BHANU PRATAP REDDY

CLUSTERING AND PCA
GRADED PROJECT

JANUARY 13 **2022**

TABLE OF CONTENTS

TABLE OF FIGURES

CLUSTERING

PROBLEM STATEMENT

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

CPC = Total Cost (spend) / Number of Clicks

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

Perform the following in given order:

- Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.
- Treat missing values in CPC, CTR and CPM using the formula given. You may refer to the Bank_KMeans Case Study to understand the coding behind treating the missing values using a specific formula. You have to basically create an user defined function and then call the function for imputing.
- 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).
- Perform z-score scaling and discuss how it affects the speed of the algorithm.
- Perform clustering and do the following:
- Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.
- Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.
- Print silhouette scores for up to 10 clusters and identify optimum number of clusters.
- Profile the ads based on optimum number of clusters using silhouette score and your domain understanding

DATA DICTIONARY

The data dictionary gives us information regarding the variables present tin the given dataset we are analyzing and a brief explanation and what that variable contains or means

- **Timestamp:** Date on which the ad was run
- **InventoryType:** Ad Format Label
- **Ad - Length:** Length of the Ad Block
- **Ad- Width:** Width of the ad block.
- **Ad Size:** Total pixel area of the ad block
- **Ad Type:** Ad Type Label
- **Platform:** The platform on which the particular Ad is being run on.
- **Device Type:** The targeted device type user.
- **Format:** Format of the run ad
- **Available_Impressions:** Max Potential number of impressions
- **Matched_Queries:** Exact keyword Matches
- **Impressions:** Actual recorded Impressions
- **Clicks:** Clicks on Particular Ad
- **Fee:** Fee on the Ad
- **Revenue:** Revenue Generated by the Ad
- **CTR:** CTR stands for click-through rate: a metric that measures the number of clicks advertisers receive on their ads per number of impressions.
- **CPM:** CPM stands for "cost per 1000 impressions." Advertisers running CPM ads set their desired price per 1000 ads served and pay each time their ad appears.
- **CPC:** Cost-per-click (CPC) bidding means that you pay for each click on your ads

DATA INTIALIZATION AND PREPROCESSING

Data was initialized from given excel file to a data frame and head and tail values of said data frame was obtained to validate this:

	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	2020-9-2-18	Format1	300	250	75000	Inter223	Web	Mobile	Display	1979	384	380	0	0.0	0.35	0.0	0.0000	0.0	NaN
2	2020-9-3-16	Format6	336	250	84000	Inter217	Web	Desktop	Video	1566	298	297	0	0.0	0.35	0.0	0.0000	0.0	NaN
3	2020-9-3-2	Format1	300	250	75000	Inter224	Web	Desktop	Display	643	103	102	0	0.0	0.35	0.0	0.0000	0.0	NaN
4	2020-9-3-13	Format1	300	250	75000	Inter225	Video	Mobile	Display	1550	347	345	0	0.0	0.35	0.0	0.0000	0.0	NaN

	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
25852	2020-10-1-5	Format5	720	300	216000	Inter222	Video	Desktop	Video	1	1	1	0	0.01	0.35	0.0065	NaN	NaN	NaN
25853	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
25854	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
25855	2020-9-30-4	Format7	300	600	180000	Inter228	Video	Mobile	Display	1	1	1	0	0.01	0.35	0.0065	NaN	NaN	NaN
25856	2020-10-17-3	Format5	720	300	216000	Inter225	Video	Mobile	Display	1	1	1	0	0.01	0.35	0.0065	NaN	NaN	NaN

Figure 1

Five head and tail values of dataframe

The size of the data frame was found to be of 25857 rows and 19 columns as seen along with some info regarding the data frame:

```
RangeIndex: 25857 entries, 0 to 25856
Data columns (total 19 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Timestamp              25857 non-null  object
 1   InventoryType          25857 non-null  object
 2   Ad - Length            25857 non-null  int64
 3   Ad- Width              25857 non-null  int64
 4   Ad Size                25857 non-null  int64
 5   Ad Type                25857 non-null  object
 6   Platform               25857 non-null  object
 7   Device Type            25857 non-null  object
 8   Format                 25857 non-null  object
 9   Available_Impressions  25857 non-null  int64
 10  Matched_Queries        25857 non-null  int64
 11  Impressions            25857 non-null  int64
 12  Clicks                 25857 non-null  int64
 13  Spend                  25857 non-null  float64
 14  Fee                    25857 non-null  float64
 15  Revenue                25857 non-null  float64
 16  CTR                    19392 non-null  float64
 17  CPM                    19392 non-null  float64
 18  CPC                    18330 non-null  float64
dtypes: float64(6), int64(7), object(6)
```


Figure 2

Info regarding dataframe

As evidenced from the information seen above it appears that some of the columns i.e., CTR, CPM and CPC have missing values in them to verify we check for null values in all columns and the result is given as:

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	6465
CPM	6465
CPC	7527

Figure 3
Null Values in each of the columns

It appears we will have to treat the missing values in CTR, CPM and CPC but before we do that let us have a look at the data summary before any form of preprocessing.

	count	mean	std	min	25%	50%	75%	max
Ad - Length	25857.0	3.904312e+02	2.306961e+02	120.00	120.0000	300.0000	7.200000e+02	728.00
Ad- Width	25857.0	3.321828e+02	1.942609e+02	70.00	250.0000	300.0000	6.000000e+02	600.00
Ad Size	25857.0	9.968328e+04	6.264069e+04	33600.00	72000.0000	75000.0000	8.400000e+04	216000.00
Available_Impressions	25857.0	2.169621e+06	4.542680e+06	0.00	9133.0000	330968.0000	2.208484e+06	27592861.00
Matched_Queries	25857.0	1.155322e+06	2.407244e+06	0.00	5451.0000	189449.0000	1.008171e+06	14702025.00
Impressions	25857.0	1.107525e+06	2.326648e+06	0.00	2558.0000	162162.0000	9.496930e+05	14194774.00
Clicks	25857.0	9.525881e+03	1.672169e+04	0.00	305.0000	3457.0000	1.068100e+04	143049.00
Spend	25857.0	2.414473e+03	3.932835e+03	0.00	36.0300	1173.6600	2.692280e+03	26931.87
Fee	25857.0	3.367289e-01	3.053978e-02	0.21	0.3500	0.3500	3.500000e-01	0.35
Revenue	25857.0	1.716549e+03	2.993025e+03	0.00	23.4200	762.8800	1.749982e+03	21276.18
CTR	19392.0	6.962653e-02	7.497012e-02	0.00	0.0024	0.0077	1.283000e-01	1.00
CPM	19392.0	7.252900e+00	6.538314e+00	0.00	1.6300	3.0350	1.222000e+01	81.56
CPC	18330.0	3.510606e-01	3.433338e-01	0.00	0.0900	0.1600	5.700000e-01	7.26

Figure 4
Data Summary Before preprocessing

It also appears that **there is no duplicated data**, so proceeding with the preprocessing of the data.
CTR, CPM and CPC are calculated from the data in campaign spend, Impressions and Clicks and since one of those have any missing values, we can just impute CTR, CPM and CPC by just calculating them from their formulas which are given as:

$$CPM = \frac{Total\ Campaign\ Spend(Spend)}{Number\ of\ Impressions(Impressions)}\ x\ 1000$$

$$CPC = \frac{Total\ Cost(Spend)}{Number\ of\ Clicks(Clicks)}$$

$$CTR = \frac{Total\ Measured\ Clicks(Clicks)}{Total\ Measured\ ad\ Impressions(Impressions)}\ x\ 100$$

Therefore, the data info and summary after imputations Is given as:

```
RangeIndex: 25857 entries, 0 to 25856
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Timestamp             25857 non-null  object
 1   InventoryType         25857 non-null  object
 2   Ad - Length           25857 non-null  int64
 3   Ad- Width             25857 non-null  int64
 4   Ad Size               25857 non-null  int64
 5   Ad Type               25857 non-null  object
 6   Platform              25857 non-null  object
 7   Device Type           25857 non-null  object
 8   Format                25857 non-null  object
 9   Available_Impressions 25857 non-null  int64
 10  Matched_Queries       25857 non-null  int64
 11  Impressions           25857 non-null  int64
 12  Clicks                25857 non-null  int64
 13  Spend                 25857 non-null  float64
 14  Fee                   25857 non-null  float64
 15  Revenue               25857 non-null  float64
 16  CTR                   25857 non-null  float64
 17  CPM                   25857 non-null  float64
 18  CPC                   25857 non-null  float64
dtypes: float64(6), int64(7), object(6)
```

Figure 5
Data Info after imputations

	count	mean	std	min	25%	50%	75%	max
Ad - Length	25857.0	390.43	230.70	120.00	120.00	300.00	720.00	728.00
Ad- Width	25857.0	332.18	194.26	70.00	250.00	300.00	600.00	600.00
Ad Size	25857.0	99683.28	62640.69	33600.00	72000.00	75000.00	84000.00	216000.00
Available_Impressions	25857.0	2169620.83	4542680.20	0.00	9133.00	330968.00	2208484.00	27592861.00
Matched_Queries	25857.0	1155321.80	2407243.93	0.00	5451.00	189449.00	1008171.00	14702025.00
Impressions	25857.0	1107525.30	2326647.65	0.00	2558.00	162162.00	949693.00	14194774.00
Clicks	25857.0	9525.88	16721.69	0.00	305.00	3457.00	10681.00	143049.00
Spend	25857.0	2414.47	3932.84	0.00	36.03	1173.66	2692.28	26931.87
Fee	25857.0	0.34	0.03	0.21	0.35	0.35	0.35	0.35
Revenue	25857.0	1716.55	2993.03	0.00	23.42	762.88	1749.98	21276.18
CTR	25857.0	0.07	0.09	0.00	0.00	0.01	0.13	2.00
CPM	25857.0	7.52	8.93	0.00	1.57	2.96	12.45	715.00
CPC	25857.0	0.30	0.34	0.00	0.08	0.12	0.50	7.26

Figure 6
Data Summary after imputations

As expected, there is a significant change in the data summary, let us have look at how this affects further processing.

OUTLIER IDENTIFICATION AND TREATMENT

Let us plot boxplots for numeric to see whether or not there are any outliers in our data.

