Business Report

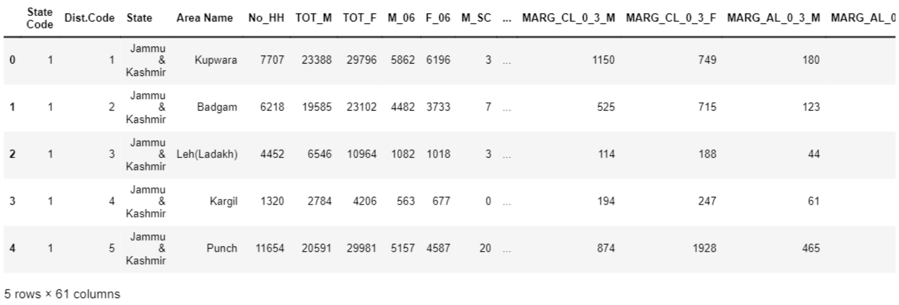
Principal Component Analysis

Ayush Sharma

Table of Contents

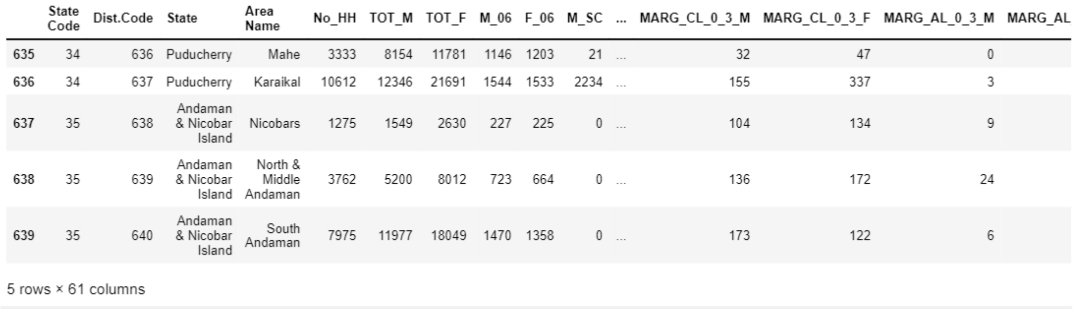
A. **Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.**

Ans:



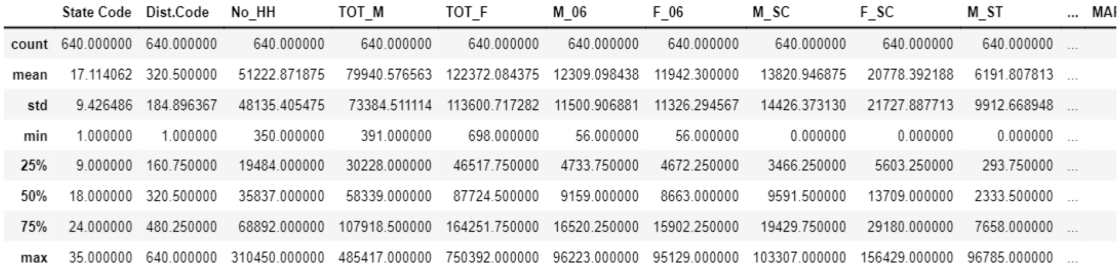
	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180	
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123	
2	1	3	Jammu & Kashmir	Leh(Ladakh)	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

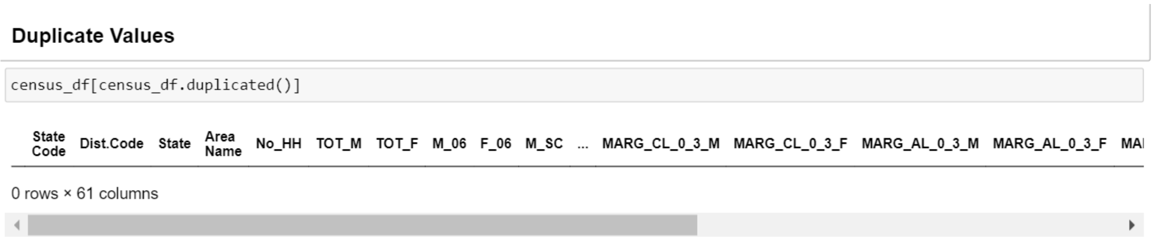


	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL
635	34	636	Puducherry	Mahe	3333	8154	11781	1146	1203	21	...	32	47	0	
636	34	637	Puducherry	Karaikal	10612	12346	21691	1544	1533	2234	...	155	337	3	
637	35	638	Andaman & Nicobar Island	Nicobars	1275	1549	2630	227	225	0	...	104	134	9	
638	35	639	Andaman & Nicobar Island	North & Middle Andaman	3762	5200	8012	723	664	0	...	136	172	24	
639	35	640	Andaman & Nicobar Island	South Andaman	7975	11977	18049	1470	1358	0	...	173	122	6	

