7/30/2023

PCA – Report

PGP-DSBA

Karthick Raj S

Table of Contents

List of Tables

List of Figures

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.

PCA

Part 2- PCA: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.

Shape:

The Shape of the dataset is (640,61).

There are **640** Rows and **61** columns in the dataset.

First Five (Head):

The First Five rows of the dataset. (The rows and columns has been transposed for easier view). *Refer Clustering jupyter workings for the output*.

	0	1	2	3	4
State Code	1	1	1	1	1
Dist.Code	1	2	3	4	5
State	Jammu & Kashmir	Jammu & Kashmir	Jammu & Kashmir	Jammu & Kashmir	Jammu & Kashmir
Area Name	Kupwara	Badgam	Leh(Ladakh)	Kargil	Punch
No_HH	7707	6218	4452	1320	11654
TOT_M	23388	19585	6546	2784	20591
TOT_F	29796	23102	10964	4206	29981
M_06	5862	4482	1082	563	5157
F_06	6196	3733	1018	677	4587
M_SC	3	7	3	0	20
F_SC	0	6	6	0	33
M_ST	1999	427	5806	2666	7670
F_ST	2598	517	9723	3968	10843
M_LIT	13381	10513	4534	1842	13243
F_LIT	11364	7891	5840	1962	13477
M_ILL	10007	9072	2012	942	7348
F_ILL	18432	15211	5124	2244	16504
TOT_WORK_M	6723	6982	2775	1002	5717
TOT_WORK_F	3752	4200	4800	1118	7692
MAINWORK_M	2763	4628	1940	491	2523
MAINWORK_F	1275	1733	2923	408	2267
MAIN_CL_M	486	1098	519	35	743
MAIN_CL_F	235	357	1205	102	766
MAIN_AL_M	407	442	36	8	254
MAIN_AL_F	143	108	71	24	237
MAIN_HH_M	78	538	19	9	35
MAIN_HH_F	86	343	55	6	64

MAIN_OT_M	1792	2550	1366	439	1491
MAIN_OT_F	811	925	1592	276	1200
MARGWORK_M	3960	2354	835	511	3194
MARGWORK_F	2477	2467	1877	710	5425
MARG_CL_M	619	384	360	135	1327
MARG_CL_F	580	661	1250	286	2462
MARG_AL_M	2052	915	44	63	1037
MARG_AL_F	641	547	157	176	1069
MARG_HH_M	142	369	15	10	62
MARG_HH_F	244	627	32	43	319
MARG_OT_M	1147	686	416	303	768
MARG_OT_F	1012	632	438	205	1575
MARGWORK_3_6_M	16665	12603	3771	1782	14874
MARGWORK_3_6_F	26044	18902	6164	3088	22289
MARG_CL_3_6_M	2810	1829	721	317	2320
MARG_CL_3_6_F	1728	1752	1689	463	3497
MARG_AL_3_6_M	439	261	316	74	862
MARG_AL_3_6_F	343	432	1161	158	1419
MARG_HH_3_6_M	1372	729	41	50	832
MARG_HH_3_6_F	389	399	123	126	767
MARG_OT_3_6_M	110	293	15	6	38
MARG_OT_3_6_F	198	449	28	33	214
MARGWORK_0_3_M	889	546	349	187	588
MARGWORK_0_3_F	798	472	377	146	1097
MARG_CL_0_3_M	1150	525	114	194	874
MARG_CL_0_3_F	749	715	188	247	1928
MARG_AL_0_3_M	180	123	44	61	465
MARG_AL_0_3_F	237	229	89	128	1043
MARG_HH_0_3_M	680	186	3	13	205
MARG_HH_0_3_F	252	148	34	50	302
MARG_OT_0_3_M	32	76	0	4	24
MARG_OT_0_3_F	46	178	4	10	105
NON_WORK_M	258	140	67	116	180
NON_WORK_F	214	160	61	59	478

Table 1 First Five Rows for PCA Data

Info:

The Info of the dataset is

```
<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
```

Summary:

The Summary of the dataset is

	count	mean	std	min	25%	50%	75%	max
No_HH	640	51222.87	48135.41	350	19484	35837	68892	310450
TOT_M	640	79940.58	73384.51	391	30228	58339	107918.5	485417
TOT_F	640	122372.1	113600.7	698	46517.75	87724.5	164251.8	750392
M_06	640	12309.1	11500.91	56	4733.75	9159	16520.25	96223
F_06	640	11942.3	11326.29	56	4672.25	8663	15902.25	95129
M_SC	640	13820.95	14426.37	0	3466.25	9591.5	19429.75	103307
F_SC	640	20778.39	21727.89	0	5603.25	13709	29180	156429
M_ST	640	6191.808	9912.669	0	293.75	2333.5	7658	96785
F_ST	640	10155.64	15875.7	0	429.5	3834.5	12480.25	130119
M_LIT	640	57967.98	55910.28	286	21298	42693.5	77989.5	403261
F_LIT	640	66359.57	75037.86	371	20932	43796.5	84799.75	571140
M_ILL	640	21972.6	19825.61	105	8590	15767.5	29512.5	105961
F_ILL	640	56012.52	47116.69	327	22367	42386	78471	254160
TOT_WORK_M	640	37992.41	36419.54	100	13753.5	27936.5	50226.75	269422
TOT_WORK_F	640	41295.76	37192.36	357	16097.75	30588.5	53234.25	257848
MAINWORK_M	640	30204.45	31480.92	65	9787	21250.5	40119	247911
MAINWORK_F	640	28198.85	29998.26	240	9502.25	18484	35063.25	226166
MAIN_CL_M	640	5424.342	4739.162	0	2023.5	4160.5	7695	29113
MAIN_CL_F	640	5486.042	5326.363	0	1920.25	3908.5	7286.25	36193
MAIN_AL_M	640	5849.109	6399.508	0	1070.25	3936.5	8067.25	40843
MAIN_AL_F	640	8925.995	12864.29	0	1408.75	3933.5	10617.5	87945
MAIN_HH_M	640	883.8938	1278.642	0	187.5	498.5	1099.25	16429
MAIN_HH_F	640	1380.773	3179.414	0	248.75	540.5	1435.75	45979
MAIN_OT_M	640	18047.1	26068.48	36	3997.5	9598	21249.5	240855
MAIN_OT_F	640	12406.04	18972.2	153	3142.5	6380.5	14368.25	209355
MARGWORK_M	640	7787.961	7410.792	35	2937.5	5627	9800.25	47553
MARGWORK_F	640	13096.91	10996.47	117	5424.5	10175	18879.25	66915
MARG_CL_M	640	1040.738	1311.547	0	311.75	606.5	1281	13201
MARG_CL_F	640	2307.683	3564.626	0	630.25	1226	2659.25	44324
MARG_AL_M	640	3304.327	3781.556	0	873.5	2062	4300.75	23719
MARG_AL_F	640	6463.281	6773.876	0	1402.5	4020.5	9089.25	45301
MARG_HH_M	640	316.7422	462.6619	0	71.75	166	356.5	4298
MARG_HH_F	640	786.6266	1198.718	0	171.75	429	962.5	15448
MARG_OT_M	640	3126.155	3609.392	7	935.5	2036	3985.25	24728
MARG_OT_F	640	3539.323	4115.191	19	1071.75	2349.5	4400.5	36377
MARGWORK_3_6_M	640	41948.17	39045.32	291	16208.25	30315	57218.75	300937
MARGWORK_3_6_F	640	81076.32	82970.41	341	26619.5	56793	107924	676450
MARG_CL_3_6_M	640	6394.988	6019.807	27	2372	4630	8167	39106
MARG_CL_3_6_F	640	10339.86	8467.473	85	4351.5	8295	15102	50065

