DATA MINING PROJECT REPORT

CODED

-by Apoorva P

Contents

Sl.no	Topic	Pg.no
1	Part 1: CLUSTERING	3
1.1	Define the Problem and Perform Exploratory Data Analysis	3-4
1.2	Data Pre-Processing	5
1.3	Hierarchical Clustering	6
1.4	K-means Clustering	6-8
1.5	Actionable Insights and Recommendations	8
2	Part 2: PCA	9
2.1	Define the Problem and Perform Exploratory Data Analysis	9-13
2.2	Data Pre-Processing	13
2.3	PCA	13-15

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.

The following three features are commonly used in digital marketing:

CPM = (Total Campaign Spend / Number of Impressions) * 1,000. Note that the Total Campaign Spend refers to the 'Spend' Column in the dataset and the Number of Impressions refers to the 'Impressions' Column in the dataset.

CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column in the dataset and the Number of Clicks refers to the 'Clicks' Column in the dataset.

CTR = Total Measured Clicks / Total Measured Ad Impressions x 100. Note that the Total Measured Clicks refers to the 'Clicks' Column in the dataset and the Total Measured Ad Impressions refers to the 'Impressions' Column in the dataset.

1.1 Define the Problem and Perform Exploratory Data Analysis

- Load the dataset and check the data
- The first 5 rows

	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	Format 1	300	250	75000	Inter 222	Video	Desktop	Display	1806	325	323	1	0.0	0.35	0.0	0.0031	0.0	0.0
1	2020-9-2-10	Format 1	300	250	75000	Inter 227	App	Mobile	Video	1780	285	285	1	0.0	0.35	0.0	0.0035	0.0	0.0
2	2020-9-1-22	Format 1	300	250	75000	Inter 222	Video	Desktop	Display	2727	356	355	1	0.0	0.35	0.0	0.0028	0.0	0.0
3	2020-9-3-20	Format 1	300	250	75000	Inter 228	Video	Mobile	Video	2430	497	495	1	0.0	0.35	0.0	0.0020	0.0	0.0
4	2020-9-4-15	Format 1	300	250	75000	Inter 217	Web	Desktop	Video	1218	242	242	1	0.0	0.35	0.0	0.0041	0.0	0.0

- The last 5 rows

	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	Format 5	720	300	216000	Inter 220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
23062	2020-11-2-7	Format 5	720	300	216000	Inter 224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.0260	NaN	NaN	NaN
23063	2020-9-14-22	Format 5	720	300	216000	Inter 218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	NaN	NaN	NaN
23064	2020-11-18-2	Format 4	120	600	72000	inter 230	Video	Mobile	Video	7	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
23065	2020-9-14-0	Format 5	720	300	216000	Inter 221	App	Mobile	Video	2	2	2	1	0.09	0.35	0.0585	NaN	NaN	NaN

- Checking the shape and information of the table

```
<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
```

- The table has 19 columns and 23066 rows
- Checking Table Description

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	3.851631e+02	2.336514e+02	120.0000	120.000000	300.0000	7.200000e+02	7.280000e+02
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.0000	250.000000	300.0000	6.000000e+02	6.000000e+02
Ad Size	23066.0	9.667447e+04	6.153833e+04	33600.0000	72000.000000	72000.0000	8.400000e+04	2.160000e+05
Available_Impressions	23066.0	2.432044e+06	4.742888e+06	1.0000	33672.250000	483771.0000	2.527712e+06	2.759286e+07
Matched_Queries	23066.0	1.295099e+06	2.512970e+06	1.0000	18282.500000	258087.5000	1.180700e+06	1.470202e+07
Impressions	23066.0	1.241520e+06	2.429400e+06	1.0000	7990.500000	225290.0000	1.112428e+06	1.419477e+07
Clicks	23066.0	1.067852e+04	1.735341e+04	1.0000	710.000000	4425.0000	1.279375e+04	1.430490e+05
Spend	23066.0	2.706626e+03	4.067927e+03	0.0000	85.180000	1425.1250	3.121400e+03	2.693187e+04
Fee	23066.0	3.351231e-01	3.196322e-02	0.2100	0.330000	0.3500	3.500000e-01	3.500000e-01
Revenue	23066.0	1.924252e+03	3.105238e+03	0.0000	55.365375	926.3350	2.091338e+03	2.127618e+04
CTR	23066.0	1.198738e-03	1.756137e-05	0.0001	0.001200	0.0012	1.200000e-03	1.200000e-03
CPM	23066.0	9.345465e-01	6.757466e-02	0.0000	0.940000	0.9400	9.400000e-01	9.400000e-01
CPC	23066.0	3.981835e-02	2.470190e-03	0.0000	0.040000	0.0400	4.000000e-02	4.000000e-02
cluster	23066.0	1.058354e+00	8.876994e-01	0.0000	1.000000	1.0000	1.000000e+00	4.000000e+00

- Checking Null values and duplicate rows
- The table has no duplicate rows

```
# check for null values
data.isnull().sum()
```

```
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
```

