MACHINE LEARNING - 1

PGP DSBA PROGRAM

by: ABHISHEK K HIREMATH

Great Learning
POWER AHEAD

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Problem 1:

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.

Sl. No	Column Name	Column Description
1	Timestamp	The Timestamp of the particular Advertisement.
2	Inventory Type	The Inventory Type of the particular Advertisement. Format 1 to 7. This is a Categorical Variable.
3	Ad - Length	The Length Dimension of the particular Advertisement.
4	Ad- Width	The Width Dimension of the particular Advertisement.
5	Ad Size	The Overall Size of the particular Advertisement. Length*Width.
6	Ad Type	The type of the particular Advertisement. This is a Categorical Variable.
7	Platform	The platform in which the particular Advertisement is displayed. Web, Video or App. This is a Categorical Variable.

8	Device Type	The type of the device which supports the particular Advertisement. This is a Categorical Variable.
9	Format	The Format in which the Advertisement is displayed. This is a Categorical Variable.
10	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.
11	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.
12	Impressions	The impression counts of the particular Advertisement out of the total available impressions.
13	Clicks	It is a marketing metric that counts the number of times users have clicked on the particular advertisement to reach an online property.
14	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.
15	Fee	The percentage of the Advertising Fees payable by Franchise Entities.
16	Revenue	It is the income that has been earned from the particular advertisement.
17	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.
18	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.
19	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

Q1.1) Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.

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

	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	0.0	0.35	0.0	0.0031	0.0	0.0
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.0	0.35	0.0	0.0035	0.0	0.0
2	2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0	0.35	0.0	0.0028	0.0	0.0
3	2020-9-3-20	Format1	300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.0	0.35	0.0	0.0020	0.0	0.0
4	2020-9-4-15	Format1	300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.0	0.35	0.0	0.0041	0.0	0.0

Fig: -1: Dataset Head

	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
23061	2020-9-13-7	Format5	720	300	216000	Inter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
23062	2020-11-2-7	Format5	720	300	216000	Inter224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.0260	NaN	NaN	NaN
23063	2020-9-14-22	Format5	720	300	216000	Inter218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	NaN	NaN	NaN
23064	2020-11-18-2	Format4	120	600	72000	inter230	Video	Mobile	Video	7	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
23065	2020-9-14-0	Format5	720	300	216000	Inter221	App	Mobile	Video	2	2	2	1	0.09	0.35	0.0585	NaN	NaN	NaN

Fig: -2: Dataset Tail

```
<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
```

Fig: -3: Data info

Shape of the dataset:

- The dataset has 23066 rows & 19 columns Checking for Null Values
- Above these are the head and tail of the dataset, and there are 6 float, 7 integer and 6 object columns

Q1.2) Treat missing values in CPC, CTR and CPM using the formula given.

```
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
```

```
Timestamp                0.000000
InventoryType            0.000000
Ad - Length              0.000000
Ad- Width                0.000000
Ad Size                  0.000000
Ad Type                  0.000000
Platform                 0.000000
Device Type              0.000000
Format                   0.000000
Available_Impressions    0.000000
Matched_Queries          0.000000
Impressions              0.000000
Clicks                   0.000000
Spend                    0.000000
Fee                      0.000000
Revenue                  0.000000
CTR                      0.205324
CPM                      0.205324
CPC                      0.205324
dtype: float64
```

Fig: -4: Null values of Dataset

- We can see that there are 3 variables wherein we have Null values
- CTR, CPM, CPC having 4736 Null values each
- Checking for Duplicate values
- There are no duplicate values in the dataset

We created three functions such as ‘calculate CPC’, ‘calculate CTR’, and ‘calculate CPM’ to treat missing values in CPC, CTR, and CPM columns using the following formula.

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.

```
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
```

```
<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                    23066 non-null  float64
 17  CPM                    23066 non-null  float64
 18  CPC                    23066 non-null  float64
dtypes: float64(6), int64(7), object(6)
memory usage: 3.3+ MB
```

Fig: -6: Dataset after treating Null values

Five Point Summary of the Dataset:

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

Fig: -7: Data describe

- As we can see there are no null values now

Q1.3) Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an analyst your judgement may be different from another analyst).

I have checked with the data and it seems that there are Outliers. Below is the Boxplot figure of Features before Treating Outliers.

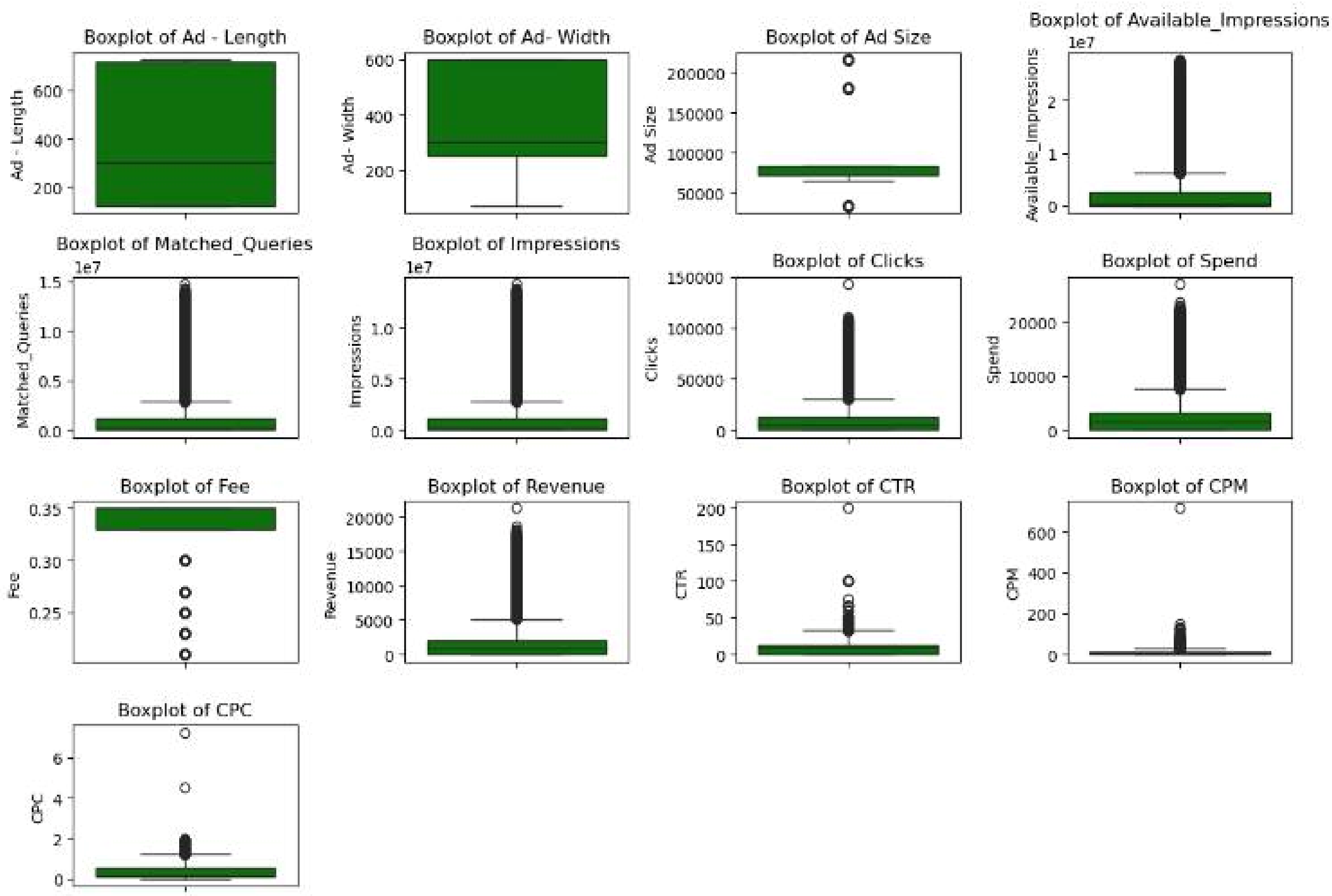


Fig: -8: Boxplot before Treating outliers

Yes. Treating Outliers is necessary for K Means Clustering. We are going to treat outliers by IQR method (Inter Quartile Range).

I have created a 'remove outlier' function using IQR formulas. We can't perform Outlier Treatment on Categorical features. Hence, I have created new dataset of Int64 and Float64 datatypes with the name of df_Num. And, applied 'remove outlier' function on it. And removed outliers.
To treat outliers, we defined a function 'treat outlier' where:

- The larger values (>upper whisker) are all equated to the 95th percentile value of the distribution.
- The smaller values (<lower whisker) are all equated to the 5th percentile value of the distribution.

  - **Upper Range = Q3 + 1.5*IQR.**
  - **Lower Range = Q1 -1.5*IQR**

Find below Boxplot diagram after treating Outliers.

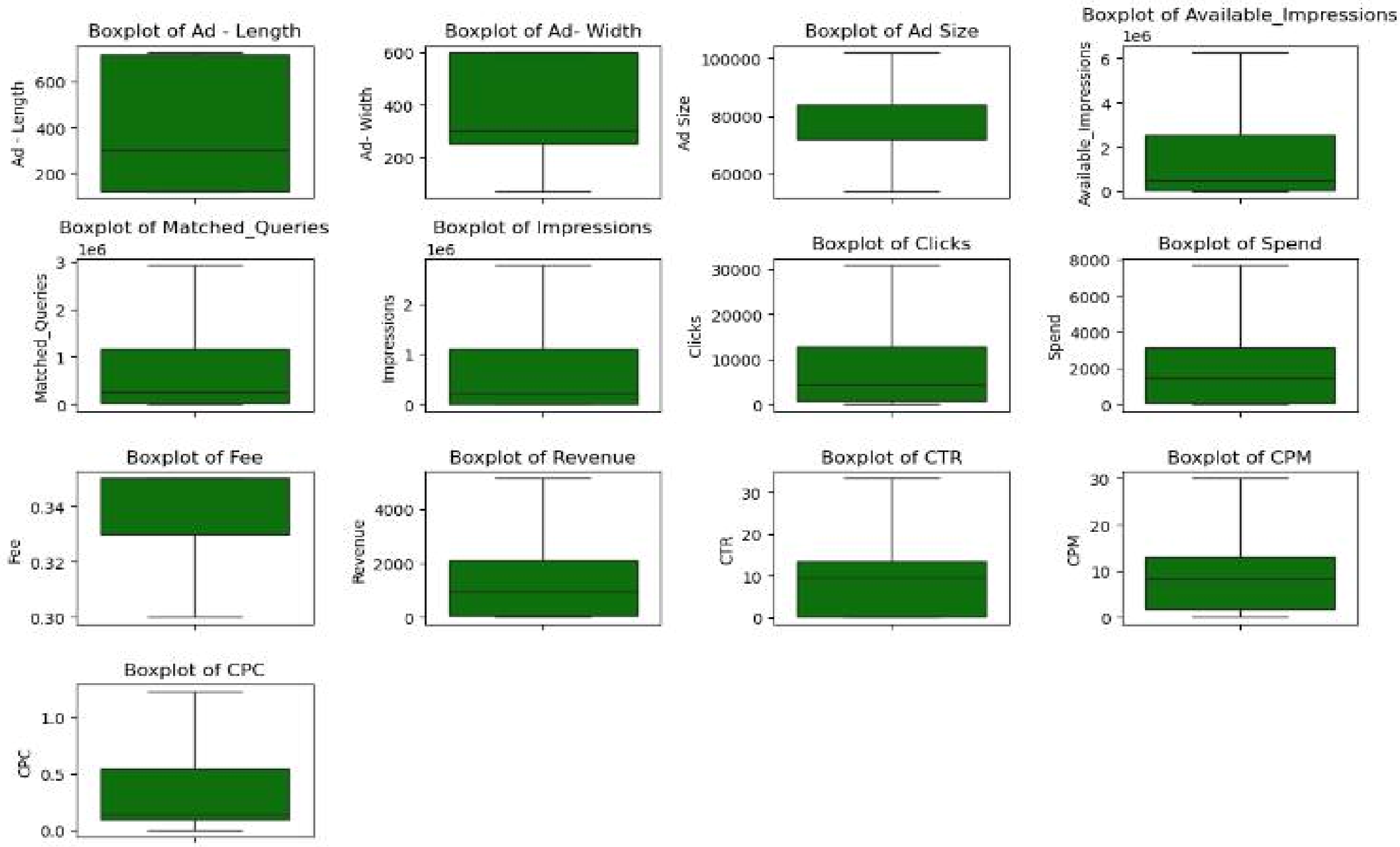


Fig: -9: Boxplot After treating outliers

Q1.4) Perform z-score scaling and discuss how it affects the speed of the algorithm?

