##### Project Report

DATA MINING

Contents

[Problem Statement 1: Clustering: Digital Ads Data 4](#_Toc116821888)

[1.1 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. 4](#_Toc116821889)

[1.2 Clustering: Treat missing values in CPC, CTR and CPM using the formula given. 4](#_Toc116821890)

[1.3 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). 5](#_Toc116821891)

[1.4 Clustering: Perform z-score scaling and discuss how it affects the speed of the algorithm. 5](#_Toc116821892)

[1.5 Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance. 5](#_Toc116821893)

[1.6 Clustering: Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm. 5](#_Toc116821894)

[1.7 Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters. 5](#_Toc116821895)

[1.8 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. Make bar plots]. 5](#_Toc116821896)

[1.9 Clustering: Conclude the project by providing summary of your learnings. 5](#_Toc116821897)

[Problem Statement 2: PCA: PCA FH (FT) 6](#_Toc116821898)

[2.1 PCA: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc. 6](#_Toc116821899)

[2.2 Analyse the datasetand list down the top 5 important variables, along with the business justifications. 6](#_Toc116821900)

[2.3 PCA: 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 7](#_Toc116821901)

[2.4 PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary? 7](#_Toc116821902)

[2.5 PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment. 7](#_Toc116821903)

[2.6 PCA: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector. 7](#_Toc116821904)

[2.7 PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot. 7](#_Toc116821905)

[2.8 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. 7](#_Toc116821906)

[2.9 PCA: Write linear equation for first PC. 7](#_Toc116821907)

##### Description

Problem Statement:

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

CPC = Total Cost (spend) / Number of Clicks

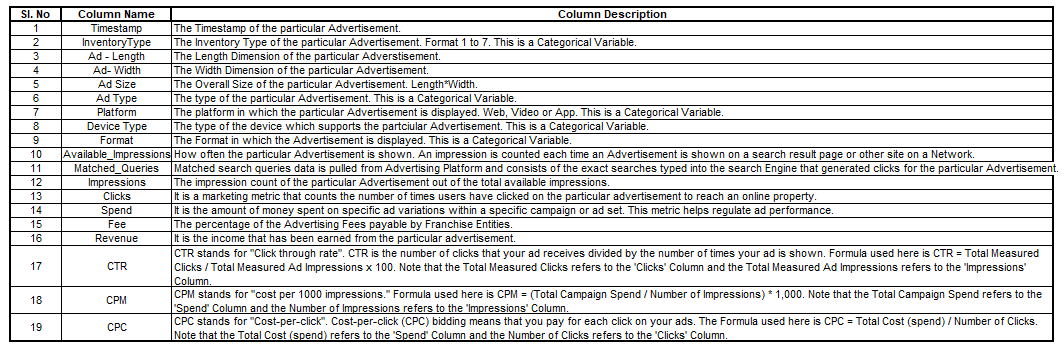
CTR = Total Measured Clicks / Total Measured Ad Impressions x 100

Perform the following in given order:

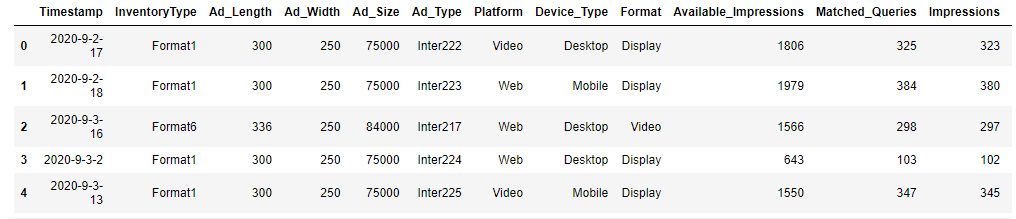
**Q1-Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.**

**Dataset has 2587 rows and 19 columns**

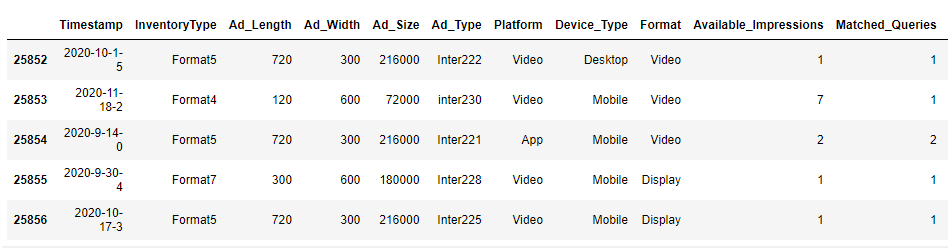
**Following are the variables present in the dataset**



