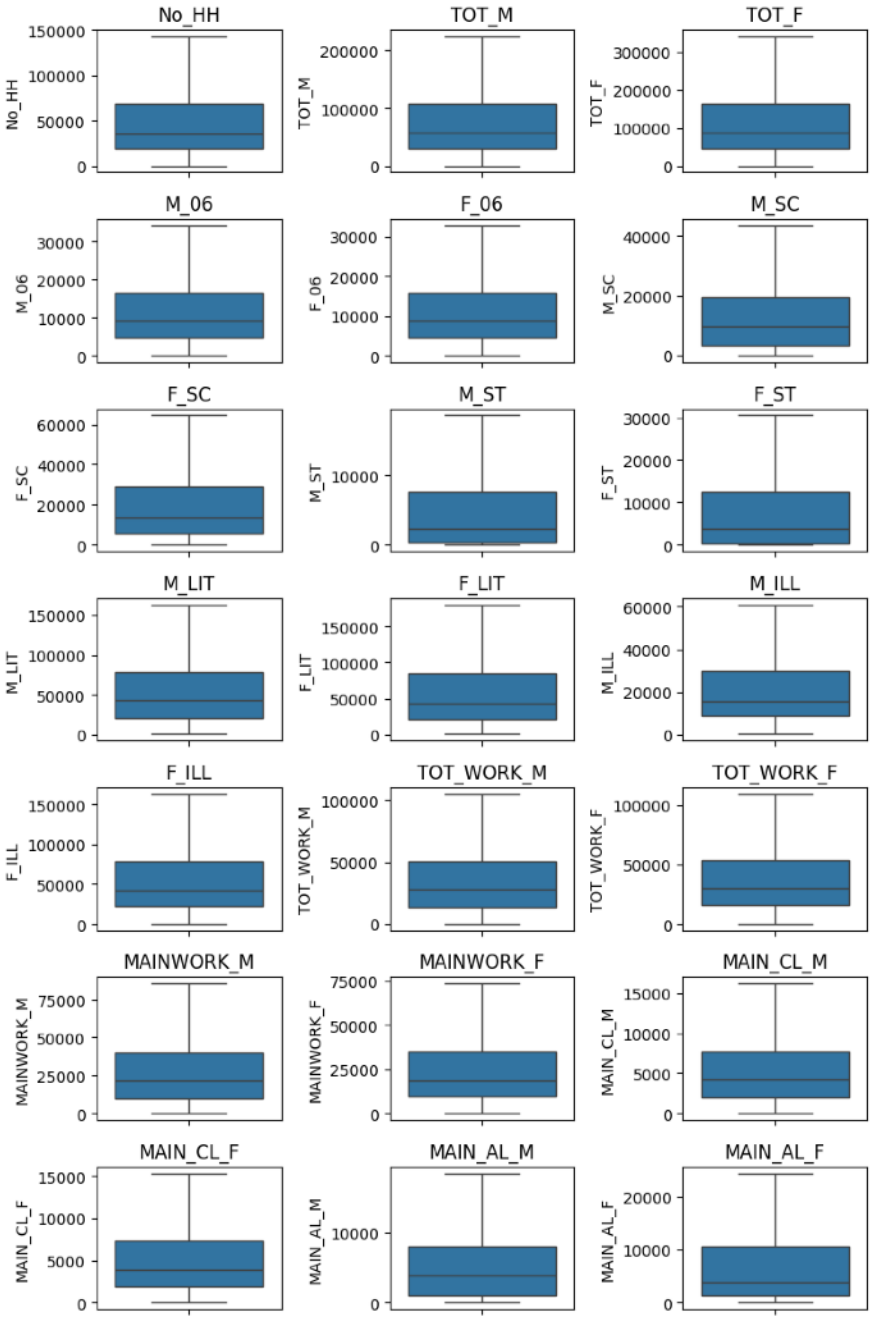

Fig. 2.11 - Boxplots of various features after treatment

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...
0	-1.038986	-0.874837	-0.937027	-0.624685	-0.561282	-1.080201	-1.079963	-0.510440	-0.574198	-0.939617	...
1	-1.076896	-0.938023	-1.009723	-0.773932	-0.835657	-1.079873	-1.079635	-0.771833	-0.782092	-1.005083	...
2	-1.121858	-1.154665	-1.141539	-1.141642	-1.138104	-1.080201	-1.079635	0.122588	0.137599	-1.141561	...
3	-1.201599	-1.217171	-1.214930	-1.197772	-1.176091	-1.080447	-1.079963	-0.399531	-0.437333	-1.203009	...
4	-0.938495	-0.921309	-0.935018	-0.700931	-0.740523	-1.078807	-1.078160	0.432534	0.249489	-0.942767	...

