DATA MINING PROJECT BUSINESS REPORT

DSBA

Written by

Priyamvada Singh

Dated: **29-01-2023**

(Format: dd-mm-yyyy)

Contents

Part 1

Clustering:

Digital Ads Data:

The ads24x7 is a Digital Marketing company which has now got seed funding of 10 Million Dollars. 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.

1. **Perform the following in given order:**

A. Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc. What is the important technical information about the dataset that a database administrator would be interested in? (Hint: Information about the size of the dataset and the nature of the variables)

- Basic analysis on the given data set:

  i. the first five rows of the dataset:

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

Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
325	323	1	0.0	0.35	0.0	0.0031	0.0	0.0
285	285	1	0.0	0.35	0.0	0.0035	0.0	0.0
356	355	1	0.0	0.35	0.0	0.0028	0.0	0.0
497	495	1	0.0	0.35	0.0	0.0020	0.0	0.0
242	242	1	0.0	0.35	0.0	0.0041	0.0	0.0

ii. the last five rows of the data set:

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

Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
1	1	0.07	0.35	0.0455	NaN	NaN	NaN
2	1	0.04	0.35	0.0260	NaN	NaN	NaN
1	1	0.05	0.35	0.0325	NaN	NaN	NaN
1	1	0.07	0.35	0.0455	NaN	NaN	NaN
2	1	0.09	0.35	0.0585	NaN	NaN	NaN

iii. total number of rows and columns (features) present:

- *There are 19 columns and 23066 rows in the data set.*

iv. datatype of each feature, number of null values, duplicated records:

- *There are six float64, seven int64, and six object data type variables.*
- *That means,* ***13 variables are numeric while 6 are non-numeric****.*

v. The following is the description of each numeric variable:

	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

- *The above information provides the mean, standard deviation, minimum, 25%, 50% (median), 75% and maximum data point values for each variable.*

vi. there are no duplicate values found in the data set.

B. Treat missing values in CPC, CTR and CPM using the formula given. You may refer to the Bank_KMeans Solution File to understand the coding behind treating the missing values using a specific formula. You have to basically create a user defined function and then call the function for imputing.

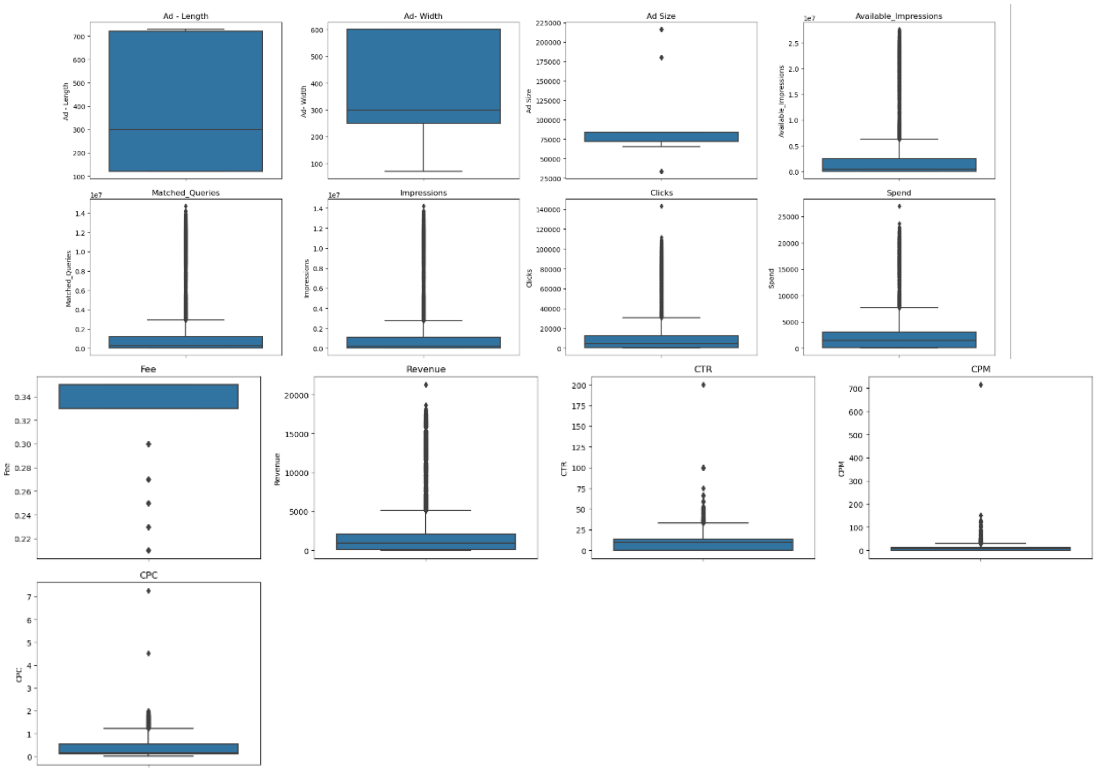
- *There were 4736 null values found in the columns CPC, CTR, and CPM.*

```
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
```

- *These missing values were imputed with the help of the formula given for CTR, CPM and CPC.*

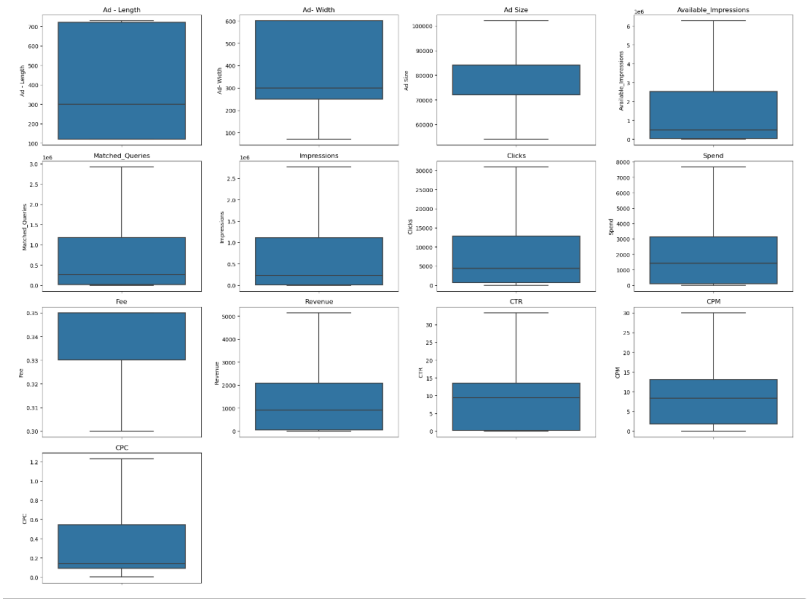
C. 3. Check if there are any outliers.

