**Data Mining Project**

Student Name : Suneel Kumar Pentapalli

Date : 04/June/2023



Contents

[***Clustering*** 3](#_Toc137405607)

[Part 1 - Clustering: Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc. 3](#_Toc137405608)

[Part 1 - Clustering: Treat missing values in CPC, CTR and CPM using the formula given. 5](#_Toc137405609)

[Part 1 - Clustering: 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). 6](#_Toc137405610)

[Part 1 - Clustering: Perform z-score scaling and discuss how it affects the speed of the algorithm. 8](#_Toc137405611)

[Part 1 - Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance. 9](#_Toc137405612)

[Part 1 - Clustering: Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm. 10](#_Toc137405613)

[Part 1 - Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters. 11](#_Toc137405614)

[Part 1 - 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. Make bar plots]. 12](#_Toc137405615)

[Part 1 - Clustering: Conclude the project by providing summary of your learnings. 12](#_Toc137405616)

[***PCA(Princial Component Analysis)*** 15](#_Toc137405617)

[PCA: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc. 15](#_Toc137405618)

[PCA: Perform detailed Exploratory analysis by creating certain questions like 18](#_Toc137405619)

[PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary? 21](#_Toc137405620)

[PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment. 22](#_Toc137405621)

[PCA: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector. 27](#_Toc137405622)

[PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot. 28](#_Toc137405623)

[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. 29](#_Toc137405624)

[PCA: Write linear equation for first PC. 33](#_Toc137405625)

# Clustering

**Digital Ads Data:**

**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. Note that the Total Campaign Spend refers to the 'Spend' Column in the dataset and the Number of Impressions refers to the 'Impressions' Column in the dataset.**

**CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column in the dataset and the Number of Clicks refers to the 'Clicks' Column in the dataset.**

**CTR = Total Measured Clicks / Total Measured Ad Impressions x 100. Note that the Total Measured Clicks refers to the 'Clicks' Column in the dataset and the Total Measured Ad Impressions refers to the 'Impressions' Column in the dataset.**

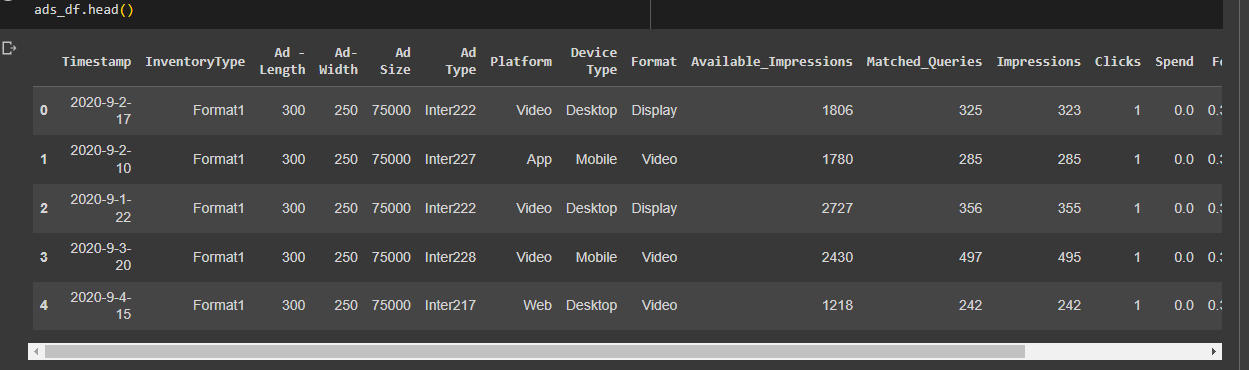
### Part 1 - Clustering: Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.

The given dataset "Clustering Clean Ads\_Data-2.xlsx" has

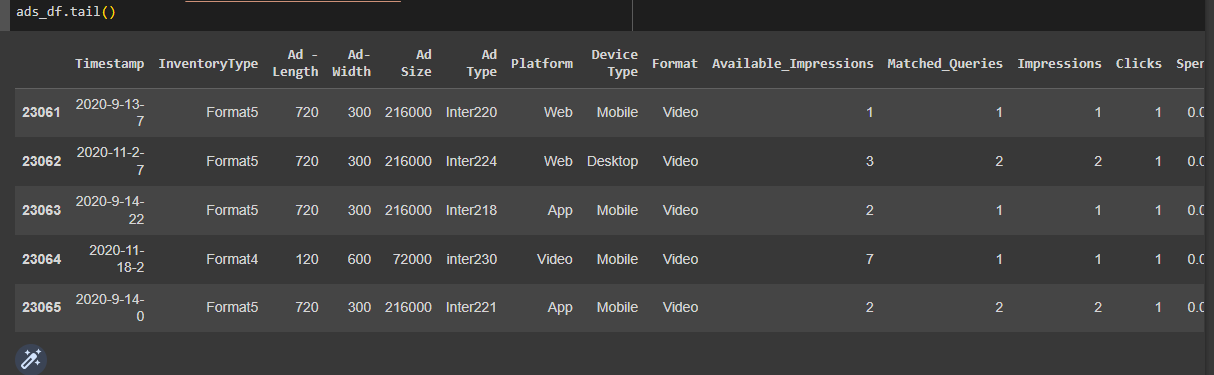
* 23066 records and 19 columns.
* 4376 Null entries in three columns CTR, CPM, CPC.
* No duplicate records.

