Principal Component Analysis
Indian Census Data 2011

Team:
Akhilesh Chauhan
Apoorv Purohit
Preyal Deep Chhabra
Rathi Sadhasivan
Renuka Prasad
Yureka M R
PGP-AIML June'23
18/11/2023

Table of Contents

Executive Summary

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. In 1881 a Census was taken for the entire country simultaneously. Since then, the 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 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.

Introduction

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.
This exercise should help the learners in exploring the data, understanding the variances in data, scaling and transforming the data and a very important unsupervised learning technique used in dimensionality reduction namely Principal Component Analysis or simply PCA.

Data Description

There are **640 records** and **61 features.**

List of columns:-

```
State Code          int64
Dist.Code           int64
State              object
Area Name          object
No_HH               int64
```

```
TOT_M              int64
TOT_F              int64
M_06               int64
F_06               int64
M_SC               int64
F_SC               int64
M_ST               int64
F_ST               int64
M_LIT              int64
F_LIT              int64
M_ILL              int64
F_ILL              int64
TOT_WORK_M         int64
TOT_WORK_F         int64
MAINWORK_M         int64
MAINWORK_F         int64
MAIN_CL_M          int64
MAIN_CL_F          int64
MAIN_AL_M          int64
MAIN_AL_F          int64
MAIN_HH_M          int64
MAIN_HH_F          int64
MAIN_OT_M          int64
MAIN_OT_F          int64
MARGWORK_M         int64
MARGWORK_F         int64
MARG_CL_M          int64
MARG_CL_F          int64
MARG_AL_M          int64
MARG_AL_F          int64
MARG_HH_M          int64
MARG_HH_F          int64
MARG_OT_M          int64
MARG_OT_F          int64
MARGWORK_3_6_M     int64
MARGWORK_3_6_F     int64
MARG_CL_3_6_M      int64
MARG_CL_3_6_F      int64
MARG_AL_3_6_M      int64
MARG_AL_3_6_F      int64
MARG_HH_3_6_M      int64
MARG_HH_3_6_F      int64
MARG_OT_3_6_M      int64
MARG_OT_3_6_F      int64
MARGWORK_0_3_M     int64
```

```
MARGWORK_0_3_F      int64
MARG_CL_0_3_M       int64
MARG_CL_0_3_F       int64
MARG_AL_0_3_M       int64
MARG_AL_0_3_F       int64
MARG_HH_0_3_M       int64
MARG_HH_0_3_F       int64
MARG_OT_0_3_M       int64
MARG_OT_0_3_F       int64
NON_WORK_M          int64
NON_WORK_F          int64
```

Out of 61 features, there are **59 INT** and **2 OBJECT** features.

The features `State` and `Area Name` are Object types and categorical in nature. Out of all the INT features, `State Code` and `Dist.Code` are also categorical in nature. So we will leave them out of our PCA calculations.

Sample of the Dataset

	State Code	Dist.Code	State	Area Name	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
294	17	295	Meghalaya	South Garo Hills	2601	5226	7273	875	885	8	7	4929	6909	3508	4272	1718	3001
396	21	397	Odisha	Nabarangapur	29737	30751	59897	5090	5282	5298	10229	15707	30721	17251	17015	13500	42882
353	20	354	Jharkhand	Dhanbad	53278	106331	145768	16863	17100	21260	29255	10729	15990	74695	72982	31636	72786
463	23	464	Madhya Pradesh	Jhabua	18254	27705	45248	7579	7627	628	1013	23216	38233	12532	12829	15173	32419
624	33	625	Tamil Nadu	Virudhunagar	90241	66704	148445	7187	6645	14191	30949	53	110	55194	78018	11510	70427
251	12	252	Arunachal Pradesh	Upper Siang	929	1187	2117	203	228	0	0	1027	1865	705	999	482	1118
631	33	632	Tamil Nadu	Coimbatore	133255	125297	239223	12101	11624	21087	40362	1095	2204	104953	146669	20344	92554
577	29	578	Karnataka	Chamarajanagar	49770	62731	107238	7822	7479	16889	26901	8370	14814	41696	42205	21035	65033
274	14	275	Manipur	Bishnupur	6864	12684	19388	1861	1820	1030	1517	171	314	9636	10427	3048	8961
115	8	116	Rajasthan	Jalor	27589	43975	67312	9456	8920	8331	11680	3980	5380	27516	20828	16459	46484

Fig 1 - sample dataset page 1

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
1885	3112	1342	2333	595	1301	108	210	25	36	614
15334	30103	8280	11387	3262	2782	2063	4255	166	357	2789
42836	39967	28937	23640	1129	1230	1280	1270	583	643	25945
9964	23467	6657	16294	3529	11545	938	2118	131	145	2059
38587	76432	35177	65612	1469	3356	5008	19321	530	3437	28170
460	1192	335	990	223	540	13	36	3	27	96
78311	101548	72903	89310	2354	4396	6881	21103	1307	3487	62361
38633	42988	31221	31290	7564	4069	11739	17302	704	1285	11214
6136	8704	5010	5313	2022	1083	321	457	129	793	2538
18482	29973	14906	17529	5868	9204	2198	3525	384	385	6456

Fig 2 - sample dataset page 2

MAIN_OT_F	MARGWORK_M	MARGWORK_F	MARG_CL_M	MARG_CL_F	MARG_AL_M	MARG_AL_F	MARG_HH_M	MARG_HH_F	MARG_OT_M	MARG_OT_F
786	543	779	215	345	101	162	17	32	210	240
3993	7054	18716	828	1380	4892	14768	183	506	1151	2062
20497	13899	16327	1560	3352	2442	4493	389	715	9508	7767
2486	3307	7173	1280	3215	1421	3181	39	71	567	706
39498	3410	10820	147	367	1069	5647	94	704	2100	4102
387	125	202	81	127	10	14	4	16	30	45
60324	5408	12238	141	372	851	3679	148	792	4268	7395
8634	7412	11698	528	841	4022	7591	221	432	2641	2834
2980	1126	3391	312	691	323	1033	91	771	400	896
4415	3576	12444	1109	4032	1192	4597	85	254	1190	3561

Fig 3 - sample dataset page 3

MARGWORK_3_6_M	MARGWORK_3_6_F	MARG_CL_3_6_M	MARG_CL_3_6_F	MARG_AL_3_6_M	MARG_AL_3_6_F	MARG_HH_3_6_M	MARG_HH_3_6_F
3341	4161	431	630	163	270	84	132
15417	29794	5964	14902	724	1131	4179	11811
63495	105801	11635	12508	1219	2367	2022	3257
17741	21781	2337	5731	763	2494	1138	2663
28117	72013	2903	8806	142	327	925	4722
727	925	81	149	47	85	10	13
46986	137675	4685	10101	133	326	736	3015
24098	64250	6510	10153	499	764	3768	6906
6548	10684	862	2674	212	554	239	799
25493	37339	2661	10165	670	3268	1000	3900

Fig 4 - sample dataset page 4

MARG_OT_3_6_M	MARG_OT_3_6_F	MARGWORK_0_3_M	MARGWORK_0_3_F	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F
12	22	172	206	112	149	52	75
140	388	921	1572	1090	3814	104	249
314	543	8080	6341	2264	3819	341	985
27	47	409	527	970	1442	517	721
67	499	1769	3258	507	2014	5	40
2	14	22	37	44	53	34	42
111	567	3705	6193	723	2137	8	46
167	317	2076	2166	902	1545	29	77
72	607	339	714	264	717	100	137
64	191	927	2806	915	2279	439	764

Fig 5 - sample dataset page 5

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
17	30	5	10	38	34
713	2957	43	118	230	490
420	1236	75	172	1428	1426
283	518	12	24	158	179
144	925	27	205	331	844
0	1	2	2	8	8
115	664	37	225	563	1202
254	685	54	115	565	668
84	234	19	164	61	182
192	697	21	63	263	755

Fig 6 - sample dataset page 6

Analyzing Data

Checking for NULL Values in the Dataset

There are no NULL values in the dataset. Please check the support file for code.