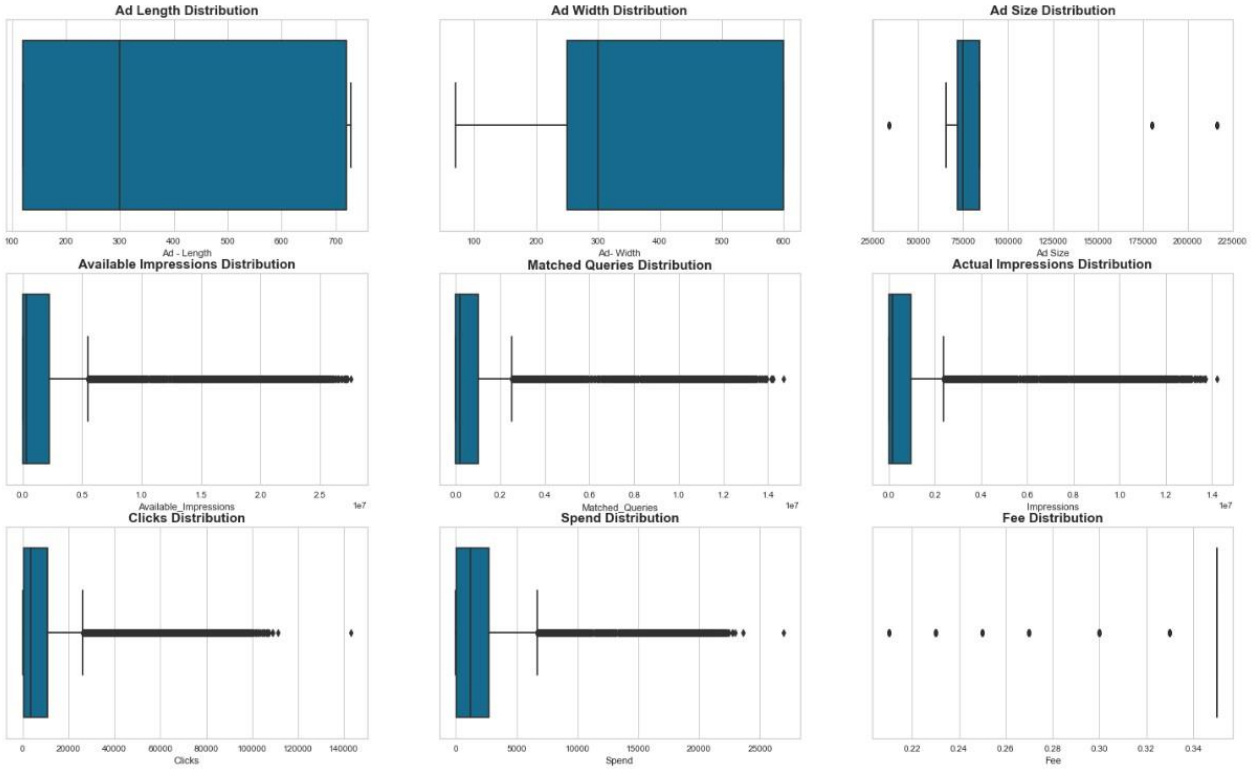

Figure 7

Boxplots before outlier Treatment - I

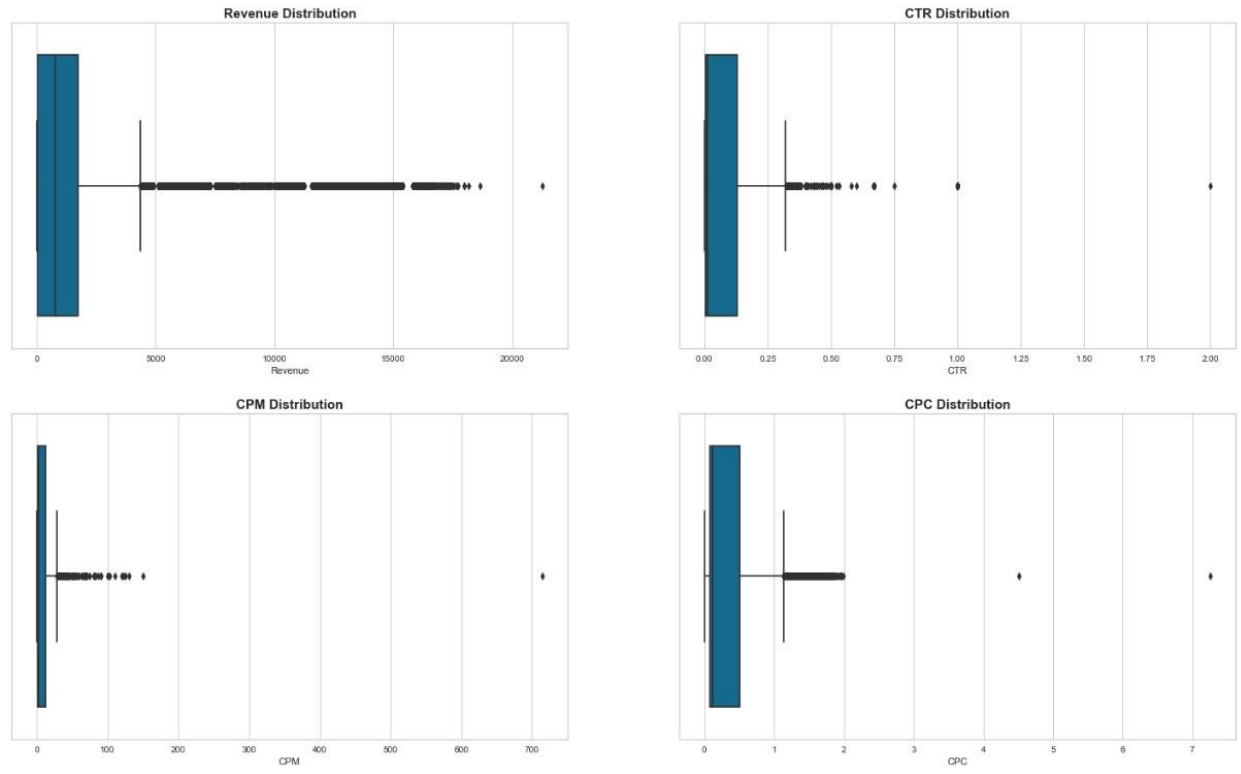

Figure 8

Boxplot before Outlier Treatment - II

As its clearly evident from the box plots it appears most of the numeric data contain outliers.

Should we treat the outliers?

It totally depends on the clustering method used, in the instance of K-Means Clustering the squared error algorithm is sensitive to outliers and it is better to get rid of them.

But there are other clustering algorithms which can handle outliers such as DBSCAN for which outlier treatment is not required.

But for our purposes we will proceed with removing outliers by moving the outliers to its closest quartile.

The result after removing outliers:

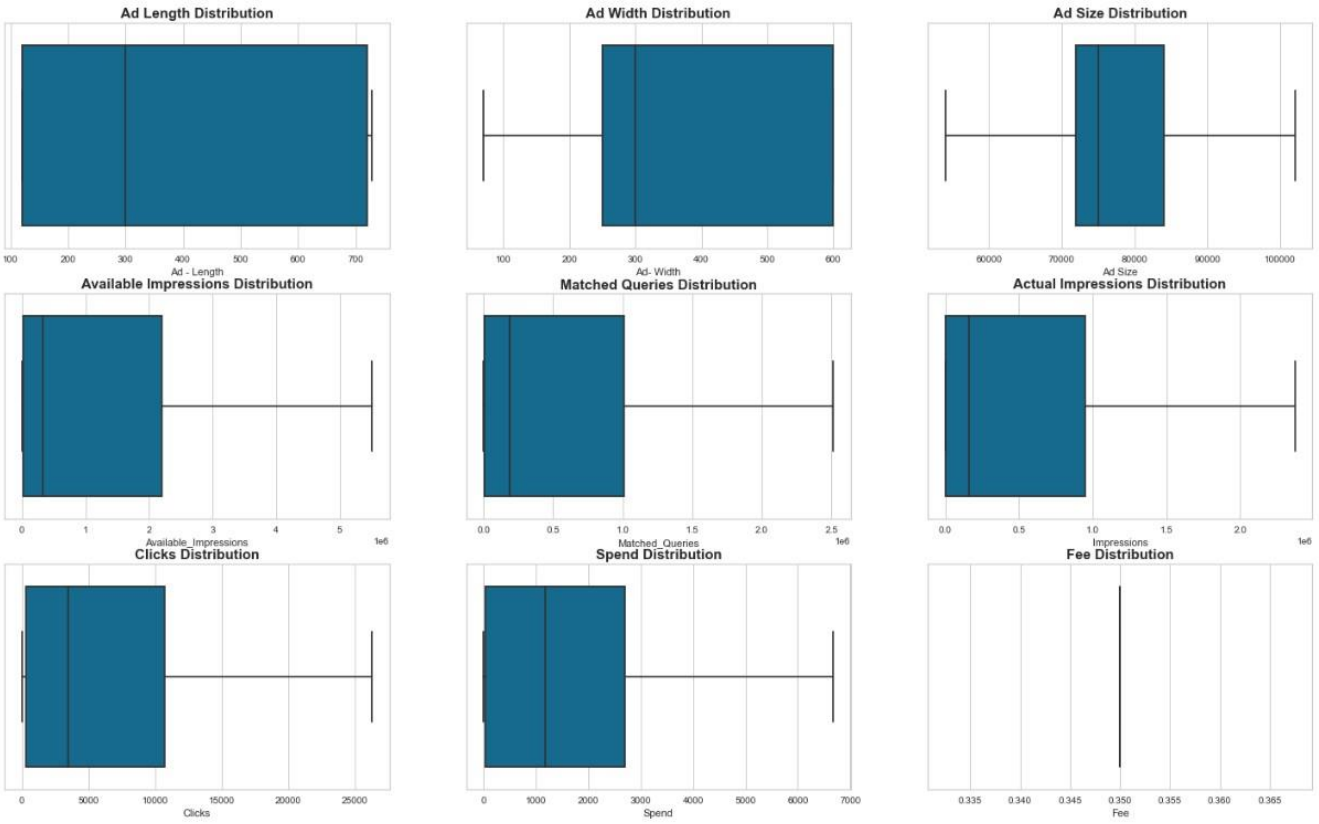

Figure 9
Boxplots after outlier Treatment - I

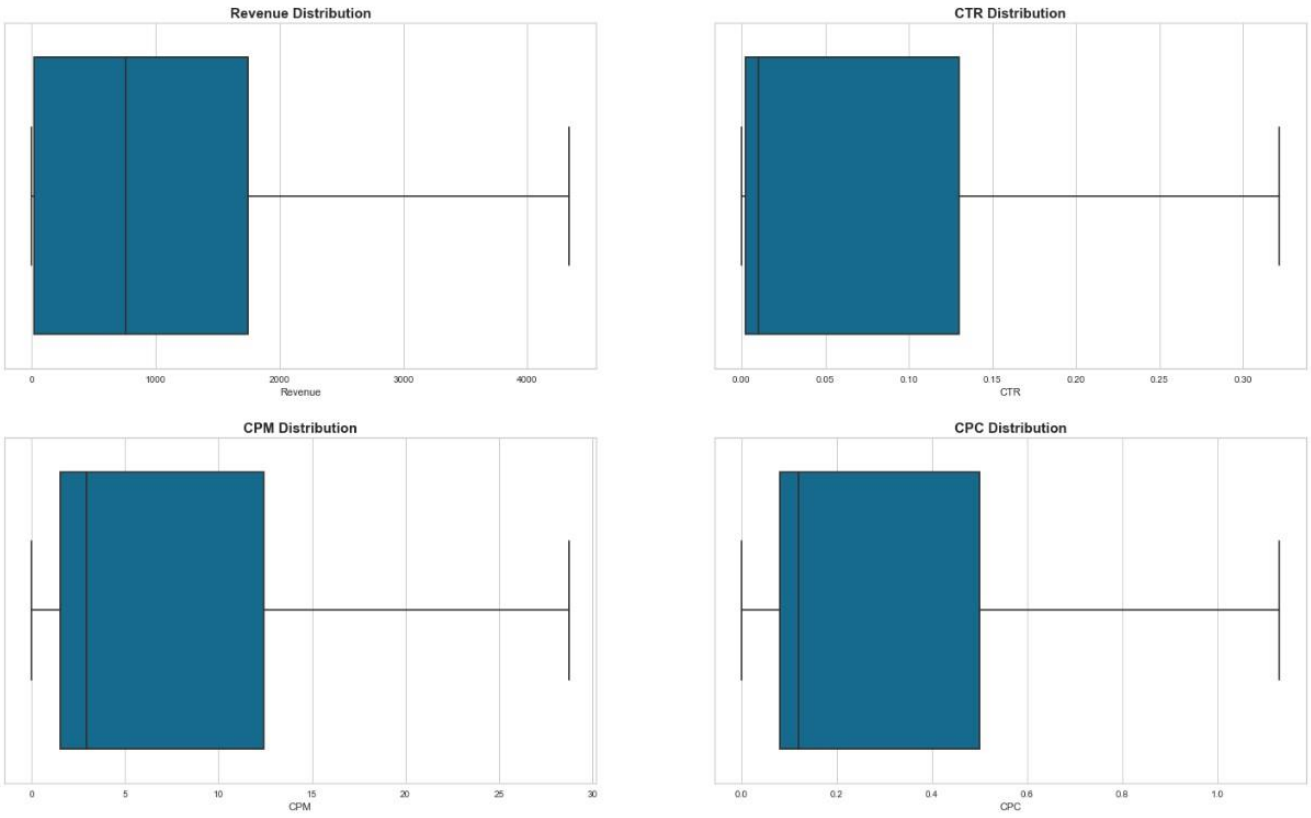

Figure 10

Boxplots after outlier Treatment - II

The data summary after outlier treatment is also given as:

	count	mean	std	min	25%	50%	75%	max
Ad - Length	25857.0	390.43	230.70	120.00	120.00	300.00	720.00	728.00
Ad- Width	25857.0	332.18	194.26	70.00	250.00	300.00	600.00	600.00
Ad Size	25857.0	77484.14	15352.88	54000.00	72000.00	75000.00	84000.00	102000.00
Available_Impressions	25857.0	1357917.55	1896821.01	0.00	9133.00	330968.00	2208484.00	5507510.50
Matched_Queries	25857.0	659959.32	885651.39	0.00	5451.00	189449.00	1008171.00	2512251.00
Impressions	25857.0	619864.63	842872.33	0.00	2558.00	162162.00	949693.00	2370395.50
Clicks	25857.0	7070.83	8602.09	0.00	305.00	3457.00	10681.00	26245.00
Spend	25857.0	1844.35	2184.31	0.00	36.03	1173.66	2692.28	6676.66
Fee	25857.0	0.35	0.00	0.35	0.35	0.35	0.35	0.35
Revenue	25857.0	1216.79	1446.34	0.00	23.42	762.88	1749.98	4339.83
CTR	25857.0	0.07	0.08	0.00	0.00	0.01	0.13	0.32
CPM	25857.0	7.36	6.93	0.00	1.57	2.96	12.45	28.77
CPC	25857.0	0.29	0.31	0.00	0.08	0.12	0.50	1.13

Figure 11

Data summary after outlier treatment.

SCALING

From the data summary we can see that the ranges of our numeric variables vary quite a bit and therefore we must scale the numeric data so that equal weights can be applied to the data. We do this mainly for two reasons:

- To generate good quality clusters and improve accuracy of the clustering algorithm by reducing any distortion.

- Improve the speed of the algorithm (in gradient bases algorithms like K Means).

How does this improve speed?

If different components of data have different scales, then the derivatives will always tend to align along the directions of higher variance, which in turn leads to slower convergence. Hence scaling the data to have equal weights will eliminate this and make the process faster.

