Data Mining – Project

Digital Ads Data

Problem Statement

The ads24x7 is a Digital Marketing company which has now got seed funding of $10 Million. They are expanding their wings in Marketing Analytics. They collected data from their Marketing Intelligence team and now wants you (their newly appointed data analyst) to segment type of ads based on the features provided. Use Clustering procedure to segment ads into homogeneous groups.

The following three features are commonly used in digital marketing:

CPM = (Total Campaign Spend / Number of Impressions) * 1,000

CPC = Total Cost (spend) / Number of Clicks

CTR = Total Measured Clicks / Total Measured Ad Impressions x 100

Solution:

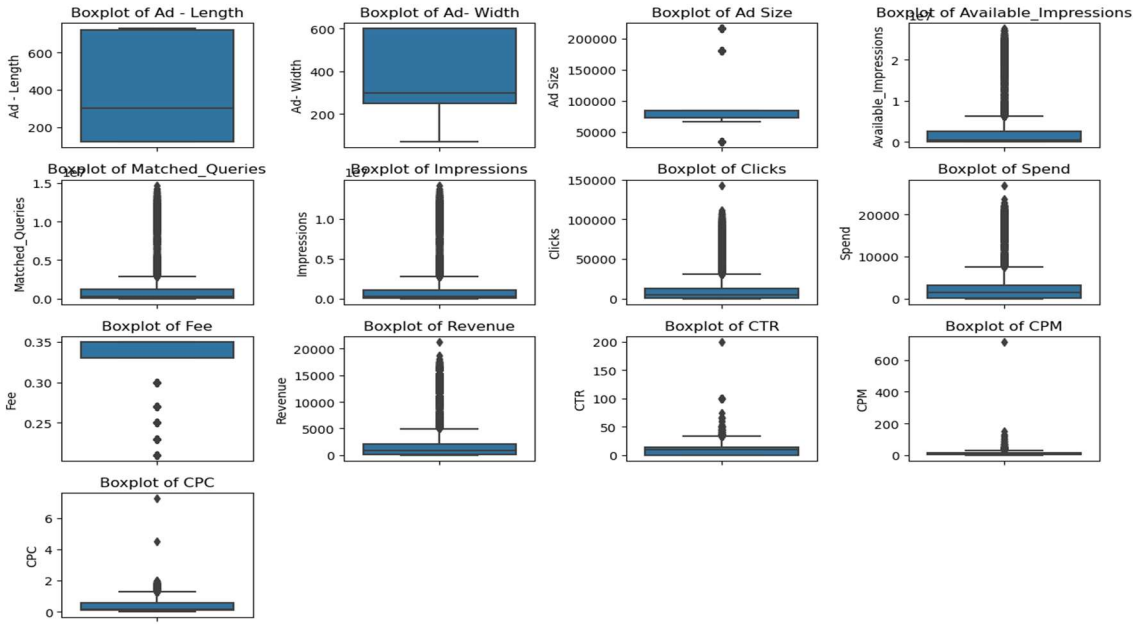
1. **Read data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc**
    - Data Summary
    - There are 25,857 Rows and 19 columns
    - There are Object, Float and Integer Datatypes
    - There are missing values under CPM (6465) , CPC (6465) and CTR(7527) which we will fill using the given formulas
    - There are no Duplicate values in the data

2. **Treat missing values in CPC, CTR and CPM using the formula given.**
    - Using below formulas , missing values for CPC , CTR and CPM were imputed
    - CPM = (Total Campaign Spend / Number of Impressions) * 1,000
    - CPC = Total Cost (spend) / Number of Clicks
    - CTR = Total Measured Clicks / Total Measured Ad Impressions x 100
    - After Imputing it still shows some null entiries under CPM CPC ans CTR

      CTR 219
      CPM 219
      CPC 2586

    - The reason being both the input data for the formulas are 0. Hence we decided to drop the Nan values
    - There are some Inf values in the data set, which we decided to convert to nan values and drop
    - Final Dataset has rows reduced to 23066

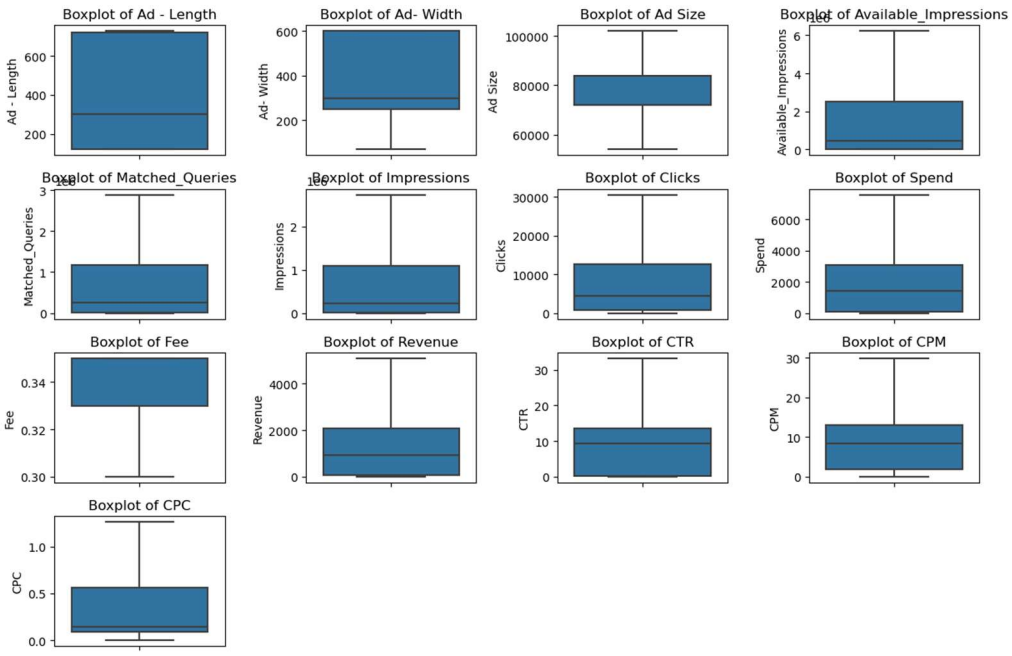
3. **Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an analyst your judgement may be different from another analyst)**



As we see there are many outliers in the data, we decide to treat them before clustering. Kemean Clustering is sensitive to outliers