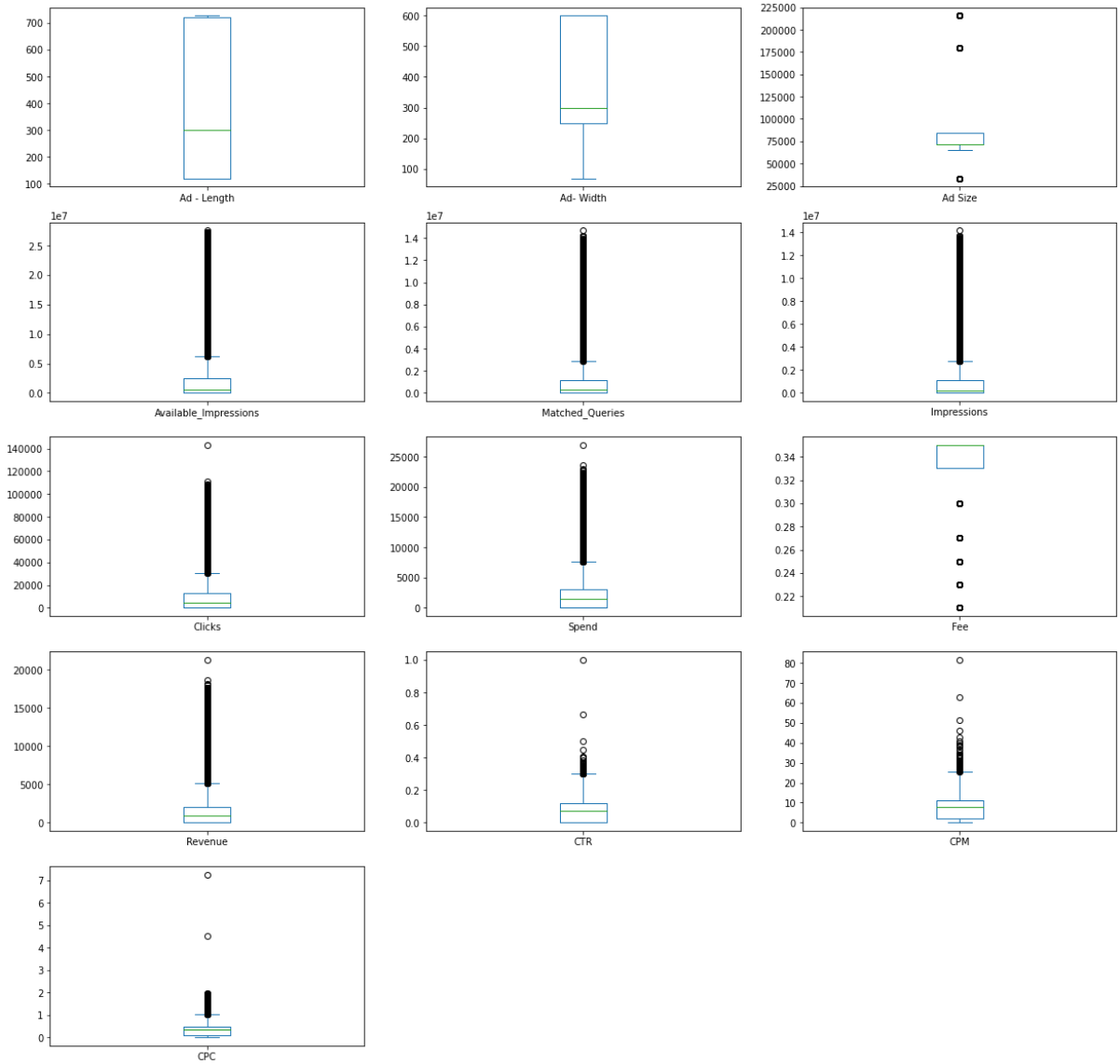
```
# check for duplicate values
data.duplicated().sum()
```

```
0
```

1.2 Data Pre-Processing

The missing values in CPC, CTR and CPM are treated by writing a user-defined function, and calling it.
CPM = (Total Campaign Spend / Number of Impressions) * 1,000
CPC = Total Cost (spend) / Number of Clicks
CTR = Total Measured Clicks / Total Measured Ad Impressions x 100
The missing values are treated using the above formulae and user defined function and calling it using return function.
The above data set has columns timestamp , inventory type, etc which are not very useful for clustering.
Also removing CPR,CPM,CPC as they are dependent