MARG_AL_3_6_M	640	789.8484	905.6393	0	235.5	480.5	986	7426
MARG_AL_3_6_F	640	1749.584	2496.542	0	497.25	985.5	2059	27171
MARG_HH_3_6_M	640	2743.636	3059.586	0	718.75	1714.5	3702.25	19343
MARG_HH_3_6_F	640	5169.85	5335.641	0	1113.75	3294	7502.25	36253
MARG_OT_3_6_M	640	245.3625	358.7286	0	58	129.5	276	3535
MARG_OT_3_6_F	640	585.8844	900.0258	0	127.75	320.5	719.25	12094
MARGWORK_0_3_M	640	2616.141	3036.964	7	755	1681.5	3320.25	20648
MARGWORK_0_3_F	640	2834.545	3327.837	14	833.5	1834.5	3610.5	25844
MARG_CL_0_3_M	640	1392.973	1489.707	4	489.5	949	1714	9875
MARG_CL_0_3_F	640	2757.05	2788.777	30	957.25	1928	3599.75	21611
MARG_AL_0_3_M	640	250.8891	453.3366	0	47	114.5	270.75	5775
MARG_AL_0_3_F	640	558.0984	1117.643	0	109	247.5	568.75	17153
MARG_HH_0_3_M	640	560.6906	762.579	0	136.5	308	642	6116
MARG_HH_0_3_F	640	1293.431	1585.378	0	298	717	1710.75	13714
MARG_OT_0_3_M	640	71.37969	107.8976	0	14	35	79	895
MARG_OT_0_3_F	640	200.7422	309.7409	0	43	113	240	3354
NON_WORK_M	640	510.0141	610.6032	0	161	326	604.5	6456
NON_WORK_F	640	704.7781	910.2092	5	220.5	464.5	853.5	10533

Table 2 Summary for PCA Data

Duplicates and Null Values:

The dataset has no null values and no Duplicates.

Part 2- PCA: Perform detailed Exploratory analysis by creating certain questions like

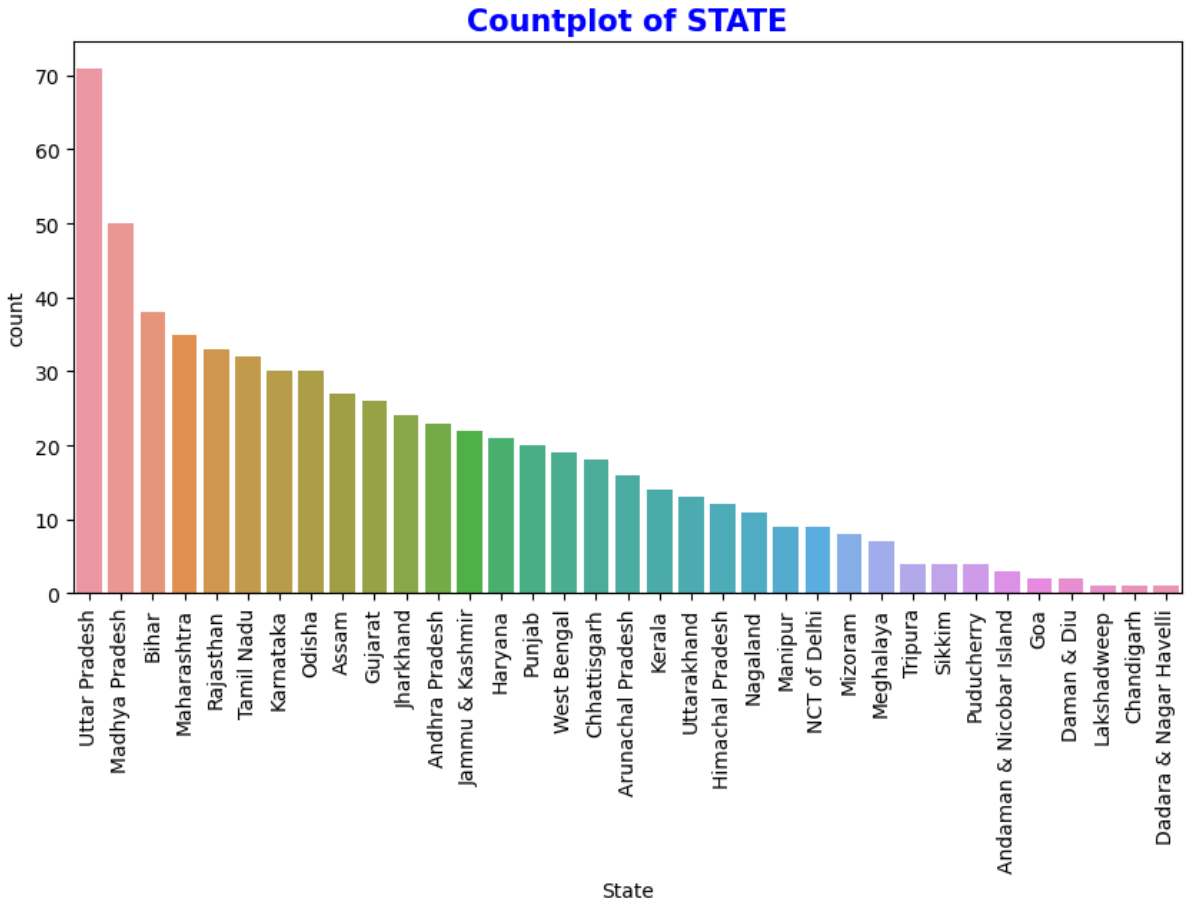


Figure A Count Plot for States

The Highest Census Data is from UT.

The Gender Ratio is Found Using the Formula = Population of Female/ Population of Male.

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

Uttar Pradesh has the highest Gender ratio. For 1 male, UT has **93** Females.

Lakshadweep has the lowest Gender ratio. For 1 male, Lakshadweep has **1** Females.

(ii) Which district has the highest & lowest gender ratio?

Raigarh District from Chhattisgarh has the highest Gender ratio. For 1 male, it has **4** Females.