Therefore, the data summary after scaling using z-score can be given as:

	count	mean	std	min	25%	50%	75%	max
Ad - Length	25857.0	1.361843e-14	1.000019	-1.172263	-1.172263	-0.392000	1.428612	1.463290
Ad- Width	25857.0	1.313640e-14	1.000019	-1.349668	-0.423062	-0.165671	1.378674	1.378674
Ad Size	25857.0	9.961410e-15	1.000019	-1.529654	-0.357213	-0.161806	0.424415	1.596856
Available_Impressions	25857.0	-2.160377e-15	1.000019	-0.715905	-0.711090	-0.541416	0.448425	2.187699
Matched_Queries	25857.0	-5.630549e-15	1.000019	-0.745183	-0.739028	-0.531269	0.393178	2.091486
Impressions	25857.0	8.653831e-15	1.000019	-0.735434	-0.732399	-0.543038	0.391322	2.076904
Clicks	25857.0	1.118393e-14	1.000019	-0.822006	-0.786548	-0.420119	0.419693	2.229056
Spend	25857.0	-7.677941e-15	1.000019	-0.844382	-0.827887	-0.307057	0.388198	2.212324
Fee	25857.0	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Revenue	25857.0	4.202798e-15	1.000019	-0.841307	-0.825114	-0.313841	0.368655	2.159308
CTR	25857.0	-5.212531e-15	1.000019	-0.894571	-0.867659	-0.772243	0.695691	3.040717
CPM	25857.0	-3.680683e-15	1.000019	-1.061370	-0.834843	-0.634287	0.734978	3.089709
CPC	25857.0	-1.745278e-14	1.000019	-0.945073	-0.685985	-0.556441	0.674227	2.714545

Figure 12
Data Summary after scaling

HIERARCHICAL CLUSTERING

We perform hierarchical clustering on our scaled numeric data, the dendrogram for the hierarchical clustering for the last 10 truncations can be seen below:

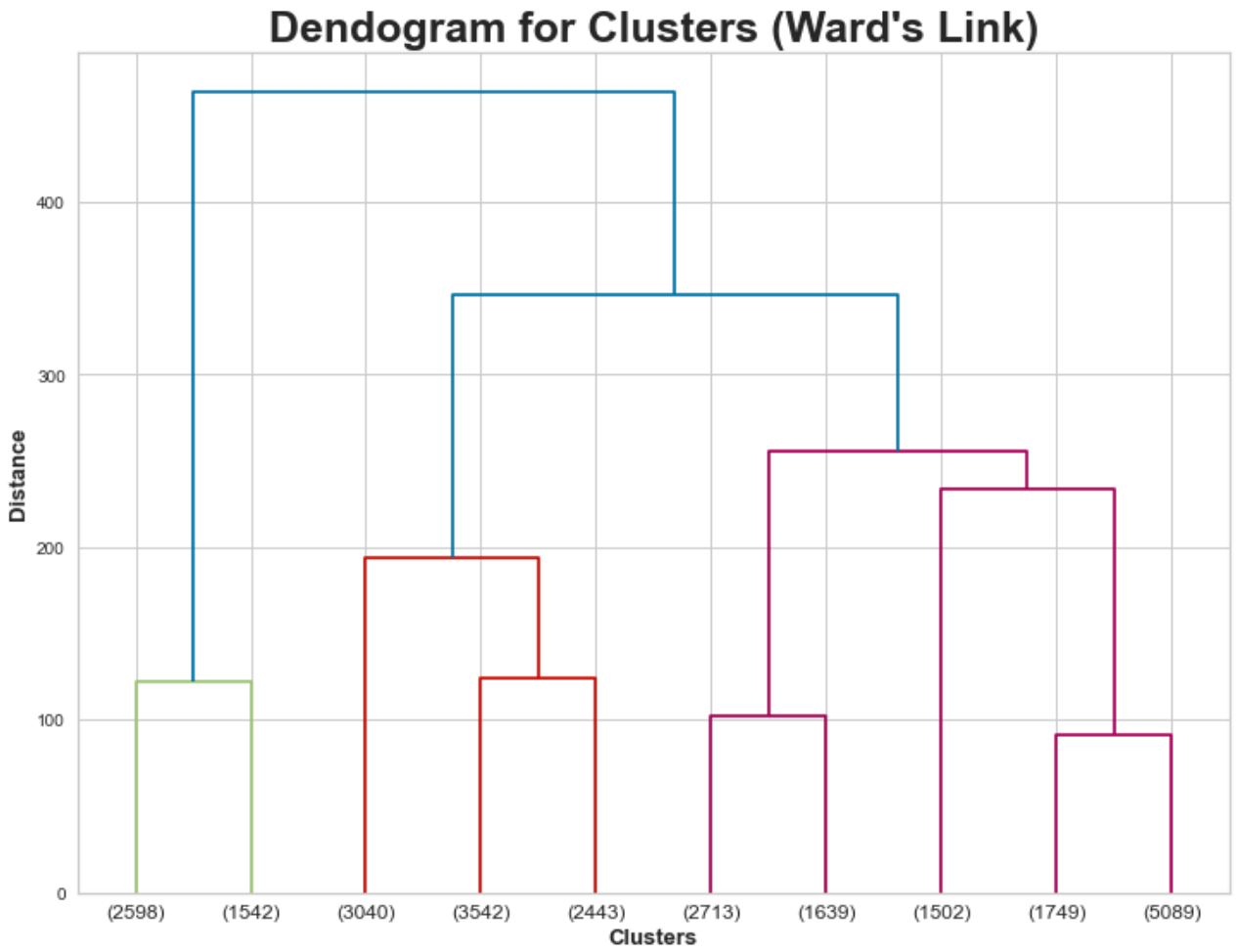

Figure 13
Dendrogram for at least 10 truncations

The dendrogram for only the last 10 truncation s was taken mainly because the entire graph would look messy and unreadable hence only the last 10 were taken.

WSS PLOT AND SILHOUETTE SCORES

The WSS plot or the elbow plot is usually utilized to determine the optimum number of clusters for a given data set using k-means algorithm. Where the steepest fall or the highest differential between consecutive K values will point towards an optimum number of clusters.

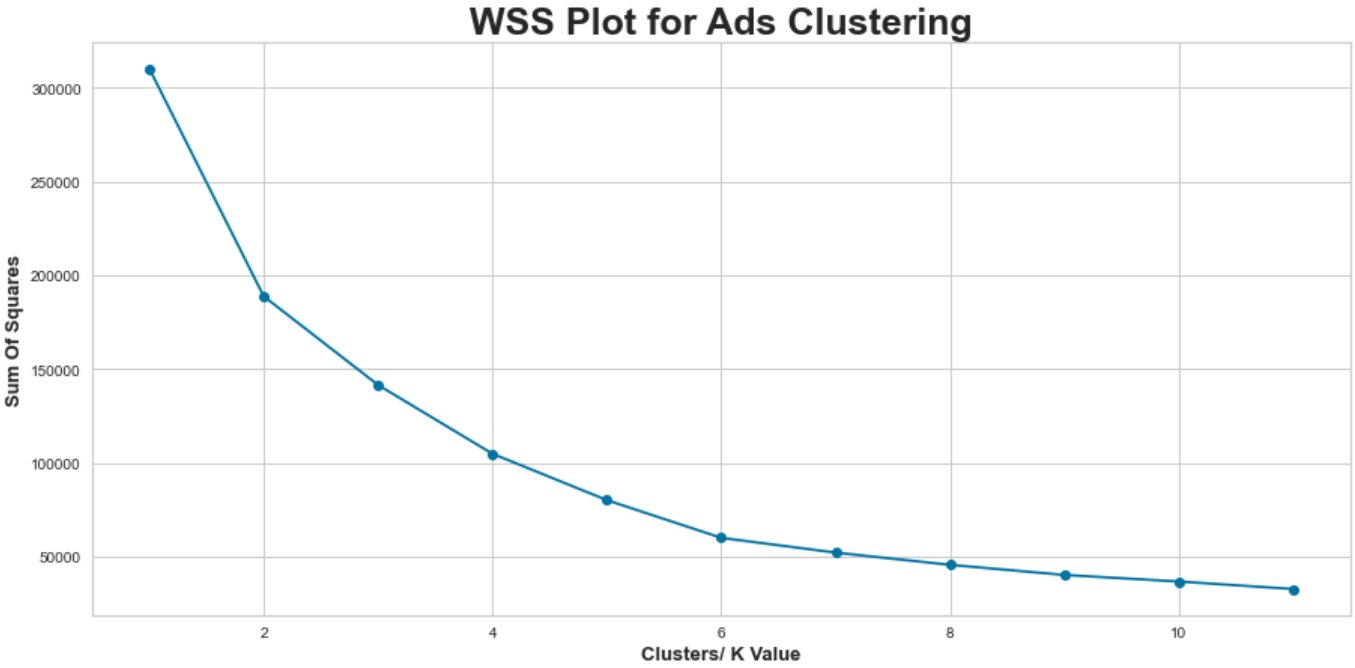

Figure 14
Elbow plot for clustering

The optimal number of clusters is determined by selecting the value of k at the “elbow” i.e., the point after which the distortion/inertia start decreasing in a linear fashion.

Looking at the graph the optimal point appears to be **k= 6.** Therefore, the optimal number of clusters from the elbow plot appears to be 6 and we can verify this further and more accurately using silhouette scores.

The silhouette scores for all values of K can be given as:

```
Silhouette Score for k =  2 is 0.37880013646163846
Silhouette Score for k =  3 is 0.3466496747547733
Silhouette Score for k =  4 is 0.4222964796199141
Silhouette Score for k =  5 is 0.4801869379109477
Silhouette Score for k =  6 is 0.48793303029582497
Silhouette Score for k =  7 is 0.4897584186020886
Silhouette Score for k =  8 is 0.5102494711675123
Silhouette Score for k =  9 is 0.5009065321401819
Silhouette Score for k =  10 is 0.5218591438221627
Silhouette Score for k =  11 is 0.4922442726730517
```

Figure 15
Silhouette score for k = 2 up to 11

The optimal number of clusters from silhouette score is determined as a +ve value that is closest to 1. Therefore, in this case the optimal value would be bet ween k=6 to 8 given marginal difference between then therefore a better way would be to visualize this using a silhouette visualizer with respect to the data.

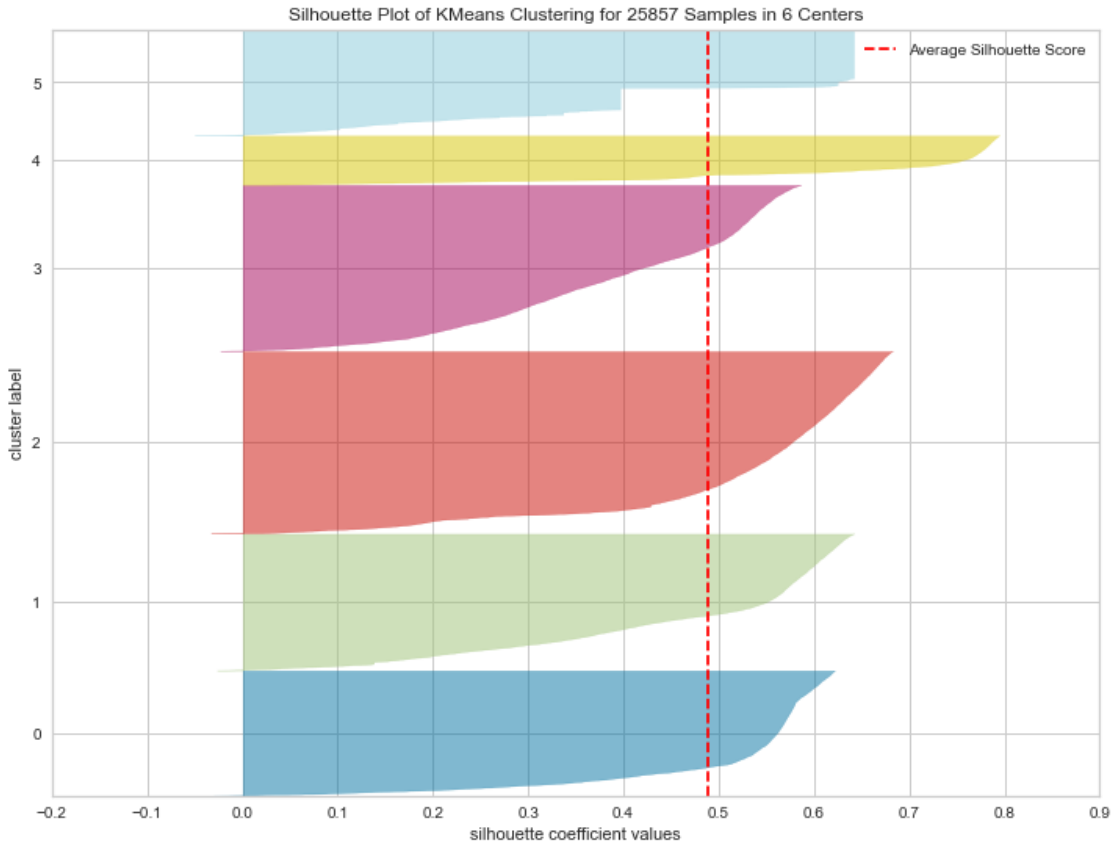

Figure 16
Silhouette Visualizer for k = 6.

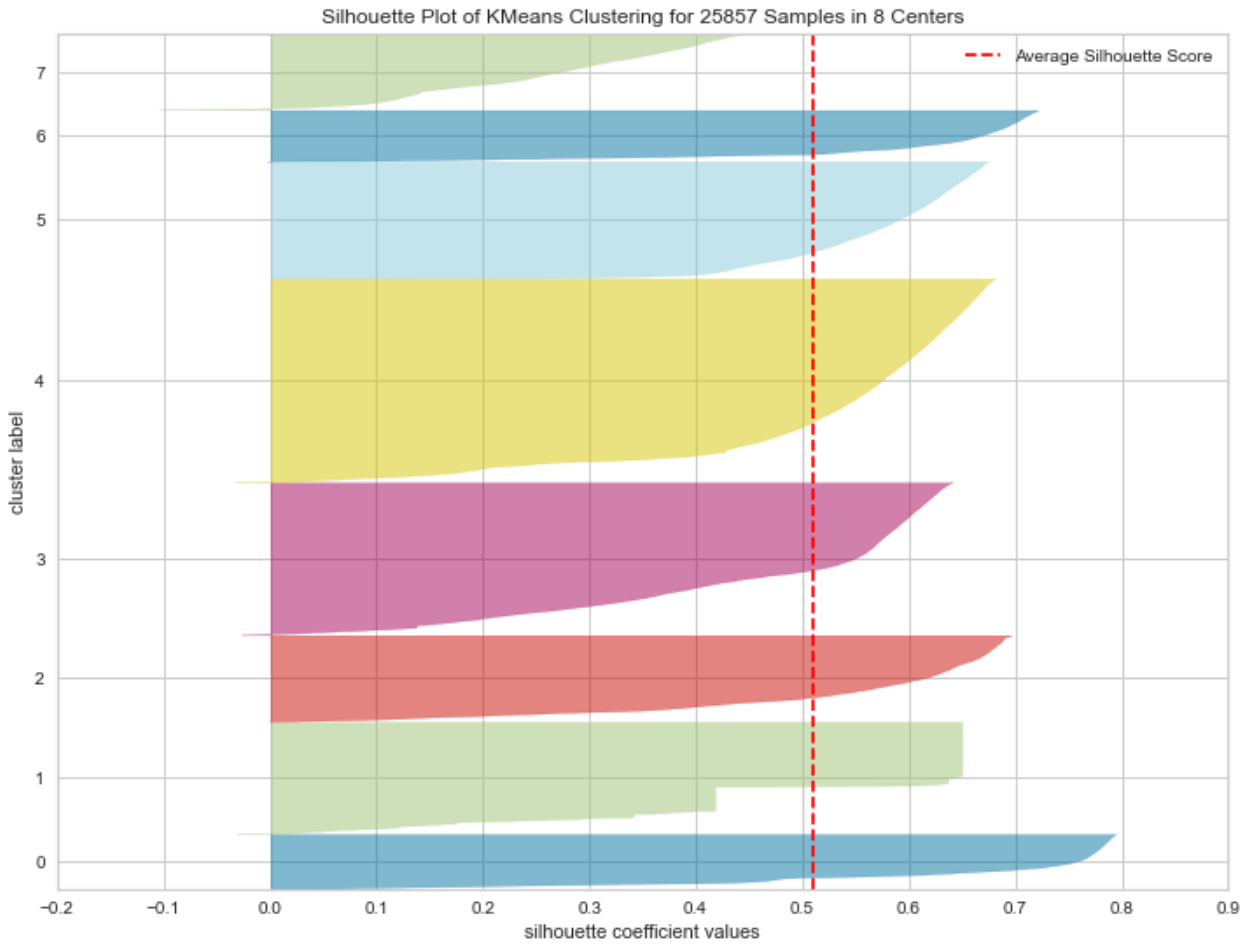

Figure 17
Silhouette Visualizer for k = 8.