Checking for Duplicate Values in the Dataset

There are no duplicate values in the dataset. Please check the support file for code.

Exploratory Data Analysis

Q. Perform detailed Exploratory analysis by creating certain questions like the given example. Pick 5 variables out of the given 20 variables below.

5 Highest Male Populated States

State	
Uttar Pradesh	9043969
Maharashtra	4196130
Bihar	4025198
West Bengal	3912553
Karnataka	3409482

Fig 7 - 5 highest male populated states

5 Lowest Male Populated States

State	
Dadara & Nagar Havelli	6982
Lakshadweep	12823
Daman & Diu	13153
Andaman & Nicobar Island	18726
Sikkim	26664

Fig 8 - 5 lowest male populated states

5 Highest Female Populated States

State	
Uttar Pradesh	12023885
Maharashtra	7138557
Andhra Pradesh	6097235
West Bengal	6016118
Tamil Nadu	5610310

Fig 9 - 5 highest female populated states

5 Lowest Female Populated States

State	
Dadara & Nagar Havelli	10831
Lakshadweep	14772
Daman & Diu	18706
Andaman & Nicobar Island	28691
Sikkim	41518

Fig 10 - 5 lowest female populated states

State-Wise Female to Male Ratio

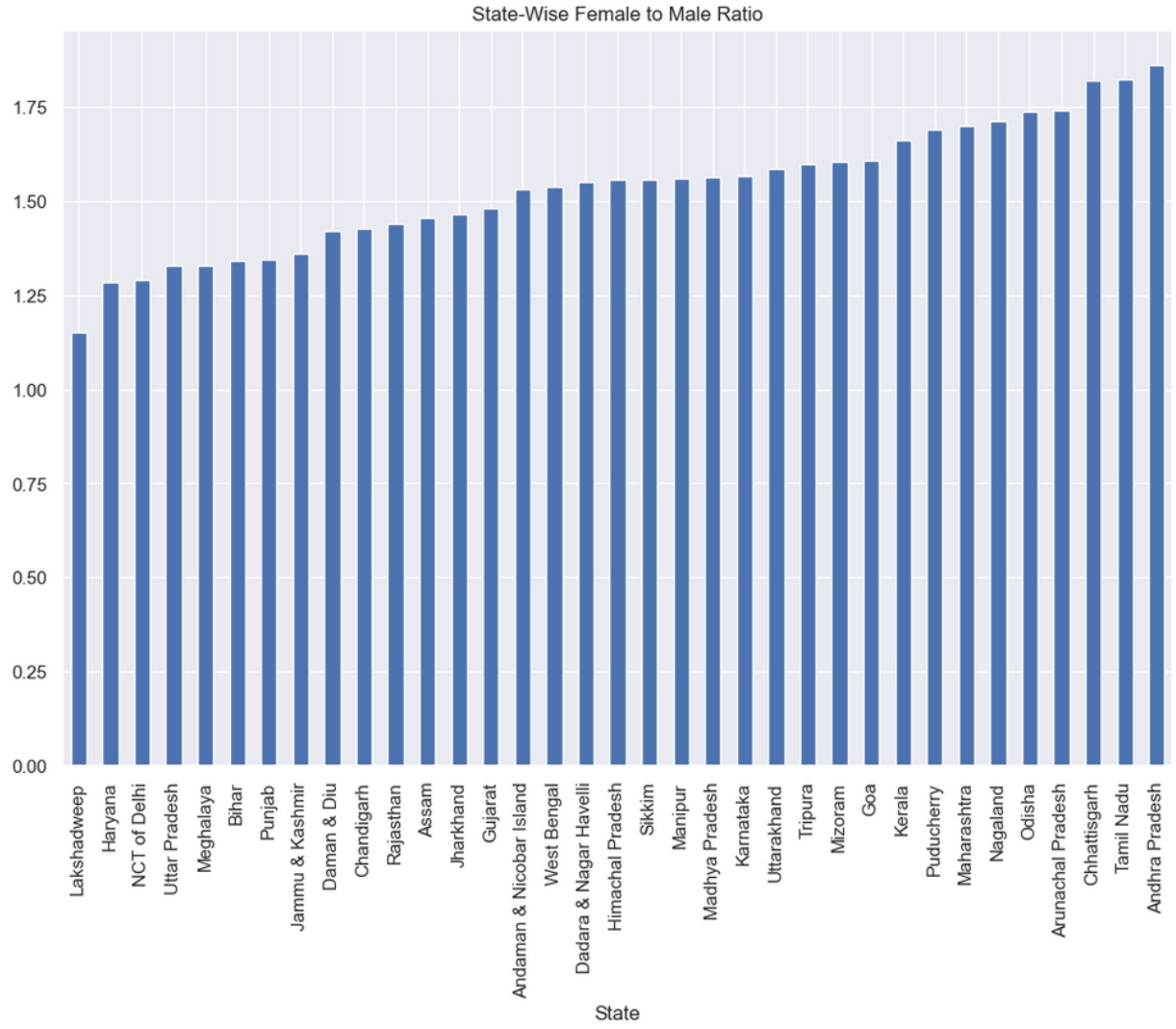

Fig 11 - state-wise female to male ratio

State-Wise Work Population Mean

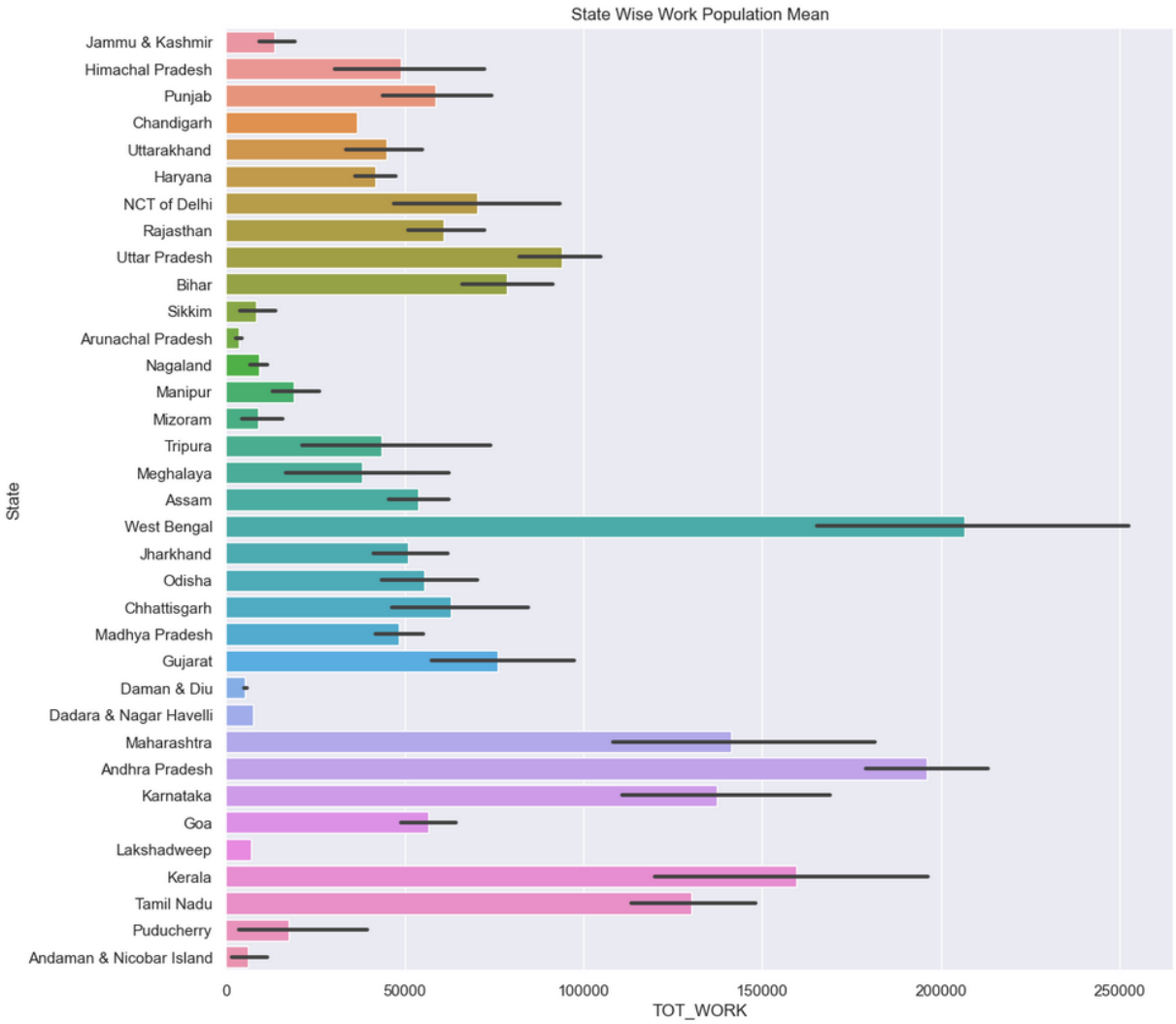

Fig 12 - state-wise population mean

Main Work Population Categories Pair Plot

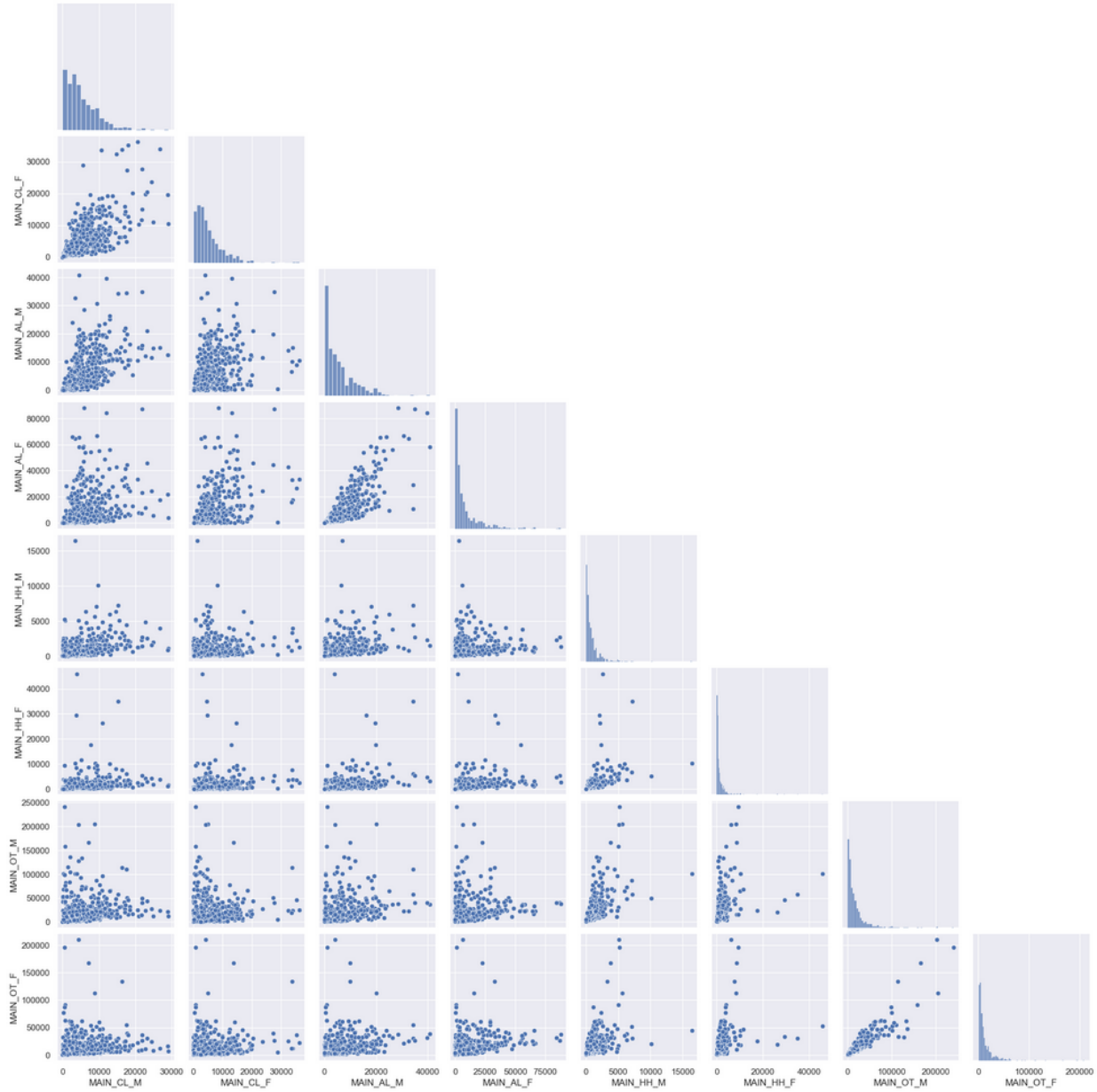

Fig 13 - pairplot for main work population categories

No_HH Univariate Plots

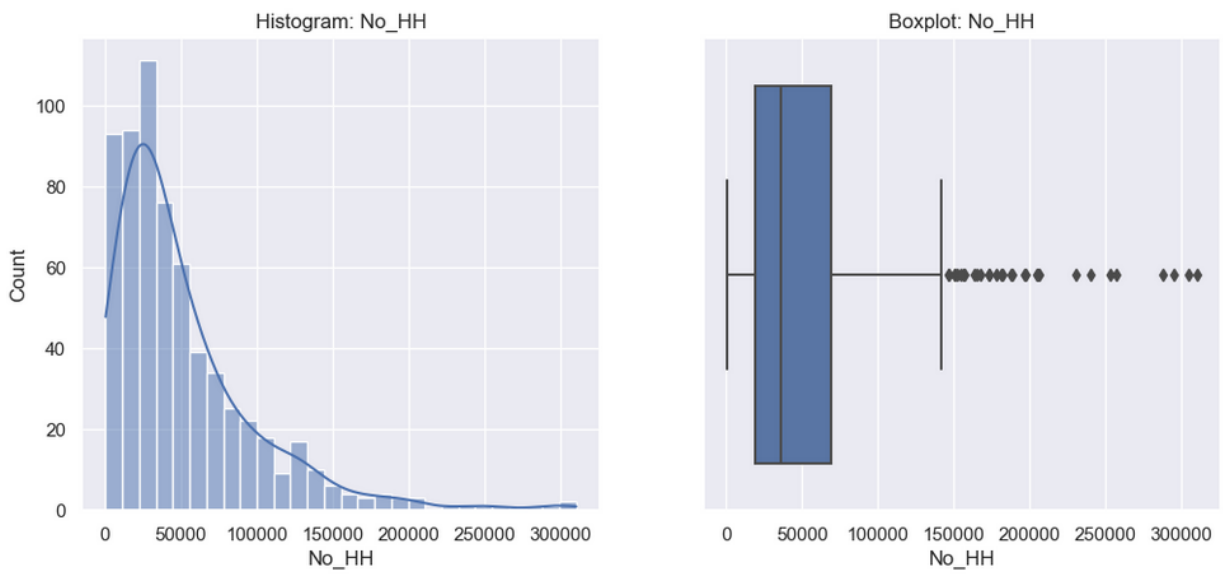

Fig 14 - histogram and boxplot for No_HH feature

Summary

- Uttar Pradesh has the highest male population, whereas Dadra & Nagar Haveli has the lowest male population.
- Similar results replicate for the Female population as well with Uttar Pradesh as the highest and Dadra & Nagar Haveli as the lowest.
- Andhra Pradesh has the highest female to male ratio and Lakshadweep has the lowest.
- West Bengal has the highest mean working population overall whereas Arunachal Pradesh has the lowest mean working population.
- Pairplot Analysis:-
    - All the data in these features is right skewed with some outliers.
    - Every Category's Male and Female population have some correlation.
    - like, Agricultural Laborers population Male and Female features seem highly correlated.
- No_HH has some outliers in the data. The data is right skewed.

Outliers

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

We are choosing not to treat outliers but outlier treatment is very important for PCA. Outliers tend to dominate in algorithms which work on squares of values and squaring an outlier which is big in magnitude shifts the result in their direction.

Scaling the Data

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

As scaling transforms the data to fit into a specific range or distribution, having outliers in the dataset impacts scaling. Since z-score scaling uses mean and standard deviation, and outliers affect mean the most by pulling the mean towards them, z-score scaling is also affected by extreme outliers. It is still a better technique to use than min-max scaling which is affected by outliers heavily.