Below is the head and tail of the given data,

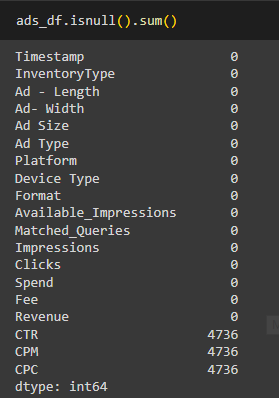
First five records based on index



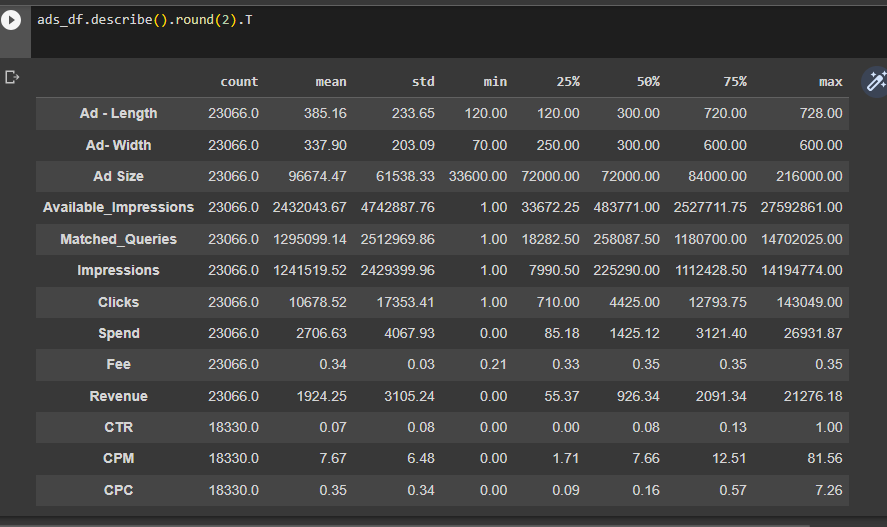
Last five records based on index



Null counts



Summary of all numeric columns



### Part 1 - Clustering: Treat missing values in CPC, CTR and CPM using the formula given.

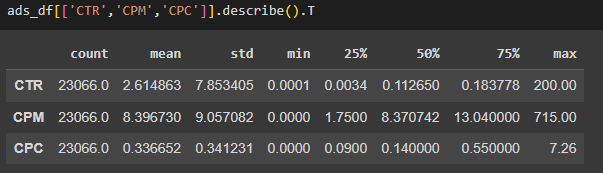
Treated all missing values with the given formulae below,

CPM = (Total Campaign Spend / Number of Impressions) \* 1,000. Note that the Total Campaign Spend refers to the 'Spend' Column in the dataset and the Number of Impressions refers to the 'Impressions' Column in the dataset.

CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column in the dataset and the Number of Clicks refers to the 'Clicks' Column in the dataset.

CTR = Total Measured Clicks / Total Measured Ad Impressions x 100. Note that the Total Measured Clicks refers to the 'Clicks' Column in the dataset and the Total Measured Ad Impressions refers to the 'Impressions' Column in the dataset.

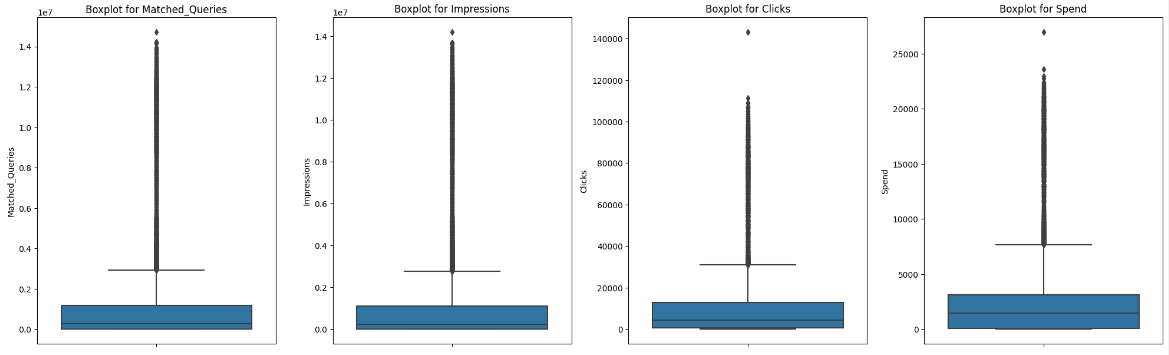
Here is the summary of CPC, CTR and CPM after treating missing values, Please refer attached Suneel\_Kumar\_04\_June\_2023\_Data\_Mining.ipynb for more details.



### Part 1 - Clustering: 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 shown below, data has outliers in the columns Ad Size, Available\_Impressions, Matched\_Queries, Impressions, Clicks, Spend, Fee, Revenue, CTR, CPM and CPC.





A picture containing diagram, screenshot, line, rectangle

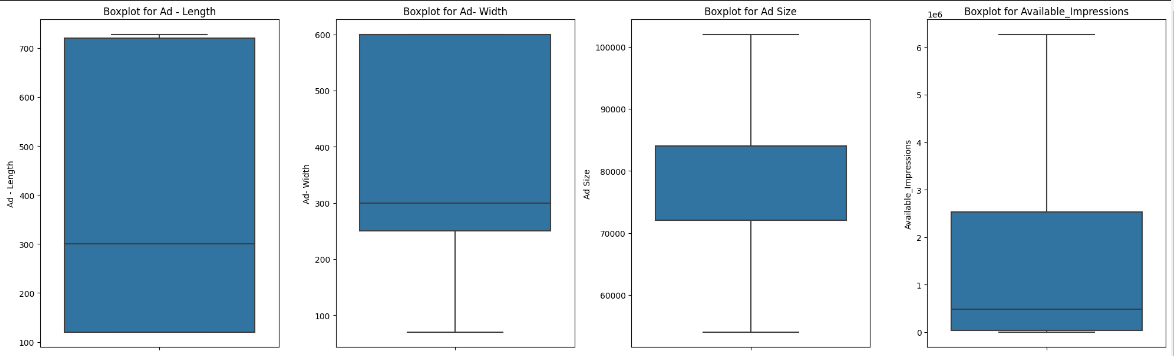
Description automatically generated

A picture containing screenshot, text, diagram, rectangle

Description automatically generated

K-Means clustering is sensitive to outliers, so its better to treat outliers.

Hence lets treat the outliers using IQR method. Below is the barplot of all the above columns after treating outliers.