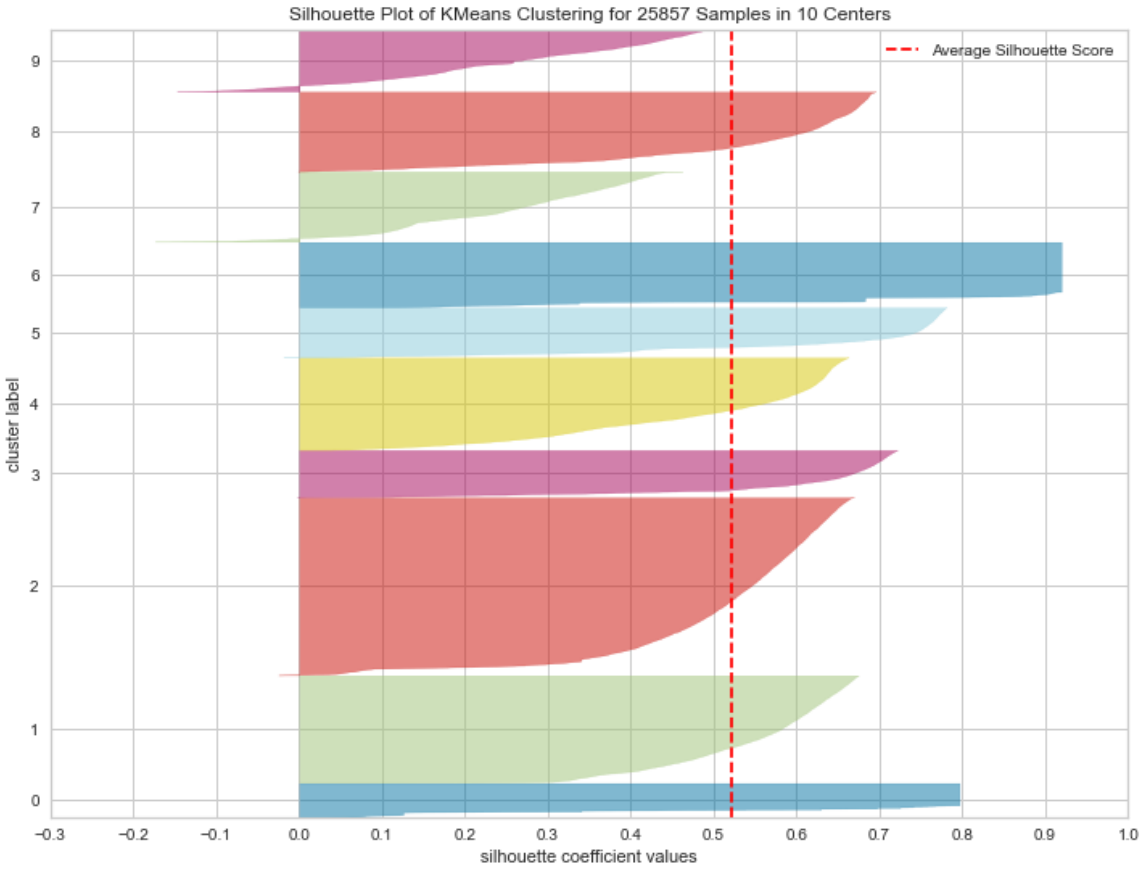

Figure 18
Silhouette Visualizer for k = 10.

We can see from the graph that both k of 8 and 10 are bad pick because they have one or more cluster labels with all points in the cluster with a below-average silhouette score. Therefore, **k=6 appears to be the optimal number of clusters** since all of its cluster labels have points above the average silhouette score and most of them are of the same size.

CLUSTERING PROFILE

The clustering profile for a cluster value of 6 was made and the averages of each numeric variable was taken with respect to the clusters and it can be given as(Red indicates maximum an d blue indicates minimum):

Cluster_Label	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Revenue	CTR	CPM	CPC	freq
0	459.843063	202.380052	73186.471283	5142266.710589	2443553.440558	2317889.235642	10961.593949	5272.363923	3510.042514	0.001860	1.577249	0.738180	4231
1	684.106494	303.387446	100764.935065	251986.073377	137958.653896	117044.801082	13833.337013	1261.017851	821.135363	0.133481	11.819506	0.089788	4620
2	136.186273	582.565309	73501.703716	27391.457732	16314.570177	10725.663313	1586.682622	166.403372	108.162137	0.160468	14.423699	0.101168	6163
3	434.316765	131.042223	62069.182255	1856547.132549	880786.020488	840840.468555	3395.492606	1537.238662	1000.667404	0.003833	1.855484	0.538619	5613
4	141.045154	573.449729	73686.935581	763383.617700	538047.037929	453769.058399	25759.151114	5794.252050	3825.374251	0.137875	15.570132	0.113179	1661
5	414.065565	295.023816	85329.784253	93296.488092	42148.968899	39409.721771	257.311011	77.853458	50.622670	0.007523	1.052482	0.048375	3569

Cluster_Label	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Revenue	CTR	CPM	CPC	freq
0	459.843063	202.380052	73186.471283	5142266.710589	2443553.440558	2317889.235642	10961.593949	5272.363923	3510.042514	0.001860	1.577249	0.738180	4231
1	684.106494	303.387446	100764.935065	251986.073377	137958.653896	117044.801082	13833.337013	1261.017851	821.135363	0.133481	11.819506	0.089788	4620
2	136.186273	582.565309	73501.703716	27391.457732	16314.570177	10725.663313	1586.682622	166.403372	108.162137	0.160468	14.423699	0.101168	6163
3	434.316765	131.042223	62069.182255	1856547.132549	880786.020488	840840.468555	3395.492606	1537.238662	1000.667404	0.003833	1.855484	0.538619	5613
4	141.045154	573.449729	73686.935581	763383.617700	538047.037929	453769.058399	25759.151114	5794.252050	3825.374251	0.137875	15.570132	0.113179	1661
5	414.065565	295.023816	85329.784253	93296.488092	42148.968899	39409.721771	257.311011	77.853458	50.622670	0.007523	1.052482	0.048375	3569

Figure 19
Clustering profile for 6 clusters

The inferences we can make looking at this is:

- Cluster label 2 has the highest average ad width with the smallest ad length and appears to be the poorest interacting cluster.
- Cluster label 0 appears to be the best interacting Cluster
- Cluster label 4 appears to be best performing cluster with having the highest average clicks and revenue generated.
- Cluster label 5 appears to be poorest performing cluster with having the lowest average clicks and revenue generated.

Now looking at a cluster profile where averages of each numeric variable was taken with respect to the clusters and Device type, we have (purple indicates maximum, green indicates, minimum):

Device Type	Cluster_Label	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Revenue	CTR	CPM	CPC	freq
Desktop	0	457.019557	202.568449	73190.691004	5137225.852999	2439856.299218	2314760.288136	11005.777705	5250.317852	3495.951585	0.001856	1.566578	0.732653	1534
	1	684.855072	303.562802	100804.347826	254772.141304	138322.144324	117394.961957	13943.804952	1256.846673	818.391240	0.133306	11.755574	0.089287	1656
	2	137.242877	581.004071	73569.877883	27608.983718	16298.029851	10699.413388	1598.035278	163.759199	106.443429	0.161126	14.428218	0.101049	2211
	3	436.109478	129.918905	61890.623416	1837836.559554	875230.134820	835277.785099	3412.110998	1534.633117	998.729590	0.003874	1.861855	0.536178	1973
	4	142.768212	569.867550	73728.476821	757784.821192	535594.200331	451786.226821	25765.617550	5784.061449	3822.659302	0.138429	15.601424	0.113063	604
	5	413.179650	294.252782	85213.831479	98495.675676	44299.110493	41450.709857	247.864070	81.989332	53.343959	0.007088	1.025700	0.050024	1258
Mobile	0	461.449017	202.272896	73184.071190	5145133.850204	2445656.301075	2319668.918799	10936.463107	5284.903291	3518.057154	0.001863	1.583319	0.741324	2697
	1	683.688259	303.289474	100742.914980	250429.484818	137755.570175	116849.164642	13771.618084	1263.348306	822.668516	0.133579	11.855226	0.090067	2964
	2	135.595142	583.438765	73463.562753	27269.759868	16323.823887	10740.349190	1580.331225	167.882690	109.123691	0.160100	14.421172	0.101235	3952
	3	433.345055	131.651099	62165.967033	1866688.879945	883797.493681	843855.626374	3386.484890	1538.650953	1001.717764	0.003811	1.852030	0.539942	3640
	4	140.060549	575.496689	73663.197729	766582.929991	539448.659413	454902.105014	25755.456008	5800.075251	3826.925650	0.137559	15.552252	0.113245	1057
	5	414.547815	295.443531	85392.903505	90466.294245	40978.532670	38298.703592	262.453483	75.602081	49.141328	0.007759	1.067062	0.047477	2311

Device Type	Cluster_Label	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Revenue	CTR	CPM	CPC	freq
Desktop	0	457.019557	202.568449	73190.691004	5137225.852999	2439856.299218	2314760.288136	11005.777705	5250.317852	3495.951585	0.001856	1.566578	0.732653	1534
	1	684.855072	303.562802	100804.347826	254772.141304	138322.144324	117394.961957	13943.804952	1256.846673	818.391240	0.133306	11.755574	0.089287	1656
	2	137.242877	581.004071	73569.877883	27608.983718	16298.029851	10699.413388	1598.035278	163.759199	106.443429	0.161126	14.428218	0.101049	2211
	3	436.109478	129.918905	61890.623416	1837836.559554	875230.134820	835277.785099	3412.110998	1534.633117	998.729590	0.003874	1.861855	0.536178	1973
	4	142.768212	569.867550	73728.476821	757784.821192	535594.200331	451786.226821	25765.617550	5784.061449	3822.659302	0.138429	15.601424	0.113063	604
	5	413.179650	294.252782	85213.831479	98495.675676	44299.110493	41450.709857	247.864070	81.989332	53.343959	0.007088	1.025700	0.050024	1258
Mobile	0	461.449017	202.272896	73184.071190	5145133.850204	2445656.301075	2319668.918799	10936.463107	5284.903291	3518.057154	0.001863	1.583319	0.741324	2697
	1	683.688259	303.289474	100742.914980	250429.484818	137755.570175	116849.164642	13771.618084	1263.348306	822.668516	0.133579	11.855226	0.090067	2964
	2	135.595142	583.438765	73463.562753	27269.759868	16323.823887	10740.349190	1580.331225	167.882690	109.123691	0.160100	14.421172	0.101235	3952
	3	433.345055	131.651099	62165.967033	1866688.879945	883797.493681	843855.626374	3386.484890	1538.650953	1001.717764	0.003811	1.852030	0.539942	3640
	4	140.060549	575.496689	73663.197729	766582.929991	539448.659413	454902.105014	25755.456008	5800.075251	3826.925650	0.137559	15.552252	0.113245	1057
	5	414.547815	295.443531	85392.903505	90466.294245	40978.532670	38298.703592	262.453483	75.602081	49.141328	0.007759	1.067062	0.047477	2311

Figure 20

Clustering profile for 6 cluster with respect to device type

The inferences we can make looking at this is:

- Cluster label 0 appears to be the best interacting Cluster and is on the mobile platform.

- Cluster label 4 appears to be best performing cluster with having the highest average clicks and revenue generated on the mobile platform.

Let us also have a look at the distribution of the clusters across different categorical variables.

