Principal Component Analysis

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

There is a 2 object and 59 null is presented in our 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_/
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180	
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123	
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	...	114	188	44	
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	...	194	247	61	
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	...	874	1928	465	

5 rows × 61 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 640 entries, 0 to 639
Data columns (total 61 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   State Code      640 non-null    int64
 1   Dist.Code       640 non-null    int64
 2   State           640 non-null    object
 3   Area Name       640 non-null    object
 4   No_HH           640 non-null    int64
 5   TOT_M           640 non-null    int64
 6   TOT_F           640 non-null    int64
 7   M_06            640 non-null    int64
 8   F_06            640 non-null    int64
 9   M_SC            640 non-null    int64
 10  F_SC            640 non-null    int64
 11  M_ST            640 non-null    int64
 12  F_ST            640 non-null    int64
 13  M_LIT           640 non-null    int64
 14  F_LIT           640 non-null    int64
 15  M_ILL           640 non-null    int64
 16  F_ILL           640 non-null    int64
 17  TOT_WORK_M      640 non-null    int64
 18  TOT_WORK_F      640 non-null    int64
 19  MAINWORK_M      640 non-null    int64
 20  MAINWORK_F      640 non-null    int64
 21  MAIN_CL_M       640 non-null    int64
```

```
 37  MARG_OT_M       640 non-null    int64
 38  MARG_OT_F       640 non-null    int64
 39  MARGWORK_3_6_M  640 non-null    int64
 40  MARGWORK_3_6_F  640 non-null    int64
 41  MARG_CL_3_6_M   640 non-null    int64
 42  MARG_CL_3_6_F   640 non-null    int64
 43  MARG_AL_3_6_M   640 non-null    int64
 44  MARG_AL_3_6_F   640 non-null    int64
 45  MARG_HH_3_6_M   640 non-null    int64
 46  MARG_HH_3_6_F   640 non-null    int64
 47  MARG_OT_3_6_M   640 non-null    int64
 48  MARG_OT_3_6_F   640 non-null    int64
 49  MARGWORK_0_3_M  640 non-null    int64
 50  MARGWORK_0_3_F  640 non-null    int64
 51  MARG_CL_0_3_M   640 non-null    int64
 52  MARG_CL_0_3_F   640 non-null    int64
 53  MARG_AL_0_3_M   640 non-null    int64
 54  MARG_AL_0_3_F   640 non-null    int64
 55  MARG_HH_0_3_M   640 non-null    int64
 56  MARG_HH_0_3_F   640 non-null    int64
 57  MARG_OT_0_3_M   640 non-null    int64
 58  MARG_OT_0_3_F   640 non-null    int64
 59  NON_WORK_M      640 non-null    int64
 60  NON_WORK_F      640 non-null    int64
dtypes: int64(59), object(2)
memory usage: 305.1+ KB
```

	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

In this summary, we can the whole feature or class has a outlier because the mean value and 50%(median value) not close, it’s far from the entire mean.

We can also visualize the outlier through boxplot but this is appropriate way to identify the outliers.

Identify the Null

There is no na and null in our dataset

```
State Code       0
Dist.Code        0
State            0
Area Name        0
No_HH            0
                ..
MARG_HH_0_3_F    0
MARG_OT_0_3_M    0
MARG_OT_0_3_F    0
NON_WORK_M       0
NON_WORK_F       0
Length: 61, dtype: int64
```

Identify the NA