Lakshadweep district has the lowest Gender ratio. For 1 male, it has **1** Females.

(iii) 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

For 5 variables, I have Choosen No_HH, M_SC, F_SC, M_ST, F_ST.

The Questions Formed are below ones:

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

(ii) Which State & District has the Zero Scheduled Castes population?

(iii) Which state has highest gender ratio and which has the lowest for Scheduled Tribes?

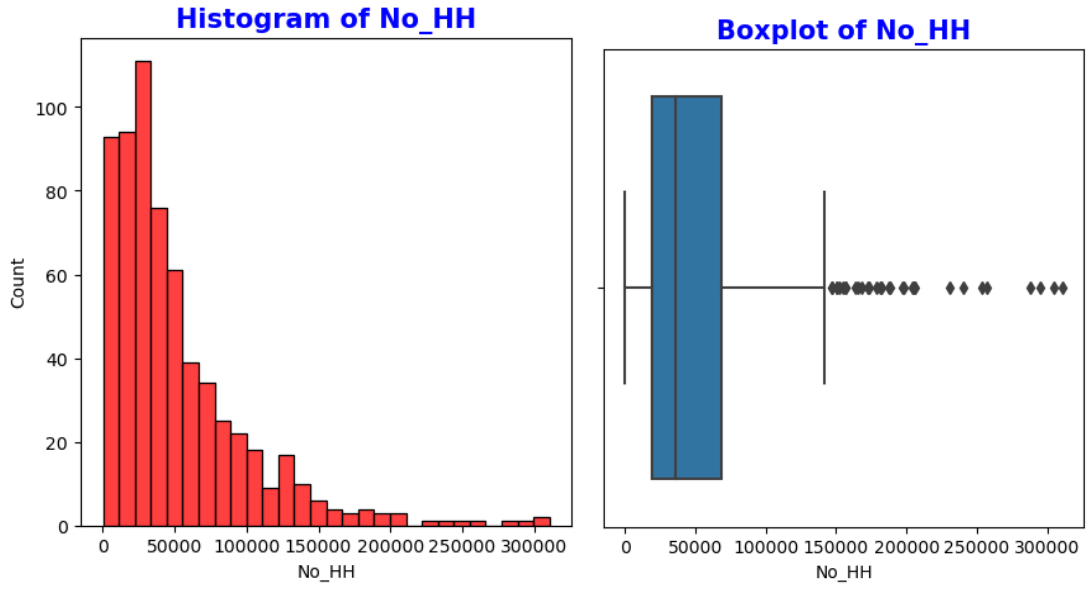
(iv) Which State & District has the Zero Scheduled Tribes population?

Univariate Analysis (Summary, Histogram and Boxplot) for these 5 variables

NO_HH:

```
Description of No_HH
--------------------------------------------------------------------------------
-
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 Distribution of No_HH
--------------------------------------------------------------------------------
-
```

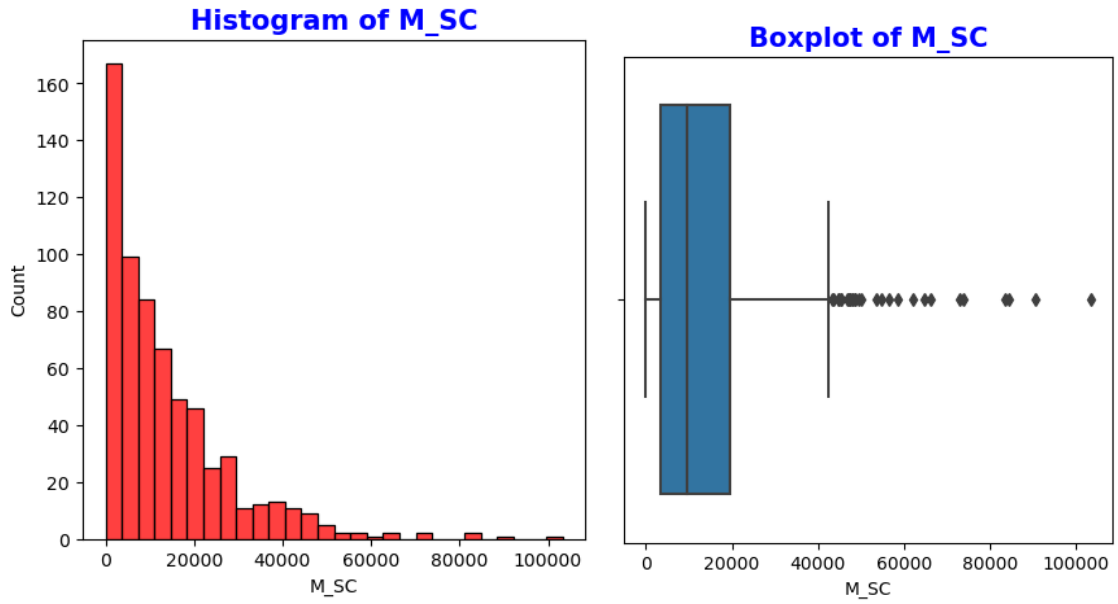
Figure B Histogram & Boxplot for NO_HH



M_SC:

```
Description of M_SC
-------------------------------------------------------------------------------
-
count       640.000000
mean      13820.946875
std       14426.373130
min           0.000000
25%        3466.250000
50%        9591.500000
75%       19429.750000
max      103307.000000
Name: M_SC, dtype: float64 Distribution of M_SC
-------------------------------------------------------------------------------
-
```

