PROJECT
ON
DATA MINING

By SHAJIL FERNANDEZ

29-07-2023

Table of Contents:

Part 1

List of Figures:

List of Tables:

Part 1

Clustering:

Digital Ads Data:

The ads24x7 is a Digital Marketing company which has now got seed funding of $10 Million. They are expanding their wings in Marketing Analytics. They collected data from their Marketing Intelligence team and now wants you (their newly appointed data analyst) to segment type of ads based on the features provided. Use Clustering procedure to segment ads into homogeneous groups.

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

Ans:

- **Table 1: Top 5 rows of the dataset**

	Timestamp	InventoryType	Ad - Length	Ad-Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0	0.35	0.0	0.0031	0.0	0.0
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.0	0.35	0.0	0.0035	0.0	0.0
2	2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0	0.35	0.0	0.0028	0.0	0.0
3	2020-9-3-20	Format1	300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.0	0.35	0.0	0.0020	0.0	0.0
4	2020-9-4-15	Format1	300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.0	0.35	0.0	0.0041	0.0	0.0

- **Table 2: Bottom 5 rows of the dataset**

	Timestamp	InventoryType	Ad - Length	Ad-Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
23061	2020-9-13-7	Format5	720	300	216000	Inter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
23062	2020-11-2-7	Format5	720	300	216000	Inter224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.0260	NaN	NaN	NaN
23063	2020-9-14-22	Format5	720	300	216000	Inter218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	NaN	NaN	NaN
23064	2020-11-18-2	Format4	120	600	72000	inter230	Video	Mobile	Video	7	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
23065	2020-9-14-0	Format5	720	300	216000	Inter221	App	Mobile	Video	2	2	2	1	0.09	0.35	0.0585	NaN	NaN	NaN

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

- **Info**:

Table 3: Basic info of the dataset

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

Observations:

- There are 23066 rows and 19 columns.

- There are 6 float64, 7 int64 and 6 object datatypes.

- There are missing values for 3 columns (CTR, CPM, CPC).

- There are **no duplicate rows** in the given dataset.

- **Null values**:

Table 4: Table showing missing value info in the dataset

```
Timestamp                 0
InventoryType             0
Ad - Length               0
Ad- Width                 0
Ad Size                   0
Ad Type                   0
Platform                  0
Device Type               0
Format                    0
Available_Impressions     0
Matched_Queries           0
Impressions               0
Clicks                    0
Spend                     0
Fee                       0
Revenue                   0
CTR                    4736
CPM                    4736
CPC                    4736
dtype: int64
```

Observation:

- There are 3 columns (CTR, CPM, CPC) with 4736 missing values each.

- **Statistical summary of numerical and categorical data:**

Table 5: Table showing statistical summary of numerical and categorical data

	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

	count	unique	top	freq
Timestamp	23066	2018	2020-11-13-22	13
InventoryType	23066	7	Format4	7165
Ad Type	23066	14	Inter224	1658
Platform	23066	3	Video	9873
Device Type	23066	2	Mobile	14806
Format	23066	2	Video	11552

Observations:

- Ad-Length ranges from minimum of 120 to maximum of 728.

- Maximum width of Ad is 600 and minimum width is 70.

- inter224 is the most selected Ad type.

- Most of the ads are shown in video format and supports Mobile.

- User clicks range from 1 to 143049.

- Maximum revenue earned for an ad is 21276.18.

1.2 Clustering: Treat missing values in CPC, CTR and CPM using the formula given.

Ans:

We can treat missing values in CTR, CPM and CPC columns by making use of an **user defined function** and the formula given below, i.e.,

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

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

CPC = Total Cost (spend) / Number of Clicks

And then, by applying the **Lambda** function on the return value of user defined function.

After performing the above procedure, there are no missing values in the dataset.

Table 6: Treating missing values

Timestamp	0
InventoryType	0
Ad - Length	0
Ad- Width	0
Ad Size	0
Ad Type	0
Platform	0
Device Type	0
Format	0
Available_Impressions	0
Matched_Queries	0
Impressions	0
Clicks	0
Spend	0
Fee	0
Revenue	0
CTR	0
CPM	0
CPC	0
dtype: int64	

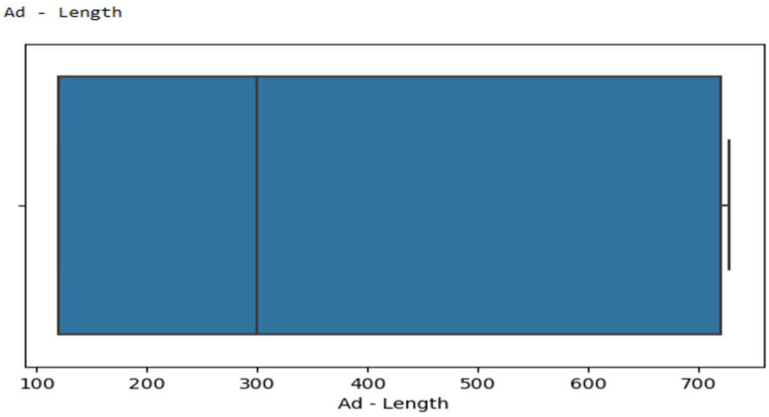
1.3 Clustering: Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an analyst your judgement may be different from another analyst).

Ans:

There are 13 numerical data in the given set. Boxplot of the same are as follows:

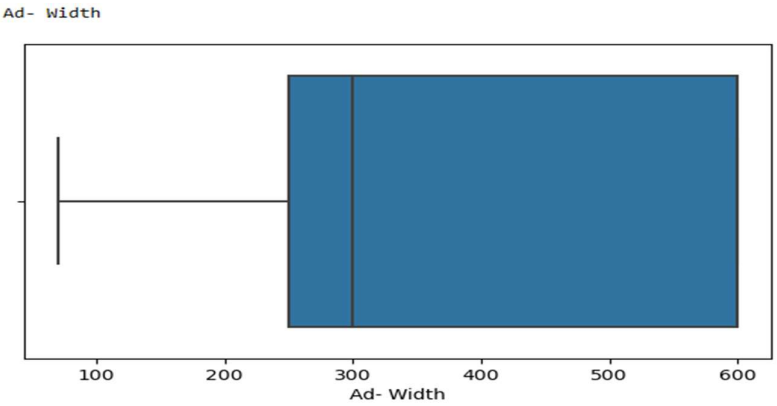
Figure 1: Boxplot on the dataset to detect outliers

i. Ad-Length:



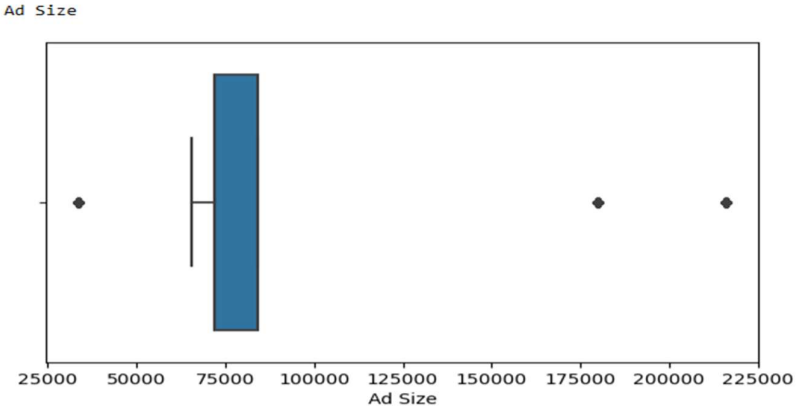
- There is no Outlier.

ii) Ad-Width:



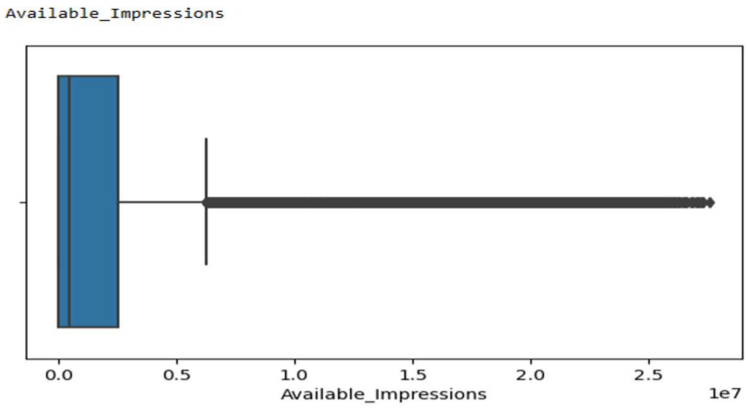
- There is no Outlier.

iii) Ad Size:



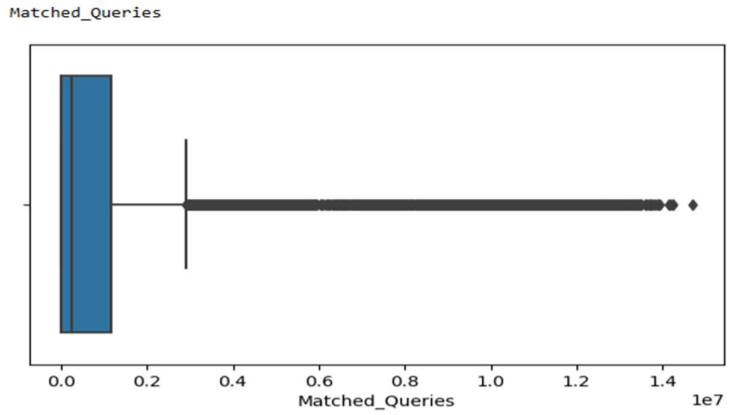
- There are Outliers in the given variable.

iv) Available Impressions:



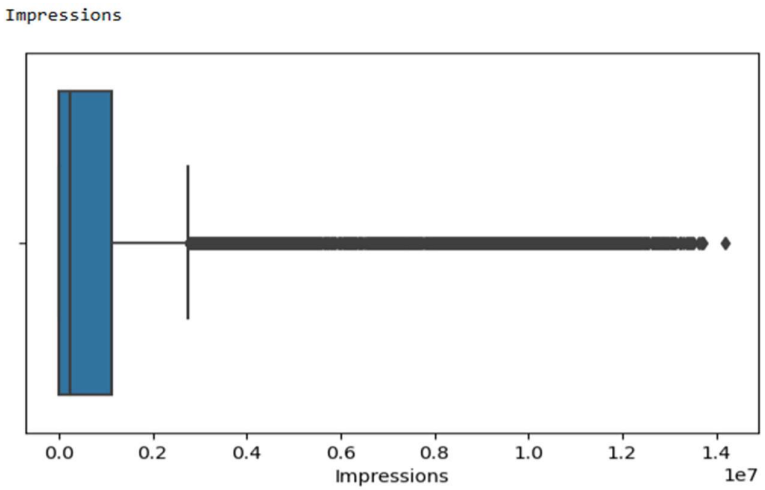
- There are Outliers in the given variable.

v) Matched Queries:



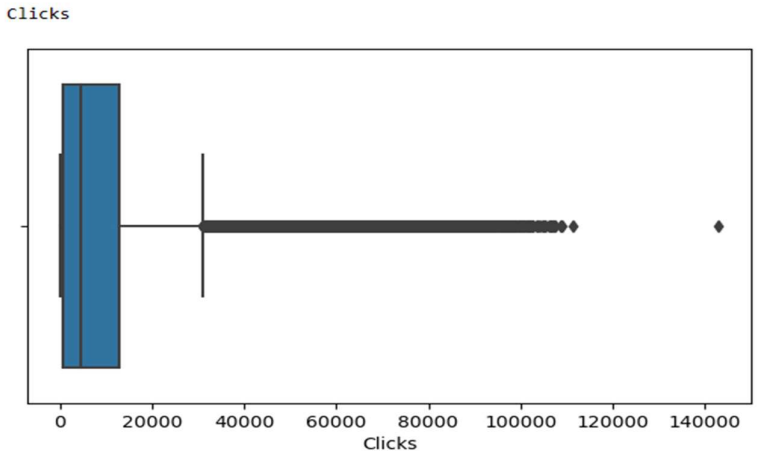
- There are Outliers in the given variable.

vi) Impressions:



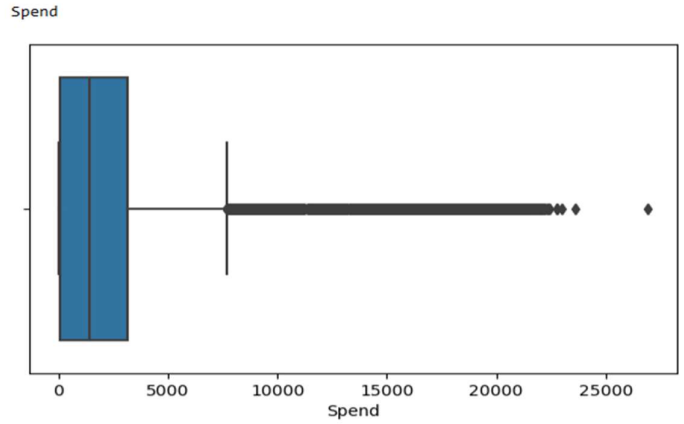
- There are Outliers in the given variable.

vii) Clicks:



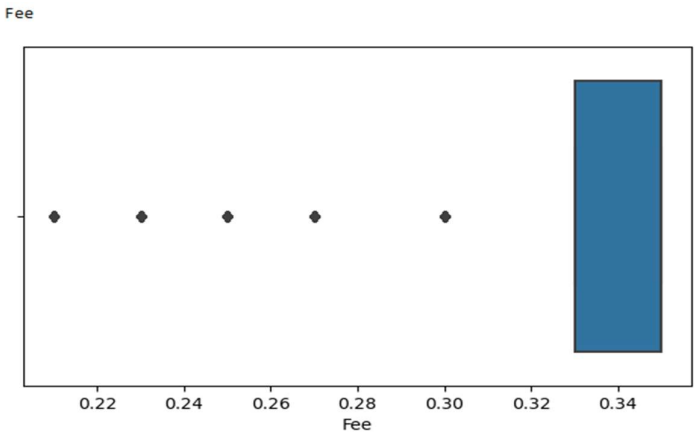
- There are Outliers in the given variable.

viii) Spend:



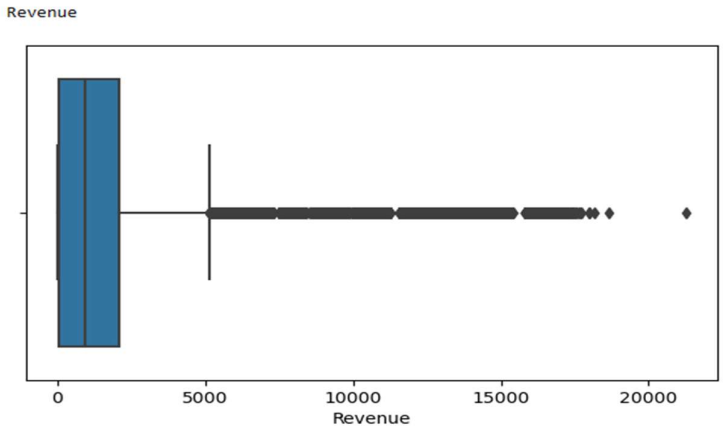
- There are Outliers in the given variable.

ix) Fee:



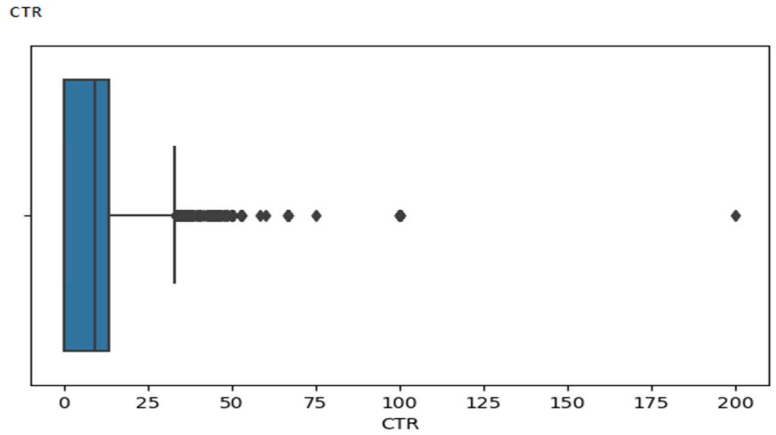
- There are Outliers in the given variable.

x) Revenue:



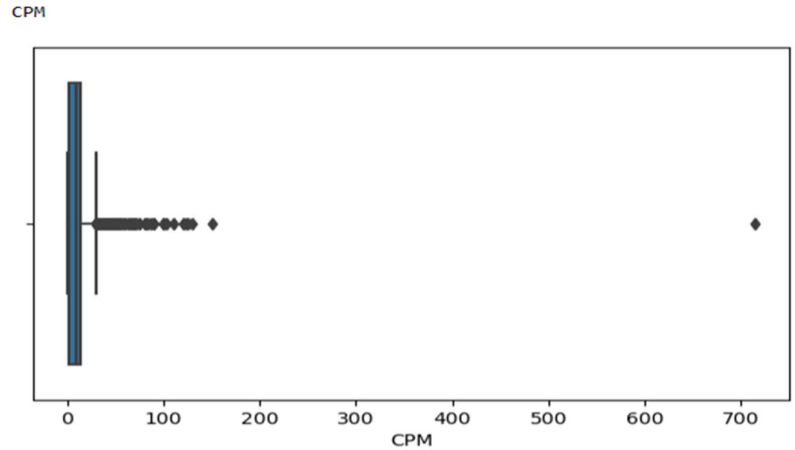
- There are Outliers in the given variable.

xi) CTR:



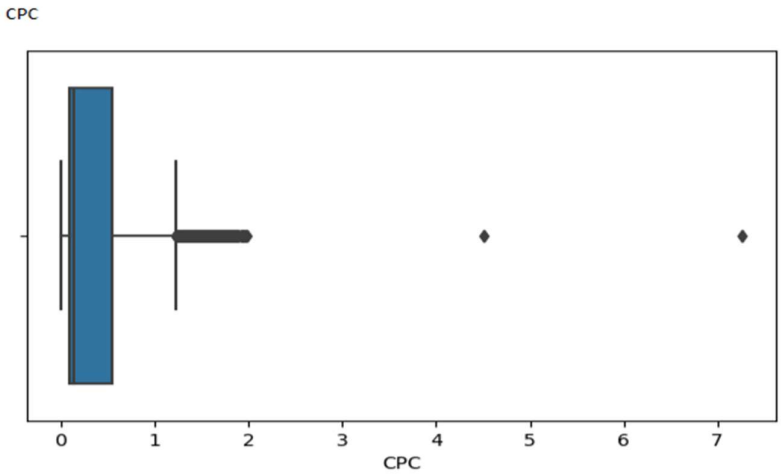
- There are Outliers in the given variable.

xii) CPM:



- There are Outliers in the given variable.

xiii) CPC:



- There are Outliers in the given variable.

As there are outliers, it is necessary to treat them in order to improve the accuracy and to avoid unnecessary movement of Median value of the given variable.

Treatment of outliers:

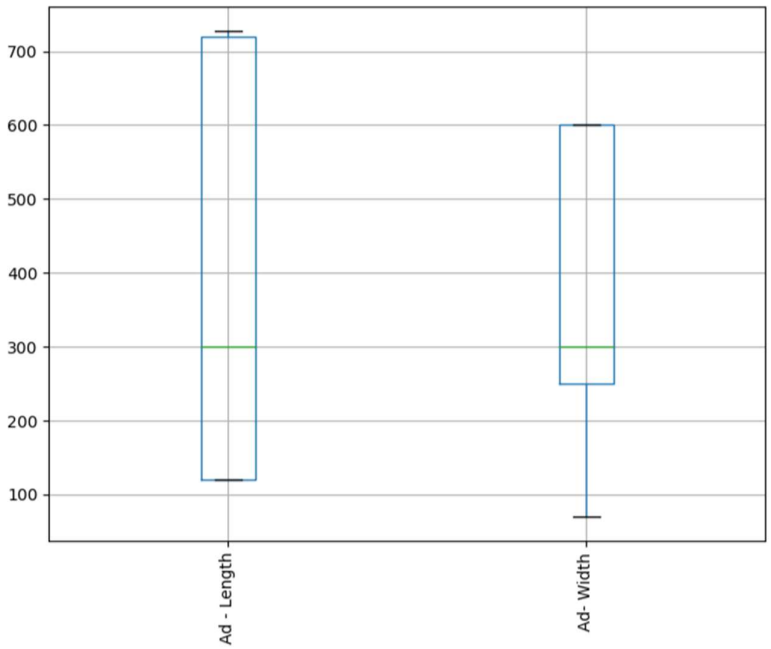
- Count of outliers:

Table 7: Table showing count of outliers

Ad - Length	0
Ad- Width	0
Ad Size	8448
Available_Impressions	2378
Matched_Queries	3192
Impressions	3269
Clicks	1691
Spend	2081
Fee	3517
Revenue	2325
CTR	275
CPM	207
CPC	585
dtype: int64	

There are only two variables without outliers as per the given data.

