Business

Report

Index-

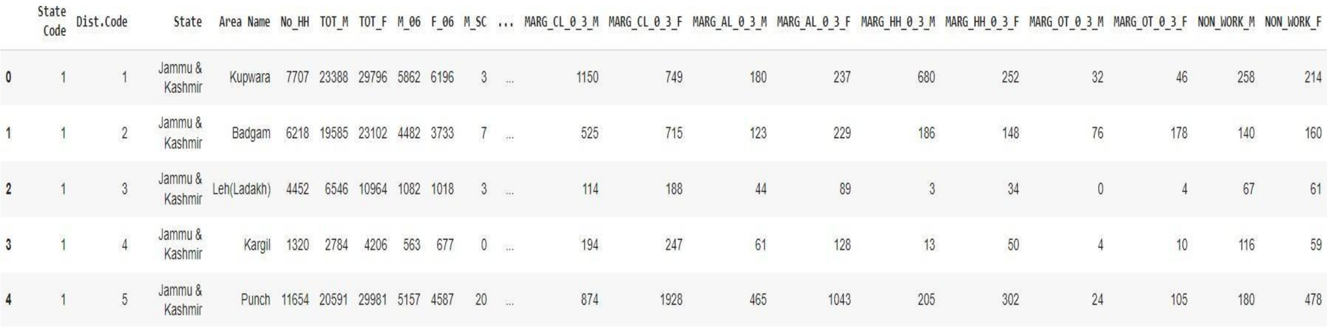
Part 2 - PCA:

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

Answer

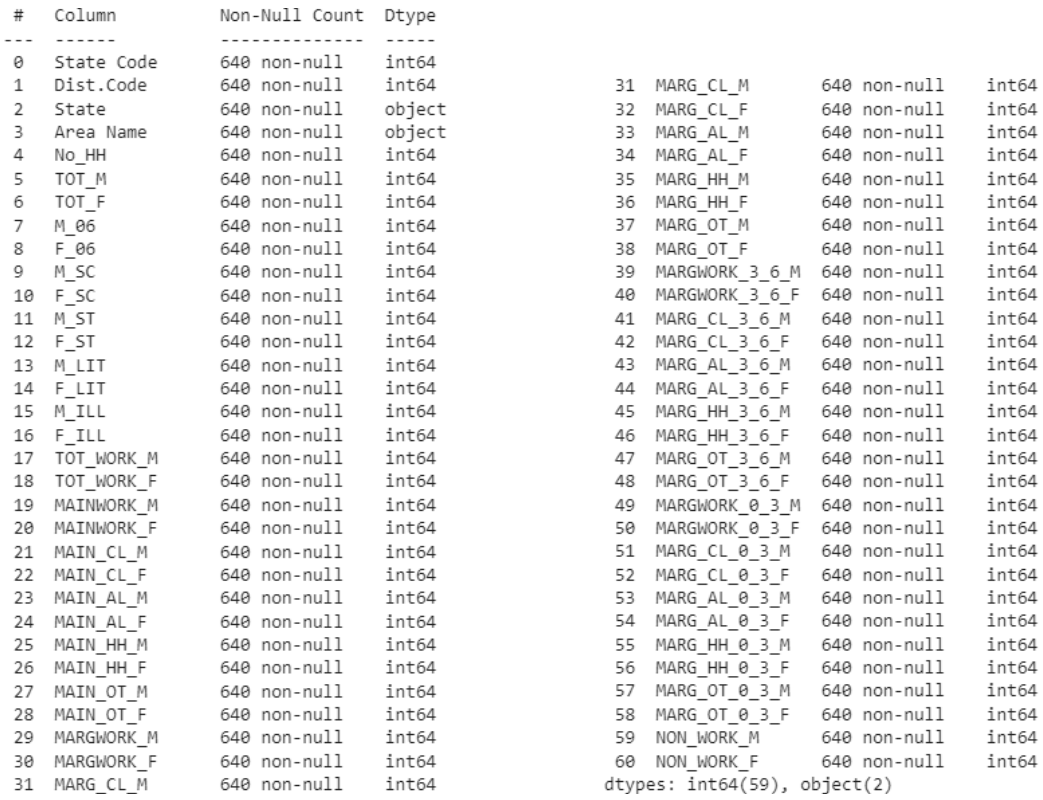
Data has 640 rows and 61 columns.

Let’s check few of the records from dataset.



	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_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180	237	680	252	32	46	258	214
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123	229	186	148	76	178	140	160
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	...	114	188	44	89	3	34	0	4	67	61
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	...	194	247	61	128	13	50	4	10	116	59
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	...	874	1928	465	1043	205	302	24	105	180	478