- *Yes, there are a significant number of outliers in the data set for each variable except Ad Length and Ad Width as shown below:*



D. 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).

- *Yes, outlier treatment is necessary for k-means clustering as it uses averaging of the observations in a cluster, i.e., finding centroids over and over till the clusters have stabilized and no movement of the point takes place between clusters.*
- *Here, I have treated outliers with the IQR method where the 25 and 75 quantile values are found as Q1 and Q3, respectively.*
- *With Q1 and Q3, I have then calculated the Interquartile Range or IQR, which is equal to Q3-Q1.*
- *Finally, I have used the formula Q1 – (1.5*IQR) and Q3 + (1.5*IQR) to find the lower range and upper range of the columns.*
- *These values are then used to impute wherever the value in a column exceeds the lower and the upper limit. If the values are below the lower limit, they get imputed with Q1 – (1.5*IQR) and if the values are above the upper limit, they are imputed with Q3 + (1.5*IQR).*
- *After performing outlier treatment, the following is the result:*



- *All the outliers in the data set have been treated successfully.*

E. Perform z-score scaling and discuss how it affects the speed of the algorithm.

- *After scaling the data set with z-score, the data set now looks like this:*

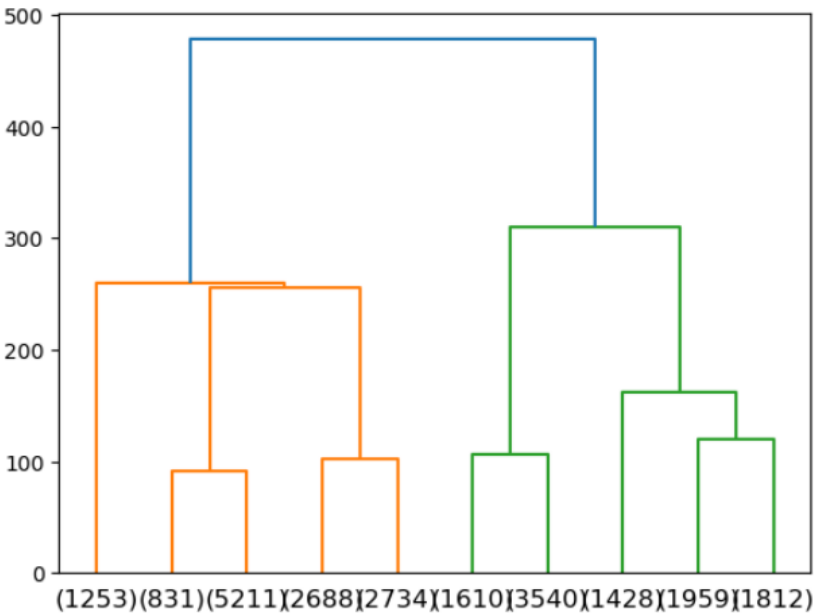
	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	-4.030447e-15	1.000022	-1.134891	-1.134891	-0.364496	1.433093	1.467332
Ad- Width	23066.0	5.390161e-15	1.000022	-1.319110	-0.432797	-0.186599	1.290590	1.290590
Ad Size	23066.0	-4.156304e-15	1.000022	-1.467840	-0.297564	-0.297564	0.482620	1.652896
Available_Impressions	23066.0	-3.617510e-15	1.000022	-0.756182	-0.740341	-0.528577	0.433059	2.193158
Matched_Queries	23066.0	1.341008e-15	1.000022	-0.779265	-0.761447	-0.527722	0.371498	2.070914
Impressions	23066.0	-1.224345e-15	1.000022	-0.768806	-0.760655	-0.538975	0.366051	2.056111
Clicks	23066.0	1.960656e-15	1.000022	-0.867488	-0.793438	-0.405431	0.468629	2.361729
Spend	23066.0	1.250852e-15	1.000022	-0.893170	-0.858046	-0.305523	0.393932	2.271900
Fee	23066.0	-2.322121e-14	1.000022	-2.222416	-0.567532	0.535724	0.535724	0.535724
Revenue	23066.0	3.136228e-15	1.000022	-0.880093	-0.846474	-0.317607	0.389803	2.244218
CTR	23066.0	1.329072e-15	1.000022	-0.995031	-0.964227	0.141524	0.635787	3.035808
CPM	23066.0	5.791296e-17	1.000022	-1.194498	-0.940303	0.022146	0.700905	3.162718
CPC	23066.0	1.987283e-15	1.000022	-1.042561	-0.759091	-0.602371	0.682987	2.846105

- *The minimum and maximum value is contained within the scale of -4 to +4.*
- *Yes, z-score standardization helps increase the speed of the algorithm. This is because it contains the value range of all dataset features from -4 to +4. This significantly decreases the processing required by the algorithm when compared to larger values that may be in hundreds or thousands before scaling.*
- *Z-score standardization also makes all the features well suited to each other for further comparison, in terms of distance and averages while clustering.*
- *If the value-range for each feature differs, the accuracy of the algorithm is compromised.*
- *Most importantly, by scaling, we are compressing the data into a smaller scale, which reduces the impact of larger values in the process and makes the algorithm faster.*

2. Perform clustering and do the following:

A. Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.

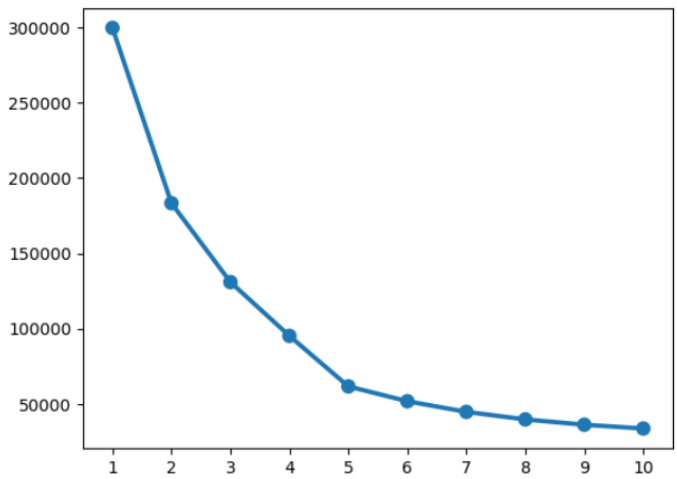
- *These are the last 10 clusters after performing hierarchical clustering using dendrogram:*



- *From the above, I have cut the cluster at 200 on the x-axis. This has left us with 5 clusters of the data set.*

B. Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm. Give justification along with presenting metrics/charts used for arriving at the conclusions.

- *Here is the Elbow plot for n=10:*



- *The optimum number of clusters in this scenario would be 5, since **the drop in Within Clusters Sum of Squares (WCSS) from cluster 4 to cluster 5 is 33594** but the drop of WCSS from cluster 5 to cluster 6 is relatively very low, i.e., 9862. According to me, 5 would be an optimum choice of k-value for k-means clustering.*

- *Below is the math performed on the WCSS scores of multiple clusters to observe the drop:*

```
299858.0000000002 - 183349.1020288608
### WCSS cluster1 - WCSS cluster 2
```

```
116508.89797113938
```

```
183349.1020288608 - 130878.34240367355
### WCSS cluster2 - WCSS cluster 3
```

```
52470.759625187246
```

```
130878.34240367355 - 95133.94481349865
### WCSS cluster3 - WCSS cluster 4
```

```
35744.39759017491
```

```
95133.94481349865 - 61539.18919785385
### WCSS cluster4 - WCSS cluster 5
```

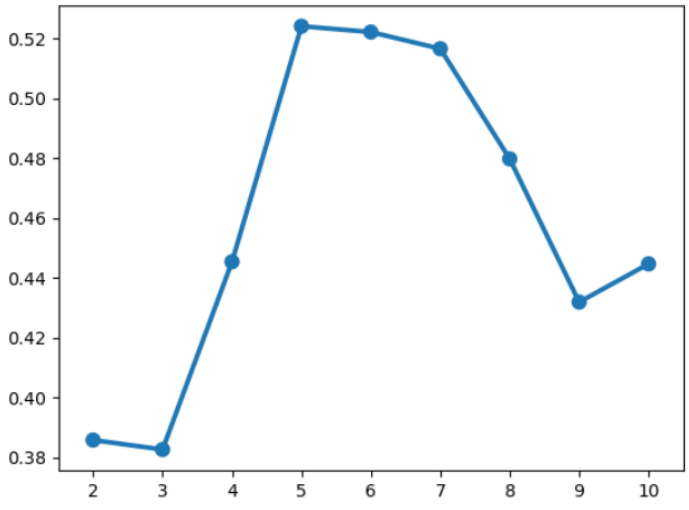
```
33594.7556156448
```

```
61539.18919785385 - 51676.896816004584
### WCSS cluster5 - WCSS cluster 6

### the drop from 5th to 6th cluster has significantly decreased
### therefore, we can take 5 as k for k-means clustering
```