Figure 2: Boxplot showing the variables without outliers before treatment



We are considering **95th percentile threshold** to **treat the outliers**, so that we can cover more data. Higher outliers will be treated at 95th percentile and Lower outliers at 5 percentile value.

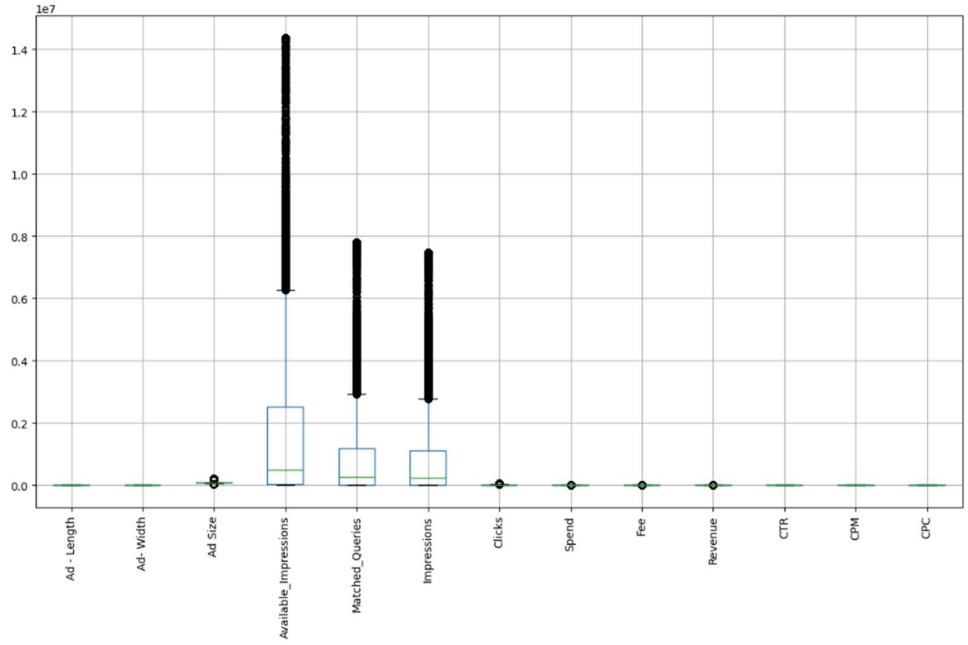
- Count of outliers at 95th percentile:

Table 8: Table showing count of outliers for quartile range of 0.95

Ad - Length	0
Ad- Width	0
Ad Size	659
Available_Impressions	1224
Matched_Queries	2038
Impressions	2115
Clicks	537
Spend	927
Fee	0
Revenue	1171
CTR	0
CPM	0
CPC	0
dtype: int64	

There are six variables without outliers after applying quartile range of 0.95. i.e., outliers are reduced after the treatment.

Figure 3: Boxplot post outlier treatment



1.4 Clustering: Perform z-score scaling and discuss how it affects the speed of the algorithm.

Ans:

Standardizing the data using z-score: (mean = 0, Variance = 1)

Table 9: Table showing scaled data after applying z-score

Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
-0.364496	-0.432797	-0.352218	-0.592761	-0.586089	-0.580978	-0.737121	-0.754487	0.481794	-0.712603	-0.999543	-1.067314	-0.908544
-0.364496	-0.432797	-0.352218	-0.592768	-0.586109	-0.580998	-0.737121	-0.754487	0.481794	-0.712603	-0.994168	-1.067314	-0.908544
-0.364496	-0.432797	-0.352218	-0.592505	-0.586073	-0.580961	-0.737121	-0.754487	0.481794	-0.712603	-1.003174	-1.067314	-0.908544
-0.364496	-0.432797	-0.352218	-0.592587	-0.586001	-0.580887	-0.737121	-0.754487	0.481794	-0.712603	-1.013545	-1.067314	-0.908544
-0.364496	-0.432797	-0.352218	-0.592925	-0.586131	-0.581021	-0.737121	-0.754487	0.481794	-0.712603	-0.986061	-1.067314	-0.908544

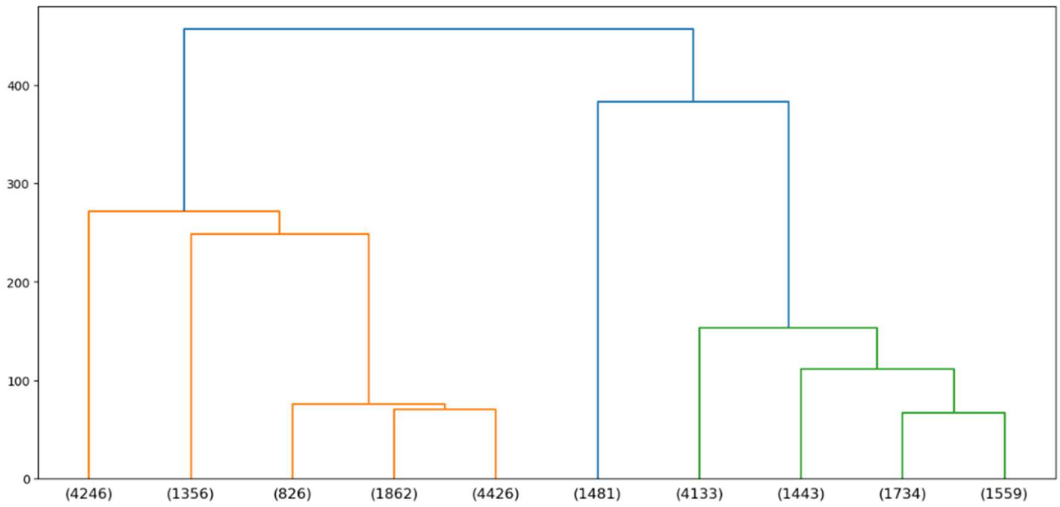
- Scaling the data increases the speed of algorithm as it compares variables with same units. It avoids the high variance between the variables, and thus, helps in deriving effective insights.
- Differently scaled data effect the accuracy of the formed clusters.
- Differently scaled data may lead to an inaccurate inference or analysis of the dataset.

1.5 Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.

Ans:

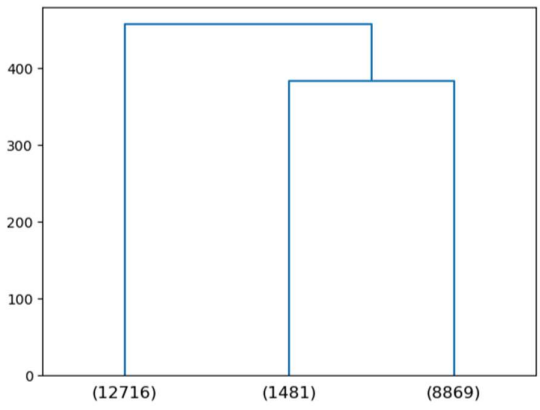
a) We need to use Linkage function on the scaled numerical data and apply WARD method. It uses Euclidean distance as default.

Figure 4: Dendrogram using Ward and Euclidean distance



b) Cutting Dendrogram with 3 clusters

Figure 5: Dendrogram with three clusters



c) Creating clusters using flcuster method

```
array([3, 3, 3, ..., 1, 1, 1], dtype=int32)
```

d) Appending clusters formed to ‘cluster’ column in the dataset.

Table 10: Appending Clusters column to the original dataset

Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	clusters
75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0	0.35	0.0	0.3096	0.0	0.0	3
75000	Inter227	App	Mobile	Video	1780	285	285	1	0.0	0.35	0.0	0.3509	0.0	0.0	3
75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0	0.35	0.0	0.2817	0.0	0.0	3

e) Calculating means of variables based on their **cluster profiles**

Table 11: Table showing mean values based on cluster profiles

clusters	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
1	341.573451	473.458635	123959.735766	2.053601e+05	1.287030e+05	1.072176e+05	11981.443929	1411.960830	0.342366	975.635423	14.220966	13.178352	0.099389
2	686.892640	114.929102	69750.222822	1.349033e+07	7.389622e+06	7.079394e+06	17713.607022	12259.319007	0.252633	9171.310774	0.194399	1.704855	0.855422
3	397.275454	180.764460	62049.971812	2.995988e+06	1.564653e+06	1.516226e+06	4494.964370	2406.726216	0.340939	1608.595654	0.359132	1.747087	0.547350

Observations:

- Cluster 1: Tier 3 ads: Has a mean spend of 1411.96 and Revenue 975.64

- Cluster 2: Tier 1 ads: Has a mean spend of 12259.32 and Revenue 9171.31

- Cluster 3: Tier 2 ads: Has a mean spend of 2406.73 and Revenue 1608.60

f) Additionally, performing **Agglomerative Clustering**: It is also known as bottom-up approach, where it considers the individual data points as a single cluster and divides into further clusters based on the distance.

- Performing Agglomerative Clustering for three clusters and by using WARD and Euclidean distance.
- Appending it to the original data set.
- Grouping based on Agglomerative Clusters

Table 12: Table showing mean values based on Agglo_clusters

Agglo_CLusters	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	Freq
0	341.573451	473.458635	123959.735766	2.053601e+05	1.287030e+05	1.072176e+05	11981.443929	1411.960830	0.342366	975.635423	14.220966	13.178352	0.099389	12716
1	397.275454	180.764460	62049.971812	2.995988e+06	1.564653e+06	1.516226e+06	4494.964370	2406.726216	0.340939	1608.595654	0.359132	1.747087	0.547350	8869
2	686.892640	114.929102	69750.222822	1.349033e+07	7.389622e+06	7.079394e+06	17713.607022	12259.319007	0.252633	9171.310774	0.194399	1.704855	0.855422	1481

Observations:

Agglo_cluster 1: CTR and CPM is highest among other 2 clusters, CPC is least compared to the other clusters.

Agglo_cluster 2: Has moderate CTR, CPM and CPC when compared with other two clusters.

Agglo_cluster 3: CTR and CPM is the least among other 2 clusters, CPC is the highest when compared with other two clusters.

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

Ans:

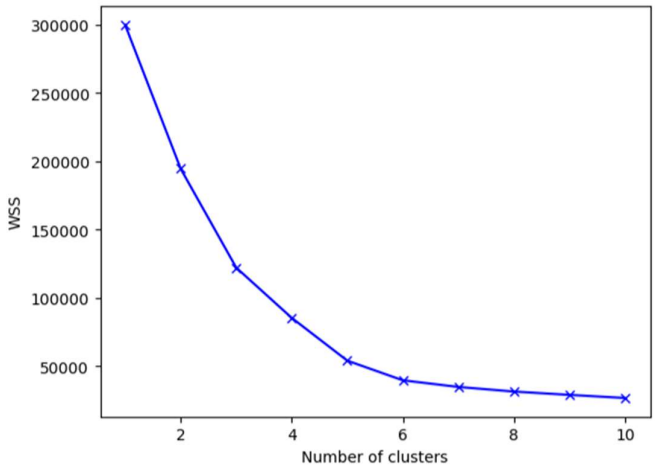
To decide the number of clusters for k-means algorithm, we need to identify the **INERTIA** (identify the distance between clusters) for each number of clusters (k=1, 2, 3, 4, …,10).

WSS values of 10 clusters (k=1, 2, 3, 4, …,10) are as follows:

Table 13: WSS values of 10 clusters

```
[299857.99999999994,
 194476.4654340704,
 122050.05691227358,
 85240.4802071002,
 53910.2239991707,
 39530.07031792419,
 34667.46892066276,
 31300.19942920031,
 28865.30782273907,
 26590.829899600332]
```

Figure 6: Elbow plot to identify optimum number of clusters



By referring the above Elbow plot, we can note that there is a significant fall up to 5 clusters so we can go ahead with k-means with k=5 i.e., with 5 clusters and group the data based on the cluster profiles.

1.7 Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters.

Ans:

The silhouette score for 2 clusters is 0.4376
The silhouette score for 3 clusters is 0.4233
The silhouette score for 4 clusters is 0.5039
The silhouette score for 5 clusters is 0.5667
The silhouette score for 6 clusters is 0.5529
The silhouette score for 7 clusters is 0.5426
The silhouette score for 8 clusters is 0.4652
The silhouette score for 9 clusters is 0.4195
The silhouette score for 10 clusters is 0.4291

By referring to the above silhouette scores, 0.5667 is the highest so the optimum number of clusters to be chosen is 5.

Also, hence, the silhouette score is 0.5667 which is in positive, we can say that the clusters are well separated between each other.

Sil-Width: Is used to check whether the mapping is correct to its cluster or not. If the value is Positive, it means mapped correctly.

However, in this data we can find 33 rows with negative values of which -0.03 is the least and it is close to 0.

Each sil-width are appended to the original data shown as follows:

Table 14: Table showing sil width appended

	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	...	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	Clus_kmeans	sil_width
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	...	323	1	0.0	0.35	0.0	0.3096	0.0	0.0	4	0.152279
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	...	285	1	0.0	0.35	0.0	0.3509	0.0	0.0	4	0.151552
2	2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	...	355	1	0.0	0.35	0.0	0.2817	0.0	0.0	4	0.152784

1.8 Clustering: 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].

Ans:

Table 15: Table showing mean values grouped by Cluster profiles

Clus_kmeans	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	Freq
0	715.911381	302.693563	215580.223881	2.459156e+05	1.345704e+05	1.142309e+05	14155.494636	1224.031686	0.349534	797.197108	13.965651	12.135436	0.089470	4288
1	676.088773	119.941253	70330.417755	1.789655e+07	9.540973e+06	9.188098e+06	17447.225849	15335.519347	0.239693	11730.937174	0.188227	1.697608	0.912393	1532
2	152.969019	558.082329	77697.074010	4.846454e+04	2.945556e+04	2.193073e+04	3019.647590	321.751120	0.349730	210.073525	15.864954	14.614030	0.101758	6972
3	142.361016	571.173644	75730.954015	8.363423e+05	5.861684e+05	4.945284e+05	67578.823610	7180.472745	0.286040	5163.174150	13.767503	15.157655	0.110092	1457
4	397.488488	180.226835	61891.241919	2.956683e+06	1.544690e+06	1.496821e+06	4464.951231	2379.853355	0.341257	1588.663940	0.356235	1.709256	0.580074	8817

Ad profile:

- Cluster 0: Tier 4 ads (second lowest revenue)

- Cluster 1: Tier 1 ads (highest revenue)

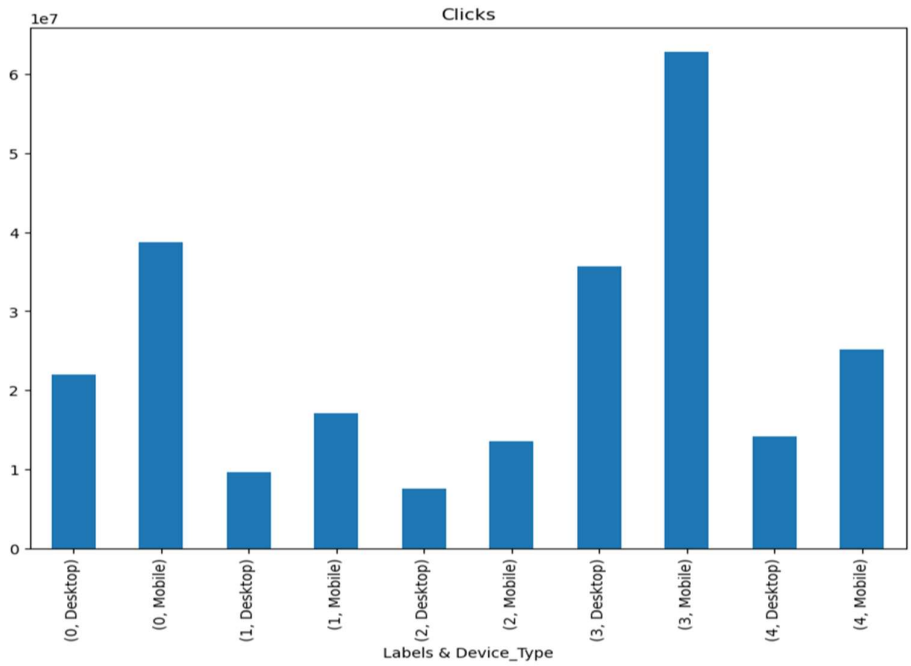
- Cluster 2: Tier 5 ads (Lowest revenue)

- Cluster 3: Tier 2 ads (second highest revenue)

- Cluster 4: Tier 3 ads (medium revenue)

a) **Clicks:**

Figure 7: Bar plots showing sum of Clicks across cluster profiles.



Observation:

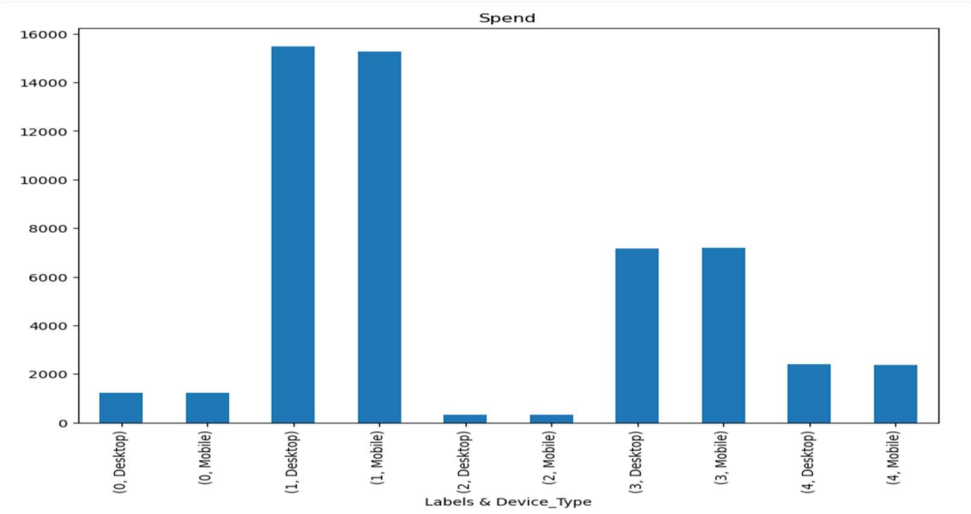
- Cluster 3 has the maximum users who clicked on advertisement via Mobile and Desktop, of which clicks on mobile ads are the most.

- Cluster 2 has the least users who clicked on advertisement via both Desktop and mobile, of which clicks on desktop ads are the least.

- Users in cluster 0 has clicked on advertisement more when compared with Cluster 1, 2 and 4.

b) **Spend:**

Figure 8: Bar plots showing mean Spends across cluster profiles



Observation:

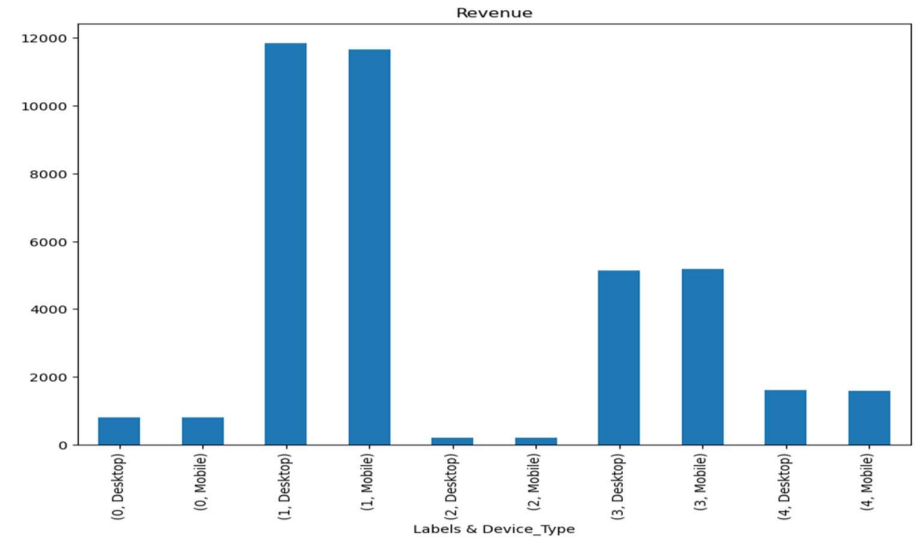
- Mean amount spent on advertisements for Cluster 1 on both Desktop and Mobile device is the highest.