**First few rows of the dataset**



**Last few rows of the dataset**



**Note-Rows are not visible completely in above two tables.**

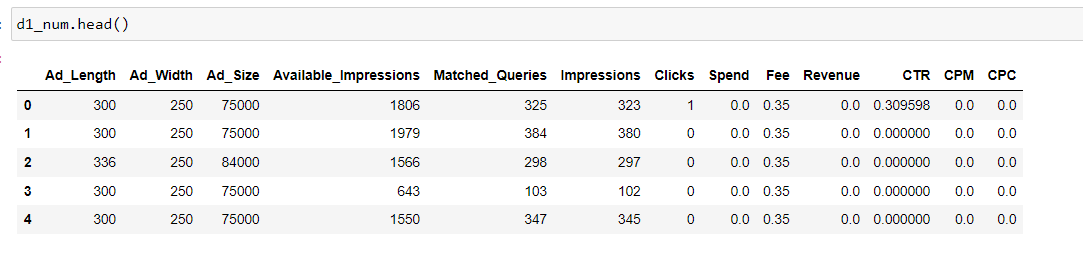
**There are 6 float64, 7 int64 and 6 object data types.**

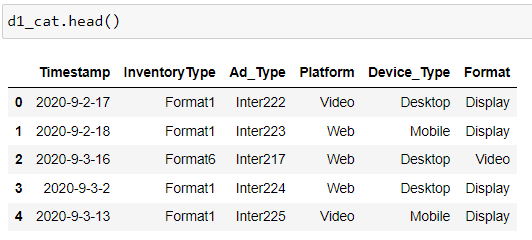
**There are 6465 missing values in CTR, 6465 missing values in CPM and 7527 in CPC.**

**There are no duplicated rows in the dataset.**

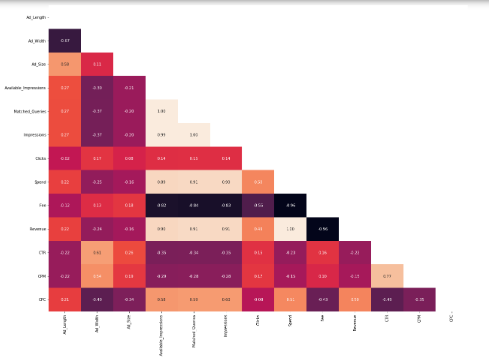
For doing further analysis we bifurcate the data into numerical column and categorical column where-

Categorical variables (d1\_cat) have Timestamp, Inventory type, Ad type, Platform, Device type and format categories, Remaining 13 variables are into numerical column(d1\_num)





**Heatmap of the variables**



**Q 2- Treat missing values in CPC, CTR and CPM using the formula given. You may refer to the Bank\_KMeans Case Study to understand the coding behind treating the missing values using a specific formula. You have to basically create an user defined function and then call the function for imputing**.

**We have treated the null values in CPC using two ways-**

* CPC= Total Cost(spend) /Number of clicks, where clicks >0
* 0, Where clicks=0

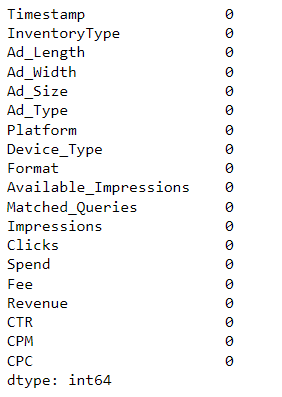
**We have treated the null values in CPM using two ways-**

* CPM= (Total Campaign Spend/Number of Impressions) x 1000, where Ad Impressions >0
* 0, Where Ad Impressions=0

**We have treated the null values in CTR using two ways-**

* CTR= Total Measured Clicks / Total Measured Ad Impressions x 100, where Ad Impressions >0
* 0, Where Ad Impressions=0