C. Print silhouette scores for up to 10 clusters and identify optimum number of clusters.

- *The following Elbow plot shows the silhouette score for each k-value from 2 to 10:*



- *Here, the silhouette scores tell us that the optimum number of k-means clusters will be 5.*
- *This can be seen due to the value 5 having the highest silhouette score.*
- *The closer the silhouette score to +1, better the segmentation of clusters.*

D. 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.]

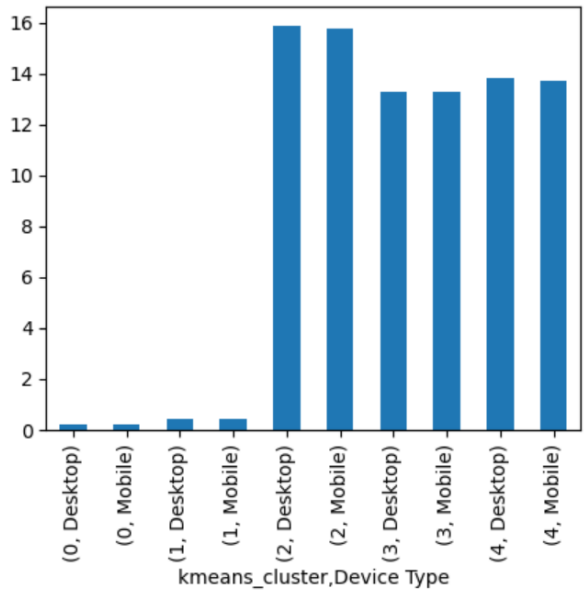
- *There are 5 clusters formed after performing k-means clustering. The records falling under each cluster are shown below:*

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

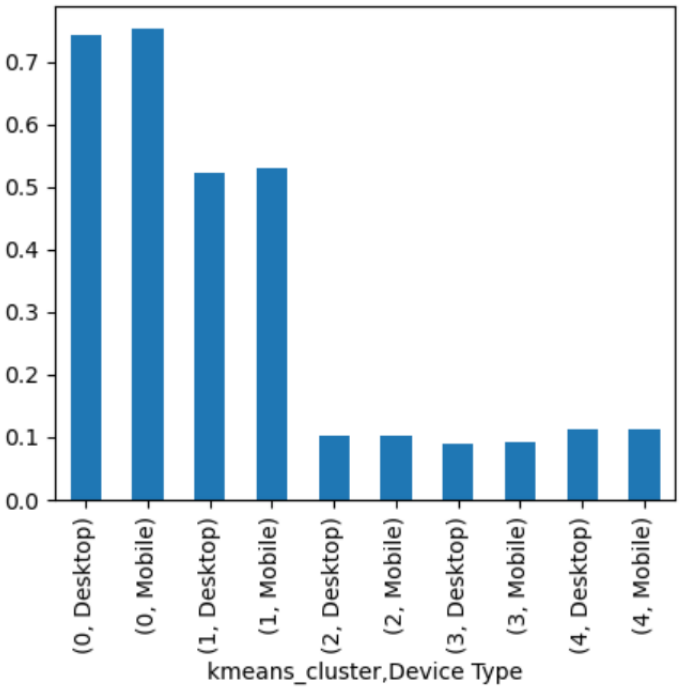
- *The final silhouette score for the clusters formed is 0.524, which is a +ve number. This indicates that the clusters were segmented properly.*
- *The minimum silhouette width for the samples is -0.037, which is below 0, indicating that some data points might not have been assigned their most-suited cluster.*
- *However, on investigating further, the total number of observations with silhouette width below 0 is 54. This is just 0.002% of the total number of observations and hence, can be neglected.*

- *The following are the observations when segmented by Device Type for each cluster:*