```
State Code       0
Dist.Code        0
State            0
Area Name        0
No_HH            0
                ..
MARG_HH_0_3_F    0
MARG_OT_0_3_M    0
MARG_OT_0_3_F    0
NON_WORK_M       0
NON_WORK_F       0
Length: 61, dtype: int64
```

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? (Example Questions). Pick 5 variables out of the given 24 variables below for EDA: No_HH, TOT_M, TOT_F, M_06, F_06, M_SC, F_SC, M_ST, F_ST, M_LIT, F_LIT, M_ILL, F_ILL, TOT_WORK_M, TOT_WORK_F, MAINWORK_M, MAINWORK_F, MAIN_CL_M, MAIN_CL_F, MAIN_AL_M, MAIN_AL_F, MAIN_HH_M, MAIN_HH_F, MAIN_OT_M, MAIN_OT_F

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

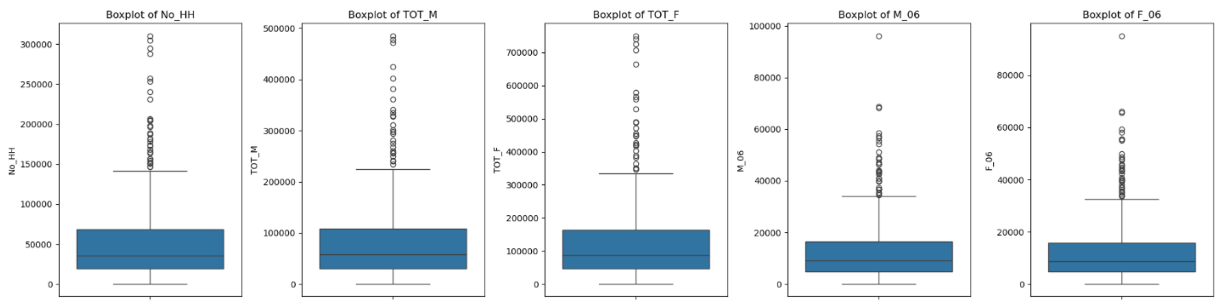
State	TOT_M	TOT_F	gender_ratio
Lakshadweep	12823	14772	0.868061
Andhra Pradesh	3274363	6097235	0.537024

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

Area Name	TOT_M	TOT_F	gender_ratio
Lakshadweep	12823	14772	0.868061
Krishna	137603	314182	0.437972

Let’s we see the boxplot for the purpose of outlier detection

<Figure size 400x500 with 0 Axes>



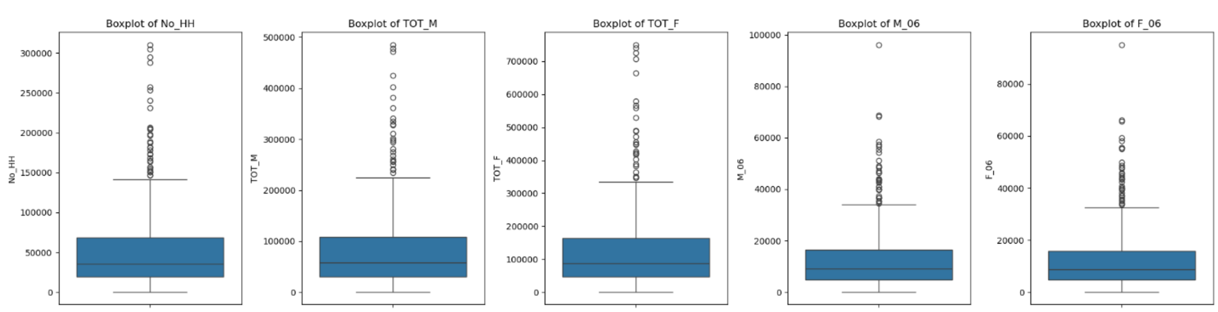
We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?

Treating outliers might affecting my scaling features. For this academic purpose e don’t treat the outlier’s

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

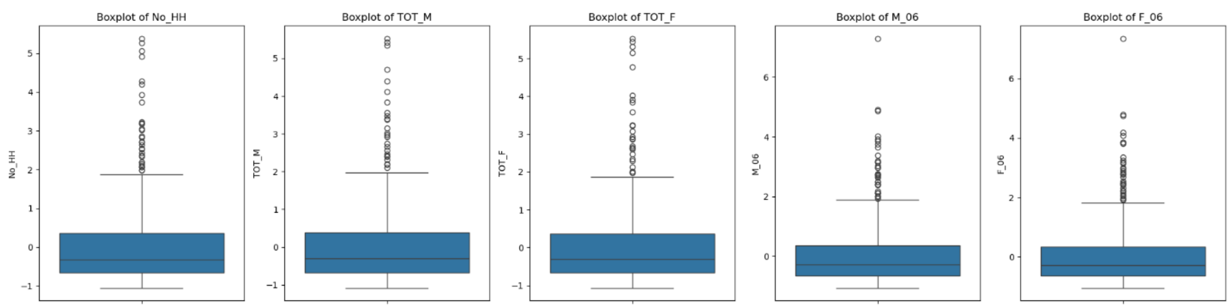
Before scaling:

<Figure size 400x500 with 0 Axes>



After scaling:

The scale of the boxplot may varies when compare to before scaling. May be the scale is changed



	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	MAR
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	
2	-0.972412	-1.000919	-0.981466	-0.976956	-0.965262	-0.958575	-0.956772	-0.038951	-0.027273	-0.956457	...	-0.859212	-0.921931	
3	-1.037530	-1.052224	-1.041001	-1.022118	-0.995393	-0.958783	-0.957049	-0.355965	-0.390060	-1.004643	...	-0.805468	-0.900758	
4	-0.822676	-0.809381	-0.813933	-0.622359	-0.649908	-0.957395	-0.955529	0.149238	0.043330	-0.800568	...	-0.348645	-0.297513	

5 rows × 57 columns

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

Before do with PCA let’s check with 2 test, one is for relationship with features and another one is adequacy (is my data have enough information or not)