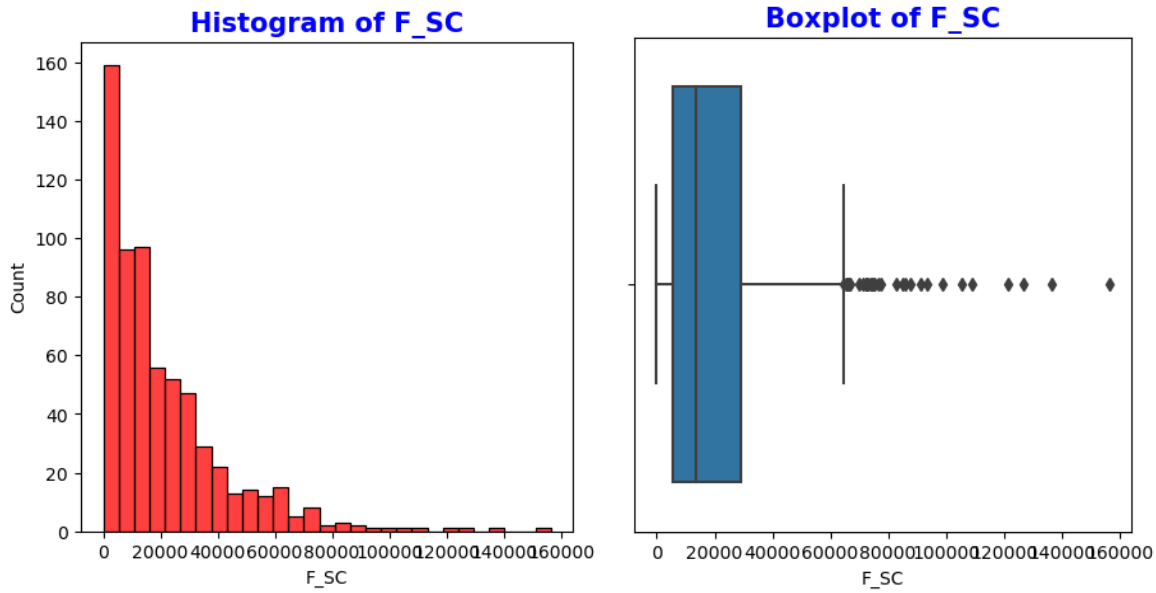
Figure C Histogram & Boxplot for M_SC



F_SC:

```
Description of F_SC
--------------------------------------------------------------------------------
-
count       640.000000
mean      20778.392188
std       21727.887713
min           0.000000
25%        5603.250000
50%       13709.000000
75%       29180.000000
max      156429.000000
Name: F_SC, dtype: float64 Distribution of F_SC
--------------------------------------------------------------------------------
-
```

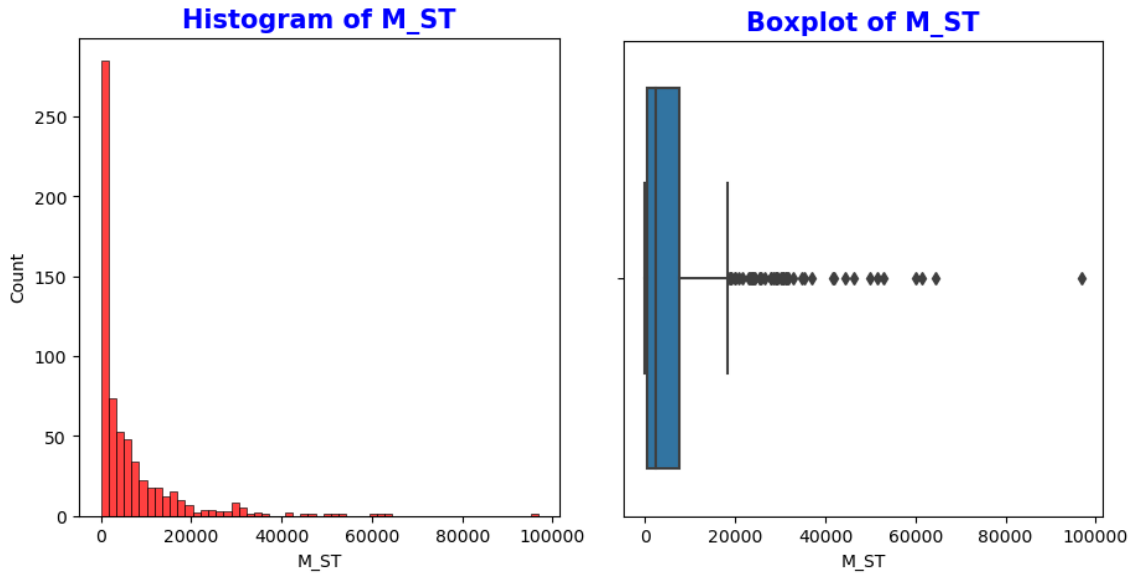
Figure D Histogram & Boxplot for F_SC



M_ST:

```
Description of M_ST
--------------------------------------------------------------------------------
-
count      640.000000
mean      6191.807813
std       9912.668948
min          0.000000
25%        293.750000
50%       2333.500000
75%       7658.000000
max      96785.000000
Name: M_ST, dtype: float64 Distribution of M_ST
--------------------------------------------------------------------------------
-
```

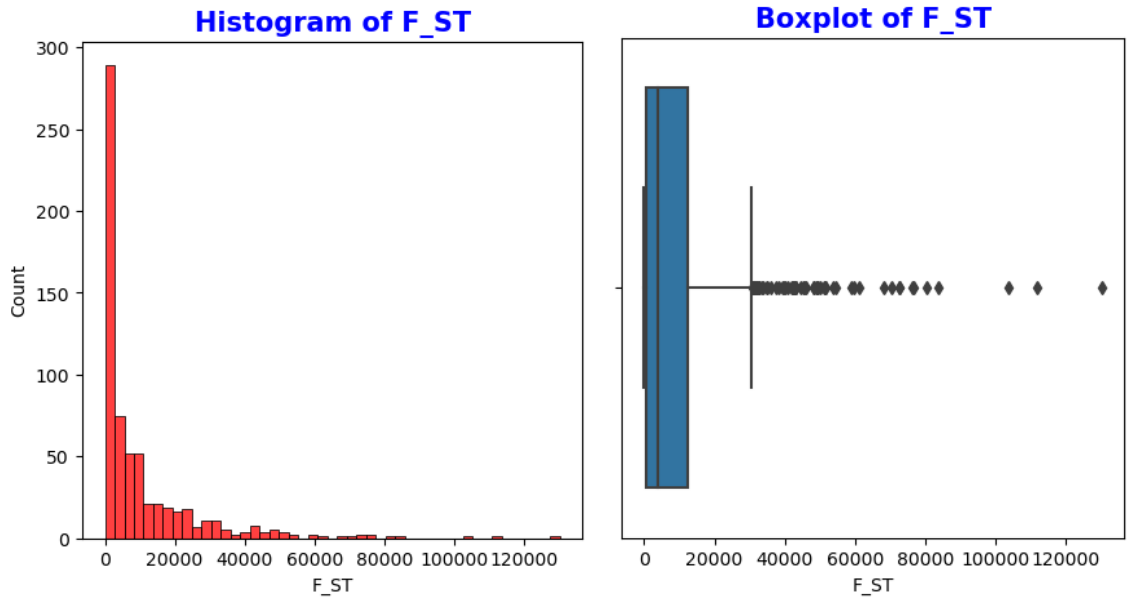
Figure E Histogram & Boxplot for M_ST



F_ST:

```
Description of F_ST
--------------------------------------------------------------------------------
-
count       640.000000
mean      10155.640625
std       15875.701488
min           0.000000
25%         429.500000
50%        3834.500000
75%       12480.250000
max      130119.000000
Name: F_ST, dtype: float64 Distribution of F_ST
--------------------------------------------------------------------------------
-
```

Figure F Histogram & Boxplot for F_ST



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

Uttar Pradesh has the highest Gender ratio. For 1 male, UT has **93** Females.
Chandigarh has the lowest Gender ratio. For 1 male, Chandigarh has **1** Females.

(ii) Which State & District has the Zero Scheduled Castes population?