A picture containing rectangle, diagram, line, screenshot

Description automatically generated

A picture containing diagram, screenshot, rectangle, line

Description automatically generated

A picture containing screenshot, diagram, text, rectangle

Description automatically generated

### Part 1 - Clustering: Perform z-score scaling and discuss how it affects the speed of the algorithm.

Time taken for Clustering before scaling is as follows,

Started at 2023-06-07 10:14:03.662651

Ends at 2023-06-07 10:14:53.838568

Time taken for Clustering after applying zscaling is as follows,

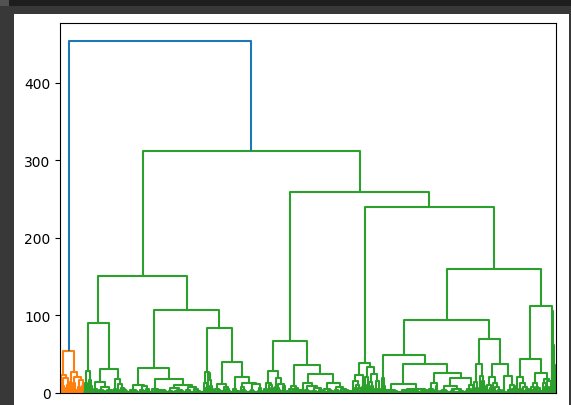
Started at 2023-06-07 10:20:00.761875

Ends at 2023-06-07 10:20:46.028056

This shows that scaled data took relatively less time than that of unscaled data. Though the time difference is very less, when applied the same to larger datasets the difference might extrapolate.

### Part 1 - Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.

Below is the dendogram of scaled data constructed using ward linkage and Euclidean distance method.



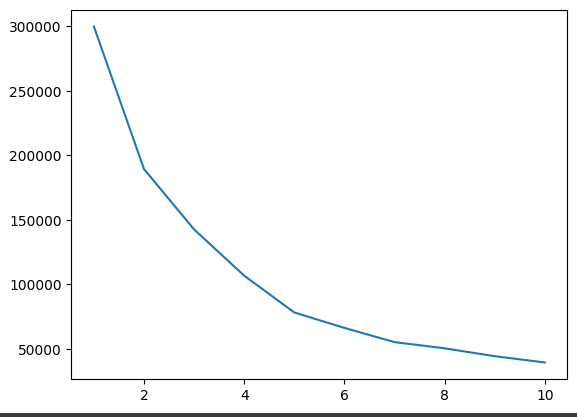
### Part 1 - Clustering: Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.

A screenshot of a computer screen

Description automatically generated with low confidence

As per above dendrogram, suggested optimum number of clusters is 2 for hierarchal and KMeans clustering.

Even the below screeplot of (Within sum of squares)wss scores shows that there is no drop in cluster variance after cluster count 2.

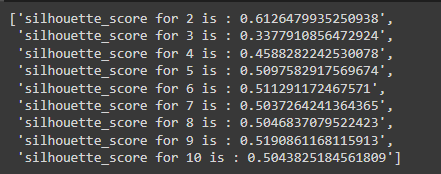


Hence we can go ahead with number of clusters =2 as both of the above graphs suggests 2 to be the optimum number of clusters.

### Part 1 - Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters.

The value of the Silhouette score varies from -1 to 1. If the score is near to 1, the cluster is dense and well-separated than other clusters

Lets see the silhouette scores of the give ads data for upto 10 clusters.

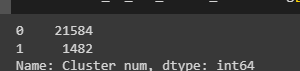


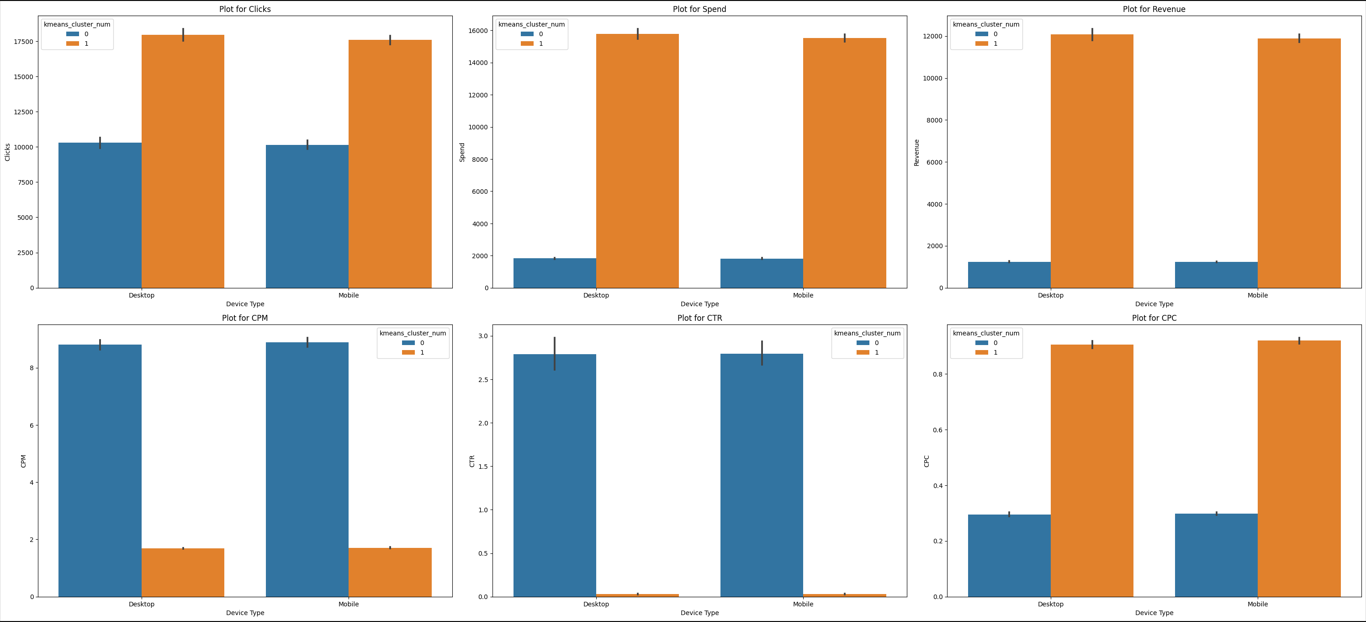
As seen above, silhouette score for 2 clusters is the maximum of all. Hence we can conclude 2 to be the optimum number of clusters for our clustering.