Fig. 2.12 - Head of scaled data using Z-score method

2.3 PCA

The number of PCA components was arbitrarily selected at 25.

```
array([[ 0.15,  0.16,  0.16, ...,  0.14,  0.15,  0.14],
       [-0.12, -0.08, -0.09, ...,  0.04, -0.05, -0.04],
       [ 0.1 , -0.04,  0.03, ..., -0.1 , -0.13, -0.03],
       ...,
       [-0.05,  0.05, -0.01, ...,  0.12,  0.1 ,  0.2 ],
       [-0.  , -0.01,  0.  , ..., -0.21,  0.15,  0.4 ],
       [-0.15,  0.03, -0.06, ...,  0.14, -0.09,  0.1 ]])
```

Fig. 2.13 - PCA components

```
array([35.64886379,  7.64357559,  3.76919551,  2.77722349,  1.90694892,
        1.1549031 ,  0.98772671,  0.46462991,  0.39670851,  0.32234689,
        0.27320737,  0.23564757,  0.18140111,  0.16924377,  0.13859233,
        0.13150585,  0.10380967,  0.09553338,  0.08585804,  0.08091387,
        0.06601791,  0.0630798 ,  0.04827561,  0.04595057,  0.04377475])
```

Fig. 2.14 - PCA explained variance

```
array([0.62444145, 0.13388829, 0.06602291, 0.04864709, 0.03340297,
       0.0202298 , 0.01730146, 0.00813867, 0.00694892, 0.00564637,
       0.00478562, 0.00412771, 0.0031775 , 0.00296455, 0.00242765,
       0.00230352, 0.00181838, 0.00167341, 0.00150393, 0.00141732,
       0.0011564 , 0.00110493, 0.00084562, 0.00080489, 0.00076678])
```

Fig. 2.15 - PCA explained variance ratio

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...
PC1	0.149222	0.159169	0.158209	0.156340	0.156814	0.143350	0.143537	0.018849	0.017878	0.155152	...
PC2	-0.115487	-0.080239	-0.093718	-0.020341	-0.014310	-0.079667	-0.087098	0.069101	0.067316	-0.105986	...
PC3	0.101528	-0.038662	0.028959	-0.074419	-0.068223	-0.037619	0.021350	0.323827	0.338705	-0.032107	...
PC4	0.076814	0.052976	0.070022	0.028520	0.016398	0.010210	0.016244	0.091143	0.079554	0.089187	...
PC5	-0.012090	-0.042344	-0.022927	-0.080339	-0.078326	-0.167893	-0.158092	0.418412	0.415965	-0.014033	...
PC6	0.082558	0.073667	0.082812	0.092379	0.080010	0.050969	0.054568	-0.231809	-0.214542	0.081378	...
PC7	0.106896	-0.124085	-0.010291	-0.200807	-0.203411	-0.040399	0.053990	-0.355238	-0.327677	-0.067062	...
PC8	-0.099515	-0.108870	-0.115276	-0.132944	-0.139342	0.189170	0.177363	-0.071632	-0.078392	-0.102886	...
PC9	0.026100	0.032856	0.036405	0.138404	0.165715	-0.531744	-0.515063	-0.113019	-0.136031	-0.017445	...
PC10	0.068124	-0.048423	-0.022468	-0.157252	-0.145040	-0.098447	-0.065840	-0.008382	-0.028617	0.000581	...
PC11	-0.058605	0.029489	-0.020147	-0.009180	-0.025584	-0.194623	-0.250356	-0.082493	-0.081426	0.023816	...
PC12	-0.021808	-0.047662	-0.042837	-0.146640	-0.144601	-0.122639	-0.114550	-0.055520	-0.051232	0.034683	...
13	-0.017069	0.005930	0.004689	0.033535	0.034778	-0.134248	-0.153439	-0.047831	-0.020856	0.033251	