We decided to remove outliers by capping to the lower and upper range

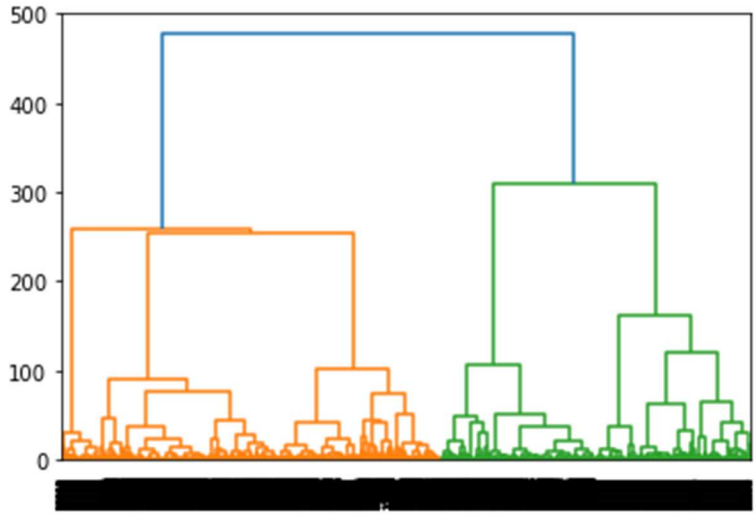
Outliers are treated Below



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

We will perform Z score on the Dataframe. Z-score normalization refers to the process of normalizing every value in a dataset such that the mean of all of the values is 0 and the standard deviation is 1. We use the following formula to perform a z-score normalization on every value in a dataset

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

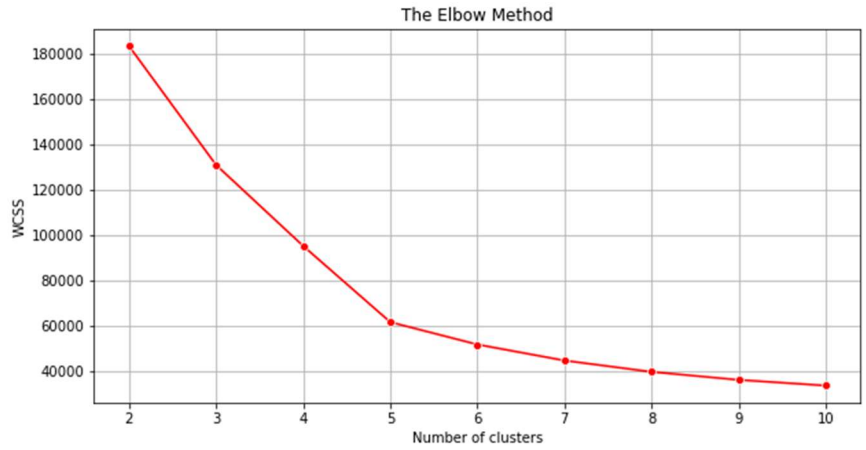


Truncating at P=200, We get 5 Clusters

Printing WSS Values upto 10 clusters

```
The WSS value for 2 clusters is 183349.10202886097
The WSS value for 3 clusters is 130878.34240367355
The WSS value for 4 clusters is 95133.9448134987
The WSS value for 5 clusters is 61539.189197853884
The WSS value for 6 clusters is 51676.89681600461
The WSS value for 7 clusters is 44598.25849746791
The WSS value for 8 clusters is 39597.84813652193
The WSS value for 9 clusters is 36061.740167829914
The WSS value for 10 clusters is 33544.28161848274
```

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



From above graph we can see 3 or 5 can be optimum number of clusters

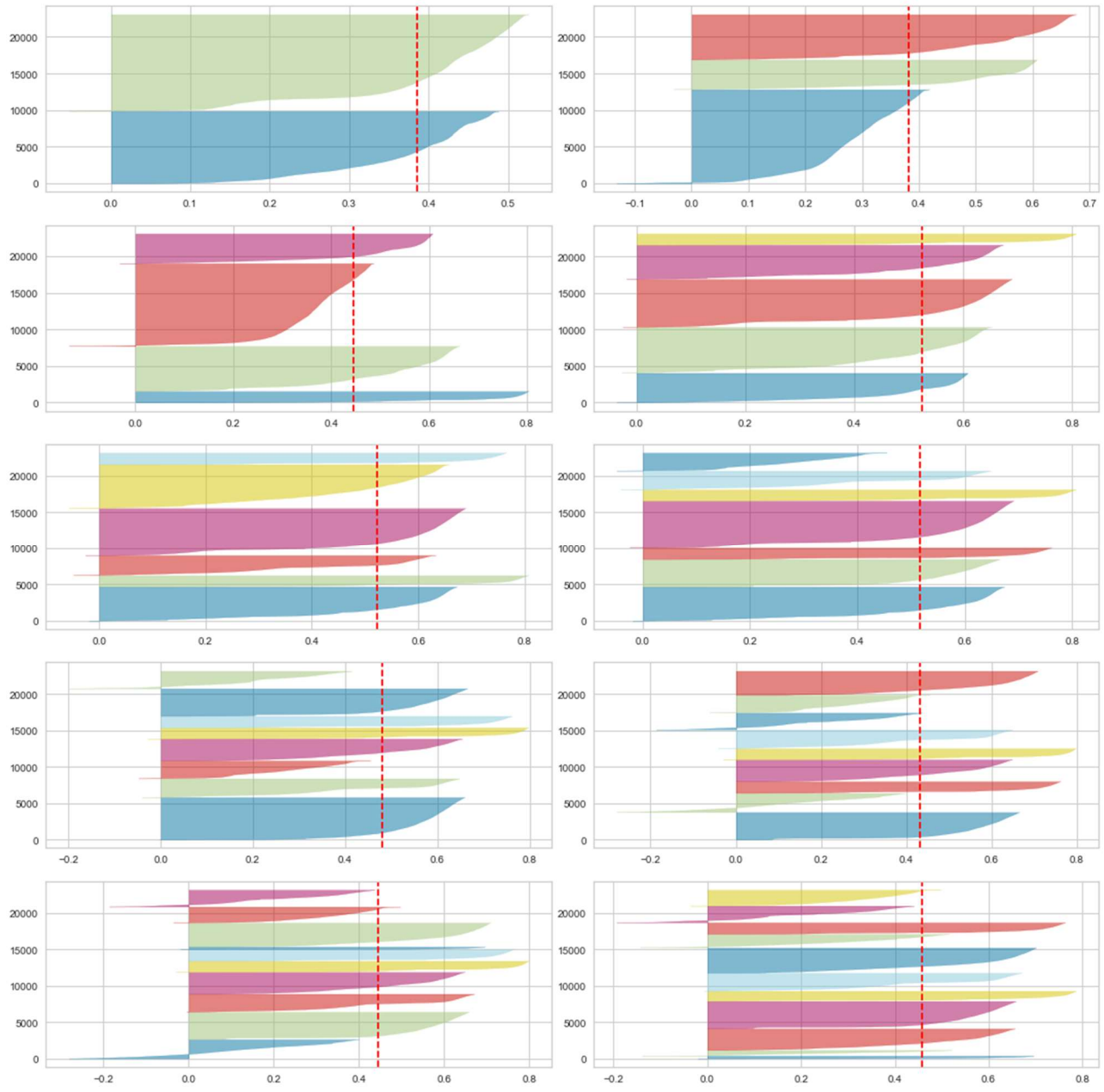
7. **Let's check with silhouette scores**