### Part 1 - 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. Make bar plots].

Lets apply KMeans clustering on the data with number of clusters as 2.

After clustering, below is the count of records in each cluster,





Ref this for clear graph: 

In the above graph, Blue bars indicate the values in cluster “0” and the orange bars represents values in cluster “1”.

Each graph represents trends in columns 'Clicks','Spend','Revenue','CPM','CTR','CPC' across device\_types “Desktop” and “Mobile”.

### Part 1 - Clustering: Conclude the project by providing summary of your learnings.

Inferences:

Now we have clusters 0(21584 records) and 1(1482 records). The trends in each cluster is as follows,

Clicks:

We can see there are more number of clicks coming from Cluster1 in both Desktop and Mobile with an average of 17963 and 17581 respectively.

This means we have successfully segmented data into two groups where Cluster1 represents ads with highest number of clicks and Cluster0 represents ads with less number of clicks with an average of 10292 clicks in Desktop and 10140 clicks in Mobile.

Spend:

Even data has been separated with respect to spends, Cluster0 with 21584 records has very less amount spent in ad variations with an average of 1820 and cluster1 with 1482 records has highest average amount of 15618 spent in ad variations.

This clearly shows that the amount spent in ad variation might have helped in higher number of clicks in Cluster 1.

Revenue:

As the number of clicks in cluster1 is high, hence the income earned from this cluster is also more, with an average of 11958.

The average revenue generated from cluster0 is 1235 which is way less than that of revenue generated from cluster1

CPM(Cost per thousand impressions):

Average cost per thousand impressions in cluster0 is 8.8

Average cost per thousand impressions in cluster1 is 1.7

Cluster1 has better spend to impression ratio which means we are spending less for a impression in cluster1 ads when compared to spends per impression in cluster0

CTR(Click through rate):

Average Click through rate in cluster0 is 2.79

Average Click through rate in cluster1 is 0.029

This means, for every time when a ad is shown, ads in cluster0 got an average of 2.79 clicks where as cluster1 got only 0.029 clicks.

CPC(Cost per click):

Average Cost per click in cluster0 is 0.29

Average Cost per click in cluster1 is 0.91

This means for every ad click we have to put 0.29 in cluster0 and 0.91 in cluster1. This is inline with the spend analysis which says the spends in cluster1 is higher when compared to spends in cluster0.

cluster0 To Conclude, we can increase the ads revenue in cluster0 by spending more in ad variations as click through rate in this is already good.

The revenue from cluster1 can further be increased if we can increase CTR somehow.

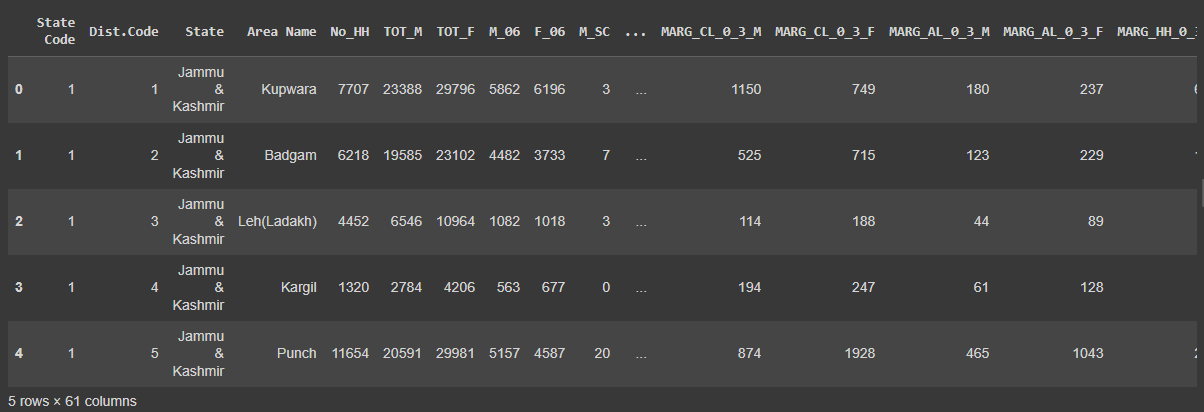
# PCA(Princial Component Analysis)

**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.**

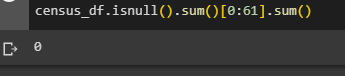
### PCA: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.

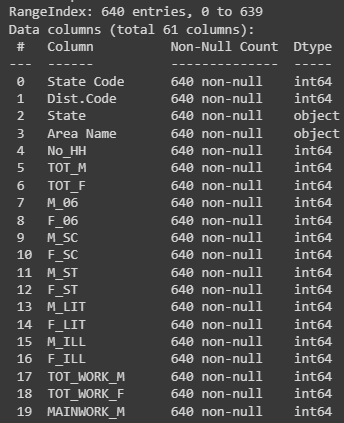
Given dataset has 640 records with 61 columns which holds census data of males and females.

Sample Data:



Of all 61 columns, only “State” and “Area Name” are of object data type, rest of the columns are all numeric type.





A screenshot of a computer

Description automatically generated

A picture containing text, screenshot, font, menu

Description automatically generated

The dataset doesn’t have any row level duplicates.

A screen shot of a computer

Description automatically generated with low confidence

None of the columns have null/invalid values.

A screen shot of a computer

Description automatically generated with low confidence

Please refer Suneel\_Kumar\_04\_June\_2023\_Data\_Mining.ipynb for more details.