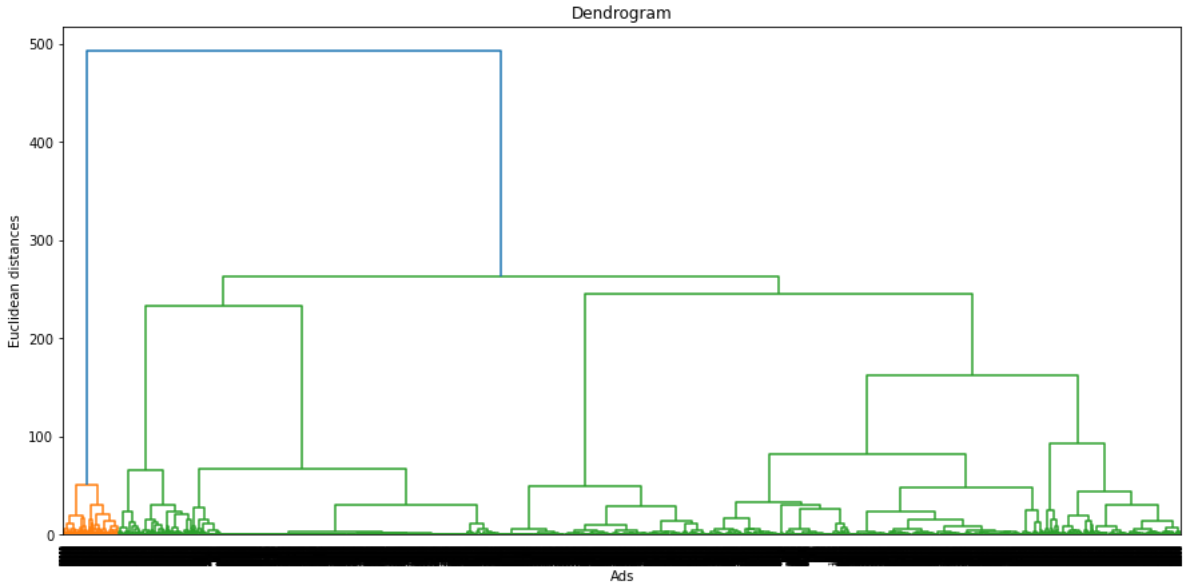
- Checking for Outliners



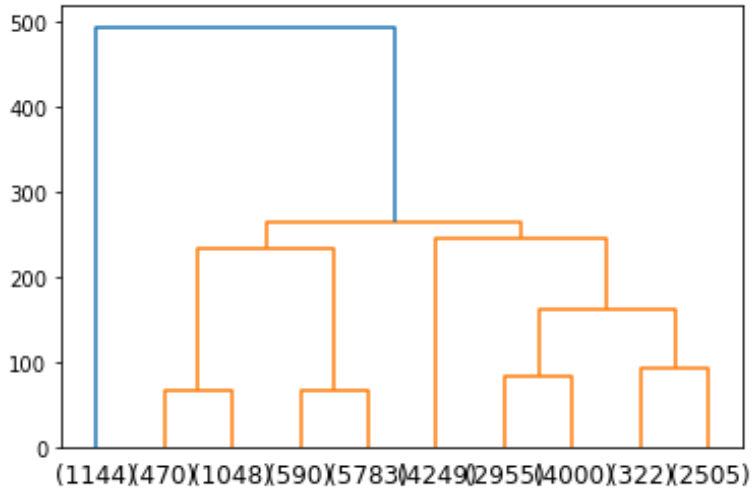
- Treating the Outliners by using an user defined function to replace the values with the mean
- Performing Z-Score Scaling on the data and storing the data in data_scaled dataframe

1.3 Hierarchical Clustering

Below Dendrogram performed for Hierarchical using WARD and Euclidean Distance on the Scaled Data i.e, data_scaled



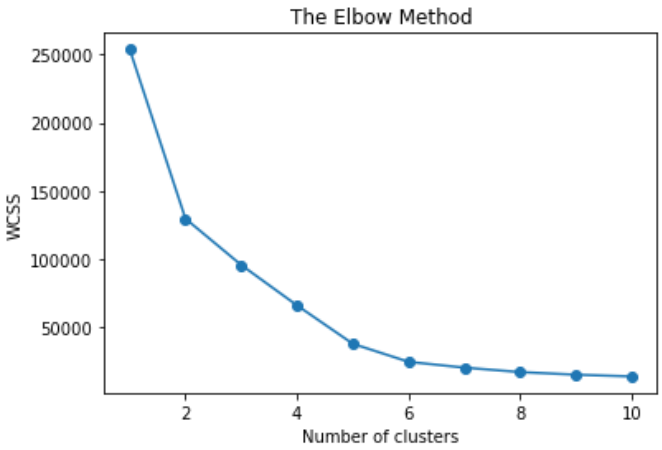
In this Dendrogram, value of P = 10, which means that only the last 10 merged clusters are shown



- According to my inference 5 clusters are the optimum no of clusters to be formed

1.4 K-means Clustering

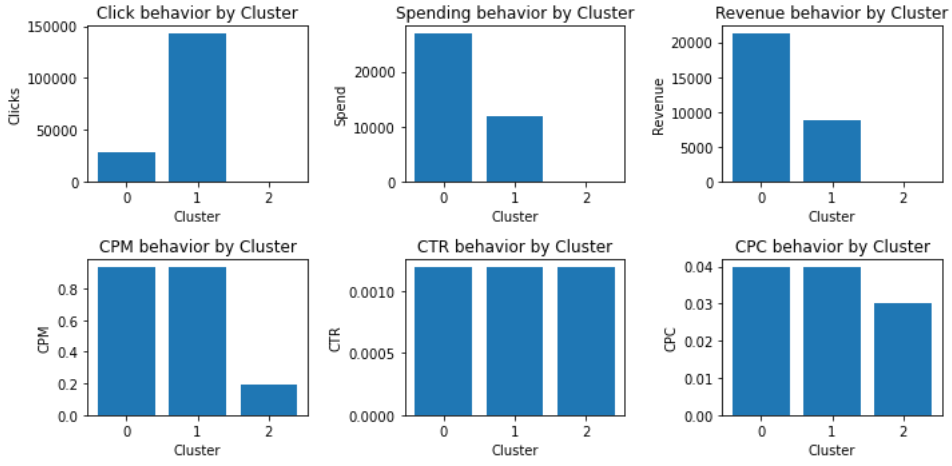
- First applying the K-mean clustering to the scaled data followed by plotting Elbow curve



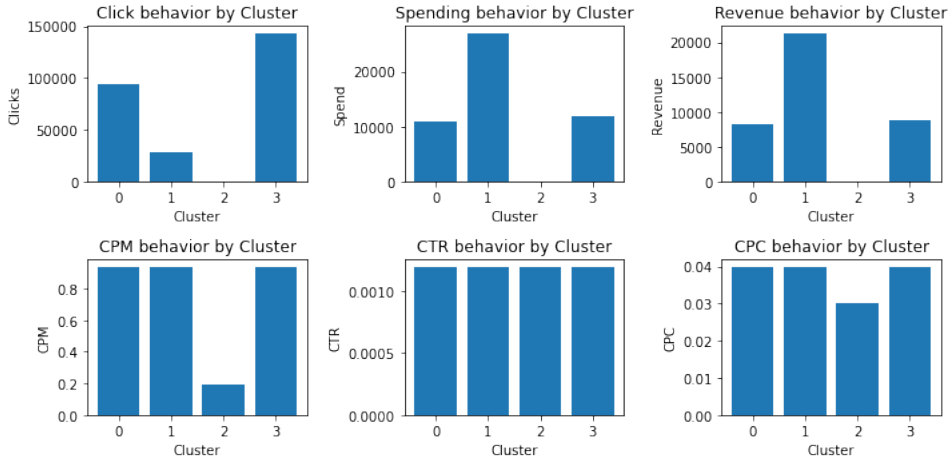
- From K=1 to K=2 there is a significant drop in the value which can also be seen from K=2 to K=3,K=3 to K=4 and K=4 to K=5.
- After K=5 the value drop is even more Gradual
- To conclude from this curve the optimal no of clusters is 5
- Next performing Silhouette scoring
- The score is 0.601
- Now performing silhouette scoring for n clusters, the result as shown below

```
For n_clusters=2, The Silhouette Coefficient is 0.7072208632531246
For n_clusters=3, The Silhouette Coefficient is 0.4110142936110454
For n_clusters=4, The Silhouette Coefficient is 0.5322626394075498
For n_clusters=5, The Silhouette Coefficient is 0.6016476899996935
For n_clusters=6, The Silhouette Coefficient is 0.6398649416102179
For n_clusters=7, The Silhouette Coefficient is 0.641752423019188
For n_clusters=8, The Silhouette Coefficient is 0.642095572616584
For n_clusters=9, The Silhouette Coefficient is 0.6613911359199435
For n_clusters=10, The Silhouette Coefficient is 0.6595910943681638
```