**No null values after treatment**

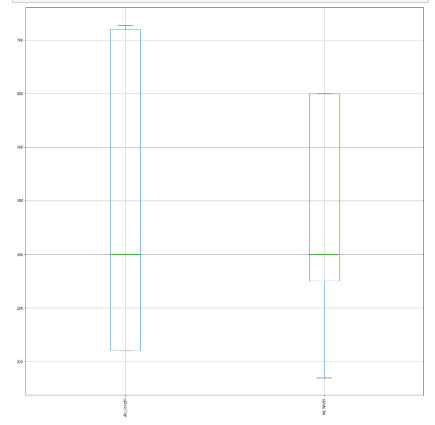


**Q - Check if there are any outliers**.

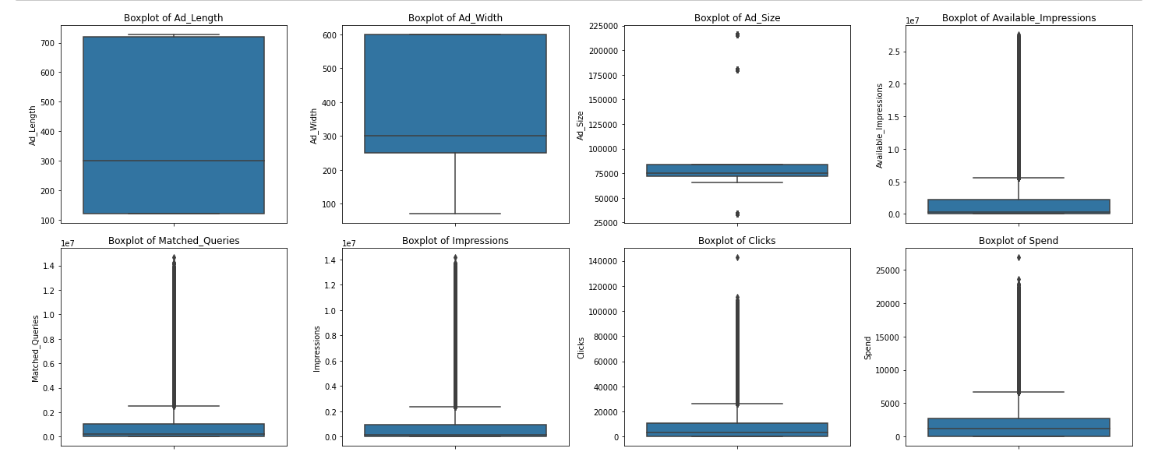
There are significant number of outliers in the following variables-

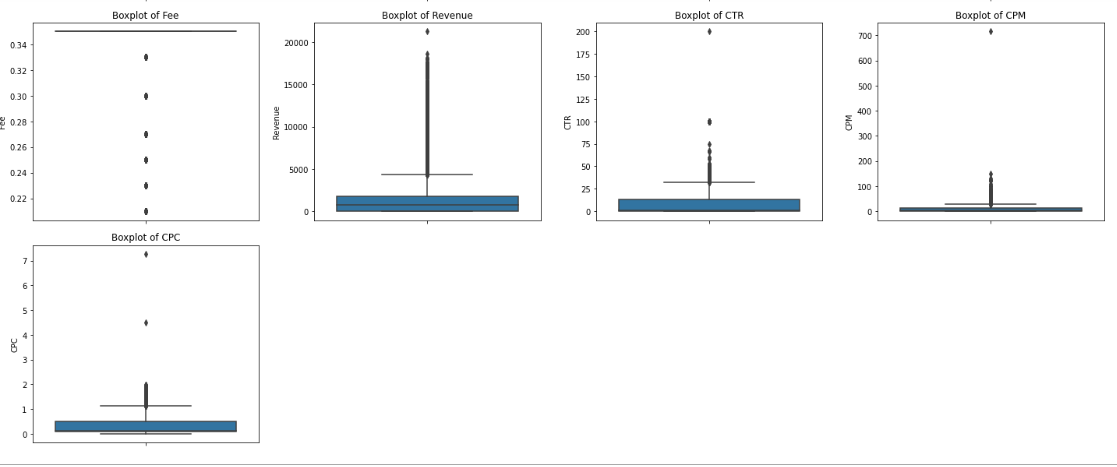
Ad size, Available Impressions, , Matched Queries, Clicks, Speed, Free, Revenue, CTR, CPM, CPC.

Only Ad length and Ad width are the two variables with no outliers.



**BOXPLOT OF DATASET SHOWING OUTLIERS**





**Q 3- 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 judgment may be different from another analyst).**

K-means algorithm are sensitive to outliers. Treating outliers is necessary for K-Means clustering , treating outliers also depend on the the demand of dataset and business domain.

Hence, we will be removing outliers considering all the above mentioned factors.

We have defined a function treat\_outlier, for the higher outliers we will treat it to get it at 95th percentile value and lower level outliers will be treated to get at 25th percentile value.

