DATA MINING

CONTENT

CLUSTERING

ANALYSIS OF DIGITAL ADS DATA

CLUSTERING :

It is an unsupervised learning Technique trying to identify groups of similar objects that are highly dissimilar with other objects.

OVERVIEW OF THE REPORT :

To analyse the digital ad data and to identify trends in clicks, spend, revenue, CPM, CTR, & CPC based on Device Type.

1) DATA ANALYSIS :

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

- The dataset contains 23099 Rows and 19 Columns.
- DUPLICATES :

  No duplicates are present.
- NULL VALUES :

  There are

4736 Null values in CPM

4736 Null values in CTR

4736 Null values in CPC

- There are 13 Numerical variables and 6 Categorical variables.

2) Treat missing values in CPC, CTR and CPM using the formula given :

FORMULAS GIVEN :

CTR = Clicks/Impresssions * 100

CPM = Spend/Impressions * 1000

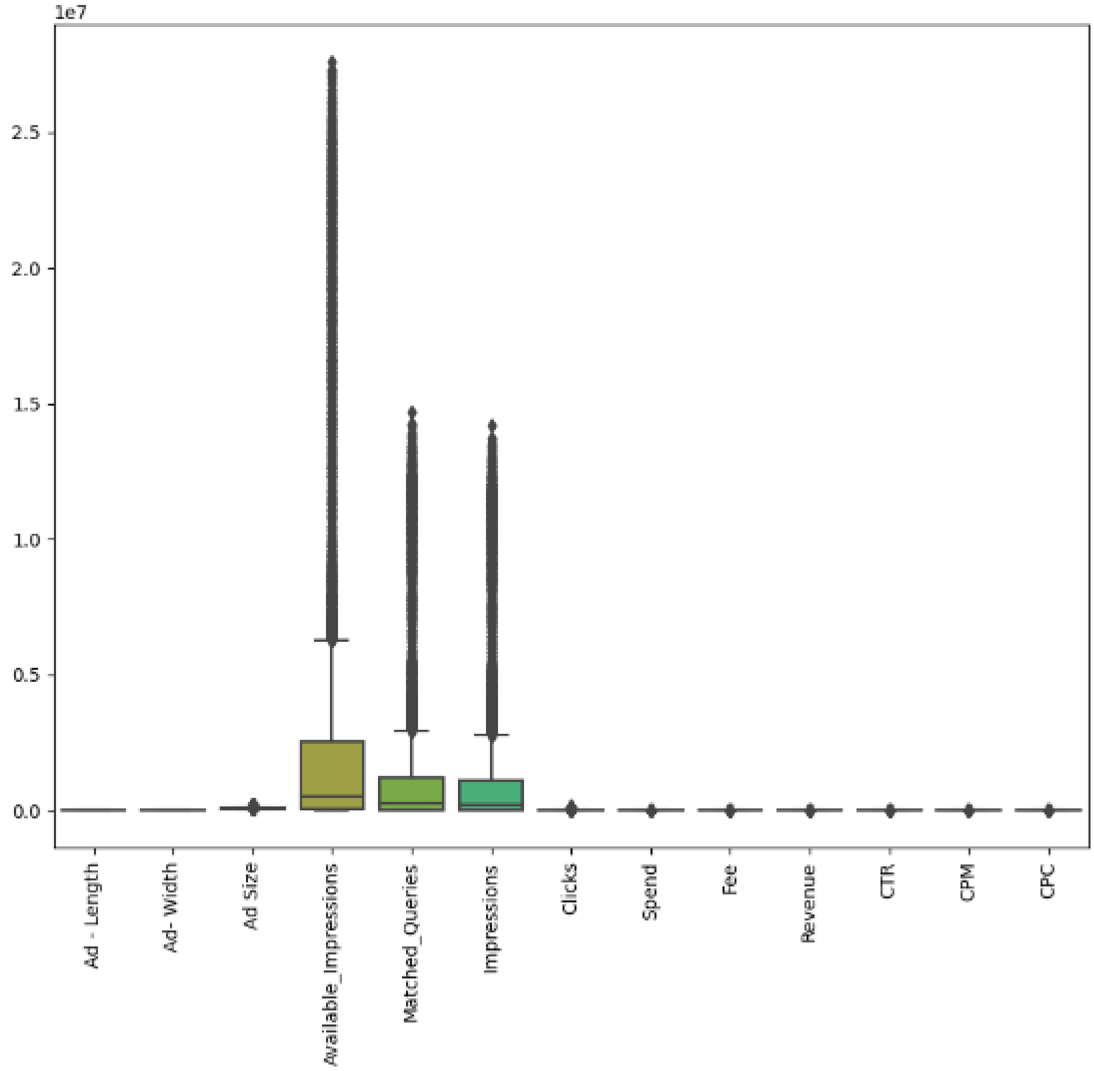
CPC = Spend/Clicks

The Missing values in dataset has to be filled using the above formula.

3) Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an

analyst your judgement may be different from another analyst).

Outliers are present in following features.

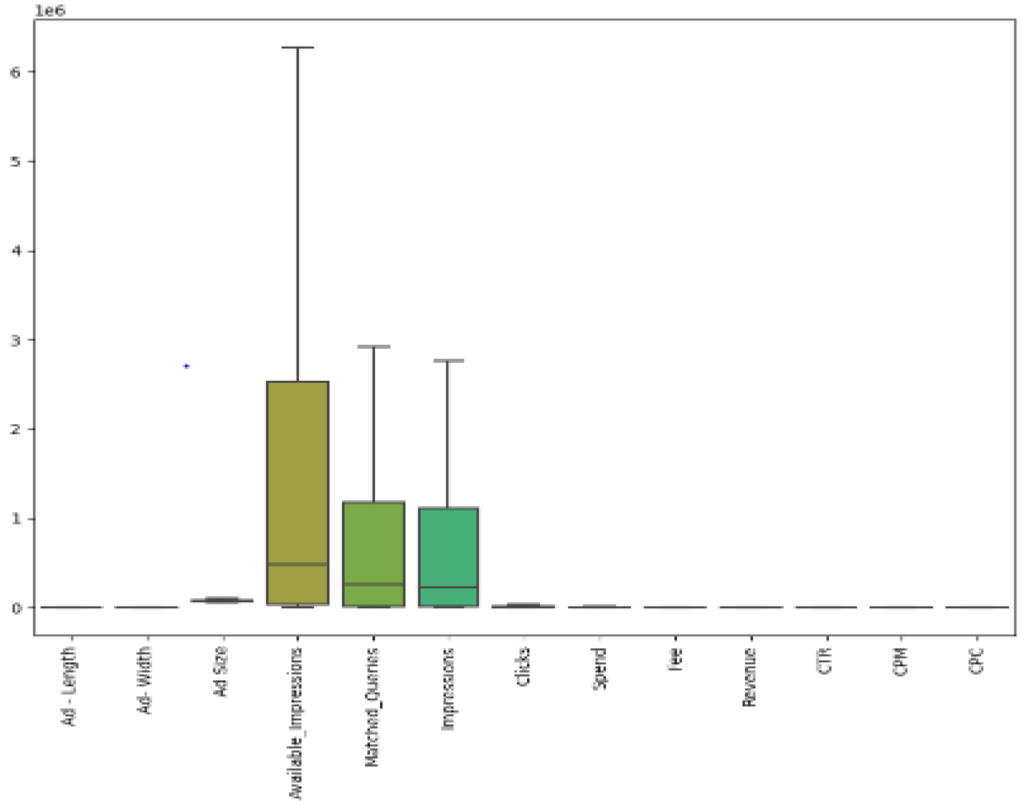


- ➢ Outlier detection is an important data analysis task
- ➢ Treating the outliers from clusters can improve the clustering accuracy. So we have to treat outlier before applying k means algorithm.

- Here we have to use BOXPLOT(QUARTILE) Method

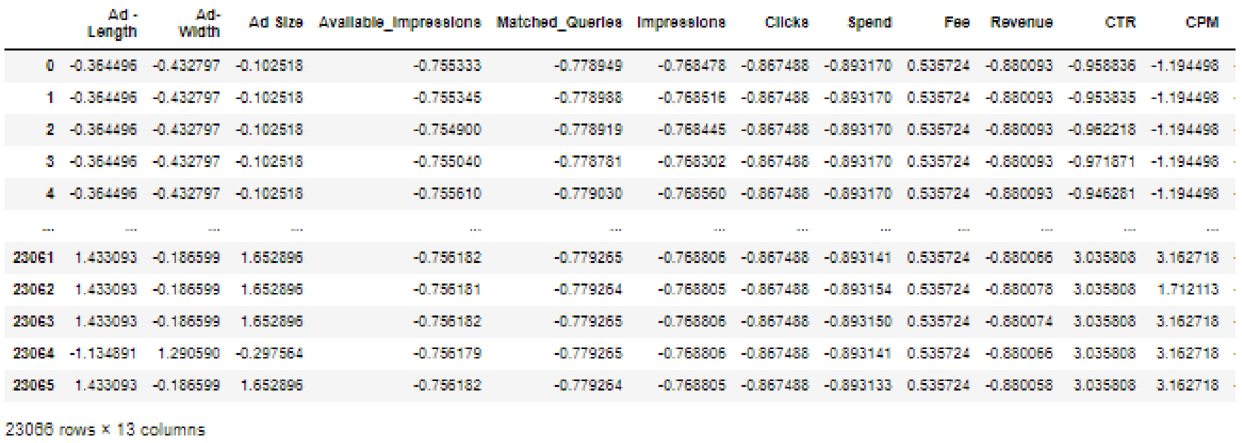
  to treat outliers.

BOXPLOT after treating outliers :



Thus using BOXPLOT Method we can treat outliers to get high accuracy while performing K-Clustering Algorithm.

4) Perform z-score scaling and discuss how it affects the speed of the algorithm.



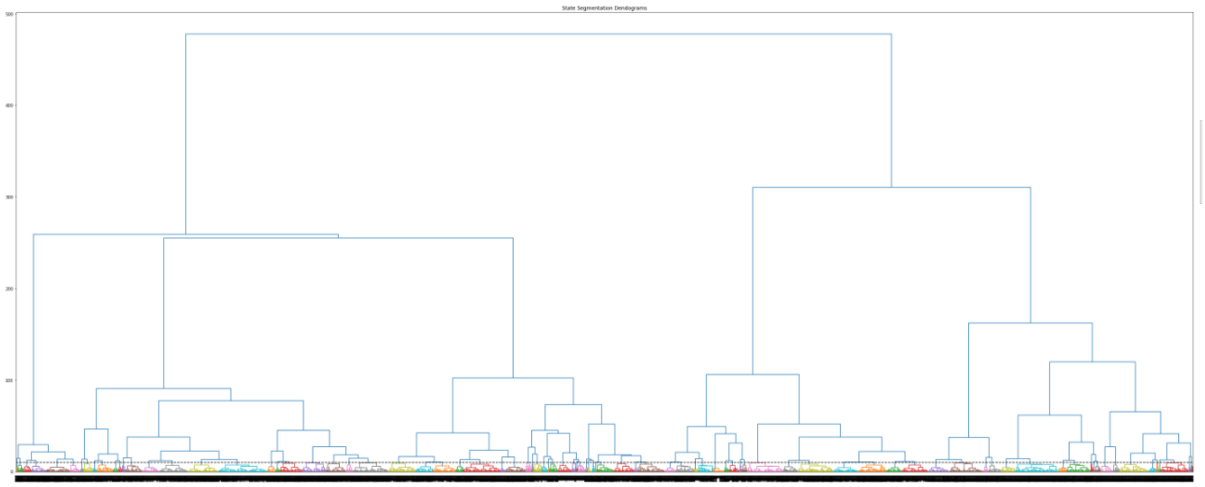
	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM
0	-0.364496	-0.432797	-0.102518	-0.755333	-0.778949	-0.768478	-0.867488	-0.893170	0.535724	-0.880093	-0.958836	-1.194498
1	-0.364496	-0.432797	-0.102518	-0.755345	-0.778988	-0.768516	-0.867488	-0.893170	0.535724	-0.880093	-0.953835	-1.194498
2	-0.364496	-0.432797	-0.102518	-0.754900	-0.778919	-0.768445	-0.867488	-0.893170	0.535724	-0.880093	-0.962218	-1.194498
3	-0.364496	-0.432797	-0.102518	-0.755040	-0.778781	-0.768302	-0.867488	-0.893170	0.535724	-0.880093	-0.971871	-1.194498
4	-0.364496	-0.432797	-0.102518	-0.755610	-0.779030	-0.768560	-0.867488	-0.893170	0.535724	-0.880093	-0.946281	-1.194498
...	...	...	...	...	...	...	...	...	...	...	...	...
23061	1.433093	-0.186599	1.652896	-0.756182	-0.779265	-0.768806	-0.867488	-0.893141	0.535724	-0.880066	3.035808	3.162718
23062	1.433093	-0.186599	1.652896	-0.756181	-0.779264	-0.768805	-0.867488	-0.893154	0.535724	-0.880078	3.035808	1.712113
23063	1.433093	-0.186599	1.652896	-0.756182	-0.779265	-0.768806	-0.867488	-0.893150	0.535724	-0.880074	3.035808	3.162718
23064	-1.134891	1.290590	-0.297564	-0.756179	-0.779265	-0.768806	-0.867488	-0.893141	0.535724	-0.880066	3.035808	3.162718
23065	1.433093	-0.186599	1.652896	-0.756182	-0.779264	-0.768805	-0.867488	-0.893133	0.535724	-0.880058	3.035808	3.162718