-The least amount spent on advertisements on both the devices is for Cluster 2.

- Amount spent on advertisements for Cluster 3 is more than Cluster 0, 2 and 4.

c) **Revenue:**

Figure 9: Bar plots showing mean Revenue across cluster profiles



Observation:

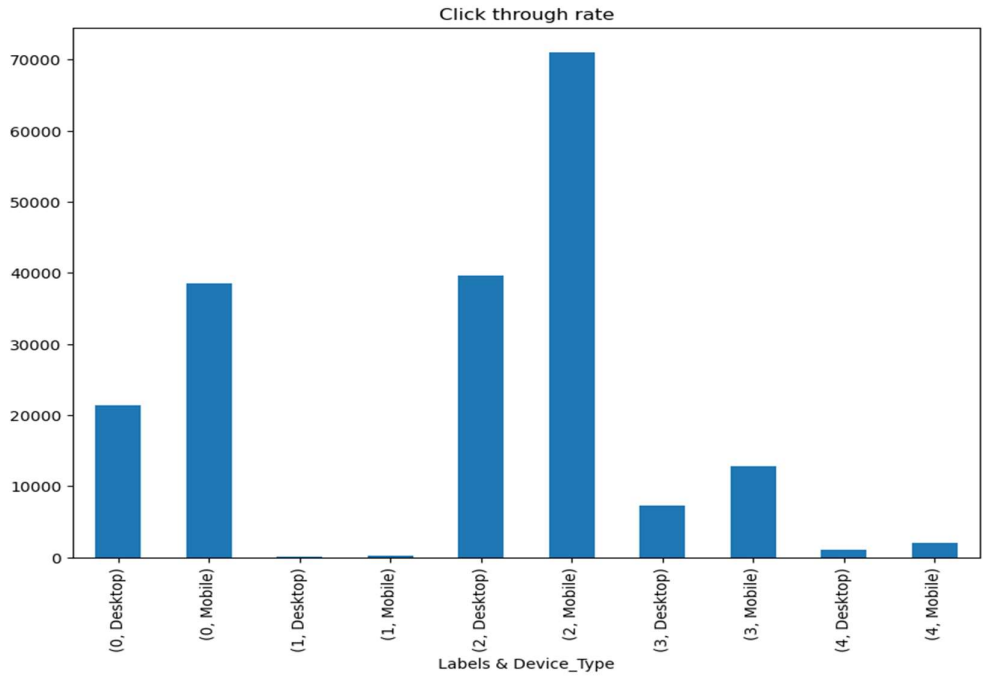
- Average revenue received from Cluster 1 on advertisements is the highest when compared with the other clusters.

- Average revenue received from Cluster 2 on advertisements via desktop and mobile devices is the least.

- Mean revenue received on advertisements from Desktop device of Cluster 1 is slightly higher when compared with the revenue from the Mobile device.

d) **CTR:**

Figure 10: Bar plots showing sum of CTR across cluster profiles



Observation:

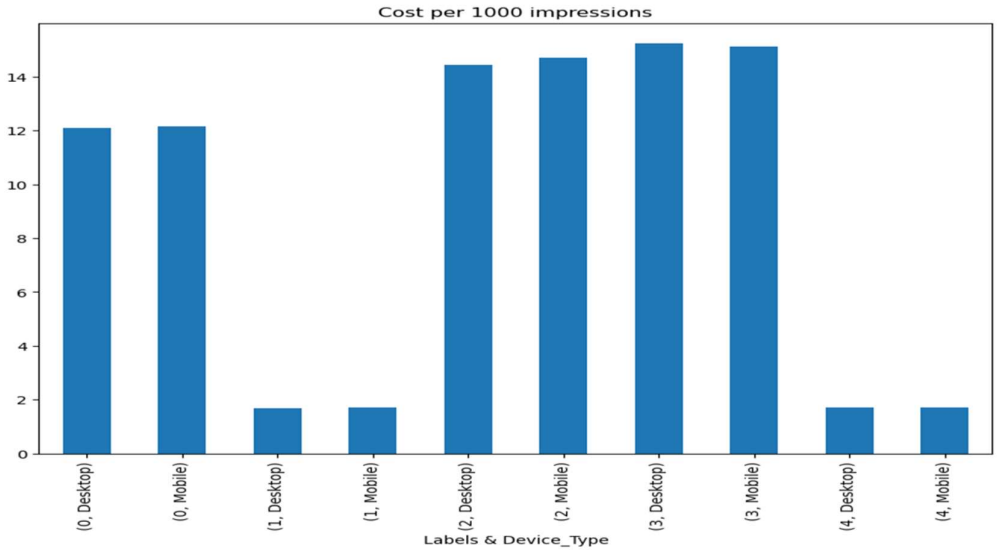
- Cluster 2 has the highest CTR among other clusters, of which CTR via mobile device is highest when compared with desktop device.

- Cluster 1 has the least CTR among all clusters on both devices type.

- Of all clusters, we can note that mobile device has more CTR rate than desktop ads.

e) **CPM:**

Figure 11: Bar plots showing mean CPM across cluster profiles



Observation:

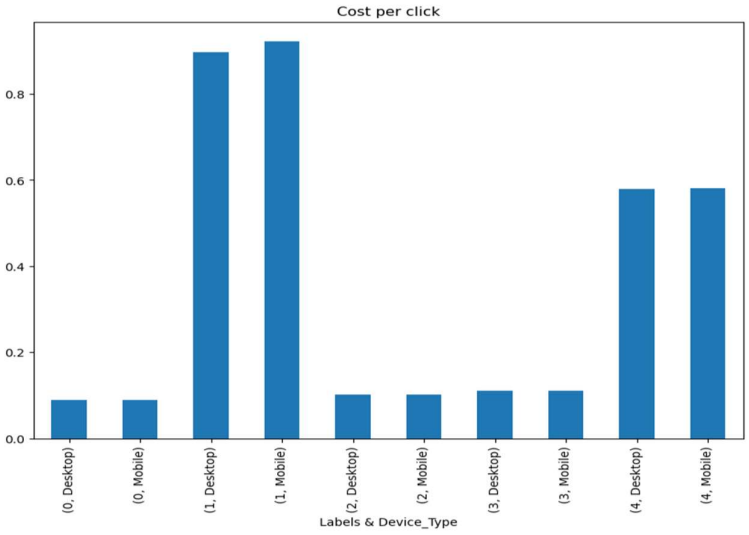
- Cluster 3 has the highest CPM among other clusters. Desktop ads CPM is slightly higher than that of mobile users.

- Average CPM of desktop ads of Cluster 1 is the least among all.

- Among mobile ads, mean CPM is least in Cluster 4.

f) **CPC:**

Figure 12: Bar plots showing mean CPC across cluster profiles



Observation:

- Average CPC is highest in Cluster 1 among other clusters, of which CPC of mobile ads are slightly higher than that of Desktop ads.

- Average CPC in Cluster 0 is the least.

- Of all clusters, Mobile ads has highest CPC than desktop ads except in cluster 3.

1.9 Clustering: **Conclude the project by providing summary of your learnings.**

Ans:

- We checked for the basic information such as size of the data, number or rows and columns, data types, missing values, duplicates, outliers and statistical summary to identify mean, standard deviation, min, max etc.
- There were missing values in CTR, CPM and CPC columns in the dataset, which we treated by applying the given formula, user defined function and Lambda function.
- We grouped the data based on the numerical and categorical data.
- We then used boxplot on grouped numerical data to identify outliers, and most of the columns had outliers. IQR for 95^{th} percentile method was used to treat outliers. Even though, there are few outliers left, most of the data has been covered.
- We then Scaled the numerical data using z-score, as there were differences in scale between the variables.
- After scaling the data, in order to perform **hierarchical clustering**, we constructed a Dendrogram by using Euclidean distance and Ward Linkage.
- We then cut the Dendrogram based on number of clusters we needed.
- Fcluster method has been used to create clusters, and then the same has been appended to the original data. We can now make inferences based on the clusters appended and create cluster profiles.
- We then perform Agglomerative clustering and append the agglomerative clusters to the original data and can start deriving insights.
- For **k-means**: we need to identify the number of clusters in advance.
- So, we check the Within Sum of Squares (WSS) by fitting the scaled data and by selecting the number of clusters k = 1, 2, 3, ….,10.
- We plot an elbow plot to select the number of clusters based on the significant fall. Based on the Elbow plot, we select the no. of clusters as 5 i.e., k=5.

- We now check for Silhouette score, to find out if the clusters are well separated or not. We check Silhouette scores of all k = 1, 2, 3, ….,10., to identify the optimum number of clusters, which is k=5, in the given dataset. And the silhouette score of the same is 0.56 which is in positive (clusters are well separated).
- We can also check the Sil-width to identify if the mapping of the observation to its centroid is correct or not. However, in this data we can find 33 rows with negative values of which -0.03 is the least and it is close to 0.
- Finally, using Silhouette score we can profile the ads based on optimum number of clusters by taking the sum or mean of mentioned columns.

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.
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.

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

Ans:

Checking basic info:

i) Checking head and tail of a dataset

Table 16: Top 5 and Bottom 5 rows of the dataset

Top 5 rows

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

MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
237	680	252	32	46	258	214
229	186	148	76	178	140	160
89	3	34	0	4	67	61
128	13	50	4	10	116	59
1043	205	302	24	105	180	478

Bottom 5 rows

	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
635	34	636	Puducherry	Mahe	3333	8154	11781	1146	1203	21	...	32	47	0
636	34	637	Puducherry	Karaikal	10612	12346	21691	1544	1533	2234	...	155	337	3
637	35	638	Andaman & Nicobar Island	Nicobars	1275	1549	2630	227	225	0	...	104	134	9
638	35	639	Andaman & Nicobar Island	North & Middle Andaman	3762	5200	8012	723	664	0	...	136	172	24
639	35	640	Andaman & Nicobar Island	South Andaman	7975	11977	18049	1470	1358	0	...	173	122	6

MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
0	0	0	0	0	32	47
14	38	130	4	23	110	170
4	2	6	17	47	76	77
44	11	21	1	4	100	103
2	17	17	2	4	148	99

ii) Size of the dataset: There are 640 rows and 61 columns in the dataset.

iii) Info of the dataset:

Table 17: Info of the dataset

```
RangeIndex: 640 entries, 0 to 639
Data columns (total 61 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   State Code      640 non-null    int64
 1   Dist.Code       640 non-null    int64
 2   State           640 non-null    object
 3   Area Name       640 non-null    object
 4   No_HH           640 non-null    int64
 5   TOT_M           640 non-null    int64
 6   TOT_F           640 non-null    int64
 7   M_06            640 non-null    int64
 8   F_06            640 non-null    int64
 9   M_SC            640 non-null    int64
 10  F_SC            640 non-null    int64
 11  M_ST            640 non-null    int64
 12  F_ST            640 non-null    int64
 13  M_LIT           640 non-null    int64
 14  F_LIT           640 non-null    int64
 15  M_ILL           640 non-null    int64
 16  F_ILL           640 non-null    int64
 17  TOT_WORK_M      640 non-null    int64
 18  TOT_WORK_F      640 non-null    int64
 19  MAINWORK_M      640 non-null    int64
 20  MAINWORK_F      640 non-null    int64
 21  MAIN_CL_M       640 non-null    int64
 22  MAIN_CL_F       640 non-null    int64
 23  MAIN_AL_M       640 non-null    int64
 24  MAIN_AL_F       640 non-null    int64
 25  MAIN_HH_M       640 non-null    int64
 26  MAIN_HH_F       640 non-null    int64
 27  MAIN_OT_M       640 non-null    int64
 28  MAIN_OT_F       640 non-null    int64
 29  MARGWORK_M      640 non-null    int64
 30  MARGWORK_F      640 non-null    int64
 31  MARG_CL_M       640 non-null    int64
 32  MARG_CL_F       640 non-null    int64
 33  MARG_AL_M       640 non-null    int64
 34  MARG_AL_F       640 non-null    int64
 35  MARG_HH_M       640 non-null    int64
 36  MARG_HH_F       640 non-null    int64
 37  MARG_OT_M       640 non-null    int64
 38  MARG_OT_F       640 non-null    int64
 39  MARGWORK_3_6_M  640 non-null    int64
 40  MARGWORK_3_6_F  640 non-null    int64
```

```
 41  MARG_CL_3_6_M   640 non-null    int64
 42  MARG_CL_3_6_F   640 non-null    int64
 43  MARG_AL_3_6_M   640 non-null    int64
 44  MARG_AL_3_6_F   640 non-null    int64
 45  MARG_HH_3_6_M   640 non-null    int64
 46  MARG_HH_3_6_F   640 non-null    int64
 47  MARG_OT_3_6_M   640 non-null    int64
 48  MARG_OT_3_6_F   640 non-null    int64
 49  MARGWORK_0_3_M  640 non-null    int64
 50  MARGWORK_0_3_F  640 non-null    int64
 51  MARG_CL_0_3_M   640 non-null    int64
 52  MARG_CL_0_3_F   640 non-null    int64
 53  MARG_AL_0_3_M   640 non-null    int64
 54  MARG_AL_0_3_F   640 non-null    int64
 55  MARG_HH_0_3_M   640 non-null    int64
 56  MARG_HH_0_3_F   640 non-null    int64
 57  MARG_OT_0_3_M   640 non-null    int64
 58  MARG_OT_0_3_F   640 non-null    int64
 59  NON_WORK_M      640 non-null    int64
 60  NON_WORK_F      640 non-null    int64
dtypes: int64(59), object(2)
memory usage: 305.1+ KB
```

Observations:

- There are 640 rows and 61 columns.

- There are 59 int64 and 2 object datatypes.

iv) Missing values:

Table 18: Table showing null values of the dataset

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

v) Duplicate values: There are no duplicate values.

vi) Statistical summary:

Table 19: Table showing statistical summary of numerical and categorical data

Numerical data

	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
M_06	640.0	12309.098438	11500.906881	56.0	4733.75	9159.0	16520.25	96223.0
F_06	640.0	11942.300000	11326.294567	56.0	4672.25	8663.0	15902.25	95129.0
M_SC	640.0	13820.946875	14426.373130	0.0	3466.25	9591.5	19429.75	103307.0
F_SC	640.0	20778.392188	21727.887713	0.0	5603.25	13709.0	29180.00	156429.0
M_ST	640.0	6191.807813	9912.668948	0.0	293.75	2333.5	7658.00	96785.0
F_ST	640.0	10155.640625	15875.701488	0.0	429.50	3834.5	12480.25	130119.0
M_LIT	640.0	57967.979688	55910.282466	286.0	21298.00	42693.5	77989.50	403261.0
F_LIT	640.0	66359.565625	75037.860207	371.0	20932.00	43796.5	84799.75	571140.0
M_ILL	640.0	21972.596875	19825.605268	105.0	8590.00	15767.5	29512.50	105961.0
F_ILL	640.0	56012.518750	47116.693769	327.0	22367.00	42386.0	78471.00	254160.0
TOT_WORK_M	640.0	37992.407813	36419.537491	100.0	13753.50	27936.5	50226.75	269422.0
TOT_WORK_F	640.0	41295.760938	37192.360943	357.0	16097.75	30588.5	53234.25	257848.0
MAINWORK_M	640.0	30204.446875	31480.915680	65.0	9787.00	21250.5	40119.00	247911.0
MAINWORK_F	640.0	28198.846875	29998.262689	240.0	9502.25	18484.0	35063.25	226166.0
MAIN_CL_M	640.0	5424.342188	4739.161969	0.0	2023.50	4160.5	7695.00	29113.0
MAIN_CL_F	640.0	5486.042188	5326.362728	0.0	1920.25	3908.5	7286.25	36193.0
MAIN_AL_M	640.0	5849.109375	6399.507966	0.0	1070.25	3936.5	8067.25	40843.0
MAIN_AL_F	640.0	8925.995312	12864.287584	0.0	1408.75	3933.5	10617.50	87945.0
MAIN_HH_M	640.0	883.893750	1278.642345	0.0	187.50	498.5	1099.25	16429.0
MAIN_HH_F	640.0	1380.773438	3179.414449	0.0	248.75	540.5	1435.75	45979.0
MAIN_OT_M	640.0	18047.101562	26068.480886	36.0	3997.50	9598.0	21249.50	240855.0
MAIN_OT_F	640.0	12406.035938	18972.202369	153.0	3142.50	6380.5	14368.25	209355.0
MARGWORK_M	640.0	7787.960938	7410.791691	35.0	2937.50	5627.0	9800.25	47553.0
MARGWORK_F	640.0	13096.914062	10996.474528	117.0	5424.50	10175.0	18879.25	66915.0
MARG_CL_M	640.0	1040.737500	1311.546847	0.0	311.75	606.5	1281.00	13201.0
MARG_CL_F	640.0	2307.682813	3564.626095	0.0	630.25	1226.0	2659.25	44324.0
MARG_AL_M	640.0	3304.326562	3781.555707	0.0	873.50	2062.0	4300.75	23719.0
MARG_AL_F	640.0	6463.281250	6773.876298	0.0	1402.50	4020.5	9089.25	45301.0
MARG_HH_M	640.0	316.742188	462.661891	0.0	71.75	166.0	356.50	4298.0
MARG_HH_F	640.0	786.626562	1198.718213	0.0	171.75	429.0	962.50	15448.0
MARG_OT_M	640.0	3126.154687	3609.391821	7.0	935.50	2036.0	3985.25	24728.0
MARG_OT_F	640.0	3539.323438	4115.191314	19.0	1071.75	2349.5	4400.50	36377.0
MARGWORK_3_6_M	640.0	41948.168750	39045.316918	291.0	16208.25	30315.0	57218.75	300937.0
MARGWORK_3_6_F	640.0	81076.323438	82970.406216	341.0	26619.50	56793.0	107924.00	676450.0
MARG_CL_3_6_M	640.0	6394.987500	6019.806644	27.0	2372.00	4630.0	8167.00	39106.0
MARG_CL_3_6_F	640.0	10339.864063	8467.473429	85.0	4351.50	8295.0	15102.00	50065.0
MARG_AL_3_6_M	640.0	789.848438	905.639279	0.0	235.50	480.5	986.00	7426.0
MARG_AL_3_6_F	640.0	1749.584375	2496.541514	0.0	497.25	985.5	2059.00	27171.0
MARG_HH_3_6_M	640.0	2743.635938	3059.586387	0.0	718.75	1714.5	3702.25	19343.0
MARG_HH_3_6_F	640.0	5169.850000	5335.640960	0.0	1113.75	3294.0	7502.25	36253.0
MARG_OT_3_6_M	640.0	245.362500	358.728567	0.0	58.00	129.5	276.00	3535.0
MARG_OT_3_6_F	640.0	585.884375	900.025817	0.0	127.75	320.5	719.25	12094.0
MARGWORK_0_3_M	640.0	2616.140625	3036.964381	7.0	755.00	1681.5	3320.25	20648.0
MARGWORK_0_3_F	640.0	2834.545312	3327.836932	14.0	833.50	1834.5	3610.50	25844.0
MARG_CL_0_3_M	640.0	1392.973438	1489.707052	4.0	489.50	949.0	1714.00	9875.0
MARG_CL_0_3_F	640.0	2757.050000	2788.776676	30.0	957.25	1928.0	3599.75	21611.0
MARG_AL_0_3_M	640.0	250.889062	453.336594	0.0	47.00	114.5	270.75	5775.0
MARG_AL_0_3_F	640.0	558.098438	1117.642748	0.0	109.00	247.5	568.75	17153.0
MARG_HH_0_3_M	640.0	560.690625	762.578991	0.0	136.50	308.0	642.00	6116.0
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

Categorical data

	count	unique	top	freq
State	640	35	Uttar Pradesh	71
Area Name	640	635	Raigarh	2

2.2 PCA: Perform detailed Exploratory analysis by creating certain questions like:

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

(ii) Which district has the highest & lowest gender ratio? (Example Questions). Pick 5 variables out of the given 24 variables below for EDA: No_HH, TOT_M, TOT_F, M_06, F_06, M_SC, F_SC, M_ST, F_ST, M_LIT, F_LIT, M_ILL, F_ILL, TOT_WORK_M, TOT_WORK_F, AINWORK_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.