5 rows × 61 columns



	State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MAI
count	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	...	
mean	17.114062	320.500000	51222.871875	79940.576563	122372.084375	12309.098438	11942.300000	13820.946875	20778.392188	6191.807813	...	
std	9.426486	184.896367	48135.405475	73384.511114	113600.717282	11500.906881	11326.294567	14426.373130	21727.887713	9912.668948	...	
min	1.000000	1.000000	350.000000	391.000000	698.000000	56.000000	56.000000	0.000000	0.000000	0.000000	...	
25%	9.000000	160.750000	19484.000000	30228.000000	46517.750000	4733.750000	4672.250000	3466.250000	5603.250000	293.750000	...	
50%	18.000000	320.500000	35837.000000	58339.000000	87724.500000	9159.000000	8663.000000	9591.500000	13709.000000	2333.500000	...	
75%	24.000000	480.250000	68892.000000	107918.500000	164251.750000	16520.250000	15902.250000	19429.750000	29180.000000	7658.000000	...	
max	35.000000	640.000000	310450.000000	485417.000000	750392.000000	96223.000000	95129.000000	103307.000000	156429.000000	96785.000000	...	



Duplicate Values

```
census_df[census_df.duplicated()]
```

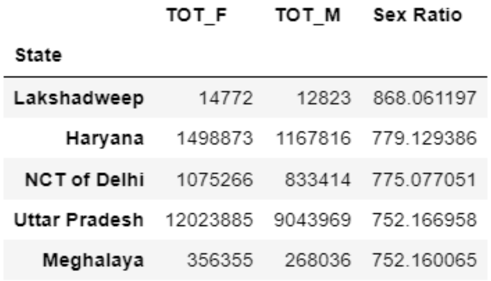
State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MAI

0 rows × 61 columns

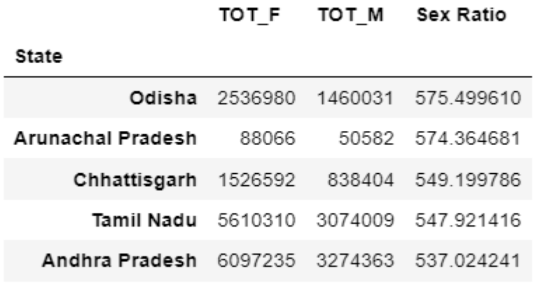
- The data consists of **640 rows** and **61 columns.**
- There is a total of **59 numeric columns** and **2 categoric columns.**
- The dataset has **no null and duplicate values.**

B. 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?

Ans:

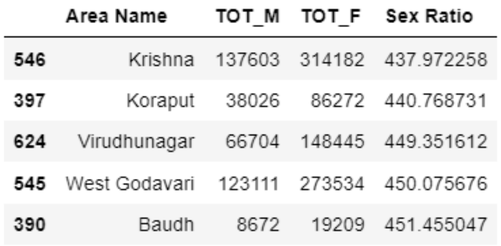


State	TOT_F	TOT_M	Sex Ratio
Lakshadweep	14772	12823	868.061197
Haryana	1498873	1167816	779.129386
NCT of Delhi	1075266	833414	775.077051
Uttar Pradesh	12023885	9043969	752.166958
Meghalaya	356355	268036	752.160065

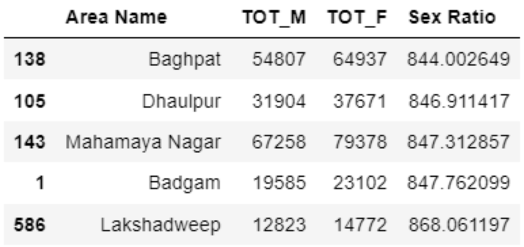


State	TOT_F	TOT_M	Sex Ratio
Odisha	2536980	1460031	575.499610
Arunachal Pradesh	88066	50582	574.364681
Chhattisgarh	1526592	838404	549.199786
Tamil Nadu	5610310	3074009	547.921416
Andhra Pradesh	6097235	3274363	537.024241

i) **Lakshadweep** has the **highest sex ratio** with **868 males for every 1000 females** while **Andhra Pradesh** has the **lowest sex ratio** with **537 males for every 1000 females.**