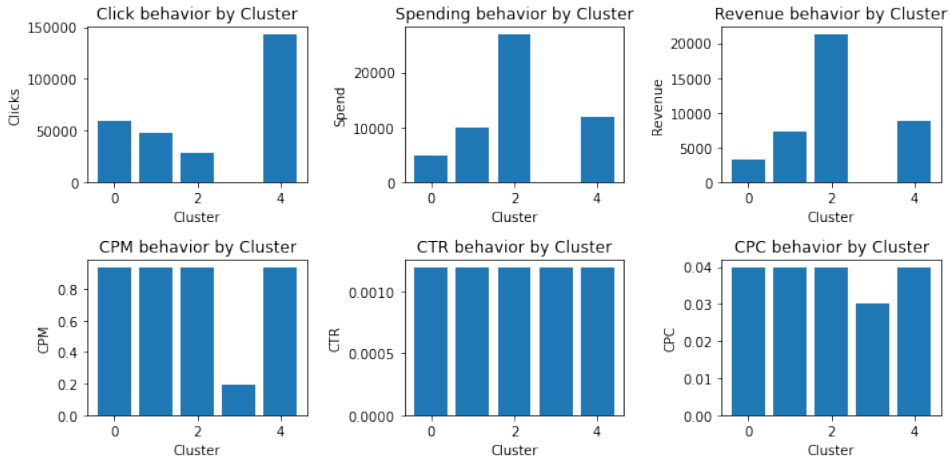
- The optimal number of clusters are 5
- Performing cluster profiling for no of cluster = 3



- Performing cluster profiling for no of cluster = 4



- Performing cluster profiling for no of cluster = 5



1.5 Actionable Insights and Recommendations

- There are 23066 rows, and 19 columns into the Dataset.
- There are no duplicate values in data frame.
- There are 4636 Null values in CTR, CPM, and CPC Columns.
- I have treated missing values in CPC, CTR, and CPM columns using the given formula
- It seems that there are Outliers into the Dataset
- We treated outliers using IQR method
- I have applied z-score method on the data frame for scaling.
- I have plotted Dendrogram for value of P = 10
- Plotted elbow plot and got optimum value is 5
- As per Elbow plot/scree-plot, we concluded that the optimal number of clusters should be 5.
- I have created 5 clusters for the Dataset.
- Based on the above analysis, it seems that k = 5 is the optimum number of clusters for this dataset.
- The code above performs basic data analysis, imputes missing values, treats outliers, and then scales the data using z-score scaling. It then runs the K-Means algorithm with k values ranging from 2-10. For each k value, it calculates the silhouette score to determine which k value results in the best clustering.
- Based on the above analysis, it seems that k = 5 is the optimum number of clusters for this dataset. This can be seen from the elbow plot and the silhouette scores. The code then profiles the ads based on the optimum number of clusters.

PART 2: PCA

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 Define the Problem and Perform Exploratory Data Analysis

- Load the data and check the first and last 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	MARG_AL_0_3
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180	2
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123	2
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	1
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	...	874	1928	465	10

5 rows × 61 columns

	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_(
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	