Let’s check what are available columns in data set and there data types-



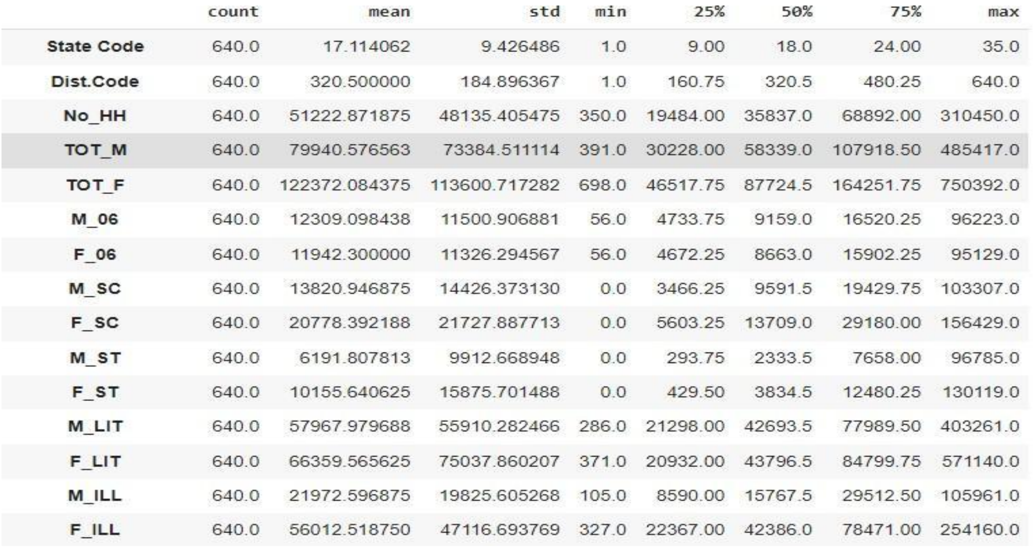
```
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   State Code     640 non-null    int64
 1   Dist.Code      640 non-null    int64               31  MARG_CL_M      640 non-null     int64
 2   State          640 non-null    object              32  MARG_CL_F   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

There are 59 columns with integer data types and 2 columns are categorical datatypes.

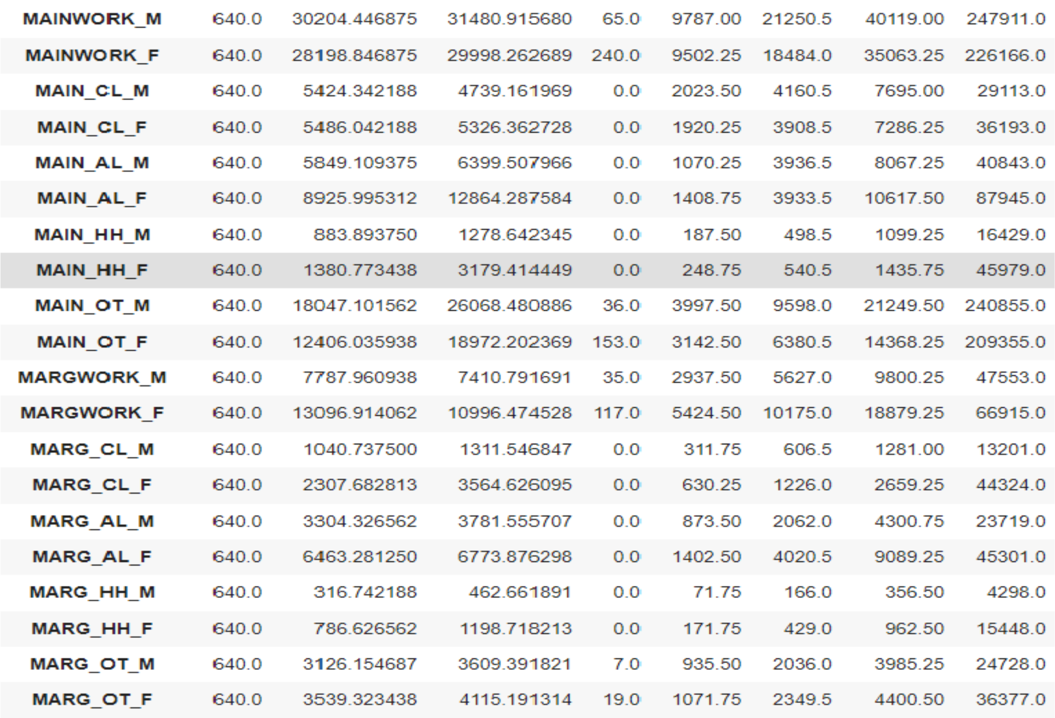
There are no null values found in dataset.

No duplicate records are there in dataset.

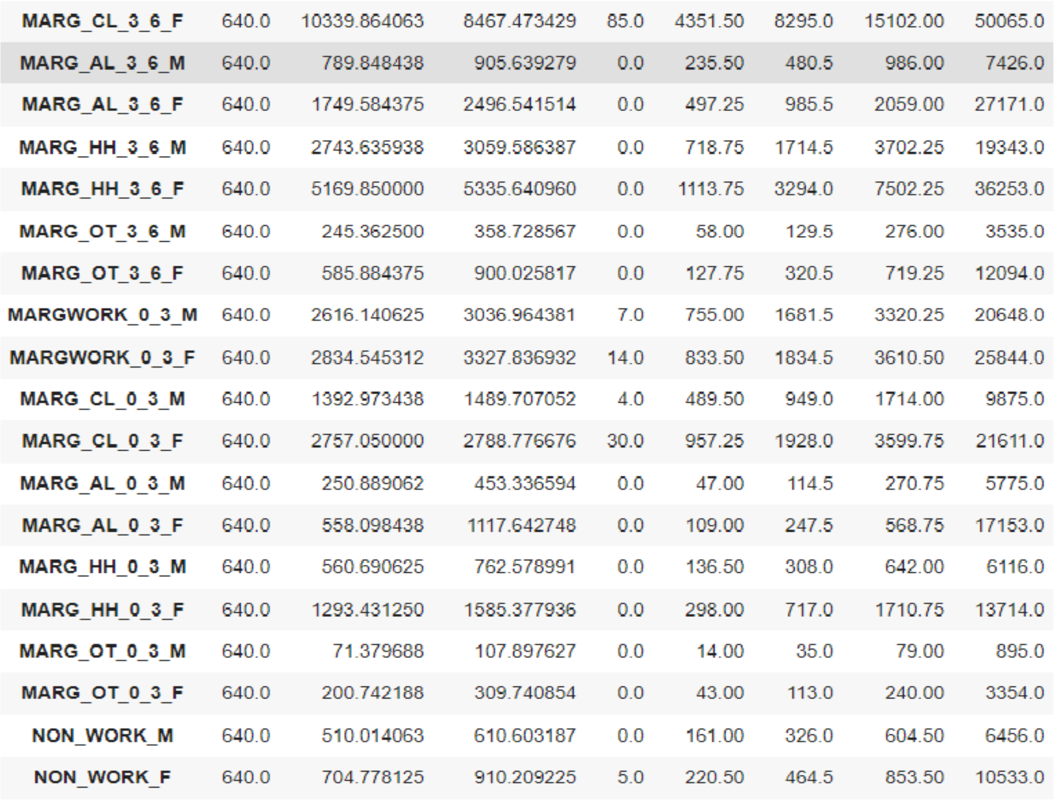
Let’s check out the summary of data –



	count	mean	std	min	25%	50%	75%	max
State Code	640.0	17.114062	9.426486	1.0	9.00	18.0	24.00	35.0
Dist.Code	640.0	320.500000	184.896367	1.0	160.75	320.5	480.25	640.0
No_HH	640.0	51222.871875	48135.405475	350.0	19484.00	35837.0	68892.00	310450.0
TOT_M	640.0	79940.576563	73384.511114	391.0	30228.00	58339.0	107918.50	485417.0
TOT_F	640.0	122372.084375	113600.717282	698.0	46517.75	87724.5	164251.75	750392.0
M_06	640.0	12309.098438	11500.906881	56.0	4733.75	9159.0	16520.25	96223.0
F_06	640.0	11942.300000	11326.294567	56.0	4672.25	8663.0	15902.25	95129.0
M_SC	640.0	13820.946875	14426.373130	0.0	3466.25	9591.5	19429.75	103307.0
F_SC	640.0	20778.392188	21727.887713	0.0	5603.25	13709.0	29180.00	156429.0
M_ST	640.0	6191.807813	9912.668948	0.0	293.75	2333.5	7658.00	96785.0
F_ST	640.0	10155.640625	15875.701488	0.0	429.50	3834.5	12480.25	130119.0
M_LIT	640.0	57967.979688	55910.282466	286.0	21298.00	42693.5	77989.50	403261.0
F_LIT	640.0	66359.565625	75037.860207	371.0	20932.00	43796.5	84799.75	571140.0
M_ILL	640.0	21972.596875	19825.605268	105.0	8590.00	15767.5	29512.50	105961.0
F_ILL	640.0	56012.518750	47116.693769	327.0	22367.00	42386.0	78471.00	254160.0



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



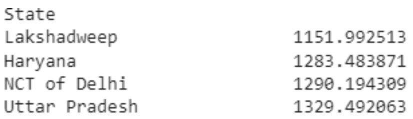
MARG_CL_3_6_F	640.0	10339.864063	8467.473429	85.0	4351.50	8295.0	15102.00	50065.0
MARG_AL_3_6_M	640.0	789.848438	905.639279	0.0	235.50	480.5	986.00	7426.0
MARG_AL_3_6_F	640.0	1749.584375	2496.541514	0.0	497.25	985.5	2059.00	27171.0
MARG_HH_3_6_M	640.0	2743.635938	3059.586387	0.0	718.75	1714.5	3702.25	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

There is no anomaly or negative found in above description.

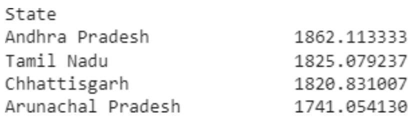
Part 2 - PCA:

2.2-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?

1. Which state has highest gender ratio and which has the lowest?