Data before Z-score Scaling is as below

	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	7.657684e+04	1.538132e+04	54000.000000	72000.000000	72000.000000	8.400000e+04	1.020000e+05
Available_Impressions	23066.0	1.607253e+06	2.125528e+06	1.000000	33672.250000	483771.000000	2.527712e+06	6.268771e+06
Matched_Queries	23066.0	7.995380e+05	1.026037e+06	1.000000	18282.500000	258087.500000	1.180700e+06	2.924326e+06
Impressions	23066.0	7.536120e+05	9.802568e+05	1.000000	7990.500000	225290.000000	1.112428e+06	2.769086e+06
Clicks	23066.0	8.306828e+03	9.574779e+03	1.000000	710.000000	4425.000000	1.279375e+04	3.091938e+04
Spend	23066.0	2.166060e+03	2.425190e+03	0.000000	85.180000	1425.125000	3.121400e+03	7.675730e+03
Fee	23066.0	3.402883e-01	1.812855e-02	0.300000	0.330000	0.350000	3.500000e-01	3.500000e-01
Revenue	23066.0	1.449389e+03	1.646894e+03	0.000000	55.365375	926.335000	2.091338e+03	5.145297e+03
CTR	23066.0	8.223203e+00	8.253522e+00	0.010874	0.265107	9.391248	1.347057e+01	3.327877e+01
CPM	23066.0	8.219181e+00	6.881016e+00	0.000000	1.749084	8.371566	1.304202e+01	2.998142e+01
CPC	23066.0	3.300346e-01	3.165682e-01	0.000000	0.089736	0.139347	5.462421e-01	1.231002e+00

Fig: -10 After applying Z-score

Here, I have applied z-score method on the 'df_Num'. And, I got the below output.

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	1.281478e-16	1.000022	-1.134891	-1.134891	-0.364496	1.433093	1.467332
Ad- Width	23066.0	-1.182903e-16	1.000022	-1.319110	-0.432797	-0.186599	1.290590	1.290590
Ad Size	23066.0	3.055833e-16	1.000022	-1.467840	-0.297564	-0.297564	0.482620	1.652896
Available_Impressions	23066.0	9.857525e-18	1.000022	-0.756182	-0.740341	-0.528577	0.433059	2.193158
Matched_Queries	23066.0	1.971505e-17	1.000022	-0.779265	-0.761447	-0.527722	0.371498	2.070914
Impressions	23066.0	0.000000e+00	1.000022	-0.768806	-0.760655	-0.538975	0.366051	2.056111
Clicks	23066.0	-1.182903e-16	1.000022	-0.867488	-0.793438	-0.405431	0.468629	2.361729
Spend	23066.0	-9.857525e-17	1.000022	-0.893170	-0.858046	-0.305523	0.393932	2.271900
Fee	23066.0	1.143473e-15	1.000022	-2.222416	-0.567532	0.535724	0.535724	0.535724
Revenue	23066.0	3.943010e-17	1.000022	-0.880093	-0.846474	-0.317607	0.389803	2.244218
CTR	23066.0	1.380054e-16	1.000022	-0.995031	-0.964227	0.141524	0.635787	3.035808
CPM	23066.0	2.464381e-17	1.000022	-1.194498	-0.940303	0.022146	0.700905	3.162718
CPC	23066.0	3.943010e-17	1.000022	-1.042561	-0.759091	-0.602371	0.682987	2.846105

Fig: -11

Scaling of variables is important for clustering to stabilize the weights of the different variables. If there is wide discrepancy in the range of variables (refer to Table 3) cluster formation may be affected by weight differential.

The features contained in a data set may have different units (e.g. feet, kilometers, and hours) that, in turn, may mean that the variables have different scales. All machine learning algorithms are dependent on the scaling of data. If there is wide discrepancy among the input values, the unscaled model may be unstable, meaning that it may suffer from poor performance during learning and sensitivity to input values resulting in higher generalization error. [2]

One of the most common forms of pre-processing consists of a simple linear rescaling of the input variables.

— Page 298, Neural Networks for Pattern Recognition, 1995.

Q1.5) Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.

Using SciPy's cluster hierarchy function, we created the below dendrogram.

Please find below Dendrogram performed for Hierarchical using WARD and Euclidean Distance on the Scaled Data such as "df_scaled".

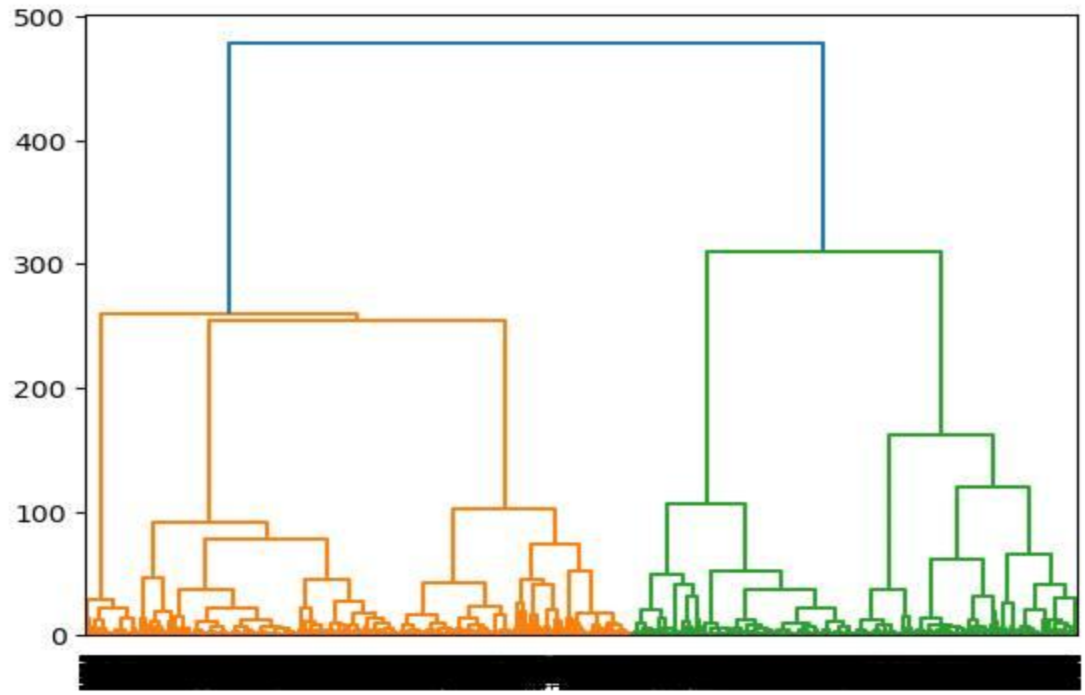
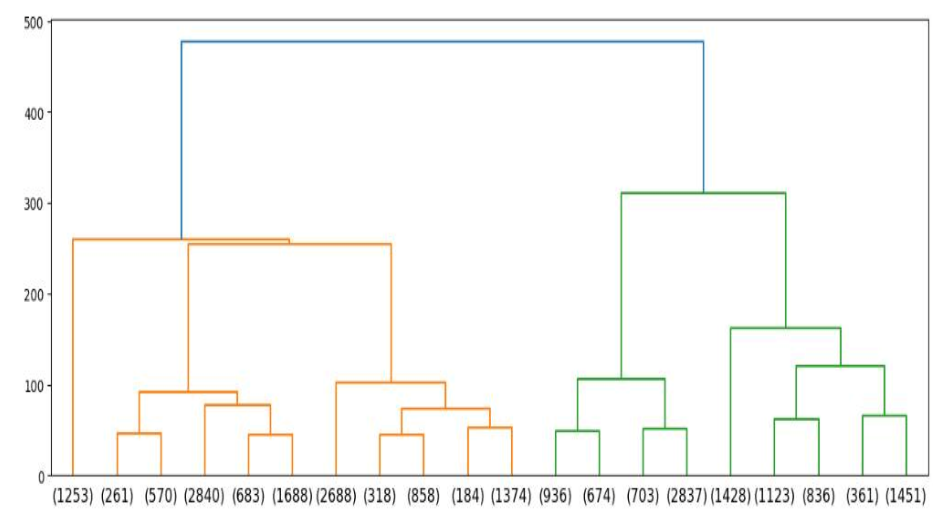


Fig: -12

In a Dendrogram, each branch is called a clade. The terminal end of each clade is called a leaf. The arrangement of the clades tells us which leaves are most similar to each other. The height of the branching points indicates how similar or different they are from each other: the greater the height, the greater the difference.

[reference - https://wheatoncollege.edu/wp-content/uploads/2012/08/How-to-Read-a-Dendrogram-Web-Ready.pdf]

Keeping the above reference as base, we can see the longest branch (tallest branch) is in blue. If we see that only blue, it will result in only 2 clusters which is not acceptable in business. If the segmentation is at the tallest green branches, separated by the yellow horizontal line, 5 clusters are identified. Alternatively, there may be 3 clusters as well, designated by the orange horizontal line. But we choose 5 Clusters using Dendrogram for this project.




```
clusters
1    1253
2    6042
3    5422
4    5150
5    1428
6    3771
Name: count, dtype: int64
```

Fig: -13: Value counts

	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
23061	2020-9-13-7	Format5	720	300	216000	Inter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
23062	2020-11-2-7	Format5	720	300	216000	Inter224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.0260	NaN	NaN	NaN
23063	2020-9-14-22	Format5	720	300	216000	Inter218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	NaN	NaN	NaN
23064	2020-11-18-2	Format4	120	600	72000	inter230	Video	Mobile	Video	7	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
23065	2020-9-14-0	Format5	720	300	216000	Inter221	App	Mobile	Video	2	2	2	1	0.09	0.35	0.0585	NaN	NaN	NaN

Fig: -14

Q1.6) Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.

