Data Mining

Business Report

Pavithra Devi
DSBA
Great Learning

Index

Part 2 - PCA: Perform all the required steps for PCA.Create the covariance Matrix Get eigen values and eigen vector.

Part 2 - PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.

Part 2 - 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.

Part 2 - PCA: Write linear equation for first PC.

Problem:1
Clustering:

Data Dictionary

1	Timestamp	The Timestamp of the particular Advertisement.
2	InventoryType	The Inventory Type of the particular Advertisement. Format 1 to 7. This is a Categorical Variable.
3	Ad - Length	The Length Dimension of the particular Adverstisement.
4	Ad- Width	The Width Dimension of the particular Advertisement.
5	Ad Size	The Overall Size of the particular Advertisement. Length*Width.
6	Ad Type	The type of the particular Advertisement. This is a Categorical Variable.
7	Platform	The platform in which the particular Advertisement is displayed. Web, Video or App. This is a Categorical Variable.
8	Device Type	The type of the device which supports the partciular Advertisement. This is a Categorical Variable.
9	Format	The Format in which the Advertisement is displayed. This is a Categorical Variable.
10	Available_Impressions	How often the particular Advertisement is shown. An impression is counted each time an Advertisement is shown on a search result page or other site on a Network.
11	Matched_Queries	Matched search queries data is pulled from Advertising Platform and consists of the exact searches typed into the search Engine that generated clicks for the particular Advertisement.
12	Impressions	The impression count of the particular Advertisement out of the total available impressions.
13	Clicks	It is a marketing metric that counts the number of times users have clicked on the particular advertisement to reach an online property.
14	Spend	It is the amount of money spent on specific ad variations within a specific campaign or ad set. This metric helps regulate ad performance.
15	Fee	The percentage of the Advertising Fees payable by Franchise Entities.
16	Revenue	It is the income that has been earned from the particular advertisement.
17	CTR	CTR stands for "Click through rate". CTR is the number of clicks that your ad receives divided by the number of times your ad is shown. Formula used here is CTR = Total Measured Clicks / Total Measured Ad Impressions x 100. Note that the Total Measured Clicks refers to the 'Clicks'

		Column and the Total Measured Ad Impressions refers to the 'Impressions' Column.
18	CPM	CPM stands for "cost per 1000 impressions." Formula used here is CPM = (Total Campaign Spend / Number of Impressions) * 1,000. Note that the Total Campaign Spend refers to the 'Spend' Column and the Number of Impressions refers to the 'Impressions' Column.
19	CPC	CPC stands for "Cost-per-click". Cost-per-click (CPC) bidding means that you pay for each click on your ads. The Formula used here is CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column and the Number of Clicks refers to the 'Clicks' Column.

Part 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.

```
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)
```

Dataset overview:
The dataset has 23066 rows and 19 columns, having datatypes: 6 float, 7 integer and 6 object variables, having 4736 null values, in the CTR, CPM, CPC columns which can be treated using the given formula and the dataset has no duplicate values.

	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

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

Given:
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.

The columns are treated using the above given formulas. Now, we have zero null values.

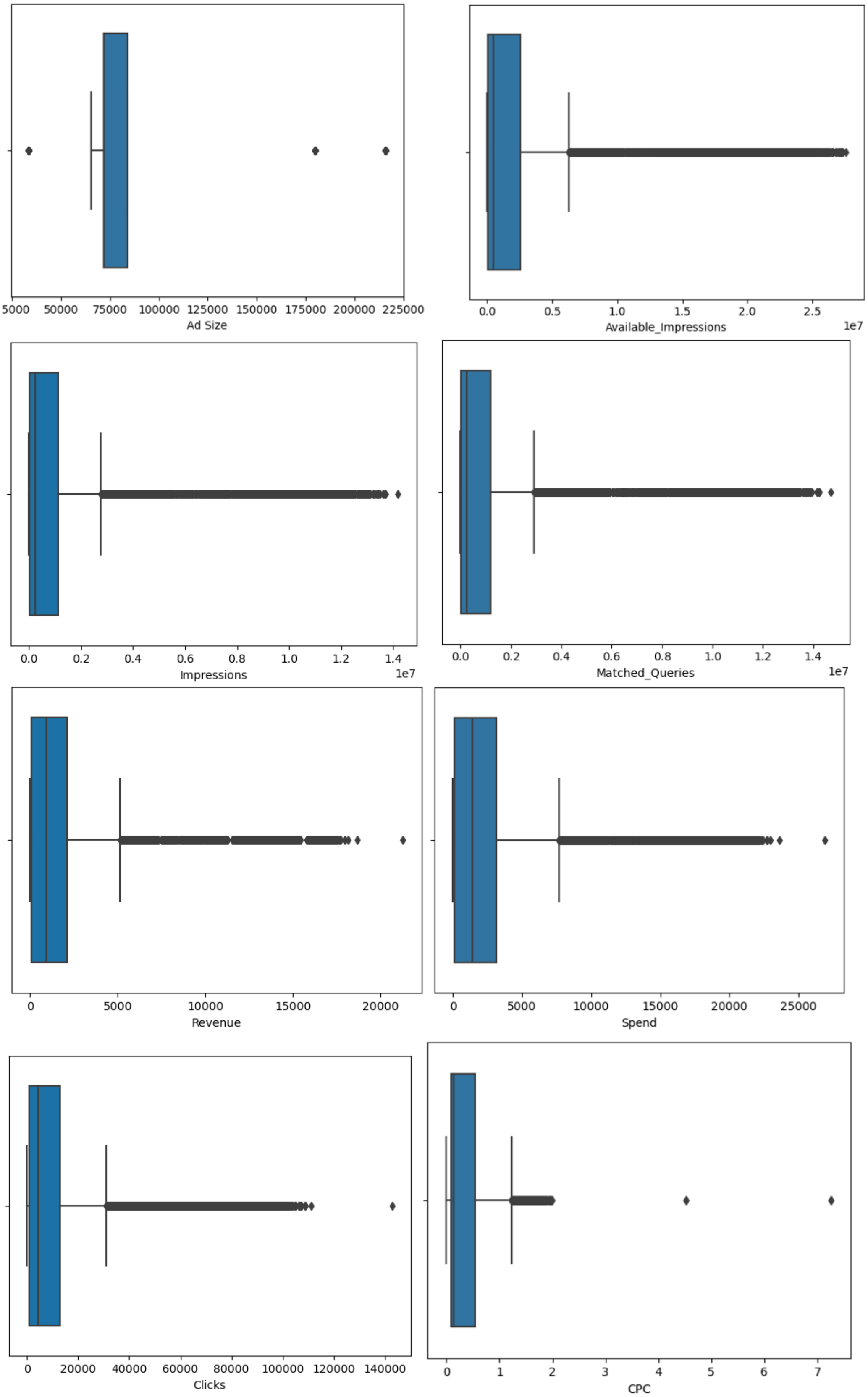
```
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   InventoryType         23066 non-null  object
 1   Ad - Length           23066 non-null  int64
 2   Ad- Width             23066 non-null  int64
 3   Ad Size               23066 non-null  int64
 4   Ad Type               23066 non-null  object
 5   Platform              23066 non-null  object
 6   Device Type           23066 non-null  object
 7   Format                23066 non-null  object
 8   Available_Impressions 23066 non-null  int64
 9   Matched_Queries       23066 non-null  int64
 10  Impressions           23066 non-null  int64
 11  Clicks                23066 non-null  int64
 12  Spend                 23066 non-null  float64
 13  Fee                   23066 non-null  float64
 14  Revenue               23066 non-null  float64
 15  CTR                   23066 non-null  float64
 16  CPM                   23066 non-null  float64
 17  CPC                   23066 non-null  float64
```

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).

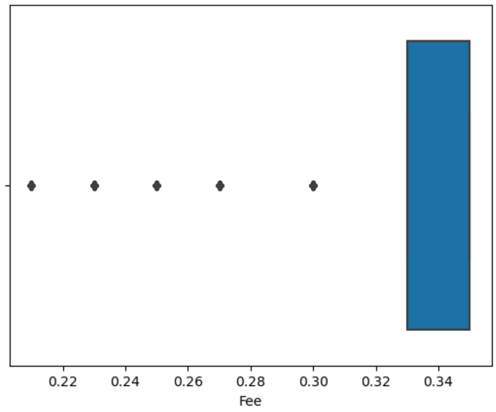
Yes, there are outliers in the given dataset. Outliers can distort the distance-based calculations that clustering algorithms like K-Means rely on. K-Means is sensitive to the scale and distribution of features. Outliers can disproportionately affect cluster centroids and distances between data points, leading to suboptimal clustering results.

