# Business Report

## DSBA Data Mining Project – Part 1
## Principal Component Analysis

# Table of Contents

## List of Figures



## List of Tables



## List of Equations

## Problem Statement

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 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 residence (rural-urban). Census 2011 covered 35 States/Union Territories containing 640 districts which in turn contained 5,924 sub-districts, 7,935 towns and 6,40,867 villages.

The data collected has so many variables 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)**.

Data file - PCA India Data Census.xlsx

## Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.

We will start analyzing the data by going thru the basic steps like:

1. Check head
2. Check info
3. Check summary
4. Check nulls
5. Check duplicates

Let us start by reading the data and extracting basic information:

*Table 1: headfirst 5 rows of the dataset*

| State Code | Dist.Code | State | Area Name | No_HH | TOT_M | TOT_F | M_06 | F_06 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | Jammu & Kashmir | Kupwara | 7707 | 23388 | 29796 | 5862 | 6196 |
| 1 | 2 | Jammu & Kashmir | Badgam | 6218 | 19585 | 23102 | 4482 | 3733 |
| 1 | 3 | Jammu & Kashmir | Leh(Ladakh ) | 4452 | 6546 | 10964 | 1082 | 1018 |
| 1 | 4 | Jammu & Kashmir | Kargil | 1320 | 2784 | 4206 | 563 | 677 |
| 1 | 5 | Jammu & Kashmir | Punch | 11654 | 20591 | 29981 | 5157 | 4587 |

(not all columns are shown)

**Checking Info about the data:**

*Table 2: Dataset info*

| | |
|---|---|
| int64 | 59 |
| object | 2 |

There are **640 rows** and **61 columns** in the dataset where the 59 columns have Integer data type and 2 columns have object data type.

**Checking summary:**

*Table 3: Dataset Summary*

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| State Code | 640 | | | | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Dist.Code | 640 | | | | | | | |
| No_HH | 640 | 51222.9 | 48135.4 | 350.0 | 19484.0 | 35837.0 | 68892.0 | 310450.0 |
| TOT_M | 640 | 79940.6 | 73384.5 | 391.0 | 30228.0 | 58339.0 | 107918.5 | 485417.0 |
| TOT_F | 640 | 122372.1 | 113600.7 | 698.0 | 46517.8 | 87724.5 | 164251.8 | 750392.0 |
| M_06 | 640 | 12309.1 | 11500.9 | 56.0 | 4733.8 | 9159.0 | 16520.3 | 96223.0 |
| F_06 | 640 | 11942.3 | 11326.3 | 56.0 | 4672.3 | 8663.0 | 15902.3 | 95129.0 |
| M_SC | 640 | 13820.9 | 14426.4 | 0.0 | 3466.3 | 9591.5 | 19429.8 | 103307.0 |
| F_SC | 640 | 20778.4 | 21727.9 | 0.0 | 5603.3 | 13709.0 | 29180.0 | 156429.0 |
| M_ST | 640 | 6191.8 | 9912.7 | 0.0 | 293.8 | 2333.5 | 7658.0 | 96785.0 |
| F_ST | 640 | 10155.6 | 15875.7 | 0.0 | 429.5 | 3834.5 | 12480.3 | 130119.0 |
| M_LIT | 640 | 57968.0 | 55910.3 | 286.0 | 21298.0 | 42693.5 | 77989.5 | 403261.0 |
| F_LIT | 640 | 66359.6 | 75037.9 | 371.0 | 20932.0 | 43796.5 | 84799.8 | 571140.0 |
| M_ILL | 640 | 21972.6 | 19825.6 | 105.0 | 8590.0 | 15767.5 | 29512.5 | 105961.0 |
| F_ILL | 640 | 56012.5 | 47116.7 | 327.0 | 22367.0 | 42386.0 | 78471.0 | 254160.0 |
| TOT_WORK_M | 640 | 37992.4 | 36419.5 | 100.0 | 13753.5 | 27936.5 | 50226.8 | 269422.0 |
| TOT_WORK_F | 640 | 41295.8 | 37192.4 | 357.0 | 16097.8 | 30588.5 | 53234.3 | 257848.0 |
| MAINWORK_M | 640 | 30204.4 | 31480.9 | 65.0 | 9787.0 | 21250.5 | 40119.0 | 247911.0 |
| MAINWORK_F | 640 | 28198.8 | 29998.3 | 240.0 | 9502.3 | 18484.0 | 35063.3 | 226166.0 |
| MAIN_CL_M | 640 | 5424.3 | 4739.2 | 0.0 | 2023.5 | 4160.5 | 7695.0 | 29113.0 |
| MAIN_CL_F | 640 | 5486.0 | 5326.4 | 0.0 | 1920.3 | 3908.5 | 7286.3 | 36193.0 |
| MAIN_AL_M | 640 | 5849.1 | 6399.5 | 0.0 | 1070.3 | 3936.5 | 8067.3 | 40843.0 |
| MAIN_AL_F | 640 | 8926.0 | 12864.3 | 0.0 | 1408.8 | 3933.5 | 10617.5 | 87945.0 |
| MAIN_HH_M | 640 | 883.9 | 1278.6 | 0.0 | 187.5 | 498.5 | 1099.3 | 16429.0 |
| MAIN_HH_F | 640 | 1380.8 | 3179.4 | 0.0 | 248.8 | 540.5 | 1435.8 | 45979.0 |
| MAIN_OT_M | 640 | 18047.1 | 26068.5 | 36.0 | 3997.5 | 9598.0 | 21249.5 | 240855.0 |
| MAIN_OT_F | 640 | 12406.0 | 18972.2 | 153.0 | 3142.5 | 6380.5 | 14368.3 | 209355.0 |
| MARGWORK_M | 640 | 7788.0 | 7410.8 | 35.0 | 2937.5 | 5627.0 | 9800.3 | 47553.0 |
| MARGWORK_F | 640 | 13096.9 | 10996.5 | 117.0 | 5424.5 | 10175.0 | 18879.3 | 66915.0 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| MARG_CL_M | 640 | 1040.7 | 1311.5 | 0.0 | 311.8 | 606.5 | 1281.0 | 13201.0 |
| MARG_CL_F | 640 | 2307.7 | 3564.6 | 0.0 | 630.3 | 1226.0 | 2659.3 | 44324.0 |
| MARG_AL_M | 640 | 3304.3 | 3781.6 | 0.0 | 873.5 | 2062.0 | 4300.8 | 23719.0 |
| MARG_AL_F | 640 | 6463.3 | 6773.9 | 0.0 | 1402.5 | 4020.5 | 9089.3 | 45301.0 |
| MARG_HH_M | 640 | 316.7 | 462.7 | 0.0 | 71.8 | 166.0 | 356.5 | 4298.0 |
| MARG_HH_F | 640 | 786.6 | 1198.7 | 0.0 | 171.8 | 429.0 | 962.5 | 15448.0 |
| MARG_OT_M | 640 | 3126.2 | 3609.4 | 7.0 | 935.5 | 2036.0 | 3985.3 | 24728.0 |
| MARG_OT_F | 640 | 3539.3 | 4115.2 | 19.0 | 1071.8 | 2349.5 | 4400.5 | 36377.0 |
| MARGWORK_3_6_M | 640 | 41948.2 | 39045.3 | 291.0 | 16208.3 | 30315.0 | 57218.8 | 300937.0 |
| MARGWORK_3_6_F | 640 | 81076.3 | 82970.4 | 341.0 | 26619.5 | 56793.0 | 107924.0 | 676450.0 |
| MARG_CL_3_6_M | 640 | 6395.0 | 6019.8 | 27.0 | 2372.0 | 4630.0 | 8167.0 | 39106.0 |
| MARG_CL_3_6_F | 640 | 10339.9 | 8467.5 | 85.0 | 4351.5 | 8295.0 | 15102.0 | 50065.0 |
| MARG_AL_3_6_M | 640 | 789.8 | 905.6 | 0.0 | 235.5 | 480.5 | 986.0 | 7426.0 |
| MARG_AL_3_6_F | 640 | 1749.6 | 2496.5 | 0.0 | 497.3 | 985.5 | 2059.0 | 27171.0 |
| MARG_HH_3_6_M | 640 | 2743.6 | 3059.6 | 0.0 | 718.8 | 1714.5 | 3702.3 | 19343.0 |
| MARG_HH_3_6_F | 640 | 5169.9 | 5335.6 | 0.0 | 1113.8 | 3294.0 | 7502.3 | 36253.0 |
| MARG_OT_3_6_M | 640 | 245.4 | 358.7 | 0.0 | 58.0 | 129.5 | 276.0 | 3535.0 |
| MARG_OT_3_6_F | 640 | 585.9 | 900.0 | 0.0 | 127.8 | 320.5 | 719.3 | 12094.0 |
| MARGWORK_0_3_M | 640 | 2616.1 | 3037.0 | 7.0 | 755.0 | 1681.5 | 3320.3 | 20648.0 |
| MARGWORK_0_3_F | 640 | 2834.5 | 3327.8 | 14.0 | 833.5 | 1834.5 | 3610.5 | 25844.0 |
| MARG_CL_0_3_M | 640 | 1393.0 | 1489.7 | 4.0 | 489.5 | 949.0 | 1714.0 | 9875.0 |
| MARG_CL_0_3_F | 640 | 2757.1 | 2788.8 | 30.0 | 957.3 | 1928.0 | 3599.8 | 21611.0 |
| MARG_AL_0_3_M | 640 | 250.9 | 453.3 | 0.0 | 47.0 | 114.5 | 270.8 | 5775.0 |
| MARG_AL_0_3_F | 640 | 558.1 | 1117.6 | 0.0 | 109.0 | 247.5 | 568.8 | 17153.0 |
| MARG_HH_0_3_M | 640 | 560.7 | 762.6 | 0.0 | 136.5 | 308.0 | 642.0 | 6116.0 |
| MARG_HH_0_3_F | 640 | 1293.4 | 1585.4 | 0.0 | 298.0 | 717.0 | 1710.8 | 13714.0 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| MARG_OT_0_3_M | 640 | 71.4 | 107.9 | 0.0 | 14.0 | 35.0 | 79.0 | 895.0 |
| MARG_OT_0_3_F | 640 | 200.7 | 309.7 | 0.0 | 43.0 | 113.0 | 240.0 | 3354.0 |
| NON_WORK_M | 640 | 510.0 | 610.6 | 0.0 | 161.0 | 326.0 | 604.5 | 6456.0 |
| NON_WORK_F | 640 | 704.8 | 910.2 | 5.0 | 220.5 | 464.5 | 853.5 | 10533.0 |