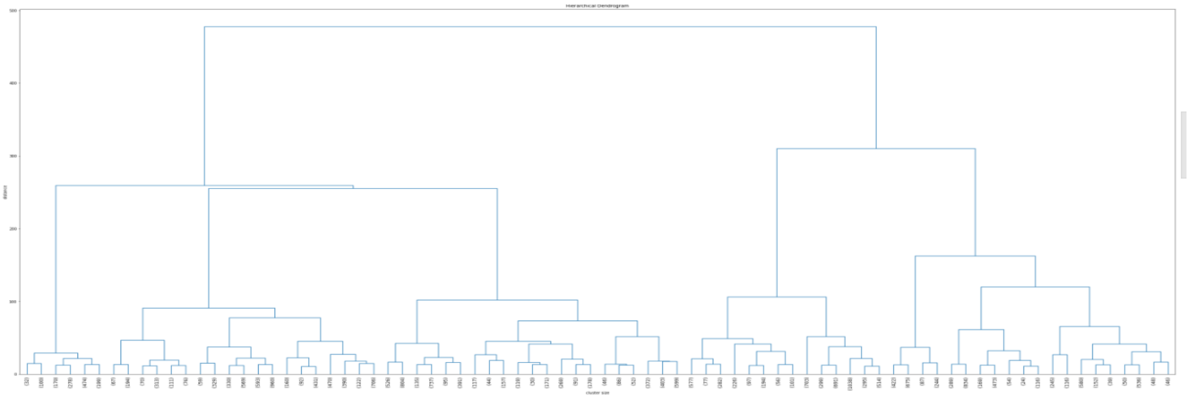
23066 rows × 13 columns

- ➢ Z-score is applied to the sorted data points as a measure to improve the selection of initial clusters.
- ➢ Because of finding initial cluster centers , it ensures
  - High accuracy
  - Reduced clustering error
  - Less computation time and
  - Less number of iterations.

5) Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.

Dendrogram :

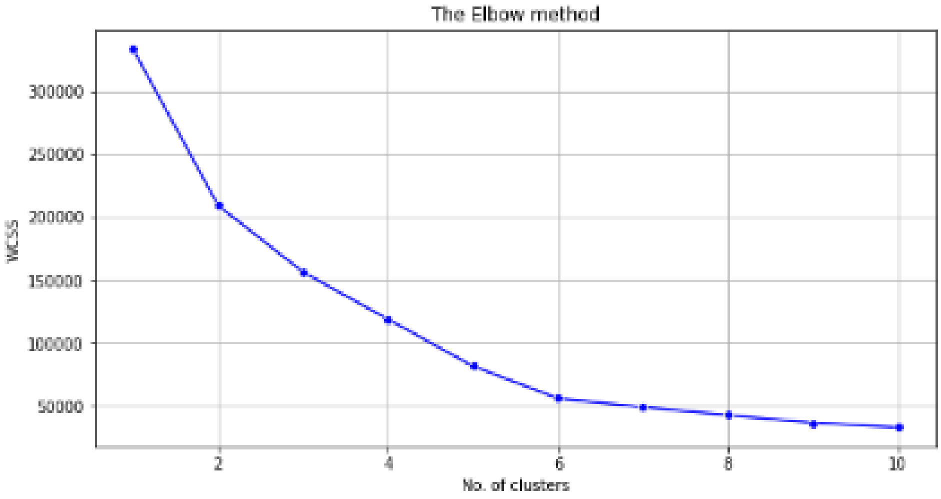




From Hierarchical clustering , we can say that the **optimum number of clusters will be 5.**

6) Clustering: Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.

Elbow plot :



FROM THE ABOVE WSS PLOT WE CAN SAY THAT THE **OPTIMUM NUMBER OF CLUSTERS ARE 5** BECAUSE AFTER 5 THE PLOT WILL BE NARROW i.e) NO MUCH DIFFERENCE EXISTS.

7) Print silhouette scores for up to 10 clusters and identify optimum number of clusters.

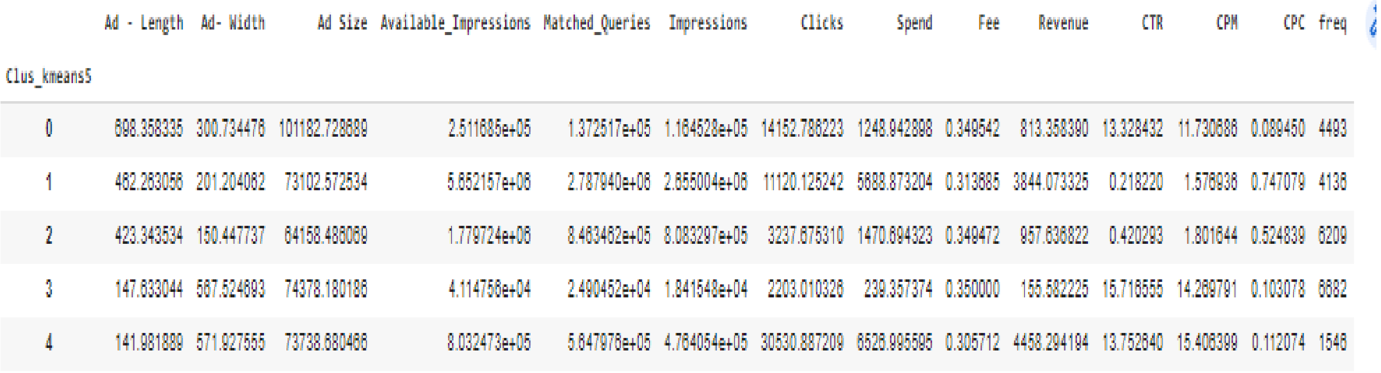
Silhouette scores

No. of clusters	Silhouette scores
2	0.4241248441233028
3	0.4284360206300488
4	0.49517949988734966
5	0.5947980661900538
6	0.5999254403934559
7	0.6363174847771058
8	0.6288118527264385

9	0.6481558775415522
10	0.6542947087856206

So we can say that the optimum number of clusters will be 5.

8) Clustering: Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type].



	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	freq
Clus_kmeans5														
0	698.358335	300.734476	101182.728689	2.511685e+05	1.372517e+05	1.164528e+05	14152.786223	1248.942898	0.349542	813.358390	13.328432	11.730686	0.089450	4493
1	462.263056	201.204082	73102.572534	5.652157e+06	2.787940e+06	2.655004e+06	11120.125242	5688.873204	0.313685	3844.073325	0.218220	1.576936	0.747079	4136
2	423.343534	150.447737	64158.486069	1.779724e+06	8.463462e+05	8.083297e+05	3237.675310	1470.694323	0.349472	957.636822	0.420293	1.801644	0.524839	6209
3	147.633044	567.524693	74378.180186	4.114756e+04	2.490452e+04	1.841548e+04	2203.010326	239.357374	0.350000	155.582225	15.716555	14.269791	0.103078	6682
4	141.981889	571.927555	73738.680466	8.032473e+05	5.647978e+05	4.764054e+05	30530.887209	6526.995595	0.305712	4458.294194	13.752840	15.406399	0.112074	1546

TRENDS :

- **CLUSTER_0 :**

- Larger Ad size
- More number of clicks
- Lesser amount spending
- Getting low revenue
- Medium amount of CTR , CPM
- Lower CPC

➢ **CLUSTER_1:**

- Smaller Ad size
- Medium number of clicks
- Larger amount spending
- Getting Higher Revenue
- Very Low CTR,CPM
- Higher CPC

➢ **CLUSTER_2 :**

- Very smaller Ad size
- Lesser number of clicks
- Average amount spending
- Average Revenue
- Low CTR,CPM
- Higher CPC

➢ **CLUSTER_3 :**

- Larger Ad size
- Lesser number of clicks
- Lesser amount spending
- Getting Very low revenue
- Higher CTR,CPM
- Lower CPC

- **CLUSTER_4 :**

  - Medium Ad size
  - Larger number of clicks
  - Large amount spending
  - Higher Revenue
  - Higher CTR,CPM
  - Average CPC

9) Conclude the project by providing summary of your learnings:

- Both Desktop and Mobile (Device type) uses Video and Web as platform

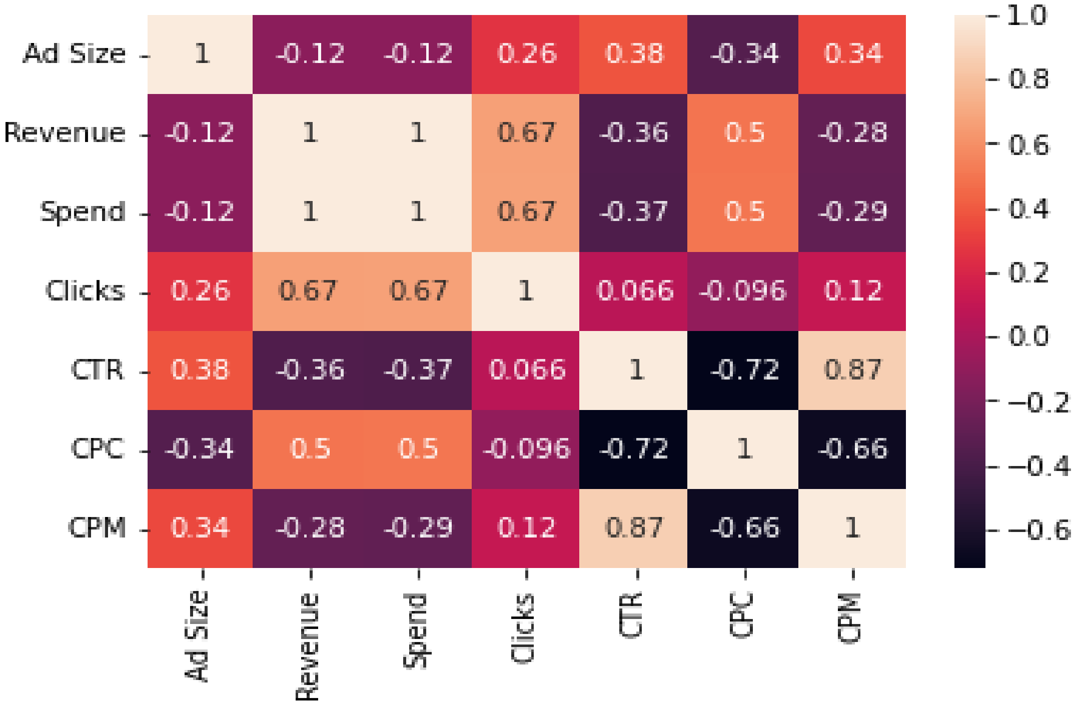
- App can be uses as a platform only through Mobile

- When Ad size is small , it yields Higher Revenue.

- When number of clicks is high , Revenue will be high

- When spending is more , Revenue will be more
- When CPC high , Revenue also high
- When number of clicks high , CPC is low
- When spending increases , both CPM AND CPC also increases
- When number of clicks increases , CTR also increases.

- When Ad size increases , Clicks,CPC ,CPM & CTR also increases.
- Positive correlation exits between Revenue,Spend,CPC,Clicks.