For checking the Optimal number of clusters, we use WSS (Within Sum of Square)

Elbow Plot for n=10 The optimum number of clusters for k-means algorithm are 5 as the drop become significant

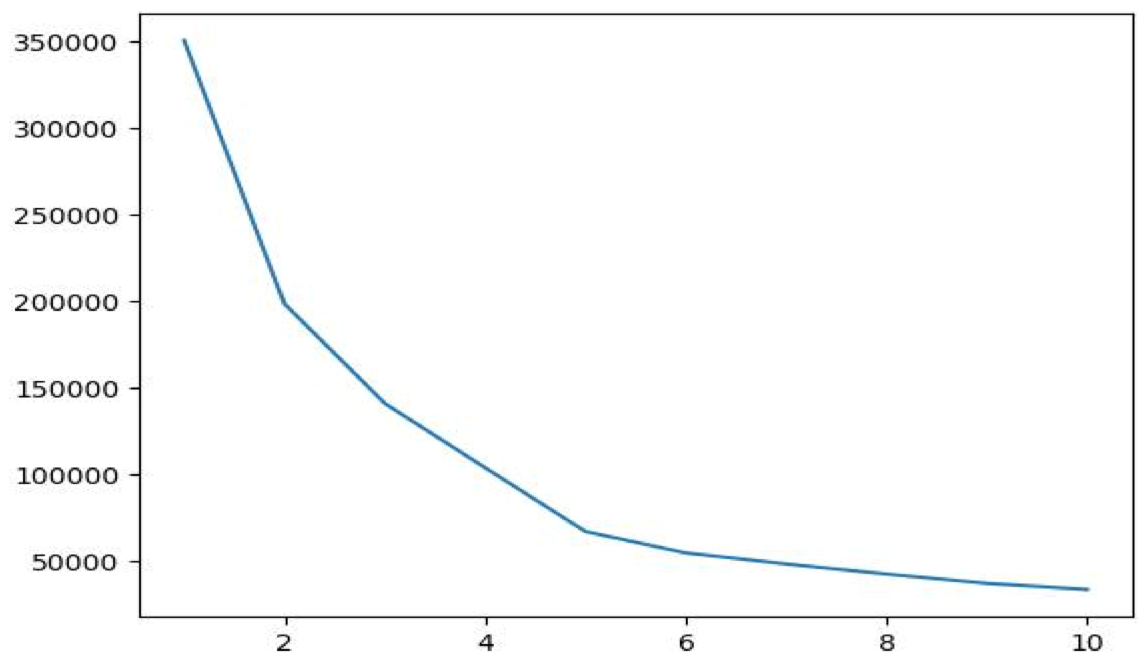


Fig: -15: Line plot

As per the check, when we move from K=1 to K=2, We see that there is a significant drop in the value. Also, when we move from k=2 to k-3, k=3 to k=4, k=4 to k=5 there is a significant drop as well. k=5 to k=6, the drop in values reduces significantly. Hence In this case, the WSS is not significantly dropping beyond 5, so 5 is optimal number of clusters

Q1.7) Print silhouette scores for up to 10 clusters and identify optimum number of clusters.

The silhouette score here is 0.5884606993260426**.**

The silhouette score for rest clusters up to 10.

Hierarchical Clustering as well as K-Means Clustering were performed. We used Elbow plot and Silhouette Score to identify optimum number of clusters in K-Means whereas in Hierarchical Clustering dendrogram was drawn. In Hierarchical method, we got 5 clusters while in K-Means, we got 5 (using elbow plot) and 6 clusters (using silhouette score). We can always try alternative approaches to clustering using other linkage types and distance metrics for an exhaustive study of the data. Please refer to the Monograph for details. We observe that the methods used in this project yielded similar results i.e. with 10 clusters. I have calculated Silhouette Score for scaled data using the silhouette_score () function.

The Silhouette Score is a measure of how similar an object is to its own cluster compared to other clusters, and it ranges from -1 to 1, with higher values indicating better clustering.

```
For no of clusters=2
257208.55083081475
The Silhouette Score is 0.4262713517981147
For no of clusters=3
166201.76151655504
The Silhouette Score is 0.4503758162120636
For no of clusters=4
114167.44887088305
The Silhouette Score is 0.5206669400279987
For no of clusters=5
75150.1956381825
The Silhouette Score is 0.578686139946828
For no of clusters=6
54527.912364221775
The Silhouette Score is 0.5884606993260426
For no of clusters=7
48910.940998052625
The Silhouette Score is 0.5639923095944765
For no of clusters=8
44431.97496451946
The Silhouette Score is 0.5533348395524547
For no of clusters=9
40022.34712205972
The Silhouette Score is 0.5083345223282044
For no of clusters=10
35988.14230793693
The Silhouette Score is 0.5257398681916272
```

Fig: -16: Clusters

Cluster_kmeans	0	1	2	3	4	5
Ad - Length	585.518868	4.652640e+02	120.310452	666.278788	4.236465e+02	138.381963
Ad- Width	324.036950	1.995169e+02	599.939634	306.380471	1.457295e+02	576.790451
Ad Size	182466.981132	7.524420e+04	72168.161435	201620.202020	5.327849e+04	75230.769231
Available_Impressions	96564.252752	1.039348e+07	31922.152984	365758.123569	1.808062e+06	811693.090849
Matched_Queries	50610.035770	5.626474e+06	19893.140738	198510.360943	8.646094e+05	571231.505305
Impressions	44466.499607	5.448141e+06	13683.866333	168224.074411	8.264839e+05	481819.978117
Clicks	2892.957547	1.125094e+04	2015.829079	20953.694949	3.256739e+03	65802.228117
Spend	320.582119	8.634991e+03	220.571908	1796.794226	1.508504e+03	7050.280007
Fee	0.350000	2.906507e-01	0.349976	0.349104	3.491436e-01	0.287487
Revenue	208.378412	6.364918e+03	143.446296	1171.041850	9.833242e+02	5064.143266
CTR	0.124121	2.174826e-03	0.146998	0.125305	4.054172e-03	0.137939
CPM	12.050131	1.560991e+00	13.916212	10.903167	1.792549e+00	15.189614
CPC	0.121555	7.535545e-01	0.105599	0.087395	5.465910e-01	0.110087
clusters	3.040487	5.648016e+00	2.001552	3.000000	4.369042e+00	1.170424
freq	2544.000000	4.057000e+03	5798.000000	2970.000000	6.189000e+03	1508.000000

Fig: -17: Cluster k means

Q1.8) Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots].

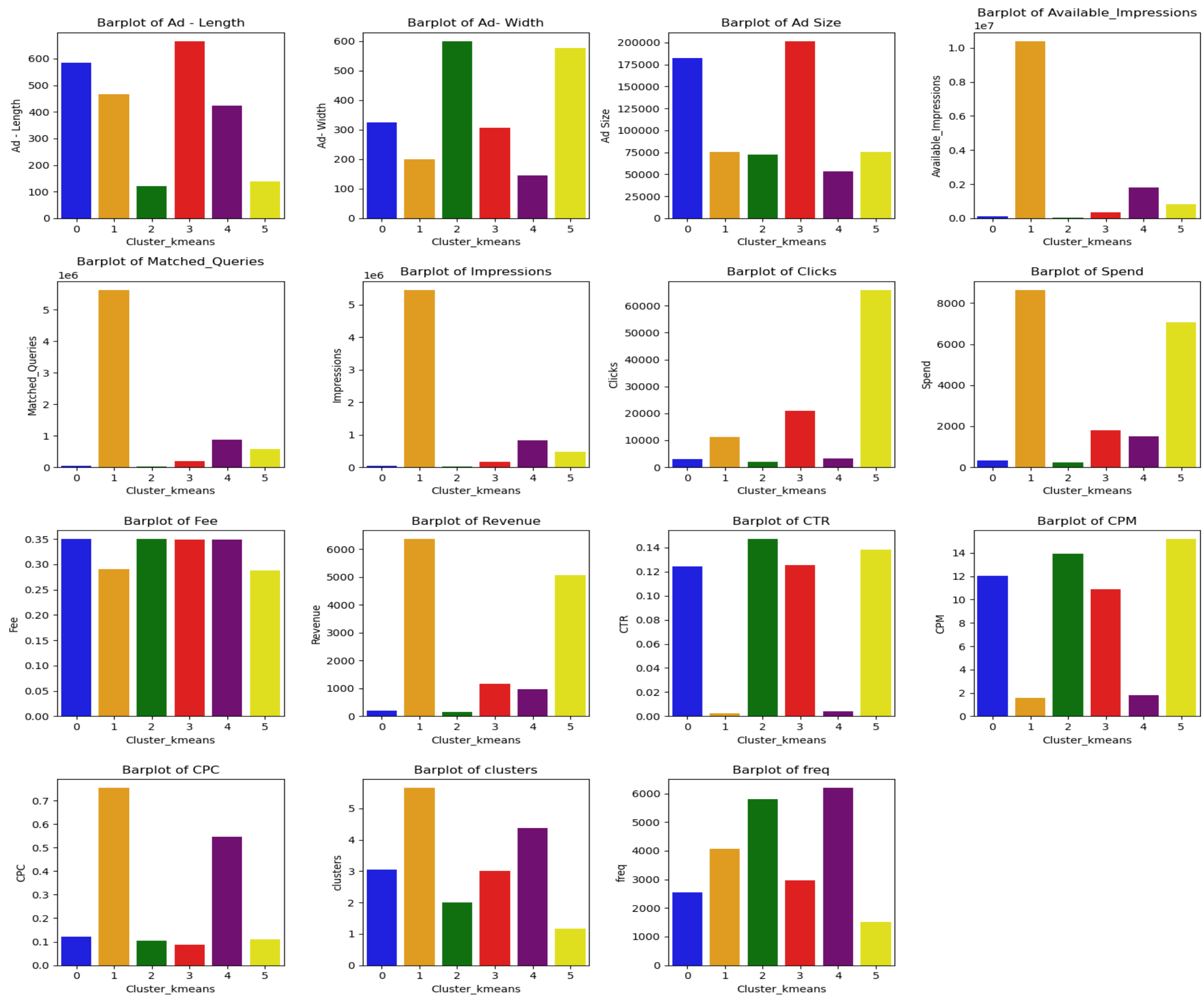


Fig: -18: Bar plot Comparisons

Observations:

- The clusters 3 contain ads that have higher mean length than other clusters.
- The clusters 2 and 5 have ads whose mean width is considerably more than the other clusters
- Cluster 3 has minimum ad size
- Available impressions is highest for cluster 1
- There is not much difference in Fee, but cluster 1 has very high mean spend and mean revenue compared to the others
- Cluster-2 have the most Click through rate (CTR).

- Cluster-1 have the highest cost per 1000 impressions (CPM)
- For cluster 1 and 4 the CPC (Cost per Click) is highest.

The data is grouped/profiled using the optimum number of clusters using silhouette score, which is five, and the mean has been taken to identify the trend in clicks, spends, revenue, CPM, CTR and CPC based on device type.