5 rows × 61 columns

- The table has 640 rows and 61 columns.
- Checking the data info

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 640 entries, 0 to 639
Data columns (total 61 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   State Code      640 non-null    int64
 1   Dist.Code       640 non-null    int64
 2   State           640 non-null    object
 3   Area Name       640 non-null    object
 4   No_HH           640 non-null    int64
 5   TOT_M           640 non-null    int64
 6   TOT_F           640 non-null    int64
 7   M_06            640 non-null    int64
 8   F_06            640 non-null    int64
 9   M_SC            640 non-null    int64
 10  F_SC            640 non-null    int64
 11  M_ST            640 non-null    int64
 12  F_ST            640 non-null    int64
 13  M_LIT           640 non-null    int64
 14  F_LIT           640 non-null    int64
 15  M_ILL           640 non-null    int64
 16  F_ILL           640 non-null    int64
 17  TOT_WORK_M      640 non-null    int64
 18  TOT_WORK_F      640 non-null    int64
 19  MAINWORK_M      640 non-null    int64
 20  MAINWORK_F      640 non-null    int64
 21  MAIN_CL_M       640 non-null    int64
 22  MAIN_CL_F       640 non-null    int64
 23  MAIN_AL_M       640 non-null    int64
 24  MAIN_AL_F       640 non-null    int64
 25  MAIN_HH_M       640 non-null    int64
 26  MAIN_HH_F       640 non-null    int64
 27  MAIN_OT_M       640 non-null    int64
 28  MAIN_OT_F       640 non-null    int64
 29  MARGWORK_M      640 non-null    int64
 30  MARGWORK_F      640 non-null    int64
 31  MARG_CL_M       640 non-null    int64
 32  MARG_CL_F       640 non-null    int64
 33  MARG_AL_M       640 non-null    int64
 34  MARG_AL_F       640 non-null    int64
 35  MARG_HH_M       640 non-null    int64
 36  MARG_HH_F       640 non-null    int64
 37  MARG_OT_M       640 non-null    int64
 38  MARG_OT_F       640 non-null    int64
 39  MARGWORK_3_6_M  640 non-null    int64
 40  MARGWORK_3_6_F  640 non-null    int64
 41  MARG_CL_3_6_M   640 non-null    int64
 42  MARG_CL_3_6_F   640 non-null    int64
 43  MARG_AL_3_6_M   640 non-null    int64
 44  MARG_AL_3_6_F   640 non-null    int64
 45  MARG_HH_3_6_M   640 non-null    int64
 46  MARG_HH_3_6_F   640 non-null    int64
 47  MARG_OT_3_6_M   640 non-null    int64
 48  MARG_OT_3_6_F   640 non-null    int64
 49  MARGWORK_0_3_M  640 non-null    int64
 50  MARGWORK_0_3_F  640 non-null    int64
 51  MARG_CL_0_3_M   640 non-null    int64
 52  MARG_CL_0_3_F   640 non-null    int64
 53  MARG_AL_0_3_M   640 non-null    int64
 54  MARG_AL_0_3_F   640 non-null    int64
 55  MARG_HH_0_3_M   640 non-null    int64
 56  MARG_HH_0_3_F   640 non-null    int64
 57  MARG_OT_0_3_M   640 non-null    int64
 58  MARG_OT_0_3_F   640 non-null    int64
 59  NON_WORK_M      640 non-null    int64
 60  NON_WORK_F      640 non-null    int64
dtypes: int64(59), object(2)
memory usage: 305.1+ KB
```

- Checking the table description and also checking for null values or duplicate values

	count	mean	std	min	25%	50%	75%	max
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

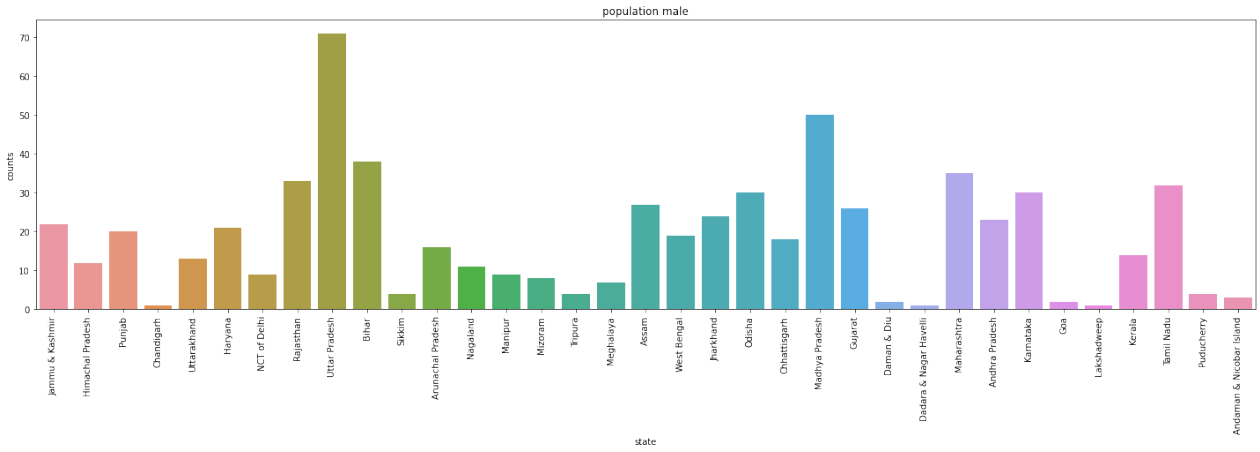
```
# check for null values
df.isnull().sum()
```