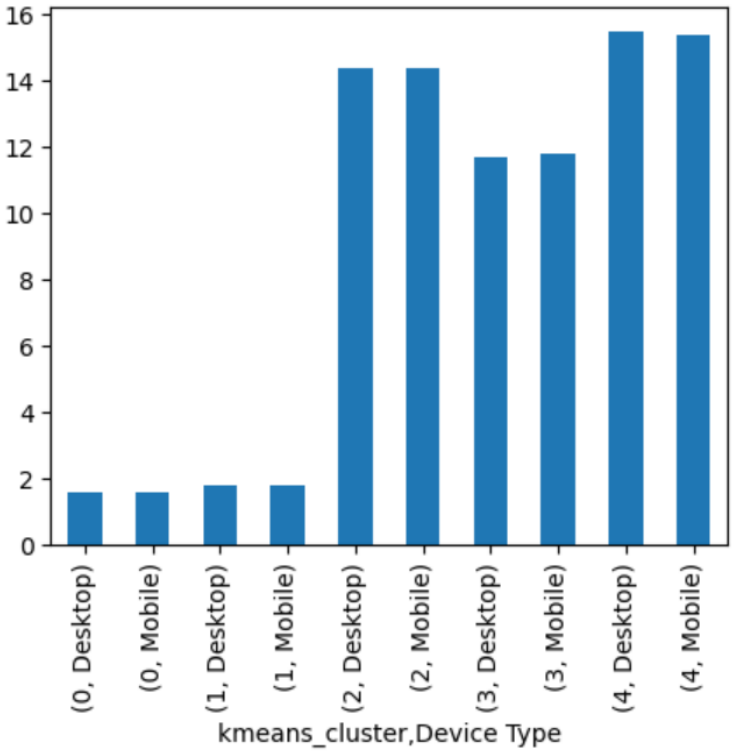
a. Average CTR



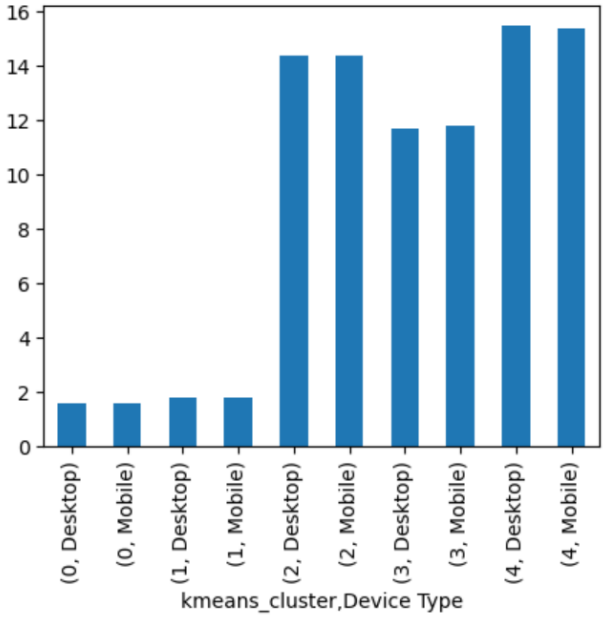
b. Average CPC



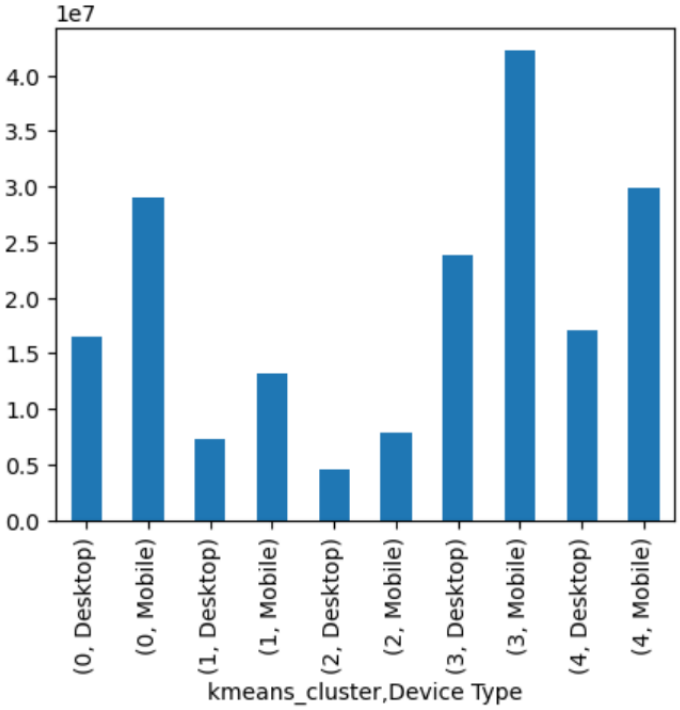
c. Average CPM



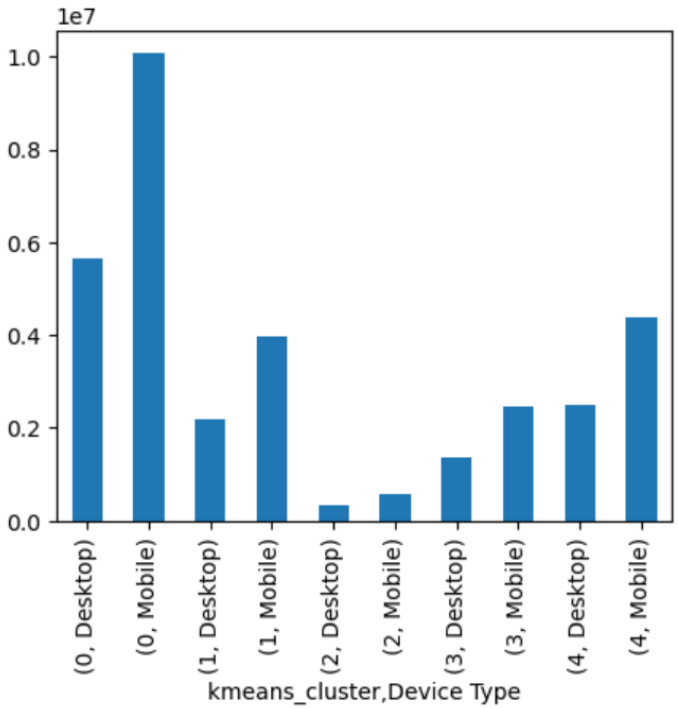
d. Total Spend



e. Total Clicks



f. Total Revenue



E. Conclude the project by providing summary of your learnings.

- *The average CTR is the highest for cluster-2, followed by cluster-4 and cluster-3.*
- *For cluster-0 and cluster-1, the ad creative and copy need work as the ad content can directly impact CTR in a positive way.*
- *The average CPC is the lowest for cluster-3, cluster-2 and cluster-4.*
- *This is a good indication that these clusters are performing well in terms of click-through rate at a relatively low cost.*
- *Cluster-0 and cluster-1 are giving very low CTRs at a relatively high cost per click.*
- *This could become a potential for the ad campaign, if the ads and targeting in these clusters is not optimised.*
- *Cluster-4 has the highest CPM, followed by cluster-2 and cluster-3.*
- *To lower the CPC further for the cluster-2, 3 and 4, the company can fine-tune its audience targeting. For example, if cluster-2 belongs to a specific geography, then we know that within that geography, we can better set the demographic targeting to lower CPM.*
- *This will serve relevant ads to relevant people within the geography, increasing clicks even more, and lowering CPC and CPM further.*
- *The lowest budget was spent by cluster-2, followed by cluster-3.*
- *Spends allocated to cluster-0 should be lower since it's giving high cost per click with very low click through rate. The budget can be increased for cluster-2 so it spends more, since its performing well on lower cost.*
- *The total number of clicks are the lowest for cluster-2. This could be because the spends are the lowest for cluster-2.*

- *Cluster-1 is generating the maximum amount of revenue. It is the most integral for revenue generation.*

Part 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.

1. **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.**

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

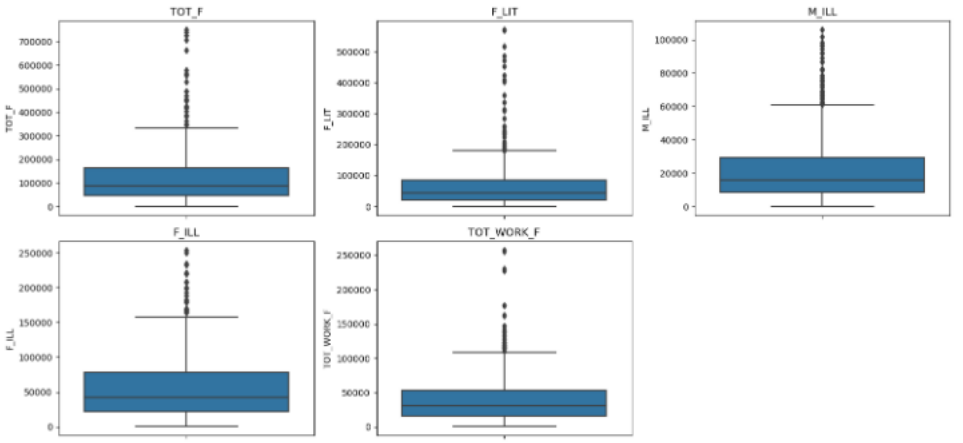
i. Basic checks performed on the data set with the following notes:

- *The data set has 640 records with 61 variables/features.*
- *It has 59 int64 and 2 object data type variables.*
- *There are no null and duplicate values in the data set.*

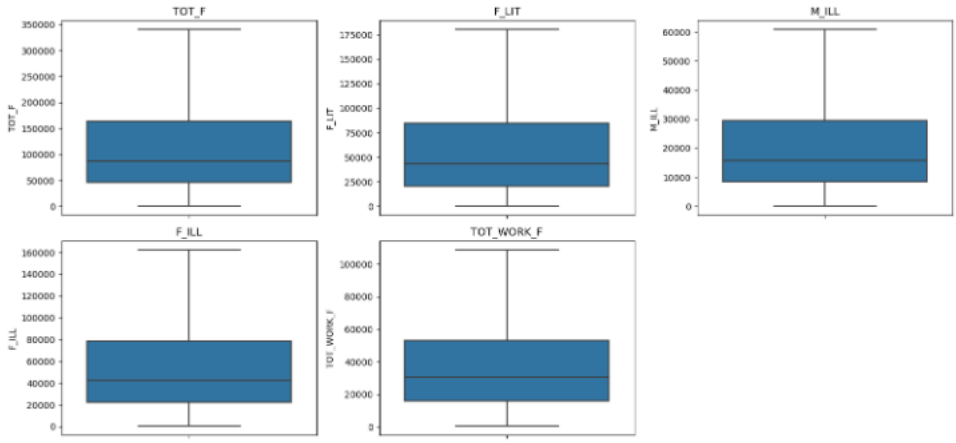
B. Perform detailed Exploratory analysis by creating certain questions like (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio? (Example Questions). Pick 5 variables out of the given 24 variables below for EDA: No_HH, TOT_M, TOT_F, 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. EDA Insights: The 5 variables chosen are TOT_F (Total female population), F_LIT (Literate female population), M_ILL (Illiterate male population), F_ILL (Illiterate female population), TOT_WORK_F (Total Working female population).