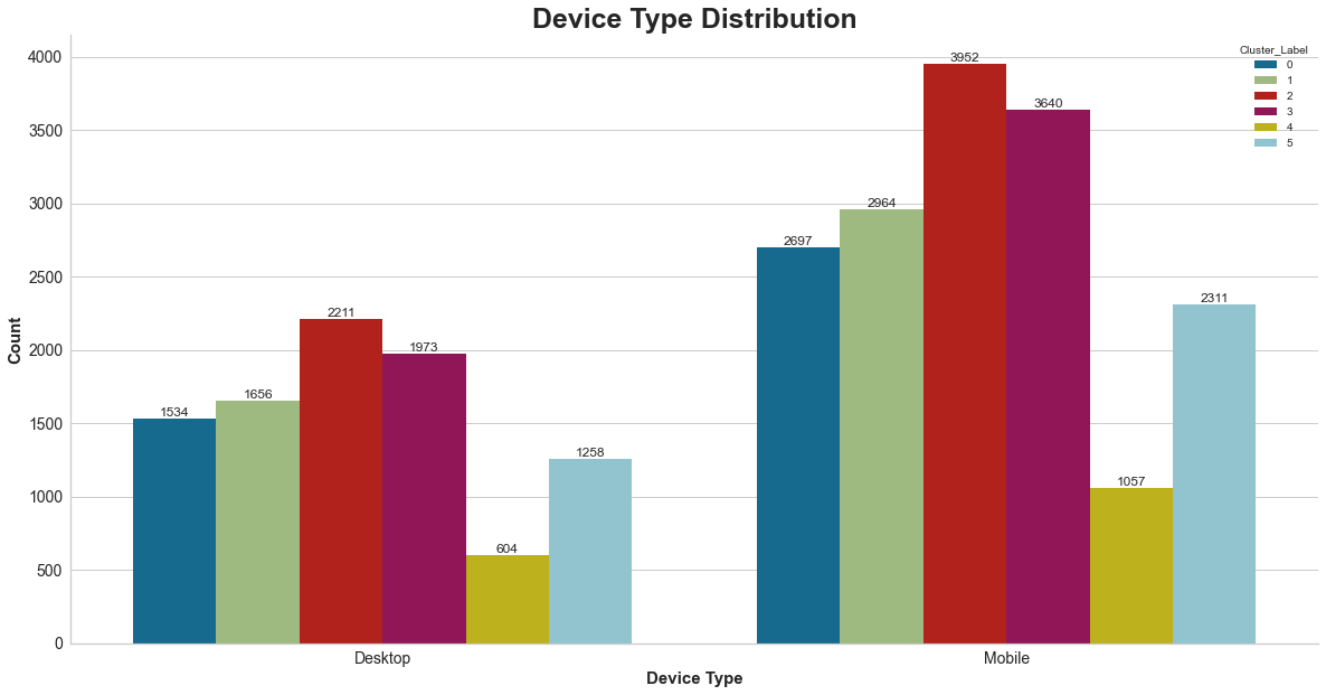

Figure 21

Distribution of clusters with respect to device type.

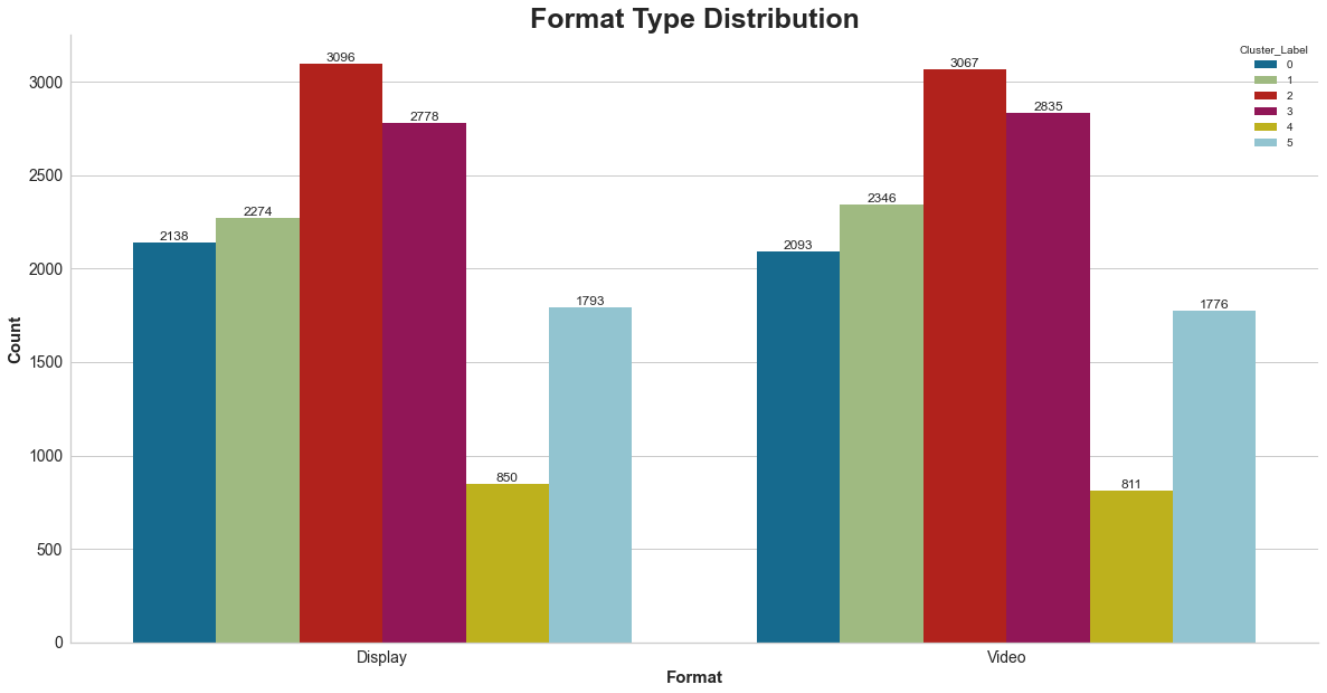

Figure 22

Distribution of clusters with respect to format type.

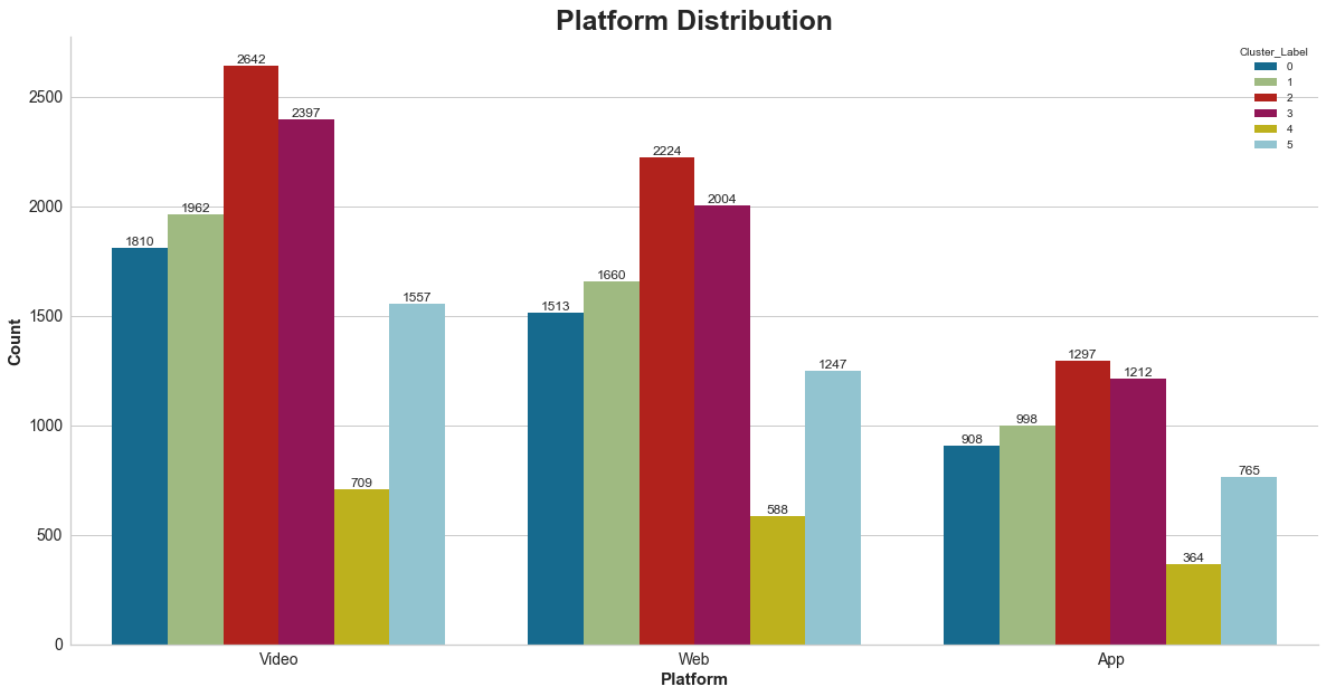

Figure 23

Distribution of clusters with respect to platform.

Given that the clusters were properly optimized we can clearly see in the bar plots the no matter the categorical variable the distribution of the cluster is similar in all. This testament that the cluster size chose was the MOST OPTIMAL one.

SUMMARY

We can summarize the clustering of the data and our process and results as follows:

- Data had null values, which were imputed by deriving it from existing data within the dataframe utilizing certain formulas for CPM, CTR and CPC.
- Since K means algorithm is very sensitive to outliers, we chose to treat the outliers by moving them to the nearest quartile.
- Based on the elbow plot and silhouette scores and visualizer we found k = 6 is the most optimal number of clusters.
- Most clusters formed identify their characteristics with at least one metric.

PCA

PROBLEM STATEMENT

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.

DATA DICTIONARY

The data dictionary gives us information regarding the variables present tin the given dataset we are analyzing and a brief explanation and what that variable contains or means

- **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
- **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

DATA INITIALIZATION AND PRE-PROCESSING

Data was initialized from given excel file to a data frame and head and tail values of said data frame was obtained to validate this:

	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_H
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180	237	680	
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123	229	186	
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	...	114	188	44	89	3	
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	...	194	247	61	128	13	
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	...	874	1928	465	1043	205	

5 rows × 61 columns

	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG
635	34	636	Puducherry	Mahe	3333	8154	11781	1146	1203	21	...	32	47	0	0	0	
636	34	637	Puducherry	Karaikal	10612	12346	21691	1544	1533	2234	...	155	337	3	14	38	
637	35	638	Andaman & Nicobar Island	Nicobars	1275	1549	2630	227	225	0	...	104	134	9	4	2	
638	35	639	Andaman & Nicobar Island	North & Middle Andaman	3762	5200	8012	723	664	0	...	136	172	24	44	11	
639	35	640	Andaman & Nicobar Island	South Andaman	7975	11977	18049	1470	1358	0	...	173	122	6	2	17	

5 rows × 61 columns

Figure 24
Head and tail values of initialized dataframe

The size of the data frame was found to be of 640 rows and 61 columns as seen:

```
No. of Rows: 640
No. of Columns: 61
```

Figure 25
Size info of datafarame

Basic info regarding the columns were found to be as:

```
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
```

Figure 26

Info of given dataframe

The data also had **NO null values** & it **did not have any duplicates.**

The data summary is as follows:

	count	mean	std	min	25%	50%	75%	max
State Code	640.0	17.114062	9.426486	1.0	9.00	18.0	24.00	35.0
Dist.Code	640.0	320.500000	184.896367	1.0	160.75	320.5	480.25	640.0
No_HH	640.0	51222.871875	48135.405475	350.0	19484.00	35837.0	68892.00	310450.0
TOT_M	640.0	79940.576563	73384.511114	391.0	30228.00	58339.0	107918.50	485417.0
TOT_F	640.0	122372.084375	113600.717282	698.0	46517.75	87724.5	164251.75	750392.0
...	...	...	...	...	...	...	...	...
MARG_HH_0_3_F	640.0	1293.431250	1585.377936	0.0	298.00	717.0	1710.75	13714.0
MARG_OT_0_3_M	640.0	71.379688	107.897627	0.0	14.00	35.0	79.00	895.0
MARG_OT_0_3_F	640.0	200.742188	309.740854	0.0	43.00	113.0	240.00	3354.0
NON_WORK_M	640.0	510.014063	610.603187	0.0	161.00	326.0	604.50	6456.0
NON_WORK_F	640.0	704.778125	910.209225	5.0	220.50	464.5	853.50	10533.0

59 rows × 8 columns

Figure 27
Data summary before any form of processing.

before we process given that most if not all columns have their own weights.

The data does look to have outliers, but we will choose not to treat them. Not treating outliers in this case, it he best course of action as treatment of outliers in this census data will introduce biases which are highly unwanted in this case. We will look at this further in the report.

EXPLORATORY DATA ANALYSIS (EDA)

To aid in our EDA of the given data we will first form a few questions which will help us to do the analysis in a more structured way i.e., we will only have to look at the variables we need to answer the questions.

i. Which state has highest gender ratio, and which has the lowest?
ii. Which district has the highest & lowest gender ratio?
iii. Which state has the highest literacy ratio among its male and female population, and which has the lowest?
iv. Which state among its literate population has the most gap among literacy between the genders and which state the lowest?

To answer the questions above we have to consider the following variables: Dist. Code, State, Area Name, TOT_M, TOT_F, M_LIT and F_LIT

As seen from the data in Fig 28 and Fig 29 where red indicates maximum and blue indicates, we can use tis data to answer question i above.