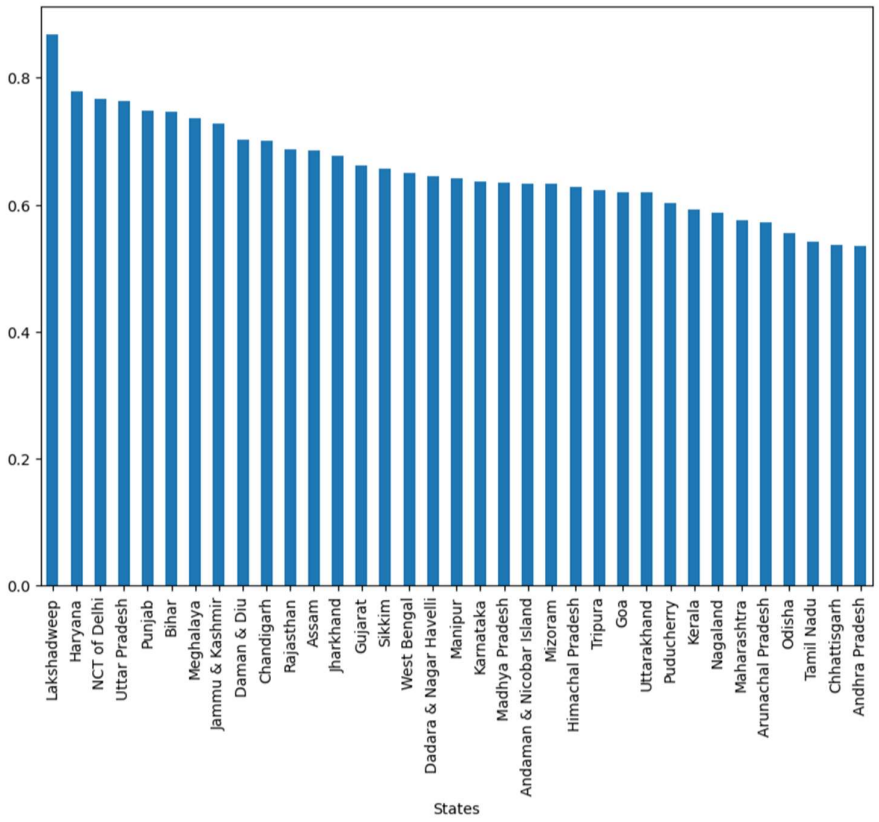
Ans:

i) Ans:

Table 20: Table showing highest to lowest Gender ratio state wise

```
State
Lakshadweep                  0.868061
Haryana                      0.777742
NCT of Delhi                 0.766436
Uttar Pradesh                0.762436
Punjab                       0.747843
Bihar                        0.746294
Meghalaya                    0.736105
Jammu & Kashmir              0.728013
Daman & Diu                  0.702191
Chandigarh                   0.700037
Rajasthan                    0.687548
Assam                        0.685665
Jharkhand                    0.677664
Gujarat                      0.662355
Sikkim                       0.657083
West Bengal                  0.649317
Dadara & Nagar Havelli       0.644631
Manipur                      0.641547
Karnataka                    0.636473
Madhya Pradesh               0.635514
Andaman & Nicobar Island     0.633861
Mizoram                      0.633117
Himachal Pradesh             0.628779
Tripura                      0.622312
Goa                          0.620158
Uttarakhand                  0.618892
Puducherry                   0.601799
Kerala                       0.591736
Nagaland                     0.587934
Maharashtra                  0.575367
Arunachal Pradesh            0.571755
Odisha                       0.555303
Tamil Nadu                   0.541956
Chhattisgarh                 0.537435
Andhra Pradesh               0.534941
Name: Gender_ratio, dtype: float64
```

Figure 13: Bar plots showing state wise Gender ratio



Observations

Lakshadweep state has the **highest** gender ratio at **0.87**

Andhra Pradesh state has the **lowest** gender ratio at **0.53**

ii) Ans:

Table 21: Table showing highest to lowest Gender ratio district wise

State	Area Name	
Lakshadweep	Lakshadweep	0.868061
Jammu & Kashmir	Badgam	0.847762
Uttar Pradesh	Mahamaya Nagar	0.847313
Rajasthan	Dhaulpur	0.846911
Uttar Pradesh	Baghpat	0.844003
		...
Odisha	Baudh	0.451455
Andhra Pradesh	West Godavari	0.450076
Tamil Nadu	Virudhunagar	0.449352
Odisha	Koraput	0.440769
Andhra Pradesh	Krishna	0.437972

Name: Gender_ratio, Length: 640, dtype: float64

Observations

Lakshadweep district of Lakshadweep state has the **highest** gender ratio at **0.87**.

Krishna district of Andhra Pradesh state has the **lowest** gender ratio at **0.44**.

EDA

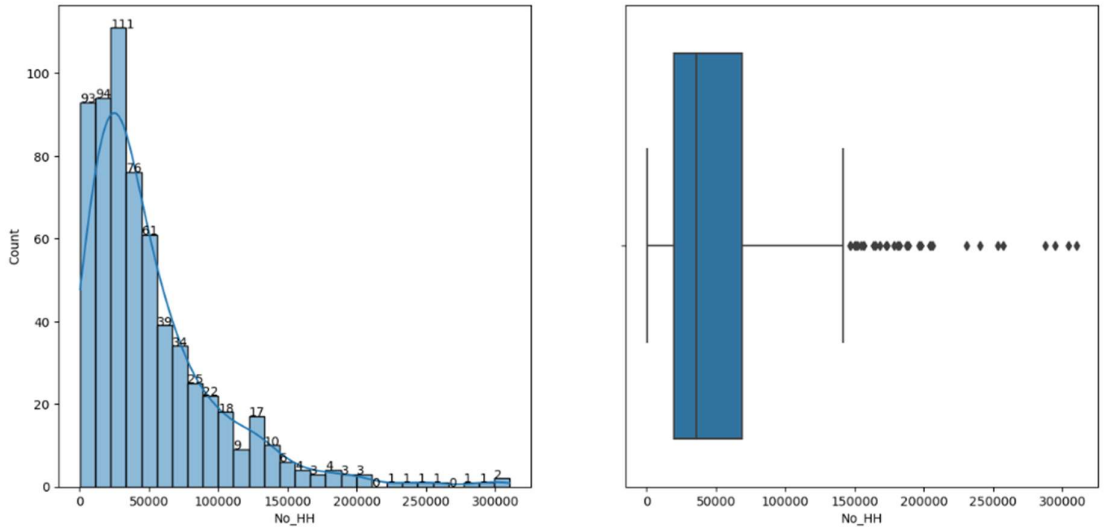
5 variables to perform EDA are **i) No_HH, ii) TOT_M, iii) TOT_F, iv) M_LIT, v) F_LIT**

There are no duplicate values.

- **Univariate analysis**:

a) Number of Household:

Figure 14: Histogram and Box plot showing Number of Household



Statistical summary of the variable

```
count       640.000000
mean      51222.871875
std       48135.405475
min         350.000000
25%       19484.000000
50%       35837.000000
75%       68892.000000
max      310450.000000
Name: No_HH, dtype: float64
```

Observations

- The average number of households of all states comes up to 51223.

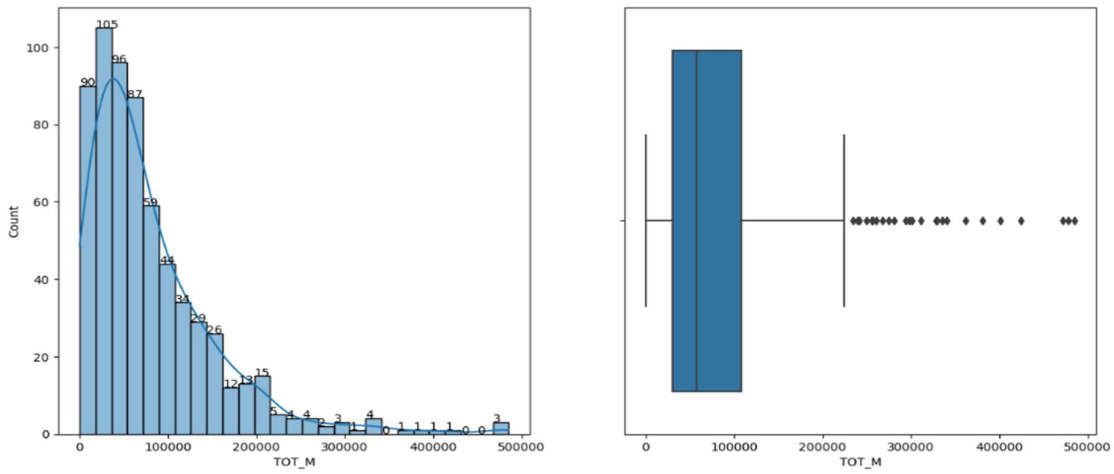
- The number of households ranges from 350 to 310450 across states.

- About 75% of the households fall under 68892.

- Also, there are outliers in the given variable.

b) Total male population:

Figure 15: Histogram and Box plot showing total male population



Statistical summary of the variable

```
count       640.000000
mean      79940.576563
std       73384.511114
min         391.000000
25%       30228.000000
50%       58339.000000
75%      107918.500000
max      485417.000000
Name: TOT_M, dtype: float64
```

Observations

- The average number of total male population across states is 79941.

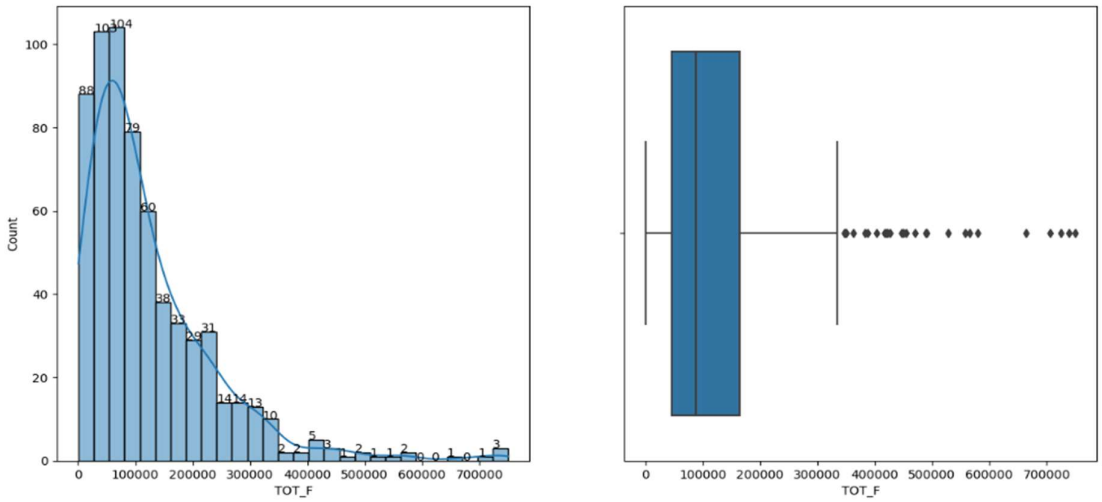
- Total male population ranges from 391 to 485417 across states.

- About 75% of the total male population fall under 107919.

- Also, there are outliers in the given variable.

c) Total female population:

Figure 16: Histogram and Box plot showing total female population



Statistical summary of the variable

```
count       640.000000
mean     122372.084375
std      113600.717282
min         698.000000
25%       46517.750000
50%       87724.500000
75%      164251.750000
max      750392.000000
Name: TOT_F, dtype: float64
```

Observations

- The average number of total female population across states is 122372.

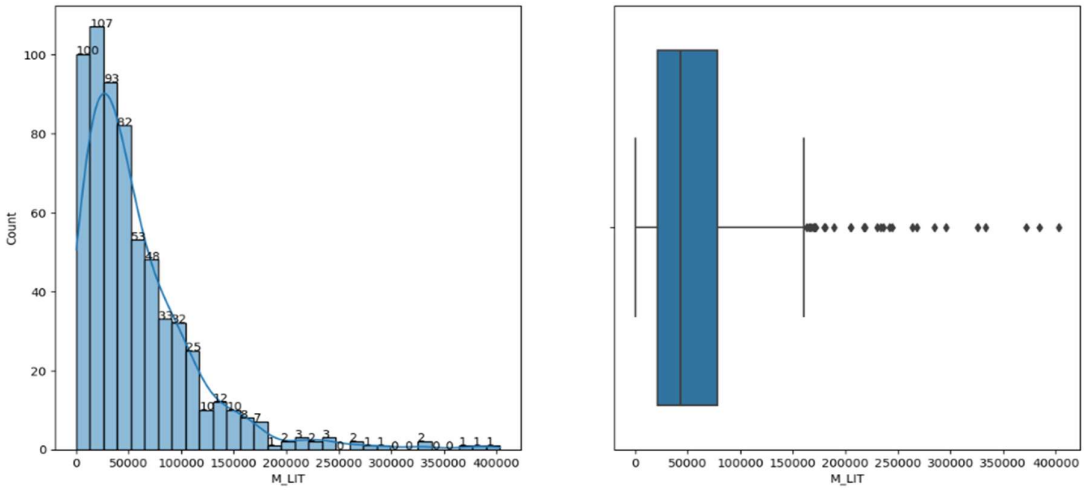
- Total female population ranges from 698 to 750392 across states.

- About 75% of the total female population fall under 164252.

- Also, there are outliers in the given variable.

d) Population male literates:

Figure 17: Histogram and Box plot showing male literates



Statistical summary of the variable

```
count       640.000000
mean      57967.979688
std       55910.282466
min         286.000000
25%       21298.000000
50%       42693.500000
75%       77989.500000
max      403261.000000
Name: M_LIT, dtype: float64
```

Observations

- The average number of male literates across states is 57968.

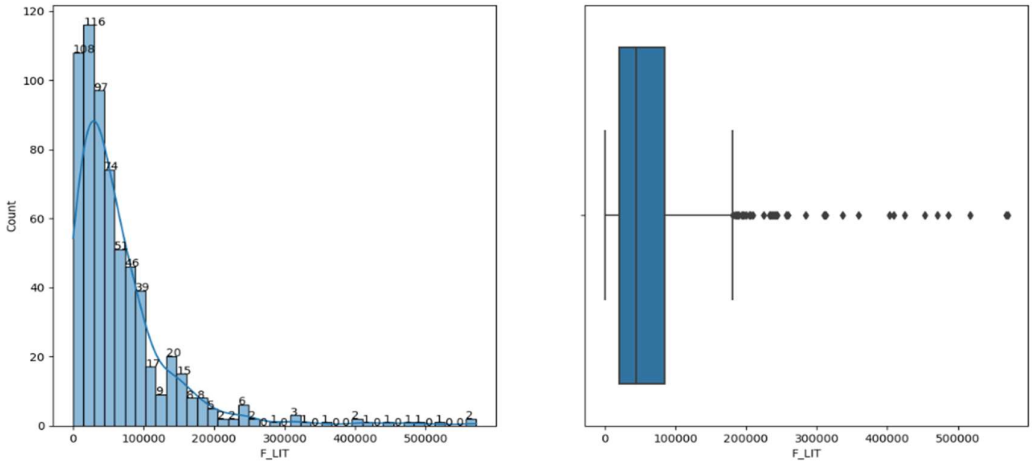
- Total male literates ranges from 286 to 403261 across states.

- About 75% of the male literates fall under 77990.

- Also, there are outliers in the given variable.

e) Population female literates:

Figure 18: Histogram and Box plot showing female literates



Statistical summary of the variable

```
count       640.000000
mean      66359.565625
std       75037.860207
min         371.000000
25%       20932.000000
50%       43796.500000
75%       84799.750000
max      571140.000000
Name: F_LIT, dtype: float64
```

Observations

- The average number of female literates across states is 66360.

- Total female literates ranges from 371 to 571140 across states.

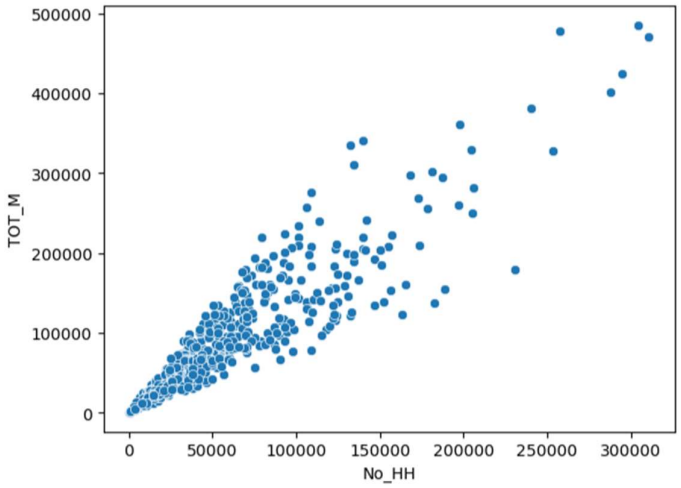
- About 75% of the female literates fall under 84800.

- Also, there are outliers in the given variable.

- **<u>Bivariate analysis and Multivariate analysis:</u>**

a) Number of Household and Total male population:

Figure 19: Scatter plot showing number of household and total male population

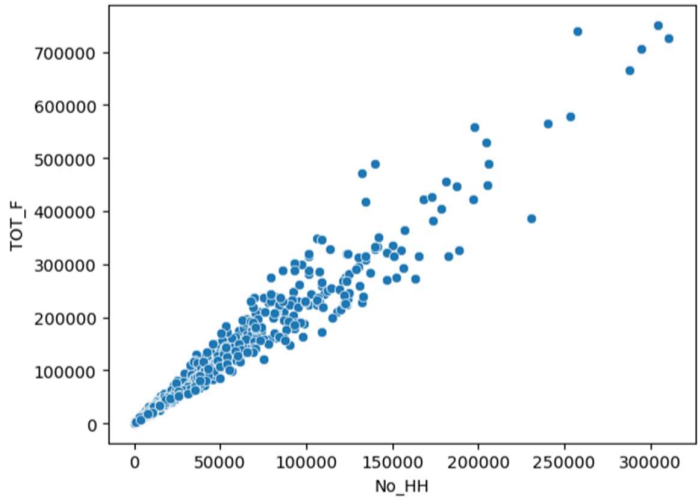


Observation

- The above scatter plot states that 'Number of Household' and 'Total male population' are positively correlated.

b) Number of Household and Total female population:

Figure 20: Scatter plot showing number of household and total female population

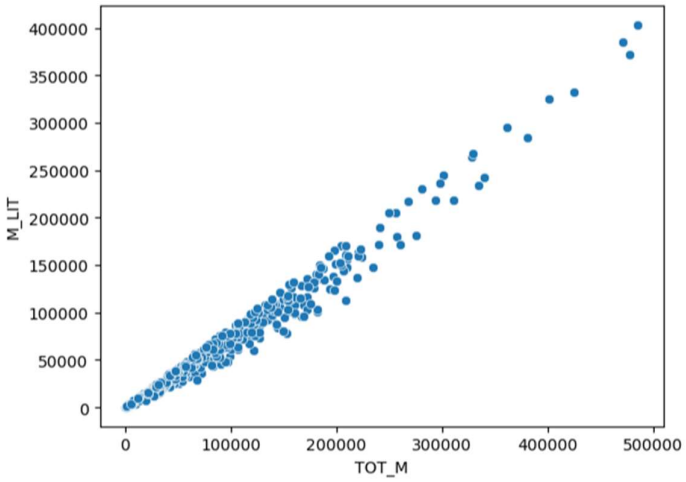


Observation

- The above scatter plot states that 'Number of Household' and 'Total female population' are positively correlated.

c) Total male population and male literates:

Figure 21: Scatter plot showing total male population and male literates



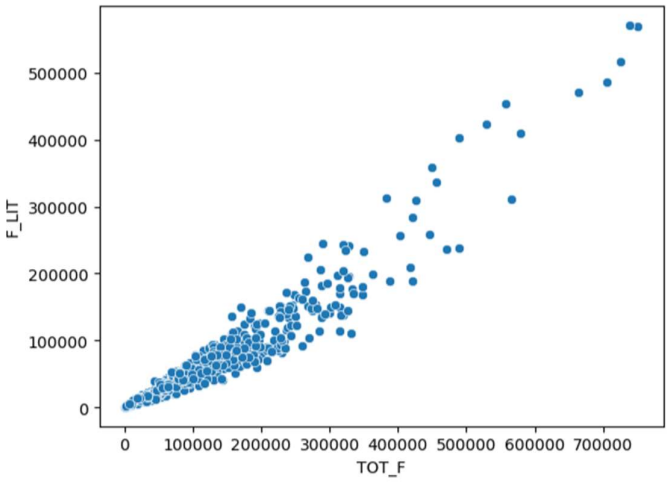
Observations

- The above scatter plot states that 'Total male population' and 'male literates' are positively correlated.