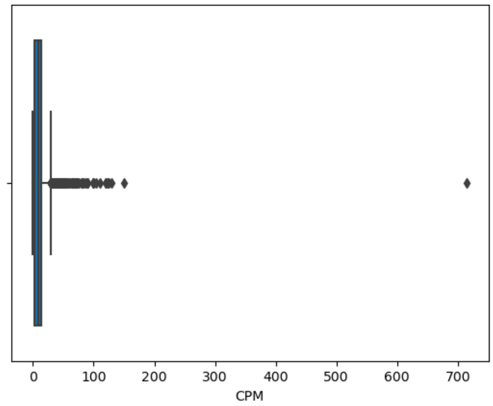
Therefore, it is necessary to treat outliers in K-Means clustering.
For the given data, the treatment of outliers is done via IQR method.

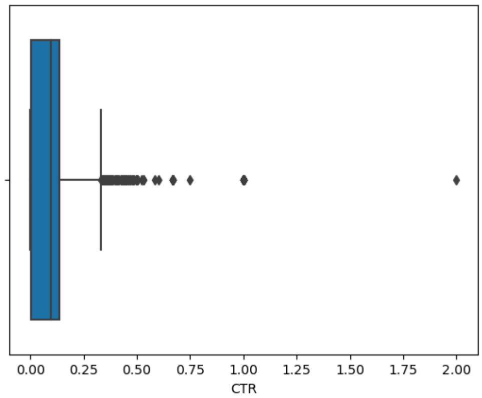
Before Treating Outliers:



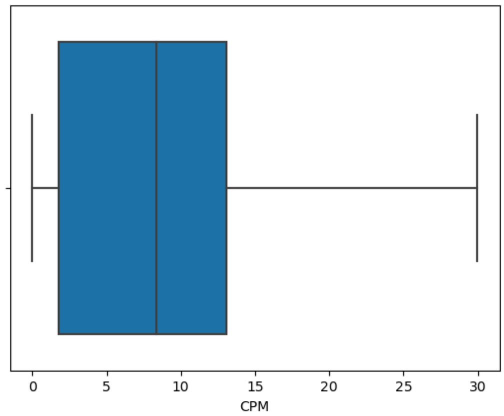
5000 50000 75000 100000 125000 150000 175000 200000 225000
Ad Size
0.0 0.5 1.0 1.5 2.0 2.5
1e7
Available_Impressions
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4
1e7
Impressions
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4
1e7
Matched_Queries
0 5000 10000 15000 20000
Revenue
0 5000 10000 15000 20000 25000
Spend
0 20000 40000 60000 80000 100000 120000 140000
Clicks
0 1 2 3 4 5 6 7
CPC

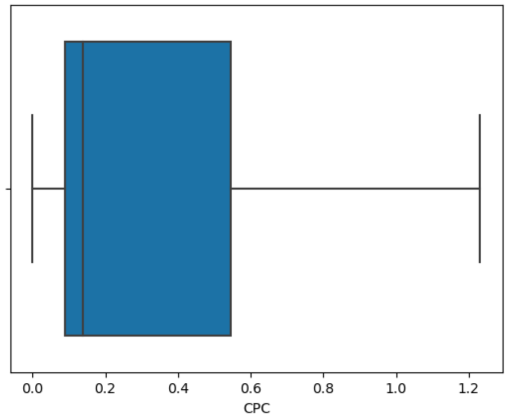


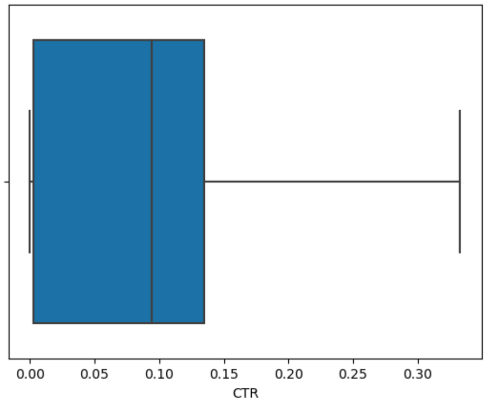


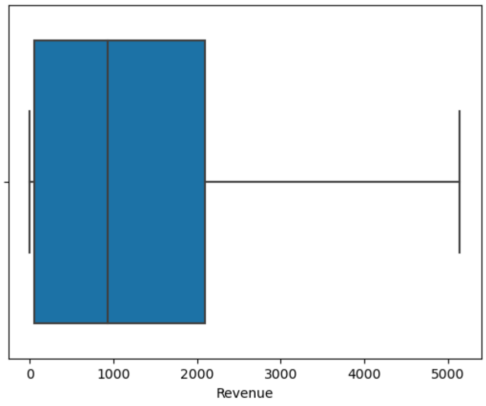


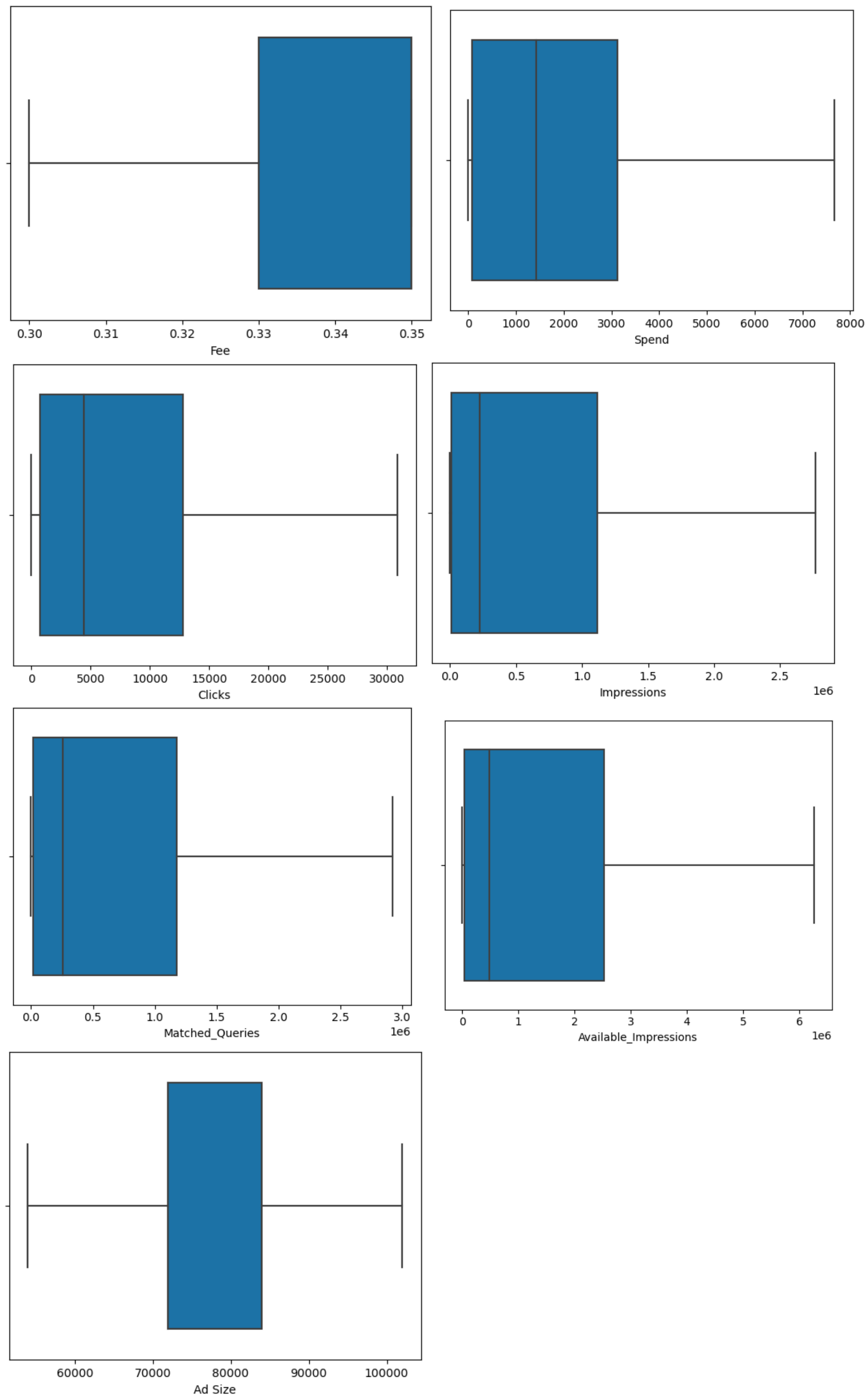
After Treating Outliers:










0.30
0.31
0.32
0.33
0.34
0.35
Fee
0
1000
2000
3000
4000
5000
6000
7000
8000
Spend
0
5000
10000
15000
20000
25000
30000
Clicks
0.0
0.5
1.0
1.5
2.0
2.5
1e6
Impressions
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1e6
Matched_Queries
0
1
2
3
4
5
6
1e6
Available_Impressions
60000
70000
80000
90000
100000
Ad Size

Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance, and identify optimum number of clusters

In this analysis, we aimed to uncover meaningful patterns and groupings within our dataset using hierarchical clustering, a technique that creates a hierarchical structure of clusters through iterative merging. To determine the ideal number of clusters, we constructed a dendrogram using the Ward linkage method with the Euclidean distance metric, offering insights into the data's intrinsic structure.