	Ad Size	Revenue	Spend	Clicks	CTR	CPC	CPM
Ad Size	1	-0.12	-0.12	0.26	0.38	-0.34	0.34
Revenue	-0.12	1	1	0.67	-0.36	0.5	-0.28
Spend	-0.12	1	1	0.67	-0.37	0.5	-0.29
Clicks	0.26	0.67	0.67	1	0.066	-0.096	0.12
CTR	0.38	-0.36	-0.37	0.066	1	-0.72	0.87
CPC	-0.34	0.5	0.5	-0.096	-0.72	1	-0.66
CPM	0.34	-0.28	-0.29	0.12	0.87	-0.66	1

PRINCIPAL COMPONENT ANALYSIS

POPULATION CENSUS ANALYSIS 2011

PCA :

Principal component analysis (**PCA**) is a popular unsupervised learning technique for analyzing large datasets containing a high number of dimensions/features per observation, increasing the interpretability of data while preserving the maximum amount of information, and enabling the visualization of multidimensional data. Formally, PCA is a statistical technique for reducing the dimensionality of a dataset.

OBJECTIVE OF PCA :

- DIMENSION REDUCTION
- PATTERN RECOGNITION
- RESOLVE MULTI-COLLINEARITY

OVERVIEW OF THE REPORT :

To analyse the population census data with respect to the features provided such as Literacy Rate , Labour Force , Gender Ratio , Population of scheduled peoples etc. by applying PCA Technique .

1) DATA ANALYSIS :

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

- The dataset contains 641 Rows & 61 Columns .

- DUPLICATES :

  NO duplicates present.

- NULL VALUES :

  Absence of Null values in the Data set.

- SUMMARY :

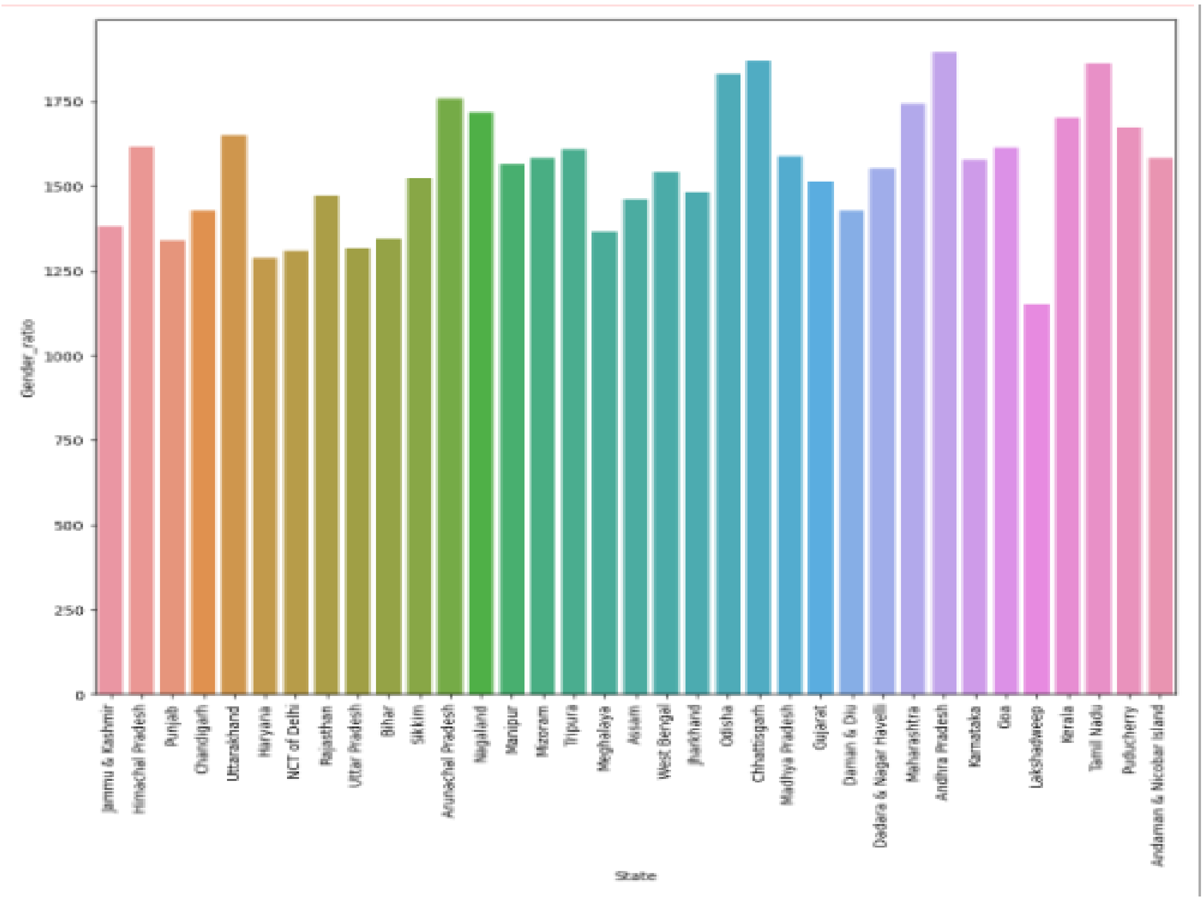
  Presence of 59 Numerical Variables and 2 Categorical Variables.

2) EDA :

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

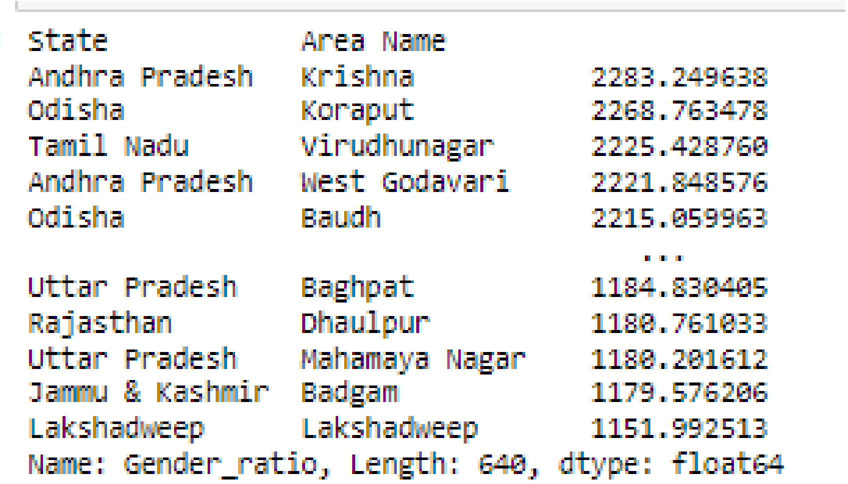
GENDER RATIO defined as number of females per 1000 males in a population

GENDER RATIO = NO. OF FEMALES / NO. OF MALES * 1000



STATE	GENDER RATIO
ANDRA PRADESH (HIGHEST)	1895 OUT OF 1000 MALES
LAKSHWADEEP (LOWEST)	1152 OUT OF 1000 MALES

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



```
State            Area Name
Andhra Pradesh   Krishna           2283.249638
Odisha           Koraput           2268.763478
Tamil Nadu       Virudhunagar      2225.428760
Andhra Pradesh   West Godavari     2221.848576
Odisha           Baudh             2215.059963
                                       ...
Uttar Pradesh    Baghpat           1184.830405
Rajasthan        Dhaulpur          1180.761033
Uttar Pradesh    Mahamaya Nagar    1180.201612
Jammu & Kashmir  Badgam            1179.576206
Lakshadweep      Lakshadweep       1151.992513
Name: Gender_ratio, Length: 640, dtype: float64
```

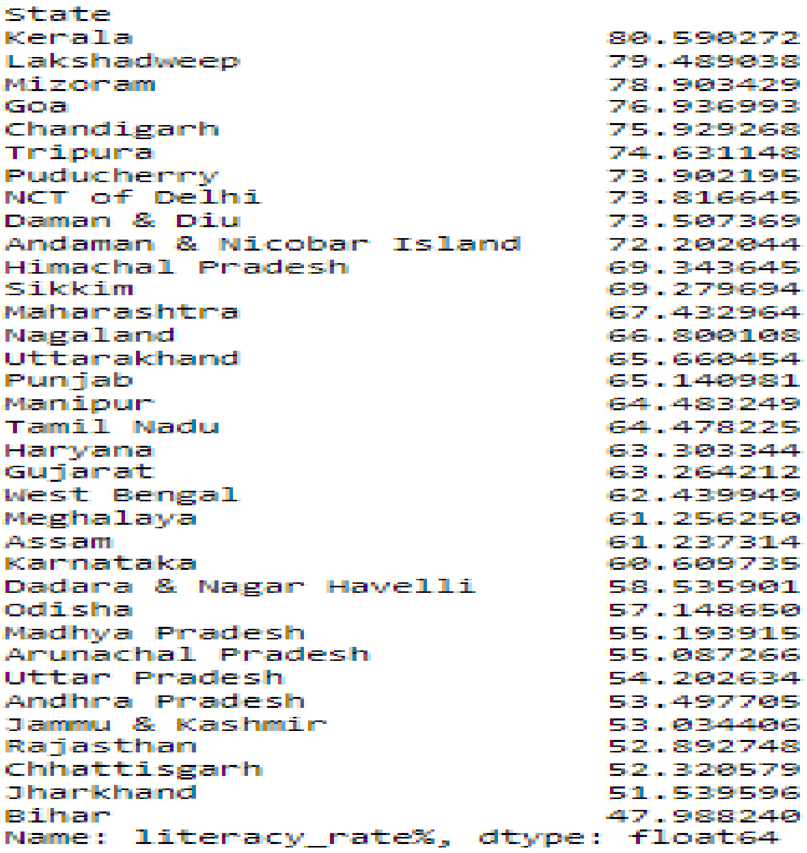
- **KRISHNA District in ANDRA PRADESH** has highest Gender Ratio .

- **LAKSHADWEEP District in LAKSHADWEEP Island** has lowest Gender Ratio.

iii) **Which state / UT has highest & lowest literacy rate ?**

FORMULA :

LITERACY RATE % = (TOTAL MALE LITERATES +TOTAL FEMALE LITERATES) / TOTAL POPULATION



```
State
Kerala                      80.590272
Lakshadweep                 79.489038
Mizoram                     78.903429
Goa                         76.936993
Chandigarh                  75.929268
Tripura                     74.631148
Puducherry                  73.902195
NCT of Delhi                73.816645
Daman & Diu                 73.507369
Andaman & Nicobar Island    72.202044
Himachal Pradesh            69.343645
Sikkim                      69.279694
Maharashtra                 67.432964
Nagaland                    66.800108
Uttarakhand                 65.660454
Punjab                      65.140981
Manipur                     64.483249
Tamil Nadu                  64.478225
Haryana                     63.303344
Gujarat                     63.264212
West Bengal                 62.439949
Meghalaya                   61.256250
Assam                       61.237314
Karnataka                   60.609735
Dadara & Nagar Havelli      58.535901
Odisha                      57.148650
Madhya Pradesh              55.193915
Arunachal Pradesh           55.087266
Uttar Pradesh               54.202634
Andhra Pradesh              53.497705
Jammu & Kashmir             53.034406
Rajasthan                   52.892748
Chhattisgarh                52.320579
Jharkhand                   51.539596
Bihar                       47.988240
Name: literacy_rate%, dtype: float64
```

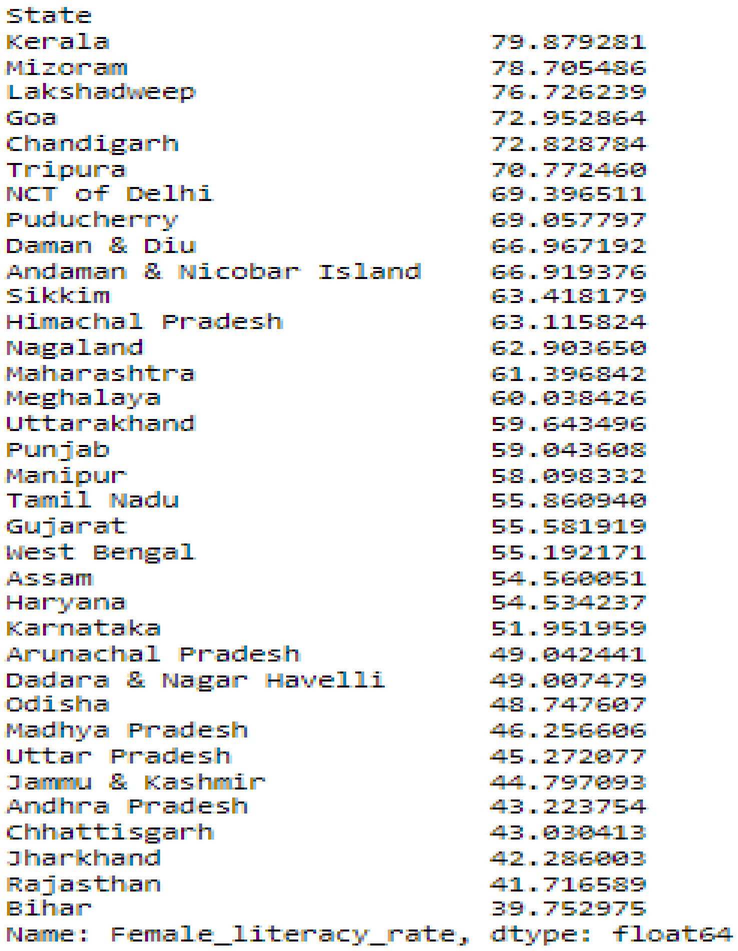
- **KERALA** has highest literacy rate of 80.59 %

- **BIHAR** has lowest literacy rate of 47.98%

iv) Which state / UT has highest & lowest female literacy rate ?