a. The outliers for these variables need treatment:



b. Outliers treated with the IQR method where the 25 and 75 quantile values are found as Q1 and Q3, respectively.
With Q1 and Q3, I have then calculated the Interquartile Range or IQR, which is equal to Q3-Q1.
Finally, I have used the formula Q1 – (1.5*IQR) and Q3 + (1.5*IQR) to find the lower range and upper range of the columns.
These values are then used to impute wherever the value in a column exceeds the lower and the upper limit. If the values are below the lower limit, they get imputed with Q1 – (1.5*IQR) and if the values are above the upper limit, they are imputed with Q3 + (1.5*IQR). The new distribution of the variables is shown below:



c. How much percentage of the total female population is working?
- Answer: 34% of the total females are working

d. Is Literate female population outnumbering Illiterate female population in female-headed households?

- Answer: YES, but there is only a minor gap between literate and illiterate females, with literate females leading by 4%.

e. Is there a major difference between the population of Illiterate female and males?
- Answer: On an average, for every 38 illiterate males, there are 100 illiterate females. The illiteracy rate amongst females is much higher than amongst males.

f. How much of the Total Female population is Illiterate?
- Answer: 47% of the total females are illiterate.

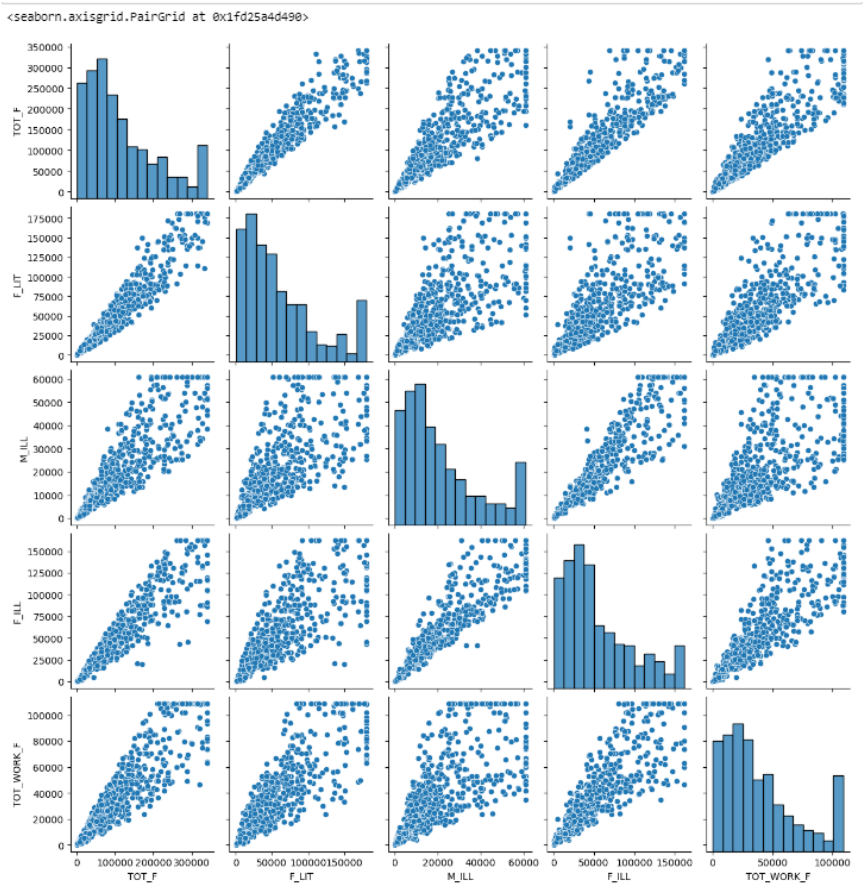
g. What is the literacy rate in the female population?
- Answer: 51% literacy rate.

h. Is the correlation between Literate and Working females stronger than that between Illiterate and Working females?
- Answer: No. There seems to be a strong correlation between illiterate and working females (0.87), which means that even females who are illiterate are able to contribute to the working female population.

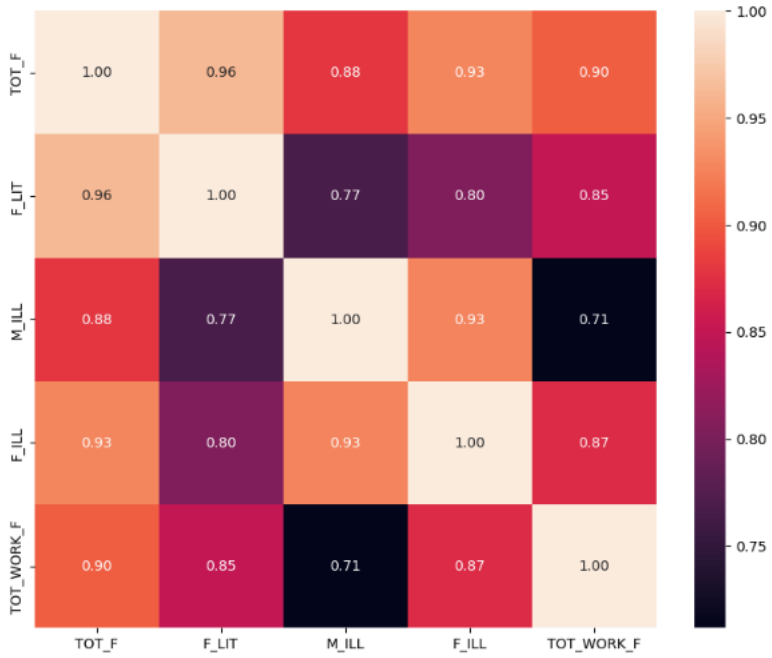
	TOT_F	F_LIT	M_ILL	F_ILL	TOT_WORK_F
TOT_F	1.00	0.96	0.88	0.93	0.90
F_LIT	0.96	1.00	0.77	0.80	0.85
M_ILL	0.88	0.77	1.00	0.93	0.71
F_ILL	0.93	0.80	0.93	1.00	0.87
TOT_WORK_F	0.90	0.85	0.71	0.87	1.00

i. Pair plot for the 5 chosen variables:

j.



k. Heatmap of the give variables:



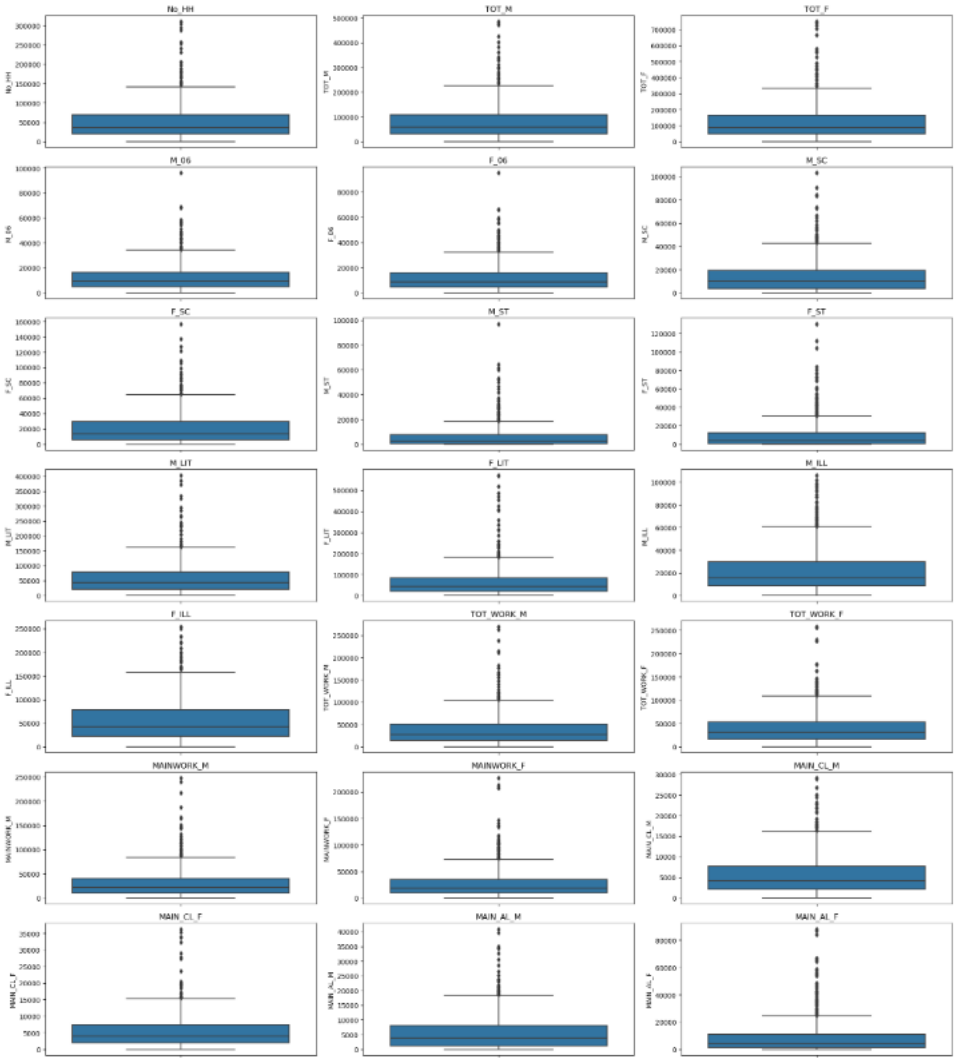
C. PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?

- Yes, treating outliers for PCA is necessary since this technique is sensitive to outliers and may give inaccurate results if the outliers are left untreated.
- After dropping columns ‘State Code’, ‘District Code’, ‘State’, ‘Area Name’, we are left with 57 numerical features, all of which have outliers that need treatment. Even after scaling with z-score, the outliers did not show any difference. Hence, it is better to treat outliers for performing PCA.

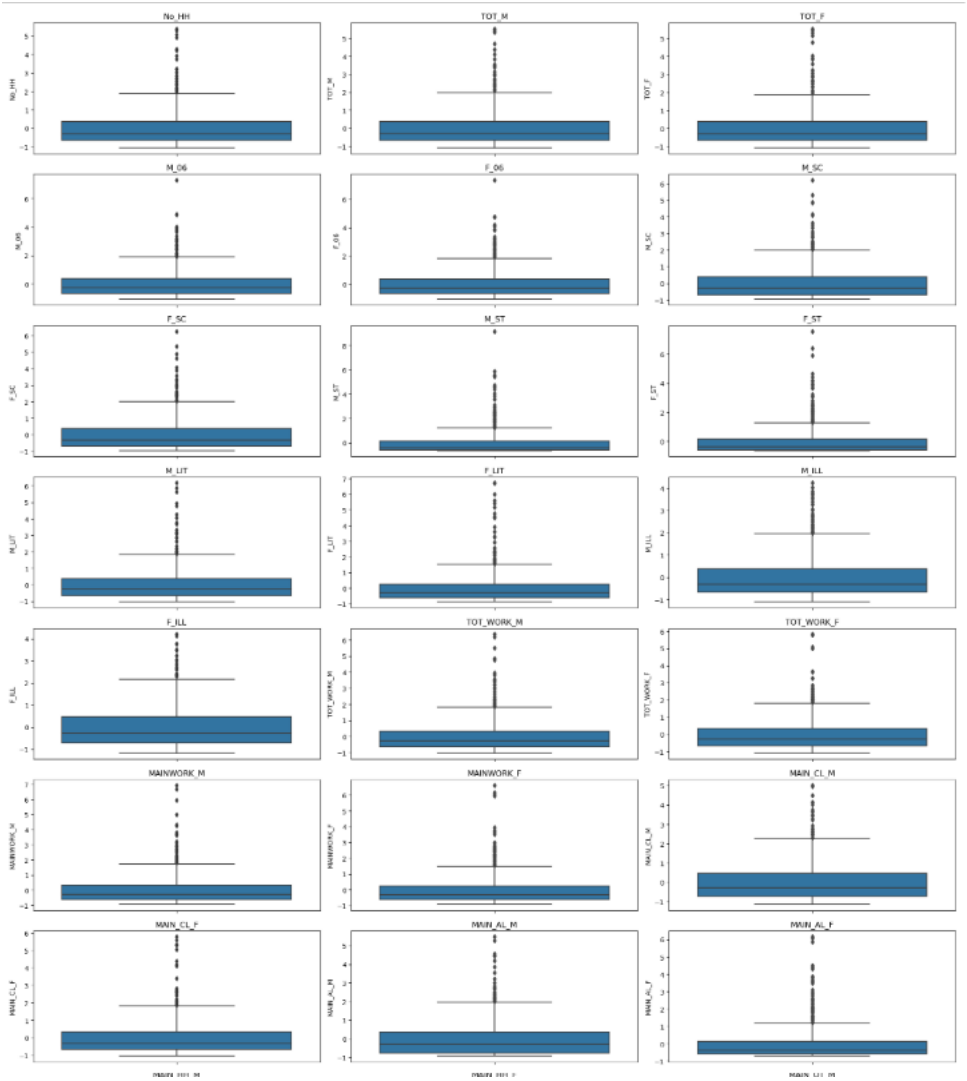
D. Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.