Average Silhouette Score is 0.54 . More than 0.5 is much better silhouette score it means a well distinguished clustering is done

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

```
The Sill Score for 1 clusters is 0.38572769619101077
The Sill Score for 2 clusters is 0.3825486036570082
The Sill Score for 3 clusters is 0.44534371698609754
The Sill Score for 4 clusters is 0.5240956940501831
The Sill Score for 5 clusters is 0.5221533662938636
The Sill Score for 6 clusters is 0.5165635029478517
The Sill Score for 7 clusters is 0.4797524035378018
The Sill Score for 8 clusters is 0.43186674723096125
The Sill Score for 9 clusters is 0.44462846563808417
The Sill Score for 10 clusters is 0.45728509161421904
```

Plotting Silhouette visualizer



.

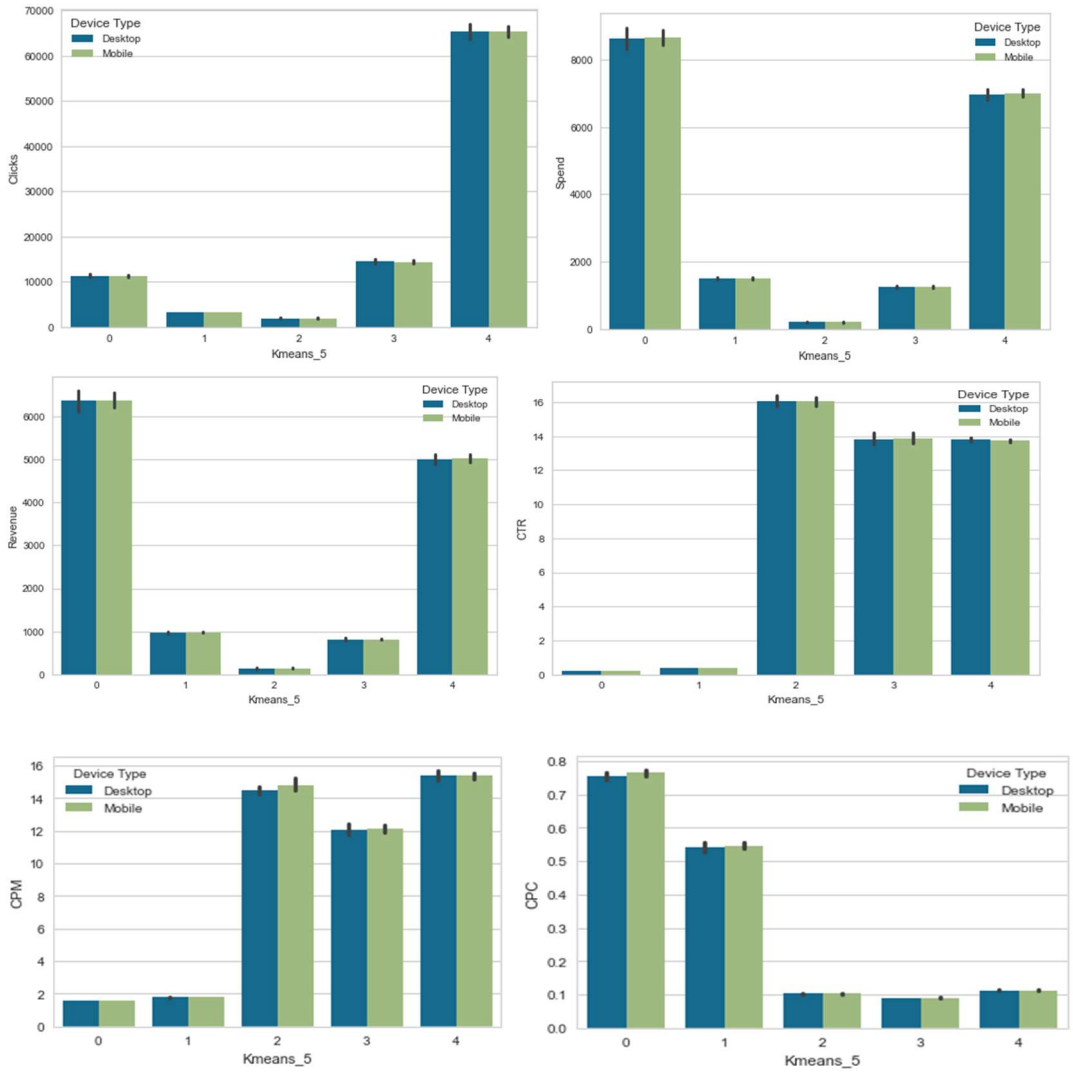
There are 2 deciding points

1. Above average silhouette score
2. The thickness of the silhouette plot representing each cluster

Silhouette score of 5,6 and 7 is above average. **F**or the plot with n_cluster 5 , the thickness is more uniform than the plot with n_cluster as 6 and 7. Thus, one can select the optimal number of clusters as 5.

From above Analysis We conclude to go ahead with 5 clusters

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



Observations:

1. Average clicks in Cluster 4 is highest for both Desktop and Mobile device type
2. Average Spend in Cluster 1 is highest for both Desktop and Mobile device type
3. Average Revenue in Cluster 1 is highest for both Desktop and Mobile device type
4. Average CTR in Cluster 3 is highest for both Desktop and Mobile device type
5. Average CPM in Cluster 5 is highest for both Desktop and Mobile device type
6. Average CPC in Cluster 1 is highest for both Desktop and Mobile device type

Problem 2
Census Data PCA

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 of 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 rural-urban residence. Census 2011 covered 35 States/Union Territories, 640 districts, 5,924 sub-districts, 7,935 Towns and 6,40,867 Villages.

The data collected has so many variables thus making it difficult to find useful details without using Data Science Techniques. You are tasked to perform detailed EDA and identify Optimum Principal Components that explains the most variance in data. Use Sklearn only.