### PCA: Perform detailed Exploratory analysis by creating certain questions like

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

**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**

The given dataset doesn’t have gender ratio column, so lets calculate the same using the below formula,

Gender\_Ratio(M/F) = male population(TOT\_M)/(female pupluation(TOT\_F))

State with least gender ratio is **Andhra Pradesh** as shown below,

A screenshot of a computer

Description automatically generated with low confidence

Similary, state with highest gender ratio is **Lakshadweep**,

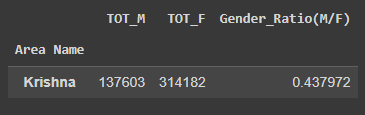
A screenshot of a computer

Description automatically generated with medium confidence

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

Afadfsfsfsgregregr

Area with least gender ratio is **Krishna** as shown below,



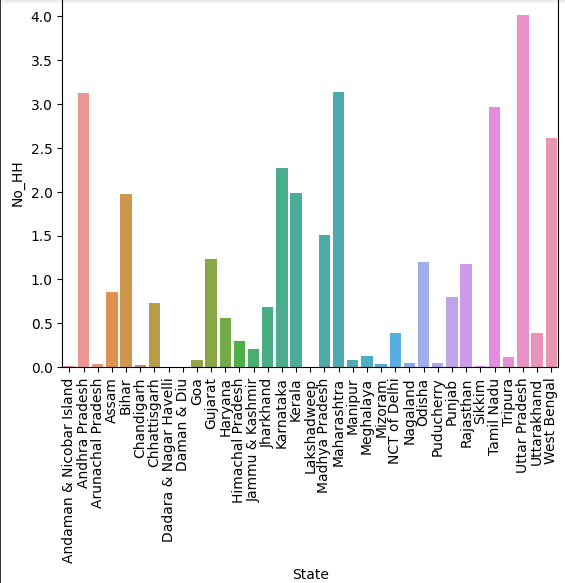
Similary, area with highest gender ratio is **Lakshadweep**,

A screenshot of a computer

Description automatically generated with medium confidence

1. **State with highest number of households**

State Uttar Pradesh has the highest number of households of 4006871



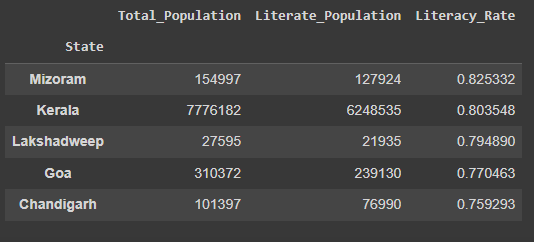
1. **States highest/lowest literacy rate**

Literacy rate is the proportion of literate people out of total population.

Literacy rate = Literate\_Population/Total\_Population

Lets add the above column to census data and see top 5 states with highest and least literacy rates,

States with highest literacy rate,



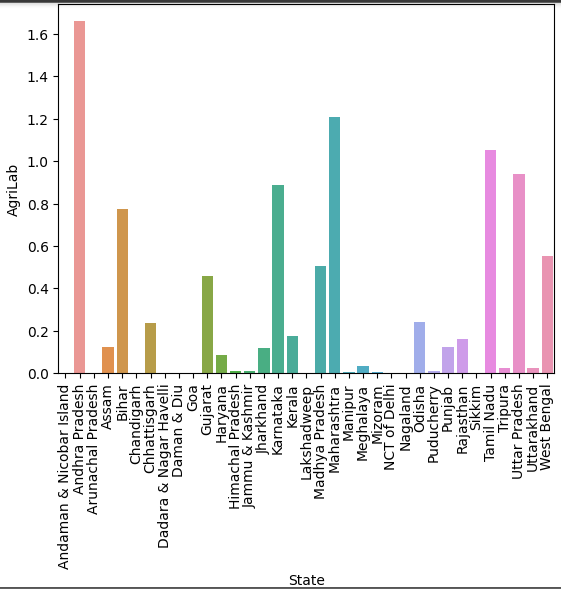
States with least literacy rates,

A screenshot of a computer

Description automatically generated with medium confidence

1. **State with most agricultural labours**

As shown below, Andhra Pradesh is the state with highest number of agricultural labours or 1659383.



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

In General, PCA is sensitive to outliers, its always better to treat outliers before proceeding with PCA. But in this case, the outliers in data are valid, as different areas will have different population based on area and other factors, hence in this case its good to proceed with outliers.

All the below numeric columns are having outliers. Please refer to this attachement for more details.

'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',

'MARGWORK\_M', 'MARGWORK\_F', 'MARG\_CL\_M', 'MARG\_CL\_F', 'MARG\_AL\_M',

'MARG\_AL\_F', 'MARG\_HH\_M', 'MARG\_HH\_F', 'MARG\_OT\_M', 'MARG\_OT\_F',

'MARGWORK\_3\_6\_M', 'MARGWORK\_3\_6\_F', 'MARG\_CL\_3\_6\_M', 'MARG\_CL\_3\_6\_F',

'MARG\_AL\_3\_6\_M', 'MARG\_AL\_3\_6\_F', 'MARG\_HH\_3\_6\_M', 'MARG\_HH\_3\_6\_F',

'MARG\_OT\_3\_6\_M', 'MARG\_OT\_3\_6\_F', 'MARGWORK\_0\_3\_M', 'MARGWORK\_0\_3\_F',

'MARG\_CL\_0\_3\_M', 'MARG\_CL\_0\_3\_F', 'MARG\_AL\_0\_3\_M', 'MARG\_AL\_0\_3\_F',

'MARG\_HH\_0\_3\_M', 'MARG\_HH\_0\_3\_F', 'MARG\_OT\_0\_3\_M', 'MARG\_OT\_0\_3\_F',

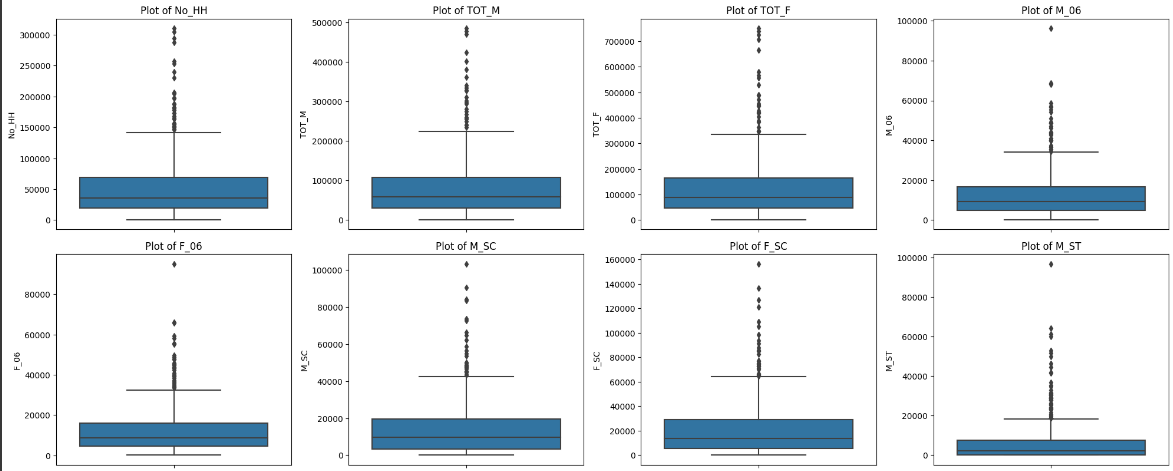
'NON\_WORK\_M', 'NON\_WORK\_F', 'Total\_Population', 'Literate\_Population'