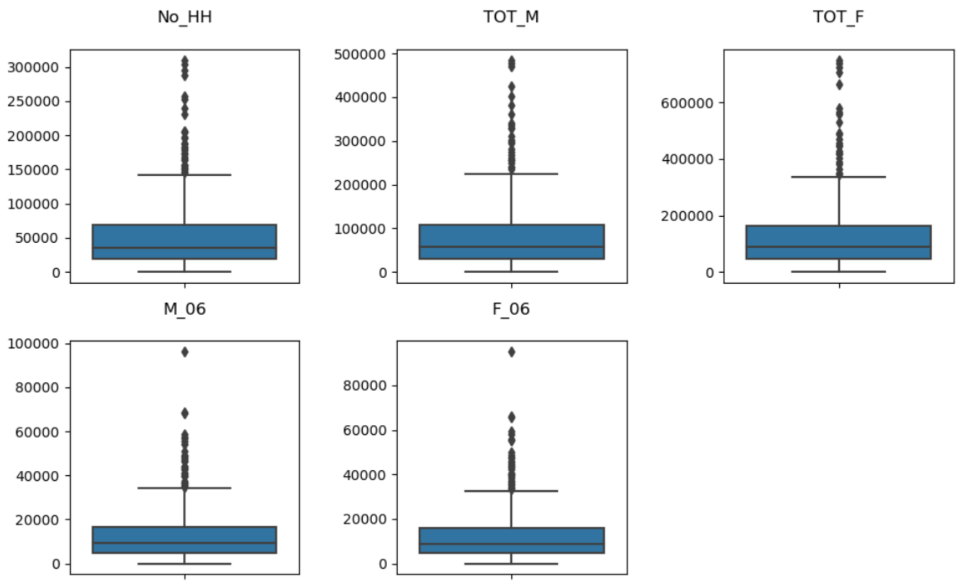


	Area Name	TOT_M	TOT_F	Sex Ratio
546	Krishna	137603	314182	437.972258
397	Koraput	38026	86272	440.768731
624	Virudhunagar	66704	148445	449.351612
545	West Godavari	123111	273534	450.075676
390	Baudh	8672	19209	451.455047



	Area Name	TOT_M	TOT_F	Sex Ratio
138	Baghpat	54807	64937	844.002649
105	Dhaulpur	31904	37671	846.911417
143	Mahamaya Nagar	67258	79378	847.312857
1	Badgam	19585	23102	847.762099
586	Lakshadweep	12823	14772	868.061197

ii) **Lakshadweep** has the **highest sex ratio** followed by the **Bagdam district** with **847 males for every 1000 females** while the **Krishna district** has the **lowest sex ratio with 437 males for every 1000 females.**



The following inferences can be gathered from the dataset:

- There is a total of **59 numeric fields** in the data
- The **average male population is 79,940** while the **average female population is 1,22,372**
- **Uttar Pradesh** has the **highest male and female populations**
- **Dadara and Nagar Havelli** has the **lowest male and female populations**
- The **male population ranges from 391 to 4,85,417** while the **female population ranges from 640 to 7,50,392**
- The **number of households ranges from 350 to 3,10,000**
- The **male population** in the **age group of 0-6 years** lies between **640 to 96,223**
- The **female population** in the **age group of 0-6 years** lies between **640 to 95,129**

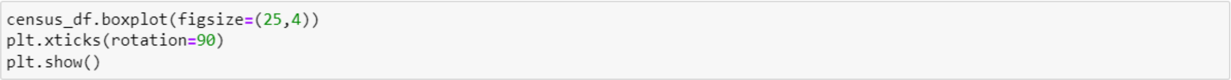
C. **We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?**

Ans: Outlier treatment is not necessary here as the variation in the population sizes is caused due to a wide variety of factors in the dataset. Treating the outliers may result in inaccuracy when determining the principal components using PCA as the effects of these factors would be nullified causing it to not be accounted for. Hence outlier treatment is not required here.