We can see there are 640 districts (as per 2011). On the average there are about 52 thousand households in each district. However, the range is between 350 and over 3 lakhs. We will explore more in the EDA section.

**Checking Nulls**

There are no missing (null) values in the dataset.

**Checking Duplicates**

There are no duplicate values in the dataset.

## Perform a detailed exploratory analysis of the variables. Since the number of variables is very large, you are asked to choose any 5 variables from the 20 important variables listed below.

**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**

**Example Question:**

While exploring the variables, it is recommended that you focus on the insights possible from each of the variables. Also provide a small discussion based on the plots or tables.

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

The state of Andhra Pradesh has the highest female to male ratio (1.89) according to 2011 census data. This means 1.89 females per male. While the Union Territory of Lakshadweep has the lowest gender ratio of 1.15. Among the states, Haryana & Uttar Pradesh have the lowest gender ratio (F to M).

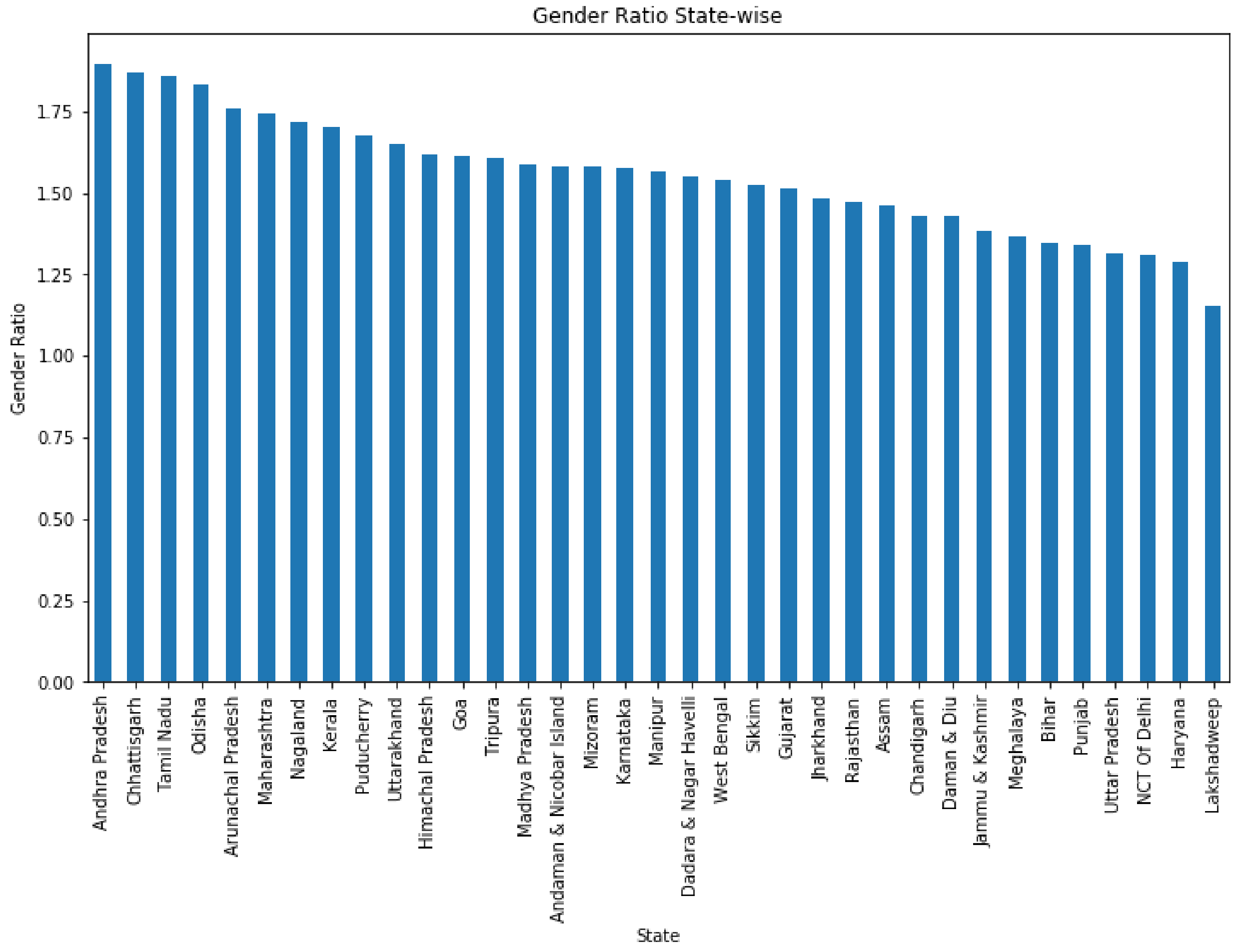



*Figure 1: Gender Ratio Statewise*

**Which district has the highest & lowest gender ratio?**

- Krishna District of Andhra Pradesh has the highest Female to Male ratio of 2.28.

Badgam District of Jammu & Kashmir has the lowest Female to Male ratio of 1.17

The below map shows Gender-Ratio as per State. You can see that 'Telangana' is white because the data is for 2011 and Telangana has been created in 2014. . You can explore to get old shape files for India before 2011.

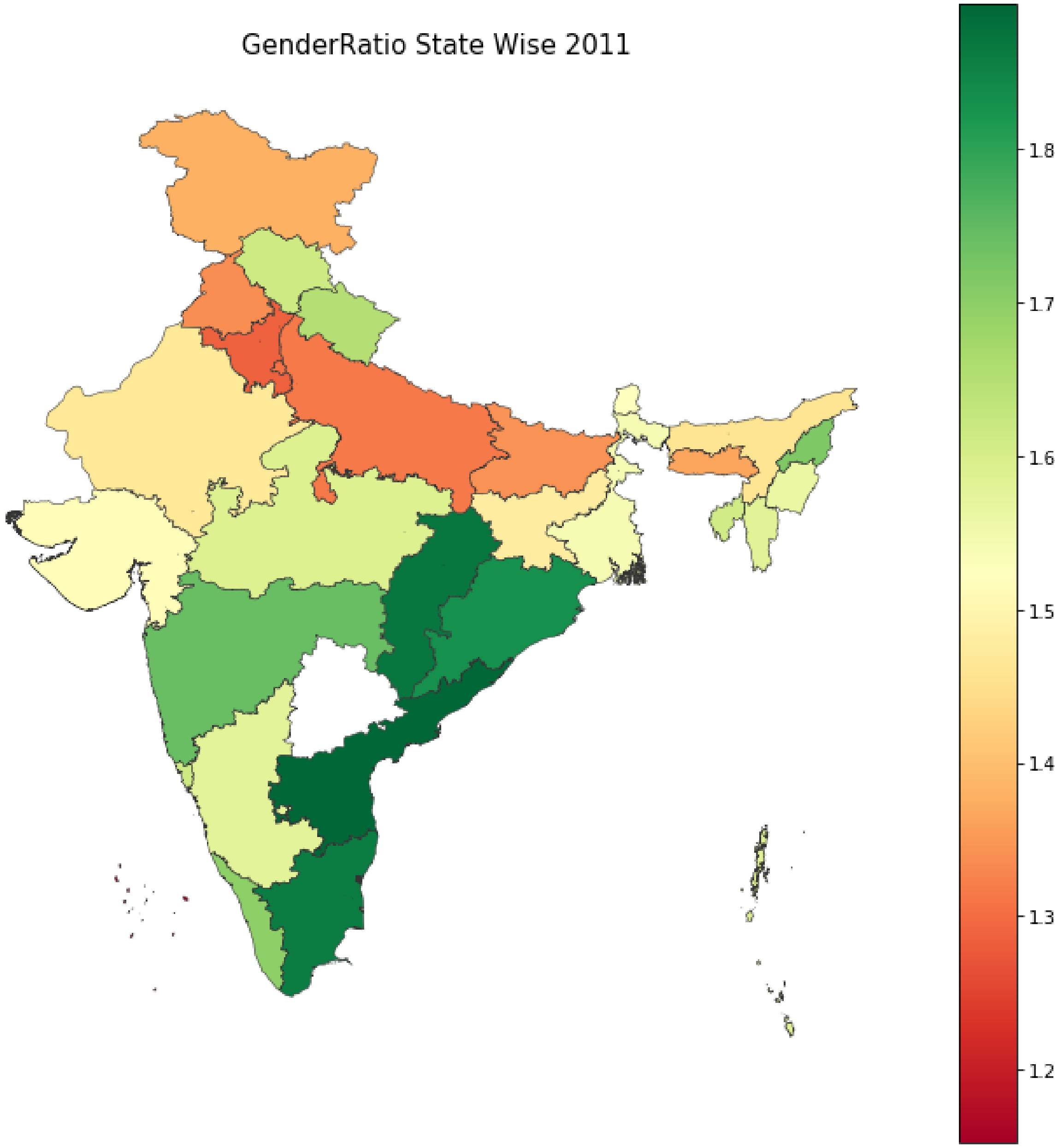



*Figure 2: India Map - with Gender Ratios*

According to the data, northern states have lower gender ratios in general.