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

Outliers can skew a probability distribution and make data scaling using standardization difficult as the calculated mean and standard deviation will be skewed by the presence of the outliers.

So, even if we scale data without treating outliers, same outliers will be projected in scaled data too as shown below.

Boxplot before scaling:



A picture containing diagram, rectangle, plan, screenshot

Description automatically generated

A picture containing diagram, text, plan, rectangle

Description automatically generated

A picture containing diagram, plan, rectangle, line

Description automatically generated

A picture containing diagram, screenshot, text, rectangle

Description automatically generated

A picture containing diagram, text, plan, screenshot

Description automatically generated

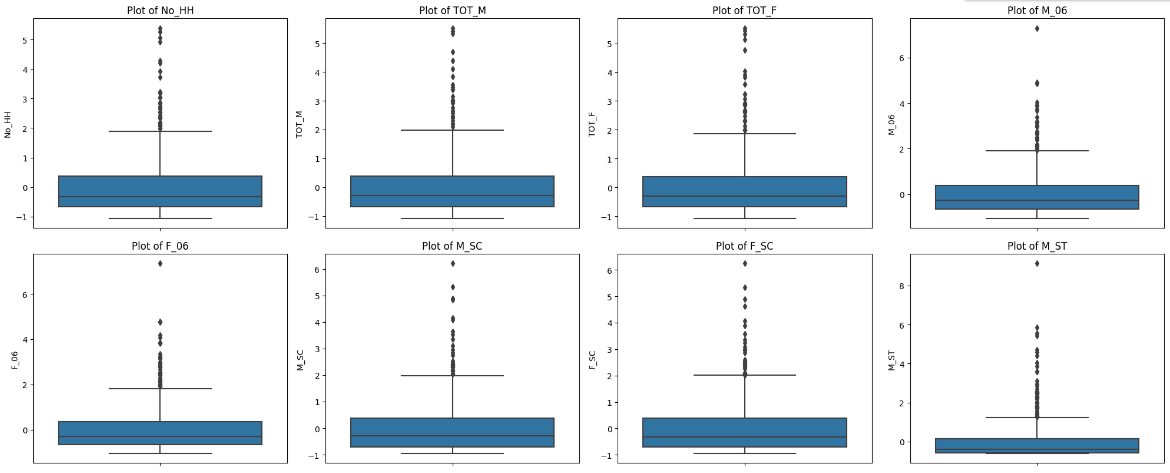
A picture containing diagram, line, screenshot, plan

Description automatically generated

A picture containing diagram, line, plot, screenshot

Description automatically generated

Scaled census data boxplot:



A picture containing diagram, plan, rectangle, screenshot

Description automatically generated

A picture containing diagram, plan, rectangle, technical drawing

Description automatically generated

A picture containing diagram, rectangle, line, plan

Description automatically generated

A picture containing diagram, plan, technical drawing, rectangle

Description automatically generated

A picture containing diagram, rectangle, plan, screenshot

Description automatically generated

A picture containing diagram, line, parallel, rectangle

Description automatically generated

A picture containing diagram, line, rectangle, plot

Description automatically generated

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

Step 1

Verify if PCA is feasable for the given dataset

Lets perform KMO test to test how good the data is to perform PCA. If MSA value greater than 0.7 then PCA will be recommended.

Confirm the adequacy of sample size.

Note: Above 0.7 is good, below 0.5 is not acceptable

From the scaled census data, we got MSA value of 0.8, which means the given data sample is adequate and good for PCA.

Step 2

Confirm the statistical significance of correlations

H0: Correlations are not significant, H1: There are significant correlations

Reject H0 if p-value < 0.05

To check the correlations, we can perform bartlett\_sphericity test and find p\_value.

Upon passing our scaled census df, we got 0 as p\_value from above test, which means we can reject null hypothesis and hence correlations are signification.

To conclude, both the above tests suggest us that the given data is good to perfom PCA and we can expect considerable reduction in dimensions.

Step 3

Apply PCA with all the 57 numeric columns. Please refer to Suneel\_Kumar\_04\_June\_2023\_Data\_Mining.ipynb for more details like exact eigen values and vectors.

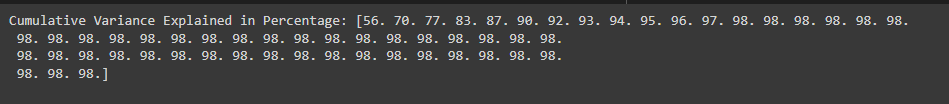
### PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.