```
State
Lakshadweep                 1151.992513
Haryana                     1283.483871
NCT of Delhi                1290.194309
Uttar Pradesh               1329.492063
```

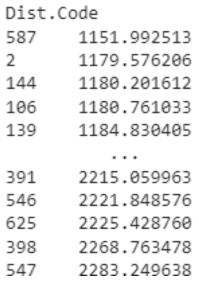


```
State
Andhra Pradesh              1862.113333
Tamil Nadu                  1825.079237
Chhattisgarh                1820.831007
Arunachal Pradesh           1741.054130
```

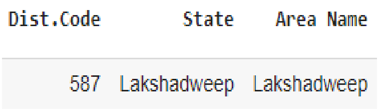
Lakshadweep has lowest gender ratio. There are 1152 females per 1000 males.

Andhra Pradesh has highest gender ratio. There are 1862 females per 1000 males.

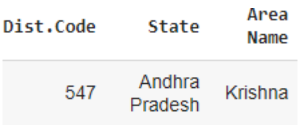
2. Which district has the highest & lowest gender ratio?



```
Dist.Code
587     1151.992513
2       1179.576206
144     1180.201612
106     1180.761033
139     1184.830405
           ...
391     2215.059963
546     2221.848576
625     2225.428760
398     2268.763478
547     2283.249638
```



Dist.Code	State	Area Name
587	Lakshadweep	Lakshadweep



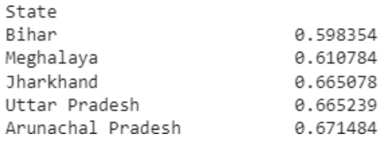
Dist.Code	State	Area Name
547	Andhra Pradesh	Krishna

Dist. Code 587 has lowest gender ratio and Dist. Code 547 has highest gender ratio.

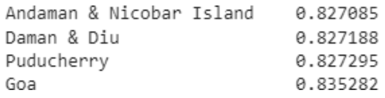
Area name – Lakshadweep has lowest gender ratio

Area name – Krishna has lowest gender ratio

3. Which state has maximum and minimum literacy rate for males and females.

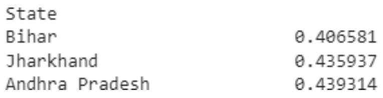


```
State
Bihar                         0.598354
Meghalaya                     0.610784
Jharkhand                     0.665078
Uttar Pradesh                 0.665239
Arunachal Pradesh             0.671484
```

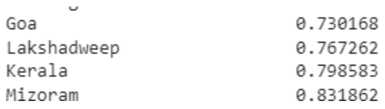


```
Andaman & Nicobar Island      0.827085
Daman & Diu                   0.827188
Puducherry                    0.827295
Goa                           0.835282
```

Bihar has lowest 0.59 and Goa has highest 0.83 literacy rate for male population.



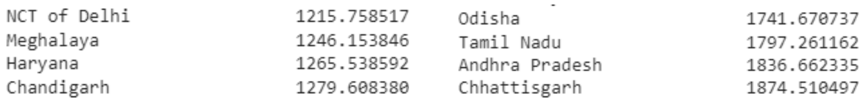
```
State
Bihar                         0.406581
Jharkhand                     0.435937
Andhra Pradesh                0.439314
```



```
Goa                           0.730168
Lakshadweep                   0.767262
Kerala                        0.798583
Mizoram                       0.831862
```

Bihar has lowest 0.40 and Mizoram has highest 0.83 literacy rate for female population.

4. Which state has highest gender ratio and which has the lowest for Scheduled Castes population?

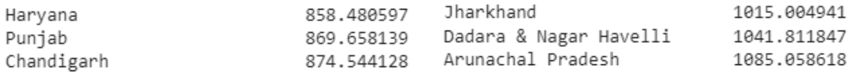


```
NCT of Delhi              1215.758517   Odisha                     1741.670737
Meghalaya                 1246.153846   Tamil Nadu                 1797.261162
Haryana                   1265.538592   Andhra Pradesh             1836.662335
Chandigarh                1279.608380   Chhattisgarh               1874.510497
```

Gender ratio (population of females / population of male)*1000

For SC population –
NCT of Delhi = 1215 has lowest gender ratio
Chhattisgarh = 1874 has highest gender ratio

5. Which state has highest gender ratio and which has the lowest for age group 0-6?

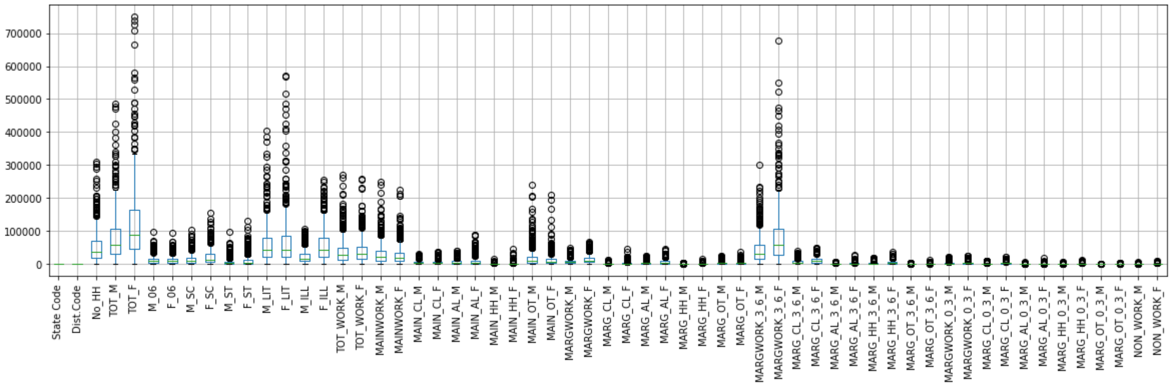