## 2. Analysis of Literacy

**Female Literacy Rate is defined as the**

Number of Literate Females / Total Literate Population *100

Kerala is at the top while Rajasthan & Bihar are at the bottom.

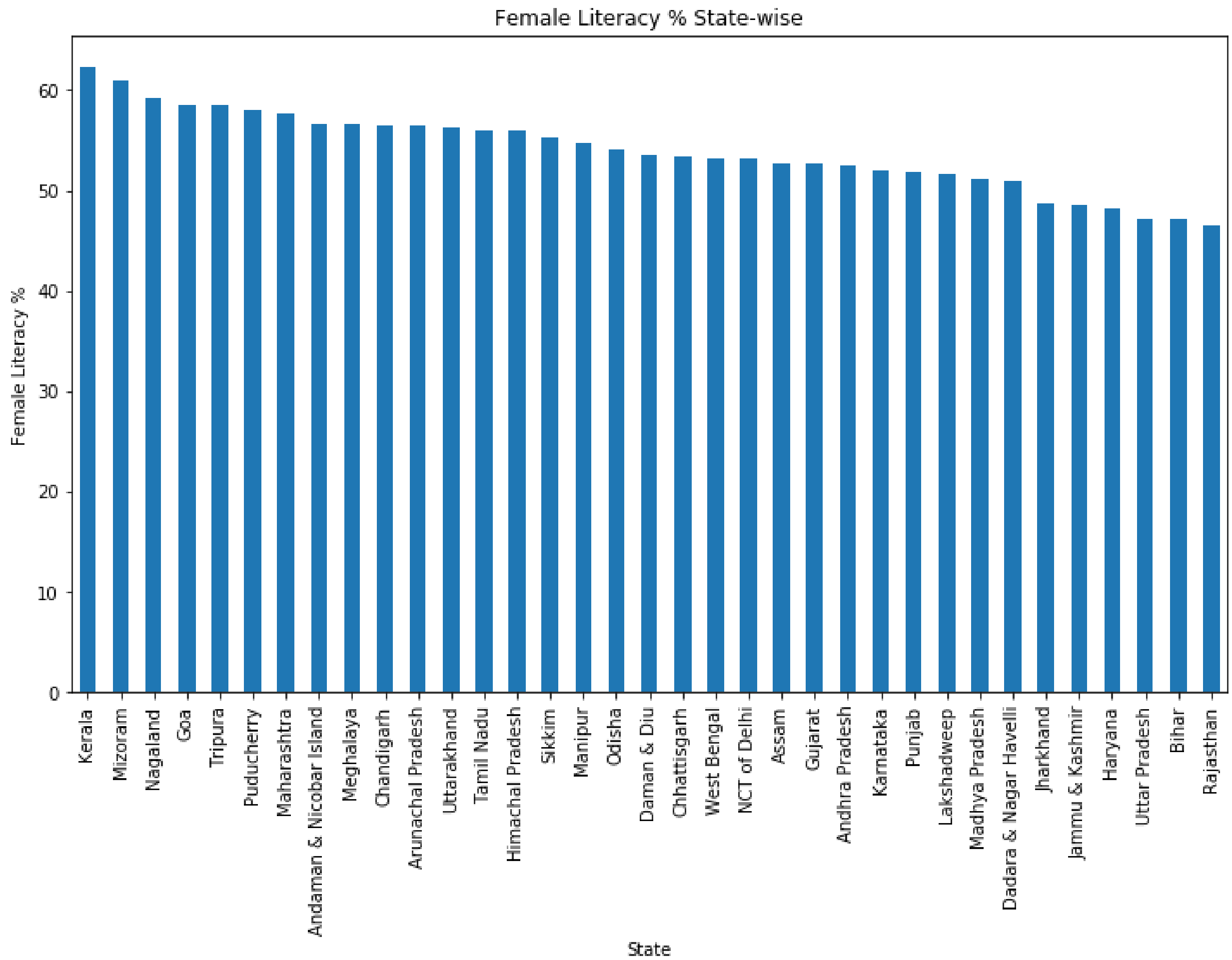



*Figure 3: Female Literacy Rate ( State wise)*

### 3. Non-working population

Uttar Pradesh has the most 'non-working' population according to the data in 2011. Kerala has most 'non-working' female population after Uttar Pradesh.

Daman & Diu and Dadra Nagar Haveli have the lowest number of non-working population for both Females & Males.

Let us now investigate non-working male and female populations separately

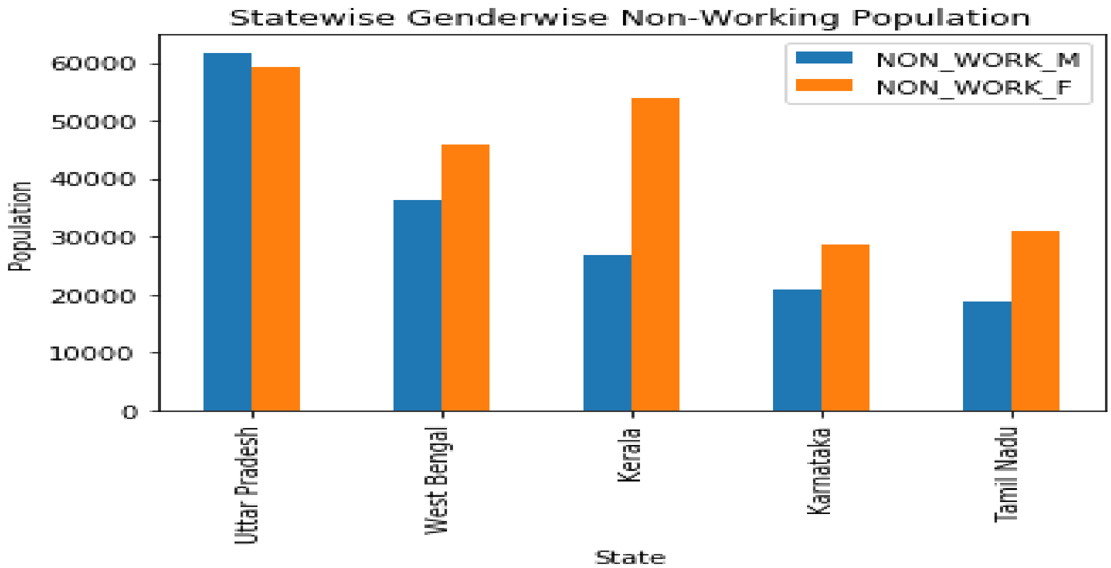



*Figure 4: Statewise Non-Working Population by gender for the top states*

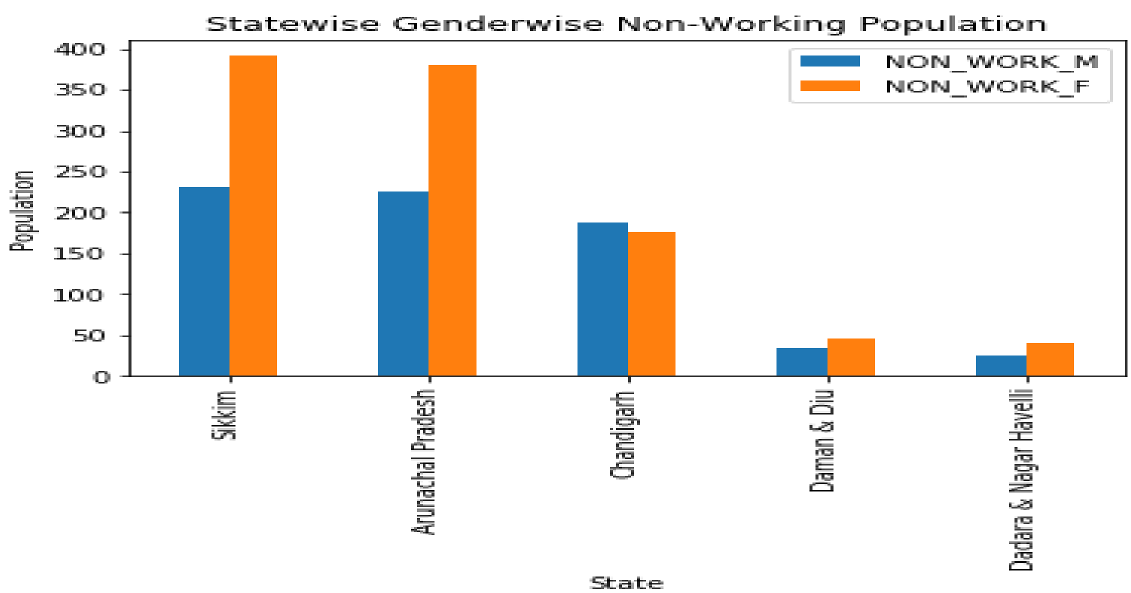



*Figure 5: Statewise Non-Working Population by gender for the bottom states*

**4. Statewise SC/ST population by gender**

Uttar Pradesh has the highest number of SC/ST population. It is also observed that SC population is significantly higher than ST population according to 2011 data. It is also noted that there are more SC Females than males.