- We can note that, male literates increase as the total male population increases.

d) Total female population and female literates:

Figure 22: Scatter plot showing total male population and female literates



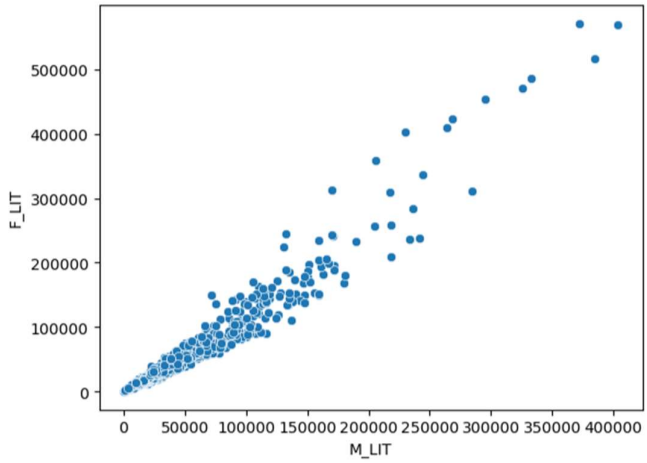
Observations

- The above scatter plot states that 'Total female population' and 'female literates' are positively correlated.

- We can note that, to some extent, female literates increase as the total female population increases.

e) Male literates and female literates:

Figure 23: Scatter plot showing male literates and female literates

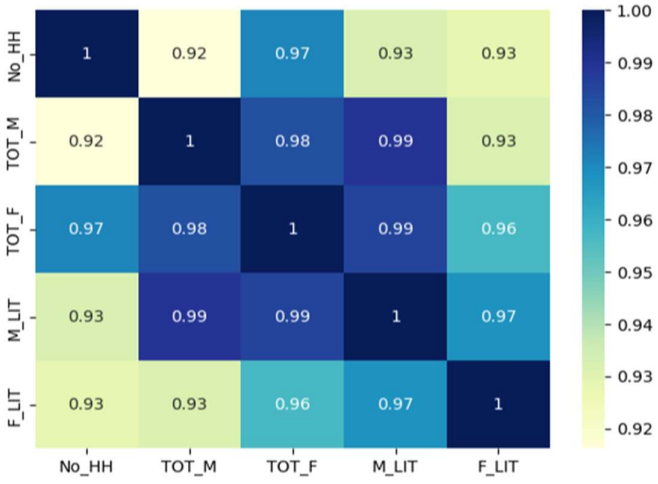


Observation

- The above scatter plot states that 'male literates' and 'female literates' are positively correlated.

f) Heatmap:

Figure 24: Heatmap to check correlation between variables



Observation

- All the variables are highly correlated to each other, as their values are close to 1.

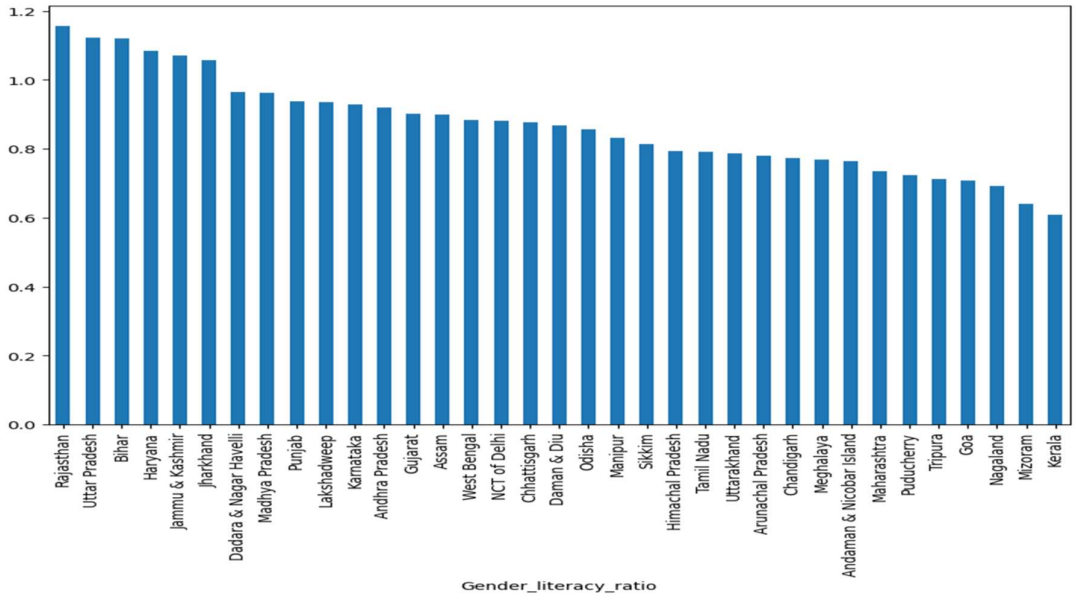
g) Gender literacy ratio:

Table 22: Table showing highest to lowest Gender literacy ratio state wise

State	
Rajasthan	1.156394
Uttar Pradesh	1.121931
Bihar	1.120135
Haryana	1.084954
Jammu & Kashmir	1.070158
Jharkhand	1.056542
Dadara & Nagar Havelli	0.964393
Madhya Pradesh	0.961363
Punjab	0.937033
Lakshadweep	0.935327
Karnataka	0.929349
Andhra Pradesh	0.919396
Gujarat	0.902603
Assam	0.898926
West Bengal	0.884096
NCT of Delhi	0.880233
Chhattisgarh	0.876549
Daman & Diu	0.868633
Odisha	0.856602
Manipur	0.831078
Sikkim	0.813240
Himachal Pradesh	0.793936
Tamil Nadu	0.791711
Uttarakhand	0.787789
Arunachal Pradesh	0.779787
Chandigarh	0.772411
Meghalaya	0.769249
Andaman & Nicobar Island	0.764699
Maharashtra	0.735532
Puducherry	0.723172
Tripura	0.711325
Goa	0.708643
Nagaland	0.691159
Mizoram	0.640936
Kerala	0.608350

Name: Gender_literacy_ratio, dtype: float64

Figure 25: Bar plots to show Gender literacy ratio



Observations:

Rajasthan state has the **highest** gender literacy ratio at **1.16**.

Kerala state has the **lowest** gender literacy ratio at **0.61**.

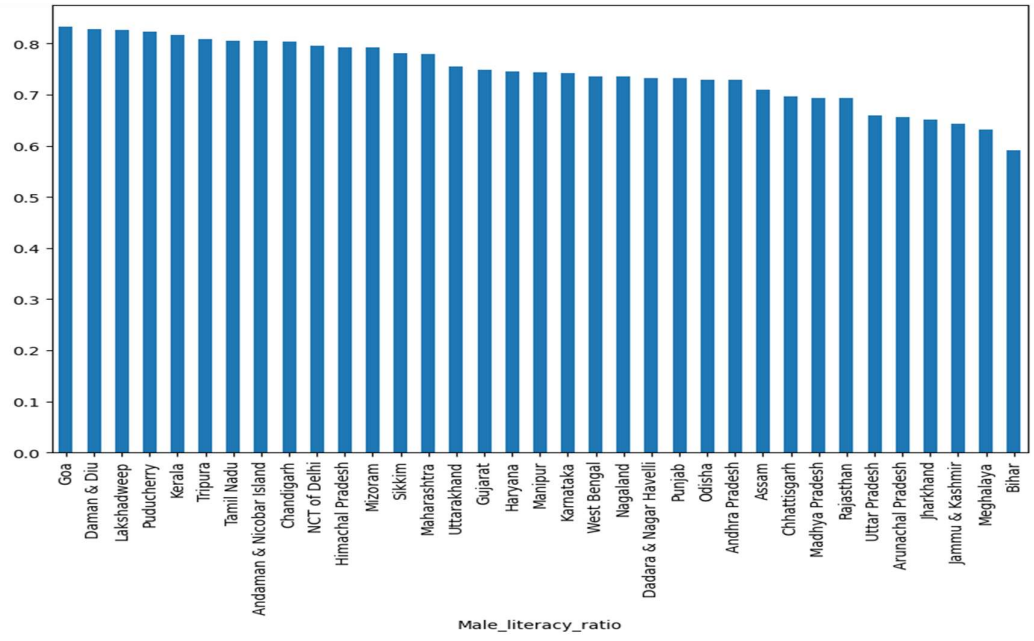
h) Male literacy ratio:

Table 23: Table showing highest to lowest Male literacy ratio state wise

State	
Goa	0.833482
Daman & Diu	0.828375
Lakshadweep	0.826718
Puducherry	0.824228
Kerala	0.816831
Tripura	0.808683
Tamil Nadu	0.806207
Andaman & Nicobar Island	0.806004
Chandigarh	0.803583
NCT of Delhi	0.795762
Himachal Pradesh	0.793232
Mizoram	0.792366
Sikkim	0.781631
Maharashtra	0.780262
Uttarakhand	0.754822
Gujarat	0.748836
Haryana	0.745162
Manipur	0.744200
Karnataka	0.742976
West Bengal	0.736261
Nagaland	0.735217
Dadara & Nagar Havelli	0.733171
Punjab	0.732630
Odisha	0.729386
Andhra Pradesh	0.729191
Assam	0.709930
Chhattisgarh	0.696946
Madhya Pradesh	0.694144
Rajasthan	0.693443
Uttar Pradesh	0.659553
Arunachal Pradesh	0.656318
Jharkhand	0.652123
Jammu & Kashmir	0.643804
Meghalaya	0.631353
Bihar	0.590680

Name: Male_literacy_ratio, dtype: float64

Figure 26: Bar plots to show Male literacy ratio



Observations:

Goa state has the **highest** male literacy ratio at **0.83**.

Bihar state has the **lowest** male literacy ratio at **0.59**.

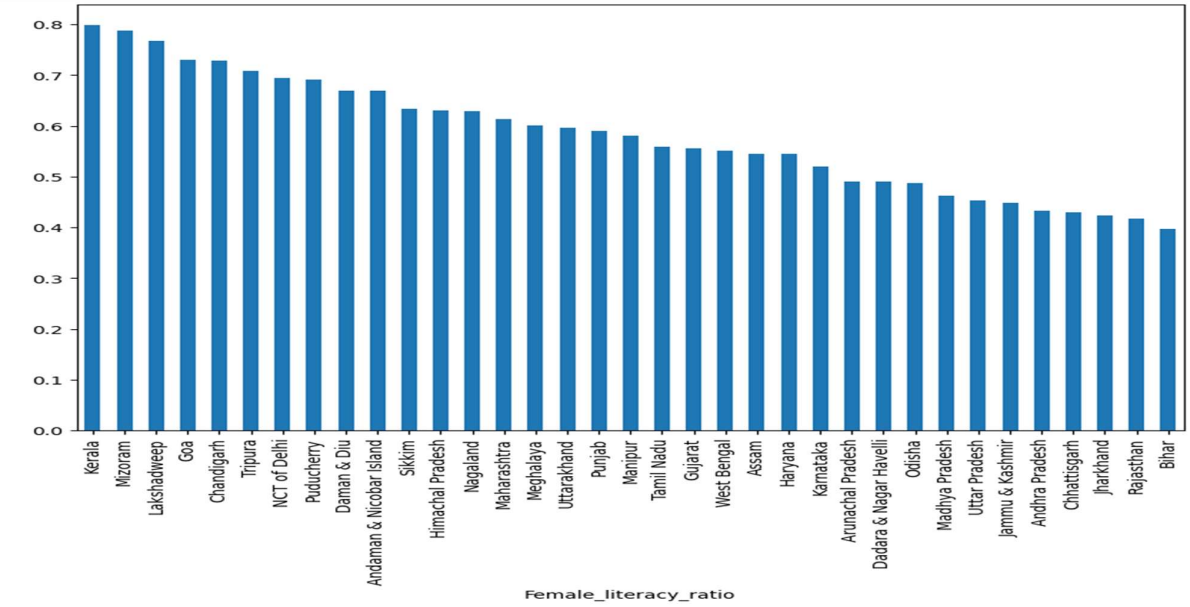
i) Female literacy ratio:

Table 24: Table showing highest to lowest Female literacy ratio state wise

State	
Kerala	0.798793
Mizoram	0.787055
Lakshadweep	0.767262
Goa	0.729529
Chandigarh	0.728288
Tripura	0.707725
NCT of Delhi	0.693965
Puducherry	0.690578
Daman & Diu	0.669672
Andaman & Nicobar Island	0.669194
Sikkim	0.634182
Himachal Pradesh	0.631158
Nagaland	0.629037
Maharashtra	0.613968
Meghalaya	0.600384
Uttarakhand	0.596435
Punjab	0.590436
Manipur	0.580983
Tamil Nadu	0.558609
Gujarat	0.555819
West Bengal	0.551922
Assam	0.545601
Haryana	0.545342
Karnataka	0.519520
Arunachal Pradesh	0.490424
Dadara & Nagar Havelli	0.490075
Odisha	0.487476
Madhya Pradesh	0.462566
Uttar Pradesh	0.452721
Jammu & Kashmir	0.447971
Andhra Pradesh	0.432238
Chhattisgarh	0.430304
Jharkhand	0.422860
Rajasthan	0.417166
Bihar	0.397530

Name: Female_literacy_ratio, dtype: float64

Figure 27: Bar plots to show Female literacy ratio



Observations:

Kerala state has the **highest** female literacy ratio at **0.80**.

Bihar state has the **lowest** female literacy ratio at **0.40**.

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

Ans: There are outliers in the data before and after scaling the data. But in this case, since we are dealing with the population census report, we can choose **NOT to treat outliers** as to avoid wrong interpretation of the data. Treating outliers may lead to wrong observation and analysis of the data. Also, it may cause in losing some sensible or valuable data.

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

Ans:

Scaled data by using z-score method

Table 25: Table showing scaled data using z-score method

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_0_3_M	MARG_CL_0_3_F
0	-0.904738	-0.771236	-0.815563	-0.561012	-0.507738	-0.958575	-0.957049	-0.423306	-0.476423	-0.798097	...	-0.163229	-0.720610
1	-0.935695	-0.823100	-0.874534	-0.681096	-0.725367	-0.958297	-0.956772	-0.582014	-0.607607	-0.849434	...	-0.583103	-0.732811
2	-0.972412	-1.000919	-0.981466	-0.976956	-0.965262	-0.958575	-0.956772	-0.038951	-0.027273	-0.956457	...	-0.859212	-0.921931
3	-1.037530	-1.052224	-1.041001	-1.022118	-0.995393	-0.958783	-0.957049	-0.355965	-0.390060	-1.004643	...	-0.805468	-0.900758
4	-0.822676	-0.809381	-0.813933	-0.622359	-0.649908	-0.957395	-0.955529	0.149238	0.043330	-0.800568	...	-0.348645	-0.297513

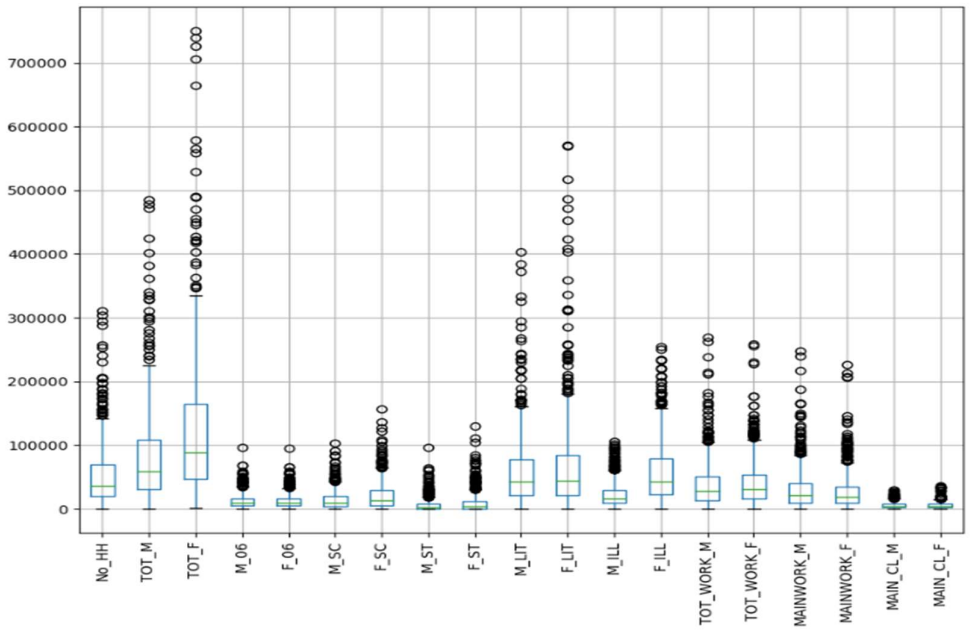
MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
-0.156494	-0.287524	0.156577	-0.657412	-0.365258	-0.499977	-0.413053	-0.539614
-0.282327	-0.294688	-0.491731	-0.723062	0.042855	-0.073481	-0.606455	-0.598988
-0.456727	-0.420050	-0.731894	-0.795026	-0.662068	-0.635680	-0.726103	-0.707839
-0.419198	-0.385127	-0.718770	-0.784926	-0.624966	-0.616294	-0.645791	-0.710038
0.472670	0.434200	-0.466796	-0.625849	-0.439461	-0.309346	-0.540895	-0.249344

Scaling does not remove outliers; it just helps in comparing variables of different scales on the same scale.