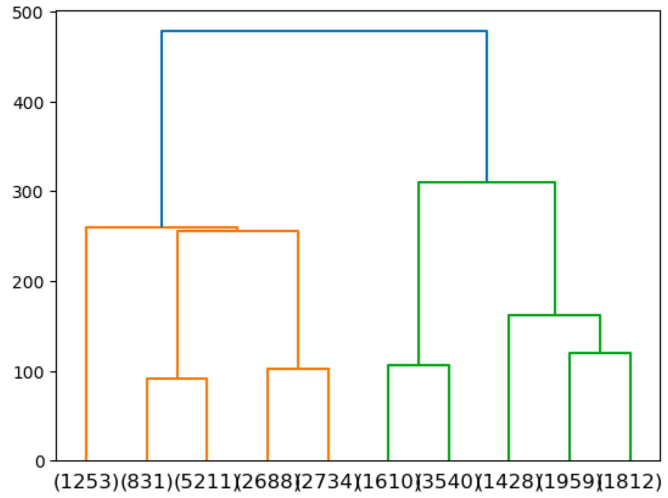
We proceeded with the following steps to achieve this:

1. Data Preparation:
We began by preparing our dataset, ensuring it was appropriately scaled and standardized. This step ensures that each feature contributes equitably to the clustering process and that no single feature dominates the analysis due to its scale.

2. Hierarchical Clustering:
Leveraging the hierarchical clustering technique, we computed the pairwise distances between data points using the Euclidean distance metric. The Ward linkage method was employed to merge clusters in a way that minimizes the variance within merged clusters. This approach helps preserve the tightness of clusters while facilitating their interpretation.

3. Dendrogram Construction:
By plotting the resulting dendrogram, we visualized the hierarchical structure of clusters, with the y-axis representing the distance between clusters. We employed the Ward linkage method to create the dendrogram, allowing us to observe the order and scale of cluster mergers.

4. Optimal Number of Clusters:
To identify the optimal number of clusters, we examined the dendrogram's structure for a point where a significant jump in distances occurs between successive merges. This jump signifies the transition from merging individual data points to merging distinct clusters. By drawing a horizontal line through the dendrogram and noting the intersections with vertical branches, we determined the number of clusters that best captures the underlying patterns in the data.



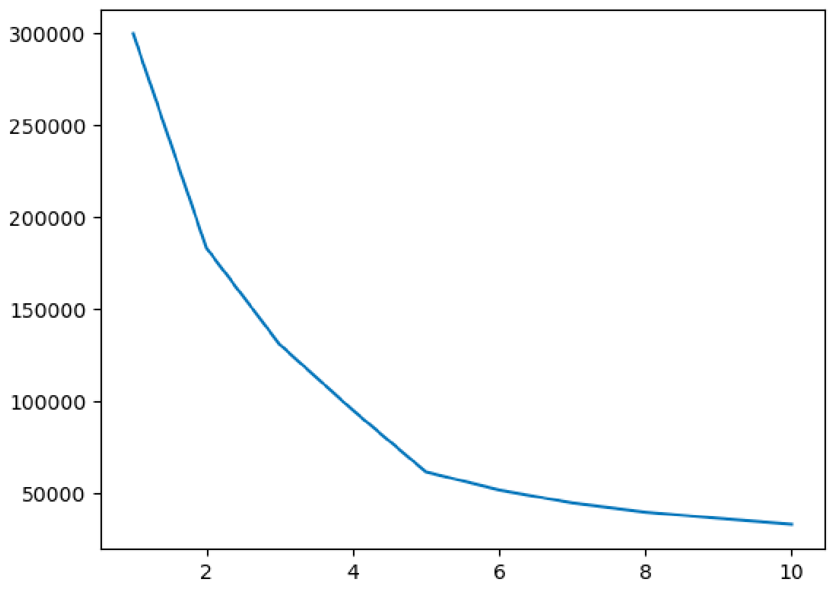
Draw a horizontal line through the dendrogram, intersecting with the tallest vertical lines. The number of intersections indicates the potential number of clusters.
Here, after the intersection we get 4 potential clusters.

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

The point at which the curve starts to level off and the rate of decrease becomes less pronounced is the "elbow" point. This point indicates the optimal number of clusters.

Here, the drop rate decreases post k= 4.
To intricate:
When k=1; Inertia= 299858
K=2;Inertia= 183349
K=3;Inertia= 140536
K=4;Inertia= 95133
K=5;Inertia= 61539
K=6;Inertia= 51676

Therefore, the optimal number of clusters is 4.



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

Silhouette score measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). Higher silhouette scores indicate better-defined clusters.

Silhouette Score: 0.44534519247649873

Clusters	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	sil_width	freq
0	145.392650	570.277240	73907.156673	8.025707e+05	5.641802e+05	4.759313e+05	30551.916344	6522.059774	0.305764	4454.785247	0.137587	15.397786	0.111971	0.681694	1551
1	423.971442	144.004543	63625.020282	1.812893e+06	8.662461e+05	8.282049e+05	3256.069609	1502.278355	0.349267	978.838843	0.004028	1.786078	0.529599	0.507415	6163
2	368.043209	461.545954	85302.904197	1.374760e+05	7.477135e+04	6.229995e+04	6954.156267	647.116196	0.349828	421.191764	0.146071	13.131697	0.100728	0.363854	11294
3	465.513061	199.411040	72981.586989	5.693098e+06	2.805313e+06	2.670479e+06	11239.989897	5736.706676	0.313297	3876.959415	0.002173	1.573353	0.748490	0.487546	4058

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].

The clusters are grouped based on the device type, the average CPC, Spend, Revenue, Clicks, CTR, CPM….

```
Clusters  Device Type
0         Mobile         990
          Desktop        561
1         Mobile        3977
          Desktop       2186
2         Mobile        7248
          Desktop       4046
3         Mobile        2591
          Desktop       1467
```

```
Clusters
0    0.111971
1    0.529599
2    0.100728
3    0.748490
Name: CPC, dtype: float64
```

```
Clusters
0    6522.059774
1    1502.278355
2     647.116196
3    5736.706676
Name: Spend, dtype: float64
```

```
Clusters
0    4454.785247
1     978.838843
2     421.191764
3    3876.959415
Name: Revenue, dtype: float64
```

```
Clusters
0    30551.916344
1     3256.069609
2     6954.156267
3    11239.989897
Name: Clicks, dtype: float64
```

```
Clusters
0    0.137587
1    0.004028
2    0.146071
3    0.002173
Name: CTR, dtype: float64
```

```
Clusters
0    15.397786
1     1.786078
2    13.131697
3     1.573353
Name: CPM, dtype: float64
```

1. CTR (Click-Through Rate):
   - Cluster 3 shows the lowest CTR on an average, suggesting that ads in this cluster have relatively lower user engagement and click-through activity.
   - Cluster 2 exhibits the highest CTR on an average, indicating that ads in this cluster are particularly effective in capturing user attention and driving clicks.

2. CPM (Cost Per Mille):
   - Cluster 3 demonstrates the lowest CPM on an average, implying that ads in this cluster are cost-effective in terms of impressions generated.
   - Cluster 0 registers the highest CPM, suggesting that ads in this cluster have a higher cost per impression.

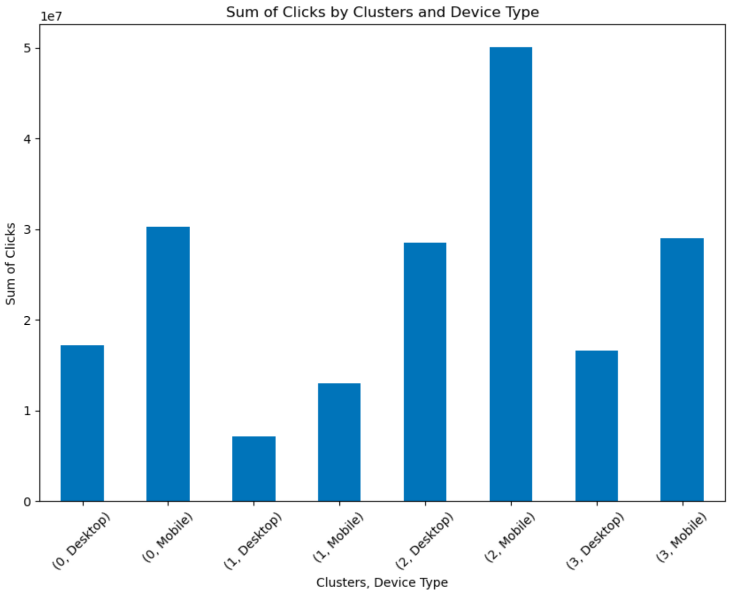
3. CPC (Cost Per Click):
   - Cluster 2 boasts the lowest CPC on an average, indicating that ads in this cluster are efficient in terms of cost per click generated.
   - Cluster 3 presents the highest CPC, suggesting that ads in this cluster are relatively costlier per click achieved.