```
Haryana                   858.480597   Jharkhand                  1015.004941
Punjab                    869.658139   Dadara & Nagar Havelli     1041.811847
Chandigarh                874.544128   Arunachal Pradesh          1085.058618
```

For age group 0-6 population –
Haryana = 858 has lowest gender ratio
Arunachal Pradesh = 1085 has highest gender ratio

Part 2 - PCA:

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



There are a lot of Outliers present in the given Dataset.

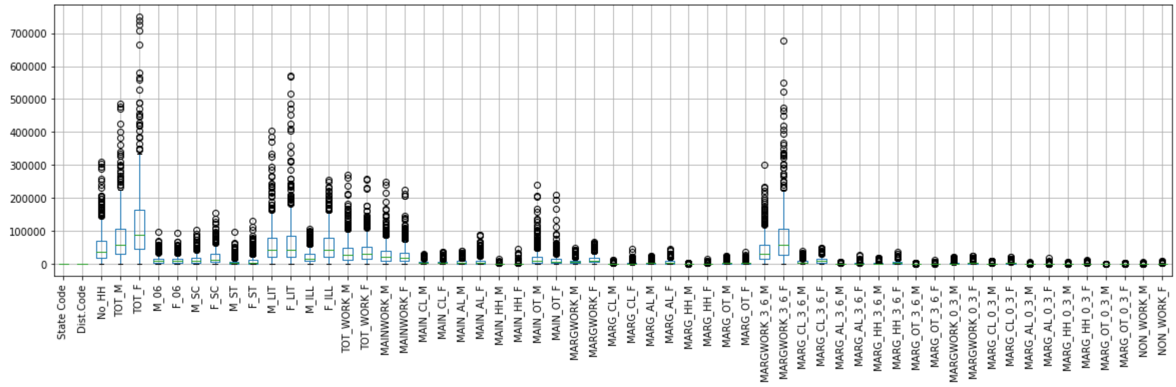
For PCA outlier treatment is important. As PCA figures out linear lines along which maximum variance is explained. If there would be outliers then direction of Principal components will be compromised.

For this case too Outlier treatment is necessary.

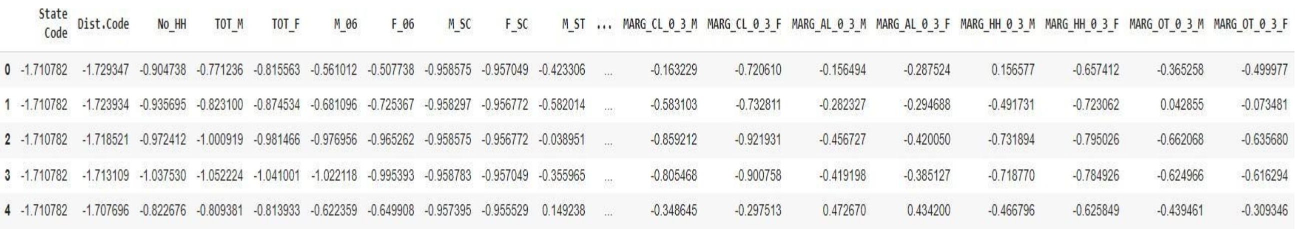
Part 2 - PCA:

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

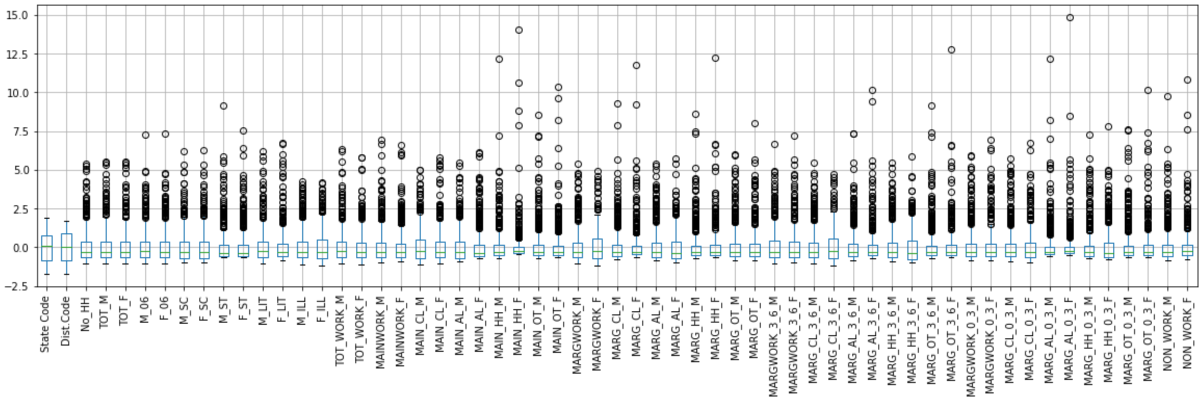
Data and outliers before scaling-



Data after applying z-score- Scaled data



	State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F
0	-1.710782	-1.729347	-0.904738	-0.771236	-0.815563	-0.561012	-0.507738	-0.958575	-0.957049	-0.423306	...	-0.163229	-0.720610	-0.156494	-0.287524	0.156577	-0.657412	-0.365258	-0.499977
1	-1.710782	-1.723934	-0.935695	-0.823100	-0.874534	-0.681096	-0.725367	-0.958297	-0.956772	-0.582014	...	-0.583103	-0.732811	-0.282327	-0.294688	-0.491731	-0.723062	0.042855	-0.073481
2	-1.710782	-1.718521	-0.972412	-1.000919	-0.981466	-0.976956	-0.965262	-0.958575	-0.956772	-0.038951	...	-0.859212	-0.921931	-0.456727	-0.420050	-0.731894	-0.795026	-0.662068	-0.635680
3	-1.710782	-1.713109	-1.037530	-1.052224	-1.041001	-1.022118	-0.995393	-0.958783	-0.957049	-0.355965	...	-0.805468	-0.900758	-0.419198	-0.385127	-0.718770	-0.784926	-0.624966	-0.616294
4	-1.710782	-1.707696	-0.822676	-0.809381	-0.813933	-0.622359	-0.649908	-0.957395	-0.955529	0.149238	...	-0.348645	-0.297513	0.472670	0.434200	-0.466796	-0.625849	-0.439461	-0.309346



There is no impact of scaling on outliers. It's just that all the features are now on same scale instead of different scales before scaling. And outliers are scaled too.

Part 2 - PCA:

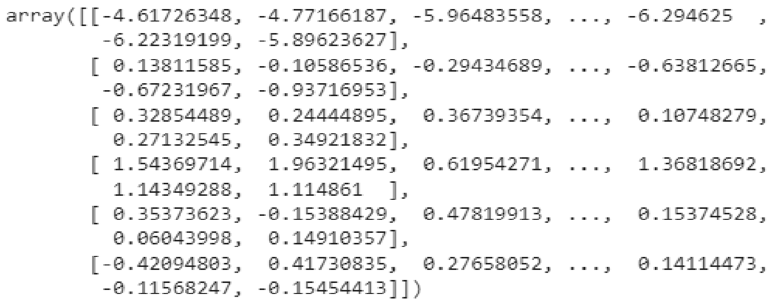
2.5- Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get Eigen values and Eigen vector.

We are dropping the columns State code and Dist. Code before proceeding for PCA. These two columns have values which should be categorical, and shall have no say in PCA.

Covariance Matrix-

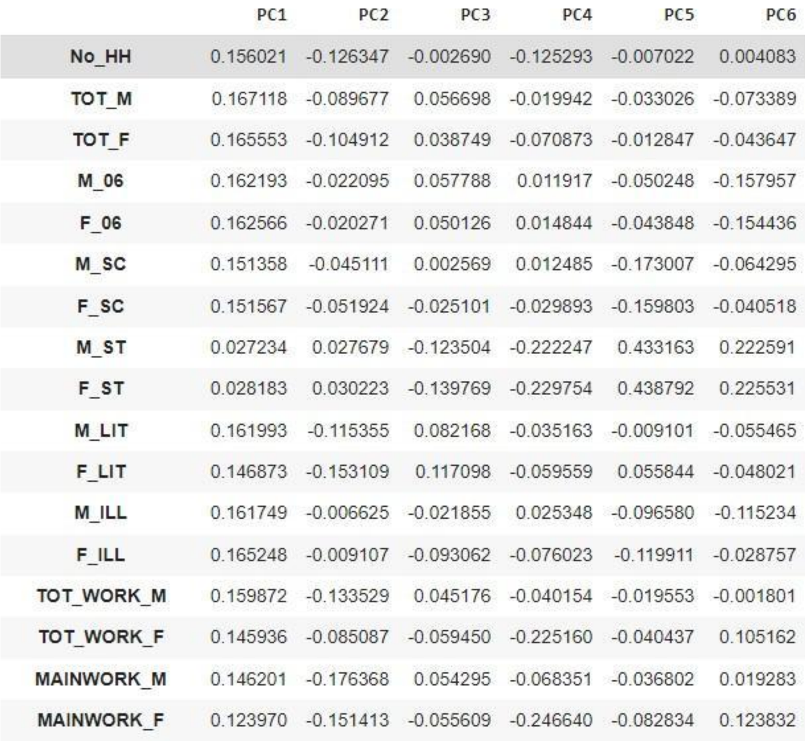
Initially we are we are generating only 6 PCA dimensions.

Covariance matrix comes out to be –

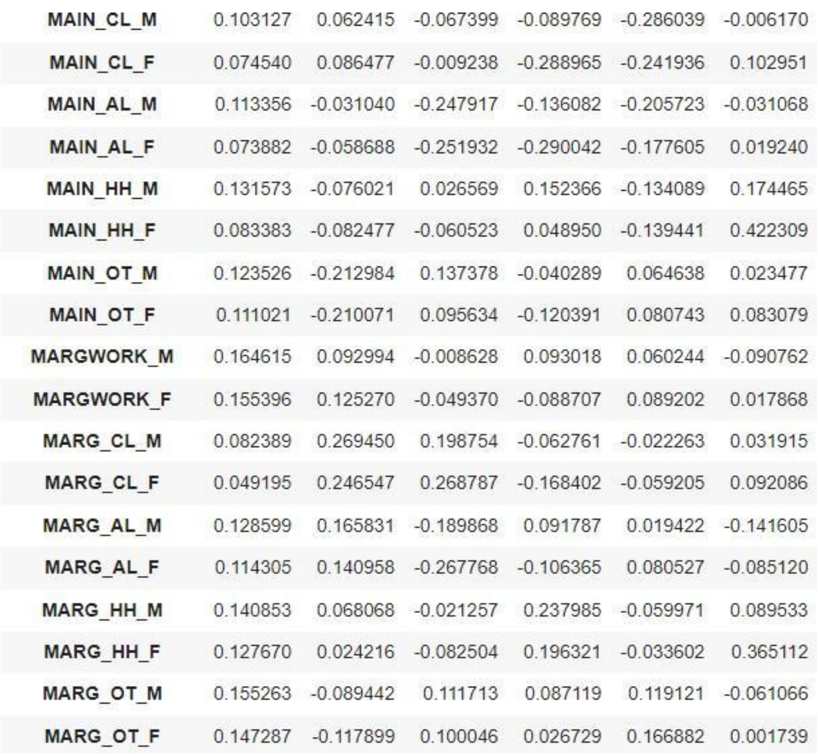