Barlett spherecity:

As per this test, we check the correlation with features. The alpha value is the indicator sign whether the correlation is presented or not.

H0: Correlations are not significant – Null Hypothesis

H1:Correlations are significant

The test should be below 0.05

P_Value is 0.0
The null hypothesis has been rejected so my model is **correlation.**

KMO Test:(Adequacy)

It tells us adequacy(the model gave enough data or not)

The test should be above 0.7 is good and below is not acceptable

Model is 0.80 so I have a enough date to process a further steps that is PCA.

Eigen values represents the quantum of information in the data while eigen vectors direction of information

Eigen Vectors

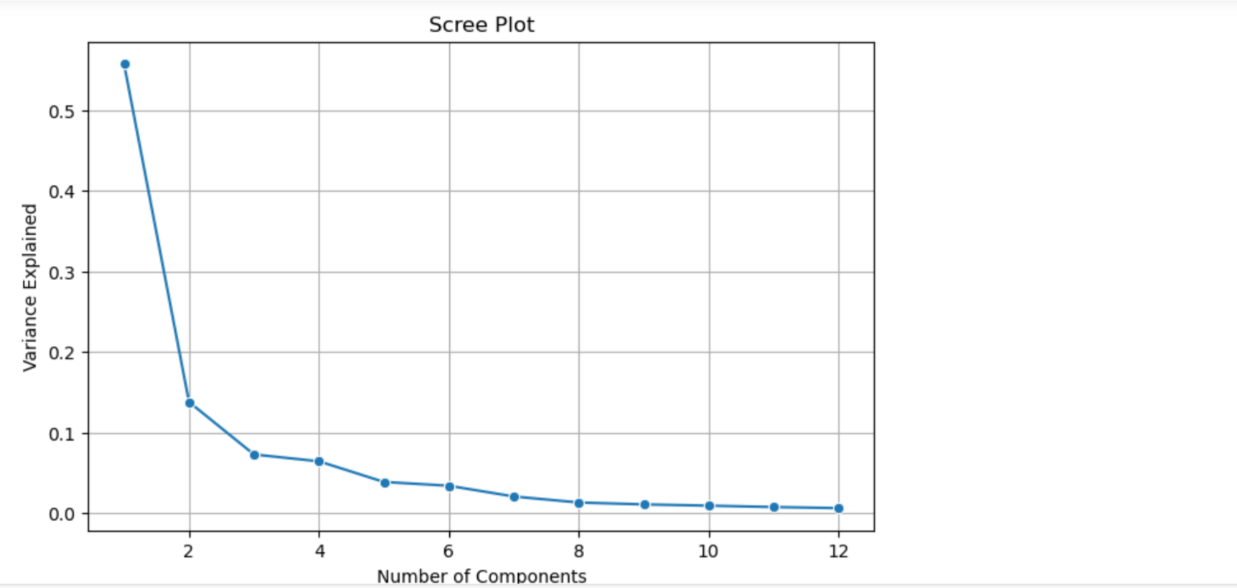
```
array([[ 1.56020579e-01,  1.67117635e-01,  1.65553179e-01,
         1.62192948e-01,  1.62566396e-01,  1.51357849e-01,
         1.51566500e-01,  2.72341946e-02,  2.81833150e-02,
         1.61992837e-01,  1.46872680e-01,  1.61749445e-01,
         1.65248187e-01,  1.59871988e-01,  1.45935804e-01,
         1.46200730e-01,  1.23970284e-01,  1.03127159e-01,
         7.45397856e-02,  1.13355712e-01,  7.38821590e-02,
         1.31572584e-01,  8.33826397e-02,  1.23526242e-01,
         1.11021264e-01,  1.64615479e-01,  1.55395618e-01,
         8.23885414e-02,  4.91953957e-02,  1.28598563e-01,
         1.14305073e-01,  1.40853227e-01,  1.27669598e-01,
         1.55262872e-01,  1.47286584e-01,  1.64971950e-01,
         1.61253433e-01,  1.65501611e-01,  1.55647049e-01,
         9.30142064e-02,  5.15358640e-02,  1.28576116e-01,
         1.10645843e-01,  1.39592763e-01,  1.24545909e-01,
         1.54293786e-01,  1.46285654e-01,  1.50125706e-01,
         1.40157047e-01,  5.25417829e-02,  4.17859530e-02,
         1.21840354e-01,  1.16011410e-01,  1.39868774e-01,
         1.32192245e-01,  1.50375578e-01,  1.31066203e-01],
       [-1.26346525e-01, -8.96765481e-02, -1.04912371e-01,
        -2.20945086e-02, -2.02705495e-02, -4.51109032e-02,
        -5.19237543e-02,  2.76790387e-02,  3.02225550e-02,
        -1.15354767e-01, -1.53109487e-01, -6.62537318e-03,
```

Eigen Values:

```
array([0.55726063, 0.13784435, 0.07275295, 0.06426418, 0.03865049,
       0.03395169, 0.02060239, 0.01315764, 0.01080859, 0.00925395,
       0.00752912, 0.00619102])
```

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

So the 6th pc component shows the 90% explained variance in 2 plots



Observations

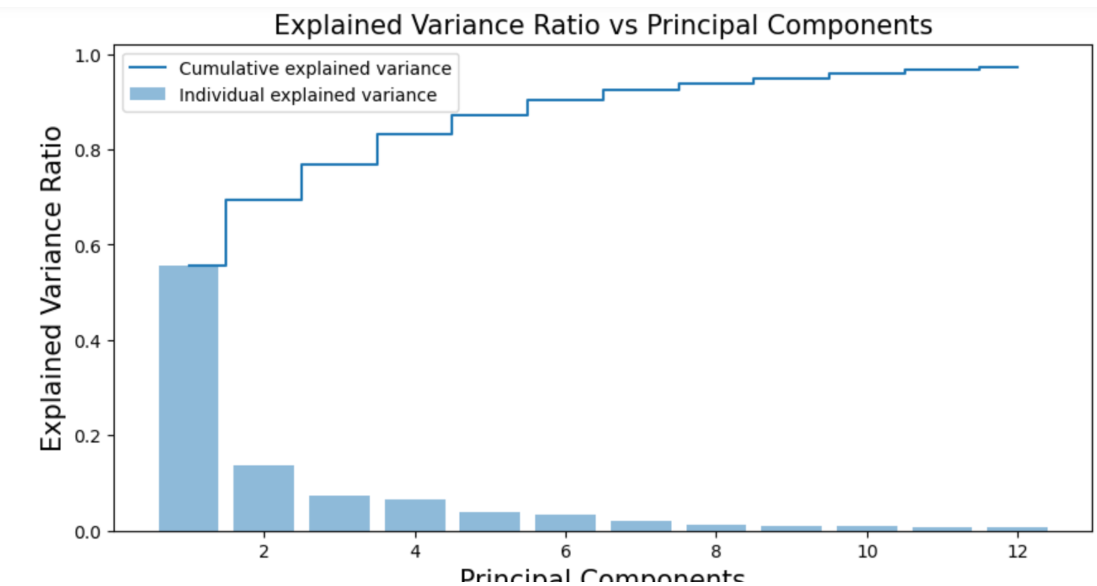
We can see that out of the 12 original features, we reduced the number of features through principal components to 6, these components explain more than 90% of the original variance.

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