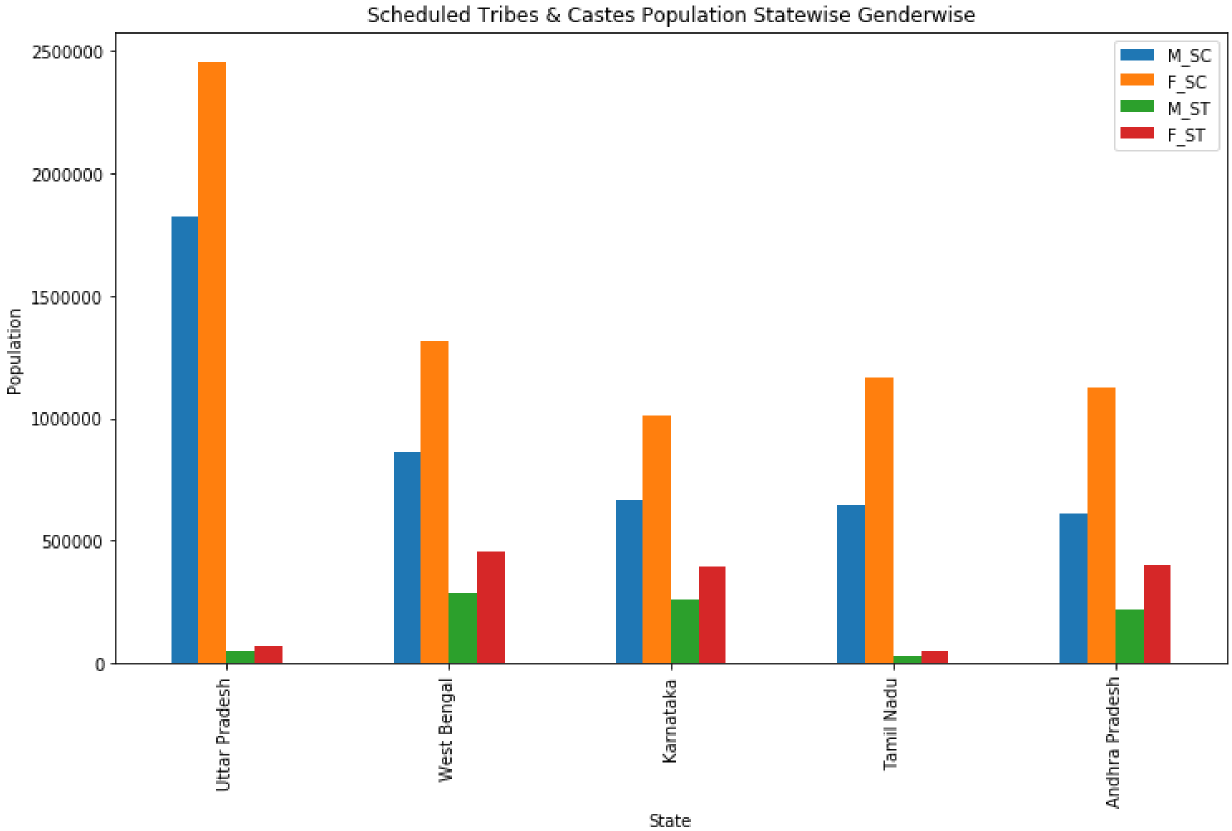



Figure number missing

There can be more exploration on this data based on your personal interest. For example – take one state or UT and dig deeper. You can create Instagram/ LinkedIn template based infographics and share them on your Social Profile to build network. **You are strictly forbidden to share this project on any public or private forum.**

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

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

Before Scaling -

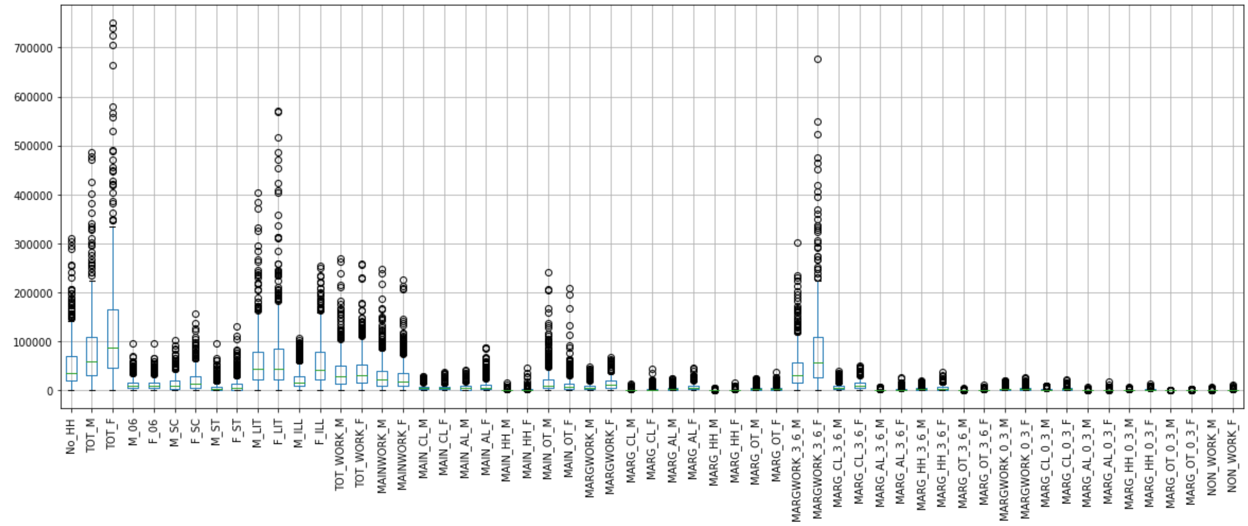


*Figure 6: Boxplot before scaling*

After scaling,

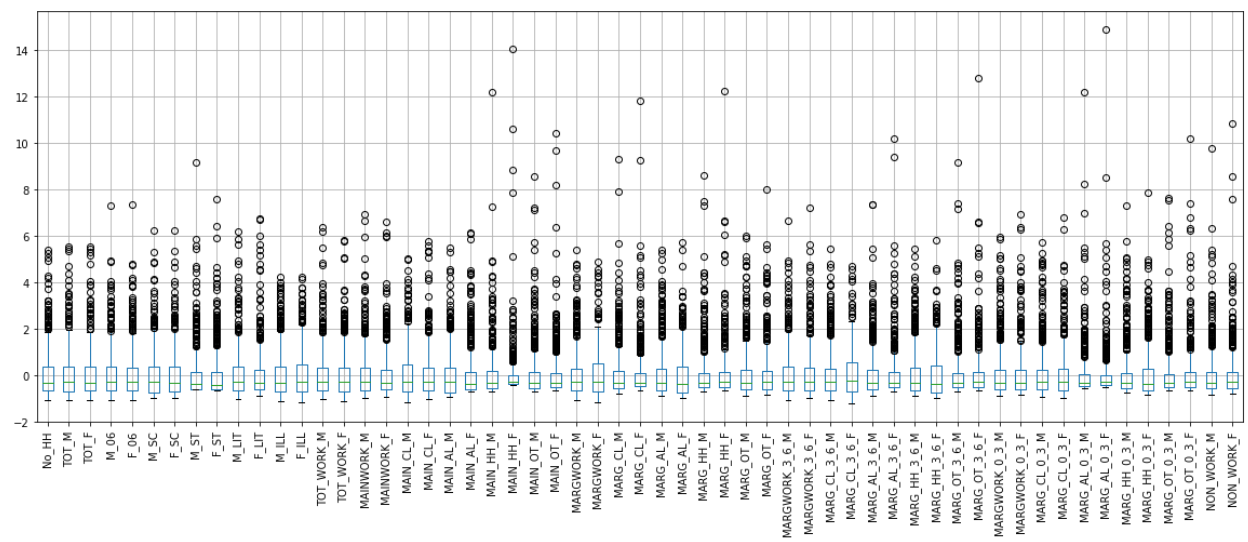


*Figure 7: Boxplots after scaling*

## Perform all the required steps for PCA (use sklearn only)

**Bartletts Test of Sphericity**

Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population. If the null hypothesis cannot be rejected, then PCA is not advisable.