Means Clicks based on Device Type:

Device Type / Cluster_kmeans	Desktop	Mobile
0	14798.108642	14467.467309
1	1932.695069	1876.040752
2	5582.584674	5540.329431
3	14552.251029	14341.473373
4	65777.215722	65707.752066
5	3260.504637	3256.474827

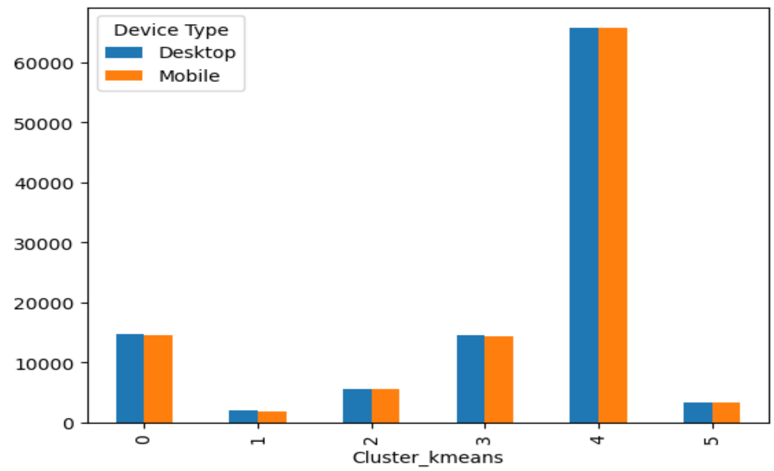


Fig: -19: Mean Clicks based on Device Type

The above bar chart, fig 19, clearly depicts the mean of cluster 4 at the peak for both the desktop and mobile devices at above 650000 clicks. Followed by Clusters 0 and 3 are at the highest, with means of above 14,798.108 and 14,467.467 clicks, respectively. And cluster 1 and cluster 5 registered with the lowest click.

Mean Spend based on Device type:

Device Type / Cluster_kmeans	Desktop	Mobile
0	12231.121901	12124.001136
1	206.234602	207.377217
2	3616.764636	3603.272831
3	1255.262499	1254.676857
4	7032.511554	7043.678864
5	1232.521184	1236.407404

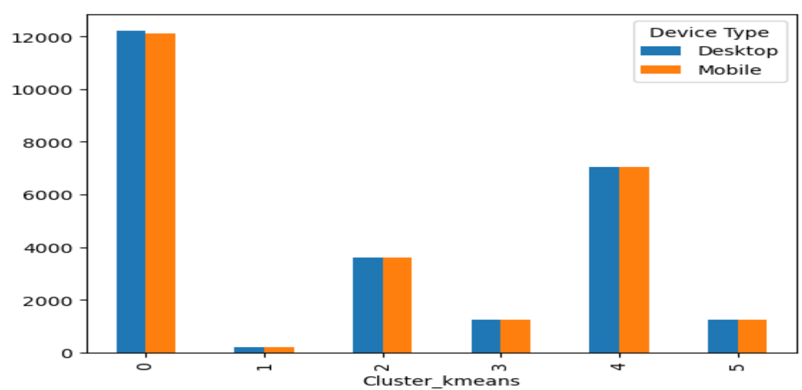


Fig: -20: Mean Spend based on Device type

The above figure 20 bar chart represents clustered data of mean spending based on device type. As we can spend mean, cluster 0 is at the top for both devices and followed the cluster 4 at the second spot. Whereases cluster 1 is at the lowest spend for both the Desktop and mobile at approx.206 and 207.

Mean revenue based on Device type:

Device Type	Desktop	Mobile
Cluster_kmeans		
0	9243.607330	9157.779886
1	134.105277	134.869976
2	2407.526612	2398.090341
3	817.803491	817.320235
4	5048.963459	5059.109156
5	801.209319	803.772784

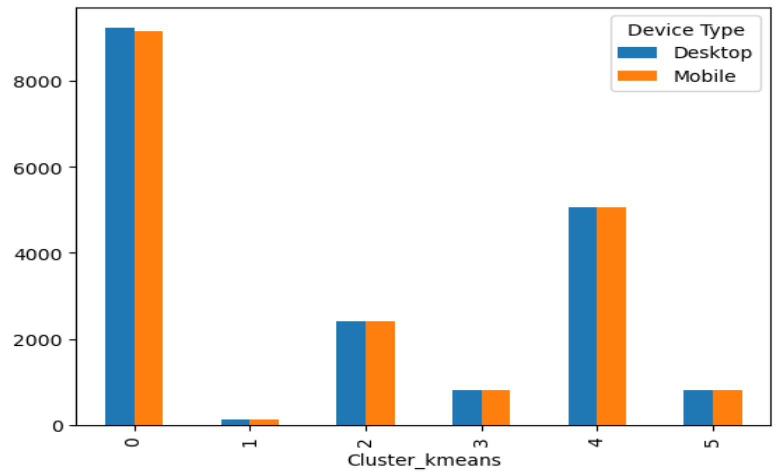


Fig: -21: Mean revenue based on Device type

The above fig 21 chart shows the clustered data of mean revenue based on device type. It is evident that the cluster 0 mean revenue is the highest with approx. 9243 and 9157 and followed by cluster 4 with approx. 5048. Cluster 2 holds the last position with the lowest mean revenue.

Mean CPM (Cost per 1000 impressions) based on Device type:

Device Type	Desktop	Mobile
Cluster_kmeans		
0	1.629837	1.657086
1	14.560995	14.831009
2	1.605973	1.614496
3	12.103770	12.147780
4	15.449728	15.364300
5	1.782974	1.784608

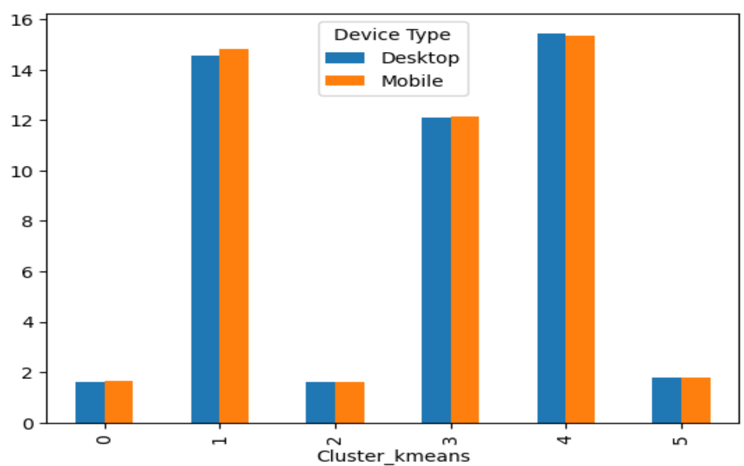


Fig: - 22: Mean CPM (Cost per 1000 impressions) based on Device type

The above chart, fig 22, shows clustered data of the mean of Cost per 1000 Impressions (CPM) based on the device type. As we can see, cluster 4 has the highest mean CPM with close to 16, followed by cluster 1 with approx. 15 for both devices. Clusters 0 and 2 are at the lowest, with approx. 2 mean CPM.

Mean CTR (Click through Rate) based on Device type:

Device Type	Desktop	Mobile
Cluster_kmeans		
0	0.202902	0.200721
1	16.123664	16.081285
2	0.228507	0.229312
3	13.821226	13.885580
4	13.789323	13.731714
5	0.449947	0.445778

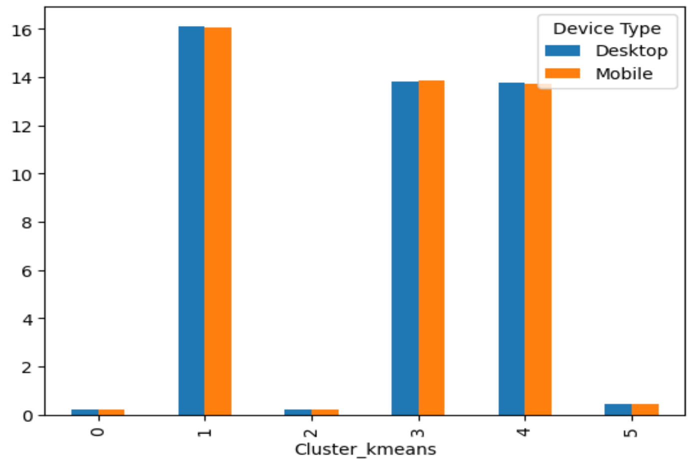


Fig: -23: Mean CTR (Click through Rate) based on Device type

The above fig 23 chart shows clustered data of mean of Click Through Rate (CTR) based on the device type. The cluster 1 is at the highest mean CTR for the both mobile and desktop device at 16. Followed by the cluster 3 and cluster 4 at mean CTR of 14 and Cluster 3 at the lowest at cluster 1 and cluster 0.

Mean CPC (Cost per Click) based on Device type:

Device Type	Desktop	Mobile
Cluster_kmeans		
0	0.831574	0.857649
1	0.101966	0.101802
2	0.765023	0.763597
3	0.089483	0.090316
4	0.112058	0.111906
5	0.453396	0.457362

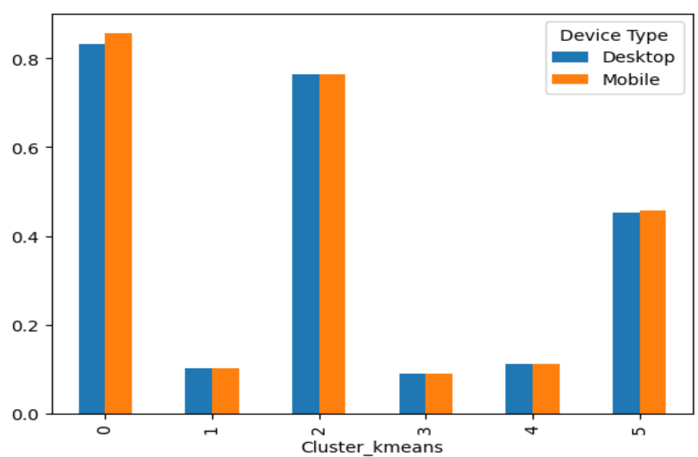


Fig: -24: Mean CPC (Cost per Click) based on Device type:

The above fig 24 chart shows clustered data of mean Spend per Click (CPC) based on the device type. As we can see, cluster 0 is at the peak for both devices at closer to 0.8 and followed by cluster 2 with a mean of 0.76 CPC. Clusters 1, 3 and 4 follow a similar trend with 0.1.

Q1.9) Clustering: Conclude the project by providing summary of your learnings.

Summary of the dataset:

The ad 24x7 marketing company have collected the data from the marketing intelligence to analysis and segmentalize the ads to target the right set of groups. The following details were found during the assessment of the dataset.

- The dataset has 23066 rows and 19 features. Out of 19 features – 6 are float type, 7 are integer and six are categorical.
- During the project, we also found 4736 missing values; those values were imputed and treated with the given formula using a user-defined function.
- The dataset also had outliers; those outliers were treated using the IQR method since the K-mean clustering is sensitive to outliers and could negatively influence the dataset.
- The dataset also has been scaled. Since the unscaled data could negatively impact the speed of the algorithm and scaling data can make the variable contribute equally to the analysis to take better business decisions.
- As per Elbow plot/scree-plot, we concluded that the optimal number of clusters should be 5.
- Plotted elbow plot and got optimum value is 5
- Hierarchal clustering has been performed with the data to find the optimum number of clusters, which is 5.

Conclusions after Clustering:

- When Click on Ads gets increases then Revenue is also increases.
- When amount of money spent on specific ad variations within a specific campaign or ad set is increases then Revenue is also increases.
- When impression count of the particular Advertisement increases then Revenue is also increases

Problem 2:

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

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	Literate population Male

F_LIT	**Literate 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 Laborers Population Male**
MAIN_AL_F	**Main Agricultural Laborers 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 Laborers Population Male**
MARG_AL_F	**Marginal Agriculture Laborers 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 laborer's Population 3-6 Male**
MARG_AL_3_6_F	**Marginal Agriculture Laboure's 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 Laboure's Population 0-3 Male**

MARG_AL_0_3_F	**Marginal Agriculture Laboure's 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

Q2.1 PCA: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.