- No, scaling with z-score method did not have any impact on outliers.
-

  - Sample snapshot of boxplot before scaling:



- Sample snapshot of boxplots after scaling:



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

- Kmo = 0.93, which means the data set size is adequate.
- The p-value = 0.0, which means that the correlations are significant and we can go ahead with performing PCA.
- Eigenvectors for PCs after performing PCA:

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

- Eigenvalues of all 57 PCs:

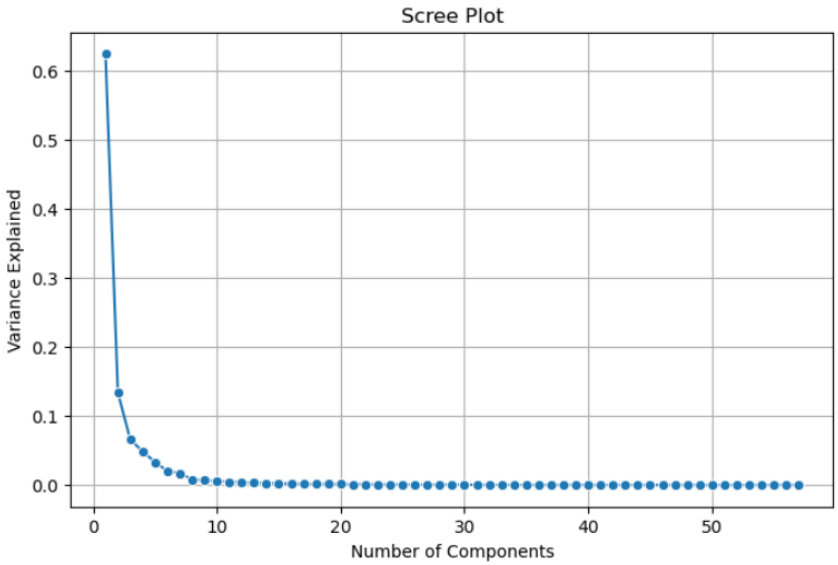
```
array([3.56488638e+01, 7.64357559e+00, 3.76919551e+00, 2.77722349e+00,
       1.90694892e+00, 1.15490310e+00, 9.87726707e-01, 4.64629906e-01,
       3.96708513e-01, 3.22346888e-01, 2.73207369e-01, 2.35647574e-01,
       1.81401107e-01, 1.69243770e-01, 1.38592325e-01, 1.31505852e-01,
       1.03809666e-01, 9.55333831e-02, 8.58580407e-02, 8.09138742e-02,
       6.60179067e-02, 6.30797999e-02, 4.82756124e-02, 4.59506197e-02,
       4.37747566e-02, 3.19339710e-02, 2.86194563e-02, 2.75481445e-02,
       2.34340044e-02, 2.20296816e-02, 1.87487040e-02, 1.59004895e-02,
       1.39957919e-02, 1.18916465e-02, 1.11133495e-02, 9.07842645e-03,
       7.25127869e-03, 6.27213692e-03, 4.95541908e-03, 4.60667097e-03,
       3.45902033e-03, 2.18408510e-03, 2.13514664e-03, 1.92111328e-03,
       1.43840980e-03, 1.09968912e-03, 9.65752052e-04, 8.62630267e-04,
       6.51634478e-04, 5.76658846e-04, 4.35790607e-04, 3.70037468e-04,
       3.06660171e-04, 2.07854170e-04, 1.38286484e-04, 8.97034441e-05,
       4.61745385e-05])
```

- Covariance matrix for PCs:

```
array([6.24441446e-01, 1.33888289e-01, 6.60229147e-02, 4.86470891e-02,
       3.34029704e-02, 2.02297994e-02, 1.73014629e-02, 8.13866529e-03,
       6.94892379e-03, 5.64637229e-03, 4.78562250e-03, 4.12770833e-03,
       3.17750294e-03, 2.96454958e-03, 2.42764517e-03, 2.30351534e-03,
       1.81837655e-03, 1.67340548e-03, 1.50392785e-03, 1.41732362e-03,
       1.15639919e-03, 1.10493400e-03, 8.45617224e-04, 8.04891611e-04,
       7.66778221e-04, 5.59369722e-04, 5.01311201e-04, 4.82545623e-04,
       4.10480504e-04, 3.85881758e-04, 3.28410688e-04, 2.78520087e-04,
       2.45156553e-04, 2.08299401e-04, 1.94666401e-04, 1.59021779e-04,
       1.27016642e-04, 1.09865556e-04, 8.68013375e-05, 8.06925096e-05,
       6.05897475e-05, 3.82574118e-05, 3.74001838e-05, 3.36510796e-05,
       2.51958296e-05, 1.92626466e-05, 1.69165450e-05, 1.51102177e-05,
       1.14143210e-05, 1.01010143e-05, 7.63350323e-06, 6.48174183e-06,
       5.37159674e-06, 3.64086663e-06, 2.42228792e-06, 1.57128566e-06,
       8.08813873e-07])
```

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

- Scree plot for PCs to identify the optimum number:



- Cumulative explained variance ratio matrix:

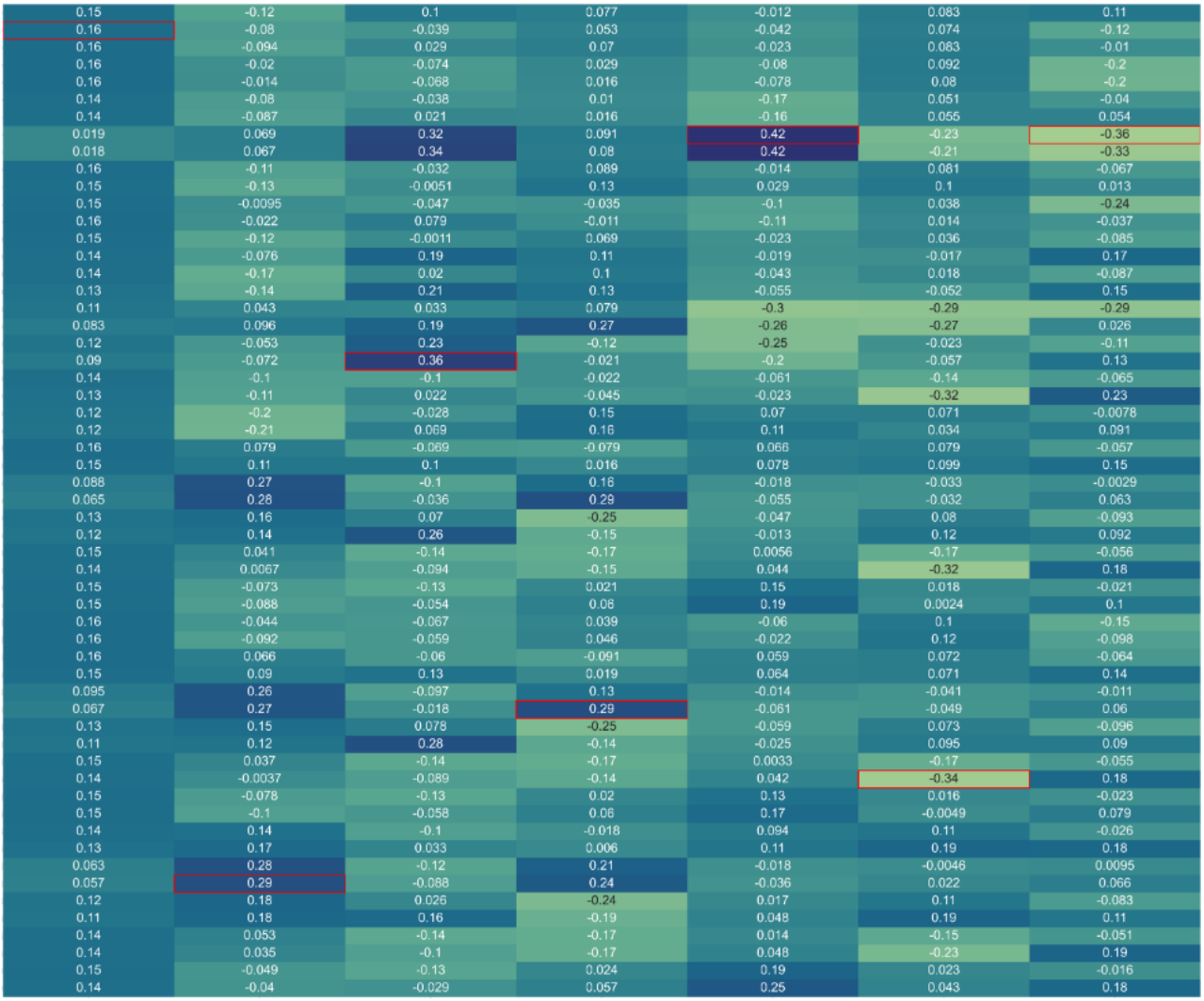
```
array([0.62444145, 0.75832974, 0.82435265, 0.87299974, 0.90640271,
       0.92663251, 0.94393397, 0.95207264, 0.95902156, 0.96466793,
       0.96945356, 0.97358126, 0.97675877, 0.97972332, 0.98215096,
       0.98445448, 0.98627285, 0.98794626, 0.98945019, 0.99086751,
       0.99202391, 0.99312884,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

- Taking 7 principal components. The value of 7 is picked with the help of the scree plot and elbow method, where the variance eigen values flatten after the 7th PC, not adding much value to the data.

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

- Here is the heatmap showing the impact of each feature on the 7 PCs through loadings:

PLEASE SCROLL FURTHER



0.15	-0.12	0.1	0.077	-0.012	0.083	0.11
0.16	-0.08	-0.039	0.053	-0.042	0.074	-0.12
0.16	-0.094	0.029	0.07	-0.023	0.083	-0.01
0.16	-0.02	-0.074	0.029	-0.08	0.092	-0.2
0.16	-0.014	-0.068	0.016	-0.078	0.08	-0.2
0.14	-0.08	-0.038	0.01	-0.17	0.051	-0.04
0.14	-0.087	0.021	0.016	-0.16	0.055	0.054
0.019	0.069	0.32	0.091	0.42	-0.23	-0.36
0.018	0.067	0.34	0.08	0.42	-0.21	-0.33
0.16	-0.11	-0.032	0.089	-0.014	0.081	-0.067
0.15	-0.13	-0.0051	0.13	0.029	0.1	0.013
0.15	-0.0095	-0.047	-0.035	-0.1	0.038	-0.24
0.16	-0.022	0.079	-0.011	-0.11	0.014	-0.037
0.15	-0.12	-0.0011	0.069	-0.023	0.036	-0.085
0.14	-0.076	0.19	0.11	-0.019	-0.017	0.17
0.14	-0.17	0.02	0.1	-0.043	0.018	-0.087
0.13	-0.14	0.21	0.13	-0.055	-0.052	0.15
0.11	0.043	0.033	0.079	-0.3	-0.29	-0.29
0.083	0.096	0.19	0.27	-0.26	-0.27	0.026
0.12	-0.053	0.23	-0.12	-0.25	-0.023	-0.11
0.09	-0.072	0.36	-0.021	-0.2	-0.057	0.13
0.14	-0.1	-0.1	-0.022	-0.061	-0.14	-0.065
0.13	-0.11	0.022	-0.045	-0.023	-0.32	0.23
0.12	-0.2	-0.028	0.15	0.07	0.071	-0.0078
0.12	-0.21	0.069	0.16	0.11	0.034	0.091
0.16	0.079	-0.089	-0.079	0.066	0.079	-0.057
0.15	0.11	0.1	0.016	0.078	0.099	0.15
0.088	0.27	-0.1	0.16	-0.018	-0.033	-0.0029
0.065	0.28	-0.036	0.29	-0.055	-0.032	0.063
0.13	0.16	0.07	-0.25	-0.047	0.08	-0.093
0.12	0.14	0.26	-0.15	-0.013	0.12	0.092
0.15	0.041	-0.14	-0.17	0.0056	-0.17	-0.056
0.14	0.0067	-0.094	-0.15	0.044	-0.32	0.18
0.15	-0.073	-0.13	0.021	0.15	0.018	-0.021
0.15	-0.088	-0.054	0.06	0.19	0.0024	0.1
0.16	-0.044	-0.067	0.039	-0.06	0.1	-0.15
0.16	-0.092	-0.059	0.046	-0.022	0.12	-0.098
0.16	0.066	-0.06	-0.091	0.059	0.072	-0.064
0.15	0.09	0.13	0.019	0.064	0.071	0.14
0.095	0.26	-0.097	0.13	-0.014	-0.041	-0.011
0.067	0.27	-0.018	0.29	-0.061	-0.049	0.06
0.13	0.15	0.078	-0.25	-0.059	0.073	-0.096
0.11	0.12	0.28	-0.14	-0.025	0.095	0.09
0.15	0.037	-0.14	-0.17	0.0033	-0.17	-0.055
0.14	-0.0037	-0.089	-0.14	0.042	-0.34	0.18
0.15	-0.078	-0.13	0.02	0.13	0.016	-0.023
0.15	-0.1	-0.058	0.06	0.17	-0.0049	0.079
0.14	0.14	-0.1	-0.018	0.094	0.11	-0.026
0.13	0.17	0.033	0.006	0.11	0.19	0.18
0.063	0.28	-0.12	0.21	-0.018	-0.0046	0.0095
0.057	0.29	-0.088	0.24	-0.036	0.022	0.066
0.12	0.18	0.026	-0.24	0.017	0.11	-0.083
0.11	0.18	0.16	-0.19	0.048	0.19	0.11
0.14	0.053	-0.14	-0.17	0.014	-0.15	-0.051
0.14	0.035	-0.1	-0.17	0.048	-0.23	0.19
0.15	-0.049	-0.13	0.024	0.19	0.023	-0.016
0.14	-0.04	-0.029	0.057	0.25	0.043	0.18

- The red rectangles show the maximum loading value for each feature across the PCs.
- The following 7 features are finalized after performing PCA on the data set which initially had 57 features:
  a. TOT_F - Total population Female
  b. M_ST - Scheduled Tribes population Male
  c. F_ST - Scheduled Tribes population Female
  d. MAIN_AL_F - Main Agricultural Labourers Population Female
  e. MARG_AL_3_6_F - Marginal Agriculture Labourers Population 3-6 Female
  f. MARG_OT_3_6_F - Marginal Other Workers Population Person 3-6 Female
  g. MARG_AL_0_3_F - Marginal Agriculture Labourers Population 0-3 Female

- Sample snap shot of the reduced dimensions:

	State	Area Name	TOT_F	MARG_AL_0_3_F	MAIN_AL_F	MARG_AL_3_6_F	F_ST	MARG_OT_3_6_F	M_ST
0	Jammu & Kashmir	Kupwara	29796.0	237.0	143.0	343.0	2598.0	198.0	1999.0
1	Jammu & Kashmir	Badgam	23102.0	229.0	108.0	432.0	517.0	449.0	427.0
2	Jammu & Kashmir	Leh(Ladakh)	10964.0	89.0	71.0	1161.0	9723.0	28.0	5806.0
3	Jammu & Kashmir	Kargil	4206.0	128.0	24.0	158.0	3968.0	33.0	2666.0
4	Jammu & Kashmir	Punch	29981.0	1043.0	237.0	1419.0	10843.0	214.0	7670.0

H. Write linear equation for first PC.

- The linear equation for the PCs is:

(0.15)*PC1 + (0.16)*PC2 + (0.16)*PC3 + (0.16)*PC4 + (0.16)*PC5 + (0.14)*PC6 + (0.14)*PC7