Engagement and Performance:

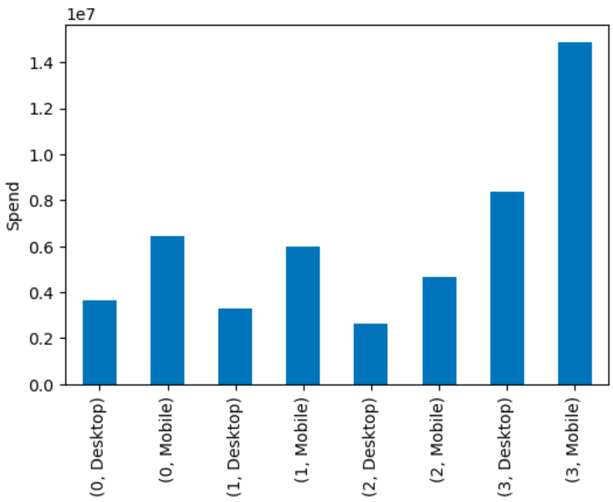
1. Engagement in Cluster 3:
   - Ads in cluster 3 demonstrate robust engagement levels across both "Mobile" and "Desktop" devices on an average.

2. Spend and Revenue:
   - Cluster 0 reflects the highest spend on an average, implying that resources are allocated to this cluster to capture a significant share of impressions.
   - Correspondingly, cluster 0 also records the highest revenue, suggesting that the higher investment in this cluster results in substantial returns.

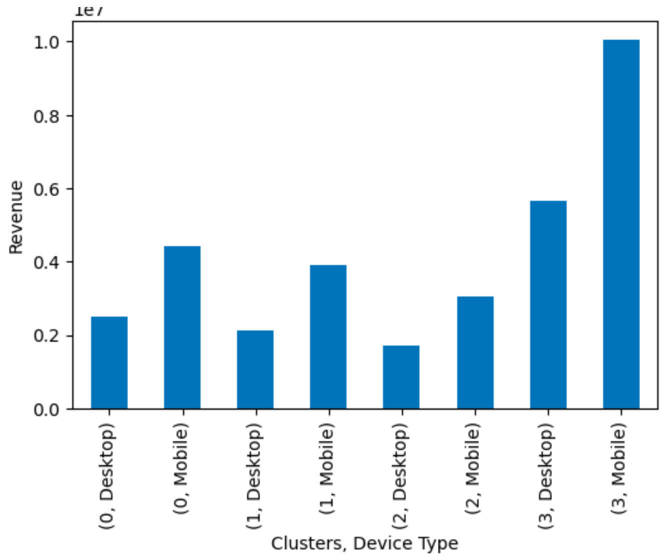
3. Clicks Distribution:
   - Cluster 0 on an average stands out with the highest number of clicks, showcasing the effectiveness of ads in generating clicks and user interest.
   - In contrast, cluster 1 records the lowest number of clicks, indicating a comparatively lower level of user engagement.



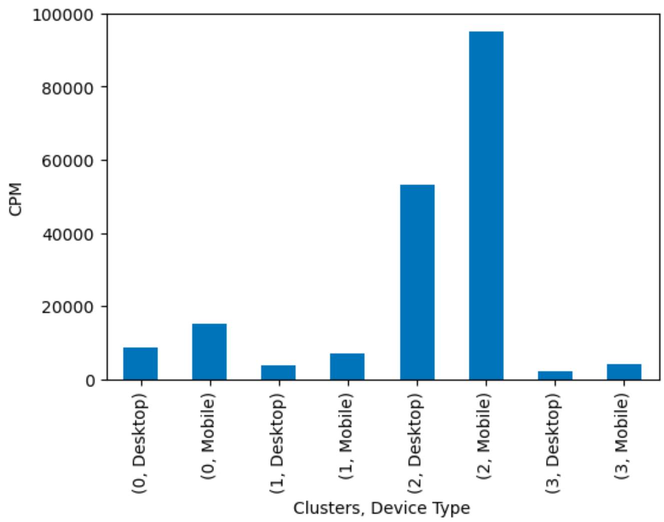
The data depicted in the above graph highlights that the cluster labeled as "2" records the highest number of clicks on both the "Mobile" and "Desktop" platforms. Conversely, the lowest number of clicks is observed within cluster "1."



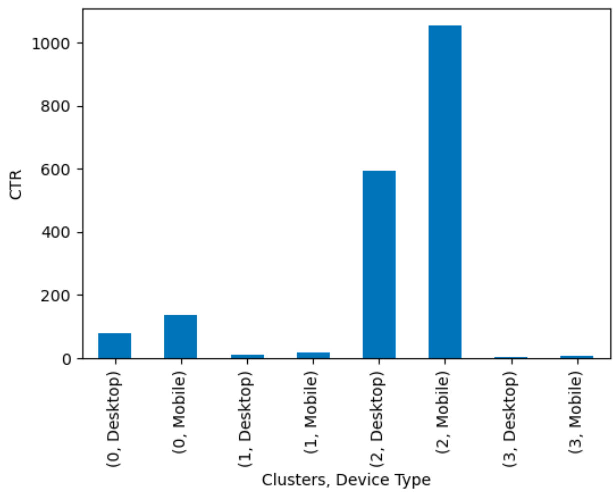
The data depicted in the above graph highlights that the cluster labeled as "3" records the highest number of Spend on both the "Mobile" and "Desktop" platforms. Conversely, the lowest number of Spend is observed within cluster "2."



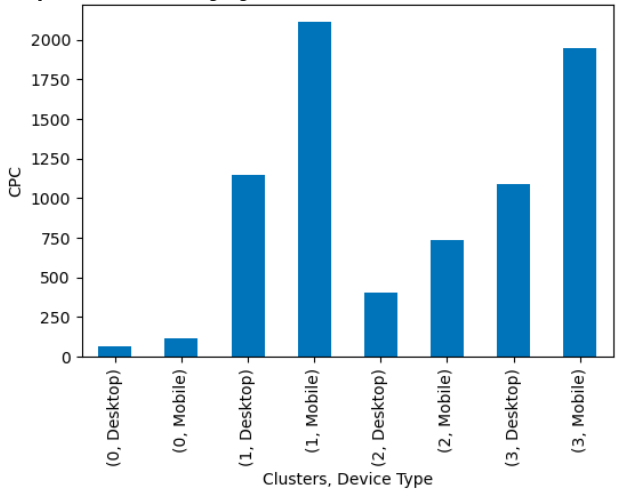
The data depicted in the above graph highlights that the cluster labeled as "3" records the highest Revenue on both the "Mobile" and "Desktop" platforms. Conversely, the lowest Revenue is observed within cluster "2."
Also, good revenue is reaped via ads through Mobile platforms.



Cluster 2 demonstrates the highest CPM value, signifying a relatively elevated cost per mille (CPM) for ads within this cluster. Conversely, cluster 3 exhibits the lowest CPM value, indicating a comparatively lower cost per mille for ads in this particular cluster.



In cluster 2, the click-through rate (CTR) reaches its peak, reflecting the highest level of engagement and interaction with ads. On the other hand, cluster 3 experiences the lowest CTR, suggesting relatively subdued engagement levels within this cluster.



Cluster 1 and cluster 3 demonstrate favorable cost-per-click (CPC) values, indicating efficient spending in relation to the number of clicks generated within these clusters. Conversely, cluster 0 registers the lowest CPC, highlighting cost-effective results in terms of clicks for this cluster.

Conclude the project by providing summary of your learnings.

Conclusions from Cluster Analysis:

The comprehensive analysis of clusters reveals valuable insights into the performance and engagement levels of different ad types within the ads24x7 Digital Marketing company's dataset. These findings offer actionable information to optimize strategies and resource allocation for marketing campaigns:

1. Clicks and Engagement:
   - Cluster 2 emerges as the standout performer, recording the highest number of clicks across both "Mobile" and "Desktop" platforms. Conversely, cluster 1 exhibits the lowest click count, suggesting a need for refining the engagement strategies in this cluster.

2. Spend and Revenue:
- Within cluster 3, the highest levels of spend are observed on both "Mobile" and "Desktop" platforms, indicating a substantial investment in this cluster. Correspondingly, cluster 3 also records the highest revenue, demonstrating the effectiveness of this investment in driving revenue generation.

3.Device-Specific Revenue:
- Notably, the analysis highlights that good revenue is generated through ads on the "Mobile" platform. This underscores the importance of optimizing strategies specifically for mobile users to further capitalize on this revenue source.