The below figure shows the first 5 Rows of 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_
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: -25: Head of the Dataset

The below figure shows the last 5 rows of the dataset:

	State Code	Dist.Code	State	Area Name	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
635	34	636	Puducherry	Mahe	3333	8154	11781	1146	1203	21	30	0	0	6916	10184	1238	1597
636	34	637	Puducherry	Karaikal	10612	12346	21691	1544	1533	2234	4155	0	0	10292	14225	2054	7466
637	35	638	Andaman & Nicobar Island	Nicobars	1275	1549	2630	227	225	0	0	1012	1750	1187	1602	362	1028
638	35	639	Andaman & Nicobar Island	North & Middle Andaman	3762	5200	8012	723	664	0	0	28	50	4206	5273	994	2739
639	35	640	Andaman & Nicobar Island	South Andaman	7975	11977	18049	1470	1358	0	0	161	264	10095	13362	1882	4687

Fig: -26: Tail of the dataset

Data Info:

```
<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)
memory usage: 305.1+ KB
```

Fig: -27: Data information

From the above data, we can see 640 rows with 61 columns. Out of 61 features – 59 columns belong to the integer data type, and 2 are object (Categorical data type).

The below fig 20 shows the data that depicts the mean, median, min and max values of the dataset. The dataset looks skewed.

	State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST
count	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000
mean	17.114062	320.500000	51222.871875	79940.576563	122372.084375	12309.098438	11942.300000	13820.946875	20778.392188	6191.807813
std	9.426486	184.896367	48135.405475	73384.511114	113600.717282	11500.906881	11326.294567	14426.373130	21727.887713	9912.668948
min	1.000000	1.000000	350.000000	391.000000	698.000000	56.000000	56.000000	0.000000	0.000000	0.000000
25%	9.000000	160.750000	19484.000000	30228.000000	46517.750000	4733.750000	4672.250000	3466.250000	5603.250000	293.750000
50%	18.000000	320.500000	35837.000000	58339.000000	87724.500000	9159.000000	8663.000000	9591.500000	13709.000000	2333.500000
75%	24.000000	480.250000	68892.000000	107918.500000	164251.750000	16520.250000	15902.250000	19429.750000	29180.000000	7658.000000
max	35.000000	640.000000	310450.000000	485417.000000	750392.000000	96223.000000	95129.000000	103307.000000	156429.000000	96785.000000

Fig: -28: Data Describe

The below fig 21 proves that there are no duplicates and null values present in the dataset.

```
State Code       0
Dist.Code        0
State            0
Area Name        0
No_HH            0
                ..
MARG_HH_0_3_F    0
MARG_OT_0_3_M    0
MARG_OT_0_3_F    0
NON_WORK_M       0
NON_WORK_F       0
Length: 61, dtype: int64
```

Fig: -29: Null values

Checking for duplicate values:

- There are no duplicate values in the given dataset Summary of the data

Q2.2 PCA: Perform detailed Exploratory analysis by creating certain questions like (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio? (Example Questions). Pick 5 variables out of the given 24 variables below for EDA: No_HH, TOT_M, TOT_F, M_06, F_06, M_SC, F_SC, M_ST, F_ST, M_LIT, F_LIT, M_ILL, F_ILL, TOT_WORK_M, TOT_WORK_F, MAINWORK_M, MAINWORK_F, MAIN_CL_M, MAIN_CL_F, MAIN_AL_M, MAIN_AL_F, MAIN_HH_M, MAIN_HH_F, MAIN_OT_M, MAIN_OT_F.

(i)Which state has highest gender ratio and which has the lowest?

I have picked 5 Variables such as 'No_HH', 'TOT_M', 'TOT_F', 'M_06', and 'F_06'. And comparing those 5 variables against 'State' and Area Name.

	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06
0	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196
1	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733
2	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018
3	Jammu & Kashmir	Kargil	1320	2784	4206	563	677
4	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587
...	...	...	...	...	...	...	...
635	Puducherry	Mahe	3333	8154	11781	1146	1203
636	Puducherry	Karaikal	10612	12346	21691	1544	1533
637	Andaman & Nicobar Island	Nicobars	1275	1549	2630	227	225
638	Andaman & Nicobar Island	North & Middle Andaman	3762	5200	8012	723	664
639	Andaman & Nicobar Island	South Andaman	7975	11977	18049	1470	1358

640 rows × 7 columns

Fig: -30: Chosen Variables

I have created the Gender Ratio Columns by calculating the Male/Female, So the new columns are 'TOT_GR' and 'GR_06'.

	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	TOT_GR	GR_06
0	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	0.784938	0.946094
1	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	0.847762	1.200643
2	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	0.597045	1.062868
3	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0.661912	0.831610
4	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	0.686802	1.124264
...	...	...	...	...	...	...	...	...	...
635	Puducherry	Mahe	3333	8154	11781	1146	1203	0.692131	0.952618
636	Puducherry	Karaikal	10612	12346	21691	1544	1533	0.569176	1.007175
637	Andaman & Nicobar Island	Nicobars	1275	1549	2630	227	225	0.588973	1.008889
638	Andaman & Nicobar Island	North & Middle Andaman	3762	5200	8012	723	664	0.649026	1.088855
639	Andaman & Nicobar Island	South Andaman	7975	11977	18049	1470	1358	0.663582	1.082474

640 rows × 9 columns

Fig: -31: Calculated gender Ratio

The below bar graph fig 24 represents the gender ratio among the Indian states. We can see that Lakshadweep has the highest gender ratio with over 0.9, and Chhattisgarh and Andra Pradesh have the lowest gender ratio with 0.5.

State	TOT_GR
Lakshadweep	0.868061
Haryana	0.777742
NCT of Delhi	0.766436
Uttar Pradesh	0.762436
Punjab	0.747843
Bihar	0.746294
Meghalaya	0.736105
Jammu & Kashmir	0.728013
Daman & Diu	0.702191
Chandigarh	0.700037
Rajasthan	0.687548
Assam	0.685665
Jharkhand	0.677664
Gujarat	0.662355
Sikkim	0.657083
West Bengal	0.649317
Dadara & Nagar Havelli	0.644631

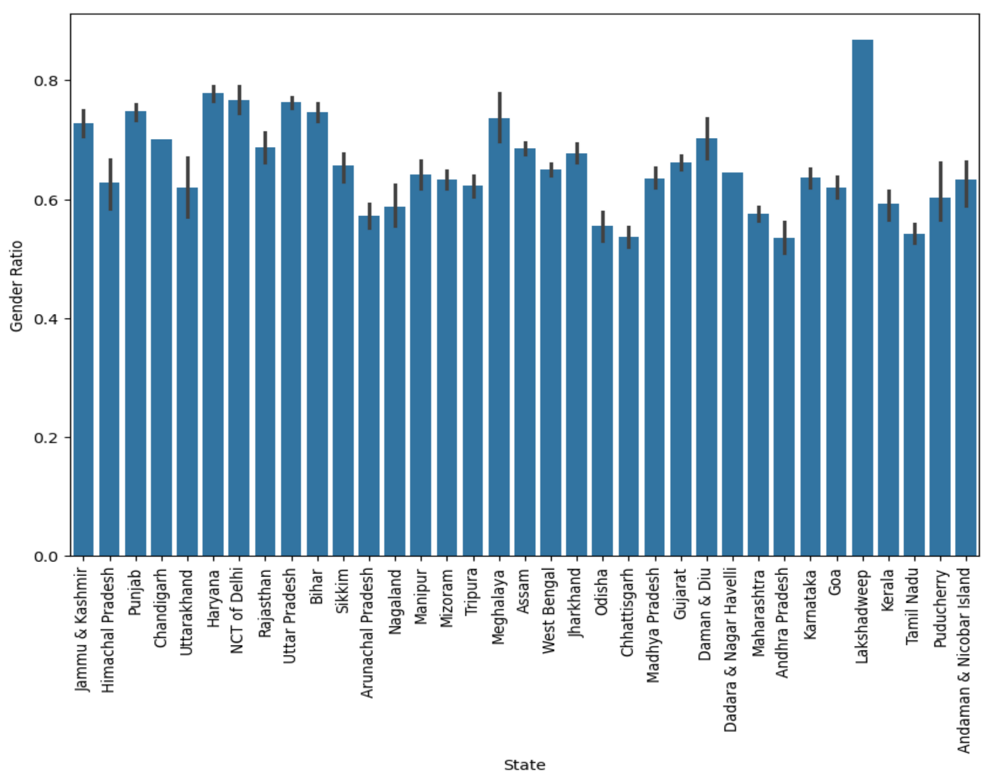


Fig: -32: Gender Ratio

(ii) Which district has the highest & lowest gender ratio?

Dist.Code	TOT_GR
587	0.868061
2	0.847762
144	0.847313
106	0.846911
139	0.844003
299	0.840393
92	0.838542
89	0.831138
160	0.817231
146	0.815491
133	0.814978
76	0.814942
202	0.814942
165	0.814889
142	0.814803
201	0.813884
143	0.812564

Dist.Code	TOT_GR
547	0.437972
398	0.440769
625	0.449352
546	0.450076
391	0.451455
...	...
139	0.844003
106	0.846911
144	0.847313
2	0.847762
587	0.868061

640 rows × 1 columns

Fig: -33: District wise Gender Ratio

As we can see the highest and lowest Gender Ratio, the above fig-33 represents the gender ratio among the Indian Districts. We can see that 587- has the highest gender ratio with over 0.9, and 547 have the lowest gender ratio with 0.5.

Q2.3 We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?

In the case of census data, outlier treatment may not be necessary for several reasons:

1. The data is usually collected from a large and representative sample of the population. This means that the data is likely to be normally distributed, and outliers are less likely to occur
2. Census data is often collected using standardized methods and questionnaires, which reduce the likelihood of errors and outliers
3. The purpose of census data is often to provide an accurate representation of the population as a whole. Outliers, by definition, are not representative of the population and may not provide any useful information.
4. Outliers may also be due to errors or anomalies in the data collection process. In the case of census data, the data collection process is typically rigorous and standardized, making it less likely that errors will occur.
5. Principal Component Analysis (PCA) is a highly flexible multivariate data dimension reduction method. In the presence of outliers, classical PCA is highly sensitive to them and may draw false conclusions.

Q2.4 PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.

Before scaling and checking the outliers, the Categorical variable has been dropped from the dataset. The below image represents the boxplot of the dataset before scaling.