$$\textbf{Z-score} = \frac{x - \mu}{\sigma}$$

Box plots before scaling

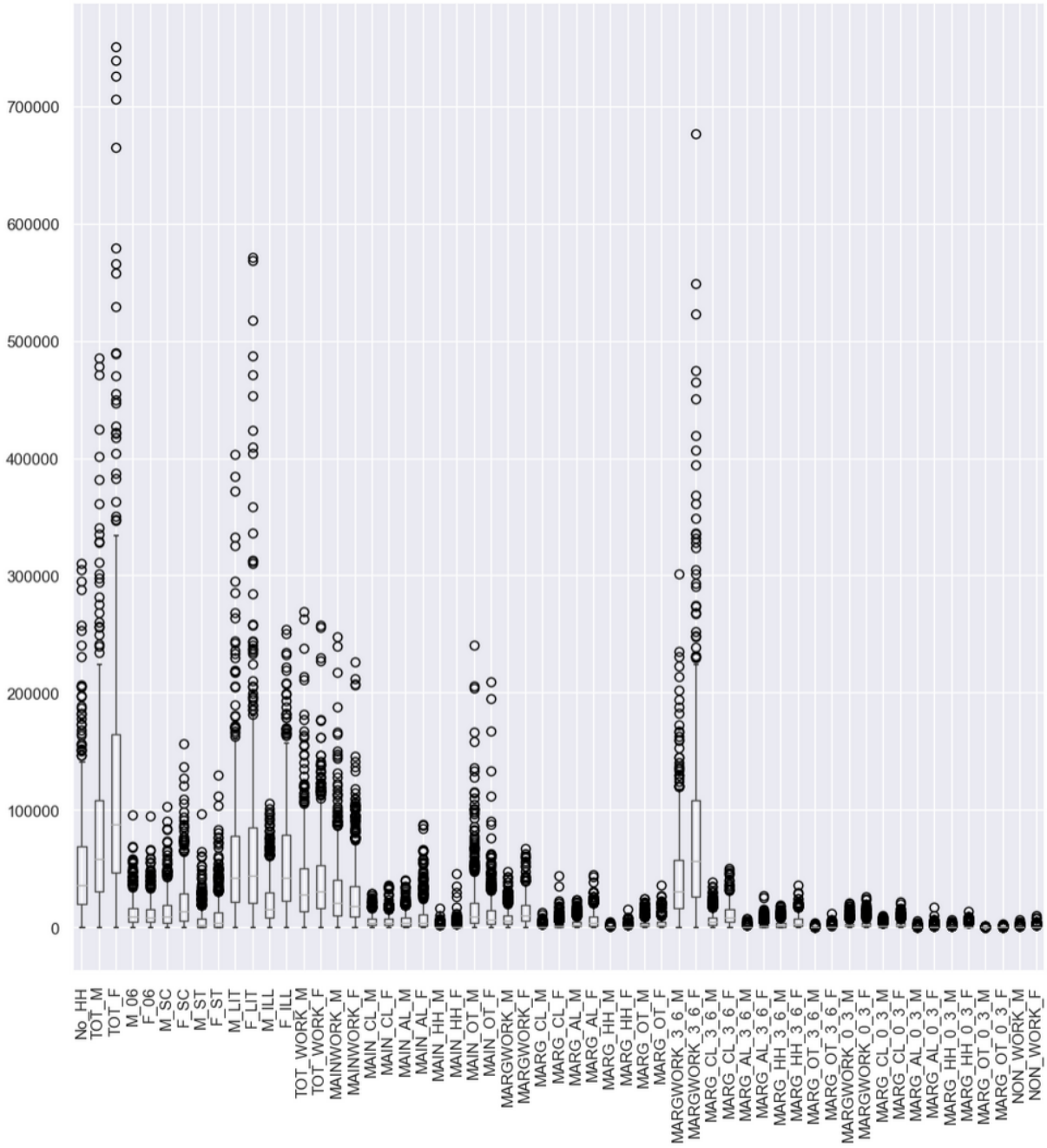

Fig 15 - box plots of numerical variables before scaling

Box plots after scaling

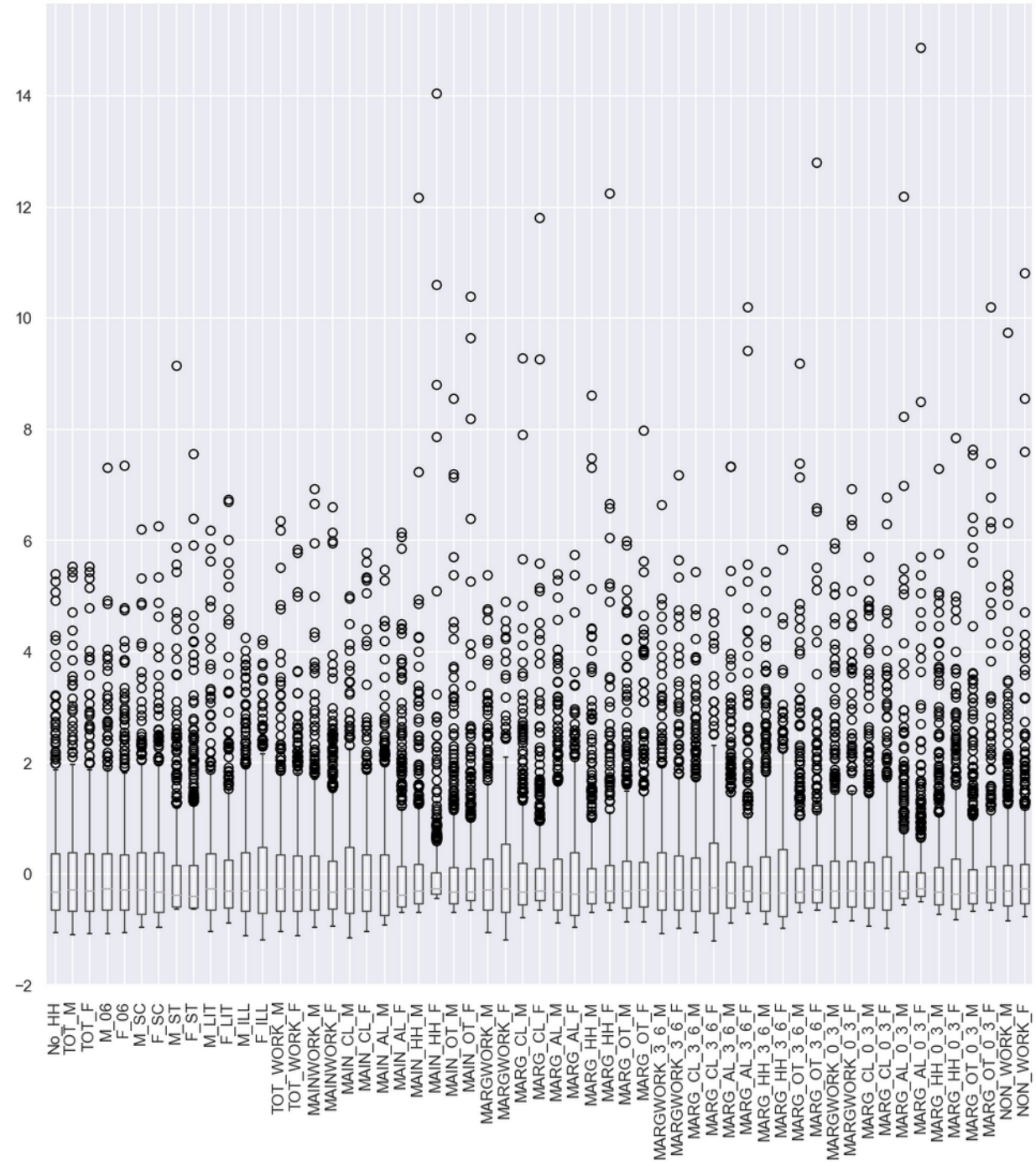

Fig 16 - box plots of numerical variables after scaling

As shown in the before and after comparison as well, the outliers are not eliminated by scaling but the data is fit for further computation now as every point is on the same scale, even the outliers.

Principal Component Analysis

We used all the 57 features in the PCA algorithm to analyze and get the ideal number of principal components for the dataset.

Initial → No of features = No of components = 57

EigenValues (or explained variance) for all 57 components

```
array([3.18135647e+01, 7.86942415e+00, 4.15340812e+00, 3.66879058e+00,
       2.20652588e+00, 1.93827502e+00, 1.17617374e+00, 7.51159086e-01,
       6.17053743e-01, 5.28300887e-01, 4.29831189e-01, 3.53440201e-01,
       2.96163013e-01, 2.81275560e-01, 1.92158325e-01, 1.36267920e-01,
       1.13389199e-01, 1.06303946e-01, 9.72885376e-02, 8.01062194e-02,
       5.76089954e-02, 4.43955966e-02, 3.78910846e-02, 2.96360194e-02,
       2.70797618e-02, 2.34458139e-02, 1.45111511e-02, 1.09852268e-02,
       9.31507853e-03, 8.13540203e-03, 7.89250253e-03, 5.02601514e-03,
       2.59771182e-03, 1.06789820e-03, 7.13559124e-04, 6.93346479e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       8.85628598e-32])
```

Fig 17 - EigenValues of 57 components

As we can see, the first component explains the maximum variance about the data ~ 31.81 and the subsequent components have decreasing values of it.

To choose the ideal number of components which can explain at least 90% of the variance in the data, we further analyzed all the components.