It would appear that Lakshadweep has the highest gender ratio among all the state i.e., the scale of population of Men in Lakshadweep compared to the population of women in Lakshadweep is the highest with respect to any other state.(0.86 Almost the amount of men equal to women)

And Andhra Pradesh has the lowest ender ratio among all the states i.e., scale of population of Men in Andhra Pradesh compared to the population of women in Andhra Pradesh is the lowest with respect to any other state. (0.53 Twice as many women than men)

State	TOT_M	TOT_F	Gender_ratio
Andaman & Nicobar Island	18726	28691	0.652679
Andhra Pradesh	3274363	6097235	0.537024
Arunachal Pradesh	50582	88066	0.574365
Assam	1437268	2093432	0.686561
Bihar	4025198	5405883	0.744596
Chandigarh	41753	59644	0.700037
Chhattisgarh	838404	1526592	0.549200
Dadara & Nagar Havelli	6982	10831	0.644631
Daman & Diu	13153	18706	0.703143
Goa	118979	191393	0.621648
Gujarat	1983685	2939472	0.674844
Haryana	1167816	1498873	0.779129
Himachal Pradesh	483381	752062	0.642741
Jammu & Kashmir	421213	572959	0.735154
Jharkhand	1202623	1763884	0.681804
Karnataka	3409482	5345675	0.637802
Kerala	2919825	4856357	0.601238
Lakshadweep	12823	14772	0.868061
Madhya Pradesh	2155608	3369745	0.639695
Maharashtra	4196130	7138557	0.587812
Manipur	145524	226963	0.641179
Meghalaya	268036	356355	0.752160
Mizoram	59534	95463	0.623634
NCT of Delhi	833414	1075266	0.775077
Nagaland	73506	125935	0.583682
Odisha	1460031	2536980	0.575500
Puducherry	70386	119074	0.591111
Punjab	1579405	2121425	0.744502
Rajasthan	2062563	2966496	0.695286
Sikkim	26664	41518	0.642227
Tamil Nadu	3074009	5610310	0.547921
Tripura	160457	256370	0.625881
Uttar Pradesh	9043969	12023885	0.752167
Uttarakhand	613924	973147	0.630865
West Bengal	3912553	6016118	0.650345

Figure 29

Gender Ratio profile by state with highlighted max values

State	TOT_M	TOT_F	Gender_ratio
Andaman & Nicobar Island	18726	28691	0.652679
Andhra Pradesh	3274363	6097235	0.537024
Arunachal Pradesh	50582	88066	0.574365
Assam	1437268	2093432	0.686561
Bihar	4025198	5405883	0.744596
Chandigarh	41753	59644	0.700037
Chhattisgarh	838404	1526592	0.549200
Dadara & Nagar Havelli	6982	10831	0.644631
Daman & Diu	13153	18706	0.703143
Goa	118979	191393	0.621648
Gujarat	1983685	2939472	0.674844
Haryana	1167816	1498873	0.779129
Himachal Pradesh	483381	752062	0.642741
Jammu & Kashmir	421213	572959	0.735154
Jharkhand	1202623	1763884	0.681804
Karnataka	3409482	5345675	0.637802
Kerala	2919825	4856357	0.601238
Lakshadweep	12823	14772	0.868061
Madhya Pradesh	2155608	3369745	0.639695
Maharashtra	4196130	7138557	0.587812
Manipur	145524	226963	0.641179
Meghalaya	268036	356355	0.752160
Mizoram	59534	95463	0.623634
NCT of Delhi	833414	1075266	0.775077
Nagaland	73506	125935	0.583682
Odisha	1460031	2536980	0.575500
Puducherry	70386	119074	0.591111
Punjab	1579405	2121425	0.744502
Rajasthan	2062563	2966496	0.695286
Sikkim	26664	41518	0.642227
Tamil Nadu	3074009	5610310	0.547921
Tripura	160457	256370	0.625881
Uttar Pradesh	9043969	12023885	0.752167
Uttarakhand	613924	973147	0.630865
West Bengal	3912553	6016118	0.650345

Figure 28

Gender Ratio profile by state with highlighted min values

```
District Code 587 has the highest gender ratio of 0.87
District Code 547 has the lowest gender ratio of 0.44
```

Figure 30
Gender ratio by district

And to answer question ii district with code 587 has the highest gender ratio and district with code 547 has the lowest gender ratio.

Creating profile on literacy, we have:

State	M_LIT	F_LIT	Literacy_Gender_ratio	Literacy_ratio_MaleTOT	Literacy_ratio_FemaleTOT
Andaman & Nicobar Island	15488	20237	0.765331	0.827085	0.705343
Andhra Pradesh	2372971	2678603	0.885899	0.724712	0.439314
Arunachal Pradesh	33965	45307	0.749663	0.671484	0.514466
Assam	1023294	1152979	0.887522	0.711972	0.550760
Bihar	2408492	2197931	1.095800	0.598354	0.406581
Chandigarh	33552	43438	0.772411	0.803583	0.728288
Chhattisgarh	614878	703825	0.873623	0.733391	0.461043
Dadara & Nagar Havelli	5119	5308	0.964393	0.733171	0.490075
Daman & Diu	10880	12520	0.869010	0.827188	0.669304
Goa	99381	139749	0.711139	0.835282	0.730168
Gujarat	1509399	1722877	0.876092	0.760907	0.586118
Haryana	874982	826676	1.058434	0.749246	0.551532
Himachal Pradesh	387845	492442	0.787595	0.802359	0.654789
Jammu & Kashmir	283106	288053	0.982826	0.672121	0.502746
Jharkhand	799838	768942	1.040180	0.665078	0.435937
Karnataka	2554163	2905370	0.879118	0.749135	0.543499
Kerala	2370331	3878204	0.611193	0.811806	0.798583
Lakshadweep	10601	11334	0.935327	0.826718	0.767262
Madhya Pradesh	1537129	1656596	0.927884	0.713084	0.491609
Maharashtra	3308633	4619012	0.716308	0.788496	0.647051
Manipur	110260	133868	0.823647	0.757676	0.589823
Meghalaya	163712	220041	0.744007	0.610784	0.617477
Mizoram	48512	79412	0.610890	0.814862	0.831862
NCT of Delhi	659926	742160	0.889196	0.791835	0.690211
Nagaland	55831	84327	0.662077	0.759543	0.669607
Odisha	1076443	1294343	0.831652	0.737274	0.510190
Puducherry	58230	78314	0.743545	0.827295	0.657692
Punjab	1171941	1296619	0.903844	0.742014	0.611202
Rajasthan	1450321	1313776	1.103933	0.703164	0.442871
Sikkim	21230	27112	0.783048	0.796205	0.653018
Tamil Nadu	2485404	3205093	0.775455	0.808522	0.571286
Tripura	130890	183251	0.714266	0.815733	0.714791
Uttar Pradesh	6016402	5574752	1.079223	0.665239	0.463640
Uttarakhand	463737	588335	0.788219	0.755365	0.604570
West Bengal	2932621	3479316	0.842873	0.749542	0.578332

Figure 31
Literacy ratio profile by state with max values highlighted

State	M_LIT	F_LIT	Literacy_Gender_ratio	Literacy_ratio_MaleTOT	Literacy_ratio_FemaleTOT
Andaman & Nicobar Island	15488	20237	0.765331	0.827085	0.705343
Andhra Pradesh	2372971	2678603	0.885899	0.724712	0.439314
Arunachal Pradesh	33965	45307	0.749663	0.671484	0.514466
Assam	1023294	1152979	0.887522	0.711972	0.550760
Bihar	2408492	2197931	1.095800	[illegible]	[illegible]
Chandigarh	33552	43438	0.772411	0.803583	0.728288
Chhattisgarh	614878	703825	0.873623	0.733391	0.461043
Dadara & Nagar Havelli	[illegible]	[illegible]	0.964393	0.733171	0.490075
Daman & Diu	10880	12520	0.869010	0.827188	0.669304
Goa	99381	139749	0.711139	0.835282	0.730168
Gujarat	1509399	1722877	0.876092	0.760907	0.586118
Haryana	874982	826676	1.058434	0.749246	0.551532
Himachal Pradesh	387845	492442	0.787595	0.802359	0.654789
Jammu & Kashmir	283106	288053	0.982826	0.672121	0.502746
Jharkhand	799838	768942	1.040180	0.665078	0.435937
Karnataka	2554163	2905370	0.879118	0.749135	0.543499
Kerala	2370331	3878204	0.611193	0.811806	0.798583
Lakshadweep	10601	11334	0.935327	0.826718	0.767262
Madhya Pradesh	1537129	1656596	0.927884	0.713084	0.491609
Maharashtra	3308633	4619012	0.716308	0.788496	0.647051
Manipur	110260	133868	0.823647	0.757676	0.589823
Meghalaya	163712	220041	0.744007	0.610784	0.617477
Mizoram	48512	79412	[illegible]	0.814862	0.831862
NCT of Delhi	659926	742160	0.889196	0.791835	0.690211
Nagaland	55831	84327	0.662077	0.759543	0.669607
Odisha	1076443	1294343	0.831652	0.737274	0.510190
Puducherry	58230	78314	0.743545	0.827295	0.657692
Punjab	1171941	1296619	0.903844	0.742014	0.611202
Rajasthan	1450321	1313776	1.103933	0.703164	0.442871
Sikkim	21230	27112	0.783048	0.796205	0.653018
Tamil Nadu	2485404	3205093	0.775455	0.808522	0.571286
Tripura	130890	183251	0.714266	0.815733	0.714791
Uttar Pradesh	6016402	5574752	1.079223	0.665239	0.463640
Uttarakhand	463737	588335	0.788219	0.755365	0.604570
West Bengal	2932621	3479316	0.842873	0.749542	0.578332

Figure 32

Literacy ratio profile by state with min values highlighted

Answering Question iii

- **Goa has the Highest Literacy** rate among all the states with respect to its **Male** population.
- **Mizoram has the Highest Literacy** rate among all the states with respect to its **Female** population.
- **Bihar has the Lowest Literacy rate** among all the states with respect **to BOTH its Male and Female population**.

Answering question iv

- Mizoram has the most gap in literacy education given that Female literate population is twice as that of the male literate population.

- Rajasthan has the smallest gap in literacy education where the Female literate population is almost same as that of the male literate population.

SCALING

We scale the data using z score method. **Scaling of data here will help reduce the effect of outliers by bringing every attribute on a common scale.**

A glimpse of the data summary of the scaled data:

	count	mean	std	min	25%	50%	75%	max
State Code	640.0	-1.121325e-15	1.000782	-1.710782	-0.861446	0.094057	0.731060	1.898897
Dist.Code	640.0	-5.169476e-17	1.000782	-1.729347	-0.864673	0.000000	0.864673	1.729347
No_HH	640.0	7.736867e-17	1.000782	-1.057697	-0.659882	-0.319887	0.367358	5.389586
TOT_M	640.0	-1.864828e-16	1.000782	-1.084858	-0.677956	-0.294592	0.381549	5.529690
TOT_F	640.0	-2.983724e-17	1.000782	-1.071906	-0.668250	-0.305233	0.368945	5.532633
...	...	...	...	...	...	...	...	...
MARG_HH_0_3_F	640.0	-1.247266e-16	1.000782	-0.816489	-0.628374	-0.363877	0.263436	7.840581
MARG_OT_0_3_M	640.0	-2.515349e-17	1.000782	-0.662068	-0.532213	-0.337432	0.070681	7.639320
MARG_OT_0_3_F	640.0	-3.053113e-17	1.000782	-0.648604	-0.509670	-0.283498	0.126843	10.188272
NON_WORK_M	640.0	-5.741935e-17	1.000782	-0.835916	-0.572036	-0.301600	0.154863	9.745505
NON_WORK_F	640.0	-1.994932e-17	1.000782	-0.769412	-0.532468	-0.264188	0.163521	10.806207

Figure 33
Data Summary After Scaling

Boxplots of before scaling:

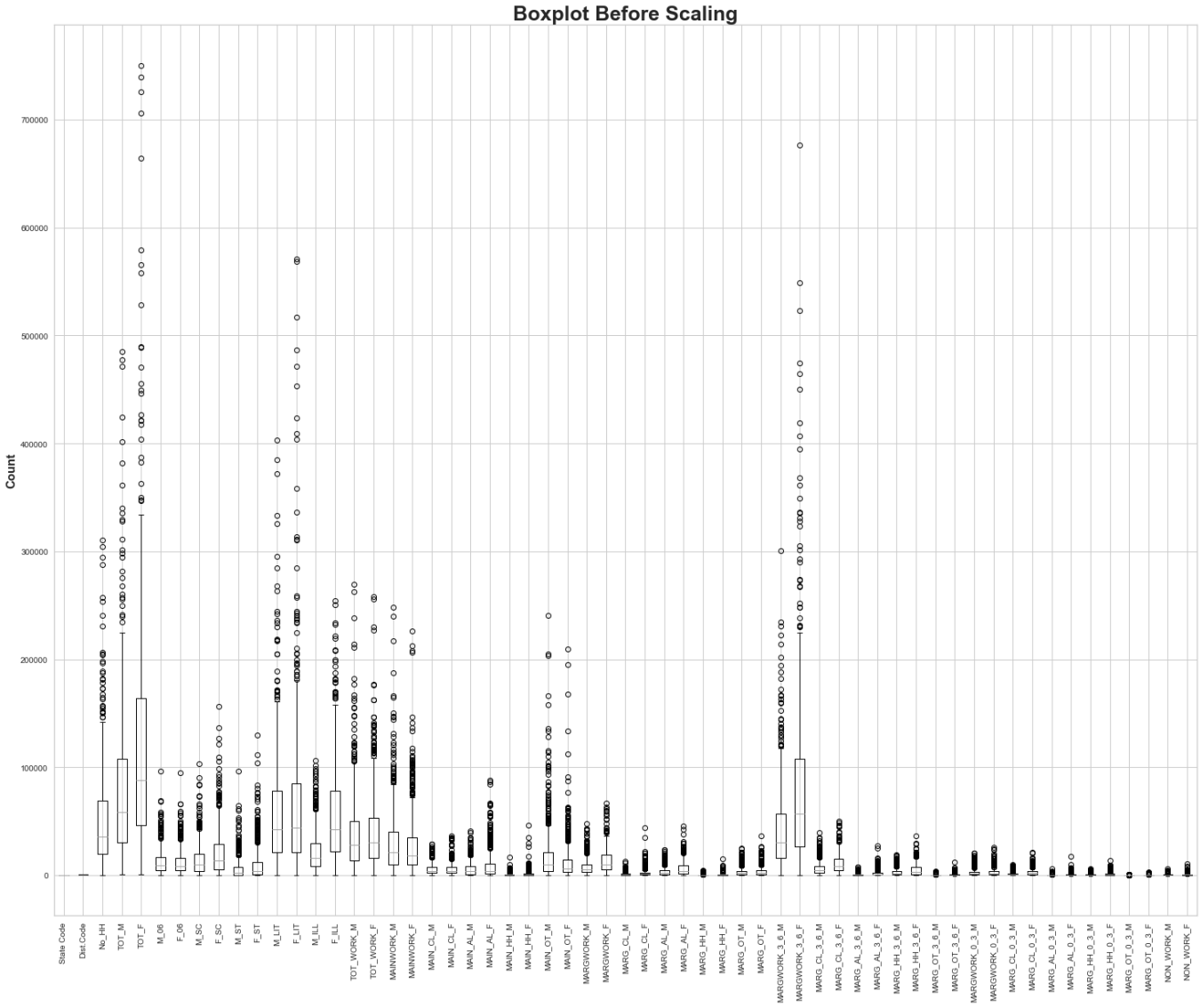

Figure 34

Box plot of al numeric data before any form of scaling.