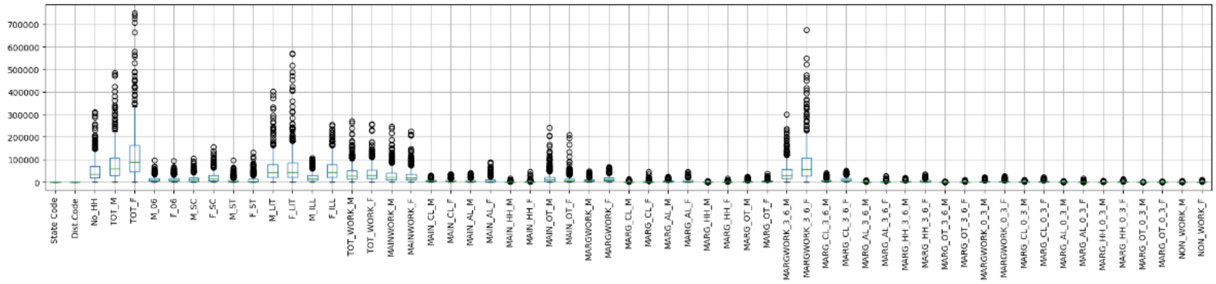
D. **Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.**

Ans:

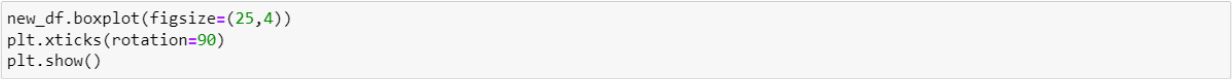
Unscaled Data



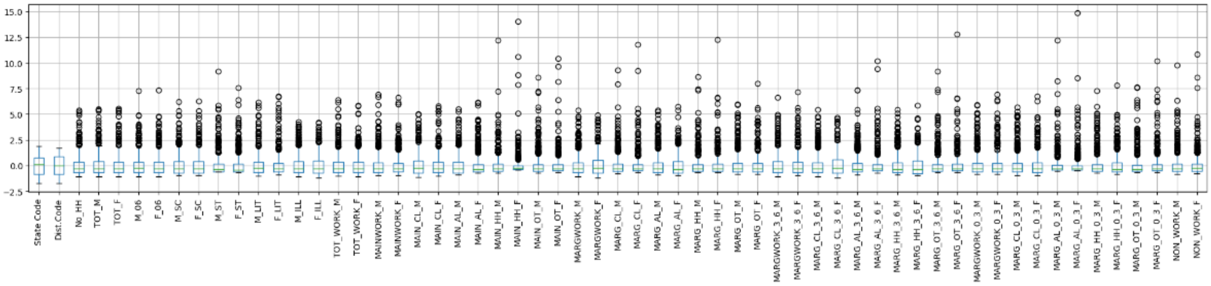
```
census_df.boxplot(figsize=(25,4))
plt.xticks(rotation=90)
plt.show()
```



Scaled data



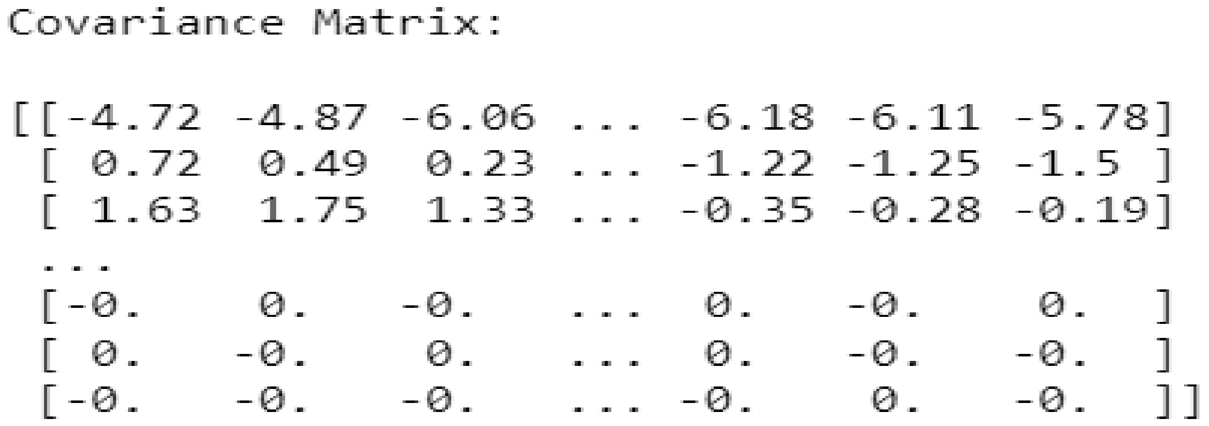
```
new_df.boxplot(figsize=(25,4))
plt.xticks(rotation=90)
plt.show()
```