Correlation Heatmap of All Numerical Features

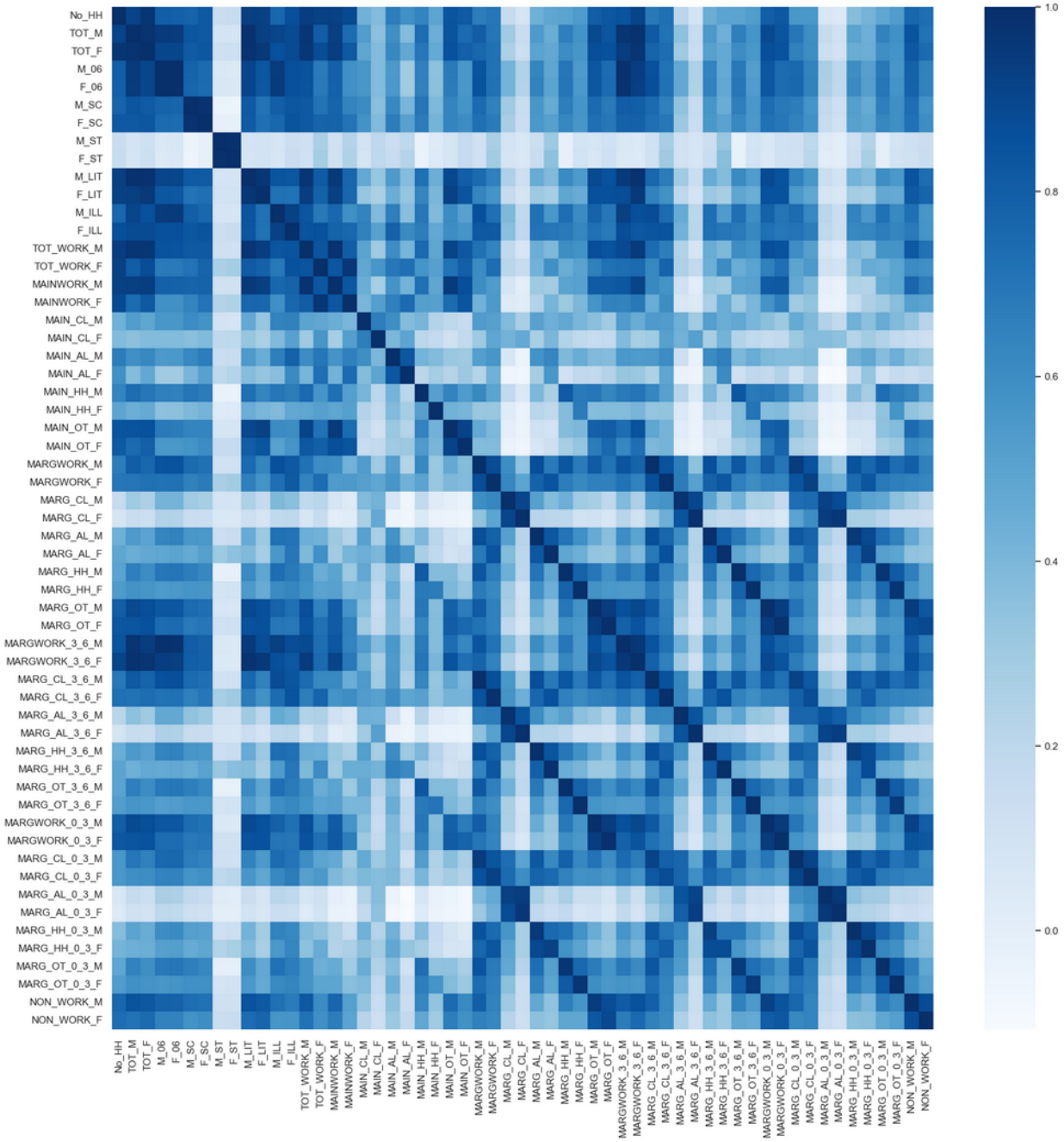

Fig 18 - Correlation Heatmap of numerical features

Cumulative Variance Plot

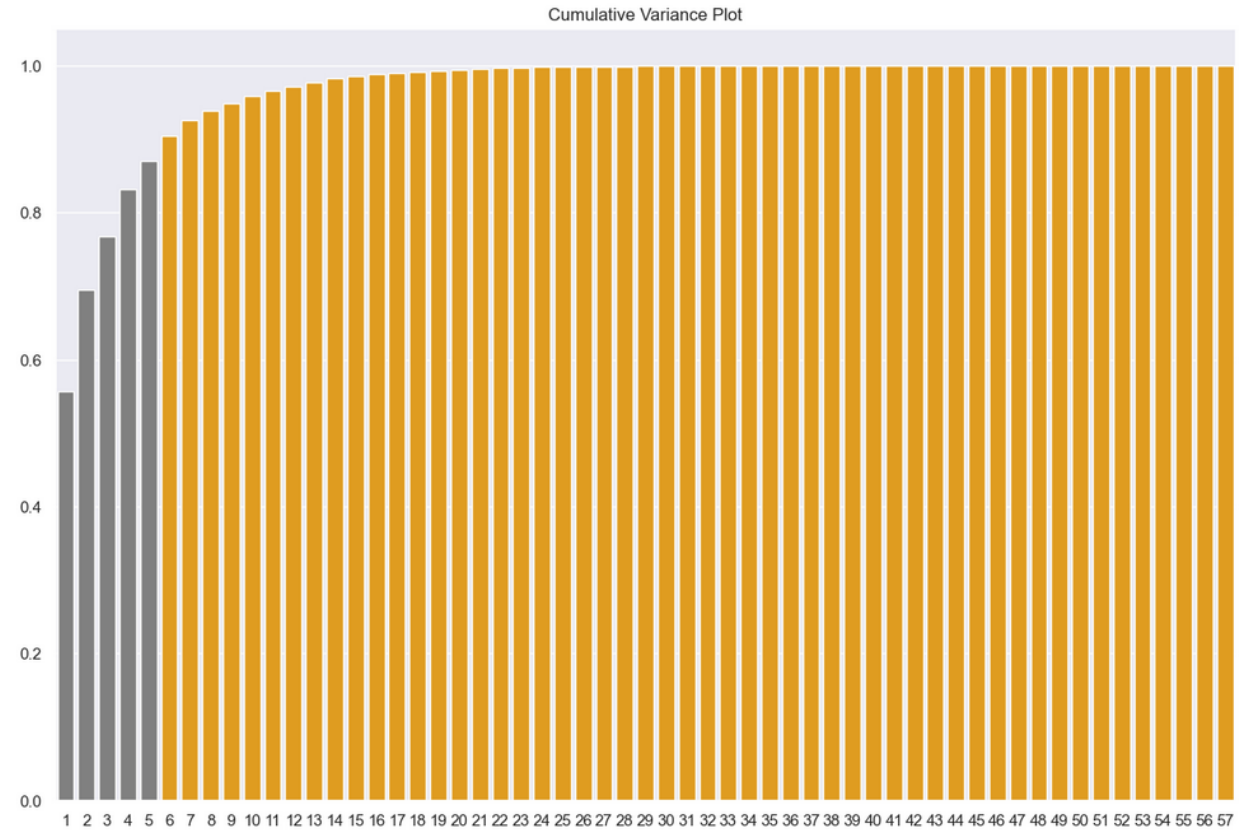

Fig 19 - Cumulative Variance Plot

As we can observe, from **Principal Component 6** onwards we can explain **at least 90% of the variance** in the data set.