```
array([[-4.61726348, -4.77166187, -5.96483558, ..., -6.294625  ,
        -6.22319199, -5.89623627],
       [ 0.13811585, -0.10586536, -0.29434689, ..., -0.63812665,
        -0.67231967, -0.93716953],
       [ 0.32854489,  0.24444895,  0.36739354, ...,  0.10748279,
         0.27132545,  0.34921832],
       [ 1.54369714,  1.96321495,  0.61954271, ...,  1.36818692,
         1.14349288,  1.114861  ],
       [ 0.35373623, -0.15388429,  0.47819913, ...,  0.15374528,
         0.06043998,  0.14910357],
       [-0.42094803,  0.41730835,  0.27658052, ...,  0.14114473,
        -0.11568247, -0.15454413]])
```

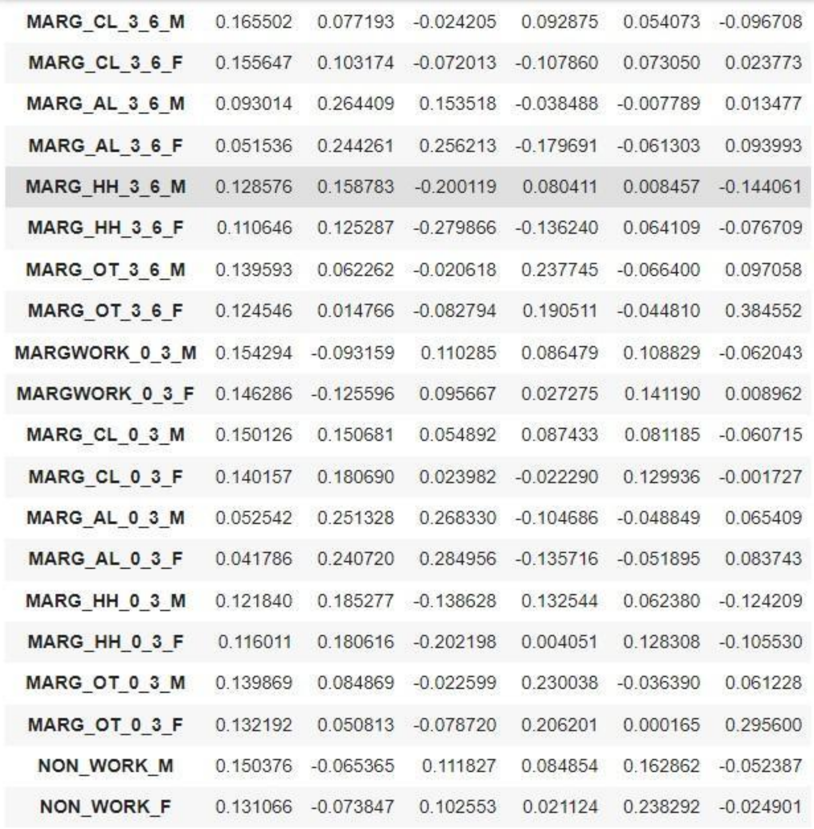
Let’s check out the Eigen vectors for these generated PC’s for each feature. We have generated a Data-frame of Eigen vectors showing PC’s and variance explained by each features in individual PC’s.



	PC1	PC2	PC3	PC4	PC5	PC6
No_HH	0.156021	-0.126347	-0.002690	-0.125293	-0.007022	0.004083
TOT_M	0.167118	-0.089677	0.056698	-0.019942	-0.033026	-0.073389
TOT_F	0.165553	-0.104912	0.038749	-0.070873	-0.012847	-0.043647
M_06	0.162193	-0.022095	0.057788	0.011917	-0.050248	-0.157957
F_06	0.162566	-0.020271	0.050126	0.014844	-0.043848	-0.154436
M_SC	0.151358	-0.045111	0.002569	0.012485	-0.173007	-0.064295
F_SC	0.151567	-0.051924	-0.025101	-0.029893	-0.159803	-0.040518
M_ST	0.027234	0.027679	-0.123504	-0.222247	0.433163	0.222591
F_ST	0.028183	0.030223	-0.139769	-0.229754	0.438792	0.225531
M_LIT	0.161993	-0.115355	0.082168	-0.035163	-0.009101	-0.055465
F_LIT	0.146873	-0.153109	0.117098	-0.059559	0.055844	-0.048021
M_ILL	0.161749	-0.006625	-0.021855	0.025348	-0.096580	-0.115234
F_ILL	0.165248	-0.009107	-0.093062	-0.076023	-0.119911	-0.028757
TOT_WORK_M	0.159872	-0.133529	0.045176	-0.040154	-0.019553	-0.001801
TOT_WORK_F	0.145936	-0.085087	-0.059450	-0.225160	-0.040437	0.105162
MAINWORK_M	0.146201	-0.176368	0.054295	-0.068351	-0.036802	0.019283
MAINWORK_F	0.123970	-0.151413	-0.055609	-0.246640	-0.082834	0.123832



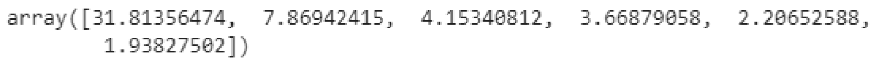
MAIN_CL_M	0.103127	0.062415	-0.067399	-0.089769	-0.286039	-0.006170
MAIN_CL_F	0.074540	0.086477	-0.009238	-0.288965	-0.241936	0.102951
MAIN_AL_M	0.113356	-0.031040	-0.247917	-0.136082	-0.205723	-0.031068
MAIN_AL_F	0.073882	-0.058688	-0.251932	-0.290042	-0.177605	0.019240
MAIN_HH_M	0.131573	-0.076021	0.026569	0.152366	-0.134089	0.174465
MAIN_HH_F	0.083383	-0.082477	-0.060523	0.048950	-0.139441	0.422309
MAIN_OT_M	0.123526	-0.212984	0.137378	-0.040289	0.064638	0.023477
MAIN_OT_F	0.111021	-0.210071	0.095634	-0.120391	0.080743	0.083079
MARGWORK_M	0.164615	0.092994	-0.008628	0.093018	0.060244	-0.090762
MARGWORK_F	0.155396	0.125270	-0.049370	-0.088707	0.089202	0.017868
MARG_CL_M	0.082389	0.269450	0.198754	-0.062761	-0.022263	0.031915
MARG_CL_F	0.049195	0.246547	0.268787	-0.168402	-0.059205	0.092086
MARG_AL_M	0.128599	0.165831	-0.189868	0.091787	0.019422	-0.141605
MARG_AL_F	0.114305	0.140958	-0.267768	-0.106365	0.080527	-0.085120
MARG_HH_M	0.140853	0.068068	-0.021257	0.237985	-0.059971	0.089533
MARG_HH_F	0.127670	0.024216	-0.082504	0.196321	-0.033602	0.365112
MARG_OT_M	0.155263	-0.089442	0.111713	0.087119	0.119121	-0.061066
MARG_OT_F	0.147287	-0.117899	0.100046	0.026729	0.166882	0.001739



MARG_CL_3_6_M	0.165502	0.077193	-0.024205	0.092875	0.054073	-0.096708
MARG_CL_3_6_F	0.155647	0.103174	-0.072013	-0.107860	0.073050	0.023773
MARG_AL_3_6_M	0.093014	0.264409	0.153518	-0.038488	-0.007789	0.013477
MARG_AL_3_6_F	0.051536	0.244261	0.256213	-0.179691	-0.061303	0.093993
MARG_HH_3_6_M	0.128576	0.158783	-0.200119	0.080411	0.008457	-0.144061
MARG_HH_3_6_F	0.110646	0.125287	-0.279866	-0.136240	0.064109	-0.076709
MARG_OT_3_6_M	0.139593	0.062262	-0.020618	0.237745	-0.066400	0.097058
MARG_OT_3_6_F	0.124546	0.014766	-0.082794	0.190511	-0.044810	0.384552
MARGWORK_0_3_M	0.154294	-0.093159	0.110285	0.086479	0.108829	-0.062043
MARGWORK_0_3_F	0.146286	-0.125596	0.095667	0.027275	0.141190	0.008962
MARG_CL_0_3_M	0.150126	0.150681	0.054892	0.087433	0.081185	-0.060715
MARG_CL_0_3_F	0.140157	0.180690	0.023982	-0.022290	0.129936	-0.001727
MARG_AL_0_3_M	0.052542	0.251328	0.268330	-0.104686	-0.048849	0.065409
MARG_AL_0_3_F	0.041786	0.240720	0.284956	-0.135716	-0.051895	0.083743
MARG_HH_0_3_M	0.121840	0.185277	-0.138628	0.132544	0.062380	-0.124209
MARG_HH_0_3_F	0.116011	0.180616	-0.202198	0.004051	0.128308	-0.105530
MARG_OT_0_3_M	0.139869	0.084869	-0.022599	0.230038	-0.036390	0.061228
MARG_OT_0_3_F	0.132192	0.050813	-0.078720	0.206201	0.000165	0.295600
NON_WORK_M	0.150376	-0.065365	0.111827	0.084854	0.162862	-0.052387
NON_WORK_F	0.131066	-0.073847	0.102553	0.021124	0.238292	-0.024901

Let’s check out the Eigen values for individual PC’s.

Eigen values explain the total amount of variance that can be explained by a given principal component.