$H_0$ : All variables in the data are uncorrelated

$H_1$: At least one pair of variables in the data are correlated

**Inference:** Since p-value: 0.00, we reject the null hypothesis is rejected.

**KMO Test**

The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is. Generally, if MSA is less than 0.5, PCA is not recommended, since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

**MSA = 0.80349**

**Considerable reduction in data dimension is expected**

**Step 1**- Create the covariance Matrix

**Covariance Matrix**

*Table 4: Covariance Matrix (part)*

| | No_HH | TOT_M | TOT_F | M_06 | F_06 | M_SC | F_SC | M_ST | F_ST |
|---|---|---|---|---|---|---|---|---|---|
| No_HH | 1 | 0.92 | 0.97 | 0.8 | 0.8 | 0.78 | 0.83 | 0.15 | 0.17 |
| TOT_M | 0.92 | 1 | 0.98 | 0.95 | 0.95 | 0.84 | 0.83 | 0.09 | 0.09 |
| TOT_F | 0.97 | 0.98 | 1 | 0.91 | 0.91 | 0.82 | 0.83 | 0.12 | 0.13 |
| M_06 | 0.8 | 0.95 | 0.91 | 1 | 1 | 0.78 | 0.75 | 0.06 | 0.04 |
| F_06 | 0.8 | 0.95 | 0.91 | 1 | 1 | 0.77 | 0.74 | 0.07 | 0.05 |
| M_SC | 0.78 | 0.84 | 0.82 | 0.78 | 0.77 | 1 | 0.99 | -0.05 | -0.05 |
| F_SC | 0.83 | 0.83 | 0.83 | 0.75 | 0.74 | 0.99 | 1 | -0.01 | -0.01 |
| M_ST | 0.15 | 0.09 | 0.12 | 0.06 | 0.07 | -0.05 | -0.01 | 1 | 0.99 |
| F_ST | 0.17 | 0.09 | 0.13 | 0.04 | 0.05 | -0.05 | -0.01 | 0.99 | 1 |
| M_LIT | 0.93 | 0.99 | 0.99 | 0.91 | 0.91 | 0.82 | 0.82 | 0.09 | 0.09 |
| F_LIT | 0.93 | 0.93 | 0.96 | 0.83 | 0.83 | 0.72 | 0.73 | 0.1 | 0.1 |
| M_ILL | 0.76 | 0.91 | 0.86 | 0.95 | 0.95 | 0.8 | 0.76 | 0.08 | 0.07 |
| F_ILL | 0.86 | 0.89 | 0.89 | 0.86 | 0.87 | 0.83 | 0.85 | 0.14 | 0.15 |
| TOT_WORK_M | 0.94 | 0.97 | 0.97 | 0.86 | 0.85 | 0.83 | 0.82 | 0.12 | 0.12 |
| TOT_WORK_F | 0.93 | 0.81 | 0.88 | 0.68 | 0.69 | 0.71 | 0.78 | 0.27 | 0.29 |
| MAINWORK_M | 0.93 | 0.93 | 0.94 | 0.79 | 0.79 | 0.78 | 0.78 | 0.11 | 0.11 |
| MAINWORK_F | 0.89 | 0.75 | 0.82 | 0.59 | 0.59 | 0.65 | 0.71 | 0.23 | 0.25 |

**Step 2**- Get eigen values and eigen vector

```
Eigenvectors: [[ 0.16  0.17  0.17  0.16  0.16  0.15  0.15  0.03  0.03  0.1
6  0.15  0.16
   0.17  0.16  0.15  0.15  0.12  0.1   0.07  0.11  0.07  0.13  0.08  0.12
   0.11  0.16  0.16  0.08  0.05  0.13  0.11  0.14  0.13  0.16  0.15  0.16
   0.16  0.17  0.16  0.09  0.05  0.13  0.11  0.14  0.12  0.15  0.15  0.15
   0.14  0.05  0.04  0.12  0.12  0.14  0.13  0.15  0.13]
 [-0.13 -0.09 -0.1  -0.02 -0.02 -0.05 -0.05  0.03  0.03 -0.12 -0.15 -0.01
  -0.01 -0.13 -0.09 -0.18 -0.15  0.06  0.09 -0.03 -0.06 -0.08 -0.08 -0.21
  -0.21  0.09  0.13  0.27  0.25  0.17  0.14  0.07  0.02 -0.09 -0.12 -0.04
  -0.11  0.08  0.1   0.26  0.24  0.16  0.13  0.06  0.01 -0.09 -0.13  0.15
   0.18  0.25  0.24  0.19  0.18  0.08  0.05 -0.07 -0.07]
 [-0.    0.06  0.04  0.06  0.05  0.   -0.03 -0.12 -0.14  0.08  0.12 -0.02
  -0.09  0.05 -0.06  0.05 -0.06 -0.07 -0.01 -0.25 -0.25  0.03 -0.06  0.14
   0.1  -0.01 -0.05  0.2   0.27 -0.19 -0.27 -0.02 -0.08  0.11  0.1   0.06
   0.08 -0.02 -0.07  0.15  0.26 -0.2  -0.28 -0.02 -0.08  0.11  0.1   0.05
   0.02  0.27  0.28 -0.14 -0.2  -0.02 -0.08  0.11  0.1 ]
```

```
Eigenvalues: [3.181e+01 7.870e+00 4.150e+00 3.670e+00 2.210e+00 1.940e+00
1.180e+00
 7.500e-01 6.200e-01 5.300e-01 4.300e-01 3.500e-01 3.000e-01 2.800e-01
 1.900e-01 1.400e-01 1.100e-01 1.100e-01 1.000e-01 8.000e-02 6.000e-02
 4.000e-02 4.000e-02 3.000e-02 3.000e-02 2.000e-02 1.000e-02 1.000e-02
 1.000e-02 1.000e-02 1.000e-02 1.000e-02 0.000e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 0.000e+00]
```

## Identify the optimum number of PCs (for this project, the optimum number is based on the explanation of at least 90% of variance )

Since the number of variables is large and value of MSA is 0.8, it is expected that  a few components will be enough to explain  90% of variation in the data.

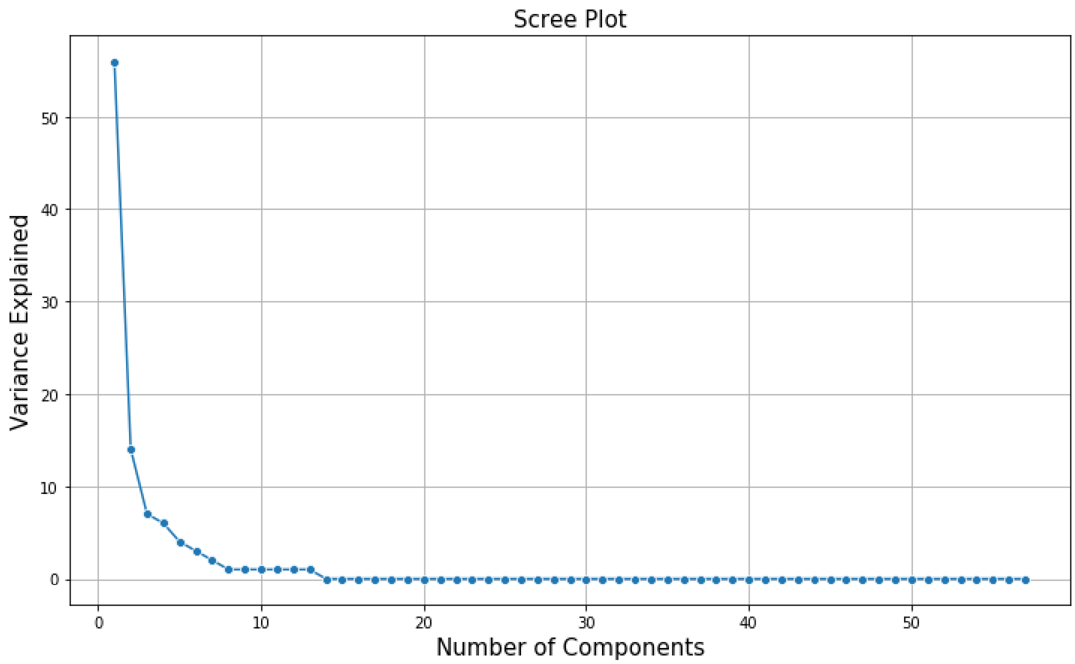



*Figure 8: Scree Plot*

From Above plot and cumulative explained variance, 6 PCs are chosen

## Compare PCs with actual variables and identify which is explaining most variance. Try to explain the PCs in terms of the original variables