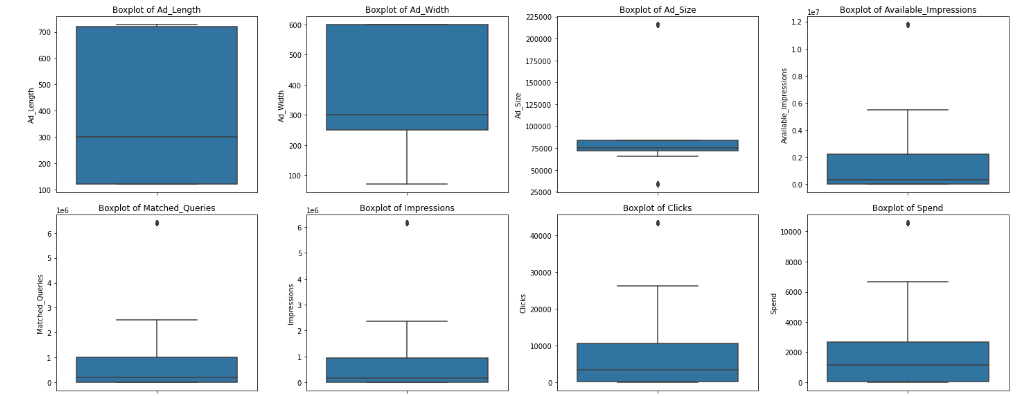
Here we must note that the lower quartile ‘q25’ is median of first half of data. The upper quartile ‘q75’ is median of second half of data. The interquartile range ‘IQR’ is difference of q75 and q5. An outlier is a point that is greater than (q75 + 1.5\*IQR) or lesser than (q25–1.5\*IQR).

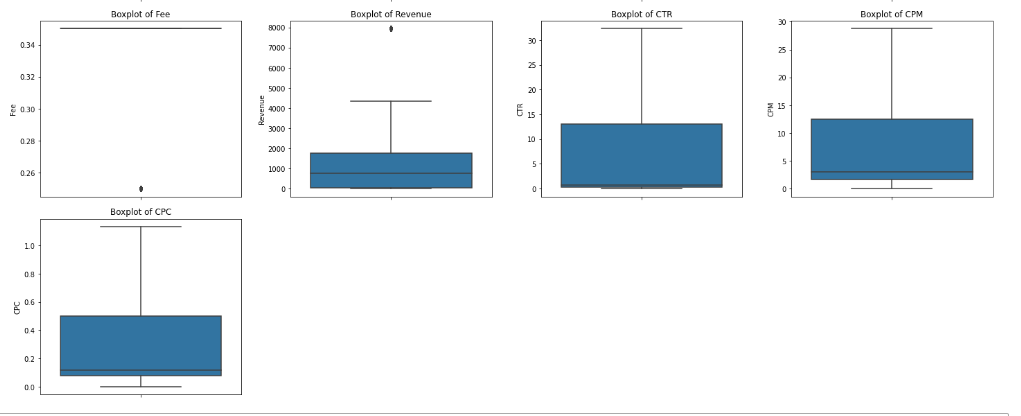
Lower bound=q25-(1.5X IQR)

Upper bound=q75+(1.5X IQR)

Where the values are smaller than lower bound will get replaced by q5 and where the values are bigger than upper bound will get replaced by q95.

**BOXPLOTS AFTER REMOVING OUTLIERS**





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

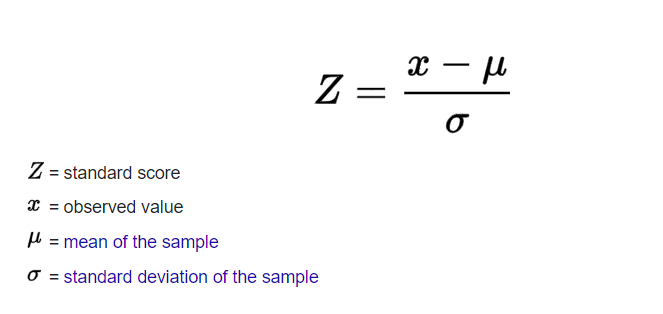
Since we can see that the observations in Ad\_lenghth, Ad\_size, Clicks, spend, fee are in hundreds, where as Available impressions, Matched queries, impressions, are very small (<5). So it is important we scale the data so that are algorithms that we apply give balanced insights and they do not get influenced by only a few columns.

Before scaling we remove the variables Ad\_size, CTR, CPR, CPM as these are dependent variables and store the new variables into data frame D1\_num.

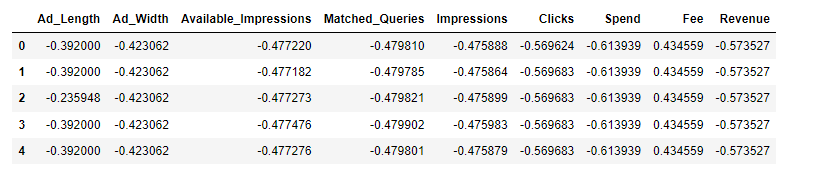
For further analysis and scaling we will be taking only numerical variables forward which are stored in D1\_num.