There are 43 District which has no scheduled castes population

District	State
Kargil	Jammu & Kashmir
Baramula	Jammu & Kashmir
Bandipore	Jammu & Kashmir
Ganderbal	Jammu & Kashmir
Pulwama	Jammu & Kashmir
Shupiyan	Jammu & Kashmir
Anantnag	Jammu & Kashmir
Tawang	Arunachal Pradesh
West Kameng	Arunachal Pradesh
East Kameng	Arunachal Pradesh
Papum Pare	Arunachal Pradesh
Upper Subansiri	Arunachal Pradesh
West Siang	Arunachal Pradesh
East Siang	Arunachal Pradesh
Upper Siang	Arunachal Pradesh
Changlang	Arunachal Pradesh
Tirap	Arunachal Pradesh
Lower Subansiri	Arunachal Pradesh
Kurung Kumey	Arunachal Pradesh
Dibang Valley	Arunachal Pradesh
Lower Dibang Valley	Arunachal Pradesh
Lohit	Arunachal Pradesh
Anjaw	Arunachal Pradesh
Mon	Nagaland

Mokokchung	Nagaland
Zunheboto	Nagaland
Wokha	Nagaland
Dimapur	Nagaland
Phek	Nagaland
Tuensang	Nagaland
Longleng	Nagaland
Kiphire	Nagaland
Kohima	Nagaland
Peren	Nagaland
Tamenglong	Manipur
Mamit	Mizoram
Champhai	Mizoram
Serchhip	Mizoram
Saiha	Mizoram
Lakshadweep	Lakshadweep
Nicobars	Andaman & Nicobar Island
North & Middle Andaman	Andaman & Nicobar Island
South Andaman	Andaman & Nicobar Island

Table 3 List of District for Zero population SC

(iii) Which state has highest gender ratio and which has the lowest for Scheduled Tribes?

Uttar Pradesh has the highest Gender ratio. For 1 male, UT has 114 Females.
Lakshadweep has the lowest Gender ratio. For 1 male, Lakshadweep has 1 Females.

(iv) Which State & District has the Zero Scheduled Tribe population?

There are 56 Districts which as zero scheduled Tribes Population.

Area Name	State
Gurdaspur	Punjab
Kapurthala	Punjab
Jalandhar	Punjab
Hoshiarpur	Punjab
Shahid Bhagat Singh Nagar	Punjab
Fatehgarh Sahib	Punjab
Ludhiana	Punjab
Moga	Punjab
Firozpur	Punjab
Muktsar	Punjab
Faridkot	Punjab
Bathinda	Punjab
Mansa	Punjab
Patiala	Punjab
Amritsar	Punjab
Tarn Taran	Punjab
Rupnagar	Punjab
Sahibzada Ajit Singh Nagar	Punjab
Sangrur	Punjab
Barnala	Punjab
Chandigarh	Chandigarh
Panchkula	Haryana
Ambala	Haryana
Yamunanagar	Haryana
Kurukshetra	Haryana
Kaithal	Haryana
Karnal	Haryana
Panipat	Haryana
Sonipat	Haryana
Jind	Haryana
Fatehabad	Haryana
Sirsa	Haryana

Hisar	Haryana
Bhiwani	Haryana
Rohtak	Haryana
Jhajjar	Haryana
Mahendragarh	Haryana
Rewari	Haryana
Gurgaon	Haryana
Mewat	Haryana
Faridabad	Haryana
Palwal	Haryana
North West	NCT of Delhi
North	NCT of Delhi
North East	NCT of Delhi
East	NCT of Delhi
New Delhi	NCT of Delhi
Central	NCT of Delhi
West	NCT of Delhi
South West	NCT of Delhi
South	NCT of Delhi
Kannauj	Uttar Pradesh
Yanam	Puducherry
Puducherry	Puducherry
Mahe	Puducherry
Karaikal	Puducherry

Table 4 List of District for Zero population ST

Part 2- PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?

Treating Outlier is important as the outlier influences the mean which affects the PCA. The PCA is based on the variance and Co variance. So, treating outlier is a necessary step.

Part 2- PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.

Before Scaling: **After Scaling:**

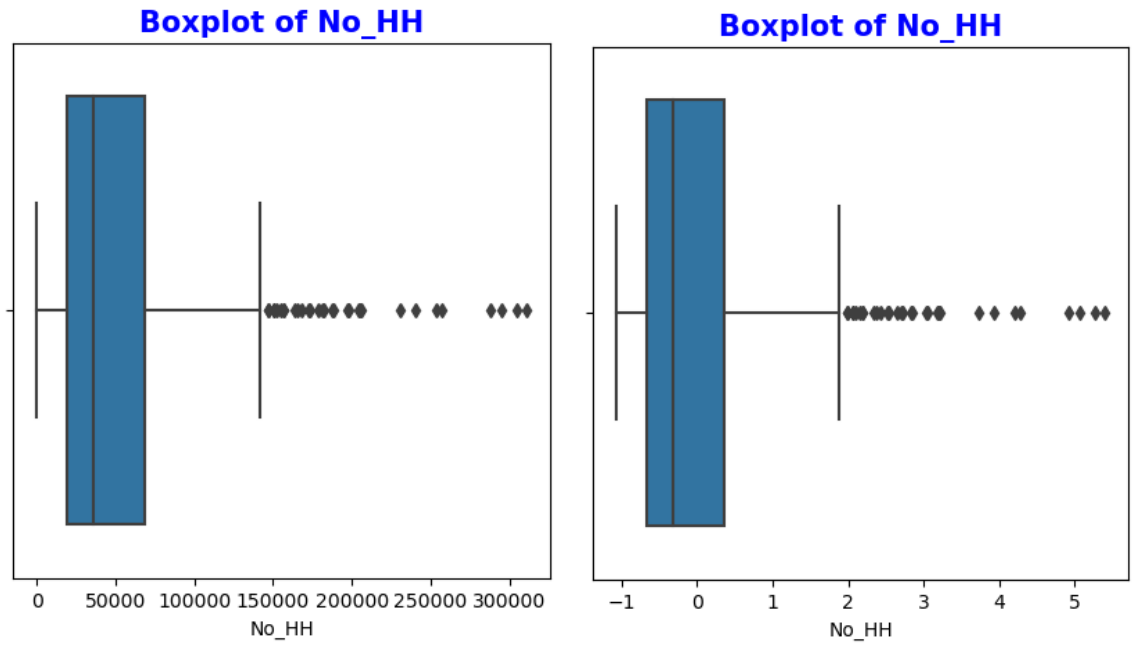


Figure G Boxplot Before and After Scaling

Refer Jupyter Workbook for all variables. Show casing only one variable (No_HH) :

Scaling Doesn't affect the outliers, only the scales have been normalised.

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

Assumptions for PCA:

The Outlier has been treated by IQR method.

Bartlett's Sphericity Test:

Null Hypothesis - The Correlation matrix is like an identity matrix i.e no correlation between variables exists for PCA

Alternate Hypothesis - Sufficient Correlations exist between variables for PCA

P- Value is 0.0. P-Value less than alpha (0.5). Reject Null Hypothesis.