```
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
```

```
# check for duplicate values
df.duplicated().sum()
```

```
0
```

- The state with highest population is Uttar Pradesh and lowest is Lakshadweep,Chandigarh and Dadara & Nagar Havelli

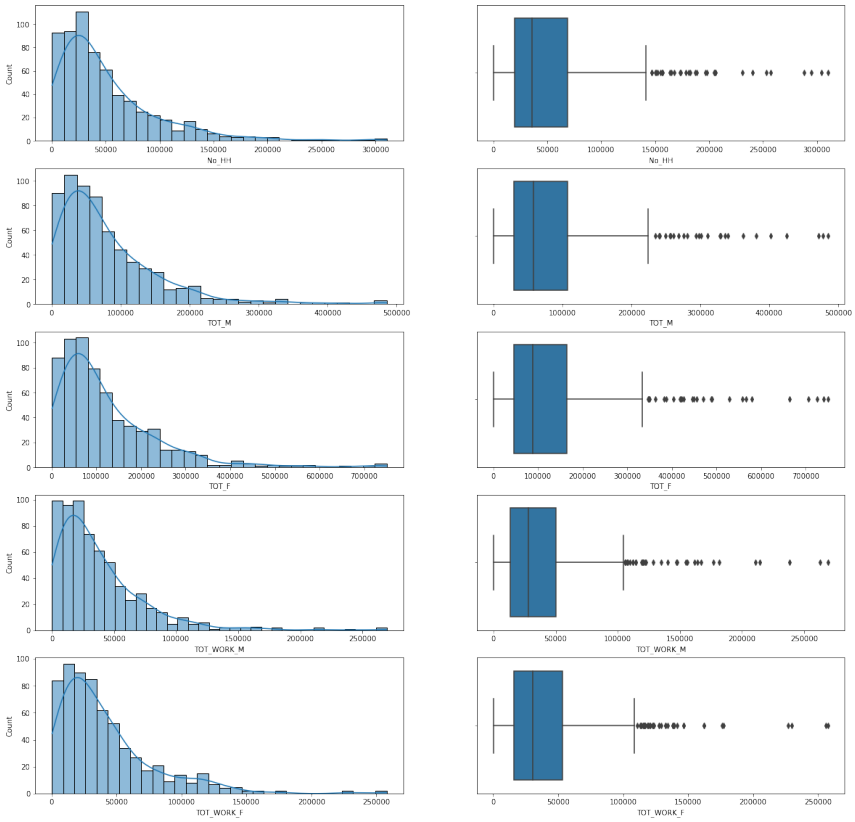


- The state with highest gender ratio is Lakshadweep with 86%
- The lowest gender ratio is a state is Andra Pradesh with 53.7%
- District Wise the highest gender ratio is Lakshadweep and lowest is Krishna in andra pradesh

- I have picked 5 Variables such as 'TOT_M', 'TOT_F', 'No_HH','TOT_WORK_M', and , 'TOT_WORK_F'. And comparing those 5 variable against 'State' and 'Dist.Code'.

No_HH	No of Household
TOT_M	Total Population of Male
TOT_F	Total Population of Female
TOT_WORK_M	Total Worker Population Male
TOT_WORK_F	Total Worker Population Female

- Doing a basic EDA on the above



- With respect to state

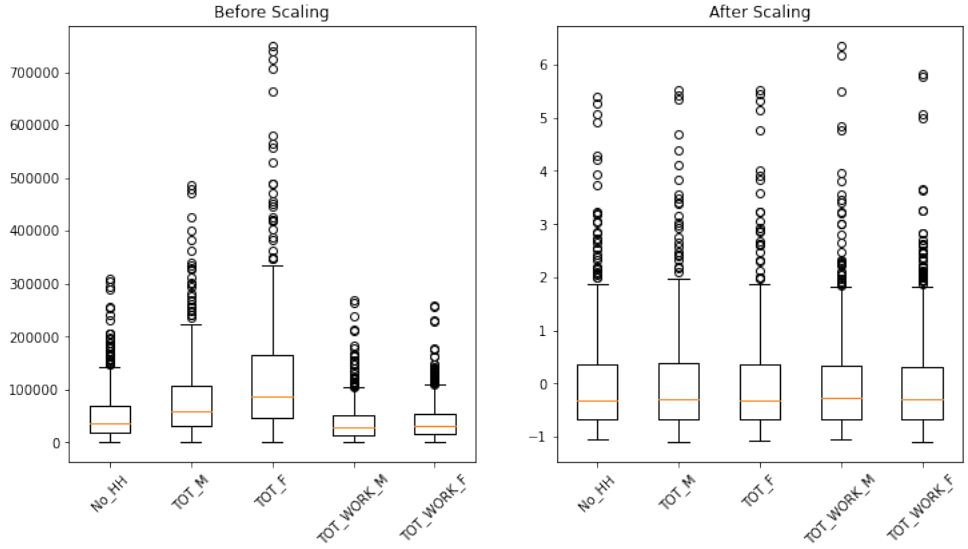
Variable	Highest	Lowest
No_HH	Uttar Pradesh	Dadara and Nagar Havelli
TOT_M	Uttar Pradesh	Dadara and Nagar Havelli
TOT_F	Uttar Pradesh	Dadara and Nagar Havelli
TOT_WORK_M	Uttar Pradesh	Dadara and Nagar Havelli
TOT_WORK_F	Uttar Pradesh	Lakshadweep

- With respect to Area

Variable	Highest	Lowest
No_HH	North Twenty Four Parganas	Dibang Valley
TOT_M	Mumbai Suburban	Dibang Valley
TOT_F	Mumbai Suburban	Dibang Valley
TOT_WORK_M	North Twenty Four Parganas	Dibang Valley
TOT_WORK_F	Bangalore	Dibang Valley

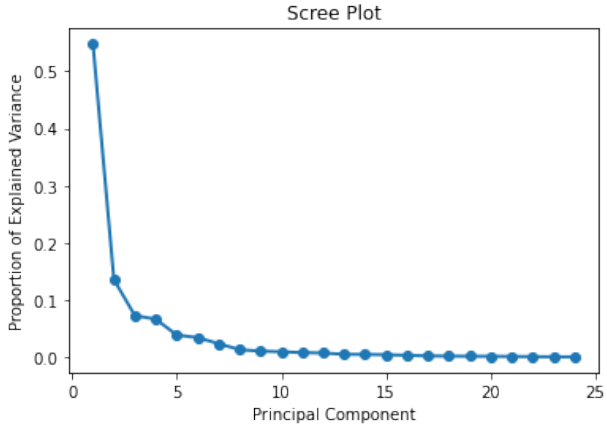
- PCA: Treating outliers: It is important to treat outliers as they can significantly affect the results of PCA. However, for this case, we have chosen not to treat outliers.
- PCA: Scaling the data: We can scale the data using the z-score method

- Checking the before and after scaling for the selected 5 attributes



PCA: Identify the optimum number of PCs

- To identify the optimum number of principal components (PCs), we can use the scree plot. The scree plot shows the eigenvalues of each PC in descending order, with the corresponding proportion of explained variance. We can plot the eigenvalues against the PCs and observe where the eigenvalues start to level off.



- PCA: Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.
- To compare the principal components with the actual columns, we can look at the loadings of each PC. The loadings represent the correlation between each variable and the PC. We can use the following code to extract the loadings and create a table that shows the contribution of each variable to each PC:

- From the loading matrix, we can see that the first principal components and compare them as

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	...	PC15	PC16	PC
TOT_M	0.942682	-0.249959	0.113013	0.052830	-0.079234	-0.103313	0.077847	0.073816	0.039333	0.041905	...	0.011650	-0.008759	-0.015
TOT_F	0.933661	-0.295559	0.039137	0.148706	-0.036109	-0.058726	-0.011894	0.063525	0.034169	0.053795	...	-0.007074	-0.004163	-0.005
MARG_CL_3_6_M	0.933658	0.218639	0.005755	-0.185935	0.063660	-0.140578	0.001590	-0.006843	-0.033012	0.022652	...	0.044961	0.024108	-0.026
F_ILL	0.932000	-0.027281	-0.212356	0.067489	-0.202466	-0.020674	0.029301	0.030235	0.003532	0.009714	...	0.014691	-0.054475	0.065
MARGWORK_3_6_M	0.930653	-0.120901	0.141509	0.017466	-0.107875	-0.183273	0.085408	0.076594	0.046398	0.138621	...	-0.020483	-0.091764	0.004
MARGWORK_M	0.928663	0.262985	0.034337	-0.174554	0.076410	-0.132222	-0.005454	-0.001592	-0.027491	0.008597	...	0.021400	0.032510	-0.026
F_06	0.917135	-0.053789	0.124568	-0.021850	-0.115721	-0.212317	0.124278	0.095558	0.019323	0.188166	...	-0.000676	-0.090456	0.028
M_06	0.915043	-0.058713	0.139059	-0.013464	-0.126648	-0.215815	0.121986	0.098551	0.017537	0.173382	...	0.000589	-0.097118	0.021
M_LIT	0.913704	-0.322756	0.148955	0.102884	-0.034556	-0.079617	0.036237	0.067911	0.055754	0.027469	...	0.001497	0.012907	-0.057
M_ILL	0.912598	-0.015016	-0.001751	-0.094595	-0.195835	-0.157884	0.185959	0.081715	-0.011644	0.077649	...	0.038900	-0.068822	0.106