To scale the data we import zscore from scipy.stats and we apply the Z score to the data set (D1\_num) and store the scaled variables in D1\_num\_scaled data set.

**Formula for zscore**



**First few rows of scaled data**



**Effect of scaling on speed of algorithm**

Scaling definitely increases the speed of algorithm since the scale of all the variables becomes the same.

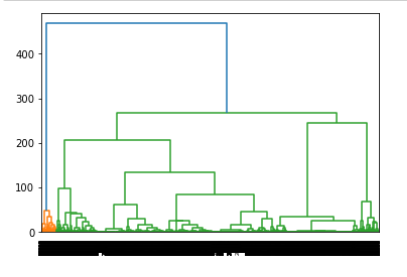
**Q 5-Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance**.

After scaling we perform statistical significance of correlations using batlett sphereicty and we get the p\_value as 0 which means that there are significant correlations in the dataset.

For performing hierarchical clustering we import dendrogram, linkage from scipy.cluster.hierarchy.

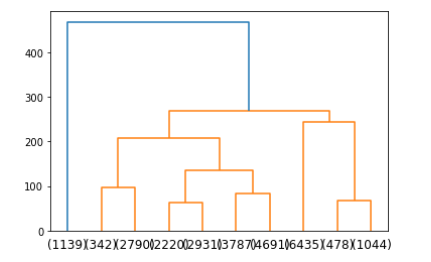
We use method ward and the default distance taken is Euclidean distance.

**Dendrogram**



Since the clusters are not visible properly we use truncate method with number of clusters as 10 for better clarity.

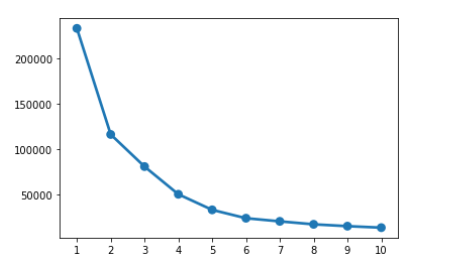
**Dendrogram with truncate method (p=10)**



**Q6-Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.**

To make elbow plot we import KMeans from sklearn.cluster and get the wss values.

Using sns.pointplot we get below elbow plot for n=10.



Optimum number of clusters for k-means algorithm will be 3 since we can see between point 1 and 2, 2 and 3 there is significant difference in the distance on Y axsis but after that there is no significant difference in the distance points are very close to each other.

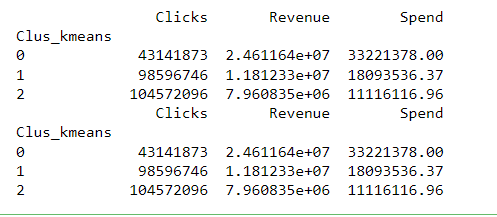
**Q7-Print silhouette scores for up to 10 clusters and identify optimum number of clusters**.

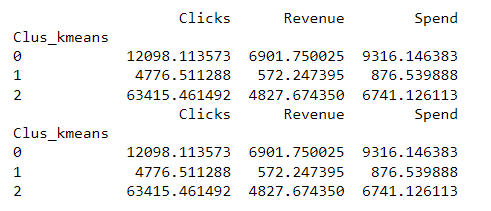
To print silhouette scores we import silhouette\_samples, shilhouette\_score from sklearn.metrics.

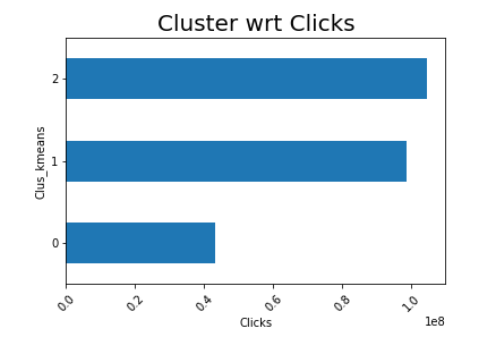
Silhouette score =0.607 (Which is positive- Indicates that we have done clustering properly)

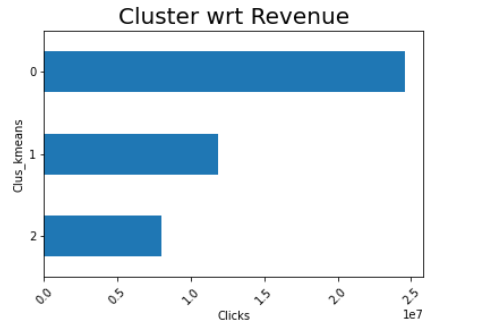
Optimum number of clusters that we take will be 3 since we can see on the elbow plot-between point 1 and 2, 2 and 3 there is significant difference in the distance on Y axsis but after that there is no significant difference in the distance, points are very close to each other.