4. CPM Variation:
- Cluster 2 exhibits the highest CPM, indicating a relatively higher cost per mille for ads in this cluster. Conversely, cluster 3 showcases the lowest CPM, suggesting that ads in this cluster achieve impressions at a relatively lower cost.

5. CTR and Engagement:
- Cluster 2 shines once again with the highest click-through rate (CTR), indicating a superior level of user engagement. Cluster 3, while registering a lower CTR, still presents meaningful engagement levels.

6. Efficient CPC:
- Clusters 1 and 3 demonstrate favorable cost-per-click (CPC) values, signifying efficient spending in relation to the clicks generated. Cluster 0, characterized by the lowest CPC, underscores the cost-effective outcomes achieved within this cluster.

In conclusion, the insights gained from this cluster analysis provide actionable recommendations for refining ad strategies and resource allocation. By focusing on high-performing clusters, optimizing engagement in clusters with lower performance, and capitalizing on device-specific revenue trends, ads24x7 can strategically enhance their marketing campaigns.

Problem 2:
PCA

DATA DICTIONARY:

Name	Description
State	State Code
District	District Code
Name	Name
TRU1	Area Name
No_HH	No of Household
TOT_M	Total population Male
TOT_F	Total population Female
M_06	Population in the age group 0-6 Male
F_06	Population in the age group 0-6 Female
M_SC	Scheduled Castes population Male
F_SC	Scheduled Castes population Female
M_ST	Scheduled Tribes population Male
F_ST	Scheduled Tribes population Female
M_LIT	Literates population Male
F_LIT	Literates population Female
M_ILL	Illiterate Male
F_ILL	Illiterate Female
TOT_WORK_M	Total Worker Population Male
TOT_WORK_F	Total Worker Population Female
MAINWORK_M	Main Working Population Male
MAINWORK_F	Main Working Population Female
MAIN_CL_M	Main Cultivator Population Male
MAIN_CL_F	Main Cultivator Population Female
MAIN_AL_M	Main Agricultural Labourers Population Male
MAIN_AL_F	Main Agricultural Labourers Population Female
MAIN_HH_M	Main Household Industries Population Male
MAIN_HH_F	Main Household Industries Population Female
MAIN_OT_M	Main Other Workers Population Male
MAIN_OT_F	Main Other Workers Population Female
MARGWORK_M	Marginal Worker Population Male
MARGWORK_F	Marginal Worker Population Female
MARG_CL_M	Marginal Cultivator Population Male
MARG_CL_F	Marginal Cultivator Population Female
MARG_AL_M	Marginal Agriculture Labourers Population Male
MARG_AL_F	Marginal Agriculture Labourers Population Female
MARG_HH_M	Marginal Household Industries Population Male
MARG_HH_F	Marginal Household Industries Population Female
MARG_OT_M	Marginal Other Workers Population Male
MARG_OT_F	Marginal Other Workers Population Female
MARGWORK_3_6_M	Marginal Worker Population 3-6 Male
MARGWORK_3_6_F	Marginal Worker Population 3-6 Female

MARG_CL_3_6_M	Marginal Cultivator Population 3-6 Male
MARG_CL_3_6_F	Marginal Cultivator Population 3-6 Female
MARG_AL_3_6_M	Marginal Agriculture Labourers Population 3-6 Male
MARG_AL_3_6_F	Marginal Agriculture Labourers Population 3-6 Female
MARG_HH_3_6_M	Marginal Household Industries Population 3-6 Male
MARG_HH_3_6_F	Marginal Household Industries Population 3-6 Female
MARG_OT_3_6_M	Marginal Other Workers Population Person 3-6 Male
MARG_OT_3_6_F	Marginal Other Workers Population Person 3-6 Female
MARGWORK_0_3_M	Marginal Worker Population 0-3 Male
MARGWORK_0_3_F	Marginal Worker Population 0-3 Female
MARG_CL_0_3_M	Marginal Cultivator Population 0-3 Male
MARG_CL_0_3_F	Marginal Cultivator Population 0-3 Female
MARG_AL_0_3_M	Marginal Agriculture Labourers Population 0-3 Male
MARG_AL_0_3_F	Marginal Agriculture Labourers Population 0-3 Female
MARG_HH_0_3_M	Marginal Household Industries Population 0-3 Male
MARG_HH_0_3_F	Marginal Household Industries Population 0-3 Female
MARG_OT_0_3_M	Marginal Other Workers Population 0-3 Male
MARG_OT_0_3_F	Marginal Other Workers Population 0-3 Female
NON_WORK_M	Non Working Population Male
NON_WORK_F	Non Working Population Female

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

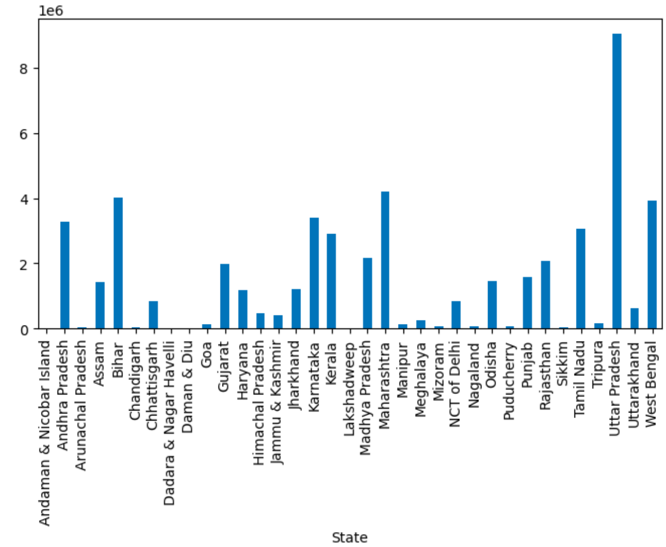
```
Data columns (total 61 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   State Code      640 non-null    int64
 1   Dist.Code       640 non-null    int64
 2   State           640 non-null    object
 3   Area Name       640 non-null    object
 4   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)
```

The dataset has 61 columns and 640 rows, having 0 null and duplicate values, with 59 intege r and 2 object values.

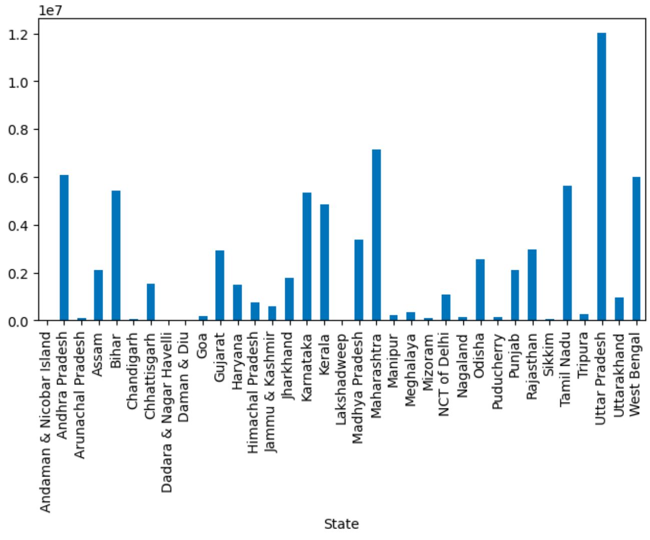
Perform detailed Exploratory analysis by creating certain questions like (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio? (Example Questions). Pick 5 variables out of the given 24 variables below for EDA: No_HH, TOT_M, TOT_F, 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

Variables taken into consideration are:
TOT_M,TOT_F,M_LIT,F_LIT,TOT_WORK_M

Which state has the highest male population?
Uttar Pradesh.



Which state has the highest female population?
Uttar Pradesh

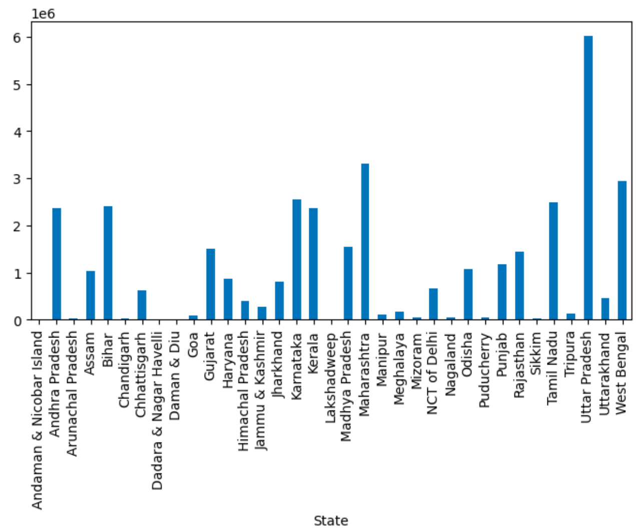


Which state has the highest and lowest gender ratio?