FORMULA:

FEMALE LITERACY RATE = TOTAL FEMALE LITERATES / TOTAL FEMALE POPULATION * 100



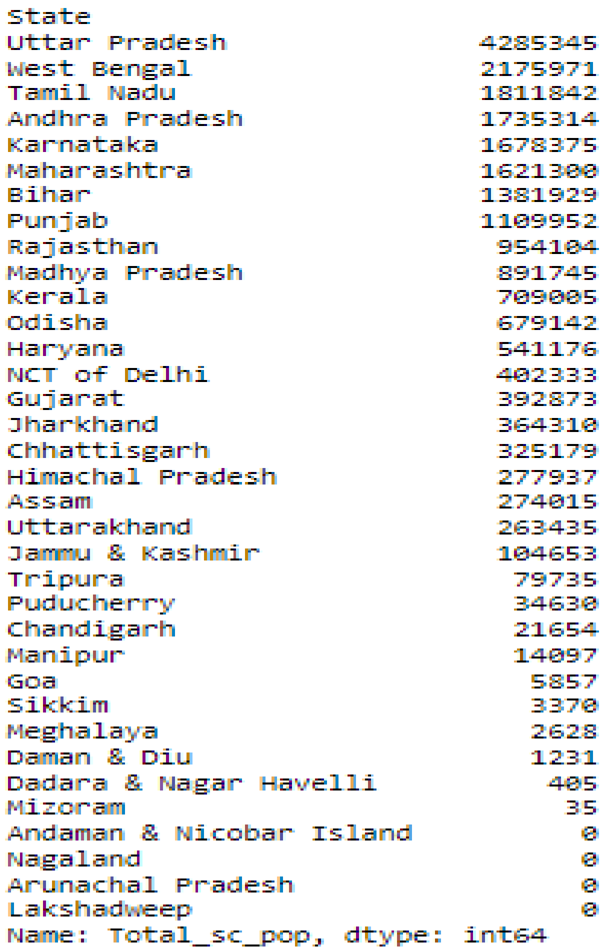
```
State
Kerala                         79.879281
Mizoram                        78.705486
Lakshadweep                    76.726239
Goa                            72.952864
Chandigarh                     72.828784
Tripura                        70.772460
NCT of Delhi                   69.396511
Puducherry                     69.057797
Daman & Diu                    66.967192
Andaman & Nicobar Island       66.919376
Sikkim                         63.418179
Himachal Pradesh               63.115824
Nagaland                       62.903650
Maharashtra                    61.396842
Meghalaya                      60.038426
Uttarakhand                    59.643496
Punjab                         59.043608
Manipur                        58.098332
Tamil Nadu                     55.860940
Gujarat                        55.581919
West Bengal                    55.192171
Assam                          54.560051
Haryana                        54.534237
Karnataka                      51.951959
Arunachal Pradesh              49.042441
Dadara & Nagar Havelli         49.007479
Odisha                         48.747607
Madhya Pradesh                 46.256606
Uttar Pradesh                  45.272077
Jammu & Kashmir                44.797093
Andhra Pradesh                 43.223754
Chhattisgarh                   43.030413
Jharkhand                      42.286003
Rajasthan                      41.716589
Bihar                          39.752975
Name: Female_literacy_rate, dtype: float64
```

Among the given states, 79.87 % of female population in KERALA are literates whereas only 39.75 % of female population are literates in BIHAR.

V) Which state / UT has more scheduled caste population ?

FORMULA :

TOTAL SC POPULATION = TOTAL SC MALES + TOTAL SC FEMALES IN EACH STATE



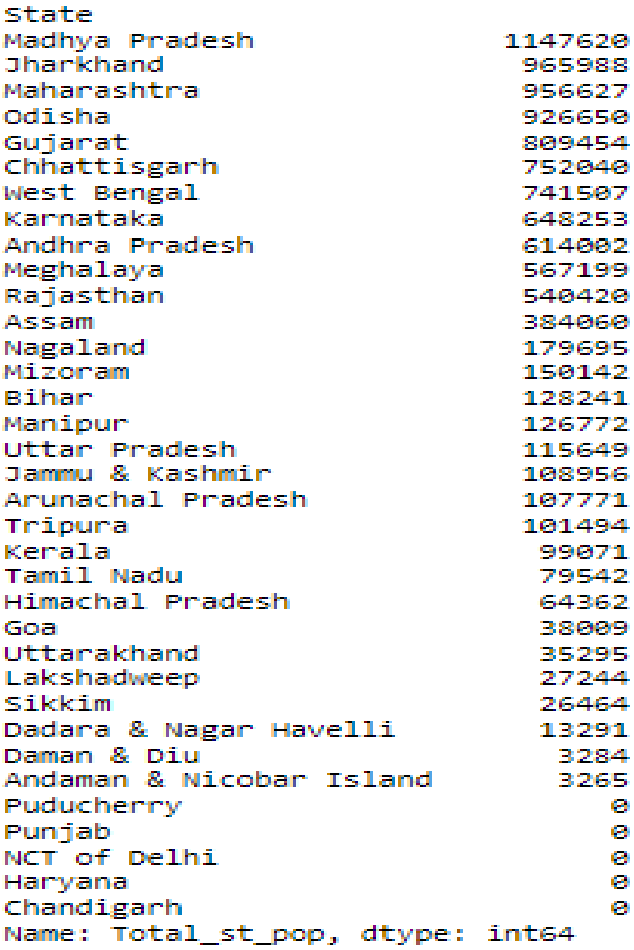
```
State
Uttar Pradesh                 4285345
West Bengal                   2175971
Tamil Nadu                    1811842
Andhra Pradesh                1735314
Karnataka                     1678375
Maharashtra                   1621300
Bihar                         1381929
Punjab                        1109952
Rajasthan                      954104
Madhya Pradesh                 891745
Kerala                         709005
Odisha                         679142
Haryana                        541176
NCT of Delhi                   402333
Gujarat                        392873
Jharkhand                      364310
Chhattisgarh                   325179
Himachal Pradesh               277937
Assam                          274015
Uttarakhand                    263435
Jammu & Kashmir                104653
Tripura                         79735
Puducherry                      34630
Chandigarh                      21654
Manipur                         14097
Goa                              5857
Sikkim                           3370
Meghalaya                        2628
Daman & Diu                      1231
Dadara & Nagar Havelli            405
Mizoram                            35
Andaman & Nicobar Island            0
Nagaland                            0
Arunachal Pradesh                   0
Lakshadweep                         0
Name: Total_sc_pop, dtype: int64
```

UTTAR PRADESH has more Scheduled caste population.

vi) Which state /UT has more scheduled tribes population ?

FORMULA :

TOTAL ST POPULATION = TOTAL ST MALES + TOTAL ST FEMALES IN EACH STATE



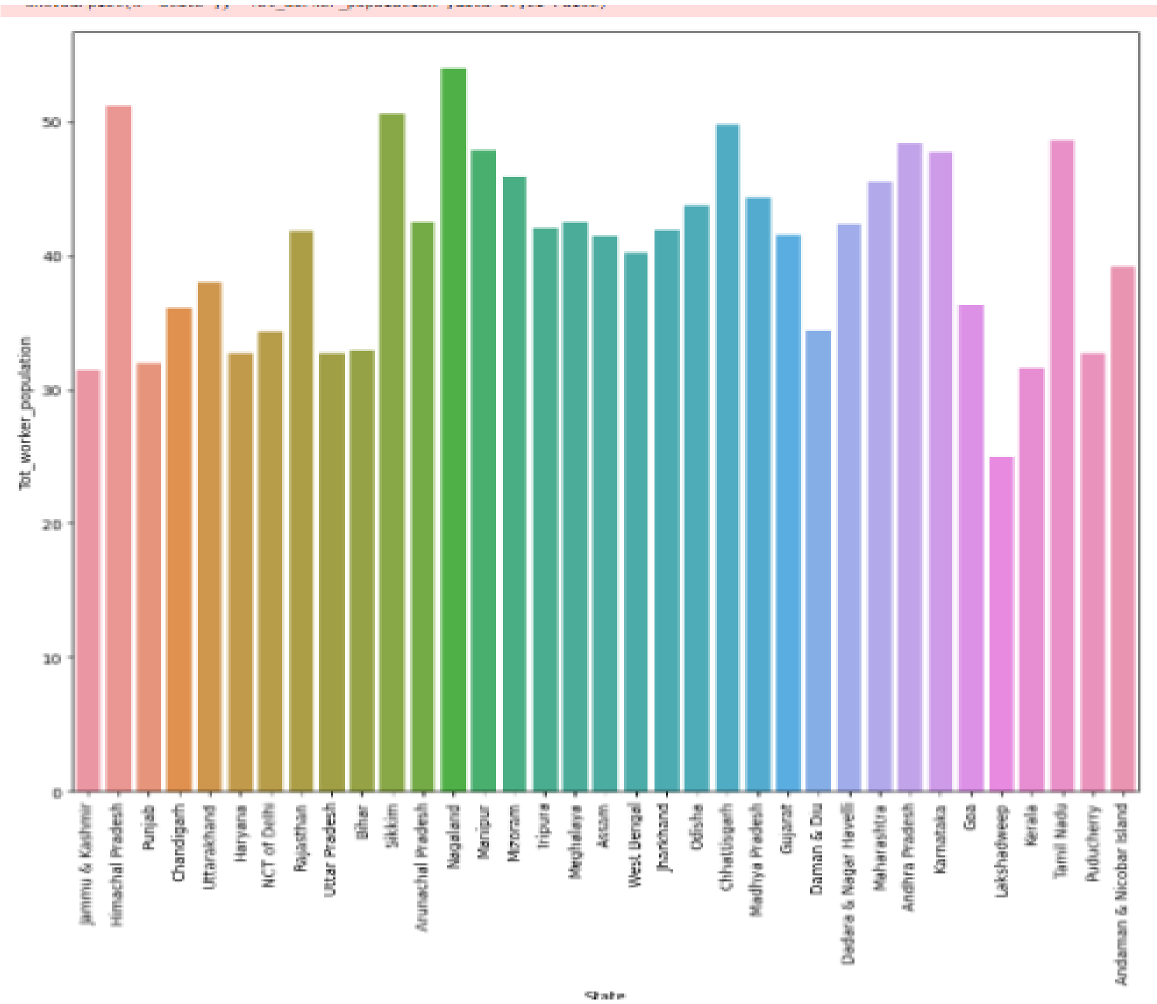
```
State
Madhya Pradesh              1147620
Jharkhand                    965988
Maharashtra                  956627
Odisha                       926650
Gujarat                      809454
Chhattisgarh                 752040
West Bengal                  741507
Karnataka                    648253
Andhra Pradesh               614002
Meghalaya                    567199
Rajasthan                    540420
Assam                        384060
Nagaland                     179695
Mizoram                      150142
Bihar                        128241
Manipur                      126772
Uttar Pradesh                115649
Jammu & Kashmir              108956
Arunachal Pradesh            107771
Tripura                      101494
Kerala                        99071
Tamil Nadu                    79542
Himachal Pradesh              64362
Goa                           38009
Uttarakhand                   35295
Lakshadweep                   27244
Sikkim                        26464
Dadara & Nagar Havelli        13291
Daman & Diu                    3284
Andaman & Nicobar Island       3265
Puducherry                        0
Punjab                            0
NCT of Delhi                      0
Haryana                           0
Chandigarh                        0
Name: Total_st_pop, dtype: int64
```

MADHYA PRADESH has more Scheduled Tribes population.

vii) Which state / UT has highest and lowest work force ?