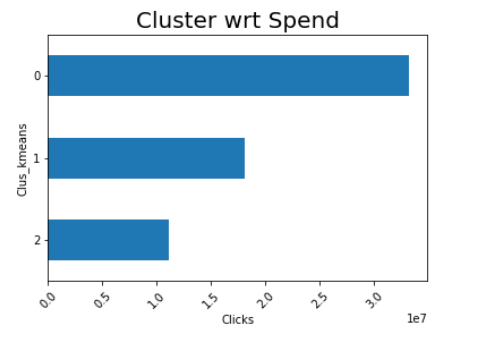
**Q8- Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots].**

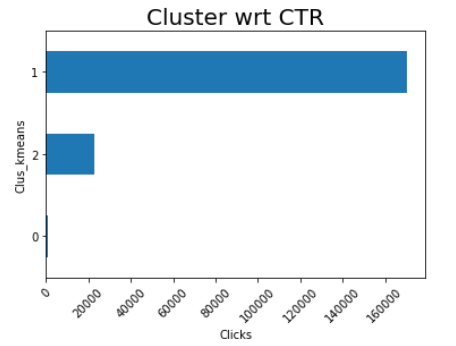


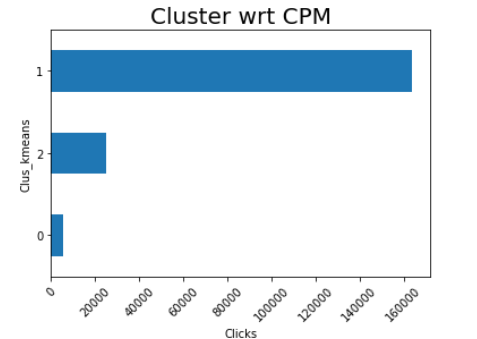


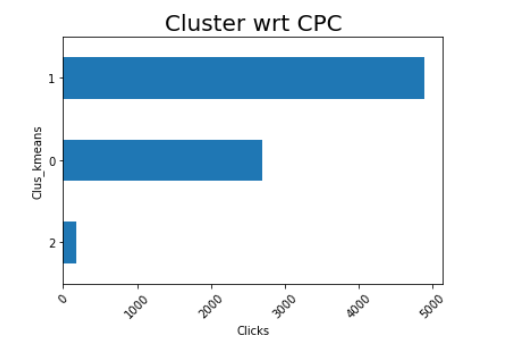




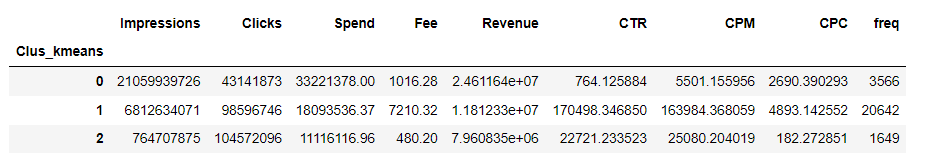








**Q9- Conclude the project by providing summary of your learnings.**

****

We can see here that Impressions have highest influence by cluster 0

Clicks by cluster 2

Spend by cluster 0

Fee by cluster 1

Revenue by cluster 2

CTR by cluster 1

CPM by cluster 1

CPC by cluster 1

**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.  
Data file - PCA India Data Census.xlsx

**Q1- Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.**

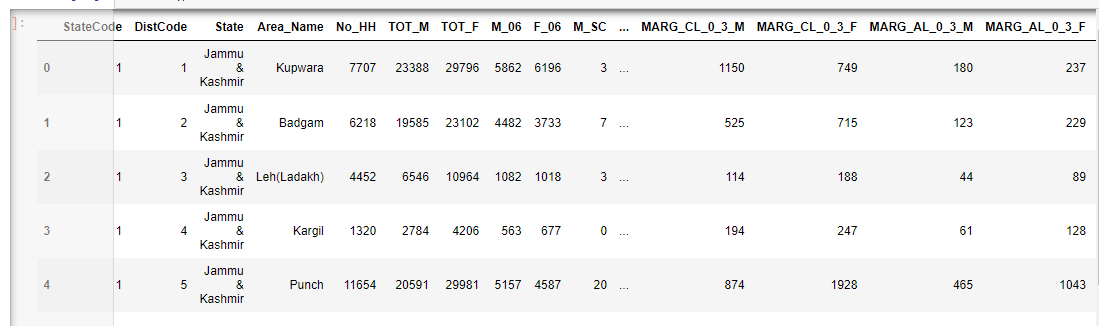
**FEATURES AND THEIR DISCRITION**

|  |  |
| --- | --- |
| Name | Description |
| |  | | --- | |  | | State | | District | | Name | | TRU1 | | 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 | | |  | | --- | | State Code | | District Code | | Name | | Area Name | | No of Household | | Total population Male | | Total population Female | | Population in the age group 0-6 Male | | Population in the age group 0-6 Female | | Scheduled Castes population Male | | Scheduled Castes population Female | | Scheduled Tribes population Male | | Scheduled Tribes population Female | | Literates population Male | | Literates population Female | | Illiterate Male | | Illiterate Female | | Total Worker Population Male | | Total Worker Population Female | | Main Working Population Male | | Main Working Population Female | | Main Cultivator Population Male | | Main Cultivator Population Female | | Main Agricultural Labourers Population Male | | Main Agricultural Labourers Population Female | | Main Household Industries