Sufficient Correlations exist between variables for PCA.

Kaiser Mayer Olkin Test:

The KMO is 0.94. Adequacy test is also suitable.

Covariance Matrix:

	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
No_HH	1.001565	0.912699	0.973013	0.812856	0.809883	0.806713	0.858562	0.116300	0.122722	0.931350	...	0.604943	0.61
TOT_M	0.912699	1.001565	0.980122	0.965044	0.960153	0.877158	0.861703	0.023439	0.013301	0.989312	...	0.739665	0.63
TOT_F	0.973013	0.980122	1.001565	0.914418	0.911167	0.857664	0.876435	0.076189	0.074248	0.983281	...	0.697119	0.65
M_06	0.812856	0.965044	0.914418	1.001565	0.999032	0.833344	0.796794	-0.006081	-0.021166	0.924761	...	0.799076	0.68
F_06	0.809883	0.960153	0.911167	0.999032	1.001565	0.823888	0.790043	0.006803	-0.007896	0.915929	...	0.805050	0.68
M_SC	0.806713	0.877158	0.857664	0.833344	0.823888	1.001565	0.984688	-0.096913	-0.099226	0.868007	...	0.647698	0.55
F_SC	0.858562	0.861703	0.876435	0.796794	0.790043	0.984688	1.001565	-0.052859	-0.048597	0.862923	...	0.620049	0.57
M_ST	0.116300	0.023439	0.076189	-0.006081	0.006803	-0.096913	-0.052859	1.001565	0.994481	0.026290	...	0.094899	0.20
F_ST	0.122722	0.013301	0.074248	-0.021166	-0.007896	-0.099226	-0.048597	0.994481	1.001565	0.017617	...	0.083930	0.20
M_LIT	0.931350	0.989312	0.983281	0.924761	0.915929	0.868007	0.862923	0.026290	0.017617	1.001565	...	0.694535	0.60
F_LIT	0.940747	0.937579	0.963424	0.844453	0.835104	0.805082	0.823245	0.047388	0.043933	0.974173	...	0.615830	0.55
M_ILL	0.782405	0.933452	0.880243	0.967971	0.972547	0.822290	0.784357	0.023378	0.010249	0.869070	...	0.781156	0.66
F_ILL	0.896107	0.917169	0.928913	0.896778	0.900544	0.842658	0.858401	0.112222	0.112487	0.877996	...	0.728973	0.70
TOT_WORK_M	0.938328	0.977458	0.974326	0.898655	0.893232	0.868242	0.866029	0.057298	0.049061	0.982191	...	0.655936	0.56
TOT_WORK_F	0.948620	0.825119	0.904224	0.732839	0.734787	0.733823	0.803562	0.250209	0.257052	0.842559	...	0.566210	0.63
MAINWORK_M	0.926588	0.936031	0.943223	0.833607	0.825308	0.838925	0.842746	0.047749	0.040740	0.954067	...	0.531633	0.44
MAINWORK_F	0.921397	0.772433	0.858357	0.650808	0.651110	0.690579	0.764551	0.217172	0.224043	0.805023	...	0.397956	0.44

Table 5 Covariance Matrix

Eigen Values:

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

Eigen Vector:

	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	MARG_A
0	0.149222	0.159169	0.158209	0.156340	0.156814	0.143350	0.143537	0.018849	0.017878	0.155152	...	0.142987	0.133784	
1	-0.115487	-0.080239	-0.093718	-0.020341	-0.014310	-0.079667	-0.087098	0.069101	0.067316	-0.105986	...	0.136839	0.166416	
2	0.101528	-0.038662	0.028959	-0.074419	-0.068223	-0.037619	0.021350	0.323827	0.338705	-0.032107	...	-0.103565	0.033423	
3	0.076814	0.052976	0.070022	0.028520	0.016398	0.010210	0.016244	0.091143	0.079554	0.089187	...	-0.018223	0.005954	
4	-0.012090	-0.042344	-0.022927	-0.080339	-0.078326	-0.167893	-0.158092	0.418412	0.415965	-0.014033	...	0.094293	0.112351	
5	0.082558	0.073667	0.082812	0.092379	0.080010	0.050969	0.054568	-0.231809	-0.214542	0.081378	...	0.111045	0.185882	
6	0.106896	-0.124085	-0.010291	-0.200807	-0.203411	-0.040399	0.053990	-0.355238	-0.327677	-0.067062	...	-0.025902	0.178500	
7	-0.099515	-0.108870	-0.115276	-0.132944	-0.139342	0.189170	0.177363	-0.071632	-0.078392	-0.102886	...	0.018271	-0.004071	
8	0.026100	0.032856	0.036405	0.138404	0.165715	-0.531744	-0.515063	-0.113019	-0.136031	-0.017445	...	-0.004772	-0.023984	
9	0.068124	-0.048423	-0.022468	-0.157252	-0.145040	-0.098447	-0.065840	-0.008382	-0.028617	0.000581	...	0.106542	0.008600	
10	-0.058605	0.029489	-0.020147	-0.009180	-0.025584	-0.194623	-0.250356	-0.082493	-0.081426	0.023816	...	-0.284163	-0.155181	
11	-0.021808	-0.047662	-0.042837	-0.146640	-0.144601	-0.122639	-0.114550	-0.055520	-0.051232	0.034683	...	-0.011059	-0.171620	
12	-0.017070	0.005930	0.004689	0.033536	0.034778	-0.134248	-0.153439	-0.047831	-0.020856	0.033251	...	0.119289	0.131845	
13	0.073049	0.001557	0.023915	-0.082495	-0.074311	0.019392	0.032860	-0.032516	-0.007417	0.027021	...	0.160843	0.117313	
14	0.026658	-0.063242	-0.061363	-0.058970	-0.046709	-0.093434	-0.095711	-0.028360	-0.021992	-0.171426	...	0.118122	0.182256	
15	0.140284	0.051650	0.112634	-0.044144	-0.080215	-0.112254	-0.075261	0.007260	0.031458	0.176838	...	0.055308	0.260353	
16	-0.159587	0.002659	-0.102147	0.049755	-0.002391	-0.063561	-0.155701	-0.011279	-0.035854	-0.008326	...	0.249665	0.225173	
17	0.112338	0.009969	0.121359	-0.012294	-0.028758	-0.107590	-0.027758	-0.010252	-0.006185	0.041564	...	-0.102858	0.010766	
18	0.038334	0.021547	0.029080	-0.031091	-0.054024	-0.044062	-0.031930	-0.013127	-0.030441	-0.029849	...	0.069282	-0.021111	