MARGWORK_3_6_F	0.909559	-0.294506	0.164685	0.031497	-0.024077	-0.143743	0.013707	0.090878	0.017783	0.082139	...	0.031371	-0.012760	-0.014
TOT_WORK_M	0.901730	-0.374043	0.076008	0.087725	-0.044003	-0.011686	0.065294	0.066622	0.029511	-0.064177	...	0.045434	0.080730	-0.035
No_HH	0.879695	-0.358693	-0.080912	0.235015	-0.012265	0.012136	-0.110078	0.061988	0.004952	-0.002112	...	-0.010595	0.025369	0.014
MARG_CL_3_6_F	0.877684	0.284607	-0.214305	0.171123	0.106373	0.012479	-0.103127	-0.062111	-0.030895	0.012312	...	0.058992	-0.043533	-0.021
MARGWORK_F	0.876292	0.347036	-0.162263	0.153222	0.135248	0.005946	-0.127029	-0.038492	-0.035653	-0.001503	...	0.029239	-0.028371	-0.005
MARG_OT_M	0.875839	-0.248111	0.263579	-0.087478	0.176454	-0.102992	-0.024562	-0.112253	-0.098084	0.009853	...	0.020425	-0.007344	-0.104
MARGWORK_0_3_M	0.870377	-0.258464	0.261425	-0.088463	0.160637	-0.101894	-0.022551	-0.114508	-0.088178	0.025008	...	0.034643	-0.024107	-0.105
F_SC	0.854837	-0.146326	-0.058922	0.030220	-0.237184	0.001321	-0.114058	-0.096645	-0.028105	-0.302575	...	0.059970	-0.133765	0.011
M_SC	0.853835	-0.124614	0.027821	-0.042449	-0.268042	-0.036254	-0.031747	-0.092091	-0.028702	-0.306076	...	0.074464	-0.147880	0.013
NON_WORK_M	0.848246	-0.181106	0.257808	-0.077105	0.244096	-0.102014	-0.033027	-0.094021	-0.141220	-0.066141	...	-0.051567	0.076492	-0.093
MARG_CL_0_3_M	0.846939	0.424755	0.147560	-0.116994	0.122869	-0.089690	-0.033559	0.019731	-0.003359	-0.048766	...	-0.075230	0.064309	-0.022
MARG_OT_F	0.830649	-0.330898	0.196129	0.030591	0.264369	-0.025028	-0.111305	-0.157133	-0.146244	0.017713	...	-0.000725	-0.023772	0.077
F_LIT	0.828272	-0.430321	0.192590	0.182751	0.072463	-0.075925	-0.036405	0.077186	0.049510	0.075341	...	-0.019934	0.027903	-0.049
MARGWORK_0_3_F	0.825024	-0.352171	0.190964	0.022969	0.225348	-0.011755	-0.084427	-0.161211	-0.097658	0.020701	...	0.020777	-0.053572	0.050
MAINWORK_M	0.824578	-0.494629	0.079848	0.142578	-0.068894	0.017606	0.076821	0.077448	0.040612	-0.076269	...	0.047524	0.085741	-0.034
TOT_WORK_F	0.822696	-0.245762	-0.247845	0.383944	-0.056582	0.141293	-0.066907	-0.008704	0.064694	-0.018929	...	-0.091591	0.015750	0.015
MARG_HH_M	0.794731	0.195465	0.086266	-0.440856	-0.029786	0.161690	0.050854	-0.220308	0.187949	0.014987	...	-0.018732	0.025851	0.002
MARG_CL_0_3_F	0.790441	0.504262	0.010865	0.084597	0.210319	-0.014444	-0.187770	0.036808	-0.046781	-0.043309	...	-0.063823	0.020305	0.043
MARG_OT_0_3_M	0.789161	0.242203	0.077318	-0.424808	-0.000804	0.119386	0.022050	-0.192136	0.177831	0.017218	...	-0.097743	0.001266	-0.020
MARG_OT_3_6_M	0.787624	0.179247	0.088004	-0.440810	-0.038175	0.172627	0.058956	-0.226347	0.188916	0.014150	...	0.005240	0.032960	0.009
MARG_OT_0_3_F	0.745660	0.143644	-0.060245	-0.394318	0.111865	0.428051	-0.029357	0.051201	-0.060124	-0.012462	...	-0.096791	-0.053211	-0.023
MAIN_HH_M	0.742297	-0.209379	0.140127	-0.265953	-0.134564	0.281772	0.103002	-0.150425	0.281247	-0.088294	...	0.114751	0.092537	0.093
NON_WORK_F	0.739096	-0.208456	0.188541	0.054330	0.371352	-0.070176	-0.194548	-0.121012	-0.304139	0.004399	...	-0.079240	0.088389	0.168
MARG_AL_M	0.725487	0.465720	-0.314798	-0.287761	-0.006813	-0.197473	-0.004817	0.111103	0.010398	0.012357	...	0.019949	0.057871	0.049
MARG_HH_3_6_M	0.725352	0.445759	-0.339375	-0.275402	-0.026807	-0.200440	-0.000093	0.113243	-0.004701	0.013465	...	0.042937	0.057097	0.053
MARG_HH_F	0.720146	0.069026	-0.070856	-0.381213	0.073309	0.524308	0.014703	0.083570	-0.111140	0.021738	...	-0.023502	-0.073057	-0.028
MARG_OT_3_6_F	0.702526	0.042499	-0.073638	-0.372023	0.059141	0.550999	0.029686	0.093684	-0.127332	0.033242	...	0.002009	-0.078990	-0.029
MAINWORK_F	0.698770	-0.431913	-0.247802	0.419854	-0.119728	0.172998	-0.036387	0.003319	0.093279	-0.022918	...	-0.124274	0.029928	0.021
MAIN_OT_M	0.696667	-0.596816	0.248270	0.153751	0.095049	0.007969	0.017305	0.124749	0.125655	-0.058114	...	0.024128	0.058718	-0.049
MARG_HH_0_3_M	0.687395	0.521008	-0.199429	-0.322024	0.073772	-0.175052	-0.023514	0.096603	0.070423	0.007254	...	-0.073347	0.057893	0.028
MARG_HH_0_3_F	0.654219	0.502832	-0.406136	-0.110825	0.167453	-0.166802	-0.153085	0.073358	0.091278	-0.004350	...	-0.045352	-0.009697	-0.016