Population Male | | Main Household Industries Population Female | | Main Other Workers Population Male | | Main Other Workers Population Female | | Marginal Worker Population Male | | Marginal Worker Population Female | | Marginal Cultivator Population Male | | Marginal Cultivator Population Female | | Marginal Agriculture Labourers Population Male | | Marginal Agriculture Labourers Population Female | | Marginal Household Industries Population Male | | Marginal Household Industries Population Female | | Marginal Other Workers Population Male | | Marginal Other Workers Population Female | | Marginal Worker Population 3-6 Male | | Marginal Worker Population 3-6 Female | | Marginal Cultivator Population 3-6 Male | | Marginal Cultivator Population 3-6 Female | | Marginal Agriculture Labourers Population 3-6 Male | | Marginal Agriculture Labourers Population 3-6 Female | | Marginal Household Industries Population 3-6 Male | | Marginal Household Industries Population 3-6 Female | | Marginal Other Workers Population Person 3-6 Male | | Marginal Other Workers Population Person 3-6 Female | | Marginal Worker Population 0-3 Male | | Marginal Worker Population 0-3 Female | | Marginal Cultivator Population 0-3 Male | | Marginal Cultivator Population 0-3 Female | | Marginal Agriculture Labourers Population 0-3 Male | | Marginal Agriculture Labourers Population 0-3 Female | | Marginal Household Industries Population 0-3 Male | | Marginal Household Industries Population 0-3 Female | | Marginal Other Workers Population 0-3 Male | | Marginal Other Workers Population 0-3 Female | | Non Working Population Male | | Non Working Population Female | |

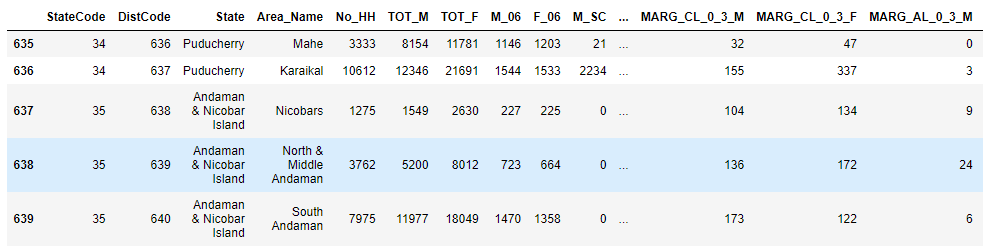
There are 640 rows and 61 columns in the data set.

There are no duplicated rows.