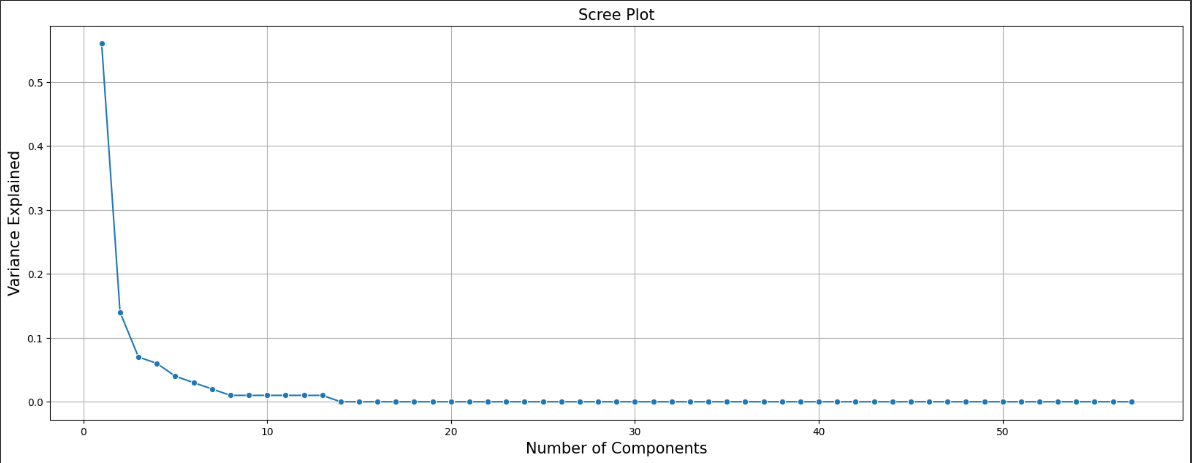
To find the optimum number of PC’s, lets fetch cumulative explained variance for all the 57 PCA’s first.

Now, we have below explained variances, but we need to consider atleast 90% explained variance. 90% explained variance can be obtained from the first six PCA components as shown below.

Hence we proceed with optimum number of clusters as 6.

Same can be observed in the screeplot of explained variances below.





### 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.

Let's identify which features have maximum loading across the components.

We will first plot the component loading on a heatmap.

For each feature, we find the maximum loading value across the components and mark the same with help of rectangular box.

Features marked with rectangular red box are the one having maximum loading on the respective component. We consider these marked features to decide the context that the component represents

A picture containing text, screenshot, parallel, line

Description automatically generated

A picture containing text, screenshot, number, colorfulness

Description automatically generated

Above is the list of columns which each PC is representing.

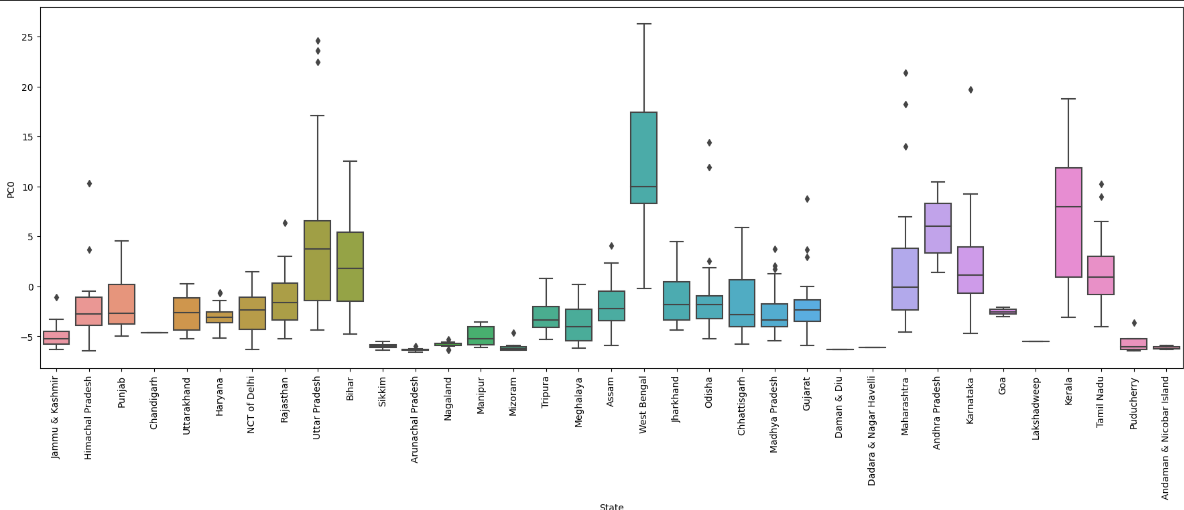
To Evaluator: From the above list I am not able to identify how I can name my PC’s, as each PC has mix of all types of columns, hence proceeding default PC name. Do let me know in the feedback on how I can name the PC’s in such scenarios.

Lets plot barplots of each PC component against states,

PC0:

As PC0 represents the columns No\_HH,TOT\_M,TOT\_F,M\_06,F\_06,M\_LIT,F\_LIT,M\_ILL,F\_ILL,TOT\_WORK\_M,MARGWORK\_M,MARGWORK\_F,MARG\_OT\_M,MARGWORK\_3\_6\_M,MARGWORK\_3\_6\_F,MARG\_CL\_3\_6\_M,MARG\_CL\_3\_6\_F,MARGWORK\_0\_3\_M,MARGWORK\_0\_3\_F,MARG\_CL\_0\_3\_M,

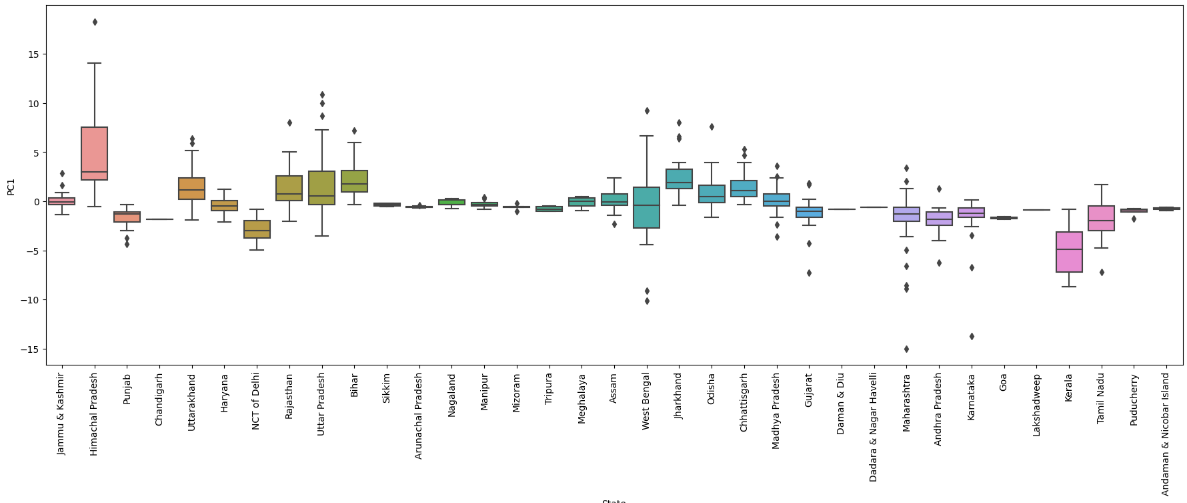
State west bengal has the highest value in the above columns when taken in average.