FORMULA :

TOTAL WORKER POPULATION = (TOTAL MALE WORKER + TOTAL FEMALE WORKER) / TOTAL POPULATION OF EACH STATE

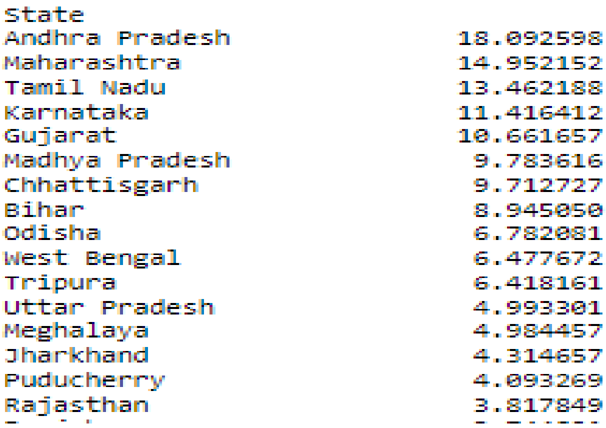


NAGALAND has 53.93 % worker population whereas as LAKSHADWEEP has only 24.98 % worker population with respect to their Total Population.

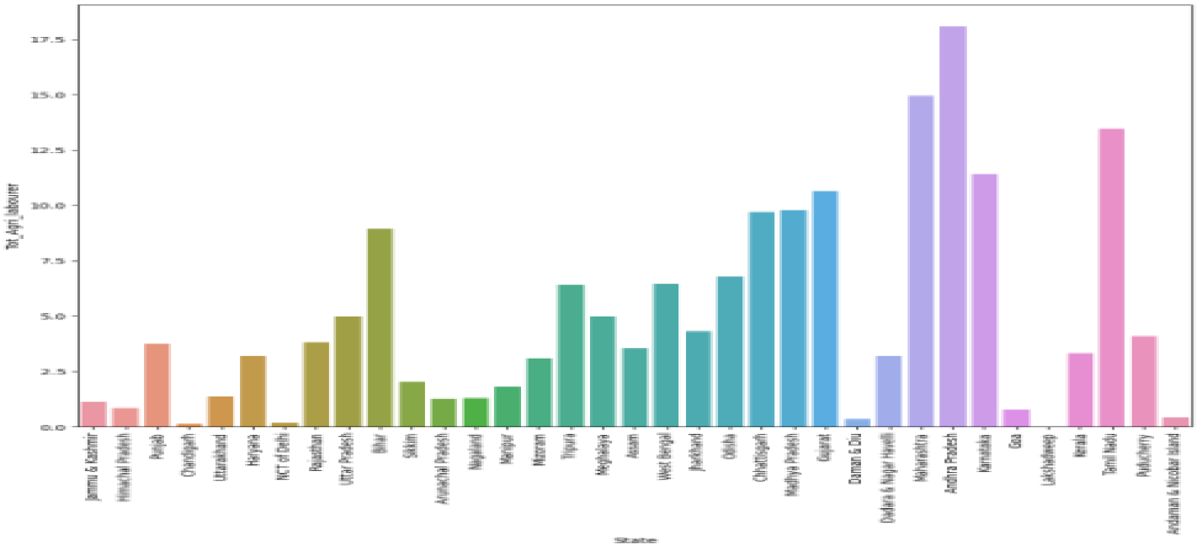
viii) Which state / UT has more agricultural labourers in India?

FORMULA :

TOTAL AGRI LABOURERS (%) = (TOTAL MALE + TOTAL FEMALE AGRI LABOURERS) / (TOTAL POPULATION OF EACH STATE) * 100



```
State
Andhra Pradesh             18.092598
Maharashtra                14.952152
Tamil Nadu                 13.462188
Karnataka                  11.416412
Gujarat                    10.661657
Madhya Pradesh              9.783616
Chhattisgarh                9.712727
Bihar                       8.945050
Odisha                      6.782081
West Bengal                 6.477672
Tripura                     6.418161
Uttar Pradesh               4.993301
Meghalaya                   4.984457
Jharkhand                   4.314657
Puducherry                  4.093269
Rajasthan                   3.817849
```

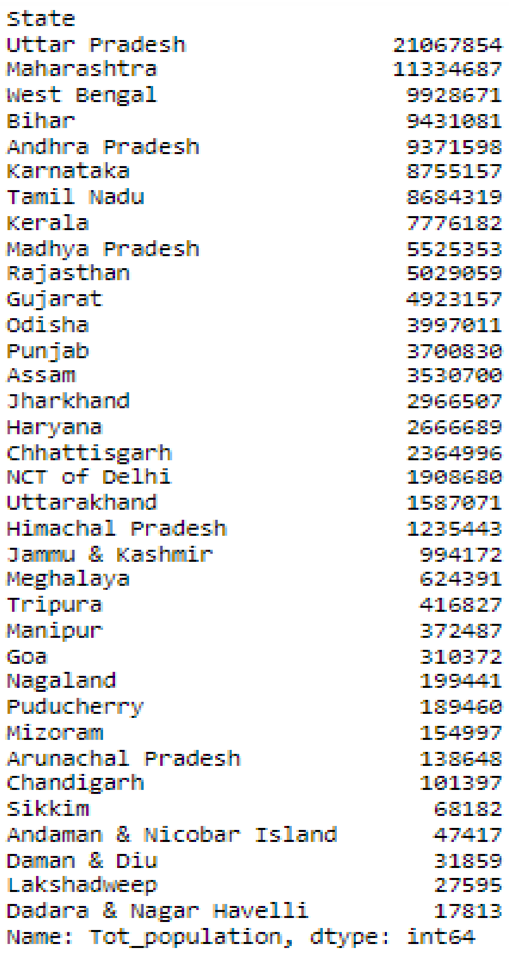


ANDRA PRADESH has 18.09 % Agricultural population who predominantly depends upon Agriculture for their livelihood whereas LAKSHADWEEP has 0 % Agricultural population.

ix) Which state/ UT has highest and lowest population ?

FORMULA :

TOTAL POPULATION OF A STATE = TOTAL MALES + TOTAL FEMALES IN RESPECTIVE STATE



```
State
Uttar Pradesh                21067854
Maharashtra                  11334687
West Bengal                   9928671
Bihar                         9431081
Andhra Pradesh                9371598
Karnataka                     8755157
Tamil Nadu                    8684319
Kerala                        7776182
Madhya Pradesh                5525353
Rajasthan                     5029059
Gujarat                       4923157
Odisha                        3997011
Punjab                        3700830
Assam                         3530700
Jharkhand                     2966507
Haryana                       2666689
Chhattisgarh                  2364996
NCT of Delhi                  1908680
Uttarakhand                   1587071
Himachal Pradesh              1235443
Jammu & Kashmir                994172
Meghalaya                      624391
Tripura                        416827
Manipur                        372487
Goa                            310372
Nagaland                       199441
Puducherry                     189460
Mizoram                        154997
Arunachal Pradesh              138648
Chandigarh                     101397
Sikkim                          68182
Andaman & Nicobar Island        47417
Daman & Diu                     31859
Lakshadweep                     27595
Dadara & Nagar Havelli          17813
Name: Tot_population, dtype: int64
```

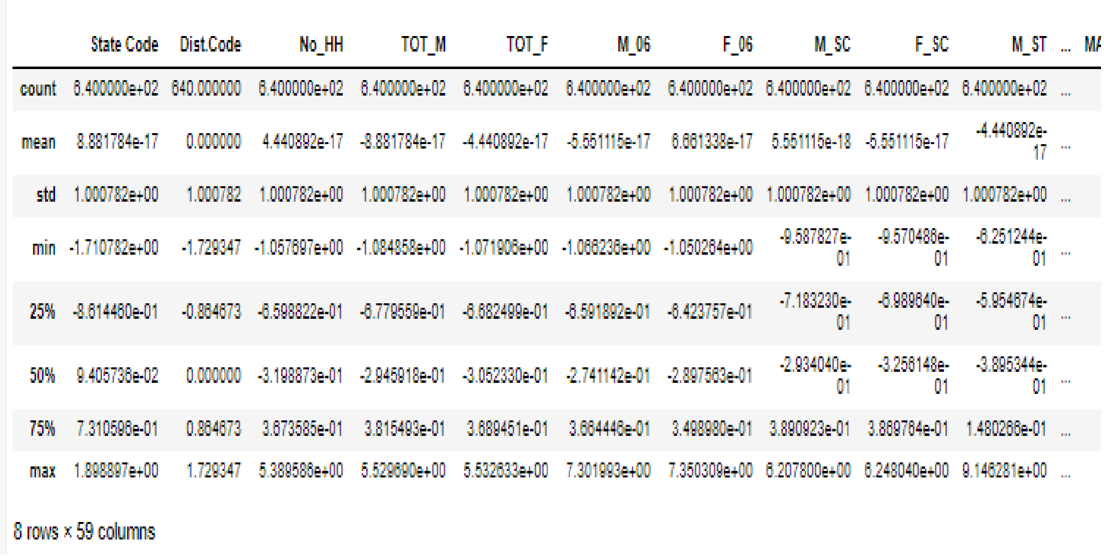
- **UTTAR PRADESH has highest population**
- **DADRA & NAGAR HAVELLI has lowest population in India**

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

Because of population data , no need to treat outliers. Some outliers represent natural variations in the population, and they should be left as it is in the dataset.

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

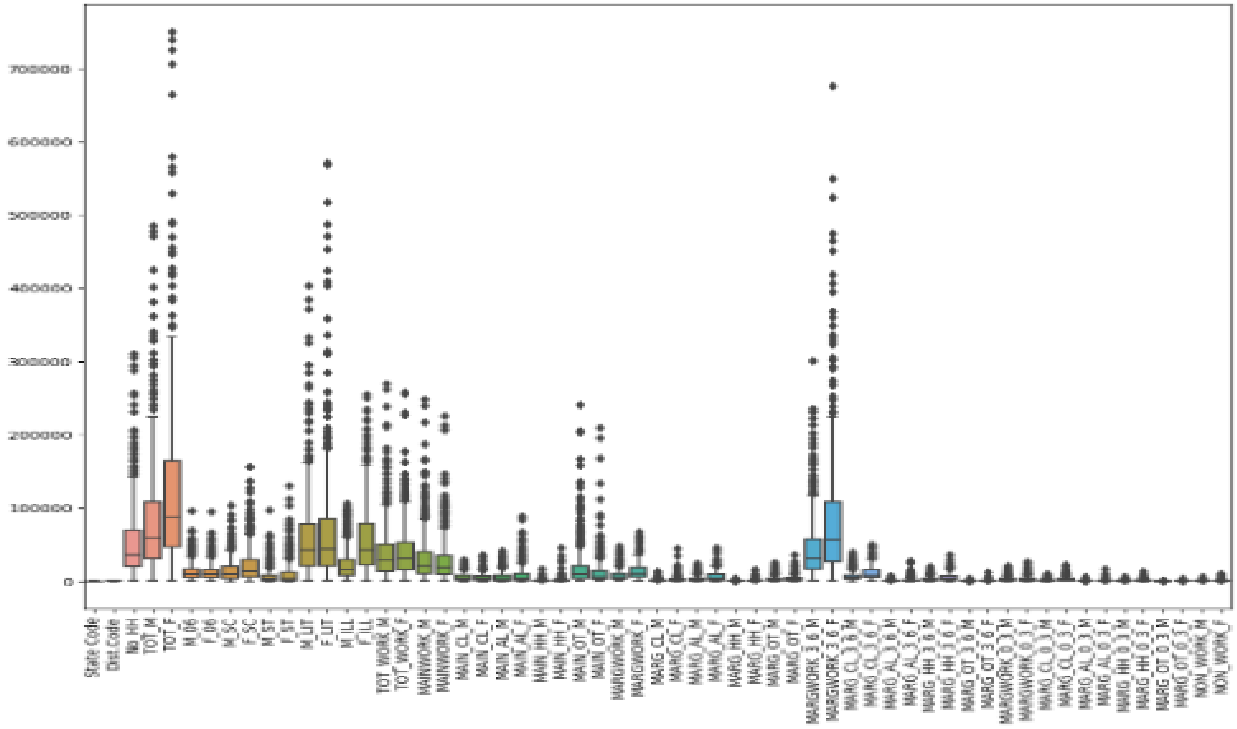
z-score scaling :