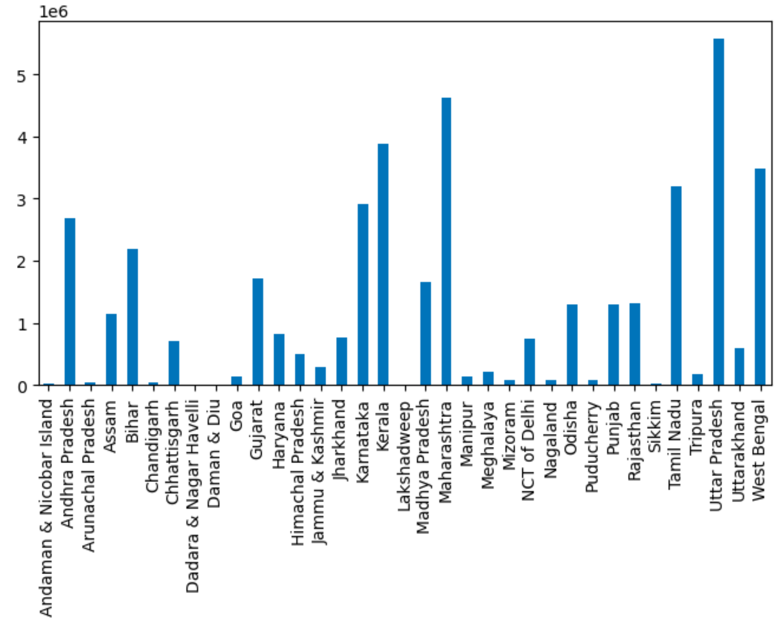
```
State
Andaman & Nicobar Island     0.652679
Andhra Pradesh               0.537024
Arunachal Pradesh            0.574365
Assam                        0.686561
Bihar                        0.744596
Chandigarh                   0.700037
Chhattisgarh                 0.549200
Dadara & Nagar Havelli       0.644631
Daman & Diu                  0.703143
Goa                          0.621648
Gujarat                      0.674844
Haryana                      0.779129
Himachal Pradesh             0.642741
Jammu & Kashmir              0.735154
Jharkhand                    0.681804
Karnataka                    0.637802
Kerala                       0.601238
Lakshadweep                  0.868061
Madhya Pradesh               0.639695
Maharashtra                  0.587812
Manipur                      0.641179
Meghalaya                    0.752160
Mizoram                      0.623634
NCT of Delhi                 0.775077
Nagaland                     0.583682
Odisha                       0.575500
Puducherry                   0.591111
Punjab                       0.744502
Rajasthan                    0.695286
Sikkim                       0.642227
Tamil Nadu                   0.547921
Tripura                      0.625881
Uttar Pradesh                0.752167
Uttarakhand                  0.630865
West Bengal                  0.650345
dtype: float64
```

Highest: Lakshwadeep
Lowest: Andra Pradesh

Which state has the highest number of literate men?
Uttar Pradesh



Which state has the highest number of literate females?
Uttar Pradesh



Which state has the highest literacy rate in total male population?

State	
Andaman & Nicobar Island	0.827085
Andhra Pradesh	0.724712
Arunachal Pradesh	0.671484
Assam	0.711972
Bihar	0.598354
Chandigarh	0.803583
Chhattisgarh	0.733391
Dadara & Nagar Havelli	0.733171
Daman & Diu	0.827188
Goa	0.835282
Gujarat	0.760907
Haryana	0.749246
Himachal Pradesh	0.802359
Jammu & Kashmir	0.672121
Jharkhand	0.665078
Karnataka	0.749135
Kerala	0.811806
Lakshadweep	0.826718
Madhya Pradesh	0.713084
Maharashtra	0.788496
Manipur	0.757676
Meghalaya	0.610784
Mizoram	0.814862
NCT of Delhi	0.791835
Nagaland	0.759543
Odisha	0.737274
Puducherry	0.827295
Punjab	0.742014
Rajasthan	0.703164
Sikkim	0.796205
Tamil Nadu	0.808522
Tripura	0.815733
Uttar Pradesh	0.665239
Uttarakhand	0.755365
West Bengal	0.749542

Highest: Goa

Lowest: Bihar

Which state has the highest literacy rate in total female population?

State	
Andaman & Nicobar Island	0.705343
Andhra Pradesh	0.439314
Arunachal Pradesh	0.514466
Assam	0.550760
Bihar	0.406581
Chandigarh	0.728288
Chhattisgarh	0.461043
Dadara & Nagar Havelli	0.490075
Daman & Diu	0.669304
Goa	0.730168
Gujarat	0.586118
Haryana	0.551532
Himachal Pradesh	0.654789
Jammu & Kashmir	0.502746
Jharkhand	0.435937
Karnataka	0.543499
Kerala	0.798583
Lakshadweep	0.767262
Madhya Pradesh	0.491609
Maharashtra	0.647051
Manipur	0.589823
Meghalaya	0.617477
Mizoram	0.831862
NCT of Delhi	0.690211
Nagaland	0.669607
Odisha	0.510190
Puducherry	0.657692
Punjab	0.611202
Rajasthan	0.442871
Sikkim	0.653018
Tamil Nadu	0.571286
Tripura	0.714791
Uttar Pradesh	0.463640
Uttarakhand	0.604570
West Bengal	0.578332

Highest: Mizoram
Lowest: Jharkhand and Andra Pradesh

Literacy ratio between men and women population:

```
State
Andaman & Nicobar Island    1.306624
Andhra Pradesh              1.128797
Arunachal Pradesh           1.333932
Assam                       1.126733
Bihar                       0.912576
Chandigarh                  1.294647
Chhattisgarh                1.144658
Dadara & Nagar Havelli      1.036921
Daman & Diu                 1.150735
Goa                         1.406194
Gujarat                     1.141432
Haryana                     0.944792
Himachal Pradesh            1.269688
Jammu & Kashmir             1.017474
Jharkhand                   0.961372
Karnataka                   1.137504
Kerala                      1.636144
Lakshadweep                 1.069144
Madhya Pradesh              1.077721
Maharashtra                 1.396048
Manipur                     1.214112
Meghalaya                   1.344074
Mizoram                     1.636956
NCT of Delhi                1.124611
Nagaland                    1.510397
Odisha                      1.202426
Puducherry                  1.344908
Punjab                      1.106386
Rajasthan                   0.905852
Sikkim                      1.277061
Tamil Nadu                  1.289566
Tripura                     1.400038
Uttar Pradesh               0.926592
Uttarakhand                 1.268682
West Bengal                 1.186419
dtype: float64
```

States with highest female literates: Mizoram and Kerala.
Lowest female literates: Rajasthan.

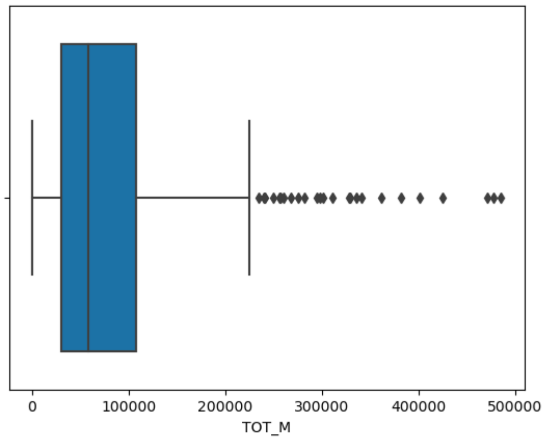
PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?

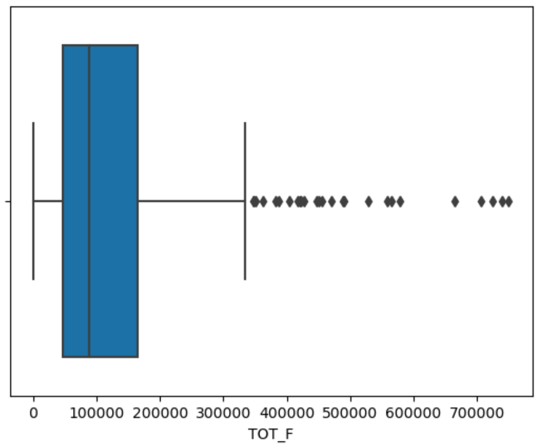
Census data is collected on a large scale and aims to capture a comprehensive snapshot of the population. Outliers in census data might be genuine representations of unique situations rather than measurement errors or anomalies.
Census data is collected meticulously, adhering to rigorous protocols and methods. This reduces the chances of significant measurement errors that often give rise to outliers in other types of data.
The presence of outliers can significantly affect statistical measures such as means and standard deviations. However, census data analysis often focuses on population-level characteristics, where individual extreme values might not have a substantial impact on overall trends.
Therefore, treating outliers for this case is not mandatory.