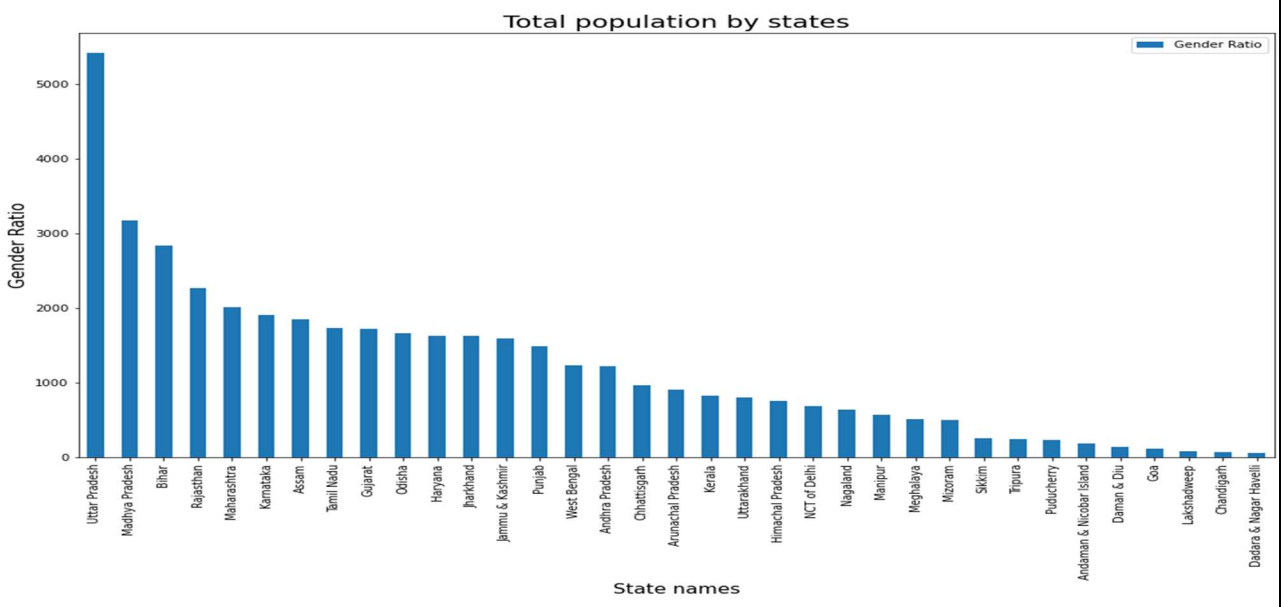
Solution:

1. **Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc**
   - Data Summary
   - There are 603 Rows and 61 columns
   - There are Object and Integer Datatypes
   - There are no null values in the dataset
   - There are no Duplicate values in the dataset
2. **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**

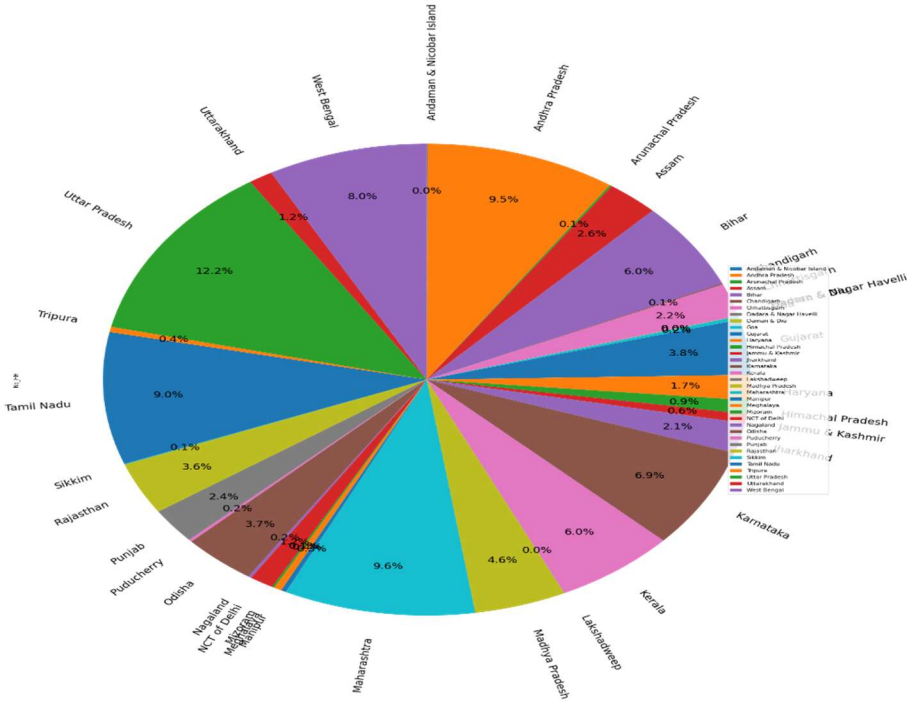
   Note : We have added gender ratio to the data frame to perform the EDA as required

   df['Gender Ratio'] = df["TOT_M"]/df["TOT_F"]*100

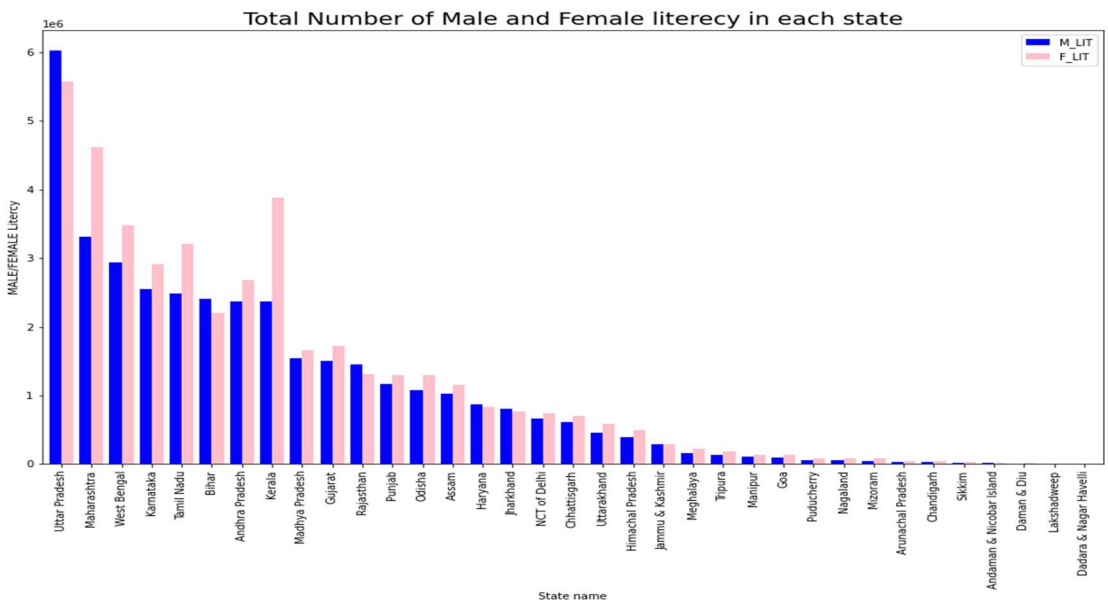
   a. **Which state has highest gender ratio and which has the lowest**