```
array([31.81356474,  7.86942415,  4.15340812,  3.66879058,  2.20652588,
        1.93827502])
```

Maximum variance is explained by PC1 = 31.81.

PC2 explains 7.86

PC3 explains 4.153

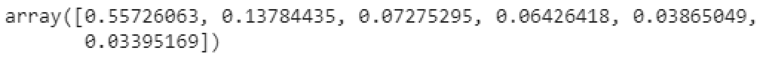
PC4 explains 3.66

PC5 explains 2.20 and

PC6 explains 1.93.

Let’s check out the percentage of variance explained by each PC that is variance explained by an individual principle component divided by total variance explained by all the PC’s.

In other words – Percentage of explained variance= Eigen value of each PC/sum of Eigen values of all PCs

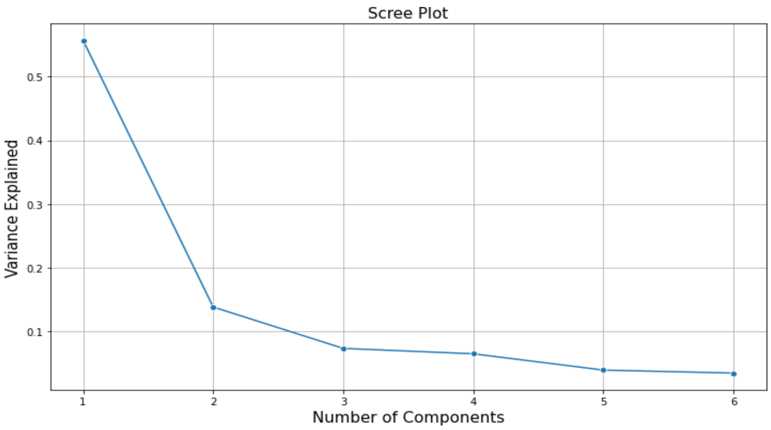