Boxplots of after scaling:

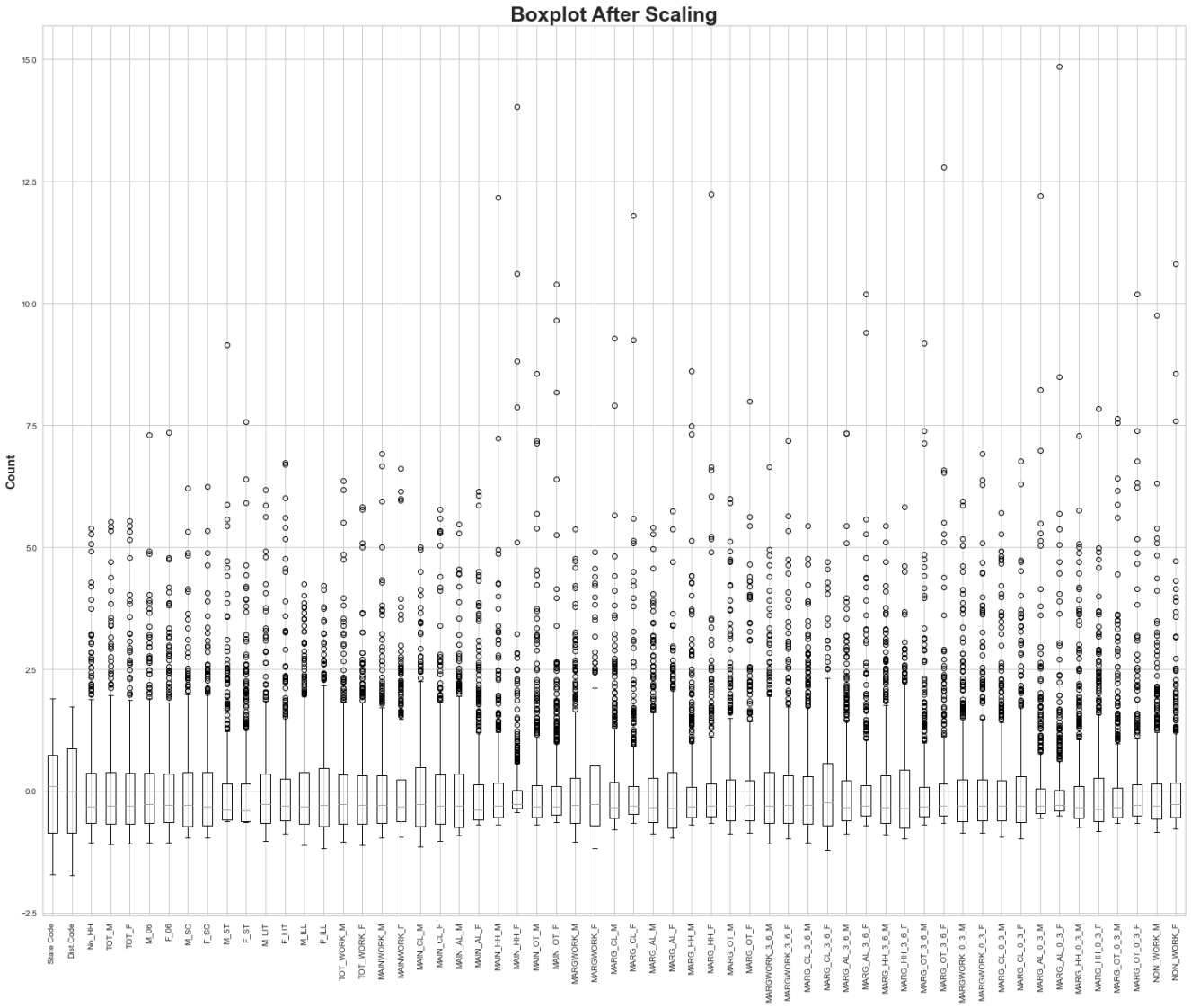

Figure 35

Box plot of al numeric data after scaling.

Heatmap of scaled data:

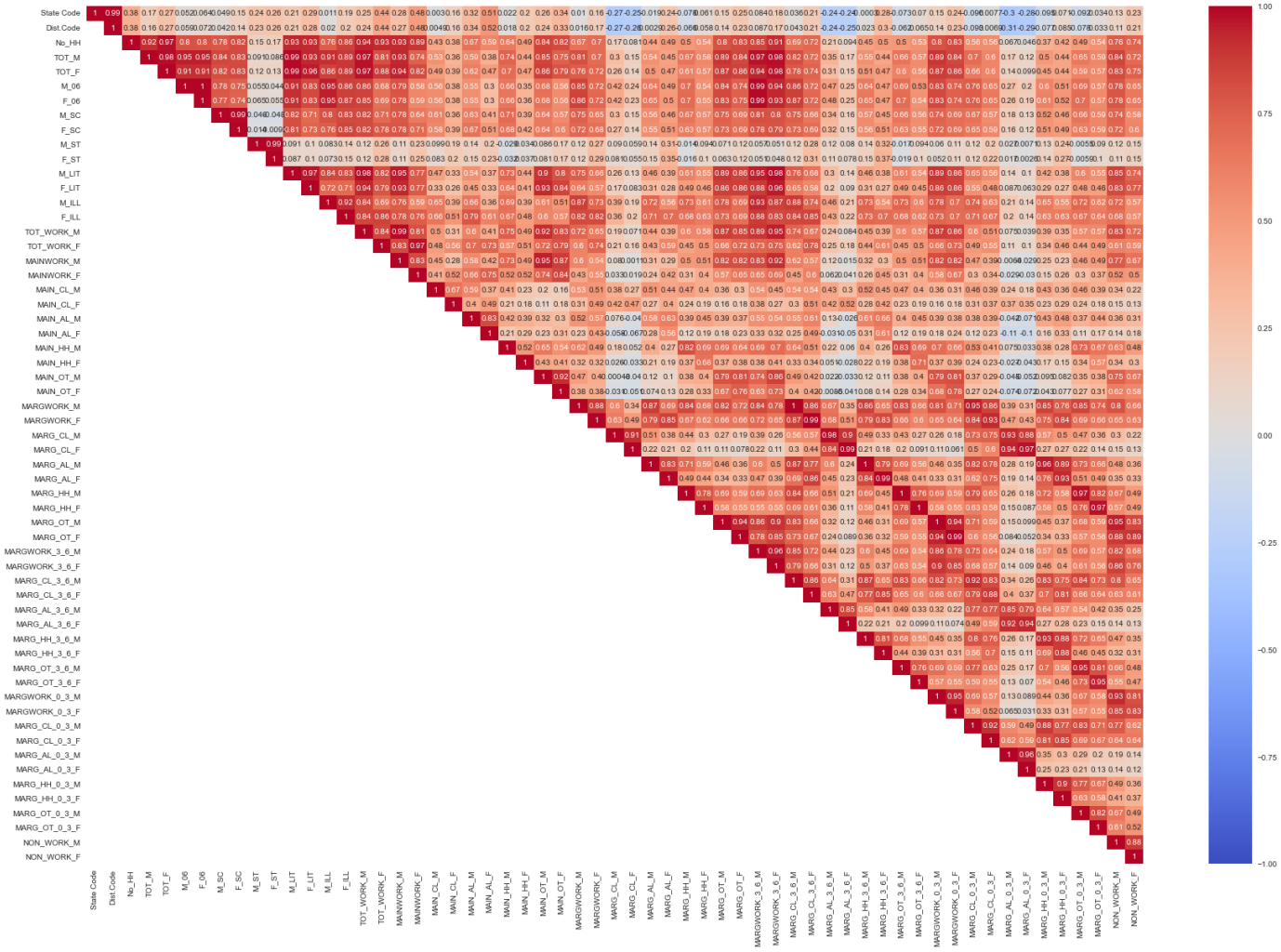

Figure 36

heatmap of correlation among the variables.

PRINCIPAL COMPONENT ANALYSIS (PCA)

To better hep data analysis we need to reduce the number of dimensions we perform analysis on and hence we use PCA for this purpose to help find the most relevant Principal component to hold onto while we remove the rest.

From initial analysis it would appear that 11 components will for sure have close to 100% explained variance, so performing the analysis with n=11, the eigen vectors are given as:

```
array([[ 3.00700521e-02,  3.00751392e-02,  1.56432451e-01,
         1.67038499e-01,  1.65701886e-01,  1.61870848e-01,
         1.62266320e-01,  1.51067631e-01,  1.51483487e-01,
         2.76635864e-02,  2.86559949e-02,  1.62028968e-01,
         1.47117900e-01,  1.61354631e-01,  1.65216191e-01,
         1.59988739e-01,  1.46484663e-01,  1.46446784e-01,
         1.24700922e-01,  1.02841551e-01,  7.46387972e-02,
         1.13762012e-01,  7.47868720e-02,  1.31280497e-01,
         8.36015471e-02,  1.23789890e-01,  1.11498595e-01,
         1.64144005e-01,  1.55258801e-01,  8.14703494e-02,
         4.84108523e-02,  1.28166982e-01,  1.14462067e-01,
         1.40274353e-01,  1.27424449e-01,  1.55154856e-01,
         1.47413552e-01,  1.64714317e-01,  1.61211005e-01,
         1.65089659e-01,  1.55618244e-01,  9.21330578e-02,
         5.07812312e-02,  1.28188765e-01,  1.10910853e-01,
         1.39029295e-01,  1.24330759e-01,  1.54196780e-01,
         1.46411774e-01,  1.49444956e-01,  1.39705021e-01,
         5.16456518e-02,  4.09693847e-02,  1.21254301e-01,
         1.15790305e-01,  1.39259946e-01,  1.31868671e-01,
         1.50219557e-01,  1.31179136e-01],
       [-1.62782525e-01, -1.58821825e-01, -1.28322211e-01,
        -8.08606182e-02, -1.01110634e-01, -1.27532047e-02,
        -1.16738305e-02, -3.56271016e-02, -4.77317054e-02,
         8.89310174e-03,  9.76501195e-03, -1.06708909e-01,
        -1.45649208e-01,  1.62458008e-03, -1.18224696e-02,
        -1.26023885e-01, -9.61653150e-02, -1.68329248e-01,
        -1.61038991e-01,  6.07840614e-02,  7.23819582e-02,
        -4.50720251e-02, -8.37820856e-02, -6.12919048e-02,
        -8.17970841e-02, -2.00257744e-01, -2.04433769e-01,
         9.57289425e-02,  1.14061542e-01,  2.70006432e-01,
         2.45991056e-01,  1.58404811e-01,  1.17724869e-01,
         7.75591059e-02,  2.77442703e-02, -7.74643230e-02,
        -1.10153459e-01, -3.44262882e-02, -9.53306829e-02,
         7.96132242e-02,  9.06783669e-02,  2.63961514e-01,
         2.42794338e-01,  1.50601318e-01,  1.00311664e-01,
         7.19234625e-02,  1.83439655e-02, -8.11258916e-02,
        -1.17213240e-01,  1.54507583e-01,  1.74434255e-01,
         2.53833841e-01,  2.42224085e-01,  1.81278106e-01,
         1.65403260e-01,  9.34463740e-02,  5.40694563e-02,
        -5.44095594e-02, -6.94741471e-02],
       [-2.50129023e-01, -2.59359844e-01, -3.34978668e-02,
         6.36304034e-02,  2.44033063e-02,  7.04531069e-02,
         6.35151136e-02,  3.53451659e-02, -9.67682787e-03,
        -2.01756454e-01, -2.20128794e-01,  7.80968550e-02,
         9.42145144e-02,  1.52867377e-02, -9.12080649e-02,
         4.91747510e-02, -1.26154826e-01,  5.32233768e-02,
        -1.19313853e-01, -7.37319713e-02, -1.21925247e-01,
        -2.41982367e-01, -3.13530915e-01,  1.02102381e-01,
        -2.48996908e-02,  1.32073828e-01,  6.23400508e-02,
         1.55720272e-02, -1.01194932e-01,  1.04666117e-01,
         1.25645506e-01, -1.42069272e-01, -2.90271087e-01,
         6.48576663e-02, -4.15682832e-03,  1.34472025e-01,
         9.97709536e-02,  7.37236777e-02,  8.99625450e-02,
         5.01551036e-03, -1.23580343e-01,  7.66976958e-02,
         1.11763378e-01, -1.53496372e-01, -3.09743511e-01,
         6.61941498e-02, -4.90542127e-03,  1.33890815e-01,
         9.96623322e-02,  5.71982572e-02, -2.38004782e-02,
         1.49588784e-01,  1.51083464e-01, -8.86563289e-02,
        -1.97795277e-01,  5.80312865e-02, -1.83333914e-03,
         1.28955424e-01,  8.67015734e-02],
       [ 1.20048868e-01,  1.10852274e-01,  1.01334853e-01,
         3.32992227e-02,  7.19475117e-02,  7.70298515e-03,
         2.41738125e-03, -2.46028665e-02,  2.28440145e-03,
         1.42128234e-01,  1.41942213e-01,  5.99044380e-02,
         1.00906949e-01, -4.56796582e-02,  1.27650599e-02,
```