Scree Plot

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

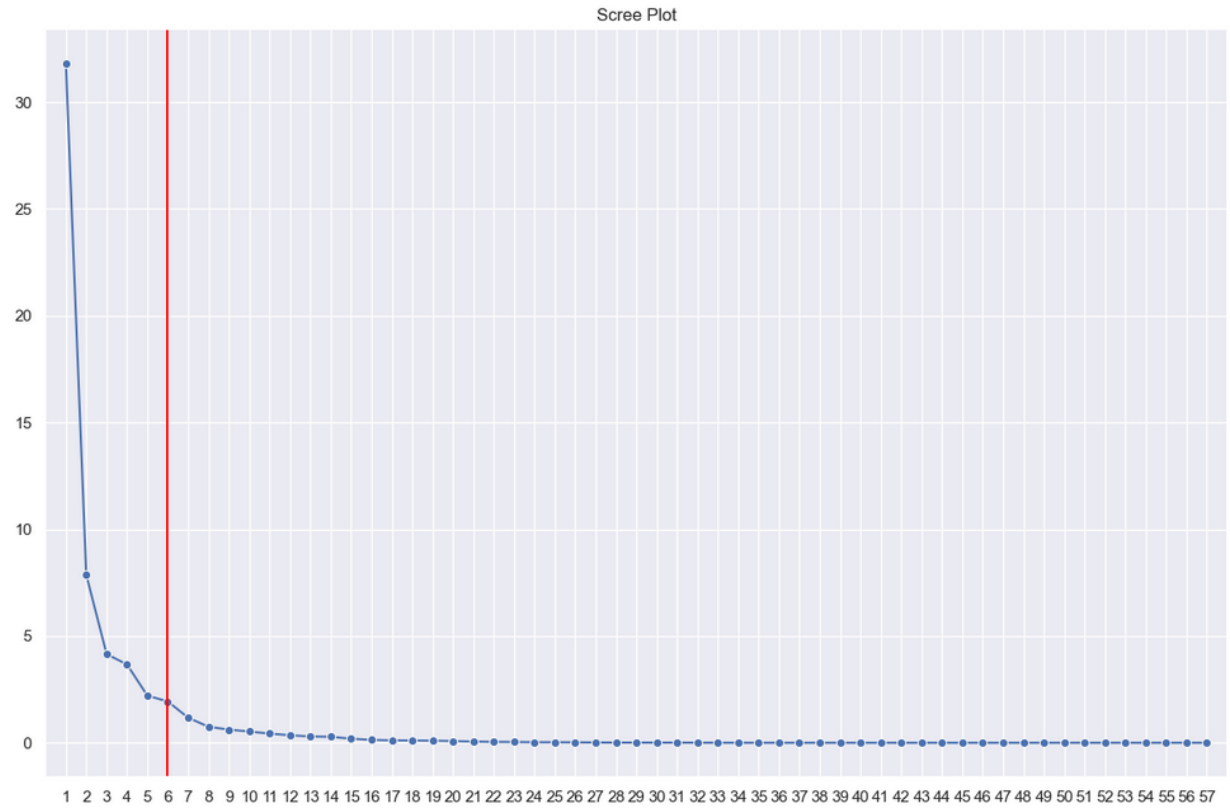

Fig 20 - Scree Plot

As we can observe from the above Scree Plot, after Principal Component 6, the distance is not significant in the subsequent principal components.

So, we will go ahead with 6 Principal Components.

Details about the Principal Components

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

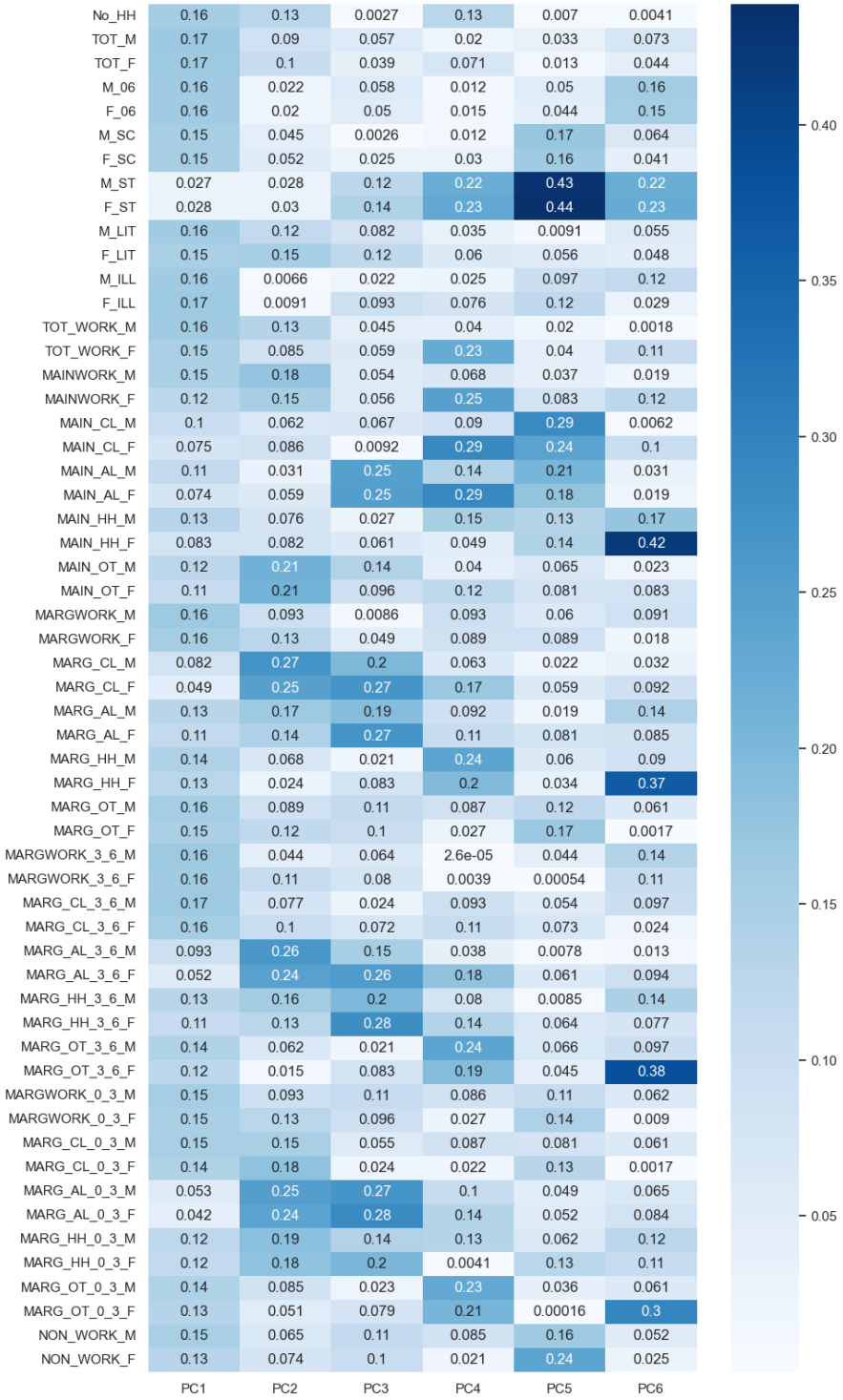

Fig 21 - Heatmap for actual features and their coefficients for each PC

- **PC1** - explains the most variance in the data 31.81. All the columns are contributing some amount to this principal component.
- **PC2** - variance is 7.86. MARG_CL_M is contributing the most to this with some other MARG_% features not far behind.
- **PC3** - variance is 4.15. Some other MARG_% columns are major contributors to the variance of this component like, MARG_AL_0_3_F, MARG_HH_3_6_F, etc.
- **PC4** - variance is 3.66. As we are moving towards more components, the number of contributors to the variance is less and are dominated by 1 or 2 features. Like, this one is dominated by MAIN_AL_F and MAIN_CL_F mostly.
- **PC5** - variance is 2.20. F_ST and M_ST are major contributors to this component.
- **PC6** - variance is 1.93. MAIN_HH_F is the major contributor to this component.