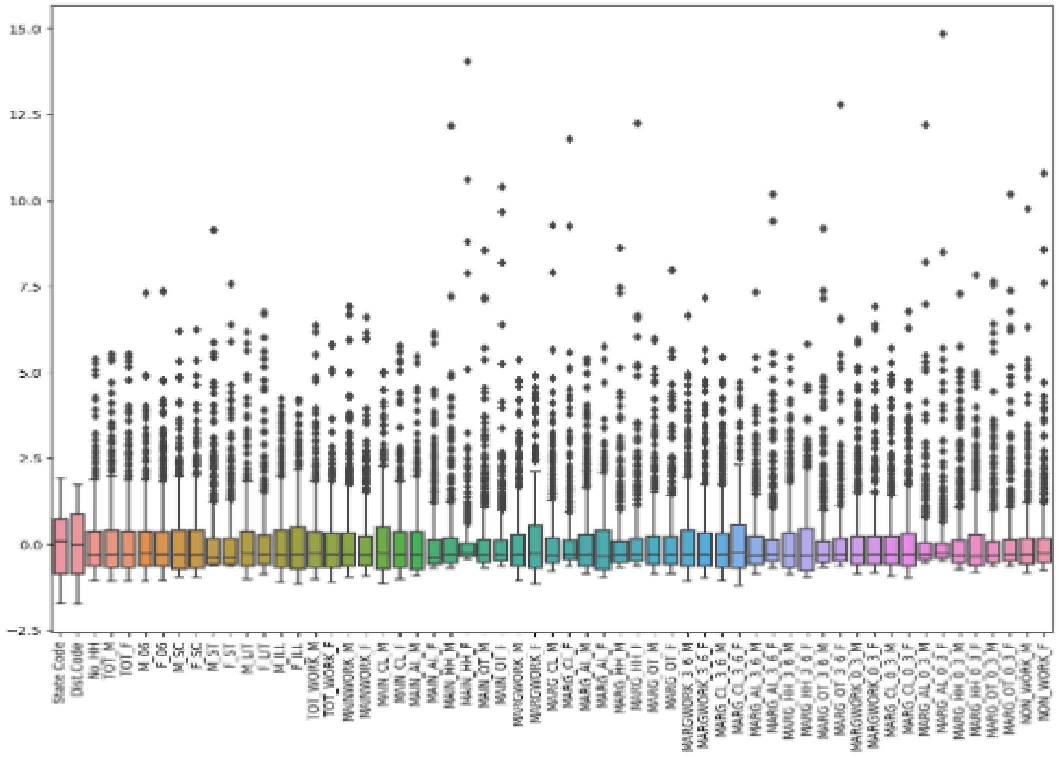
	State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MA
count	6.400000e+02	640.000000	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	...	
mean	8.881784e-17	0.000000	4.440892e-17	-8.881784e-17	-4.440892e-17	-5.551115e-17	6.661338e-17	5.551115e-18	-5.551115e-17	-4.440892e-17	...	
std	1.000782e+00	1.000782	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	...	
min	-1.710782e+00	-1.729347	-1.057697e+00	-1.084858e+00	-1.071906e+00	-1.066236e+00	-1.050264e+00	-9.587827e-01	-9.570486e-01	-6.251244e-01	...	
25%	-8.614460e-01	-0.864673	-6.598822e-01	-6.779559e-01	-6.682499e-01	-6.591892e-01	-6.423757e-01	-7.183230e-01	-6.989640e-01	-5.954674e-01	...	
50%	9.405736e-02	0.000000	-3.198873e-01	-2.945918e-01	-3.052330e-01	-2.741142e-01	-2.897563e-01	-2.934040e-01	-3.256148e-01	-3.895344e-01	...	
75%	7.310596e-01	0.864673	3.673585e-01	3.815493e-01	3.689451e-01	3.664446e-01	3.498980e-01	3.890923e-01	3.869784e-01	1.480268e-01	...	
max	1.898897e+00	1.729347	5.389586e+00	5.529690e+00	5.532633e+00	7.301993e+00	7.350309e+00	6.207800e+00	6.248040e+00	9.146281e+00	...	

8 rows × 59 columns

BOXPLOT BEFORE SCALING :



BOXPLOT AFTER SCALING :



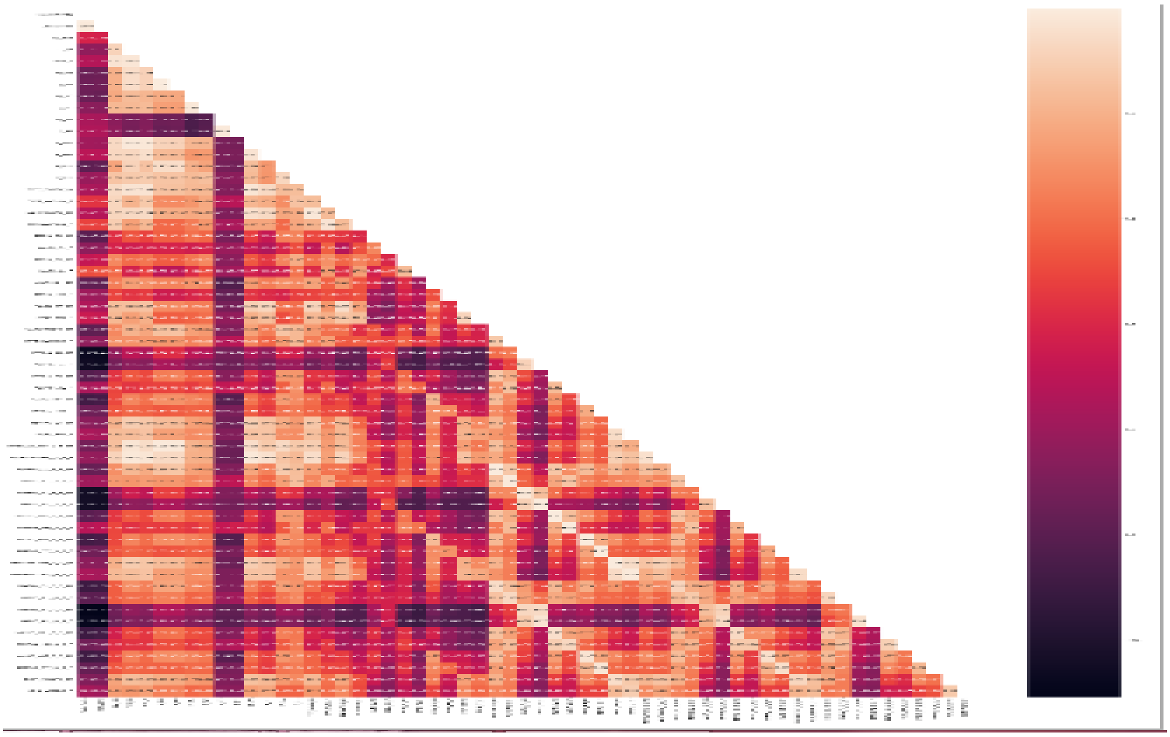
OBSERVATION :

- Scaling shrinks the range of values as shown in the figure while keeping the outliers in.
- However, the outliers have an influence only when computing the empirical mean and standard deviation.
- To compare the boxplot before and after scaling , the only difference is the distances between marginal outliers and inliers are shrunk.

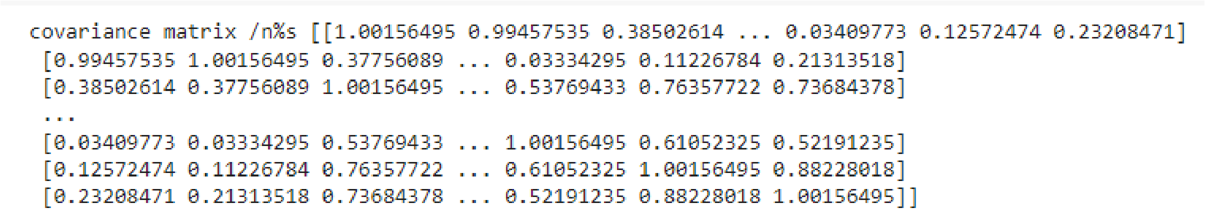
5) PCA: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.

STEP 1 : TO FIND CORRELATION

HEATMAP:



Covariance Matrix :



```
covariance matrix /n%s [[1.00156495 0.99457535 0.38502614 ... 0.03409773 0.12572474 0.23208471]
 [0.99457535 1.00156495 0.37756089 ... 0.03334295 0.11226784 0.21313518]
 [0.38502614 0.37756089 1.00156495 ... 0.53769433 0.76357722 0.73684378]
 ...
 [0.03409773 0.03334295 0.53769433 ... 1.00156495 0.61052325 0.52191235]
 [0.12572474 0.11226784 0.76357722 ... 0.61052325 1.00156495 0.88228018]
 [0.23208471 0.21313518 0.73684378 ... 0.52191235 0.88228018 1.00156495]]
```

Calculate BARTLETT SPHERICITY :

To confirm the statistical significance of correlation :

H0: All variables in the data are uncorrelated

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

Hence p_value = 0 ,

we can reject H0 because p_value < 0.05 and can say that correlations are significant.

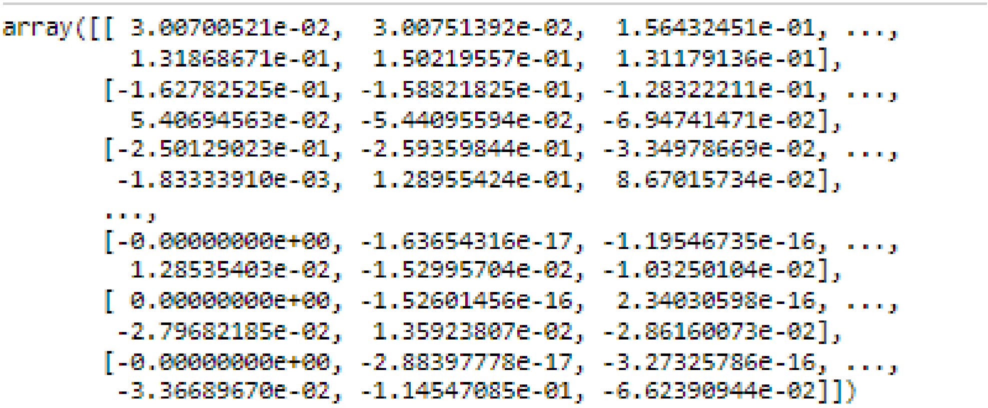
Calculate kmo :

To confirm adequacy of sample size :

kmo value is 0.80 which is above 0.7 , so we can say that there are adequate sample size to perform PCA.