Figure 37

First 4 Eigen vectors after PCA

The above image represents **only up to first 4 vectors.**

Eigen Values:

```
array([31.86742634,  8.18907061,  4.54275124,  3.84336785,  2.27105793,
        1.95992589,  1.37548006,  0.88734267,  0.71989796,  0.61405955,
        0.49439968])
```

Figure 38
Eigen Values

Contribution of Each Principal Component:

```
array([53.92819235, 13.85809353,  7.68754779,  6.50400438,  3.84323628,
        3.3167178 ,  2.32767944,  1.50162068,  1.21825953,  1.03915268,
        0.83665625])
```

Figure 39
Contribution of Each PC

Cumulative Sum:

```
Cumulative Variance Explained
 [0.53928192 0.67786286 0.75473834 0.81977838 0.85821074 0.89137792
 0.91465472 0.92967092 0.94185352 0.95224504 0.96061161]
```

The Scree plot for the above analysis is given as:

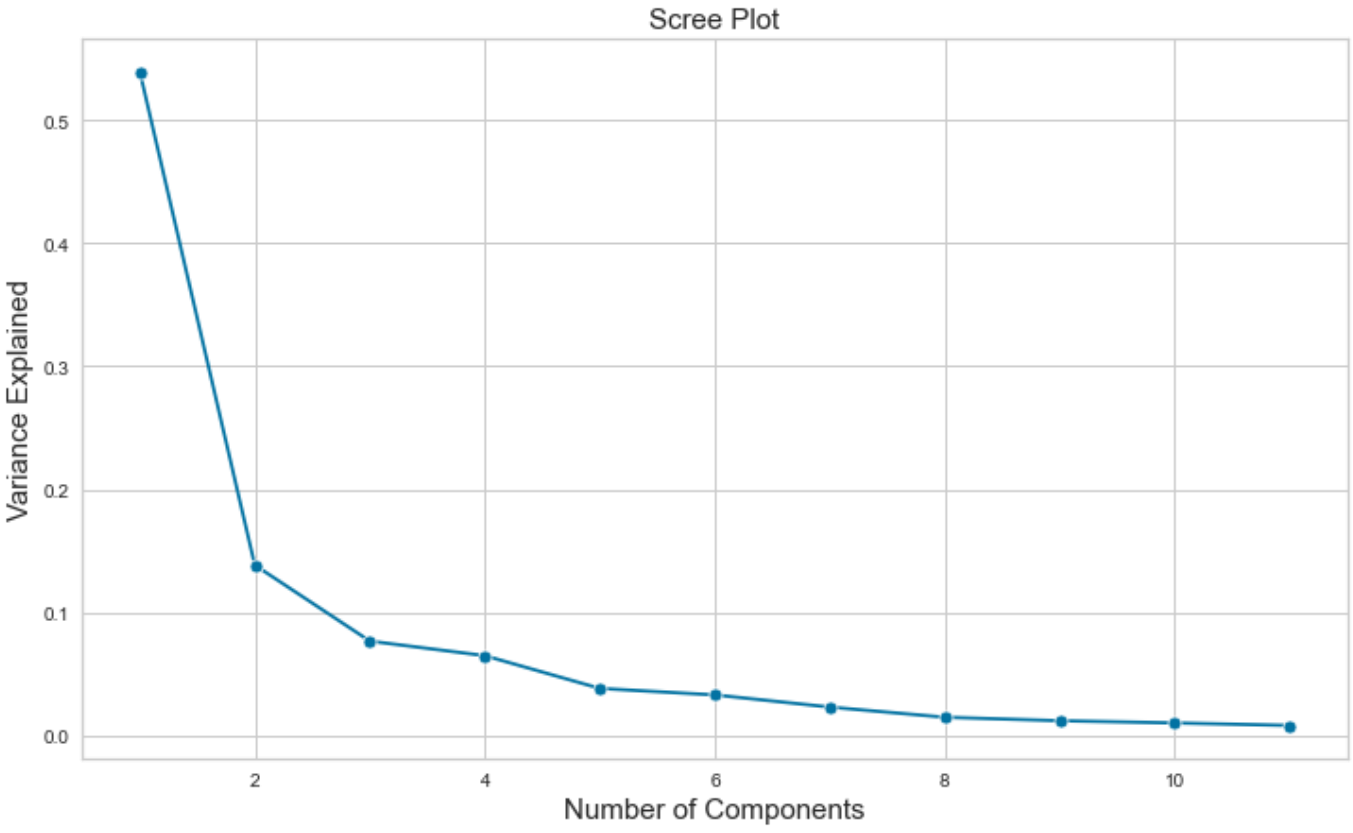

Figure 40
Scree plot of contribution of all 11 Principal Components

To understand it better let us also plot the Explained variance ratio vs principal components:

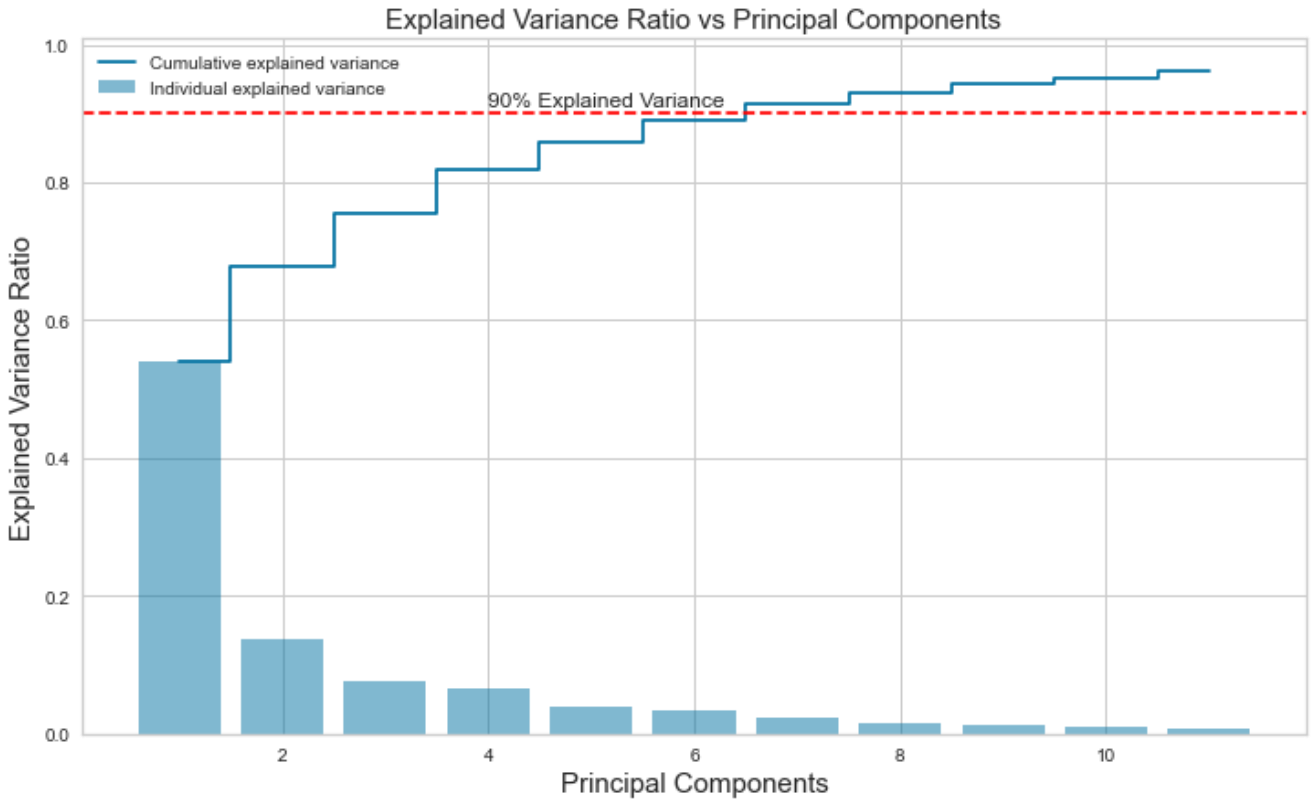

Figure 41
Cumulative sum plot of PC

Looking at the graph we can clearly state that for a minimum of 90% explained variance we require at least 7 principal components.

Visualizing how each Principal component interacts/varies with respect to the variables (refer to notebook for a clearer image):

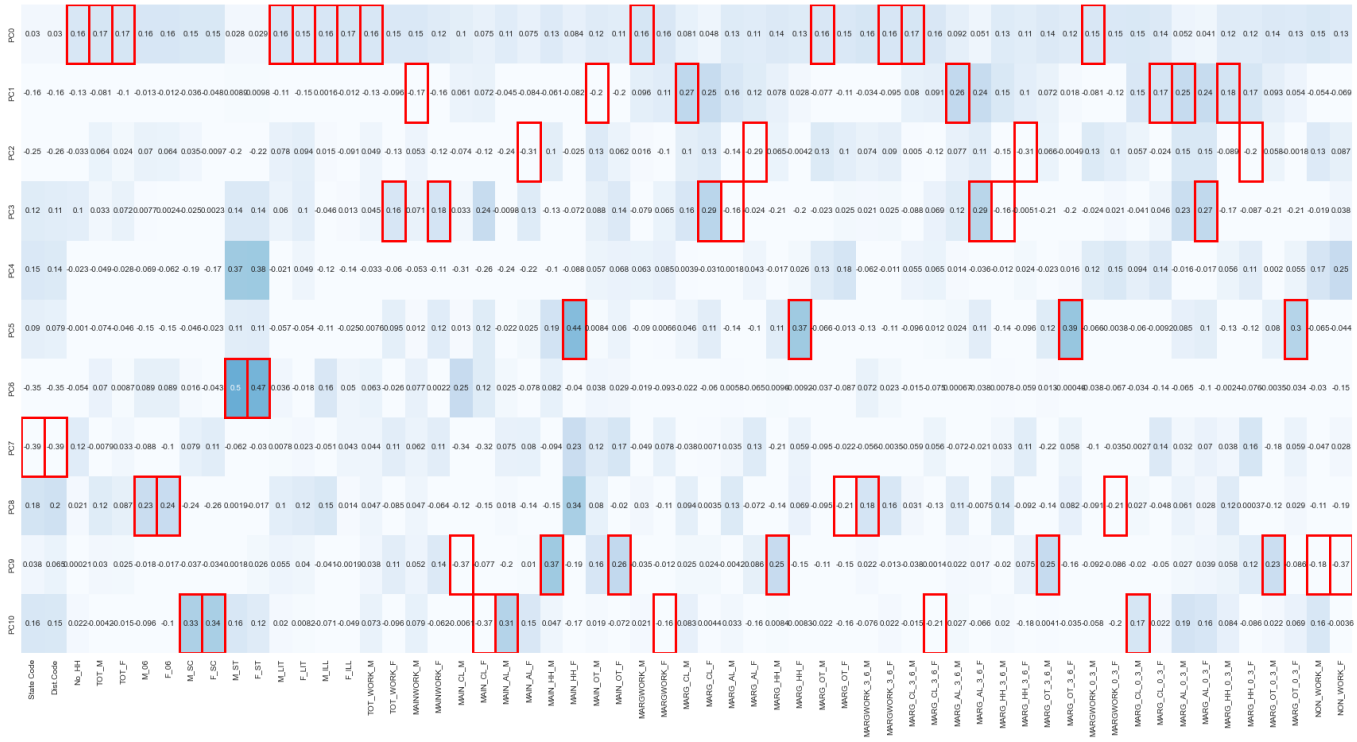

Figure 42

Rectangular plot of all PC with respect to all the variables

The Rectangular plot can be summarized as follows:

- Variables related to literacy are mostly loaded on principal component PC0 along with parameters related to total population.

- Population parameters related to Scheduled tribes are very heavily loaded on principal component PC6.

- Population parameters related to Scheduled caste are very heavily loaded on principal component PC10.

The Linear Equation for the first Principal Component can be written as:

0.03 * State Code 0.03 * Dist.Code 0.156 * No_HH 0.167 * TOT_M 0.166 * TOT_F 0.162 * M_06 0.162 * F_06 0.151 * M_SC 0.151 * F_SC 0.028 * M_ST 0.029 * F_ST 0.162 * M_LIT 0.147 * F_LIT 0.161 * M_ILL 0.165 * F_ILL 0.16 * TOT_WORK_M 0.146 * TOT_WORK_F 0.146 * MAINWORK_M 0.125 * MAINWORK_F 0.103 * MAIN_CL_M 0.075 * MAIN_CL_F 0.114 * MAIN_AL_M 0.075 * MAIN_AL_F 0.131 * MAIN_HH_M 0.084 * MAIN_HH_F 0.124 * MAIN_OT_M 0.111 * MAIN_OT_F 0.164 * MARGWORK_M 0.155 * MARGWORK_F 0.081 * MARG_CL_M 0.048 * MARG_CL_F 0.128 * MARG_AL_M 0.114 * MARG_AL_F 0.14 * MARG_HH_M 0.127 * MARG_HH_F 0.155 * MARG_OT_M 0.147 * MARG_OT_F 0.165 * MARGWORK_3_6_M 0.161 * MARGWORK_3_6_F 0.165 * MARG_CL_3_6_M 0.156 * MARG_CL_3_6_F 0.092 * MARG_AL_3_6_M 0.051 * MARG_AL_3_6_F 0.128 * MARG_HH_3_6_M 0.111 * MARG_HH_3_6_F 0.139 * MARG_OT_3_6_M 0.124 * MARG_OT_3_6_F 0.154 * MARGWORK_0_3_M 0.146 * MARGWORK_0_3_F 0.149 * MARG_CL_0_3_M 0.14 * MARG_CL_0_3_F 0.052 * MARG_AL_0_3_M 0.041 * MARG_AL_0_3_F 0.121 * MARG_HH_0_3_M 0.116 * MARG_HH_0_3_F 0.139 * MARG_OT_0_3_M 0.132 * MARG_OT_0_3_F 0.15 * NON_WORK_M 0.131 * NON_WORK_F