Equation of PC1

Q. Write the linear equation for the first PC.

```
[0.15602057858567925 x No_HH + 0.16711763488533515 x TOT_M +
0.16555317909064896 x TOT_F + 0.16219294820465543 x M_06 +
0.1625663956573483 x F_06 + 0.15135784909060582 x M_SC +
0.1515665001920887 x F_SC + 0.027234194571004233 x M_ST +
0.028183315015872696 x F_ST + 0.1619928373362916 x M_LIT +
0.14687268030140294 x F_LIT + 0.1617494446347163 x M_ILL +
0.1652481873683337 x F_ILL + 0.15987198816201292 x TOT_WORK_M +
0.14593580377247625 x TOT_WORK_F + 0.14620072976305987 x MAINWORK_M +
0.12397028357273655 x MAINWORK_F + 0.10312715883019864 x MAIN_CL_M +
0.0745397855548368 x MAIN_CL_F + 0.11335571218156727 x MAIN_AL_M +
0.07388215903155891 x MAIN_AL_F + 0.13157258402275596 x MAIN_HH_M +
0.08338263967435766 x MAIN_HH_F + 0.12352624192253081 x MAIN_OT_M +
0.11102126391320132 x MAIN_OT_F + 0.1646154785601101 x MARGWORK_M +
0.15539561810834127 x MARGWORK_F + 0.08238854140704541 x MARG_CL_M +
0.04919539567873822 x MARG_CL_F + 0.12859856294668556 x MARG_AL_M +
0.11430507278919892 x MARG_AL_F + 0.14085322696185132 x MARG_HH_M +
0.1276695980147536 x MARG_HH_F + 0.155262871623116 x MARG_OT_M +
0.14728658356523394 x MARG_OT_F + 0.16497194993714456 x
MARGWORK_3_6_M + 0.16125343257531358 x MARGWORK_3_6_F +
0.1655016110258062 x MARG_CL_3_6_M + 0.1556470491448339 x
MARG_CL_3_6_F + 0.09301420640192848 x MARG_AL_3_6_M +
0.0515358639701522 x MARG_AL_3_6_F + 0.12857611642867817 x
MARG_HH_3_6_M + 0.11064584323696926 x MARG_HH_3_6_F +
0.13959276252158825 x MARG_OT_3_6_M + 0.12454590917258752 x
MARG_OT_3_6_F + 0.15429378578916028 x MARGWORK_0_3_M +
0.14628565406214417 x MARGWORK_0_3_F + 0.15012570610262055 x
MARG_CL_0_3_M + 0.1401570468901039 x MARG_CL_0_3_F +
```

```
0.05254178285396341 x MARG_AL_0_3_M + 0.04178595301201031 x
MARG_AL_0_3_F + 0.12184035387925014 x MARG_HH_0_3_M +
0.11601141016824106 x MARG_HH_0_3_F + 0.13986877411042797 x
MARG_OT_0_3_M + 0.13219224458196 53 x MARG_OT_0_3_F +
0.15037557804411297 x NON_WORK_M + 0.13106620313207334 x NON_WORK_F]
```

Note - For Covariance Matrix, EigenVectors and EigenValues, Please check the support file.

Sample of the Dataset with Principal Components

	State Code	Dist.Code	State	Area Name	PC1	PC2	PC3	PC4	PC5	PC6
32	2	33	Himachal Pradesh	Shimla	-2.083485	3.397777	3.911474	-1.874783	-1.659165	1.030022
177	9	178	Uttar Pradesh	Ambedkar Nagar	5.020271	3.330486	-0.417023	2.256841	-0.547847	-0.796067
103	8	104	Rajasthan	Alwar	2.780606	4.745162	4.322923	-3.371014	-1.181092	0.827915
456	23	457	Madhya Pradesh	Balaghat	2.095755	3.616121	-3.144440	-2.263671	2.175663	-0.402013
301	18	302	Assam	Goalpara	-2.478949	-0.186919	-0.170837	0.341265	0.982491	0.621702
276	14	277	Manipur	Imphal West	-3.578426	-0.824271	0.412773	1.521102	0.168941	1.340295
635	34	636	Puducherry	Mahe	-6.262088	-0.854414	0.242575	1.174113	0.063816	-0.159470
170	9	171	Uttar Pradesh	Chitrakoot	-4.204357	-0.106023	0.030358	0.632765	-0.559789	-0.290528
279	14	280	Manipur	Chandel	-5.941127	-0.222743	0.190224	0.769291	0.321824	0.371604
142	9	143	Uttar Pradesh	Aligarh	6.369499	-0.663564	0.255701	2.430008	-1.391714	-0.685808

Fig 22 - Sample Dataset with Principal Components

Correlation Heatmap of Principal Components

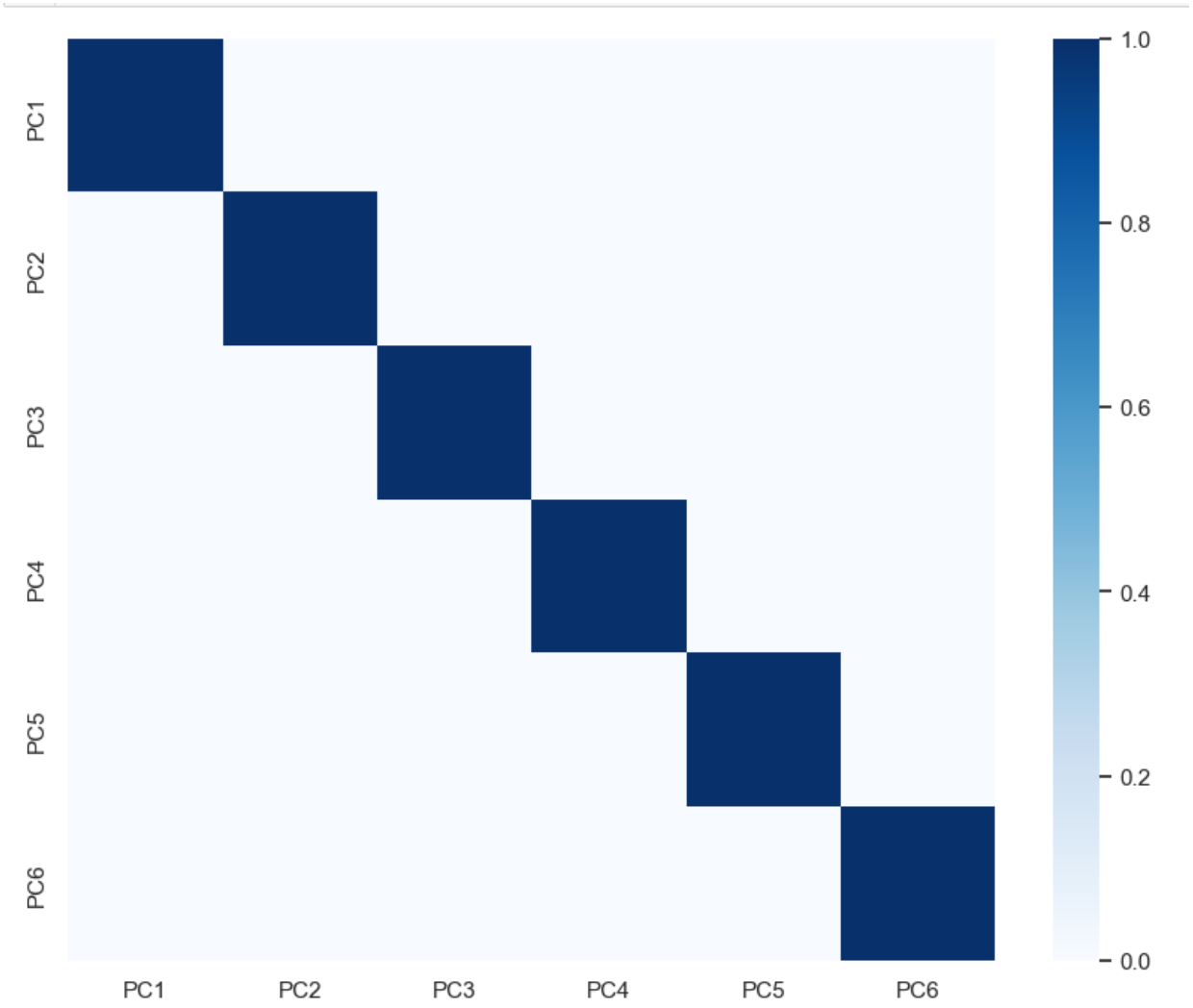

Fig 23 - Correlation Heatmap of Principal Components

THE END