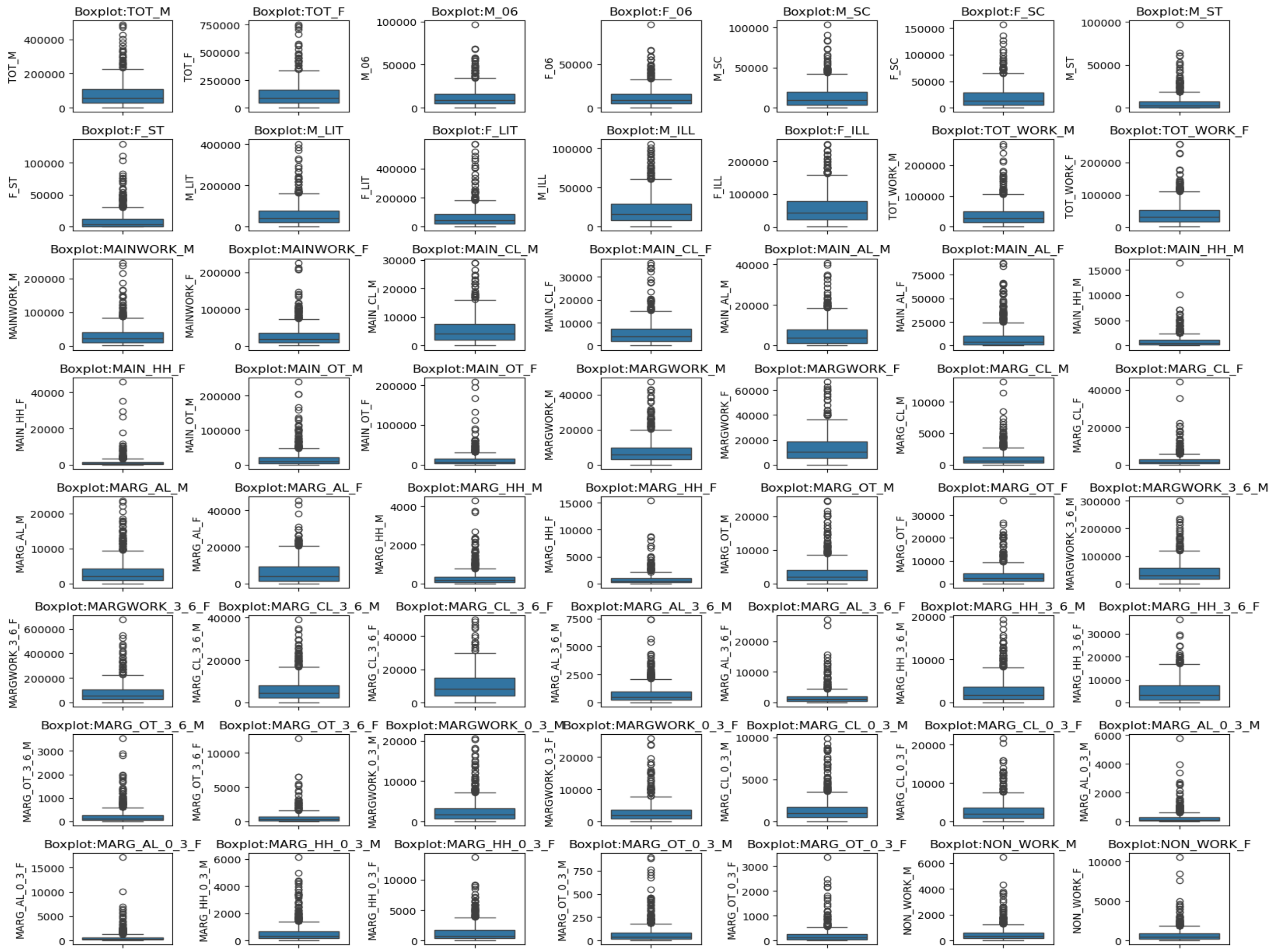


Fig: -34: Boxplot before scaling

As we have applied the Z-score for the Dataset, below we can see the Dataset fig-35 the values tend to be below zero.

	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	F_LIT	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M
0	-0.771236	-0.815563	-0.561012	-0.507738	-0.958575	-0.957049	-0.423306	-0.476423	-0.798097	-0.733477	...	-0.163229	-0.720610	-0.156494
1	-0.823100	-0.874534	-0.681096	-0.725367	-0.958297	-0.956772	-0.582014	-0.607607	-0.849434	-0.779797	...	-0.583103	-0.732811	-0.282327
2	-1.000919	-0.981466	-0.976956	-0.965262	-0.958575	-0.956772	-0.038951	-0.027273	-0.956457	-0.807151	...	-0.859212	-0.921931	-0.456727
3	-1.052224	-1.041001	-1.022118	-0.995393	-0.958783	-0.957049	-0.355965	-0.390060	-1.004643	-0.858872	...	-0.805468	-0.900758	-0.419198
4	-0.809381	-0.813933	-0.622359	-0.649908	-0.957395	-0.955529	0.149238	0.043330	-0.800568	-0.705296	...	-0.348645	-0.297513	0.472670

Fig: -35: Dataset after applying Z-score

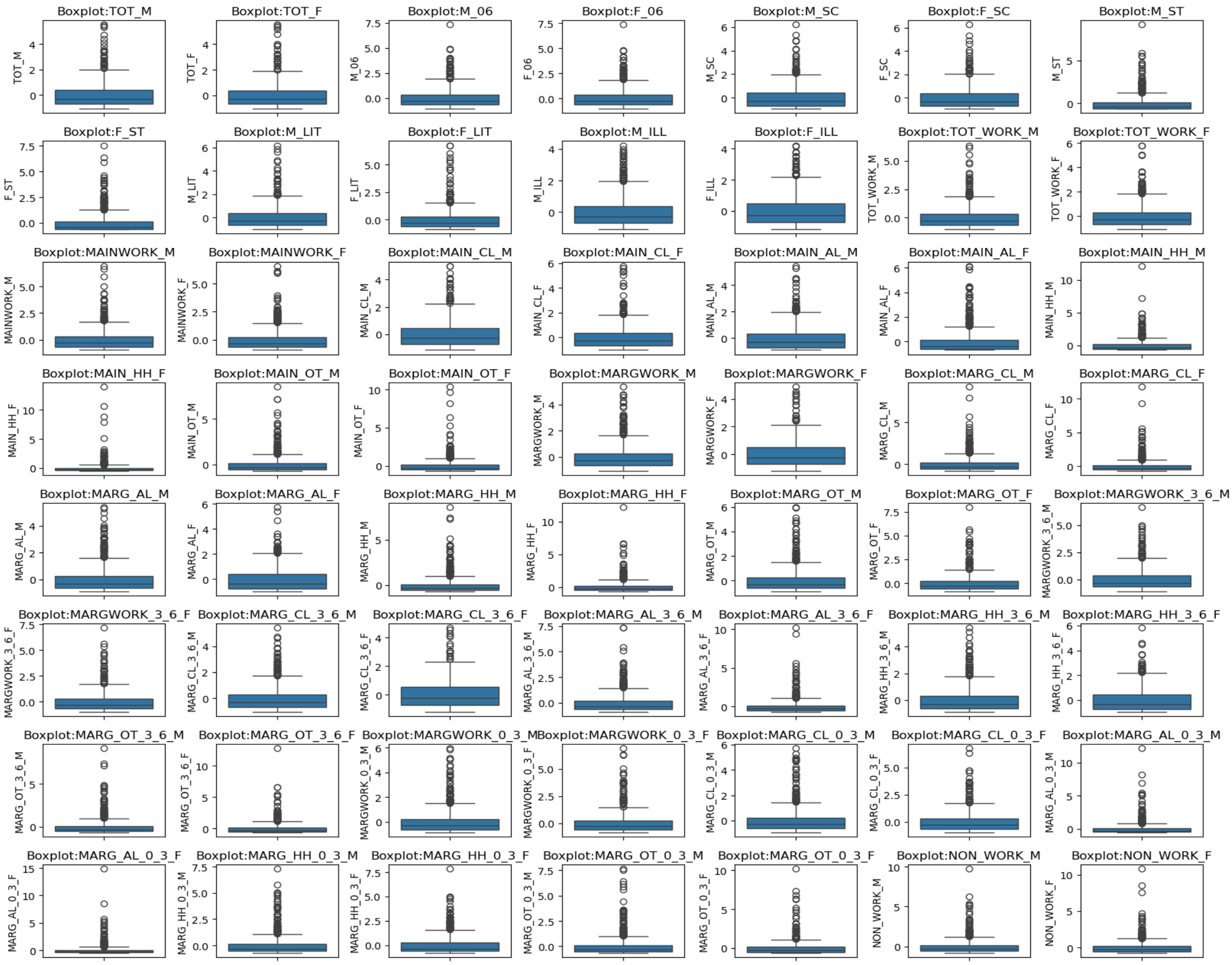


Fig: -36: Boxplot After Scaling

Apart from the scaling adjustment, there are absolutely no changes when we compare the box plot before and after scaling.

Q2.5. PCA: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.

Checking the Correlation.

Statistical test is done before performing PCA. Though we have seen few correlations in the dataset. The *Bartletts test of Sphericity* is performed to understand correlation significance in the population.

The Null hypothesis an alternative hypothesis will be defined.

H0: All variables in the data are uncorrelated
Ha: At least one pair of variables in the dataset are correlated.

We Decide the significance level Here we select
$\alpha = 0.05$. – we got **'0'**

The p-value is 0, which is less than the significant level. Therefore, the null hypothesis will be rejected. Hence, it proves that at least one pair of variables in the dataset is correlated, and PCA will be performed.

And the next step would KMO test: (Kaiser-Meyer-Olkin)

The KMO test will be conducted to measure the sample adequacy (MSA) of the dataset.

If MSA is less than 0.5, PCA will not be suggested. Alternatively, if it is greater than 0.7, it gives a substantial reduction in dimension and extracts significant components.

KMO is 0.80, Since it is greater 0.7 – the PCA is recommend.