MARG_AL_F	0.644434	0.388690	-0.590782	0.045309	0.080833	-0.142300	-0.128934	0.008875	0.047299	-0.003267	...	0.036382	-0.008165	-0.051
MAIN_AL_M	0.639179	-0.091388	-0.540639	0.075114	-0.344913	-0.010566	0.008958	0.072030	-0.161281	-0.140293	...	0.009399	0.054915	0.022
MAIN_OT_F	0.625945	-0.591890	0.114970	0.284475	0.126619	0.081419	-0.014784	0.073733	0.193638	-0.079085	...	-0.063016	0.029080	0.075
MARG_HH_3_6_F	0.623754	0.344056	-0.629354	0.090451	0.052866	-0.131096	-0.118202	-0.010529	0.032927	-0.002855	...	0.059664	-0.007485	-0.060
MAIN_CL_M	0.581926	0.177091	-0.142994	0.071701	-0.478415	0.011366	0.375222	-0.228416	-0.279502	0.026298	...	0.139318	0.147447	-0.015
MARG_AL_3_6_M	0.524838	0.743094	0.273266	0.165755	-0.009844	0.016043	0.063154	0.045589	0.017317	0.015612	...	0.035556	0.035139	-0.009
MAIN_HH_F	0.470172	-0.232815	-0.103558	-0.107407	-0.085858	0.619687	-0.024431	0.362173	-0.138014	0.119768	...	0.075366	0.072227	0.007
MARG_CL_M	0.464888	0.757027	0.345867	0.239651	-0.023707	0.048657	0.032725	0.057302	0.018313	-0.019452	...	0.013798	0.027931	-0.003
MAIN_CL_F	0.420158	0.236660	-0.172222	0.510388	-0.373114	0.170321	0.182975	-0.311135	-0.039961	0.299281	...	-0.150121	-0.011780	-0.012
MAIN_AL_F	0.416156	-0.174709	-0.650505	0.374740	-0.290227	0.059663	-0.132769	-0.061690	-0.017404	-0.090323	...	-0.153330	0.013928	-0.057
MARG_AL_0_3_M	0.296487	0.705659	0.454719	0.362201	-0.048920	0.108722	-0.031487	0.074705	0.018387	-0.087464	...	-0.031113	0.010610	0.009
MAIN_AL_F	0.416156	-0.174709	-0.650505	0.374740	-0.290227	0.059663	-0.132769	-0.061690	-0.017404	-0.090323	...	-0.153330	0.013928	-0.057
MARG_AL_0_3_M	0.296487	0.705659	0.454719	0.362201	-0.048920	0.108722	-0.031487	0.074705	0.018387	-0.087464	...	-0.031113	0.010610	0.009
MARG_AL_3_6_F	0.290721	0.684091	0.390205	0.490583	-0.073908	0.139534	0.004689	-0.007041	0.000924	0.008280	...	0.044146	-0.031765	-0.000
MARG_CL_F	0.277528	0.690732	0.419509	0.479452	-0.066238	0.141334	-0.023305	0.017690	0.006337	-0.026189	...	0.029801	-0.019995	-0.000
MARG_AL_0_3_F	0.235752	0.674939	0.466365	0.433328	-0.046169	0.139090	-0.084801	0.072148	0.018147	-0.102025	...	-0.003564	0.007184	0.000
F_ST	0.158581	0.076989	-0.413163	0.354246	0.605545	0.087975	0.528371	-0.014776	0.037930	-0.066527	...	0.031639	-0.036490	0.007
M_ST	0.153317	0.071110	-0.371110	0.342409	0.588985	0.076489	0.585951	-0.008593	0.015850	-0.082385	...	0.005110	-0.027354	0.010
gender_ratio	0.039821	0.113584	0.388384	-0.443099	-0.281056	-0.250388	0.478539	0.161333	-0.134384	-0.234927	...	-0.235137	-0.001928	-0.012

58 rows × 24 columns

PC1 =

+0.259 X No_HH
-0.215 X TOT_M
-0.215 X TOT_F
+0.253 X M_06
+ 0.254 X F_06
-0.246 X M_SC
-0.245 X F_SC
-0.251 X M_ST
-0.252 X F_ST
-0.195 X M_LIT
-0.195 X F_LIT
+0.195 X M_ILL
+0.196 X F_ILL
-0.202 X TOT_WORK_M
-0.202 X TOT_WORK_F
-0.198 X MAINWORK_M
-0.198 X MAINWORK_F
-0.196 X MAIN_CL_M
-0.196 X MAIN_CL_F
-0.195 X MAIN_AL_M
-0.195 X MAIN_AL_F
-0.195 X MAIN_HH_M
-0.195 X MAIN_HH_F
-0.195 X MAIN_OT_M
-0.195 X MAIN_OT_F

- This equation represents the combination of the original variables that make up the first principal component, where the coefficients represent the weights assigned to each variable in the linear combination. The sign of the coefficient indicates the direction of the relationship between the variable and the PC1 score. In this case, positive coefficients indicate a positive relationship with PC1, while negative coefficients indicate a negative relationship.
- We can interpret the first principal component as representing a combination of variables that are related to overall population and household size, with a negative influence from the number of male and female population and a positive influence from the number of households, population in the age group 0-6, and illiterate population. This component also shows a negative influence from the number of Scheduled Castes and Scheduled Tribes population and a negative influence from the number of workers, with a particularly strong negative influence from the number of cultivators.