```
array([0.55726063, 0.13784435, 0.07275295, 0.06426418, 0.03865049,
       0.03395169])
```

55% of total variance is explained by PC1. 13.7% of total variance is explained by PC2.

7.2% of total variance is explained by PC3. 6.4% of total variance is explained by PC4.

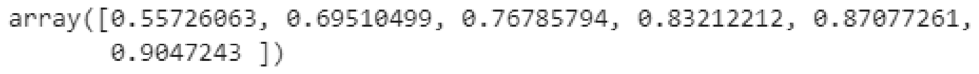
3.8% of total variance is explained by PC5. 3.4% of total variance is explained by PC6.



Part 2 - PCA:

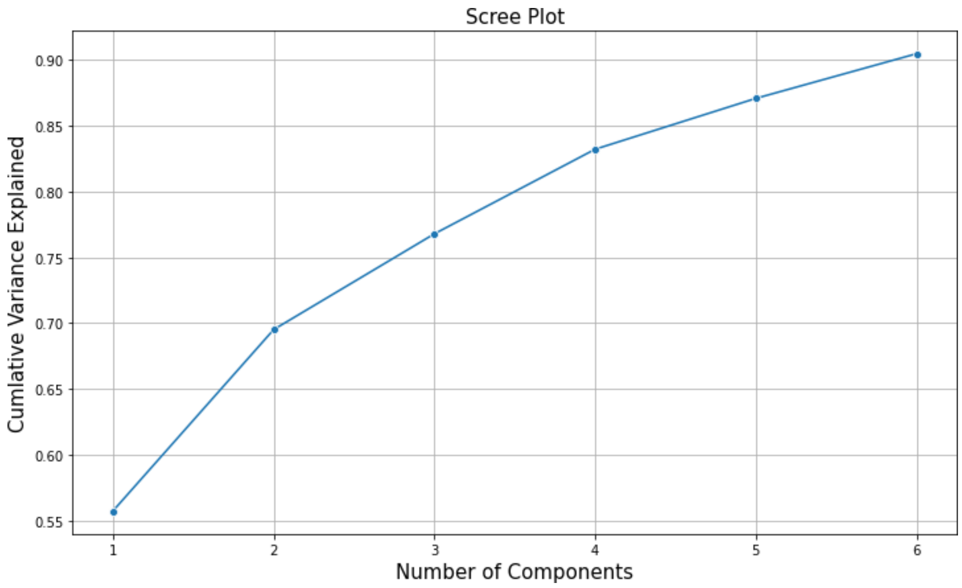
2.6- Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.

Cumulative sum of variance explained by all the PC’s