It can be observed that scaling has changed the outlier distribution for the variables. Earlier, the outlier distribution was varied for different variables not to mention the difference in their population ranges. Scaling has standardized both the outlier distribution along with the data ranges.

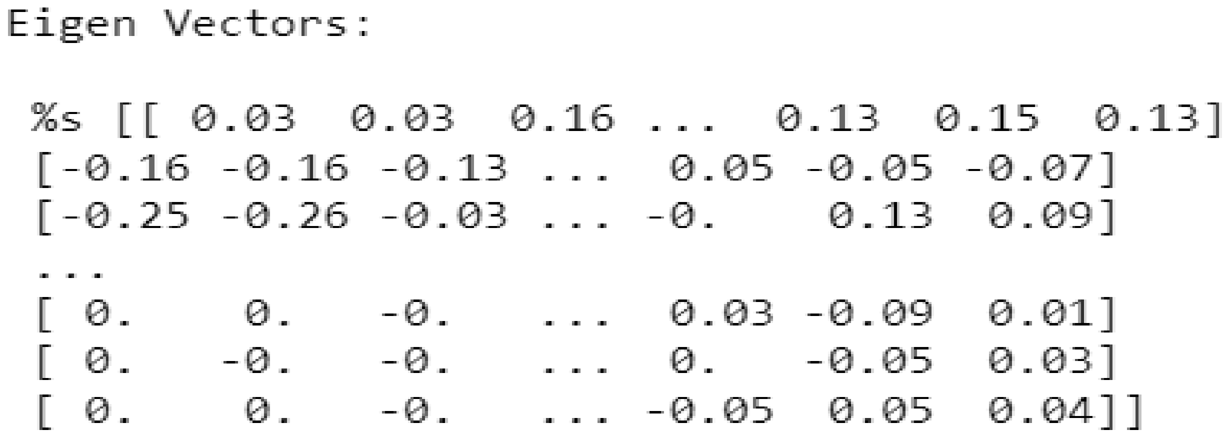
E. Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.

Ans:



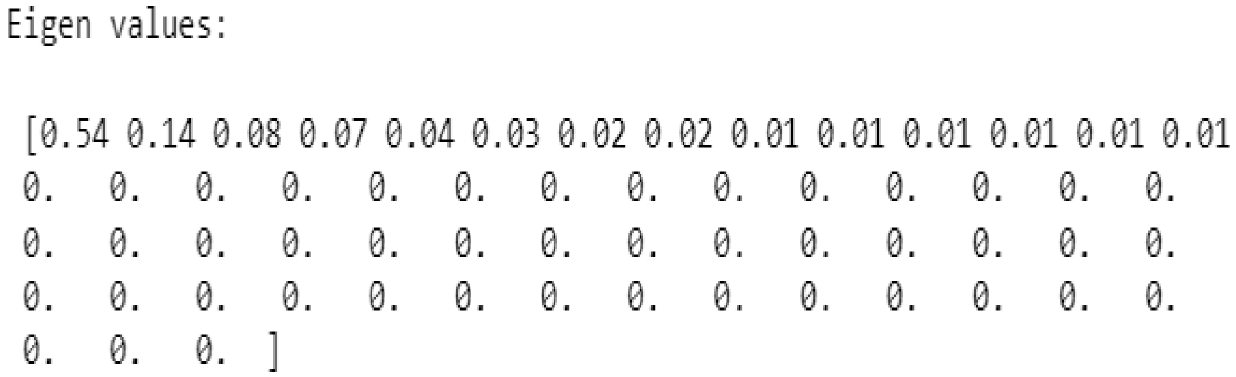
```
Covariance Matrix:

[[-4.72 -4.87 -6.06 ... -6.18 -6.11 -5.78]
 [ 0.72  0.49  0.23 ... -1.22 -1.25 -1.5 ]
 [ 1.63  1.75  1.33 ... -0.35 -0.28 -0.19]
 ...
 [-0.    0.   -0.   ...  0.   -0.    0.  ]
 [ 0.   -0.    0.   ...  0.   -0.   -0.  ]
 [-0.   -0.   -0.   ... -0.    0.   -0.  ]]
```



```
Eigen Vectors:

 %s [[ 0.03  0.03  0.16 ...  0.13  0.15  0.13]
 [-0.16 -0.16 -0.13 ...  0.05 -0.05 -0.07]
 [-0.25 -0.26 -0.03 ... -0.    0.13  0.09]
 ...
 [ 0.    0.   -0.   ...  0.03 -0.09  0.01]
 [ 0.   -0.   -0.   ...  0.   -0.05  0.03]
 [ 0.    0.   -0.   ... -0.05  0.05  0.04]]
```