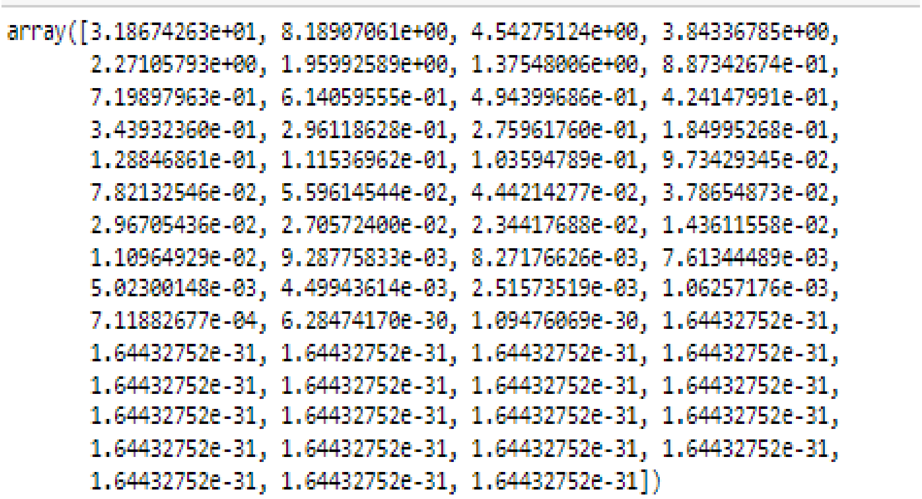
STEP 2 : APPLY PCA TECHNIQUE

EIGEN VECTORS :



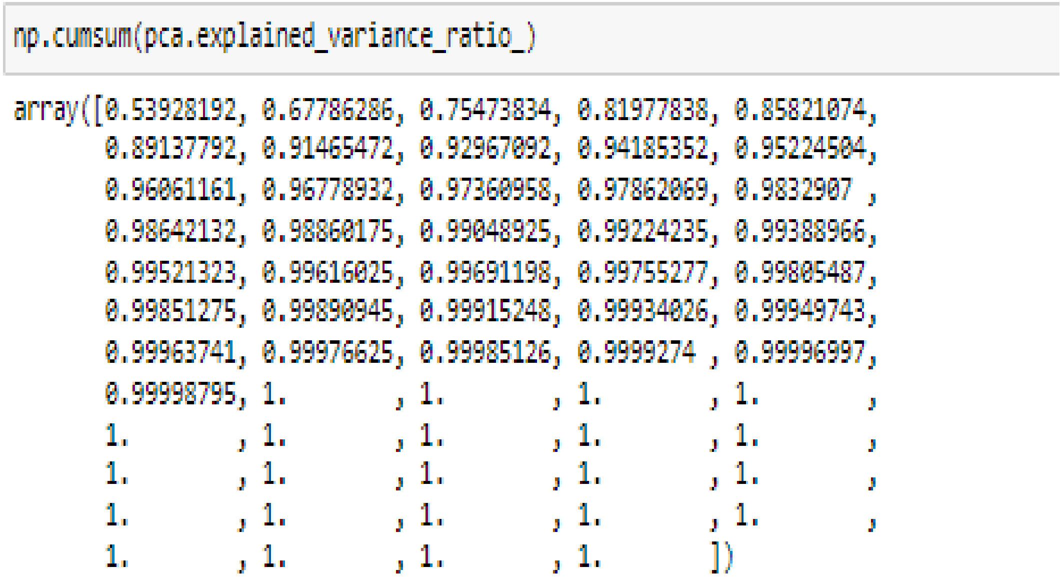
```
array([[ 3.00700521e-02,  3.00751392e-02,  1.56432451e-01, ...,
         1.31868671e-01,  1.50219557e-01,  1.31179136e-01],
       [-1.62782525e-01, -1.58821825e-01, -1.28322211e-01, ...,
         5.40694563e-02, -5.44095594e-02, -6.94741471e-02],
       [-2.50129023e-01, -2.59359844e-01, -3.34978669e-02, ...,
        -1.83333910e-03,  1.28955424e-01,  8.67015734e-02],
       ...,
       [-0.00000000e+00, -1.63654316e-17, -1.19546735e-16, ...,
         1.28535403e-02, -1.52995704e-02, -1.03250104e-02],
       [ 0.00000000e+00, -1.52601456e-16,  2.34030598e-16, ...,
        -2.79682185e-02,  1.35923807e-02, -2.86160073e-02],
       [-0.00000000e+00, -2.88397778e-17, -3.27325786e-16, ...,
        -3.36689670e-02, -1.14547085e-01, -6.62390944e-02]])
```

EIGEN VALUES :



```
array([3.18674263e+01, 8.18907061e+00, 4.54275124e+00, 3.84336785e+00,
       2.27105793e+00, 1.95992589e+00, 1.37548006e+00, 8.87342674e-01,
       7.19897963e-01, 6.14059555e-01, 4.94399686e-01, 4.24147991e-01,
       3.43932360e-01, 2.96118628e-01, 2.75961760e-01, 1.84995268e-01,
       1.28846861e-01, 1.11536962e-01, 1.03594789e-01, 9.73429345e-02,
       7.82132546e-02, 5.59614544e-02, 4.44214277e-02, 3.78654873e-02,
       2.96705436e-02, 2.70572400e-02, 2.34417688e-02, 1.43611558e-02,
       1.10964929e-02, 9.28775833e-03, 8.27176626e-03, 7.61344489e-03,
       5.02300148e-03, 4.49943614e-03, 2.51573519e-03, 1.06257176e-03,
       7.11882677e-04, 6.28474170e-30, 1.09476069e-30, 1.64432752e-31,
       1.64432752e-31, 1.64432752e-31, 1.64432752e-31, 1.64432752e-31,
       1.64432752e-31, 1.64432752e-31, 1.64432752e-31, 1.64432752e-31,
       1.64432752e-31, 1.64432752e-31, 1.64432752e-31, 1.64432752e-31,
       1.64432752e-31, 1.64432752e-31, 1.64432752e-31, 1.64432752e-31,
       1.64432752e-31, 1.64432752e-31, 1.64432752e-31])
```

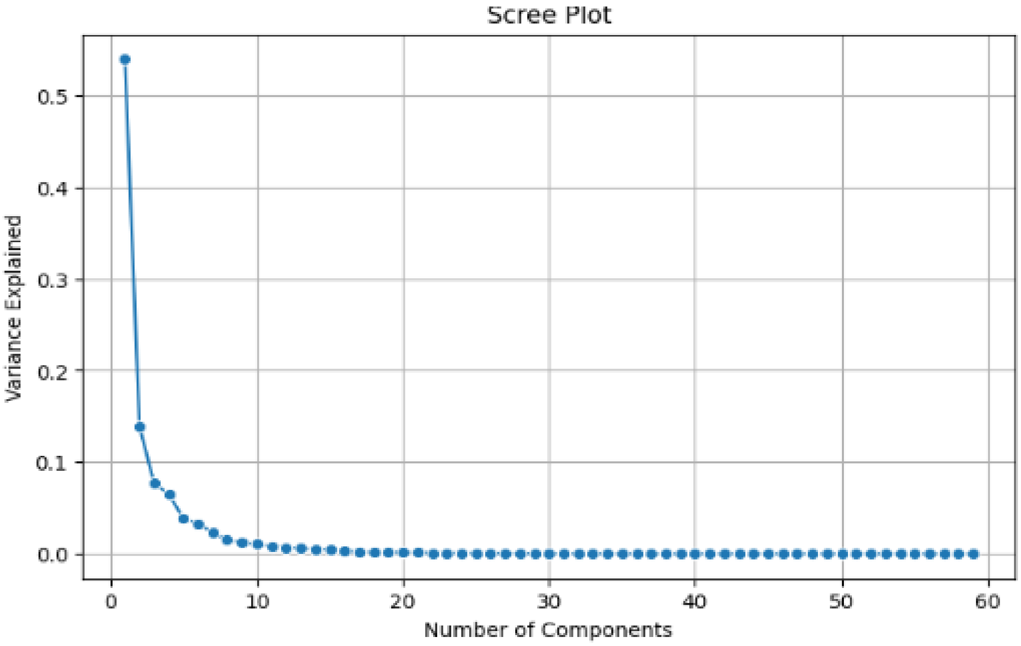
6) PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.



```
np.cumsum(pca.explained_variance_ratio_)
```

```
array([0.53928192, 0.67786286, 0.75473834, 0.81977838, 0.85821074,
       0.89137792, 0.91465472, 0.92967092, 0.94185352, 0.95224504,
       0.96061161, 0.96778932, 0.97360958, 0.97862069, 0.9832907 ,
       0.98642132, 0.98860175, 0.99048925, 0.99224235, 0.99388966,
       0.99521323, 0.99616025, 0.99691198, 0.99755277, 0.99805487,
       0.99851275, 0.99890945, 0.99915248, 0.99934026, 0.99949743,
       0.99963741, 0.99976625, 0.99985126, 0.9999274 , 0.99996997,
       0.99998795, 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        ])
```

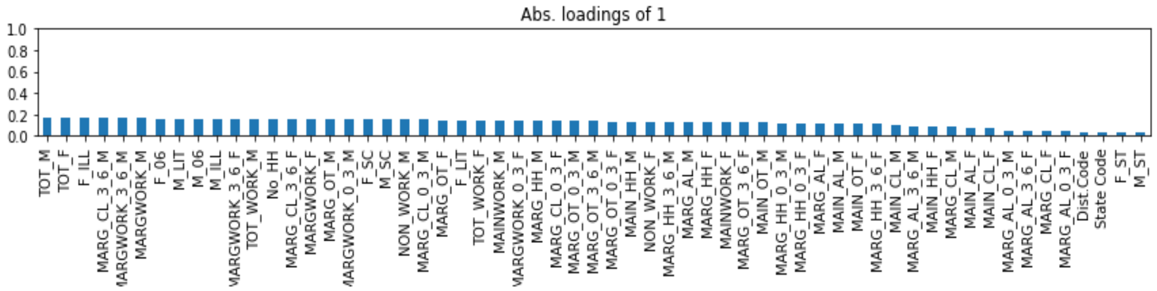
SCREE PLOT :



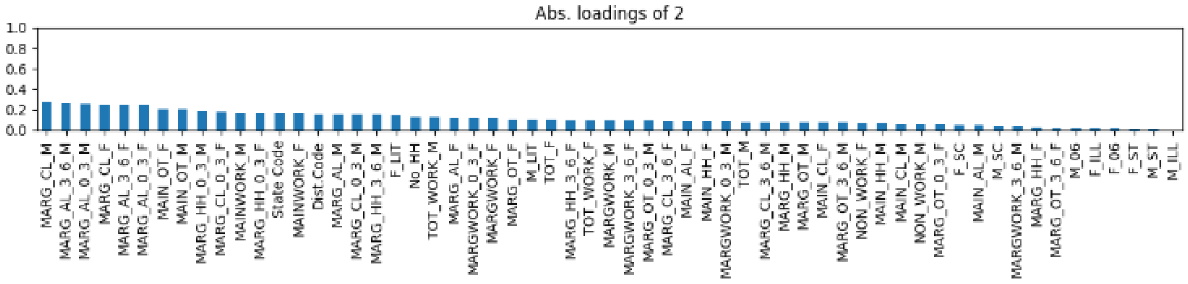
From the above Scree plot and explained variance ratio, we can say that the **Optimum number of PCs are 7** which has 91.4 % explained variance.