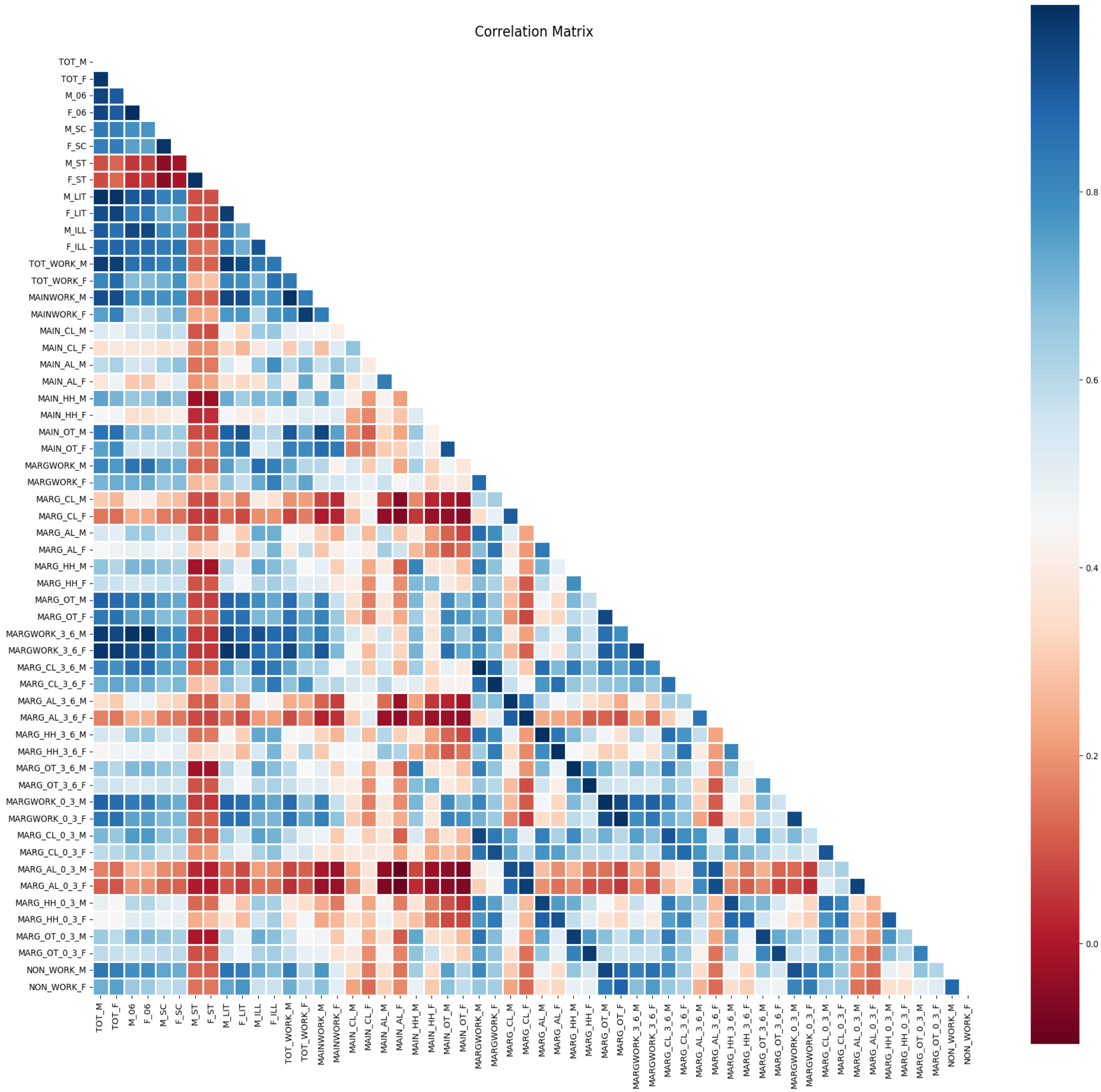


Fig: -37: Heatmap

Covariance Matrix:

```
array([[1.00156495, 0.98417823, 0.95231299, ..., 0.5891007 , 0.84621844,
        0.71718181],
       [0.98417823, 1.00156495, 0.90939623, ..., 0.572748  , 0.82894851,
        0.74775097],
       [0.95231299, 0.90939623, 1.00156495, ..., 0.56591416, 0.78618919,
        0.65216231],
       ...,
       [0.5891007 , 0.572748  , 0.56591416, ..., 1.00156495, 0.61052325,
        0.52191235],
       [0.84621844, 0.82894851, 0.78618919, ..., 0.61052325, 1.00156495,
        0.88228018],
       [0.71718181, 0.74775097, 0.65216231, ..., 0.52191235, 0.88228018,
        1.00156495]])
```

	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	F_LIT	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F
TOT_M	1.00	0.98	0.95	0.95	0.84	0.83	0.09	0.09	0.99	0.93	...	0.70	0.60	0.17	0.12
TOT_F	0.98	1.00	0.91	0.91	0.82	0.83	0.12	0.13	0.99	0.96	...	0.66	0.60	0.14	0.10
M_06	0.95	0.91	1.00	1.00	0.78	0.75	0.06	0.04	0.91	0.83	...	0.76	0.65	0.27	0.20
F_06	0.95	0.91	1.00	1.00	0.77	0.74	0.07	0.05	0.91	0.83	...	0.76	0.65	0.26	0.19
M_SC	0.84	0.82	0.78	0.77	1.00	0.99	-0.05	-0.05	0.82	0.72	...	0.67	0.57	0.18	0.13
F_SC	0.83	0.83	0.75	0.74	0.99	1.00	-0.01	-0.01	0.82	0.73	...	0.65	0.59	0.16	0.12
M_ST	0.09	0.12	0.06	0.07	-0.05	-0.01	1.00	0.99	0.09	0.10	...	0.12	0.20	0.03	0.01
F_ST	0.09	0.13	0.04	0.05	-0.05	-0.01	0.99	1.00	0.09	0.10	...	0.12	0.22	0.02	0.00
M_LIT	0.99	0.99	0.91	0.91	0.82	0.82	0.09	0.09	1.00	0.97	...	0.65	0.56	0.14	0.10
F_LIT	0.93	0.96	0.83	0.83	0.72	0.73	0.10	0.10	0.97	1.00	...	0.55	0.49	0.09	0.06
M_ILL	0.91	0.86	0.95	0.95	0.80	0.76	0.08	0.07	0.84	0.72	...	0.75	0.63	0.21	0.14
F_ILL	0.89	0.89	0.86	0.87	0.83	0.85	0.14	0.15	0.84	0.72	...	0.71	0.67	0.20	0.14
TOT_WORK_M	0.97	0.97	0.86	0.85	0.83	0.82	0.12	0.12	0.98	0.94	...	0.60	0.51	0.07	0.04
TOT_WORK_F	0.81	0.88	0.68	0.69	0.71	0.78	0.27	0.29	0.82	0.79	...	0.49	0.55	0.12	0.10
MAINWORK_M	0.93	0.94	0.79	0.79	0.78	0.78	0.11	0.11	0.95	0.93	...	0.47	0.39	-0.01	-0.03
MAINWORK_F	0.75	0.82	0.59	0.59	0.65	0.71	0.23	0.25	0.77	0.77	...	0.30	0.34	-0.03	-0.03
MAIN_CL_M	0.53	0.49	0.56	0.56	0.61	0.58	0.10	0.08	0.47	0.33	...	0.47	0.39	0.24	0.18

Fig: -38: Covariance matrix

Eigen Values:

The below fig 39 represents the Eigen value of all 12 principal components.

```
array([31.04602689,  7.74229066,  4.15338002,  3.6086627 ,  2.20641038,
        1.93824124,  1.15914355,  0.74854534,  0.6170419 ,  0.52808406,
        0.42978387,  0.35091506])
```

Fig: -39: Eigen value

Maximum variance is explained by PC1 = 31.81.

PC2 explains 7.86

PC3 explains 4.153

PC4 explains 3.66

PC5 explains 2.20 and PC6 explains 1.93.

The below figure shows the Eigen Vector of all the principal components. It derived by using Components

```
array([[ 1.68547147e-01,  1.66605242e-01,  1.64165243e-01,
         1.64562632e-01,  1.52996391e-01,  1.52940303e-01,
         2.74810453e-02,  2.83878596e-02,  1.63102824e-01,
         1.47366621e-01,  1.63910507e-01,  1.66998116e-01,
         1.60838746e-01,  1.46469761e-01,  1.46635940e-01,
         1.23739954e-01,  1.04870439e-01,  7.54411527e-02,
         1.14068301e-01,  7.34727501e-02,  1.33145711e-01,
         8.38711544e-02,  1.23482797e-01,  1.10599756e-01,
         1.67517738e-01,  1.57829907e-01,  8.50399103e-02,
         5.09848671e-02,  1.31481028e-01,  1.16285596e-01,
         1.43745975e-01,  1.29795621e-01,  1.56867596e-01,
         1.48361985e-01,  1.66756931e-01,  1.62454535e-01,
         1.68319155e-01,  1.57919009e-01,  9.58488238e-02,
         5.33359286e-02,  1.31385264e-01,  1.12415081e-01,
         1.42433411e-01,  1.26569034e-01,  1.55867751e-01,
         1.47332261e-01,  1.53177967e-01,  1.42857457e-01,
         5.45496809e-02,  4.34724137e-02,  1.24863483e-01,
         1.18519203e-01,  1.42828454e-01,  1.34541425e-01,
         1.52032961e-01,  1.32101728e-01],
```

Fig: -40: Eigen vector

Q2.6 PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.

The Explained variance ratio explain the proposition of variance of the principal components.
Let's check out the percentage of variance explained by each PC that is variance explained by an individual principal component divided by total variance explained by all the PC's.

In other words – Percentage of explained variance= Eigen value of each PC/sum of Eigen values of all PCs

```
array([0.5535271 , 0.13803917, 0.07405161, 0.06433972, 0.03933862,
       0.03455737, 0.02066665, 0.013346  , 0.01100139, 0.00941534,
       0.00766272, 0.00625655])
```

Fig: -41: Explained variance

55% of total variance is explained by PC1. 13.7% of total variance is explained by PC2.

7.2% of total variance is explained by PC3. 6.4% of total variance is explained by PC4.

3.8% of total variance is explained by PC5. 3.4% of total variance is explained by PC6.

Cumulative explained variance ratio:

The cumulative explained variance ratio to find a cut off for selecting the number of Principal components. (PCs)

```
array([0.5535271 , 0.69156626, 0.76561788, 0.8299576 , 0.86929622,
       0.90385359, 0.92452024, 0.93786623, 0.94886762, 0.95828296,
       0.96594568, 0.97220223])
```

Fig: -42: Cumulative Explained variance ratio

We can see from the above fig.42 that Cumulative explained variance ratio of 6 pcs is more than 90%. Therefore, we can conclude by saying the optimum number of PCs is 6 and the below fig 43 Scree plot supports the same.

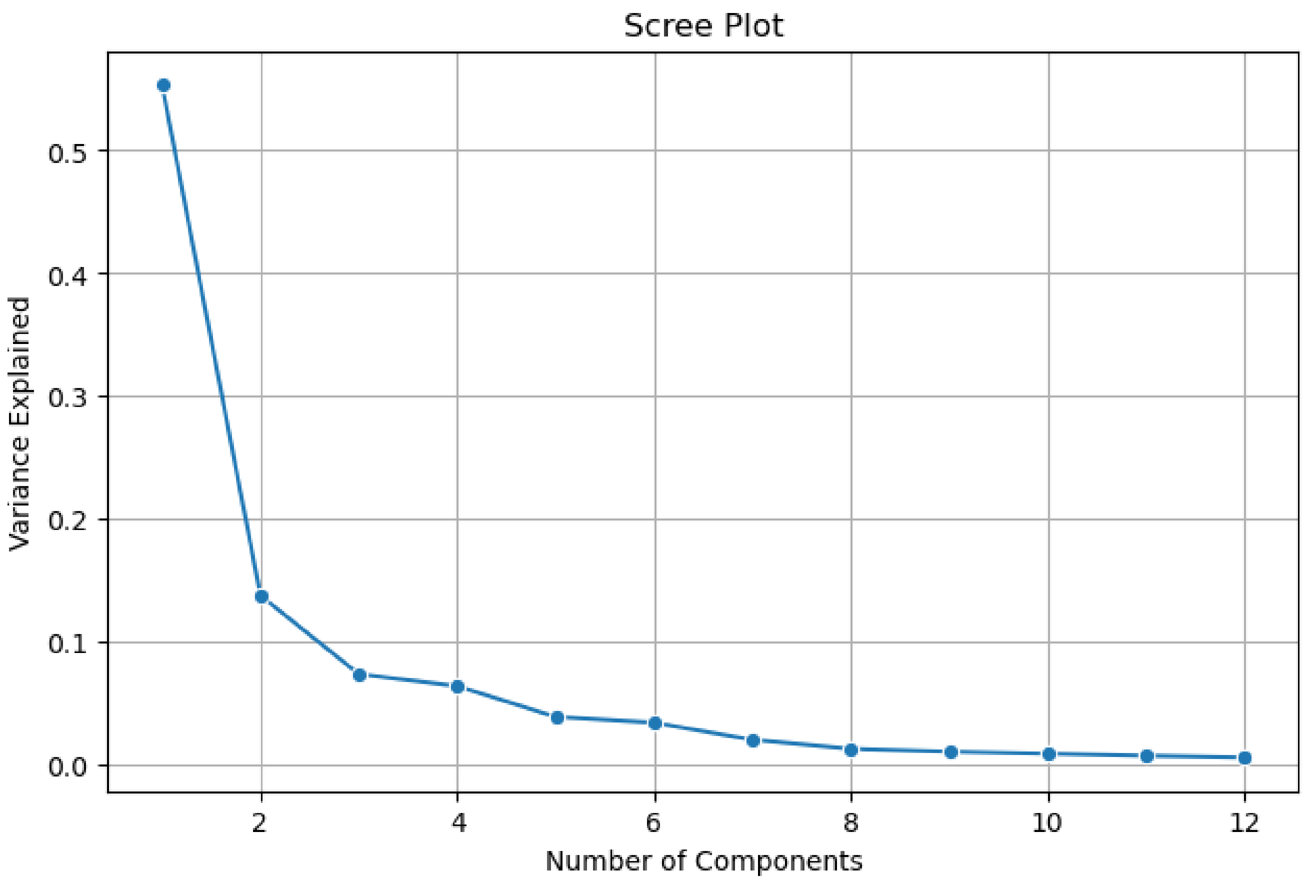


Fig: -42: Scree plot

The dots on the Scree plots are the 12 principal components. We can see a formation of the elbow at 6 PCs. It is evident that post 6 PCs, the drop is not that significant. therefore, the optimum number of PCs is 6.