```
array([0.55726063, 0.13784435, 0.07275295, 0.06426418, 0.03865049,
       0.03395169, 0.02060239, 0.01315764, 0.01080859, 0.00925395,
       0.00752912, 0.00619102])
```

As per the explained variance, the PC1, PC2, PC3, PC4, PC5 and PC6 is the most explained variables of the above PCA's

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
No_HH	0.156021	-0.126347	-0.002690	-0.125293	-0.007022	0.004083	-0.118110	0.057239	0.004263	0.019988	0.010595	0.086181
TOT_M	0.167118	-0.089677	0.056698	-0.019942	-0.033026	-0.073389	0.089554	0.111431	0.018872	-0.024502	0.011145	0.018851
TOT_F	0.165553	-0.104912	0.038749	-0.070873	-0.012847	-0.043647	-0.002124	0.088355	0.014911	-0.038040	0.007735	0.093546
M_06	0.162193	-0.022095	0.057788	0.011917	-0.050248	-0.157957	0.165067	0.169595	-0.056772	-0.153575	0.081251	0.104358
F_06	0.162566	-0.020271	0.050126	0.014844	-0.043848	-0.154436	0.169082	0.169458	-0.059322	-0.169568	0.081963	0.105285
M_SC	0.151358	-0.045111	0.002569	0.012485	-0.173007	-0.064295	-0.001566	-0.129301	0.037481	0.448516	-0.228822	-0.076361
F_SC	0.151567	-0.051924	-0.025101	-0.029893	-0.159803	-0.040518	-0.084658	-0.144352	0.041232	0.446968	-0.213023	-0.010992
M_ST	0.027234	0.027679	-0.123504	-0.222247	0.433163	0.222591	0.405505	0.021982	0.018632	0.160418	0.067589	0.014768
F_ST	0.028183	0.030223	-0.139769	-0.229754	0.438792	0.225531	0.357800	0.014874	0.043866	0.134863	0.053348	0.022338
M_LIT	0.161993	-0.115355	0.082168	-0.035163	-0.009101	-0.055465	0.045934	0.099423	0.045194	-0.005752	-0.030218	0.075911
F_LIT	0.146873	-0.153109	0.117098	-0.059559	0.055844	-0.048021	-0.021064	0.110360	0.021997	-0.040669	-0.033356	0.192890
M_ILL	0.161749	-0.006625	-0.021855	0.025348	-0.096580	-0.115234	0.201947	0.132080	-0.057596	-0.074472	0.126472	-0.144300
F_ILL	0.165248	-0.009107	-0.093062	-0.076023	-0.119910	-0.028757	0.028425	0.037270	0.000918	-0.026946	0.071772	-0.081651
TOT_WORK_M	0.159872	-0.133529	0.045176	-0.040154	-0.019553	-0.001801	0.045053	0.076869	0.045257	0.080154	-0.031362	-0.104655



Write linear equation for first PC.

PCA Equation:

```
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 * 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 * 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 * MARG_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 * MAR
G_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 (+)
```

Conclusion:

The first 6 PCA Components is the important to take further analysis and the maximum amount data is available only upto the 6 th PCA compenents