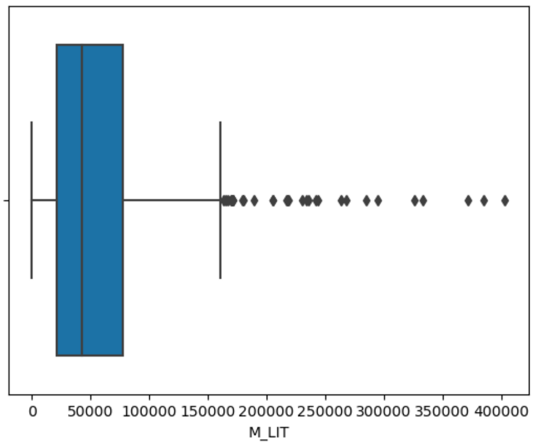
PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.

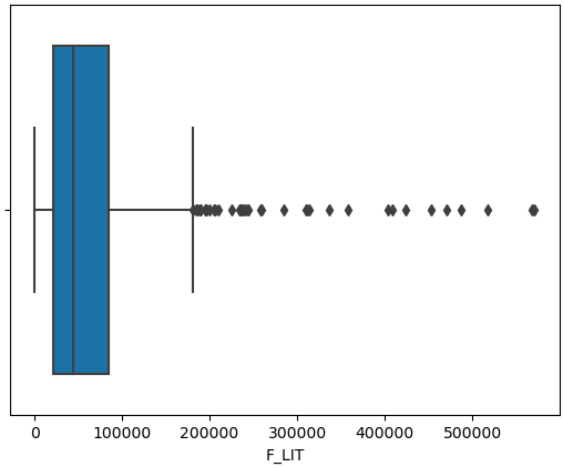
Z-score scaling has a significant impact on outliers by standardizing the data and making boxplots more consistent across variables. It enhances the ability to visualize patterns and relationships within the data while reducing the influence of extreme values. However, the actual outlier values are not removed from the data; they are only transformed in terms of their standardized values i.e., While scaling reduces the impact of outliers visually, it doesn't remove the actual outliers from the data.

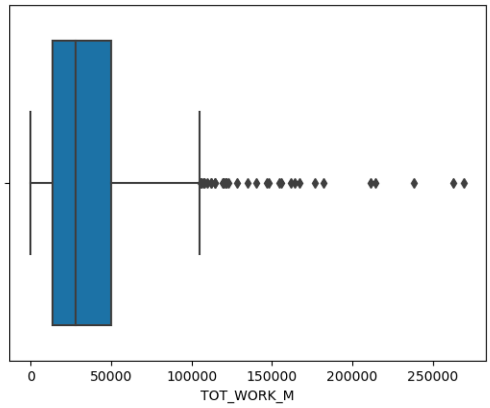
Let’s look at the boxplots before and after scaling.

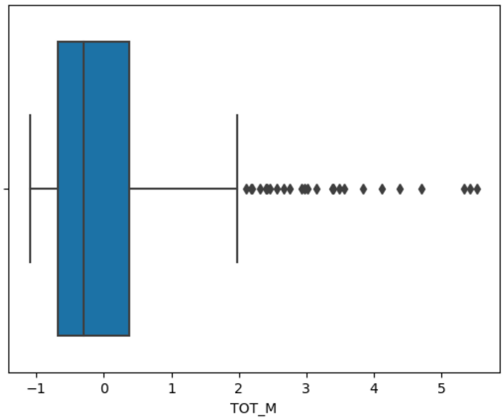


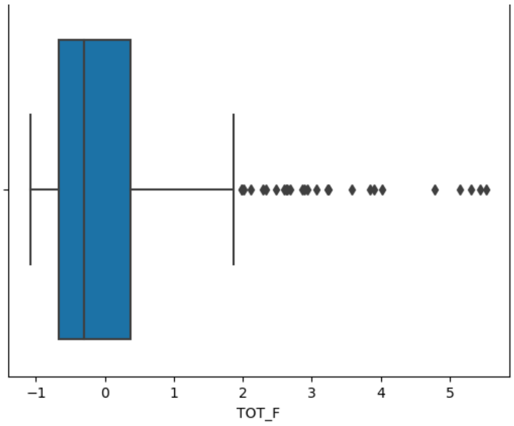


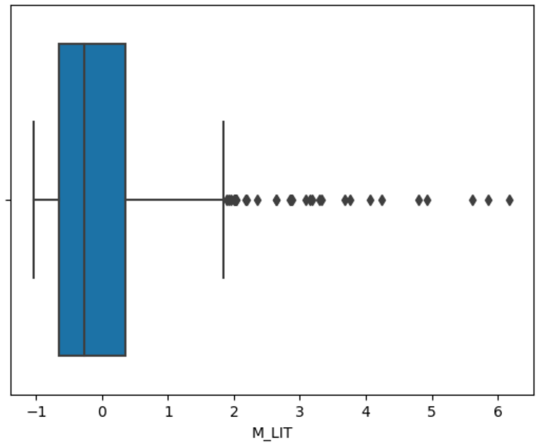


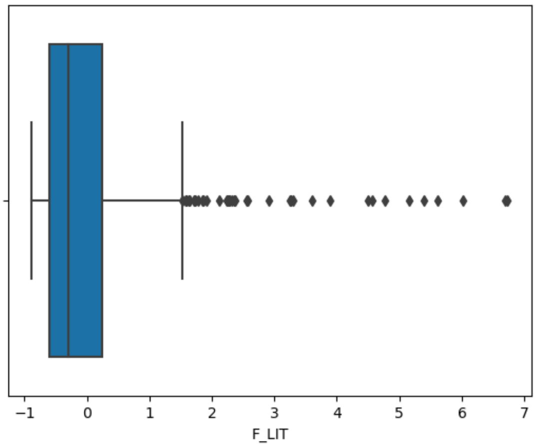


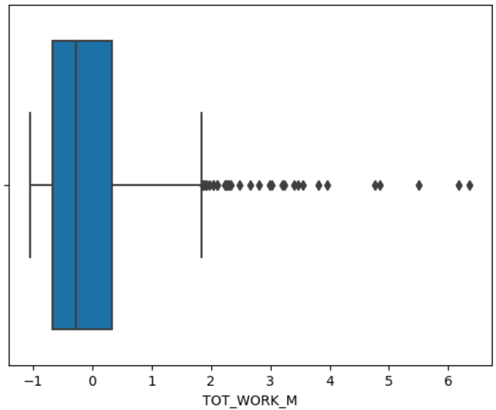












It is very evident that the boxplot visualization remains the same, only the scale is standardized/shrinked.

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

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.        ,  0.37643683,  0.15058437, ...,  0.03363703,
        -0.07959556, -0.02552519],
       [-0.        ,  0.2448199 ,  0.09383958, ..., -0.02638552,
        -0.01672564,  0.03567243],
       [-0.        , -0.09325898, -0.0110033 , ...,  0.01165739,
        -0.01279215, -0.00377366]])
```

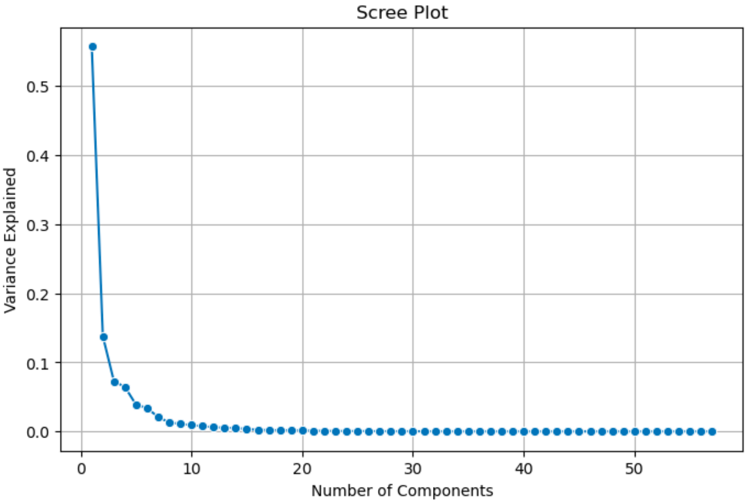
Eigen Values

```
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 = (eigen value of each PC)/(sum of eigen values of all PCs)

```
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, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33])
```