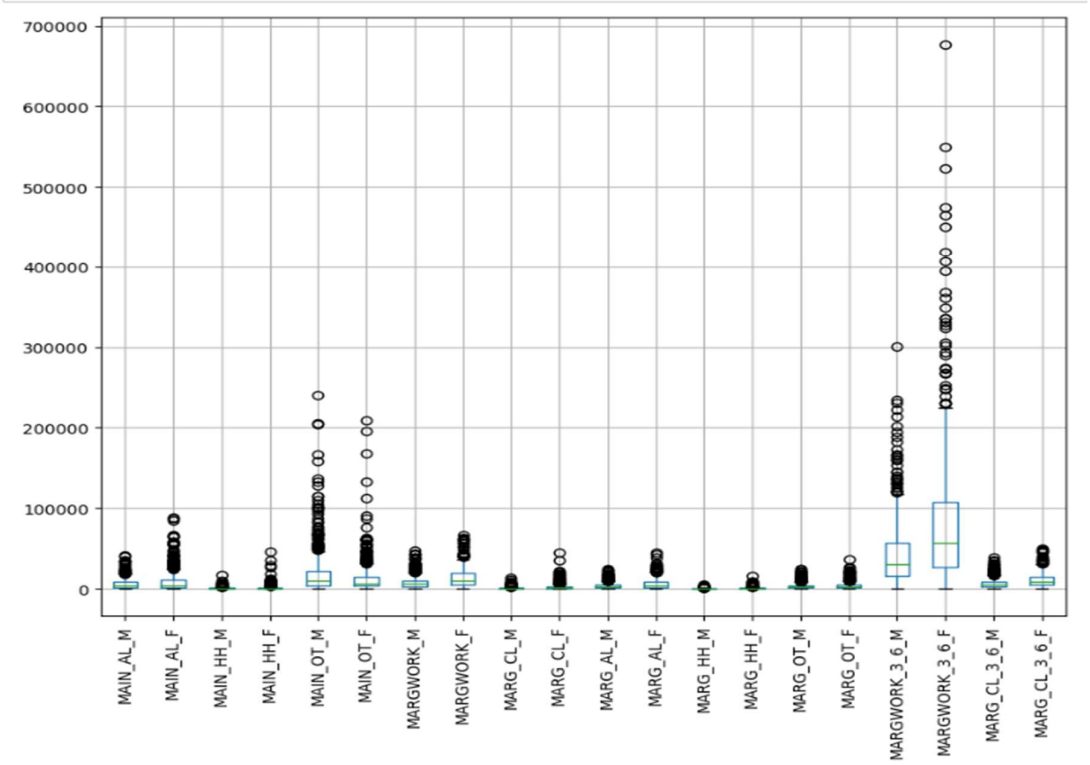
Boxplots before scaling the data:

Figure 28: Box plots before scaling the data

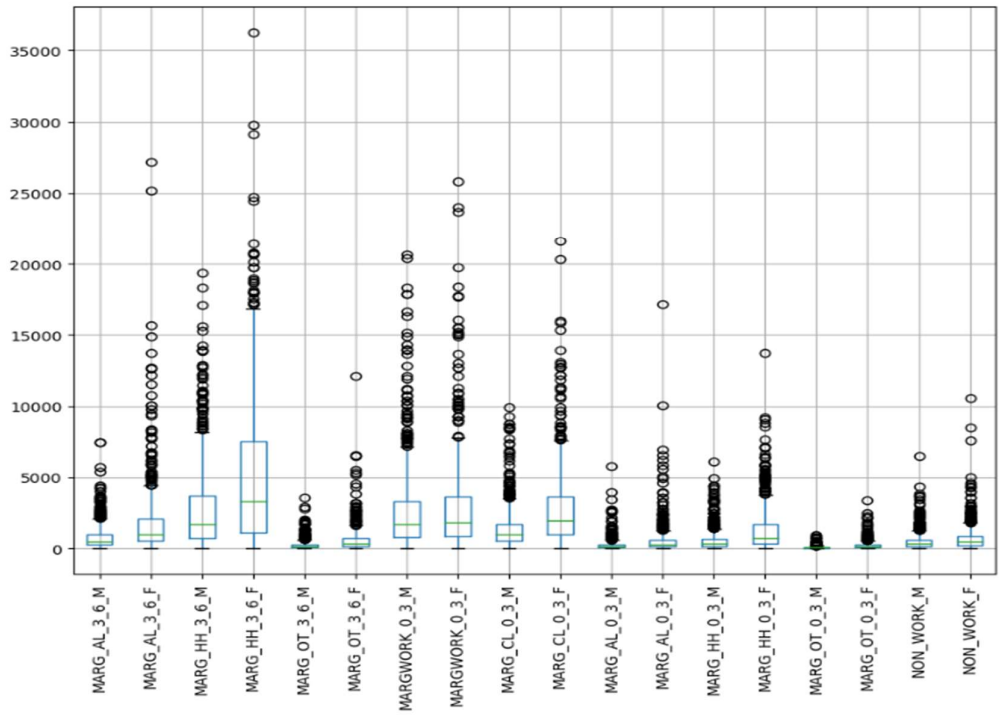
- Columns 1 - 19



- Columns 20 - 39



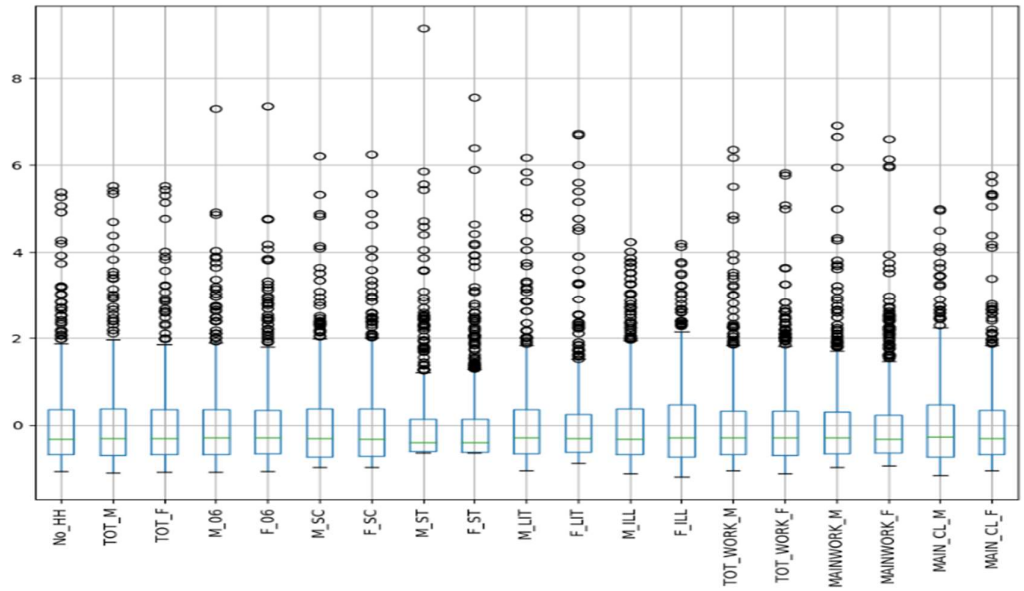
- Columns 40 - 57



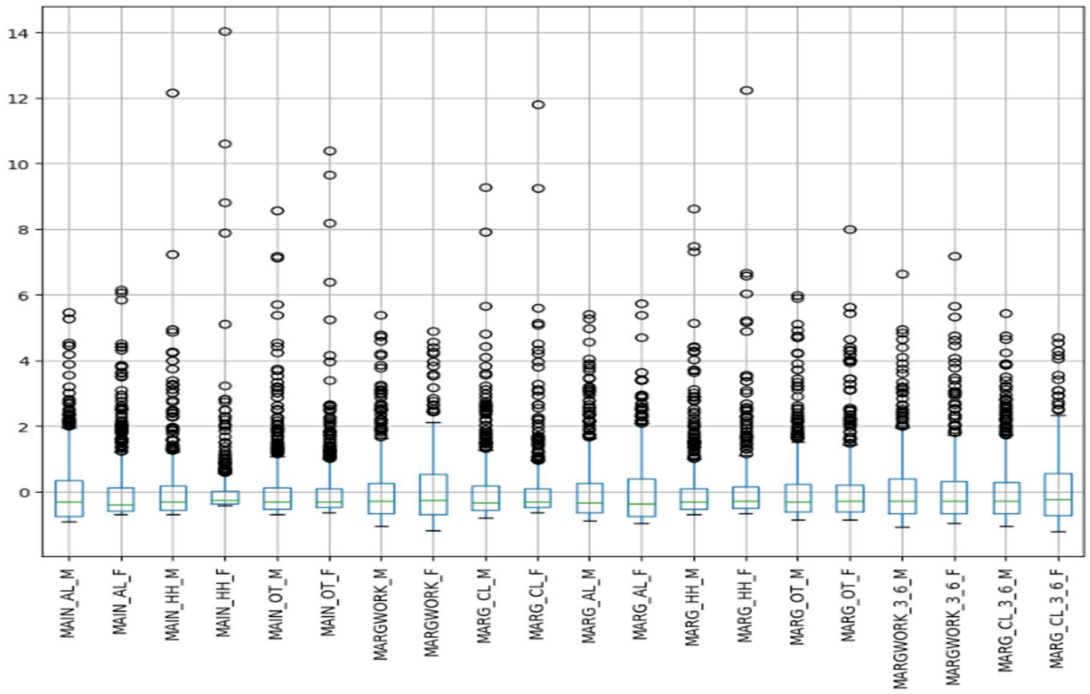
Boxplots after scaling the data:

Figure 29: Box plots after scaling the data

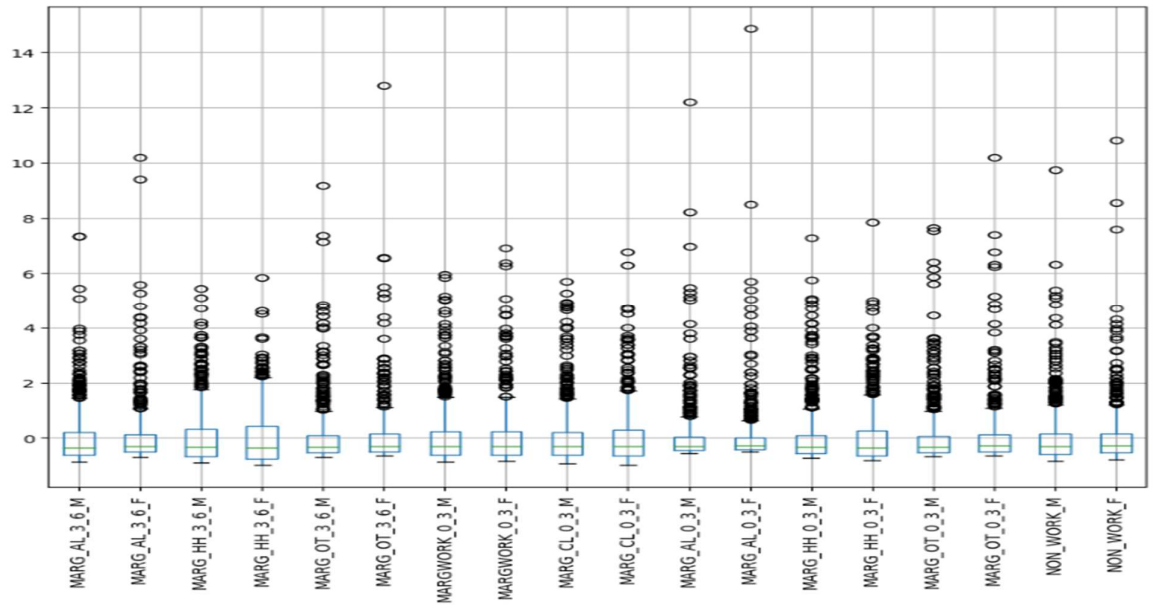
- Columns 1 - 19



- Columns 20 - 39



- Columns 40 - 57



Observations:

- After comparing boxplots before and after scaling, we can note that the outliers still exist. It does not remove outliers.
- Scaling helps in comparing variables of different scales on the same scale.

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

Ans:

PCA steps

a) Bartletts Test of Sphericity

H0: All variables in the data are uncorrelated

Ha: At least one pair of variables in the data are correlated

As the **p value (0.00) is < 0.05**, we can reject Null hypothesis, i.e., Atleast one pair or all variables are correlated.

b) **Adequacy of sample size - The Kaiser-Meyer-Olkin (KMO)**

The kmo value (0.80) is above 0.7, so we can say that adequacy of sample size is good.

c) **Covariance matrix**

Table 26: Table showing covariance matrix of the data

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_0_3_M	MARG_CL_0_3_F
No_HH	1.001565	0.917604	0.972109	0.798807	0.797619	0.776522	0.825137	0.149861	0.165361	0.933397	...	0.557813	0.556412
TOT_M	0.917604	1.001565	0.984178	0.952313	0.949275	0.841240	0.827592	0.091565	0.086315	0.990860	...	0.699403	0.596628
TOT_F	0.972109	0.984178	1.001565	0.909396	0.907975	0.818238	0.834059	0.123819	0.128848	0.986983	...	0.656373	0.599888
M_06	0.798807	0.952313	0.909396	1.001565	0.999713	0.782342	0.748700	0.055361	0.044017	0.914186	...	0.761800	0.648011
F_06	0.797619	0.949275	0.907975	0.999713	1.001565	0.774345	0.742846	0.065240	0.054748	0.909062	...	0.764809	0.650851
M_SC	0.776522	0.841240	0.818238	0.782342	0.774345	1.001565	0.986612	-0.045738	-0.047900	0.819765	...	0.674687	0.570470
F_SC	0.825137	0.827592	0.834059	0.748700	0.742846	0.986612	1.001565	-0.014144	-0.009204	0.815424	...	0.651473	0.586607
M_ST	0.149861	0.091565	0.123819	0.055361	0.065240	-0.045738	-0.014144	1.001565	0.989593	0.090682	...	0.123160	0.196878
F_ST	0.165361	0.086315	0.128848	0.044017	0.054748	-0.047900	-0.009204	0.989593	1.001565	0.087512	...	0.121601	0.217081
M_LIT	0.933397	0.990860	0.986983	0.914186	0.909062	0.819765	0.815424	0.090682	0.087512	1.001565	...	0.653528	0.560942

MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
0.067531	0.046200	0.369168	0.418100	0.487508	0.537694	0.763577	0.736844
0.167666	0.115761	0.496704	0.441049	0.652624	0.589101	0.846218	0.717182
0.138980	0.099593	0.451717	0.443825	0.594735	0.572748	0.828949	0.747751
0.267091	0.198649	0.602031	0.514860	0.691682	0.565914	0.786189	0.652162
0.259074	0.189865	0.612525	0.524089	0.699729	0.575077	0.784954	0.652458
0.184620	0.129953	0.524269	0.462621	0.665958	0.592362	0.736550	0.580927
0.163683	0.116410	0.508801	0.489422	0.629453	0.590268	0.721226	0.601028
0.027262	0.007088	0.126534	0.239186	-0.005491	0.090264	0.123178	0.147184
0.017232	0.002560	0.136610	0.273735	-0.005889	0.100140	0.114309	0.151105
0.144292	0.101514	0.422422	0.382324	0.601059	0.553235	0.853533	0.739827

d) PCA:

- **Eigen vectors:**

Table 27: Table showing Eigen vectors

```
array([[ 0.15602058,  0.16711763,  0.16555318, ...,  0.13219224,
         0.15037558,  0.1310662 ],
       [-0.12634653, -0.08967655, -0.10491237, ...,  0.05081332,
        -0.06536455, -0.07384742],
       [-0.00269025,  0.05669762,  0.03874947, ..., -0.07871987,
         0.11182732,  0.1025525 ],
       ...,
       [-0.        , -0.17278849, -0.09520952, ...,  0.00987322,
        -0.04362296, -0.0207041 ],
       [-0.        , -0.0116324 , -0.0814326 , ...,  0.04647201,
        -0.17212428,  0.03763315],
       [ 0.        ,  0.18260602,  0.03874463, ..., -0.00370238,
        -0.05681626, -0.03729932]])
```

- **Eigen values:**

Table 28: Table showing Eigen values (descending order)

```
array([3.18135647e+01, 7.86942415e+00, 4.15340812e+00, 3.66879058e+00,
       2.20652588e+00, 1.93827502e+00, 1.17617374e+00, 7.51159086e-01,
       6.17053743e-01, 5.28300887e-01, 4.29831189e-01, 3.53440201e-01,
       2.96163013e-01, 2.81275560e-01, 1.92158325e-01, 1.36267920e-01,
       1.13389199e-01, 1.06303946e-01, 9.72885376e-02, 8.01062194e-02,
       5.76089954e-02, 4.43955966e-02, 3.78910846e-02, 2.96360194e-02,
       2.70797618e-02, 2.34458139e-02, 1.45111511e-02, 1.09852268e-02,
       9.31507853e-03, 8.13540203e-03, 7.89250253e-03, 5.02601514e-03,
       2.59771182e-03, 1.06789820e-03, 7.13559124e-04, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31])
```

- **Explained variance for each PC:**

Table 29: Table showing explained variance for each PC

```
array([5.57260632e-01, 1.37844354e-01, 7.27529548e-02, 6.42641771e-02,
       3.86504944e-02, 3.39516923e-02, 2.06023855e-02, 1.31576386e-02,
       1.08085894e-02, 9.25395468e-03, 7.52911540e-03, 6.19101667e-03,
       5.18772384e-03, 4.92694855e-03, 3.36593119e-03, 2.38692984e-03,
       1.98617593e-03, 1.86206747e-03, 1.70414955e-03, 1.40317638e-03,
       1.00910494e-03, 7.77653131e-04, 6.63717190e-04, 5.19117774e-04,
       4.74341222e-04, 4.10687364e-04, 2.54183814e-04, 1.92422147e-04,
       1.63167083e-04, 1.42503342e-04, 1.38248605e-04, 8.80379297e-05,
       4.55026824e-05, 1.87057826e-05, 1.24990208e-05, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33])
```

- **Cumulative variance in percentage:**

Table 30: Table showing Cumulative variance in percentage

```
Cumulative variance in percentage: [ 55.73  69.51  76.79  83.21  87.08  90.47  92.53  93.85  94.93  95.85
  96.61  97.23  97.75  98.24  98.57  98.81  99.01  99.2   99.37  99.51
  99.61  99.69  99.75  99.81  99.85  99.89  99.92  99.94  99.96  99.97
  99.98  99.99 100.   100.   100.   100.   100.   100.   100.   100.
 100.   100.   100.   100.   100.   100.   100.   100.   100.   100.
 100.   100.   100.   100.   100.   100.   100.  ]
```

- **Principal components:**

Table 31: Table showing principal components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	...	PC48	PC49
No_HH	0.156021	-0.126347	-0.002690	-0.125293	-0.007022	0.004083	-0.118110	0.057238	0.004265	0.019985	...	0.000000e+00	0.000000e+00
TOT_M	0.167118	-0.089677	0.056698	-0.019942	-0.033026	-0.073389	0.089554	0.111431	0.018872	-0.024501	...	1.790399e-01	2.116730e-01
TOT_F	0.165553	-0.104912	0.038749	-0.070873	-0.012847	-0.043647	-0.002124	0.088355	0.014911	-0.038041	...	3.049004e-01	3.599700e-01
M_06	0.162193	-0.022095	0.057788	0.011917	-0.050248	-0.157957	0.165067	0.169595	-0.056773	-0.153574	...	1.128224e-15	1.236131e-15
F_06	0.162566	-0.020271	0.050126	0.014844	-0.043848	-0.154436	0.169082	0.169459	-0.059323	-0.169567	...	-3.053702e-16	-1.127955e-15
M_SC	0.151358	-0.045111	0.002569	0.012485	-0.173007	-0.064295	-0.001566	-0.129301	0.037480	0.448517	...	-1.946360e-15	-2.317591e-15
F_SC	0.151567	-0.051924	-0.025101	-0.029893	-0.159803	-0.040518	-0.084658	-0.144352	0.041232	0.446968	...	2.161504e-15	2.347179e-15
M_ST	0.027234	0.027679	-0.123504	-0.222247	0.433163	0.222591	0.405505	0.021982	0.018632	0.160418	...	2.576064e-16	6.834810e-16
F_ST	0.028183	0.030223	-0.139769	-0.229754	0.438792	0.225531	0.357800	0.014873	0.043866	0.134862	...	-7.979728e-17	-5.429684e-16
M_LIT	0.161993	-0.115355	0.082168	-0.035163	-0.009101	-0.055465	0.045934	0.099423	0.045193	-0.005752	...	-2.403377e-02	-2.298470e-01

PC50	PC51	PC52	PC53	PC54	PC55	PC56	PC57
0.000000e+00	-0.000000e+00	-0.000000e+00	0.000000e+00	0.000000e+00	-0.000000e+00	-0.000000e+00	0.000000e+00
-8.102590e-02	2.134714e-02	5.997863e-02	-1.346169e-01	-8.916366e-02	-1.727885e-01	-1.163240e-02	1.826060e-01
-5.161556e-01	3.162317e-01	7.458124e-02	2.244617e-01	1.163855e-02	-9.520952e-02	-8.143260e-02	3.874463e-02
-2.437023e-15	-3.881605e-16	-2.204057e-15	-7.092846e-16	1.806623e-15	-4.634998e-16	8.556932e-16	1.678060e-17
1.935852e-15	-2.310314e-16	2.616085e-15	5.302943e-16	-1.328301e-15	5.786993e-16	-2.470046e-16	3.410647e-16
1.004405e-15	-1.537046e-15	-1.605487e-15	5.516421e-16	3.382711e-16	1.356554e-15	2.594062e-15	-1.578598e-15
-1.077155e-15	1.762324e-15	1.554672e-15	-7.695957e-16	-3.306910e-16	-1.345778e-15	-2.857240e-15	1.414827e-15
3.365364e-16	3.690624e-16	2.814589e-16	-6.245005e-17	7.979728e-17	4.484260e-16	3.292722e-16	-3.226586e-16
-4.145989e-16	-3.885781e-16	-2.432950e-16	1.148712e-16	-1.474515e-16	-3.243933e-16	-4.427882e-16	3.269954e-16
1.799804e-02	3.454980e-02	6.417461e-04	-1.147567e-01	-1.393562e-02	-3.532632e-01	4.256257e-02	7.258125e-02