b. **What is the State wise population**



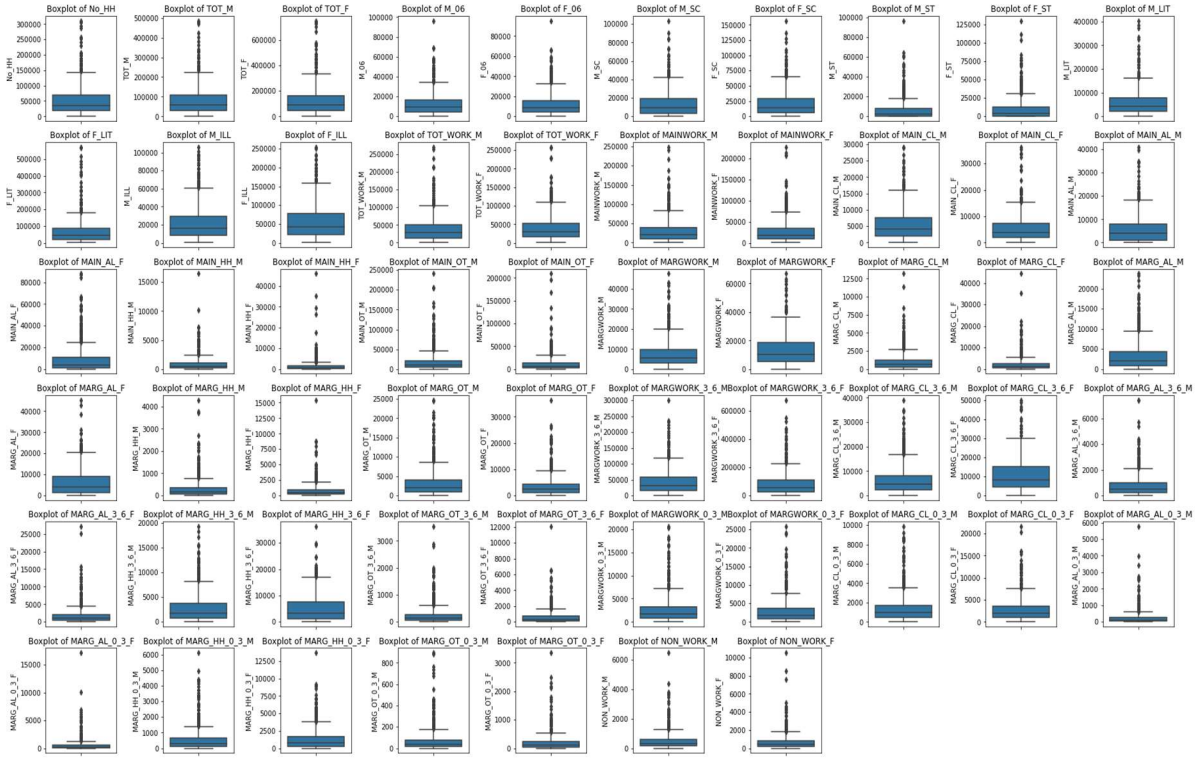
c. Total Literacy rate(M/F) in each state



Observations

1. Added Gender Ratio column and came up with the output that Aurangabad has the highest Gender Ratio - 138.055960
2. Analysing State wise using bar plot and observed that Uttarpradesh has highest Gender ratio Value.
3. Using Pyplot coming to the conclusion that Uttarpradesh has greatest percentage of Number of households like 12.2% because of this reason Gender ration is highest in Uttarpradesh.
4. Literacy rate is highest in uttarpradesh and Male literacy is higher than female literacy in UP.
5. At the same time the Bottom 4 state in literacy is Andaman and Nicobar Island, Daman and Diu, Lakshadweep, Dadra and Nagar Havelli

3. Checking for Outliers

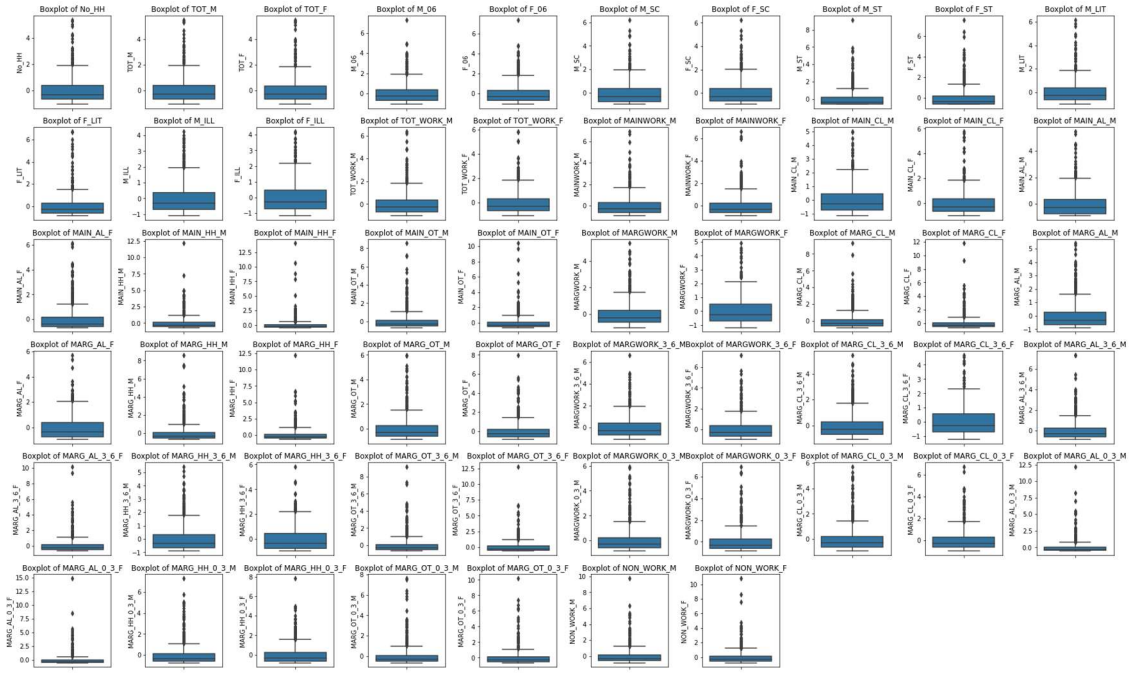


We are not treating outliers as our opinion that for dimensionality reduction treating outliers is not mandatory