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



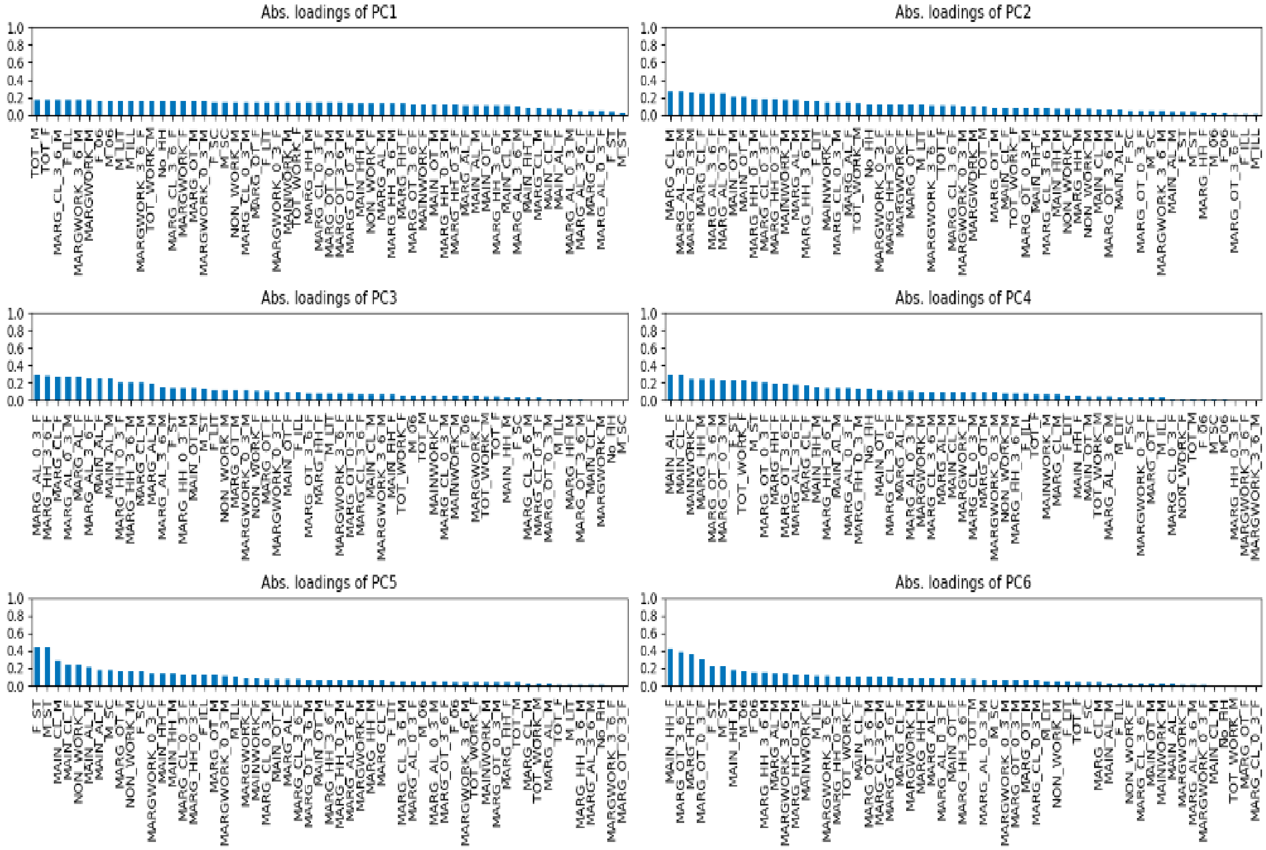
The above clearly depicts that the plot is stable from the 7th point(not much of variance).

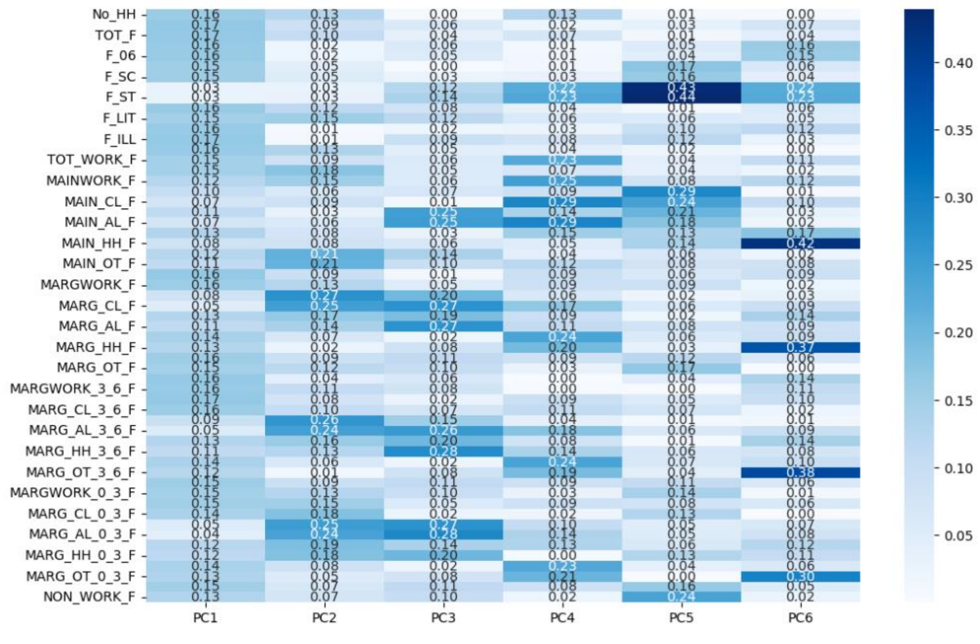
```
array([0.55726063, 0.69510499, 0.76785794, 0.83212212, 0.87077261,
       0.9047243 , 0.92532669, 0.93848433, 0.94929292, 0.95854687,
       0.96607599, 0.97226701, 0.97745473, 0.98238168, 0.98574761,
       0.98813454, 0.99012071, 0.99198278, 0.99368693, 0.99509011,
       0.99609921, 0.99687687, 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.        ])
```

The above array depicts the cumulative explained variance ratio.
90% explained variance is covered in the first 6 PC's.

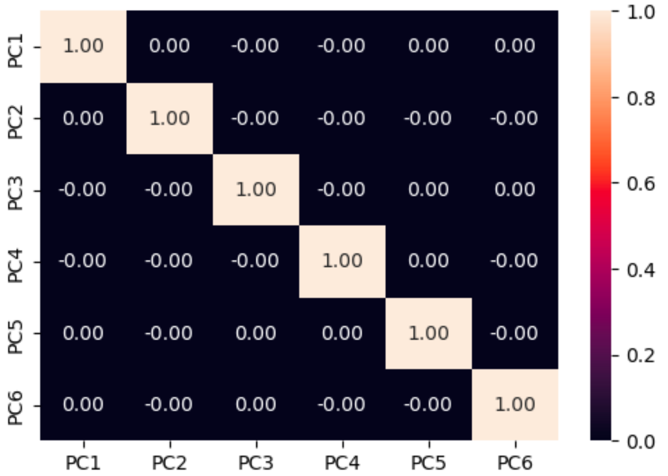
PCA: Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.

PC1 is very important as it captures the highest explained variance ratio.
Each and every PC has its own characteristics.


Abs. loadings of PC1
Abs. loadings of PC2
Abs. loadings of PC3
Abs. loadings of PC4
Abs. loadings of PC5
Abs. loadings of PC6


No_HH
TOT_F
F_06
F_SC
F_ST
F_LIT
F_ILL
TOT_WORK_F
MAINWORK_F
MAIN_CL_F
MAIN_AL_F
MAIN_HH_F
MAIN_OT_F
MARGWORK_F
MARG_CL_F
MARG_AL_F
MARG_HH_F
MARG_OT_F
MARGWORK_3_6_F
MARG_CL_3_6_F
MARG_AL_3_6_F
MARG_HH_3_6_F
MARG_OT_3_6_F
MARGWORK_0_3_F
MARG_CL_0_3_F
MARG_AL_0_3_F
MARG_HH_0_3_F
MARG_OT_0_3_F
NON_WORK_F
PC1
PC2
PC3
PC4
PC5
PC6

	PC1	PC2	PC3	PC4	PC5	PC6
0	-4.617263	0.138116	0.328545	1.543697	0.353736	-0.420948
1	-4.771662	-0.105865	0.244449	1.963215	-0.153884	0.417308
2	-5.964836	-0.294347	0.367394	0.619543	0.478199	0.276581
3	-6.280796	-0.500384	0.212701	1.074515	0.300799	0.051157
4	-4.478566	0.894154	1.078277	0.535557	0.804065	0.341678
5	-3.319963	2.823865	3.058460	-0.447904	0.742445	0.634676
6	-5.021393	-0.346359	0.650378	0.981072	-0.059778	-0.246957
7	-4.608709	0.022370	0.398755	1.576995	0.171316	-0.139444
8	-5.186703	-0.059097	0.184397	1.735440	0.169174	0.455039
9	-4.226190	-1.335080	0.697838	1.470509	0.269146	-0.002576



PCA: Write linear equation for first PC.

$PC1 = a1 \cdot X1 + a2 \cdot X2 + \ldots + an \cdot Xn$

```
( 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_W
ORK_M + ( 0.15 ) * TOT_WORK_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 ) * MARGWORK_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 ) * MARG
WORK_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 +
```