Table 6 Eigen Vector

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

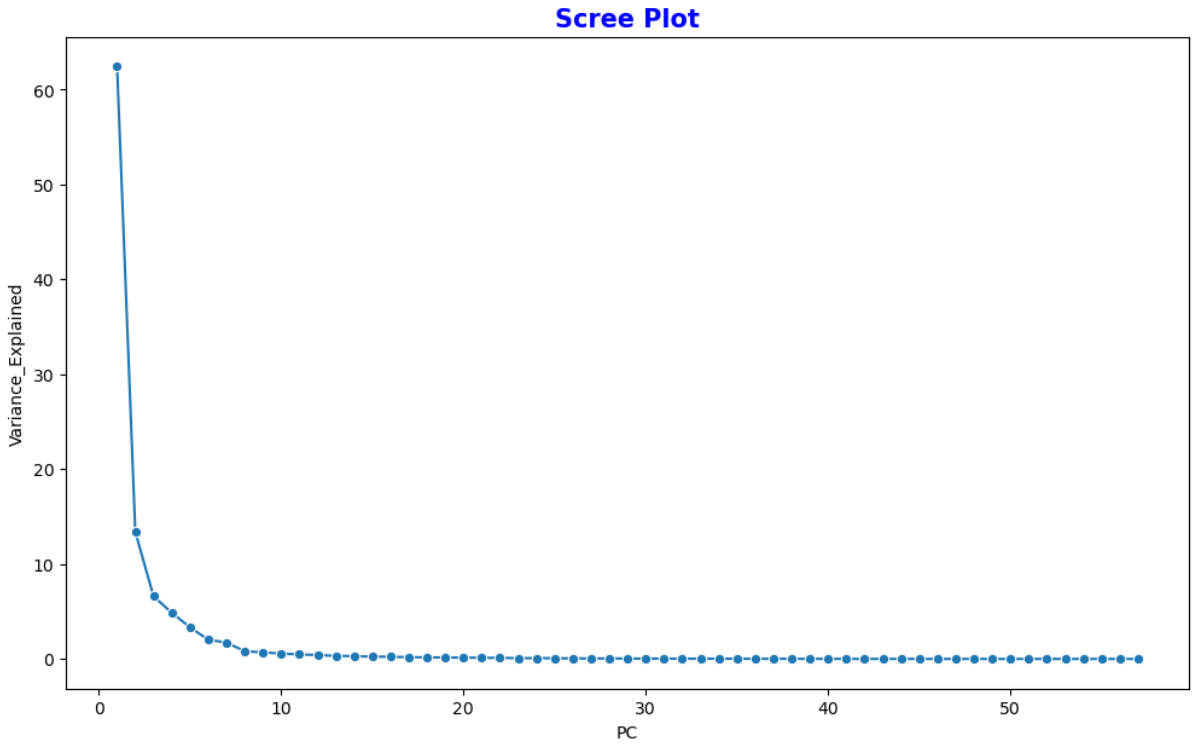


Figure H Scree Plot

PC	Eigen Values	Variance Explained	Cumulative Variance Explained
1	35.64886	62.44415	62.444145
2	7.643576	13.38883	75.832974
3	3.769196	6.602291	82.435265
4	2.777223	4.864709	87.299974
5	1.906949	3.340297	90.640271

Table 7 Variance & Cumulative Variance

From the Scree plot and the Cumulative variance, 5 is the optimum number of PC. It also has 90% variance.

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.

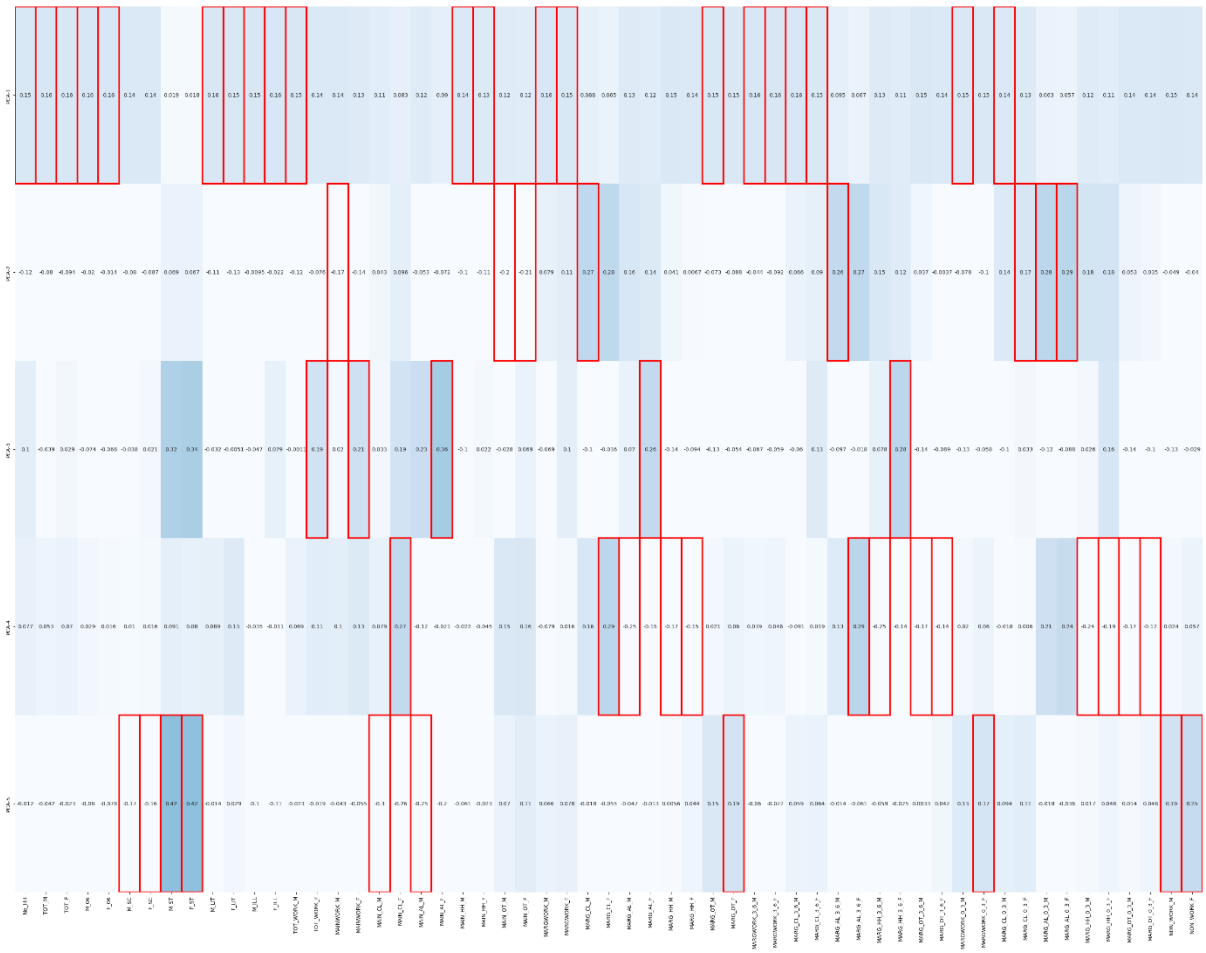


Figure I PCA Variable Components

PC Vs Actual Variables:

PCA-1	PCA-2	PCA-3	PCA-4	PCA-5
No_HH	MAINWORK_M	TOT_WORK_F	MAIN_CL_F	M_SC
TOT_M	MAIN_OT_M	MAINWORK_F	MARG_CL_F	F_SC
TOT_F	MAIN_OT_F	MAIN_AL_F	MARG_AL_M	M_ST
M_06	MARG_CL_M	MARG_AL_F	MARG_HH_M	F_ST
F_06	MARG_AL_3_6_M	MARG_HH_3_6_F	MARG_HH_F	MAIN_CL_M
M_LIT	MARG_CL_0_3_F		MARG_AL_3_6_F	MAIN_AL_M
F_LIT	MARG_AL_0_3_M		MARG_HH_3_6_M	MARG_OT_F
M_ILL	MARG_AL_0_3_F		MARG_OT_3_6_M	MARGWORK_0_3_F
F_ILL			MARG_OT_3_6_F	NON_WORK_M
TOT_WORK_M			MARG_HH_0_3_M	NON_WORK_F
MAIN_HH_M			MARG_HH_0_3_F	
MAIN_HH_F			MARG_OT_0_3_M	
MARGWORK_M			MARG_OT_0_3_F	
MARGWORK_F				
MARG_OT_M				
MARGWORK_3_6_M				
MARGWORK_3_6_F				
MARG_CL_3_6_M				
MARG_CL_3_6_F				
MARGWORK_0_3_M				
MARG_CL_0_3_M				

Table 8 List of PCA Variable Components

PC1:

PC 1 describes more about the total household, population and Literates and Illiterate population.

PCA-1 Variable Components Description
No of Household
Total population Male
Total population Female
Population in the age group 0-6 Male
Population in the age group 0-6 Female
Literates population Male
Literates population Female
Illiterate Male
Illiterate Female
Total Worker Population Male
Main Household Industries Population Male
Main Household Industries Population Female
Marginal Worker Population Male
Marginal Worker Population Female
Marginal Other Workers Population Male
Marginal Worker Population 3-6 Male
Marginal Worker Population 3-6 Female
Marginal Cultivator Population 3-6 Male
Marginal Cultivator Population 3-6 Female
Marginal Worker Population 0-3 Male
Marginal Cultivator Population 0-3 Male

Table 9 List of PCA 1 Variable Components

PC2:

PC 2 describes more about the .

PCA-2 Variable Components Description
Main Working Population Male
Main Other Workers Population Male
Main Other Workers Population Female
Marginal Cultivator Population Male
Marginal Agriculture Labourers Population 3-6 Male
Marginal Cultivator Population 0-3 Female
Marginal Agriculture Labourers Population 0-3 Male
Marginal Agriculture Labourers Population 0-3 Female

Table 10 List of PCA 2 Variable Components

PC3:

PC 3 describes more about the Female Working and labourers Population.

PCA-3 Variable Components Description
Total Worker Population Female
Main Working Population Female
Main Agricultural Labourers Population Female
Marginal Agriculture Labourers Population Female
Marginal Household Industries Population 3-6 Female

Table 11 List of PCA 3 Variable Components

PC4:

PC 4 describes more about the agriculture & household Industries and Workers Population.

PCA-4 Variable Components Description
Main Cultivator Population Female
Marginal Cultivator Population Female
Marginal Agriculture Labourers Population Male
Marginal Household Industries Population Male
Marginal Household Industries Population Female
Marginal Agriculture Labourers Population 3-6 Female
Marginal Household Industries Population 3-6 Male
Marginal Other Workers Population Person 3-6 Male
Marginal Other Workers Population Person 3-6 Female
Marginal Household Industries Population 0-3 Male
Marginal Household Industries Population 0-3 Female
Marginal Other Workers Population 0-3 Male
Marginal Other Workers Population 0-3 Female

Table 12 List of PCA 4 Variable Components

PC5:

PC 5 describes more about the Scheduled Casts and Tribe & Non-working population.

PCA-5 Variable Components Description
Scheduled Castes population Male
Scheduled Castes population Female
Scheduled Tribes population Male
Scheduled Tribes population Female
Main Cultivator Population Male
Main Agricultural Labourers Population Male
Marginal Other Workers Population Female
Marginal Worker Population 0-3 Female
Non-Working Population Male
Non-Working Population Female

Table 13 List of PCA 5 Variable Components

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

PC 1 = (0.15) * No_HH + (0.16) * TOT_M + (0.16) * TOT_F + (0.16) * M_06 + (0.16) * F_06 + (0.14) * M_SC + (0.14) * F_SC + (0.02) * M_ST +

(0.02) * F_ST + (0.16) * M_LIT + (0.15) * F_LIT + (0.15) * M_ILL + (0.16) * F_ILL + (0.15) * TOT_WORK_M + (0.14) * TOT_WORK_F +

(0.14) * MAINWORK_M + (0.13) * MAINWORK_F + (0.11) * MAIN_CL_M +

(0.08) * MAIN_CL_F + (0.12) * MAIN_AL_M + (0.09) * MAIN_AL_F +

(0.14) * MAIN_HH_M + (0.13) * MAIN_HH_F + (0.12) * MAIN_OT_M +

(0.12) * MAIN_OT_F + (0.16) * MARGWORK_M + (0.15) * MARGWORK_F +

(0.09) * MARG_CL_M + (0.07) * MARG_CL_F + (0.13) * MARG_AL_M +

(0.12) * MARG_AL_F + (0.15) * MARG_HH_M + (0.14) * MARG_HH_F +

(0.15) * MARG_OT_M + (0.15) * MARG_OT_F + (0.16) * MARGWORK_3_6_M + (0.16) * MARGWORK_3_6_F + (0.16) * MARG_CL_3_6_M +

(0.15) * MARG_CL_3_6_F + (0.09) * MARG_AL_3_6_M +

(0.07) * MARG_AL_3_6_F + (0.13) * MARG_HH_3_6_M +

(0.11) * MARG_HH_3_6_F + (0.15) * MARG_OT_3_6_M +

(0.14) * MARG_OT_3_6_F + (0.15) * MARGWORK_0_3_M +

(0.15) * MARGWORK_0_3_F + (0.14) * MARG_CL_0_3_M +

(0.13) * MARG_CL_0_3_F + (0.06) * MARG_AL_0_3_M +

(0.06) * MARG_AL_0_3_F + (0.12) * MARG_HH_0_3_M +

(0.11) * MARG_HH_0_3_F + (0.14) * MARG_OT_0_3_M +

(0.14) * MARG_OT_0_3_F + (0.15) * NON_WORK_M + (0.14) * NON_WORK_F.