*Table 5: Correlations between PCs and original variables*

| | PC0 | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|---|
| **No_HH** | 0.16 | -0.13 | -0 | -0.13 | -0.01 | 0 |
| **TOT_M** | 0.17 | -0.09 | 0.06 | -0.02 | -0.03 | -0.07 |
| **TOT_F** | 0.17 | -0.1 | 0.04 | -0.07 | -0.01 | -0.04 |
| **M_06** | 0.16 | -0.02 | 0.06 | 0.01 | -0.05 | -0.16 |

| PC0 | PC1 | PC2 | PC3 | PC4 | PC5 | |
|---|---|---|---|---|---|---|
| **F_06** | 0.16 | -0.02 | 0.05 | 0.01 | -0.04 | -0.15 |
| **M_SC** | 0.15 | -0.05 | 0 | 0.01 | -0.17 | -0.06 |
| **F_SC** | 0.15 | -0.05 | -0.03 | -0.03 | -0.16 | -0.04 |
| **M_ST** | 0.03 | 0.03 | -0.12 | -0.22 | 0.43 | 0.22 |
| **F_ST** | 0.03 | 0.03 | -0.14 | -0.23 | 0.44 | 0.23 |
| **M_LIT** | 0.16 | -0.12 | 0.08 | -0.04 | -0.01 | -0.06 |
| **F_LIT** | 0.15 | -0.15 | 0.12 | -0.06 | 0.06 | -0.05 |
| **M_ILL** | 0.16 | -0.01 | -0.02 | 0.03 | -0.1 | -0.12 |
| **F_ILL** | 0.17 | -0.01 | -0.09 | -0.08 | -0.12 | -0.03 |
| **TOT_WORK_M** | 0.16 | -0.13 | 0.05 | -0.04 | -0.02 | -0 |
| **TOT_WORK_F** | 0.15 | -0.09 | -0.06 | -0.23 | -0.04 | 0.11 |
| **MAINWORK_M** | 0.15 | -0.18 | 0.05 | -0.07 | -0.04 | 0.02 |
| **MAINWORK_F** | 0.12 | -0.15 | -0.06 | -0.25 | -0.08 | 0.12 |
| **MAIN_CL_M** | 0.1 | 0.06 | -0.07 | -0.09 | -0.29 | -0.01 |
| **MAIN_CL_F** | 0.07 | 0.09 | -0.01 | -0.29 | -0.24 | 0.1 |
| **MAIN_AL_M** | 0.11 | -0.03 | -0.25 | -0.14 | -0.21 | -0.03 |
| **MAIN_AL_F** | 0.07 | -0.06 | -0.25 | -0.29 | -0.18 | 0.02 |

| PC0 | PC1 | PC2 | PC3 | PC4 | PC5 | |
|---|---|---|---|---|---|---|
| **MAIN_HH_M** | 0.13 | -0.08 | 0.03 | 0.15 | -0.13 | 0.17 |
| **MAIN_HH_F** | 0.08 | -0.08 | -0.06 | 0.05 | -0.14 | 0.42 |
| **MAIN_OT_M** | 0.12 | -0.21 | 0.14 | -0.04 | 0.06 | 0.02 |
| **MAIN_OT_F** | 0.11 | -0.21 | 0.1 | -0.12 | 0.08 | 0.08 |
| **MARGWORK_M** | 0.16 | 0.09 | -0.01 | 0.09 | 0.06 | -0.09 |
| **MARGWORK_F** | 0.16 | 0.13 | -0.05 | -0.09 | 0.09 | 0.02 |
| **MARG_CL_M** | 0.08 | 0.27 | 0.2 | -0.06 | -0.02 | 0.03 |
| **MARG_CL_F** | 0.05 | 0.25 | 0.27 | -0.17 | -0.06 | 0.09 |
| **MARG_AL_M** | 0.13 | 0.17 | -0.19 | 0.09 | 0.02 | -0.14 |
| **MARG_AL_F** | 0.11 | 0.14 | -0.27 | -0.11 | 0.08 | -0.09 |
| **MARG_HH_M** | 0.14 | 0.07 | -0.02 | 0.24 | -0.06 | 0.09 |
| **MARG_HH_F** | 0.13 | 0.02 | -0.08 | 0.2 | -0.03 | 0.37 |
| **MARG_OT_M** | 0.16 | -0.09 | 0.11 | 0.09 | 0.12 | -0.06 |
| **MARG_OT_F** | 0.15 | -0.12 | 0.1 | 0.03 | 0.17 | 0 |
| **MARGWORK_3_6_M** | 0.16 | -0.04 | 0.06 | -0 | -0.04 | -0.14 |
| **MARGWORK_3_6_F** | 0.16 | -0.11 | 0.08 | 0 | 0 | -0.11 |
| **MARG_CL_3_6_M** | 0.17 | 0.08 | -0.02 | 0.09 | 0.05 | -0.1 |

| PC0 | PC1 | PC2 | PC3 | PC4 | PC5 | |
|---|---|---|---|---|---|---|
| **MARG_CL_3_6_F** | 0.16 | 0.1 | -0.07 | -0.11 | 0.07 | 0.02 |
| **MARG_AL_3_6_M** | 0.09 | 0.26 | 0.15 | -0.04 | -0.01 | 0.01 |
| **MARG_AL_3_6_F** | 0.05 | 0.24 | 0.26 | -0.18 | -0.06 | 0.09 |
| **MARG_HH_3_6_M** | 0.13 | 0.16 | -0.2 | 0.08 | 0.01 | -0.14 |
| **MARG_HH_3_6_F** | 0.11 | 0.13 | -0.28 | -0.14 | 0.06 | -0.08 |
| **MARG_OT_3_6_M** | 0.14 | 0.06 | -0.02 | 0.24 | -0.07 | 0.1 |
| **MARG_OT_3_6_F** | 0.12 | 0.01 | -0.08 | 0.19 | -0.04 | 0.38 |
| **MARGWORK_0_3_M** | 0.15 | -0.09 | 0.11 | 0.09 | 0.11 | -0.06 |
| **MARGWORK_0_3_F** | 0.15 | -0.13 | 0.1 | 0.03 | 0.14 | 0.01 |
| **MARG_CL_0_3_M** | 0.15 | 0.15 | 0.05 | 0.09 | 0.08 | -0.06 |
| **MARG_CL_0_3_F** | 0.14 | 0.18 | 0.02 | -0.02 | 0.13 | -0 |
| **MARG_AL_0_3_M** | 0.05 | 0.25 | 0.27 | -0.1 | -0.05 | 0.07 |
| **MARG_AL_0_3_F** | 0.04 | 0.24 | 0.28 | -0.14 | -0.05 | 0.08 |
| **MARG_HH_0_3_M** | 0.12 | 0.19 | -0.14 | 0.13 | 0.06 | -0.12 |
| **MARG_HH_0_3_F** | 0.12 | 0.18 | -0.2 | 0 | 0.13 | -0.11 |
| **MARG_OT_0_3_M** | 0.14 | 0.08 | -0.02 | 0.23 | -0.04 | 0.06 |
| **MARG_OT_0_3_F** | 0.13 | 0.05 | -0.08 | 0.21 | 0 | 0.3 |

| PC0 | PC1 | PC2 | PC3 | PC4 | PC5 | |
|---|---|---|---|---|---|---|
| **NON_WORK_M** | 0.15 | -0.07 | 0.11 | 0.08 | 0.16 | -0.05 |
| **NON_WORK_F** | 0.13 | -0.07 | 0.1 | 0.02 | 0.24 | -0.02 |

**Observations**:

The first Principal component is positively correlated with Number of Household, Total Male & Female population, Literacy & Illiteracy Numbers among M & F, Number of SC in Males & Females, Working population, etc. These variables explain the most variance in the data i.e. 56%

The Second Principal component is correlated with Marginal Cultivator Male/Female population and Marginal Agriculture (Male & Female) population etc. The Second PC explains about 14% of variation in the data.

The Third Principal Component explains about 7% variation in the data. It positively correlates with Marginal Agriculture 0-3 Female, and 3-6 M&F Population.

The Fourth Principal Component correlated positively with Marginal Households Male, Marginal Other (0-3,3-6) Workers Male population. It explains about 6% of variation in the data.

The Fifth Principal Component explains about 4% variation in data. It is positively correlated with Scheduled Tribes Population Male& Female, Non-working Male& Female population.

The Sixth Principal Component explains about 3% variation in data. It is positively correlated with Female Marginal Other workers (0-3,3-6), Main & Marginal Households Female population.

Overall the first 6 PCs explain 90% variation in the data. Each PCs correlates with a different set of variables explaining how different aspects of population contribute to the variation in data.