Fig. 2.16 - Extracted loadings

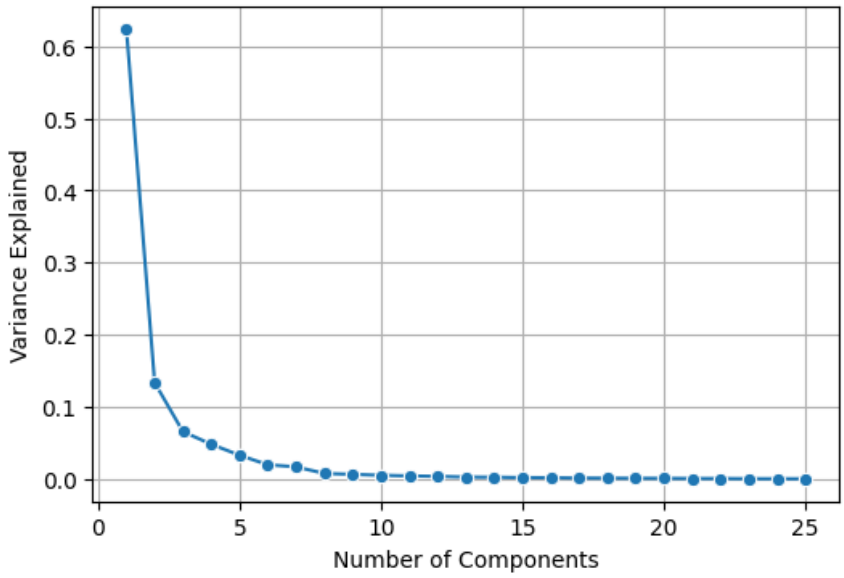

Fig. 2.17 - Scree plot of Components and Explained Variance

```
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])
```

Fig. 2.18 - Cumulative sum of explained variance ratio

We see that 90% variance is being explained with just 5 components.

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...
PC1	0.149222	0.159169	0.158209	0.156340	0.156814	0.143350	0.143537	0.018849	0.017878	0.155152	...
PC2	-0.115487	-0.080239	-0.093718	-0.020341	-0.014310	-0.079667	-0.087098	0.069101	0.067316	-0.105986	...
PC3	0.101528	-0.038662	0.028959	-0.074419	-0.068223	-0.037619	0.021350	0.323827	0.338705	-0.032107	...
PC4	0.076814	0.052976	0.070022	0.028520	0.016398	0.010210	0.016244	0.091143	0.079554	0.089187	...
PC5	-0.012090	-0.042344	-0.022927	-0.080339	-0.078326	-0.167893	-0.158092	0.418412	0.415965	-0.014033	...