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

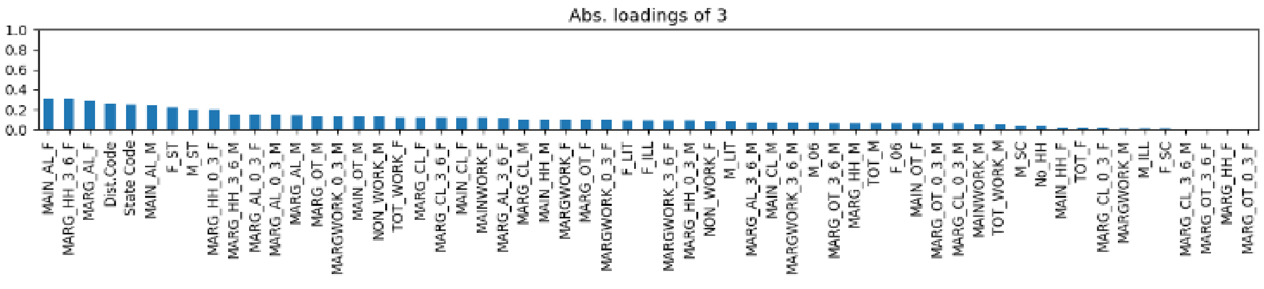
PC1 :



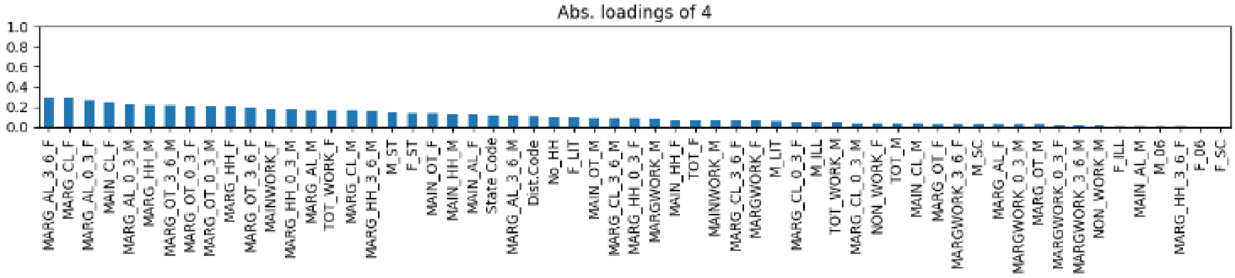
PC2 :



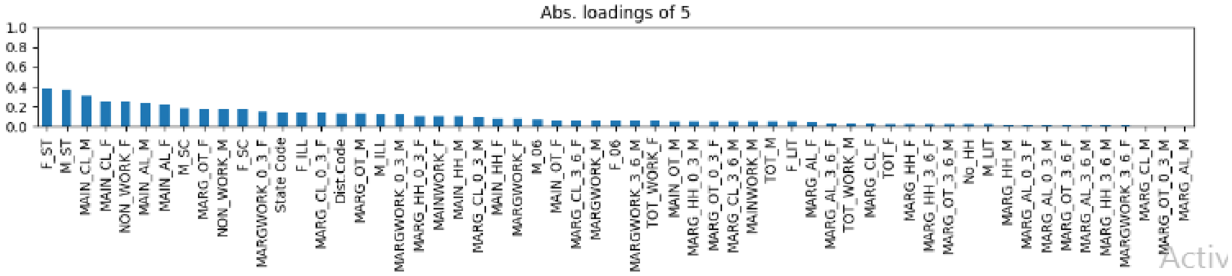
PC3 :



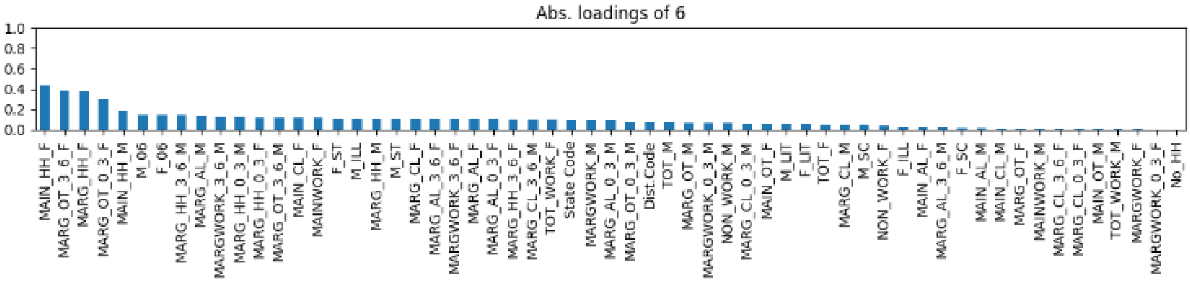
PC4 :


Abs. loadings of 4
1.0
0.8
0.6
0.4
0.2
0.0
MARG_AL_3_6_F
MARG_CL_F
MARG_AL_0_3_F
MAIN_CL_F
MARG_AL_0_3_M
MARG_HH_M
MARG_OT_3_6_M
MARG_OT_0_3_F
MARG_OT_0_3_M
MARG_HH_F
MARG_OT_3_6_F
MAINWORK_F
MARG_HH_0_3_M
MARG_AL_M
TOT_WORK_F
MARG_CL_M
MARG_HH_3_6_M
M_ST
F_ST
MAIN_OT_F
MAIN_HH_M
MAIN_AL_F
State Code
MARG_AL_3_6_M
Dist.Code
No_HH
F_LIT
MAIN_OT_M
MARG_CL_3_6_M
MARG_HH_0_3_F
MARGWORK_M
MAIN_HH_F
TOT_F
MAINWORK_M
MARG_CL_3_6_F
MARGWORK_F
M_LIT
MARG_CL_0_3_F
M_ILL
TOT_WORK_M
MARG_CL_0_3_M
NON_WORK_F
TOT_M
MAIN_CL_M
MARG_OT_F
MARGWORK_3_6_F
M_SC
MARG_AL_F
MARGWORK_0_3_M
MARG_OT_M
MARGWORK_0_3_F
MARGWORK_3_6_M
NON_WORK_M
F_ILL
MAIN_AL_M
M_06
MARG_HH_3_6_F
F_06
F_SC

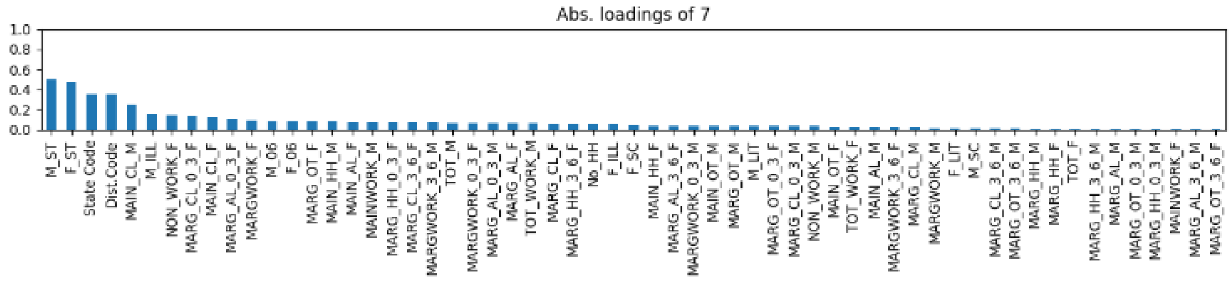
PC5 :


Abs. loadings of 5
1.0
0.8
0.6
0.4
0.2
0.0
F_ST
M_ST
MAIN_CL_M
MAIN_CL_F
NON_WORK_F
MAIN_AL_M
MAIN_AL_F
M_SC
MARG_OT_F
NON_WORK_M
F_SC
MARGWORK_0_3_F
State Code
F_ILL
MARG_CL_0_3_F
Dist.Code
MARG_OT_M
M_ILL
MARGWORK_0_3_M
MARG_HH_0_3_F
MAINWORK_F
MAIN_HH_M
MARG_CL_0_3_M
MAIN_HH_F
MARGWORK_F
M_06
MAIN_OT_F
MARG_CL_3_6_F
MARGWORK_M
F_06
MARGWORK_3_6_M
TOT_WORK_F
MAIN_OT_M
MARG_HH_0_3_M
MARG_OT_0_3_F
MARG_CL_3_6_M
MAINWORK_M
TOT_M
F_LIT
MARG_AL_F
MARG_AL_3_6_F
TOT_WORK_M
MARG_CL_F
TOT_F
MARG_HH_F
MARG_HH_3_6_F
MARG_OT_3_6_M
No_HH
M_LIT
MARG_HH_M
MARG_AL_0_3_F
MARG_AL_0_3_M
MARG_OT_3_6_F
MARG_AL_3_6_M
MARG_HH_3_6_M
MARGWORK_3_6_F
MARG_CL_M
MARG_OT_0_3_M
MARG_AL_M

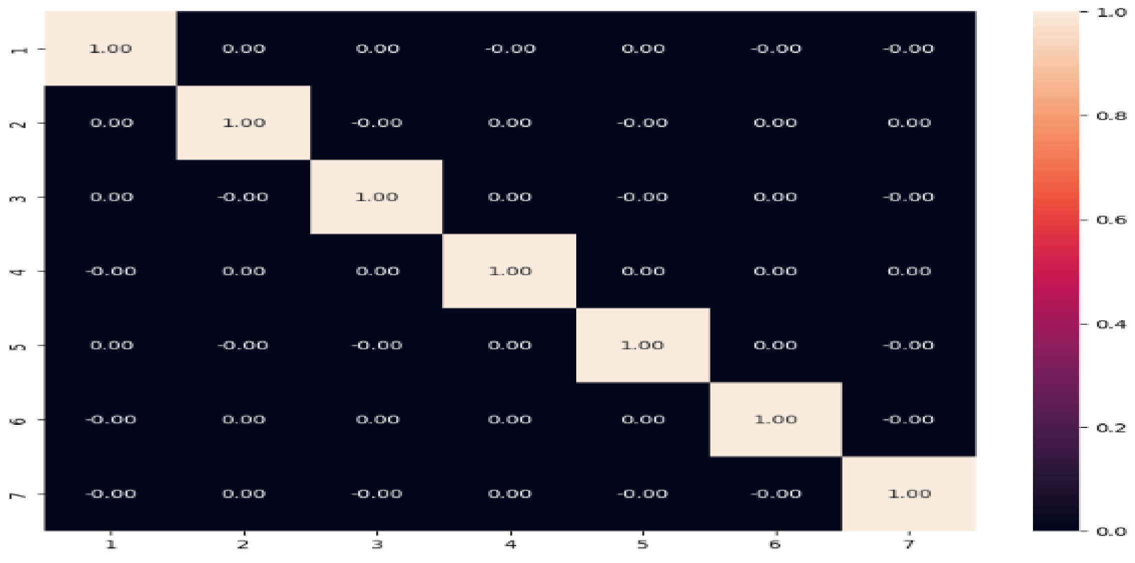
PC6 :


Abs. loadings of 6
1.0
0.8
0.6
0.4
0.2
0.0
MAIN_HH_F
MARG_OT_3_6_F
MARG_HH_F
MARG_OT_0_3_F
MAIN_HH_M
M_06
F_06
MARG_HH_3_6_M
MARG_AL_M
MARGWORK_3_6_M
MARG_HH_0_3_M
MARG_HH_0_3_F
MARG_OT_3_6_M
MAIN_CL_F
MAINWORK_F
F_ST
M_ILL
MARG_HH_M
M_ST
MARG_CL_F
MARG_AL_3_6_F
MARGWORK_3_6_F
MARG_AL_F
MARG_AL_0_3_F
MARG_HH_3_6_F
MARG_CL_3_6_M
TOT_WORK_F
State Code
MARGWORK_M
MARG_AL_0_3_M
MARG_OT_0_3_M
Dist.Code
TOT_M
MARG_OT_M
MARGWORK_0_3_M
NON_WORK_M
MARG_CL_0_3_M
MAIN_OT_F
M_LIT
F_LIT
TOT_F
MARG_CL_M
M_SC
NON_WORK_F
F_ILL
MAIN_AL_F
MARG_AL_3_6_M
F_SC
MAIN_AL_M
MAIN_CL_M
MARG_OT_F
MAINWORK_M
MARG_CL_3_6_F
MARG_CL_0_3_F
MAIN_OT_M
TOT_WORK_M
MARGWORK_F
MARGWORK_0_3_F
No_HH

PC7 :



From the above we can say that PC1 exhibit maximum variance when compared to others. Finally we got the heatmap with no correlation exists among the PCs.



8) **PCA: Write linear equation for first PC.**

In General, Linear equation for PC1 :

PC1= a1x1+a2x2+….+anxn

a1,a2,…an : co-efficient / Eigen vectors / Factor loadings

x1,x2,…xn : observed data

PC1 = 0.030070 (State CODE) + 0.030075 (Dist.CODE)

+ 0.156432 (NO_ HH) + …..+ 0.131179(NON_WORK_F)

INSIGHTS :

- **UTTAR PRADESH** has highest population
- **DADRA & NAGAR HAVELLI** has lowest population in India
- **ANDRA PRADESH** has highest Gender ratio of 1895 Females per 1000 Males.
- **LAKSHWADEEP** has lowest Gender ratio of 1152 Females per 1000 Males.
- **KERALA** has highest literacy rate of 80.59 %
- **BIHAR** has lowest literacy rate of 47.98%
- 79.87 % of female population in **KERALA** are literates (highest**).**
- 39.75 % of female population are literates in **BIHAR** (lowest).
- **UTTAR PRADESH** has more Scheduled caste population.
- **MADHYA PRADESH** has more Scheduled Tribes population.
- **NAGALAND** has 53.93 % worker population
- **LAKSHADWEEP** has only 24.98 % worker population with respect to their Total Population.
- **ANDRA PRADESH** has 18.09 % Agricultural population who predominantly depends upon Agriculture for their livelihood
- **LAKSHADWEEP** has 0 % Agricultural population.

THANK YOU