## Write explicitly the linear equation for the first PC

*Equation 1: Linear Equation for First PC*

```
( 0.16 ) * No_HH + ( 0.17 ) * TOT_M + ( 0.17 ) * TOT_F + ( 0.16 ) * M_06 +
( 0.16 ) * F_06 + ( 0.15 ) * M_SC + ( 0.15 ) * F_SC + ( 0.03 ) * M_ST + (
0.03 ) * F_ST + ( 0.16 ) * M_LIT + ( 0.15 ) * F_LIT + ( 0.16 ) * M_ILL + (
0.17 ) * F_ILL + ( 0.16 ) * TOT_WORK_M + ( 0.15 ) * TOT_WORK_F + ( 0.15 )
* MAINWORK_M + ( 0.12 ) * MAINWORK_F + ( 0.1 ) * MAIN_CL_M + ( 0.07 ) * MA
IN_CL_F + ( 0.11 ) * MAIN_AL_M + ( 0.07 ) * MAIN_AL_F + ( 0.13 ) * MAIN_HH
_M + ( 0.08 ) * MAIN_HH_F + ( 0.12 ) * MAIN_OT_M + ( 0.11 ) * MAIN_OT_F +
( 0.16 ) * MARGWORK_M + ( 0.16 ) * MARGWORK_F + ( 0.08 ) * MARG_CL_M + ( 0
.05 ) * MARG_CL_F + ( 0.13 ) * MARG_AL_M + ( 0.11 ) * MARG_AL_F + ( 0.14 )
```

```
* MARG_HH_M + ( 0.13 ) * MARG_HH_F + ( 0.16 ) * MARG_OT_M + ( 0.15 ) * MAR
G_OT_F + ( 0.16 ) * MARGWORK_3_6_M + ( 0.16 ) * MARGWORK_3_6_F + ( 0.17 )
* MARG_CL_3_6_M + ( 0.16 ) * MARG_CL_3_6_F + ( 0.09 ) * MARG_AL_3_6_M + (
0.05 ) * MARG_AL_3_6_F + ( 0.13 ) * MARG_HH_3_6_M + ( 0.11 ) * MARG_HH_3_6
_F + ( 0.14 ) * MARG_OT_3_6_M + ( 0.12 ) * MARG_OT_3_6_F + ( 0.15 ) * MARG
WORK_0_3_M + ( 0.15 ) * MARGWORK_0_3_F + ( 0.15 ) * MARG_CL_0_3_M + ( 0.14
) * MARG_CL_0_3_F + ( 0.05 ) * MARG_AL_0_3_M + ( 0.04 ) * MARG_AL_0_3_F +
( 0.12 ) * MARG_HH_0_3_M + ( 0.12 ) * MARG_HH_0_3_F + ( 0.14 ) * MARG_OT_0
_3_M + ( 0.13 ) * MARG_OT_0_3_F + ( 0.15 ) * NON_WORK_M + ( 0.13 ) * NON_W
ORK_F
```

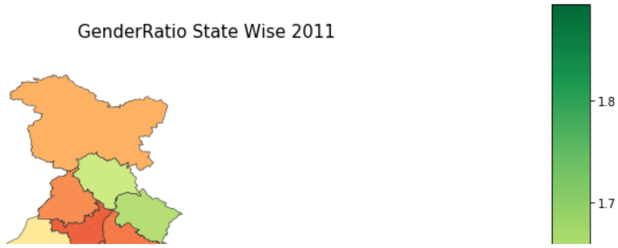
The variable names are indicative of their scaled form.

# Appendix

Code:

```
In [1]: import pandas as pd
        import numpy as np
        import seaborn as sns
        import matplotlib.pyplot as plt

        from factor_analyzer import FactorAnalyzer

In [2]: # reading data

        df = pd.read_excel('PCA India Data Census.xlsx')

In [3]: df.head().T

In [12]: from matplotlib import pyplot as plt

In [14]: # Which state has highest gender ratio and which has the lowest?

In [15]: eda= df.copy(deep=True)

In [16]: eda['GenderRatio'] = eda['TOT_F']/eda['TOT_M']

In [17]: plt.title('Gender Ratio State-wise')
         plt.ylabel('Gender Ratio')
         eda.groupby(['State','Area Name']).mean()['GenderRatio'].sort_values(ascending=False)#.plot(kind='bar',figsize=(12,7));

Out[17]: State           Area Name
         Andhra Pradesh  Krishna                     2.283250
         Odisha          Koraput                     2.268763
         Tamil Nadu      Virudhunagar                2.225429
         Andhra Pradesh  West Godavari               2.221849

In [18]: # Which state district has the highest gender ratio?

In [19]: # Which state district has the highest gender ratio?

In [20]: #!pip install geopandas

In [21]: import geopandas as gpd

In [22]: shapes = gpd.read_file('../Downloads/India Map Shape Files/India States/Indian_states.shp')

In [23]: plot_data = eda.copy(deep=True)

         plot_data= plot_data[['State','GenderRatio']]

         plot_data1= plot_data.groupby(['State']).mean()['GenderRatio'].reset_index()

         plot_data2 = pd.merge(shapes,plot_data1, right_on=plot_data1.State,left_on=shapes.st_nm, how='left')

         plot_data2= plot_data2.set_index('st_nm')[['geometry','GenderRatio']].dropna()
```

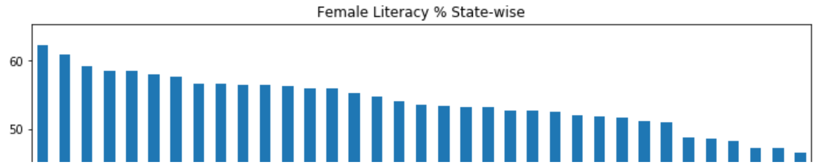
In [24]:
```python
variable ='GenderRatio'
fig, ax = plt.subplots(1, figsize=(12, 12))
ax.axis('off')
ax.set_title(' GenderRatio State Wise 2011',fontdict={'fontsize': '15', 'fontweight' : '3'})
fig = plot_data2.plot(variable,cmap='RdYlGn', linewidth=0.5, ax=ax, edgecolor='0.2',legend=True)
# due to data being old some states not visible and delhi is missing probably because of spelling
```




**Literacy**

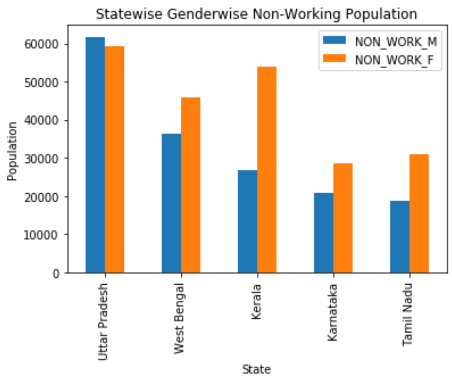
In [25]:
```python
#Female Literacy

eda['f_lit_r'] = eda['F_LIT'] / (eda['M_LIT']+eda['F_LIT'] ) *100
```

In [26]:
```python
plt.title('Female Literacy % State-wise')
plt.ylabel('Female Literacy %')
eda.groupby(['State']).mean()['f_lit_r'].sort_values(ascending=False).plot(kind='bar',figsize=(12,7));
```

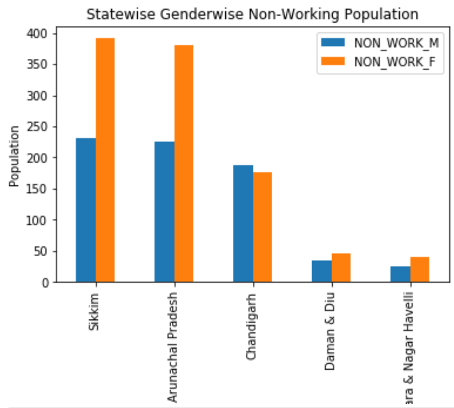



In [28]:
```python
import seaborn as sns
```

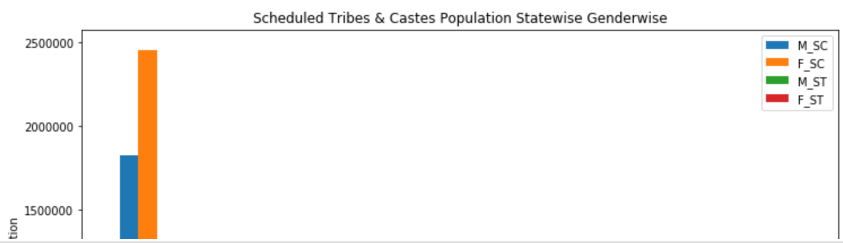
In [29]:
```python
eda.groupby('State').sum()[['NON_WORK_M', 'NON_WORK_F']].sort_values(by=['NON_WORK_M', 'NON_WORK_F'],ascending=False).head(5).pl
plt.title('Statewise Genderwise Non-Working Population')
plt.ylabel('Population')
plt.show()
```



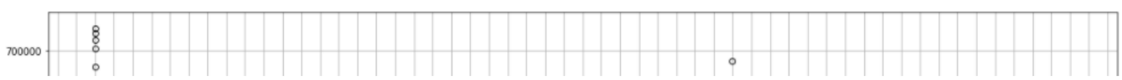
In [30]:
```
eda.groupby('State').sum()[['NON_WORK_M', 'NON_WORK_F']].sort_values(by=['NON_WORK_M', 'NON_WORK_F'],
                                                                     ascending=False).tail(5).plot(kind='bar')

plt.title('Statewise Genderwise Non-Working Population')
plt.ylabel('Population')
plt.show()
```




In [31]:
```
eda.groupby('State').sum()[['M_SC', 'F_SC', 'M_ST','F_ST']].sort_values(
    by=['M_SC', 'F_SC', 'M_ST','F_ST'],ascending=False).head(5).plot(kind='bar',figsize=(12,7))
plt.title('Scheduled Tribes & Castes Population Statewise Genderwise')
plt.ylabel('Population')
plt.show()
```




In [38]:
```
df_num.boxplot(figsize=(20,7))
plt.xticks(rotation=90)
plt.show()
```

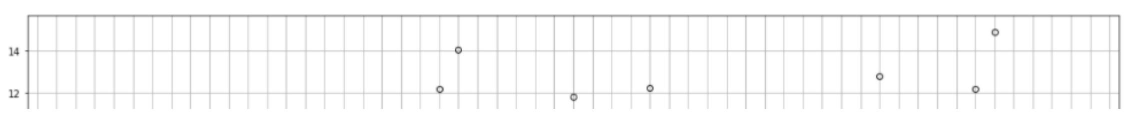


In [42]:
```
from scipy.stats import zscore
df_num_scaled=df_num.apply(zscore)
df_num_scaled.head()
```

Out[42]:

| | 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_AL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | -0.904738 | -0.771236 | -0.815563 | -0.561012 | -0.507738 | -0.958575 | -0.957049 | -0.423306 | -0.476423 | -0.798097 | ... | -0.163229 | -0.720610 | -( |
| 1 | -0.935695 | -0.823100 | -0.874534 | -0.681096 | -0.725367 | -0.958297 | -0.956772 | -0.582014 | -0.607607 | -0.849434 | ... | -0.583103 | -0.732811 | -( |

In [44]:
```
df_num_scaled.boxplot(figsize=(20,7))
plt.xticks(rotation=90)
plt.show()
```



## Bartletts Test of Sphericity

**Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.**

> $H_O$: All variables in the data are uncorrelated

> $H_A$: At least one pair of variables in the data are correlated

If the null hypothesis cannot be rejected, then PCA is not advisable.