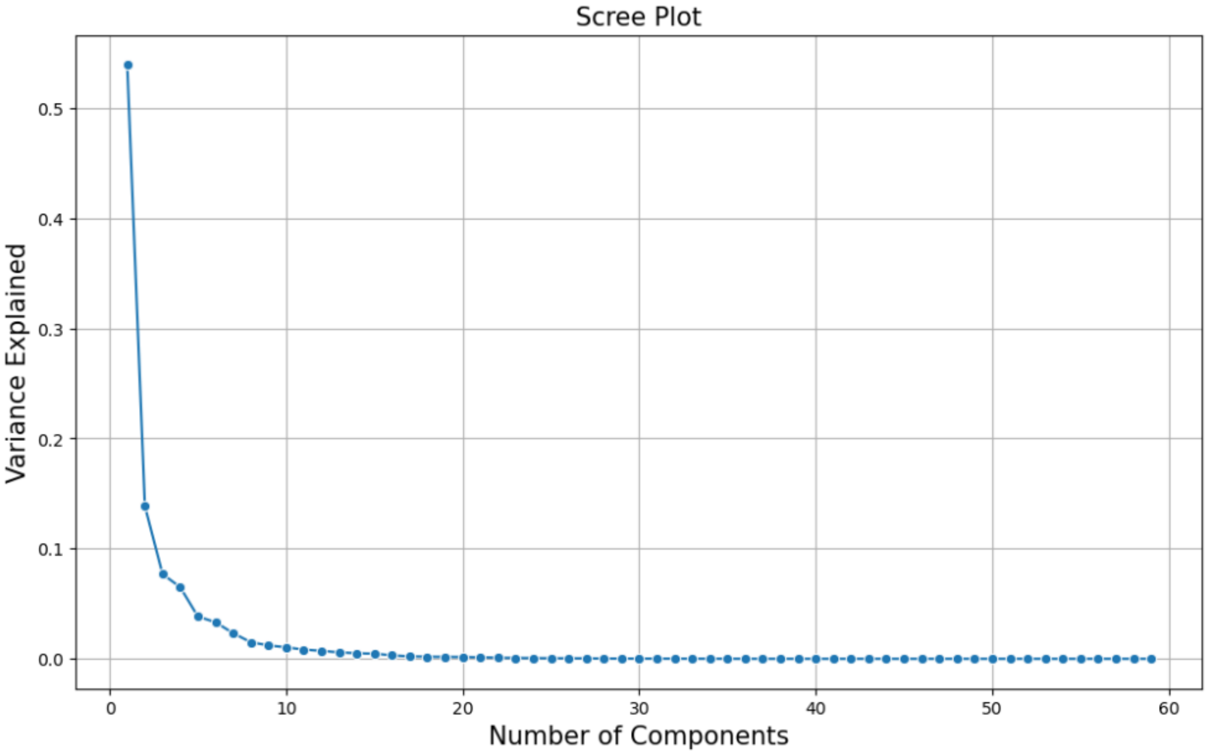


```
Eigen values:

 [0.54 0.14 0.08 0.07 0.04 0.03 0.02 0.02 0.01 0.01 0.01 0.01 0.01 0.01
 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.   0.  ]
```

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

Ans:



The number of components can be decided upon the explained variance. It can be observed from the cumulative variance values and from the scree plot that at least 90% of the explained variance is captured by having 7 principal components.