4. Scale the Data using z-score method

from scipy.stats import zscore we scale the Data

We can see that after applying Zscore there is no impact on the outliers



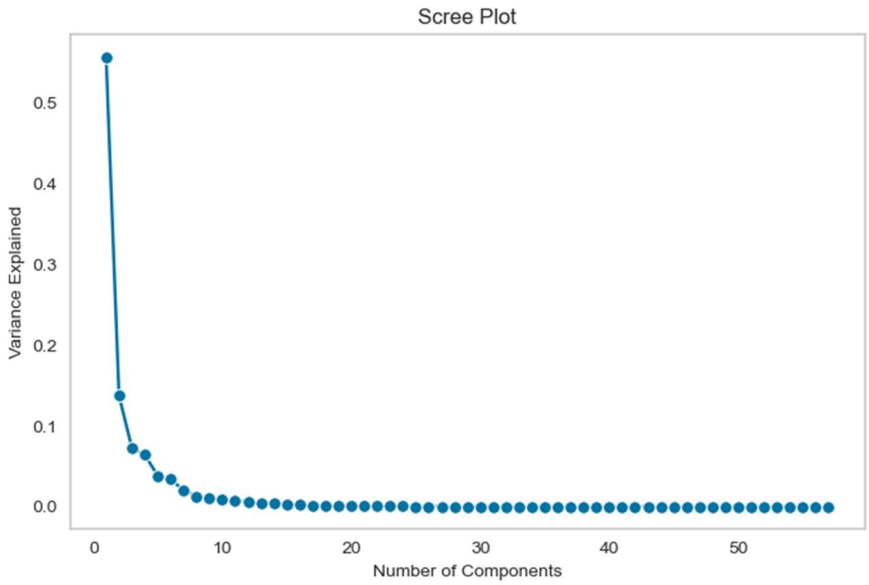
5. Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector

We will do some prerequisites before proceeding with PCA

- P_Value using bartlett sphericity test=0.0. This means There is significate correlation between the variables to work on
- Finding sample adequacy . Kmo test Value is 0.8. . Value >0.7 which means we have adequate PCA

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

- We will start with PCA considering all the 57 components of the data and perfrom fit transform on PCA using given data
- We get the PCA Components Array
- Check with Eigen values by calculating Explained Variance
- Checked the explained variance for PC and put it in the dataframe " df_extracted_loadings"
- Creating Scree plot to see the explained variance ration



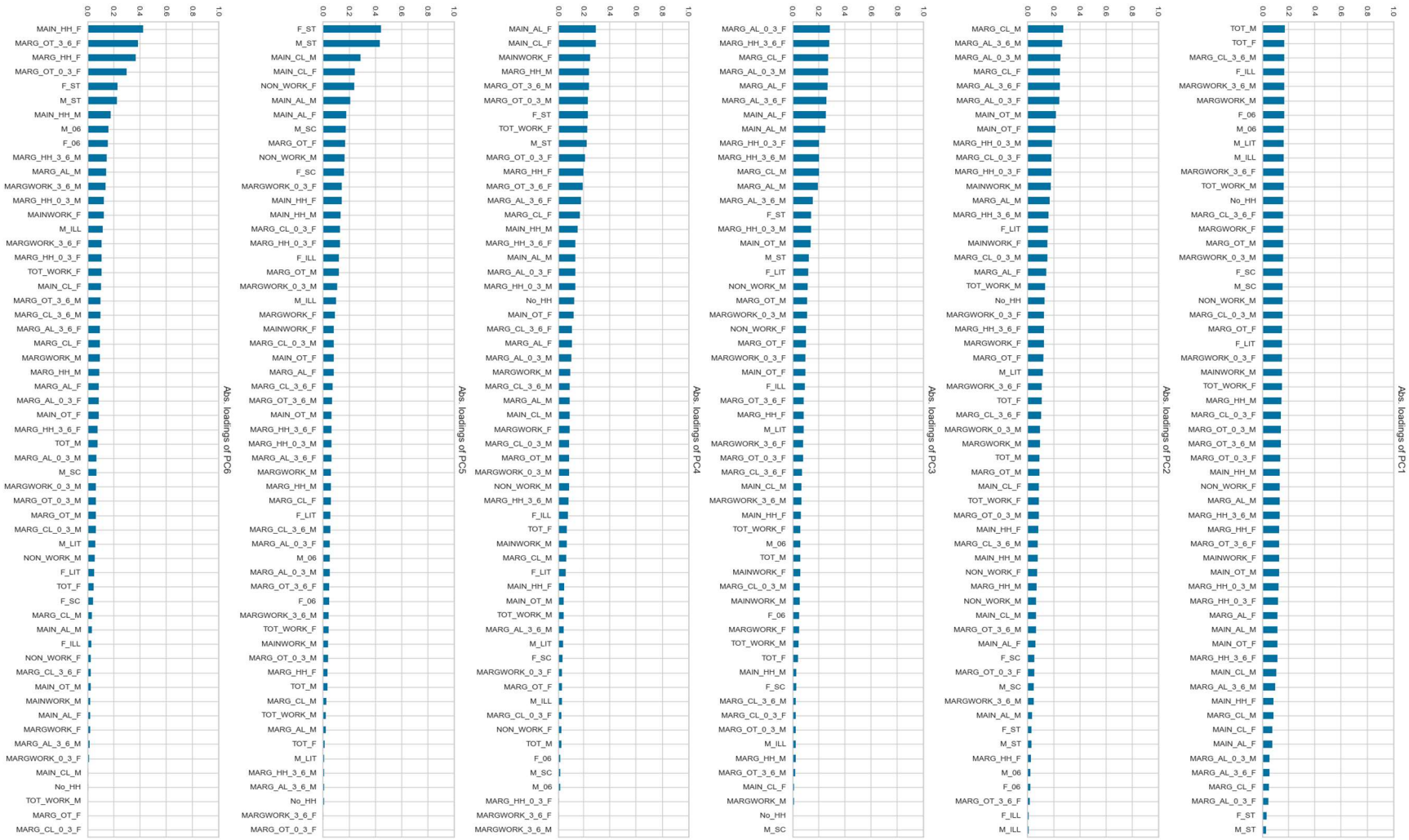
- Check the cumulative explained variance ratio

```
array([0.55726063, 0.69510499, 0.76785794, 0.83212212, 0.87077261,
       0.9047243 , 0.92532669, 0.93848433, 0.94929292, 0.95854687,
       0.96607599, 0.97226701, 0.97745473, 0.98238168, 0.98574761,
       0.98813454, 0.99012071, 0.99198278, 0.99368693, 0.99509011,
       0.99609921, 0.99687687, 0.99754058, 0.9980597 , 0.99853404,
       0.99894473, 0.99919891, 0.99939134, 0.9995545 , 0.99969701,
       0.99983525, 0.99992329, 0.9999688 , 0.9999875 , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        ])
```