In [46]:
```
from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
chi_square_value,p_value=calculate_bartlett_sphericity(df_num_scaled)
p_value
```

```
C:\Users\Vimesh\Anaconda3\lib\site-packages\factor_analyzer\factor_analyzer.py:111: RuntimeWarning: divide by zero encountered
in log
  statistic = -np.log(corr_det) * (n - 1 - (2 * p + 5) / 6)
```

Out[46]:
```
0.0
```

## KMO Test

**The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.**

> Generally, if MSA is less than *0.5*, PCA is not recommended, since no reduction is expected. On the other hand, *MSA > 0.7* is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

In [47]:
```
from factor_analyzer.factor_analyzer import calculate_kmo
kmo_all,kmo_model=calculate_kmo(df_num_scaled)
kmo_model
```

```
C:\Users\Vimesh\Anaconda3\lib\site-packages\factor_analyzer\utils.py:249: UserWarning: The inverse of the variance-covariance m
atrix was calculated using the Moore-Penrose generalized matrix inversion, due to its determinant being at or very close to zer
o.
  warnings.warn('The inverse of the variance-covariance matrix '
```

Out[47]:
```
0.8034956686157672
```

## Step 1- Create the covariance Matrix

In [48]:
```
pd.set_option('display.max_rows', 200)
pd.set_option('display.max_columns', 200)

pd.set_option('display.expand_frame_repr', True)
pd.get_option("display.max_rows")
np.set_printoptions(threshold=np.inf)
```

In [49]:
```
from sklearn.decomposition import PCA
pca = PCA(random_state=123)
df_pca = pca.fit_transform(df_num_scaled)
```

In [50]:
```
pd.DataFrame(np.round(pca.get_covariance(),2),columns=df_num_scaled.columns,index=df_num_scaled.columns) #cov matrix
```

## Step 2- Get eigen values and eigen vector

In [51]:
```
eigenvec=pca.components_
print('Eigenvectors:',np.round(eigenvec,2))
```

In [52]:
```
eigenvalues=pca.explained_variance_
print('Eigenvalues:',np.round(eigenvalues,2))
```

```
Eigenvalues: [3.181e+01 7.870e+00 4.150e+00 3.670e+00 2.210e+00 1.940e+00 1.180e+00
 7.500e-01 6.200e-01 5.300e-01 4.300e-01 3.500e-01 3.000e-01 2.800e-01
 1.900e-01 1.400e-01 1.100e-01 1.100e-01 1.000e-01 8.000e-02 6.000e-02
 4.000e-02 4.000e-02 3.000e-02 3.000e-02 2.000e-02 1.000e-02 1.000e-02
```

```
0.000e+00]
```

```
In [53]: var_exp=np.round(pca.explained_variance_ratio_,2)*100
```

```
In [54]: var_exp
```

```
Out[54]: array([56., 14.,  7.,  6.,  4.,  3.,  2.,  1.,  1.,  1.,  1.,  1.,  1.,
                 0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
                 0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
                 0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
                 0.,  0.,  0.,  0.,  0.])
```

### Step 3 View Scree Plot to identify the number of components to be built

```
In [55]: plt.figure(figsize=(12,7))
         sns.lineplot(y=var_exp,x=range(1,len(var_exp)+1),marker='o')
         plt.xlabel('Number of Components',fontsize=15)
         plt.ylabel('Variance Explained',fontsize=15)
         plt.title('Scree Plot',fontsize=15)
         plt.grid()
         plt.show()
```

Scree Plot

```
In [56]: # Step 4 Apply PCA for the number of decided components to get the loadings and component output

         from sklearn.decomposition import PCA
         pca = PCA(n_components=6,random_state=123)
         df_pca = pca.fit_transform(df_num_scaled)
         df_pca.transpose() # Component output
```

```
df_pca_loading = pd.DataFrame(pca.components_,columns=list(df_num_scaled),index=['PC0','PC1','PC2','PC3','PC4','PC5'])
df_pca_loading.shape
```

```
(6, 57)
```

```
df_pca_loading = np.round(df_pca_loading,2)
```

```
df_pca_loading.style.highlight_max(color = 'lightgreen', axis = 0)
```

| | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PC0 | 0.16 | 0.17 | 0.17 | 0.16 | 0.16 | 0.15 | 0.15 | 0.03 | 0.03 | 0.16 | 0.15 | 0.16 | 0.17 | 0.16 | 0.15 | 0.15 | 0. |
| PC1 | -0.13 | -0.09 | -0.1 | -0.02 | -0.02 | -0.05 | -0.05 | 0.03 | 0.03 | -0.12 | -0.15 | -0.01 | -0.01 | -0.13 | -0.09 | -0.18 | -0. |
| PC2 | -0 | 0.06 | 0.04 | 0.06 | 0.05 | 0 | -0.03 | -0.12 | -0.14 | 0.08 | 0.12 | -0.02 | -0.09 | 0.05 | -0.06 | 0.05 | -0. |

```
In [60]: # linear equation of first PC
```

```
In [61]: for i in range(0,57):
             print("(",np.round(pca.components_[0][i],2),")",'*',df_num_scaled.columns[i], end=' + ')
```

```
( 0.16 ) * No_HH + ( 0.17 ) * TOT_M + ( 0.17 ) * TOT_F + ( 0.16 ) * M_06 + ( 0.16 ) * F_06 + ( 0.15 ) * M_SC + ( 0.15 ) * F_SC
+ ( 0.03 ) * M_ST + ( 0.03 ) * F_ST + ( 0.16 ) * M_LIT + ( 0.15 ) * F_LIT + ( 0.16 ) * M_ILL + ( 0.17 ) * F_ILL + ( 0.16 ) * TO
T_WORK_M + ( 0.15 ) * TOT_WORK_F + ( 0.15 ) * MAINWORK_M + ( 0.12 ) * MAINWORK_F + ( 0.1 ) * MAIN_CL_M + ( 0.07 ) * MAIN_CL_F +
( 0.11 ) * MAIN_AL_M + ( 0.07 ) * MAIN_AL_F + ( 0.13 ) * MAIN_HH_M + ( 0.08 ) * MAIN_HH_F + ( 0.12 ) * MAIN_OT_M + ( 0.11 ) * M
AIN_OT_F + ( 0.16 ) * MARGWORK_M + ( 0.16 ) * MARGWORK_F + ( 0.08 ) * MARG_CL_M + ( 0.05 ) * MARG_CL_F + ( 0.13 ) * MARG_AL_M +
( 0.11 ) * MARG_AL_F + ( 0.14 ) * MARG_HH_M + ( 0.13 ) * MARG_HH_F + ( 0.16 ) * MARG_OT_M + ( 0.15 ) * MARG_OT_F + ( 0.16 ) * M
ARGWORK_3_6_M + ( 0.16 ) * MARGWORK_3_6_F + ( 0.17 ) * MARG_CL_3_6_M + ( 0.16 ) * MARG_CL_3_6_F + ( 0.09 ) * MARG_AL_3_6_M + (
0.05 ) * MARG_AL_3_6_F + ( 0.13 ) * MARG_HH_3_6_M + ( 0.11 ) * MARG_HH_3_6_F + ( 0.14 ) * MARG_OT_3_6_M + ( 0.12 ) * MARG_OT_3_
6_F + ( 0.15 ) * MARGWORK_0_3_M + ( 0.15 ) * MARGWORK_0_3_F + ( 0.15 ) * MARG_CL_0_3_M + ( 0.14 ) * MARG_CL_0_3_F + ( 0.05 ) *
MARG_AL_0_3_M + ( 0.04 ) * MARG_AL_0_3_F + ( 0.12 ) * MARG_HH_0_3_M + ( 0.12 ) * MARG_HH_0_3_F + ( 0.14 ) * MARG_OT_0_3_M + (
0.13 ) * MARG_OT_0_3_F + ( 0.15 ) * NON_WORK_M + ( 0.13 ) * NON_WORK_F +
```