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

Ans:

- **Scree plot**

Figure 30: Scree plot to identify optimum number of PCs

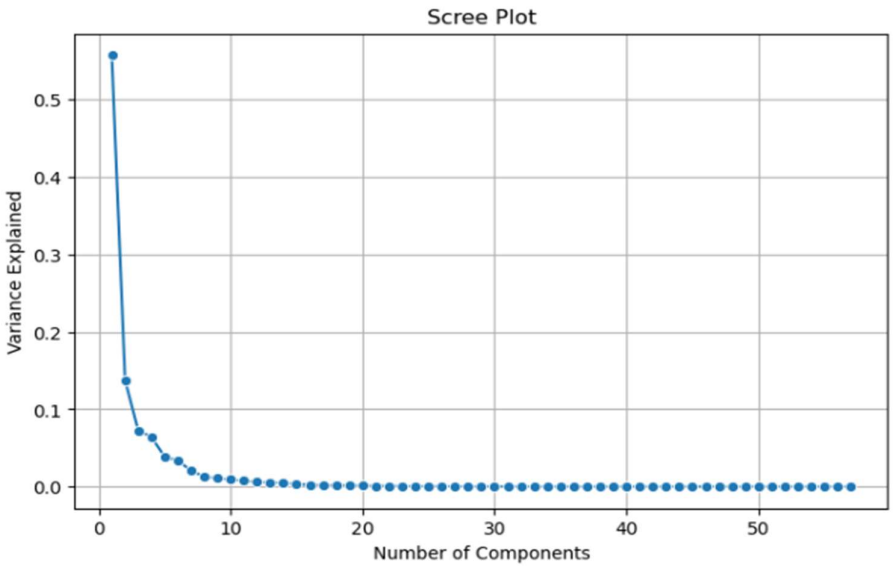


Table 32: Table showing cumulative sum of explained variance

```
array([0.55726063, 0.69510499, 0.76785794, 0.83212212, 0.87077261,
       0.9047243 , 0.92532669, 0.93848433, 0.94929292, 0.95854687,
       0.96607599, 0.97226701, 0.97745473, 0.98238168, 0.98574761,
       0.98813454, 0.99012071, 0.99198278, 0.99368693, 0.99509011,
       0.99609921, 0.99687687, 0.99754058, 0.9980597 , 0.99853404,
       0.99894473, 0.99919891, 0.99939134, 0.9995545 , 0.99969701,
       0.99983525, 0.99992329, 0.9999688 , 0.9999875 , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        ])
```

We are proceeding with **7 optimum** number of PCs, as it contains **92.53%** explained variance, after which there is minute variances between the variables.

- **Extracted PCs from covariance matrix:**

Table 33: Table showing Extracted PCs from covariance matrix

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
No_HH	0.156021	-0.126347	-0.002690	-0.125293	-0.007022	0.004083	-0.118110
TOT_M	0.167118	-0.089677	0.056698	-0.019942	-0.033026	-0.073389	0.089554
TOT_F	0.165553	-0.104912	0.038749	-0.070873	-0.012847	-0.043647	-0.002124
M_06	0.162193	-0.022095	0.057788	0.011917	-0.050248	-0.157957	0.165067
F_06	0.162566	-0.020271	0.050126	0.014844	-0.043848	-0.154436	0.169082
M_SC	0.151358	-0.045111	0.002569	0.012485	-0.173007	-0.064295	-0.001566
F_SC	0.151567	-0.051924	-0.025101	-0.029893	-0.159803	-0.040518	-0.084658
M_ST	0.027234	0.027679	-0.123504	-0.222247	0.433163	0.222591	0.405505
F_ST	0.028183	0.030223	-0.139769	-0.229754	0.438792	0.225531	0.357800
M_LIT	0.161993	-0.115355	0.082168	-0.035163	-0.009101	-0.055465	0.045934
F_LIT	0.146873	-0.153109	0.117098	-0.059559	0.055844	-0.048021	-0.021064
M_ILL	0.161749	-0.006625	-0.021855	0.025348	-0.096580	-0.115234	0.201947
F_ILL	0.165248	-0.009107	-0.093062	-0.076023	-0.119911	-0.028757	0.028425
TOT_WORK_M	0.159872	-0.133529	0.045176	-0.040154	-0.019553	-0.001801	0.045053
TOT_WORK_F	0.145936	-0.085087	-0.059450	-0.225160	-0.040437	0.105162	-0.119424
MAINWORK_M	0.146201	-0.176368	0.054295	-0.068351	-0.036802	0.019283	0.047367
MAINWORK_F	0.123970	-0.151413	-0.055609	-0.246640	-0.082834	0.123832	-0.090431
MAIN_CL_M	0.103127	0.062415	-0.067399	-0.089769	-0.286039	-0.006170	0.385792
MAIN_CL_F	0.074540	0.086477	-0.009238	-0.288965	-0.241936	0.102951	0.207882
MAIN_AL_M	0.113356	-0.031040	-0.247917	-0.136082	-0.205724	-0.031068	-0.013077
MAIN_AL_F	0.073882	-0.058688	-0.251932	-0.290042	-0.177605	0.019240	-0.158334
MAIN_HH_M	0.131573	-0.076021	0.026569	0.152366	-0.134089	0.174465	0.119825
MAIN_HH_F	0.083383	-0.082477	-0.060523	0.048950	-0.139441	0.422309	-0.139294
MAIN_OT_M	0.123526	-0.212984	0.137378	-0.040289	0.064638	0.023477	-0.015601
MAIN_OT_F	0.111021	-0.210071	0.095634	-0.120391	0.080743	0.083079	-0.070645
MARGWORK_M	0.164615	0.092994	-0.008628	0.093018	0.060244	-0.090761	0.020195
MARGWORK_F	0.155396	0.125270	-0.049370	-0.088707	0.089202	0.017868	-0.157222
MARG_CL_M	0.082389	0.269450	0.198754	-0.062761	-0.022263	0.031915	0.029072
MARG_CL_F	0.049195	0.246547	0.268787	-0.168402	-0.059205	0.092086	-0.045884
MARG_AL_M	0.128599	0.165831	-0.189868	0.091787	0.019422	-0.141605	0.020298
MARG_AL_F	0.114305	0.140958	-0.267768	-0.106365	0.080527	-0.085120	-0.150712
MARG_HH_M	0.140853	0.068068	-0.021257	0.237985	-0.059971	0.089533	0.108604
MARG_HH_F	0.127670	0.024216	-0.082504	0.196321	-0.033602	0.365112	-0.049472
MARG_OT_M	0.155263	-0.089442	0.111713	0.087119	0.119121	-0.061066	-0.004288
MARG_OT_F	0.147287	-0.117899	0.100046	0.026729	0.166882	0.001739	-0.117886
MARGWORK_3_6_M	0.164972	-0.043995	0.064423	-0.000026	-0.043834	-0.136253	0.126291
MARGWORK_3_6_F	0.161253	-0.105502	0.079704	0.003894	0.000537	-0.106900	0.050625
MARG_CL_3_6_M	0.165502	0.077193	-0.024205	0.092875	0.054073	-0.096708	0.026671
MARG_CL_3_6_F	0.155647	0.103174	-0.072013	-0.107860	0.073050	0.023773	-0.138021
MARG_AL_3_6_M	0.093014	0.264409	0.153518	-0.038488	-0.007789	0.013477	0.063274
MARG_AL_3_6_F	0.051536	0.244261	0.256213	-0.179691	-0.061303	0.093993	-0.019221
MARG_HH_3_6_M	0.128576	0.158783	-0.200119	0.080411	0.008457	-0.144061	0.021451
MARG_HH_3_6_F	0.110646	0.125287	-0.279866	-0.136240	0.064109	-0.076708	-0.146121
MARG_OT_3_6_M	0.139593	0.062262	-0.020618	0.237745	-0.066400	0.097057	0.115068
MARG_OT_3_6_F	0.124546	0.014766	-0.082794	0.190511	-0.044810	0.384552	-0.042239
MARGWORK_0_3_M	0.154294	-0.093159	0.110285	0.086479	0.108829	-0.062043	-0.001204
MARGWORK_0_3_F	0.146286	-0.125596	0.095667	0.027275	0.141190	0.008962	-0.091060
MARG_CL_0_3_M	0.150126	0.150681	0.054892	0.087433	0.081185	-0.060715	-0.007316
MARG_CL_0_3_F	0.140157	0.180690	0.023982	-0.022290	0.129936	-0.001727	-0.200877
MARG_AL_0_3_M	0.052542	0.251328	0.268330	-0.104686	-0.048849	0.065409	-0.042295
MARG_AL_0_3_F	0.041786	0.240720	0.284956	-0.135716	-0.051895	0.083743	-0.103407
MARG_HH_0_3_M	0.121840	0.185277	-0.138628	0.132544	0.062380	-0.124209	0.014587
MARG_HH_0_3_F	0.116011	0.180616	-0.202198	0.004051	0.128308	-0.105530	-0.152175
MARG_OT_0_3_M	0.139869	0.084869	-0.022599	0.230038	-0.036390	0.061228	0.083122
MARG_OT_0_3_F	0.132192	0.050813	-0.078720	0.206201	0.000165	0.295600	-0.068722
NON_WORK_M	0.150376	-0.065365	0.111827	0.084854	0.162862	-0.052386	-0.019354
NON_WORK_F	0.131066	-0.073847	0.102553	0.021124	0.238292	-0.024901	-0.200053

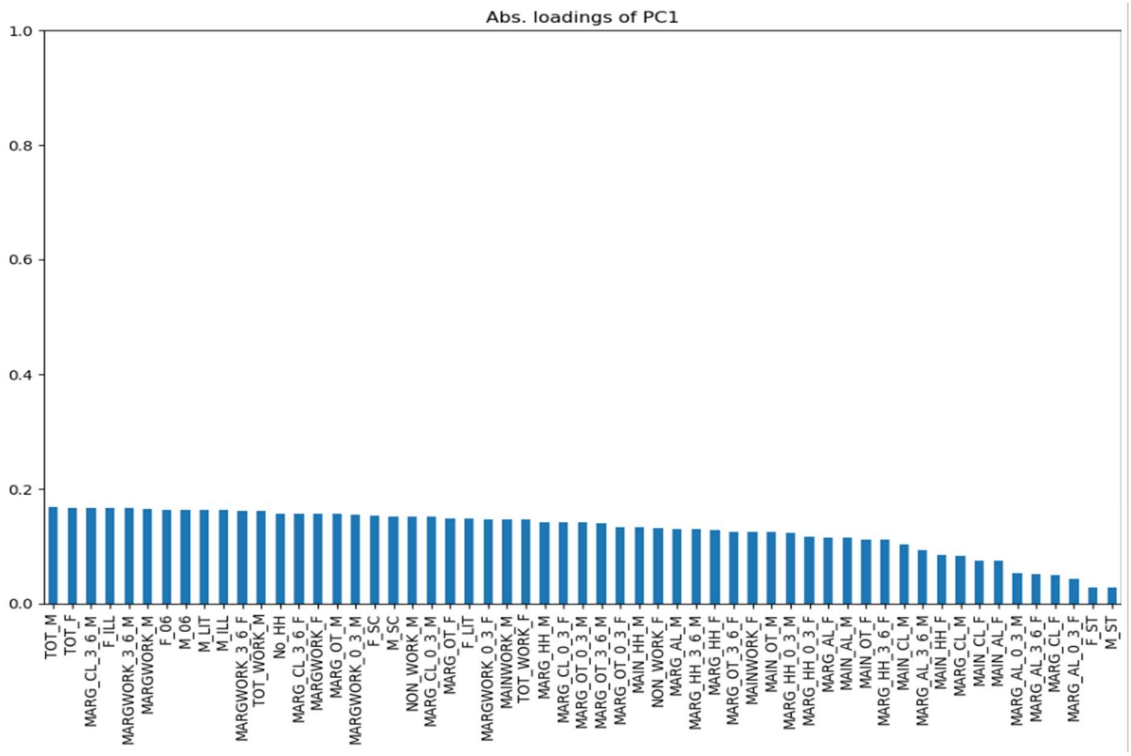
2.7 PCA: Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.

Ans:

- How the original features matter to each PC

Figure 31: Bar plots showing how the original features matter to each PC

PC1:

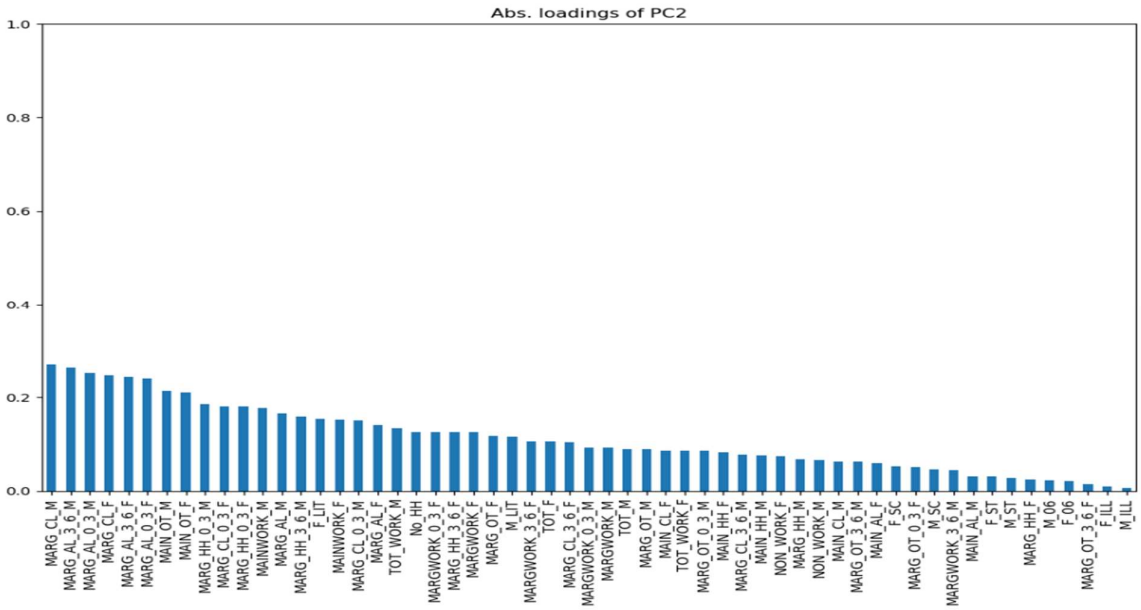


Observations:

- PC1 has a high value of ‘Total Male population’ followed by ‘Total Female population’.

- The values of ‘Scheduled Tribes male’ and ‘Scheduled Tribes female’ are at the bottom in PC1, of which ‘Scheduled Tribes’ male is the least.

PC2:

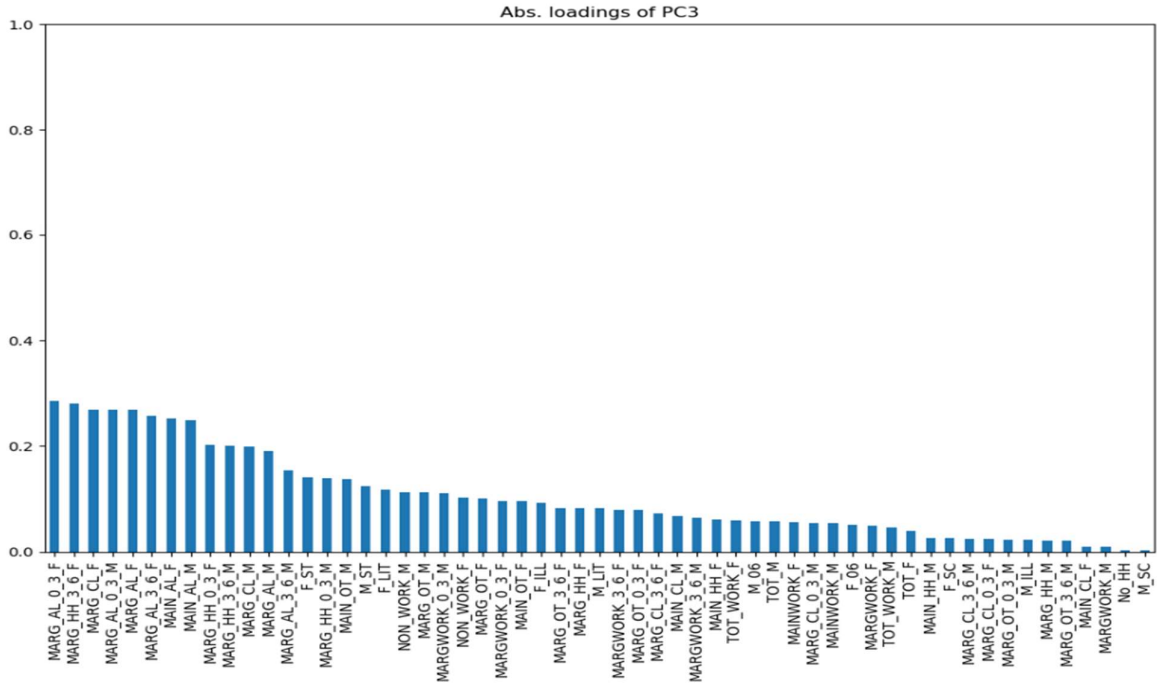


Observations:

- PC2 has a high value of ‘Marginal Cultivator Population Male’ followed by ‘Marginal Agriculture Labourers Population 3-6 Male’.

- The values of ‘Illiterate Female and Male’ are at the bottom in PC2, of which ‘Illiterate male’ is the least.

PC3:

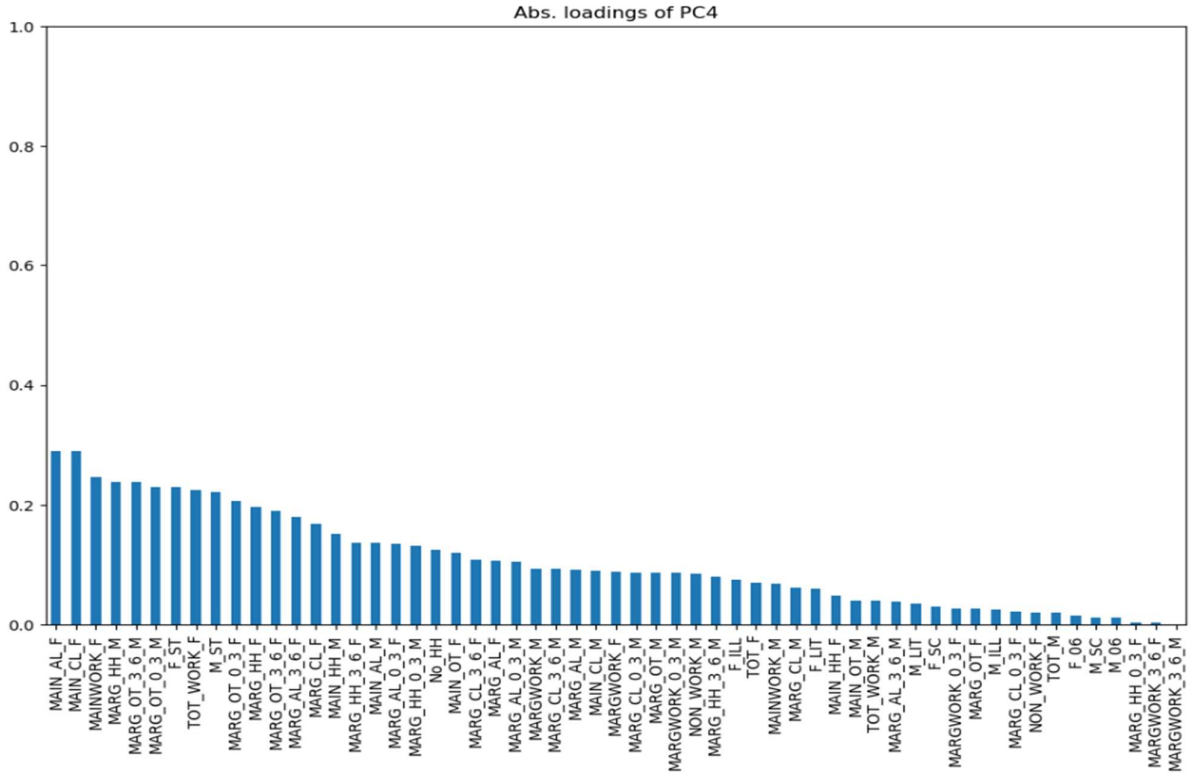


Observations:

- PC3 has a high value of ‘Marginal Agriculture Labourers Population 0-3 Female’ followed by ’Marginal Household Industries Population 3-6 Female’.