Q2.7 PCA: Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.

To categorize the pattern, the components are loaded against each feature in a new data frame. we have 12 principal components and one co-efficient each for all 56 variables.

	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	F_LIT	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M
PC1	0.168547	0.166605	0.164165	0.164563	0.152996	0.152940	0.027481	0.028388	0.163103	0.147367	...	0.153178	0.142857	0.054550
PC2	-0.094738	-0.108873	-0.027200	-0.025412	-0.049895	-0.055754	0.029232	0.032066	-0.120102	-0.157002	...	0.146628	0.179272	0.253535
PC3	0.056601	0.038782	0.057654	0.049985	0.002433	-0.025125	-0.123022	-0.139260	0.082100	0.117085	...	0.054719	0.024117	0.268735
PC4	-0.027099	-0.077830	0.006567	0.009531	0.007256	-0.034768	-0.227647	-0.234733	-0.042616	-0.066948	...	0.089502	-0.018582	-0.097636
PC5	-0.033485	-0.013307	-0.050635	-0.044227	-0.173308	-0.160092	0.432293	0.437965	-0.009562	0.055421	...	0.081417	0.130170	-0.048741

Fig: -43: Selected components

To analysis the variable that has the highest loading among the principal components. The component has to be loaded on a heatmap. For each variable with maximum loading, the heatmap shows a blue rectangle box that is marked across the components. Ref Fig. 44

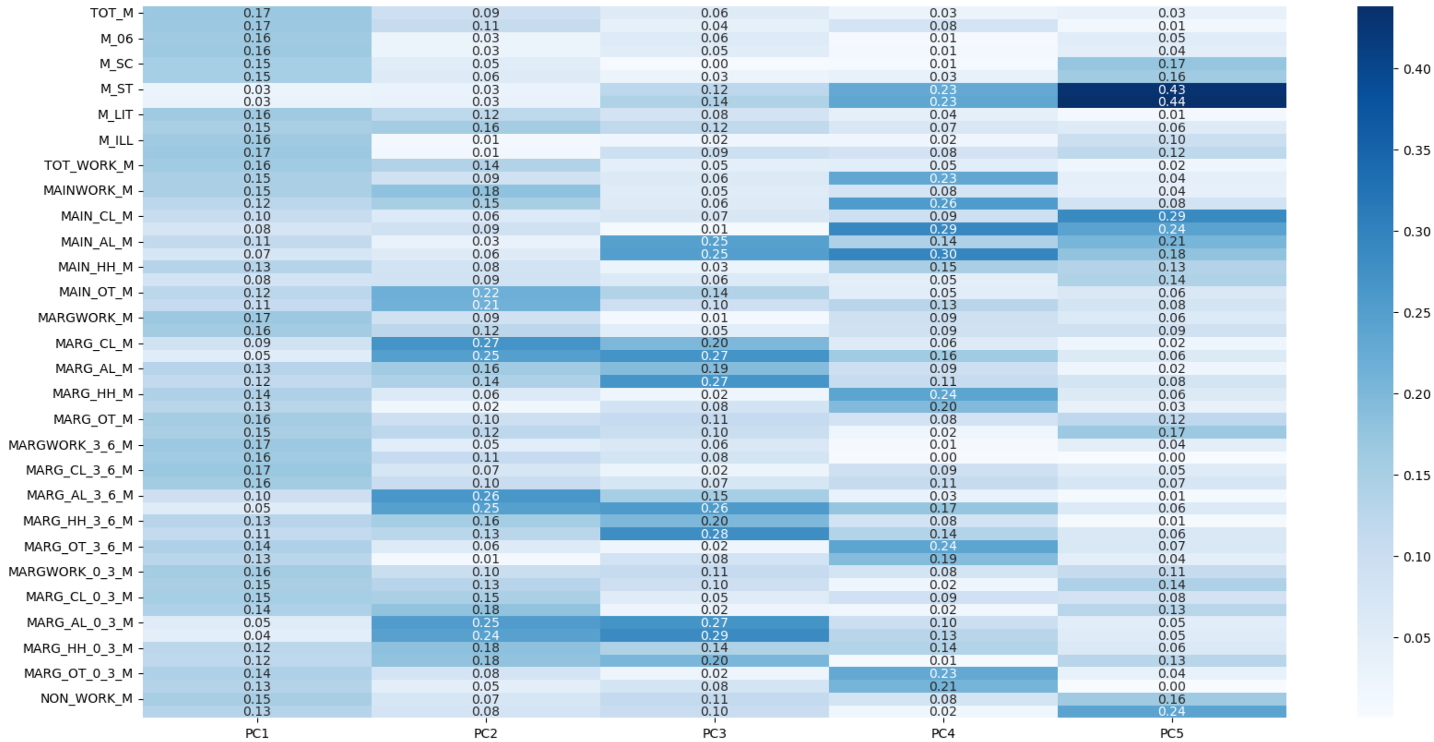


Fig: -44: Heatmap

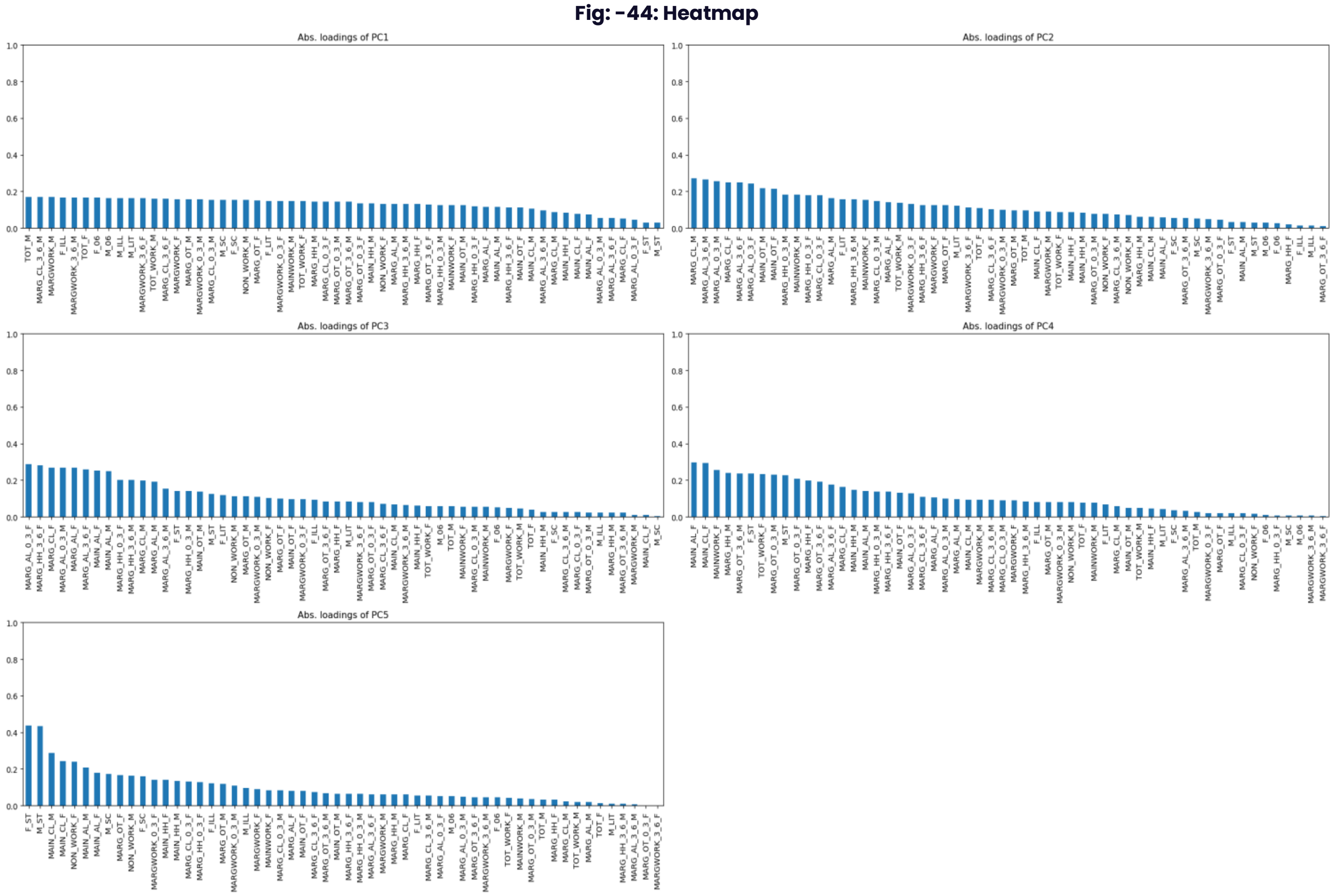


Fig: -45: Comparison

The above table shows the variables that contributes maximum towards the respective PCs:

Q2.8 PCA: Write linear equation for first PC.

The below fig depicts the linear equation for the first PC (PC1):
The value in the Parentheses is the coefficient and those multiplied by variables.

```
PC1 = 0.355 * x1 + 0.554 * x2 + 0.052 * x3 + 0.051 * x4 + 0.058 * x5 + 0.088 * x6 + 0.005 * x7 + 0.009 * x8 + 0.271 * x9 + 0.352 * x10 + 0.084 * x11 + 0.
201 * x12 + 0.174 * x13 + 0.155 * x14 + 0.146 * x15 + 0.117 * x16 + 0.011 * x17 + 0.009 * x18 + 0.019 * x19 + 0.027 * x20 + 0.004 * x21 + 0.007 * x22 +
0.111 * x23 + 0.074 * x24 + 0.028 * x25 + 0.038 * x26 + 0.002 * x27 + 0.002 * x28 + 0.009 * x29 + 0.015 * x30 + 0.001 * x31 + 0.003 * x32 + 0.016 * x33 +
0.017 * x34 + 0.181 * x35 + 0.399 * x36 + 0.023 * x37 + 0.030 * x38 + 0.001 * x39 + 0.002 * x40 + 0.007 * x41 + 0.011 * x42 + 0.001 * x43 + 0.002 * x44 +
0.013 * x45 + 0.014 * x46 + 0.005 * x47 + 0.008 * x48 + 0.000 * x49 + 0.001 * x50 + 0.002 * x51 + 0.003 * x52 + 0.000 * x53 + 0.001 * x54 + 0.003 * x55 +
0.003 * x56
```