Fig. 2.19 - Extracted loadings of top 5 components

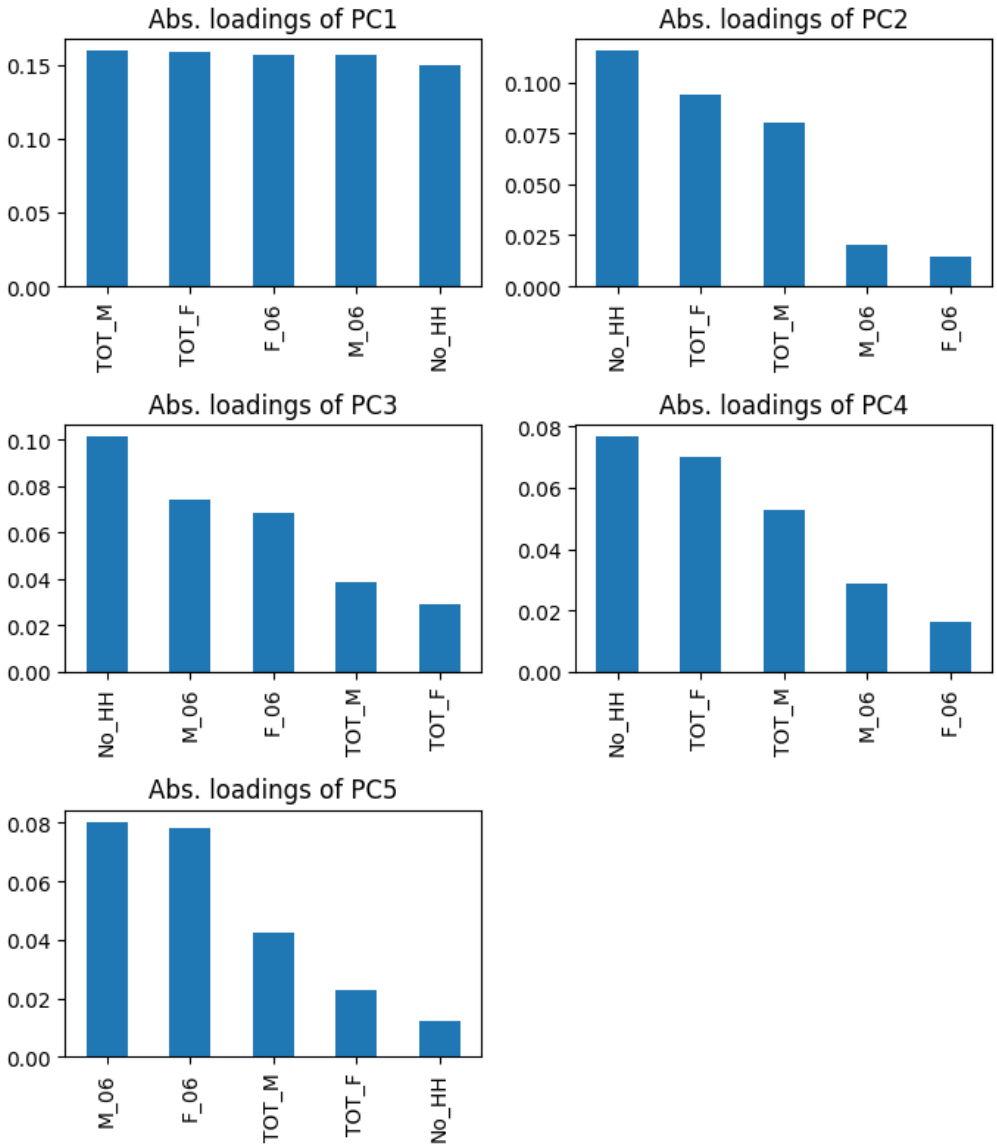

Fig. 2.20 - Absolute loadings of the 5 components

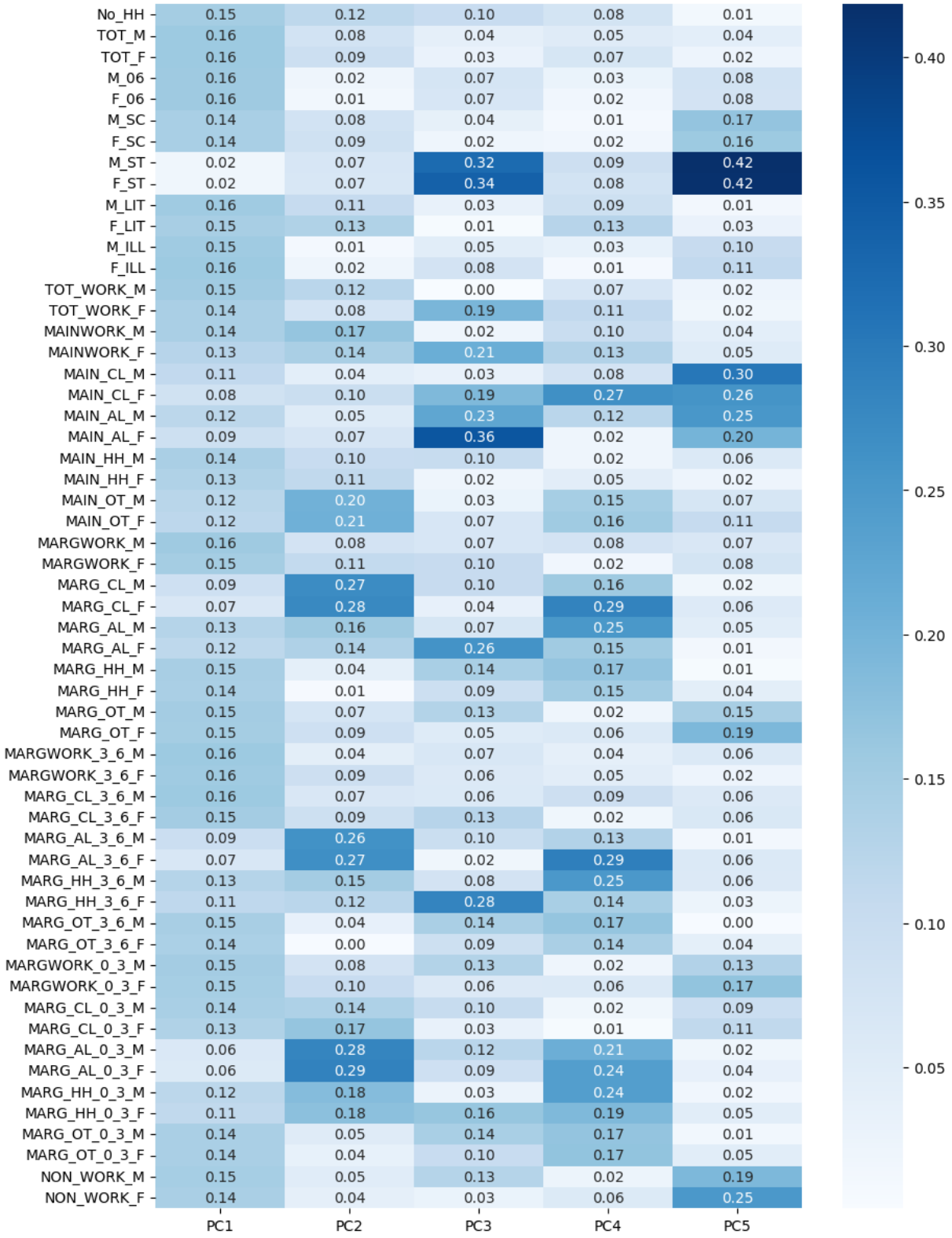

Fig. 2.21 - Absolute loadings of all features

	PC1	PC2	PC3	PC4	PC5
0	-5.528161	0.430378	-1.473827	-1.278049	0.376358
1	-5.492016	-0.106110	-2.015641	-1.750168	-0.006857