PC1:

As PC1 represents the columns MAINWORK\_M,MAIN\_OT\_M,MAIN\_OT\_F,MARG\_CL\_M,MARG\_AL\_3\_6\_M,MARG\_CL\_0\_3\_F,MARG\_HH\_0\_3\_M

Below chart shows Himachal Pradesh has the average highest values in above columns.



PC2:

As PC2 represents the columns MAIN\_AL\_M,MARG\_CL\_F,MARG\_AL\_M,MARG\_AL\_F,MARG\_AL\_3\_6\_F,MARG\_HH\_3\_6\_M,MARG\_HH\_3\_6\_F,MARG\_AL\_0\_3\_M,MARG\_AL\_0\_3\_F,MARG\_HH\_0\_3\_F

Below chart shows Himachal Pradesh has the average highest values in above columns.

A picture containing diagram, line, plot

Description automatically generated

PC3:

On an average, state west Bengal and uttar pradesh has the highest TOT\_WORK\_F,MAINWORK\_F,MAIN\_CL\_F,MAIN\_AL\_F,MARG\_HH\_M,MARG\_OT\_3\_6\_M,MARG\_OT\_0\_3\_M

A picture containing diagram, line, plot, plan

Description automatically generated

PC4:

On an average, State kerala has the highest values in columns M\_SC,F\_SC,M\_ST,F\_ST,MAIN\_CL\_M,MARG\_OT\_F,NON\_WORK\_M,NON\_WORK\_F

A picture containing diagram, line, plot, plan

Description automatically generated

PC5:

On an average, State west Bengal and Himachal Pradesh has the highest values in columns MAIN\_HH\_M,MAIN\_HH\_F,MARG\_HH\_F,MARG\_OT\_3\_6\_F,MARG\_OT\_0\_3\_F

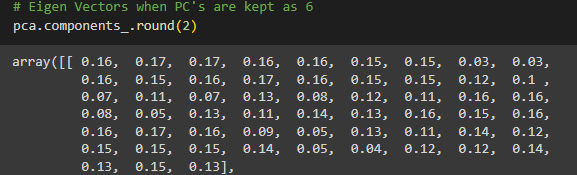
A picture containing diagram, line, plot, text

Description automatically generated

### PCA: Write linear equation for first PC.

Below is the linear algebraic equation for the First PCA component.

Eigen vectors of PC0 are as below,



Linear equation using the above eigen vectors for PC0,

( 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 ) \* 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

Upon solving the equation by passing our scaled df values for first row, we will get value of PC0 for first row, which is equal to -4.62

( 0.16 ) \* -0.904737544272622 + ( 0.17 ) \* -0.771236453673318 + ( 0.17 ) \* -0.815562582036951 + ( 0.16 ) \* -0.561011510675832 + ( 0.16 ) \* -0.507738389707144 + ( 0.15 ) \* -0.958574597291519 + ( 0.15 ) \* -0.957048592525633 + ( 0.03 ) \* -0.42330550019368 + ( 0.03 ) \* -0.476423165654139 + ( 0.16 ) \* -0.798097454907794 + ( 0.15 ) \* -0.733477489550826 + ( 0.16 ) \* -0.604014646963687 + ( 0.17 ) \* -0.798229034600876 + ( 0.16 ) \* -0.859260382156752 + ( 0.15 ) \* -1.01023773336172 + ( 0.15 ) \* -0.872366958468956 + ( 0.12 ) \* -0.898215543071509 + ( 0.1 ) \* -1.04284361377049 + ( 0.07 ) \* -0.986630022534231 + ( 0.11 ) \* -0.851060119084042 + ( 0.07 ) \* -0.683276475397626 + ( 0.13 ) \* -0.630765981349575 + ( 0.08 ) \* -0.40755497183694 + ( 0.12 ) \* -0.624041570867164 + ( 0.11 ) \* -0.611637223937407 + ( 0.16 ) \* -0.516942750671118 + ( 0.16 ) \* -0.966511637782547 + ( 0.08 ) \* -0.321808839601663 + ( 0.05 ) \* -0.485053439616662 + ( 0.13 ) \* -0.331426055233964 + ( 0.11 ) \* -0.860192154179759 + ( 0.14 ) \* -0.37798415705244 + ( 0.13 ) \* -0.453026391793353 + ( 0.16 ) \* -0.548763563204698 + ( 0.15 ) \* -0.614625185724889 + ( 0.16 ) \* -0.648040441222254 + ( 0.16 ) \* -0.663795326568343 + ( 0.17 ) \* -0.595997806921944 + ( 0.16 ) \* -1.0178478914467 + ( 0.09 ) \* -0.387707189232455 + ( 0.05 ) \* -0.563853855728662 + ( 0.13 ) \* -0.448658286439496 + ( 0.11 ) \* -0.896722522644112 + ( 0.14 ) \* -0.377634759674543 + ( 0.12 ) \* -0.431307366748112 + ( 0.15 ) \* -0.569151073215756 + ( 0.15 ) \* -0.612451352674452 + ( 0.15 ) \* -0.163229061031329 + ( 0.14 ) \* -0.720610096696927 + ( 0.05 ) \* -0.156494116071813 + ( 0.04 ) \* -0.287524426316545 + ( 0.12 ) \* 0.156577477469688 + ( 0.12 ) \* -0.657411585675081 + ( 0.14 ) \* -0.365258163281536 + ( 0.13 ) \* -0.499976737575042 + ( 0.15 ) \* -0.413052510952805 + ( 0.13 ) \* -0.539614389540428