- The values of ‘No of Household’ and ‘Scheduled Castes population Male’ are at the bottom in PC3, of which ‘Scheduled Castes population Male’ is the least.

PC4:

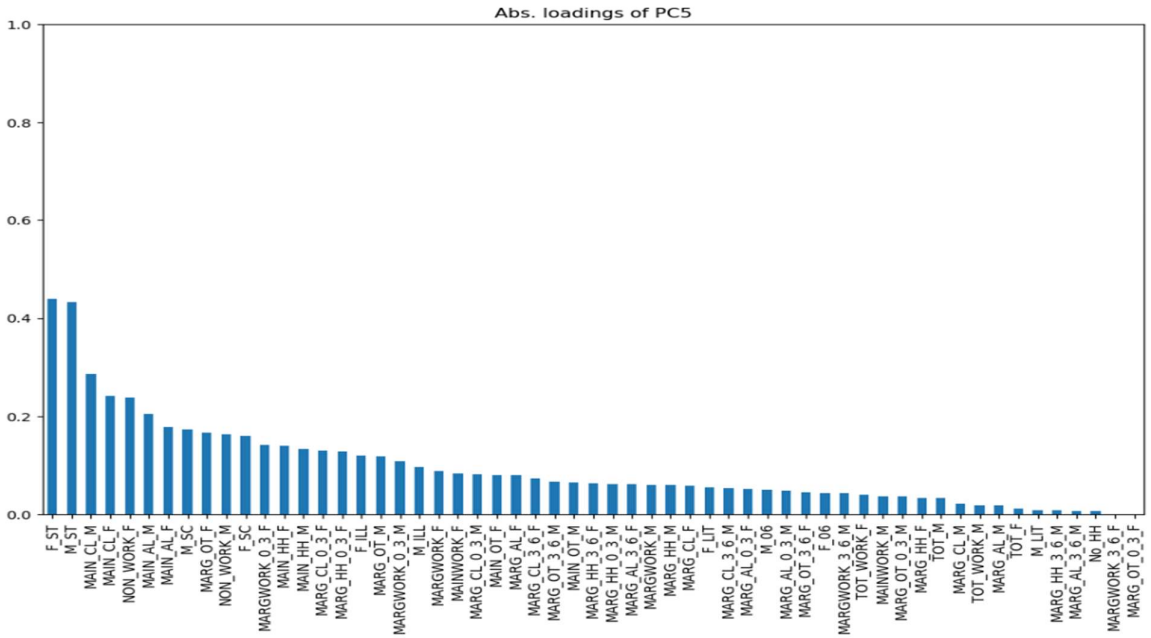


Observations:

- PC4 has a high value of ‘Main Agricultural Labourers Population Female’ followed by ‘Main Cultivator Population Female’.

- The values of ‘Marginal Worker Population 3-6 Female’ and ‘Marginal Worker Population 3-6 Male’ are at the bottom in PC4, of which ‘Marginal Worker Population 3-6 Male’ is the least.

PC5:

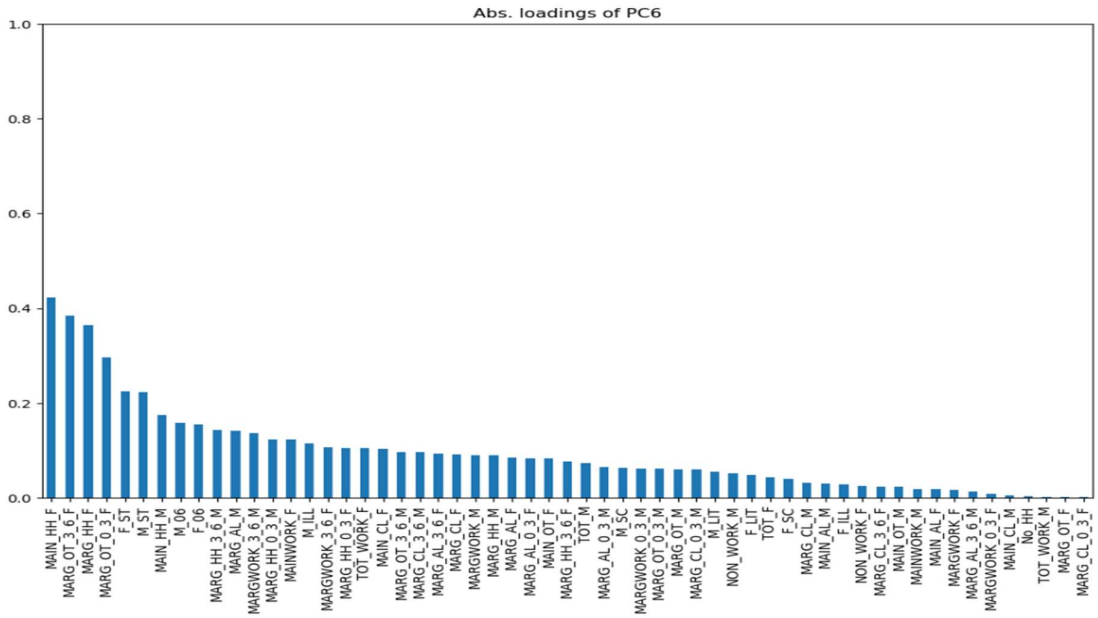


Observations:

- PC5 has a high value of ‘Scheduled Tribes population Female’ followed by ‘Scheduled Tribes population Male’.

- The values of ‘Marginal Worker Population 3-6 Female’ and ‘Marginal Other Workers Population 0-3 Female’ are at the bottom in PC5, of which ‘Marginal Other Workers Population 0-3 Female’ is the least.

PC6:

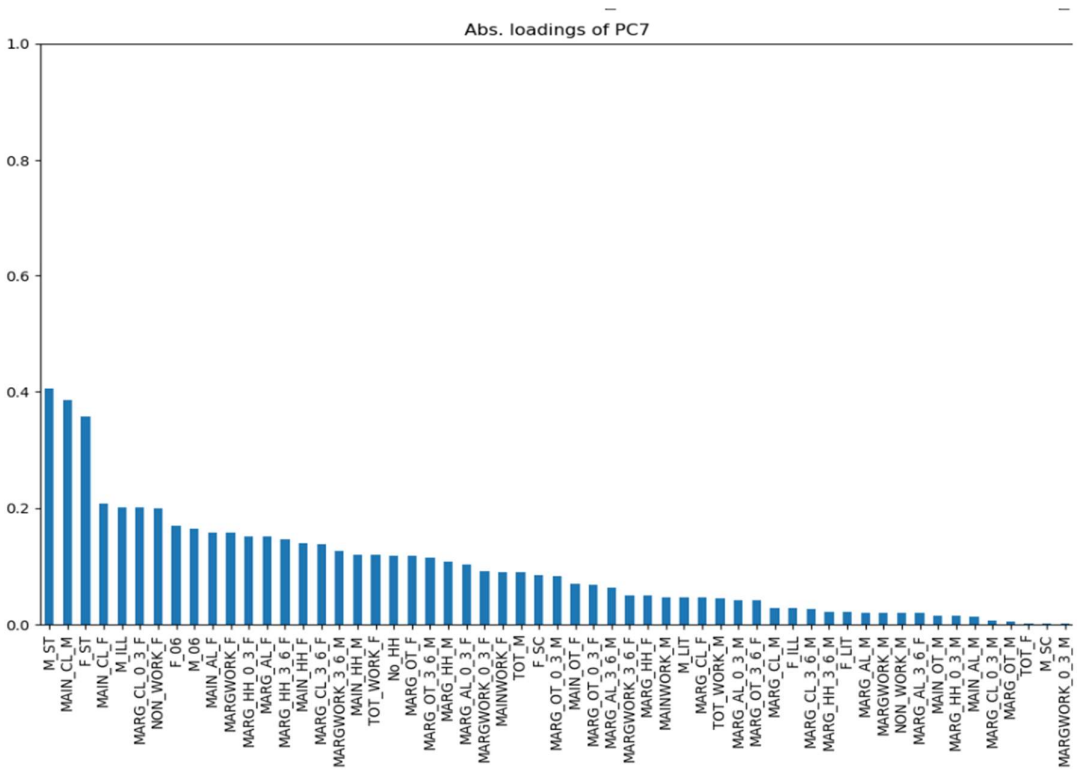


Observations:

- PC6 has a high value of 'Main Household Industries Population Female' followed by 'Marginal Other Workers Population Person 3-6 Female'.

- The values of 'Marginal Other Workers Population Female' and 'Marginal Cultivator Population 0-3 Female' are at the bottom in PC6, of which 'Marginal Cultivator Population 0-3 Female' is the least.

PC7:



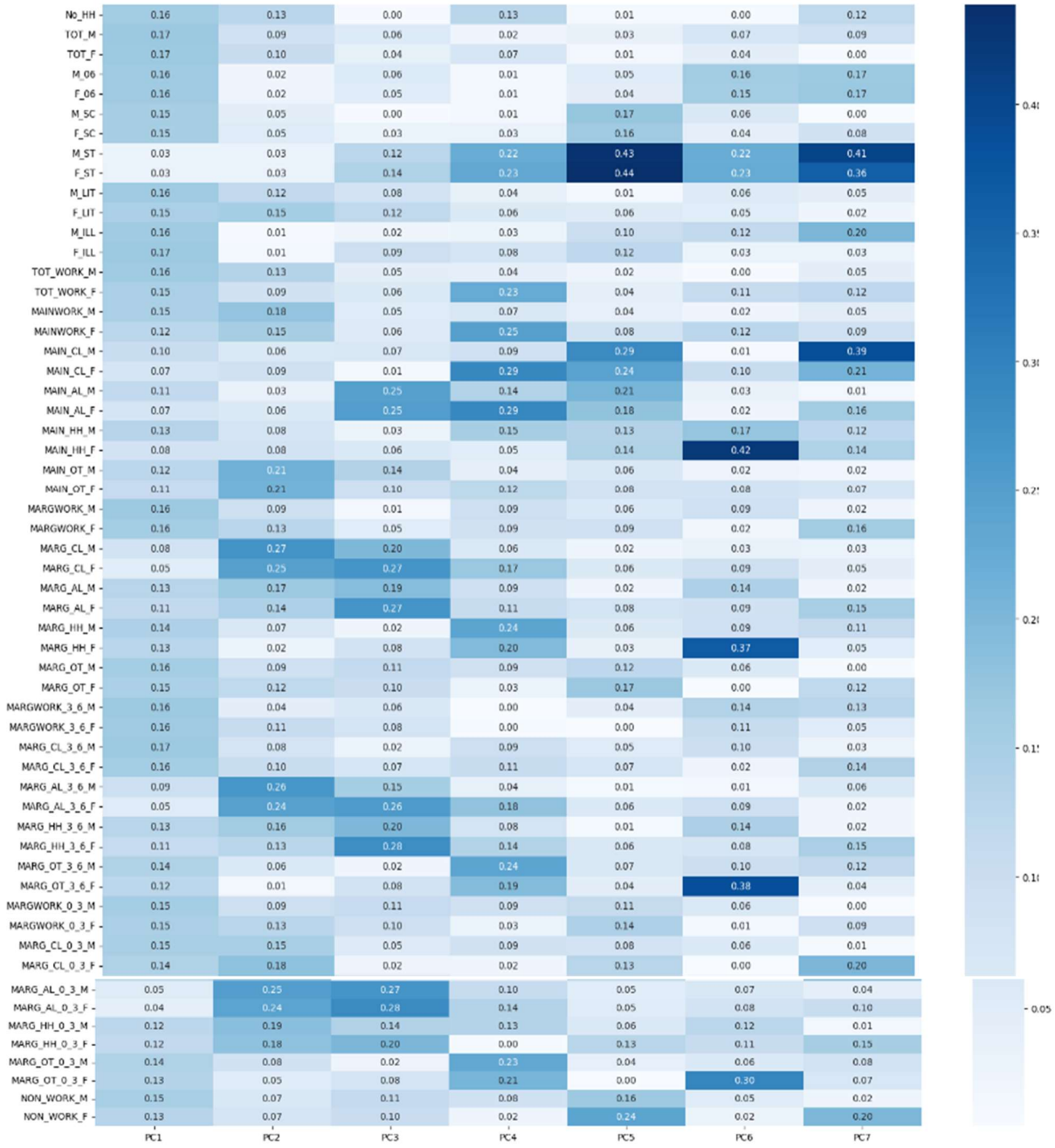
Observations:

- PC7 has a high value of 'Scheduled Tribes population Male' followed by 'Main Cultivator Population Male'.

- The values of 'Scheduled Castes population Male' and 'Marginal Worker Population 0-3 Male' are at the bottom in PC7, of which 'Marginal Worker Population 0-3 Male' is the least.

- Heatmap to show how the original features influence various PCs:

Figure 32: Heatmap showing how the original features influence various PCs



Observations:

- From the above heatmap, we can state that PC1 explains the most variance, as the weightage of PC1 alone comprises 55.73% of information of the data. PC1 has the highest magnitude of 'Total Male population' (0.167).

- 'Marginal Cultivator Population Male' (0.269) has the highest magnitude in PC2.

- 'Marginal Agriculture Labourers Population 0-3 Female' (0.284) has the highest magnitude in PC3.

- ‘Main Agricultural Labourers Population Female’ (0.290) has the highest magnitude in PC4.

- ‘Scheduled Tribes population Female’ (0.438), ‘Scheduled Tribes population Male’ (0.43) are the variables that have highest magnitudes in PC5.

- ‘Main Household Industries Population Female’ (0.422) has the highest magnitude in PC6.

- ‘Scheduled Tribes population Male’ (0.406) has the highest magnitude in PC7.

Even though there are few variables which has highest magnitude in different PCs, still priority should be given to PC1 as it contains 55.73% of information of the data.

Additionally,

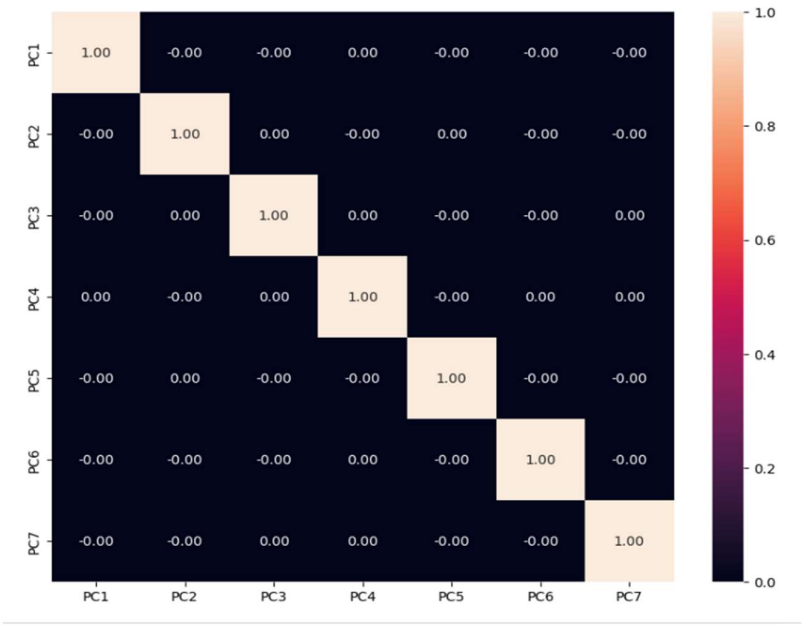
- PC scores

Table 34: Table showing PC scores

```
array([[-4.61726348,  0.13811585,  0.32854494, ...,  0.35373666,
        -0.42094718, -0.01038594],
       [-4.77166187, -0.10586534,  0.24444916, ..., -0.15388355,
         0.41731047, -0.02312096],
       [-5.96483558, -0.29434688,  0.36739364, ...,  0.47819919,
         0.27658083,  0.06955481],
       ...,
       [-6.294625  , -0.63812664,  0.10748287, ...,  0.15374553,
         0.14114567, -0.34450018],
       [-6.22319199, -0.67231967,  0.27132551, ...,  0.0604402 ,
        -0.11568169, -0.38303353],
       [-5.89623627, -0.93716953,  0.3492184 , ...,  0.14910383,
        -0.15454297, -0.38450817]])
```

- After creating a data frame of the PC scores, next step is to check if there are any correlation between the PCs.

Figure 33: Heatmap showing correlation between the PCs



Dimensions reduced from 57 variables to 7 PCs and correlation between the PCs are almost 0.

2.8 PCA: Write linear equation for first PC.

Ans:

Linear equation for first PC

PC1 = a1X1 + a2X2 + a3X3 + a4X4 + ...+ a57X57

i.e., **a** = eigen vectors, **X** = Variables after scaling the data

PC1 = (0.16) * No_HH + (0.17) * TOT_M + (0.17) * TOT_F+……+ (0.13) * NON_WORK_F

Table 35: linear equation for first PC

```
( 0.16 ) * No_HH+( 0.17 ) * TOT_M+( 0.17 ) * TOT_F+( 0.16 ) * M_06+( 0.16 ) * F_06+( 0.15 ) * M_SC+( 0.15 ) * F_SC+( 0.03 ) * M
_ST+( 0.03 ) * F_ST+( 0.16 ) * M_LIT+( 0.15 ) * F_LIT+( 0.16 ) * M_ILL+( 0.17 ) * F_ILL+( 0.16 ) * TOT_WORK_M+( 0.15 ) * TOT_WO
RK_F+( 0.15 ) * MAINWORK_M+( 0.12 ) * MAINWORK_F+( 0.1 ) * MAIN_CL_M+( 0.07 ) * MAIN_CL_F+( 0.11 ) * MAIN_AL_M+( 0.07 ) * MAIN_
AL_F+( 0.13 ) * MAIN_HH_M+( 0.08 ) * MAIN_HH_F+( 0.12 ) * MAIN_OT_M+( 0.11 ) * MAIN_OT_F+( 0.16 ) * MARGWORK_M+( 0.16 ) * MARGW
ORK_F+( 0.08 ) * MARG_CL_M+( 0.05 ) * MARG_CL_F+( 0.13 ) * MARG_AL_M+( 0.11 ) * MARG_AL_F+( 0.14 ) * MARG_HH_M+( 0.13 ) * MARG_
HH_F+( 0.16 ) * MARG_OT_M+( 0.15 ) * MARG_OT_F+( 0.16 ) * MARGWORK_3_6_M+( 0.16 ) * MARGWORK_3_6_F+( 0.17 ) * MARG_CL_3_6_M+(
0.16 ) * MARG_CL_3_6_F+( 0.09 ) * MARG_AL_3_6_M+( 0.05 ) * MARG_AL_3_6_F+( 0.13 ) * MARG_HH_3_6_M+( 0.11 ) * MARG_HH_3_6_F+( 0.
14 ) * MARG_OT_3_6_M+( 0.12 ) * MARG_OT_3_6_F+( 0.15 ) * MARGWORK_0_3_M+( 0.15 ) * MARGWORK_0_3_F+( 0.15 ) * MARG_CL_0_3_M+( 0.
14 ) * MARG_CL_0_3_F+( 0.05 ) * MARG_AL_0_3_M+( 0.04 ) * MARG_AL_0_3_F+( 0.12 ) * MARG_HH_0_3_M+( 0.12 ) * MARG_HH_0_3_F+( 0.14
) * MARG_OT_0_3_M+( 0.13 ) * MARG_OT_0_3_F+( 0.15 ) * NON_WORK_M+( 0.13 ) * NON_WORK_F+
```

----------THE END----------