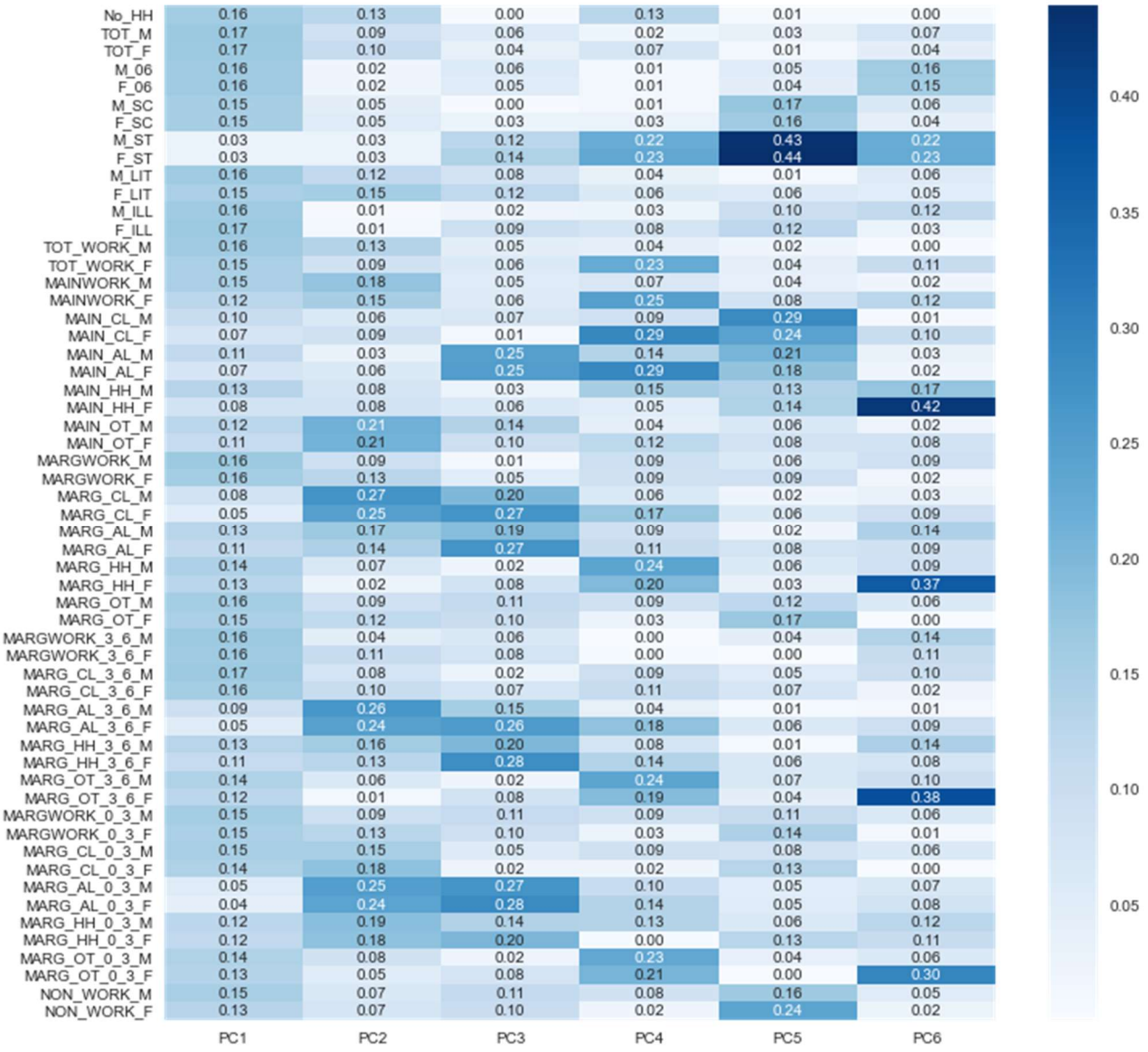
With above analysis with PC6 components we have captured 90% of the data

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

   - We will plot to look in to the magnitude of the coefficients by each PC
   - PCs with different characteristics can be studied from the bar plot below