```
array([0.55726063, 0.69510499, 0.76785794, 0.83212212, 0.87077261,
       0.9047243 ])
```

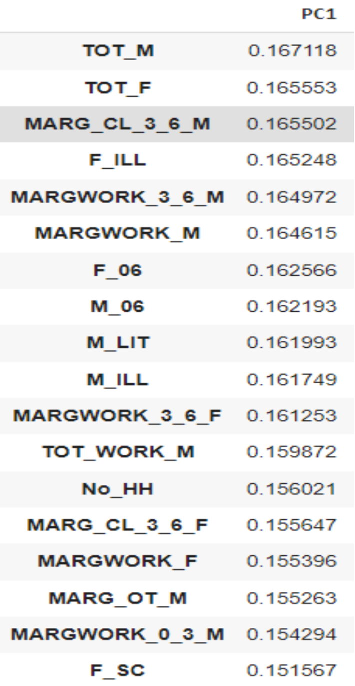
Variance explained by all the 6 PC’s is around 90.47%



Part 2 - PCA:

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

For PC1-



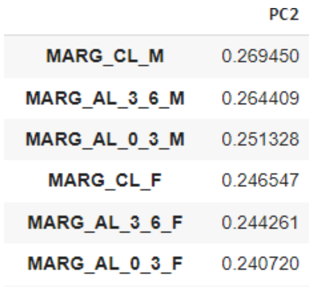
	PC1
TOT_M	0.167118
TOT_F	0.165553
MARG_CL_3_6_M	0.165502
F_ILL	0.165248
MARGWORK_3_6_M	0.164972
MARGWORK_M	0.164615
F_06	0.162566
M_06	0.162193
M_LIT	0.161993
M_ILL	0.161749
MARGWORK_3_6_F	0.161253
TOT_WORK_M	0.159872
No_HH	0.156021
MARG_CL_3_6_F	0.155647
MARGWORK_F	0.155396
MARG_OT_M	0.155263
MARGWORK_0_3_M	0.154294
F_SC	0.151567

Inferences- for PC1 these are the features that are explaining most of the variance.
Total population Male
Total population Female
Marginal Cultivator Population 3-6 Male
Illiterate Female
Marginal Worker Population 3-6 Male
Marginal Worker Population Male
Population in the age group 0-6 Female
Population in the age group 0-6 Male
Literate Male
Illiterate Male
Marginal Worker Population 3-6 Female
Total Worker Population Male

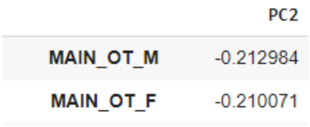
PC1 could be named as -TOT_ILL/MARGWORK

(Total illetrate population and Total Worker population)

For PC2-



	PC2
MARG_CL_M	0.269450
MARG_AL_3_6_M	0.264409
MARG_AL_0_3_M	0.251328
MARG_CL_F	0.246547
MARG_AL_3_6_F	0.244261
MARG_AL_0_3_F	0.240720

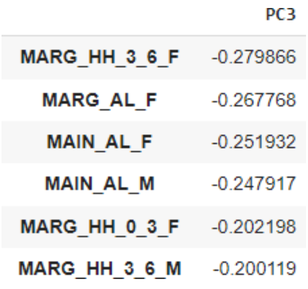


	PC2
MAIN_OT_M	-0.212984
MAIN_OT_F	-0.210071

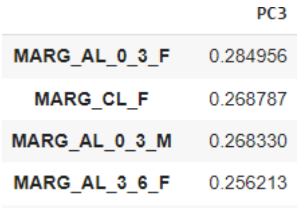
Inferences- for PC2 these are the features that are explaining most of the variance.
Marginal Cultivator Population Male = 0.26
Marginal Agriculture Laborers Population 3-6 Male =0.26
Marginal Agriculture Laborers Population 0-3 Male = 0.25
Marginal Cultivator Population Female = 0.24
Marginal Agriculture Laborers Population 3-6 Female =0.24
Marginal Agriculture Laborers Population 0-3 Female = 0.24

This PC2 can be named as MARG CL/AL 0-6
(Marginal Cultivator Population Female/Male/ Marginal Agriculture Laborers Population 3-6 Female/Male)

For PC3-



	PC3
MARG_HH_3_6_F	-0.279866
MARG_AL_F	-0.267768
MAIN_AL_F	-0.251932
MAIN_AL_M	-0.247917
MARG_HH_0_3_F	-0.202198
MARG_HH_3_6_M	-0.200119



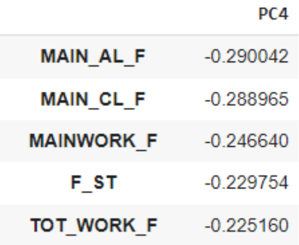
	PC3
MARG_AL_0_3_F	0.284956
MARG_CL_F	0.268787
MARG_AL_0_3_M	0.268330
MARG_AL_3_6_F	0.256213

Inferences- for PC3 these are the features that are explaining most of the variance.
Marginal Agriculture Labourers Population 0-3 Female
Marginal Household Industries Population 3-6 Female
Marginal Cultivator Population Female
Marginal Agriculture Labourers Population 0-3 Male
Marginal Agriculture Labourers Population 0-3 Female
Marginal Agriculture Labourers Population Female
Marginal Agriculture Labourers Population Male
Marginal Household Industries Population 0-3 Female
Marginal Household Industries Population 3-6 Male

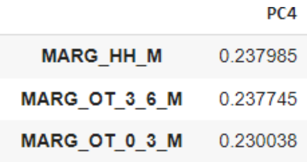
This PC3 can be named as MARG_AL_HH_0-6

(Marginal Agriculture Labourers population Female/Male and Marginal Household Industries Population 0-6 Male/Female)

For PC4



	PC4
MAIN_AL_F	-0.290042
MAIN_CL_F	-0.288965
MAINWORK_F	-0.246640
F_ST	-0.229754
TOT_WORK_F	-0.225160



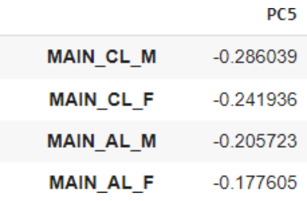
	PC4
MARG_HH_M	0.237985
MARG_OT_3_6_M	0.237745
MARG_OT_0_3_M	0.230038

Inferences- for PC4 these are the features that are explaining most of the variance.
Main Agricultural Labourers Population Female
Main Cultivator Population Female
Marginal Worker Population Female
Marginal Household Industries Population Male
Marginal Other Workers Population 0-6 Male
Total Worker Population Female

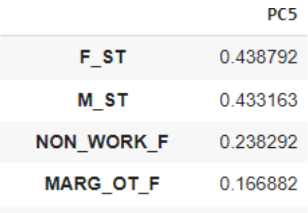
This PC4 can be named as -MARG_TOT_WORK_F

(Main and Marginal and total worker population female)

For PC5



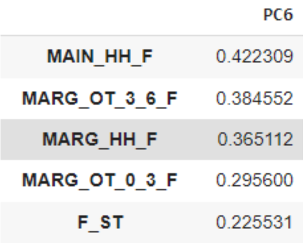
	PC5
MAIN_CL_M	-0.286039
MAIN_CL_F	-0.241936
MAIN_AL_M	-0.205723
MAIN_AL_F	-0.177605



	PC5
F_ST	0.438792
M_ST	0.433163
NON_WORK_F	0.238292
MARG_OT_F	0.166882

Inferences- for PC5 these are the features that are explaining most of the variance.
Scheduled Tribes population Female
Scheduled Tribes population Male
Main Cultivator Population Male
Main Cultivator Population Female
This PC5 can be named as - SC_TOT_MAIN_CL
(SC male and female / main cultivator Male and female)

For PC6-



	PC6
MAIN_HH_F	0.422309
MARG_OT_3_6_F	0.384552
MARG_HH_F	0.365112
MARG_OT_0_3_F	0.295600
F_ST	0.225531

Inferences- for PC6 these are the features that are explaining most of the variance.

Main Household Industries Population Female
Marginal Other Workers Population Person 3-6 Female
Marginal Household Industries Population Female
Marginal Other Workers Population Person 0-3 Female

This PC6 can be named as -MAIN_MARG_HH_MARG_OT_0-6_F

(Main and Marginal Industries, Marginal other worker population 0-6 female)

Part 2 - PCA:

2.8- Write linear equation for first PC.

Equation for PC1 is as follows-

Equation= 'TOT_F' * 0.17 + 'MARG_CL_3_6_M' * 0.17 + 'TOT_M' * 0.17 + 'F_ILL' * 0.17 + 'No_HH' * 0.16 + 'M_ILL' * 0.16 + 'MARG_CL_3_6_F' * 0.16 + 'MARGWORK_3_6_F' * 0.16 + 'MARGWORK_3_6_M' * 0.16 + 'MARG_OT_M' * 0.16 + 'MARGWORK_F' * 0.16 + 'TOT_WORK_M' * 0.16 + 'M_LIT' * 0.16 + 'F_06' * 0.16 + 'M_06' * 0.16 + 'MARGWORK_M' * 0.16 + 'F_LIT' * 0.15 + 'TOT_WORK_F' * 0.15 + 'MAINWORK_M' * 0.15 + 'MARG_OT_F' * 0.15 + 'NON_WORK_M' * 0.15 + 'MARG_CL_0_3_M' * 0.15 + 'MARGWORK_0_3_F' * 0.15 + 'MARGWORK_0_3_M' * 0.15 + 'F_SC' * 0.15 + 'M_SC' * 0.15 + 'MARG_OT_0_3_M' * 0.14 + 'MARG_OT_3_6_M' * 0.14 + 'MARG_CL_0_3_F' * 0.14 + 'MARG_HH_M' * 0.14 + 'MARG_HH_3_6_M' * 0.13 + 'MARG_OT_0_3_F' * 0.13 + 'NON_WORK_F' * 0.13 + 'MARG_HH_F' * 0.13 + 'MARG_AL_M' * 0.13 + 'MAIN_HH_M' * 0.13 + 'MARG_HH_0_3_M' * 0.12 + 'MAIN_OT_M' * 0.12 + 'MAINWORK_F' * 0.12 + 'MARG_OT_3_6_F' * 0.12 + 'MARG_HH_0_3_F' * 0.12 + 'MAIN_OT_F' * 0.11 + 'MARG_HH_3_6_F' * 0.11 + 'MARG_AL_F' * 0.11 + 'MAIN_AL_M' * 0.11 + 'MAIN_CL_M' * 0.1 + 'MARG_AL_3_6_M' * 0.09 + 'MARG_CL_M' * 0.08 + 'MAIN_HH_F' * 0.08 + 'MAIN_AL_F' * 0.07 + 'MAIN_CL_F' * 0.07 + 'MARG_AL_0_3_M' * 0.05 + 'MARG_AL_3_6_F' * 0.05 + 'MARG_CL_F' * 0.05 + 'MARG_AL_0_3_F' * 0.04 + 'F_ST' * 0.03 + 'M_ST' * 0.03

END