2	-7.474643	-0.217194	-0.247428	0.006079	0.556282
3	-7.919737	-0.652311	-0.659220	-0.735550	0.272465
4	-5.175695	2.304059	-1.157327	1.060796	1.080249
5	-3.647563	4.598733	-1.742810	3.301079	1.212639

Fig. 2.22 - Head of final PCA Dataframe

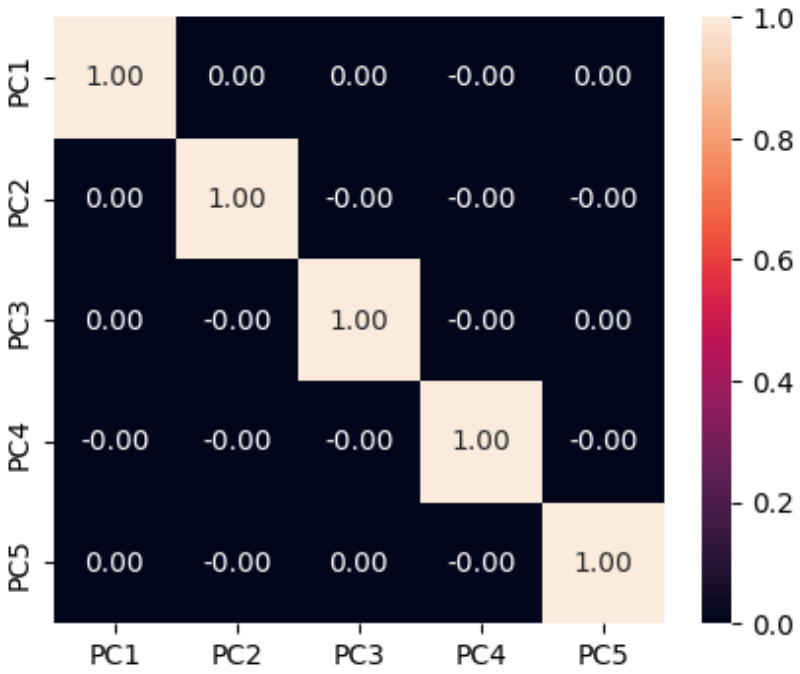

Fig. 2.23 - Heatmap showing no correlation between components