MAIN_HH_F
MARG_OT_3_6_F
MARG_HH_F
MARG_OT_0_3_F
F_ST
M_ST
MAIN_HH_M
M_06
F_06
MARG_HH_3_6_M
MARG_AL_M
MARGWORK_3_6_M
MARG_HH_0_3_M
MAINWORK_F
M_ILL
MARGWORK_3_6_F
MARG_HH_0_3_F
TOT_WORK_F
MAIN_CL_F
MARG_OT_3_6_M
MARG_CL_3_6_M
MARG_AL_3_6_F
MARG_CL_F
MARGWORK_M
MARG_HH_M
MARG_AL_F
MARG_AL_0_3_F
MAIN_OT_F
MARG_HH_3_6_F
TOT_M
MARG_AL_0_3_M
M_SC
MARGWORK_0_3_M
MARG_OT_0_3_M
MARG_OT_M
MARG_CL_0_3_M
M_LIT
NON_WORK_M
F_LIT
TOT_F
F_SC
MARG_CL_M
MAIN_AL_M
F_ILL
NON_WORK_F
MARG_CL_3_6_F
MAIN_OT_M
MAINWORK_M
MAIN_AL_F
MARGWORK_F
MARG_AL_3_6_M
MARGWORK_0_3_F
MAIN_CL_M
No_HH
TOT_WORK_M
MARG_OT_F
MARG_CL_0_3_F
Abs. loadings of PC6
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
MAIN_HH_F
MAIN_HH_M
MARG_CL_0_3_F
MARG_HH_0_3_F
F_ILL
MARG_OT_M
MARGWORK_0_3_M
M_ILL
MARGWORK_F
MAINWORK_F
MARG_CL_0_3_M
MAIN_OT_F
MARG_AL_F
MARG_CL_3_6_F
MARG_OT_3_6_M
MAIN_OT_M
MARG_HH_3_6_F
MARG_HH_0_3_M
MARG_AL_3_6_F
MARGWORK_M
MARG_HH_M
MARG_CL_F
F_LIT
MARG_CL_3_6_M
MARG_AL_0_3_F
M_06
MARG_AL_0_3_M
MARG_OT_3_6_F
F_06
MARGWORK_3_6_M
TOT_WORK_F
MAINWORK_M
MARG_OT_0_3_M
MARG_HH_F
TOT_M
MARG_CL_M
TOT_WORK_M
MARG_AL_M
TOT_F
M_LIT
MARG_HH_3_6_M
MARG_AL_3_6_M
No_HH
MARGWORK_3_6_F
MARG_OT_0_3_F
Abs. loadings of PC5
MAIN_AL_F
MAIN_CL_F
MAINWORK_F
MARG_HH_M
MARG_OT_3_6_M
MARG_OT_0_3_M
F_ST
TOT_WORK_F
M_ST
MARG_OT_0_3_F
MARG_HH_F
MARG_OT_3_6_F
MARG_AL_3_6_F
MARG_CL_F
MAIN_HH_M
MARG_HH_3_6_F
MAIN_AL_M
MARG_AL_0_3_F
MARG_HH_0_3_M
No_HH
MAIN_OT_F
MARG_CL_3_6_F
MARG_AL_F
MARG_AL_0_3_M
MARGWORK_M
MARG_CL_3_6_M
MARG_AL_M
MAIN_CL_M
MARGWORK_F
MARG_CL_0_3_M
MARG_OT_M
MARGWORK_0_3_M
NON_WORK_M
MARG_HH_3_6_M
F_ILL
TOT_F
MAINWORK_M
MARG_CL_M
F_LIT
MAIN_HH_F
MAIN_OT_M
TOT_WORK_M
MARG_AL_3_6_M
M_LIT
F_SC
MARGWORK_0_3_F
MARG_OT_F
M_ILL
MARG_CL_0_3_F
NON_WORK_F
TOT_M
F_06
M_SC
M_06
MARG_HH_0_3_F
MARGWORK_3_6_F
MARGWORK_3_6_M
Abs. loadings of PC4
MARG_AL_0_3_F
MARG_HH_3_6_F
MARG_CL_F
MARG_AL_0_3_M
MARG_AL_F
MARG_AL_3_6_F
MAIN_AL_F
MAIN_AL_M
MARG_HH_0_3_F
MARG_HH_3_6_M
MARG_CL_M
MARG_AL_M
MARG_AL_3_6_M
F_ST
MARG_HH_0_3_M
MAIN_OT_M
M_ST
F_LIT
NON_WORK_M
MARG_OT_M
MARGWORK_0_3_M
NON_WORK_F
MARG_OT_F
MARGWORK_0_3_F
MAIN_OT_F
F_ILL
MARG_OT_3_6_F
MARG_HH_F
M_LIT
MARGWORK_3_6_F
MARG_OT_0_3_F
MARG_CL_3_6_F
MAIN_CL_M
MARGWORK_3_6_M
MAIN_HH_F
TOT_WORK_F
M_06
TOT_M
MAINWORK_F
MARG_CL_0_3_M
MAINWORK_M
F_06
MARGWORK_F
TOT_WORK_M
TOT_F
MAIN_HH_M
F_SC
MARG_CL_3_6_M
MARG_CL_0_3_F
MARG_OT_0_3_M
M_ILL
MARG_HH_M
MARG_OT_3_6_M
MAIN_CL_F
MARGWORK_M
No_HH
M_SC
Abs. loadings of PC3
MARG_CL_M
MARG_AL_3_6_M
MARG_AL_0_3_M
MARG_CL_F
MARG_AL_3_6_F
MARG_AL_0_3_F
MAIN_OT_M
MAIN_OT_F
MARG_HH_0_3_M
MARG_CL_0_3_F
MARG_HH_0_3_F
MAINWORK_M
MARG_AL_M
MARG_HH_3_6_M
F_LIT
MAINWORK_F
MARG_CL_0_3_M
MARG_AL_F
TOT_WORK_M
No_HH
MARGWORK_0_3_F
MARG_HH_3_6_F
MARGWORK_F
MARG_OT_F
M_LIT
MARGWORK_3_6_F
TOT_F
MARG_CL_3_6_F
MARGWORK_0_3_M
MARGWORK_M
TOT_M
MARG_OT_M
MAIN_CL_F
TOT_WORK_F
MARG_OT_0_3_M
MAIN_HH_F
MARG_CL_3_6_M
MAIN_HH_M
NON_WORK_F
MARG_HH_M
NON_WORK_M
MAIN_CL_M
MARG_OT_3_6_M
MAIN_AL_F
F_SC
MARG_OT_0_3_F
M_SC
MARGWORK_3_6_M
MAIN_AL_M
F_ST
M_ST
MARG_HH_F
M_06
F_06
MARG_OT_3_6_F
F_ILL
M_ILL
Abs. loadings of PC2
TOT_M
TOT_F
MARG_CL_3_6_M
F_ILL
MARGWORK_3_6_M
MARGWORK_M
F_06
M_06
M_LIT
M_ILL
MARGWORK_3_6_F
TOT_WORK_M
No_HH
MARG_CL_3_6_F
MARGWORK_F
MARG_OT_M
MARGWORK_0_3_M
F_SC
M_SC
NON_WORK_M
MARG_CL_0_3_M
MARG_OT_F
F_LIT
MARGWORK_0_3_F
MAINWORK_M
TOT_WORK_F
MARG_HH_M
MARG_CL_0_3_F
MARG_OT_0_3_M
MARG_OT_3_6_M
MARG_OT_0_3_F
MAIN_HH_M
MAIN_HH_F
MARG_HH_3_6_M
MARG_AL_M
NON_WORK_F
MAIN_OT_M
MARG_HH_0_3_M
MARG_HH_0_3_F
MARG_AL_F
MAIN_AL_M
MAIN_OT_F
MARG_HH_3_6_F
MAIN_CL_M
MARG_AL_3_6_M
MAIN_HH_F
MARG_CL_M
MAIN_CL_F
MAIN_AL_F
MARG_AL_0_3_M
MARG_AL_3_6_F
MARG_CL_F
MARG_AL_0_3_F
F_ST
M_ST
Abs. loadings of PC1

- Using heat map we can compare how the original features influence various PCs



Observations

- PC1 has all the features of almost equal value
- PC2 has high of Marginal Cultivator Population Male and Marginal Agriculture Labourers Population 3-6 male
- PC3 has high values of Marginal Agriculture Labourers Population 0-3 Male & female, Marginal House hold 3-6 male & Female. Marginal Agriculture Labourers Population 3-6 Female
- PC4 has high vale of Marginal Other Workers Population 0-3 Male & Female. Marginal Other Workers Population 3-6 Male & Female. Marginal House hold Male & Female , Total Work force female, Main Work force Female, Main Cultivator Population Female, Main Agriculture Labourers Population Female

- PC 5 had high values of Scheduled Castes population Male & Female, Main Cultivator Population Male Non Working Population Male & female
- PC6 has high vales of Main Household Industries Population Female, Marginal Household Industries Population Female, Marginal Other Workers Population Person 3-6 Female, Marginal Other Workers Population 0-3 Female

7. **PCA: Write linear equation for first PC.**

Transforming Data in to Principle component (PC Scores) for

Input

```
pc_score = np.dot(df_selected['PC1'], df_pca_scaled.iloc[0])
print(round(pc_score, 6), end = ' ')
```

Output : -4.617263

Using heatmap on the transferred data, we can see co-relation is significantly reduced