First few rows of the dataset



Last few rows of the dataset



Basic information about the dataset

|  |
| --- |
| <class 'pandas.core.frame.DataFrame'>  RangeIndex: 640 entries, 0 to 639  Data columns (total 61 columns):  # Column Non-Null Count Dtype  --- ------ -------------- -----  0 StateCode 640 non-null int64  1 DistCode 640 non-null int64  2 State 640 non-null object  3 Area\_Name 640 non-null object  4 No\_HH 640 non-null int64  5 TOT\_M 640 non-null int64  6 TOT\_F 640 non-null int64  7 M\_06 640 non-null int64  8 F\_06 640 non-null int64  9 M\_SC 640 non-null int64  10 F\_SC 640 non-null int64  11 M\_ST 640 non-null int64  12 F\_ST 640 non-null int64  13 M\_LIT 640 non-null int64  14 F\_LIT 640 non-null int64  15 M\_ILL 640 non-null int64  16 F\_ILL 640 non-null int64  17 TOT\_WORK\_M 640 non-null int64  18 TOT\_WORK\_F 640 non-null int64  19 MAINWORK\_M 640 non-null int64  20 MAINWORK\_F 640 non-null int64  21 MAIN\_CL\_M 640 non-null int64  22 MAIN\_CL\_F 640 non-null int64  23 MAIN\_AL\_M 640 non-null int64  24 MAIN\_AL\_F 640 non-null int64  25 MAIN\_HH\_M 640 non-null int64  26 MAIN\_HH\_F 640 non-null int64  27 MAIN\_OT\_M 640 non-null int64  28 MAIN\_OT\_F 640 non-null int64  29 MARGWORK\_M 640 non-null int64  30 MARGWORK\_F 640 non-null int64  31 MARG\_CL\_M 640 non-null int64  32 MARG\_CL\_F 640 non-null int64  33 MARG\_AL\_M 640 non-null int64  34 MARG\_AL\_F 640 non-null int64  35 MARG\_HH\_M 640 non-null int64  36 MARG\_HH\_F 640 non-null int64  37 MARG\_OT\_M 640 non-null int64  38 MARG\_OT\_F 640 non-null int64  39 MARGWORK\_3\_6\_M 640 non-null int64  40 MARGWORK\_3\_6\_F 640 non-null int64  41 MARG\_CL\_3\_6\_M 640 non-null int64  42 MARG\_CL\_3\_6\_F 640 non-null int64  43 MARG\_AL\_3\_6\_M 640 non-null int64  44 MARG\_AL\_3\_6\_F 640 non-null int64  45 MARG\_HH\_3\_6\_M 640 non-null int64  46 MARG\_HH\_3\_6\_F 640 non-null int64  47 MARG\_OT\_3\_6\_M 640 non-null int64  48 MARG\_OT\_3\_6\_F 640 non-null int64  49 MARGWORK\_0\_3\_M 640 non-null int64  50 MARGWORK\_0\_3\_F 640 non-null int64  51 MARG\_CL\_0\_3\_M 640 non-null int64  52 MARG\_CL\_0\_3\_F 640 non-null int64  53 MARG\_AL\_0\_3\_M 640 non-null int64  54 MARG\_AL\_0\_3\_F 640 non-null int64  55 MARG\_HH\_0\_3\_M 640 non-null int64  56 MARG\_HH\_0\_3\_F 640 non-null int64  57 MARG\_OT\_0\_3\_M 640 non-null int64  58 MARG\_OT\_0\_3\_F 640 non-null int64  59 NON\_WORK\_M 640 non-null int64  60 NON\_WORK\_F 640 non-null int64  dtypes: int64(59), object(2)  memory usage: 305.1+ KB |

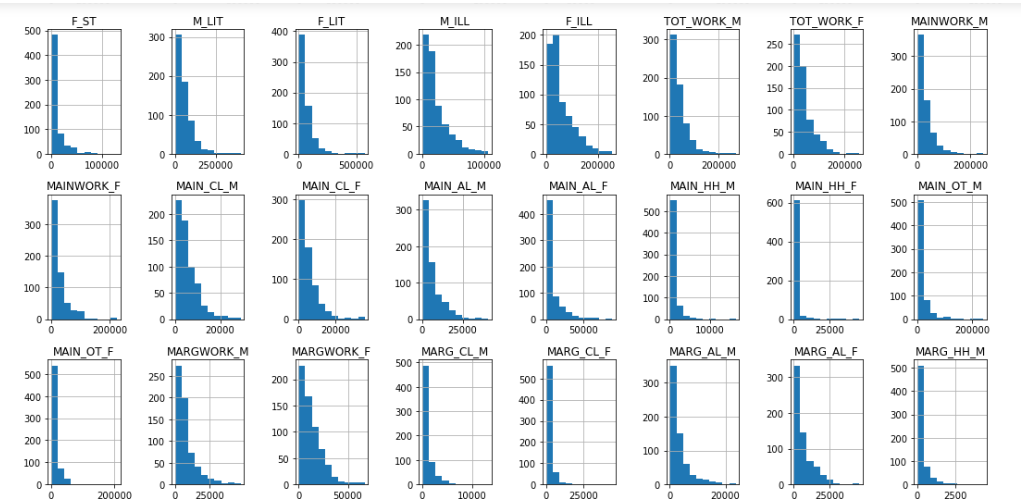
Summary of the data

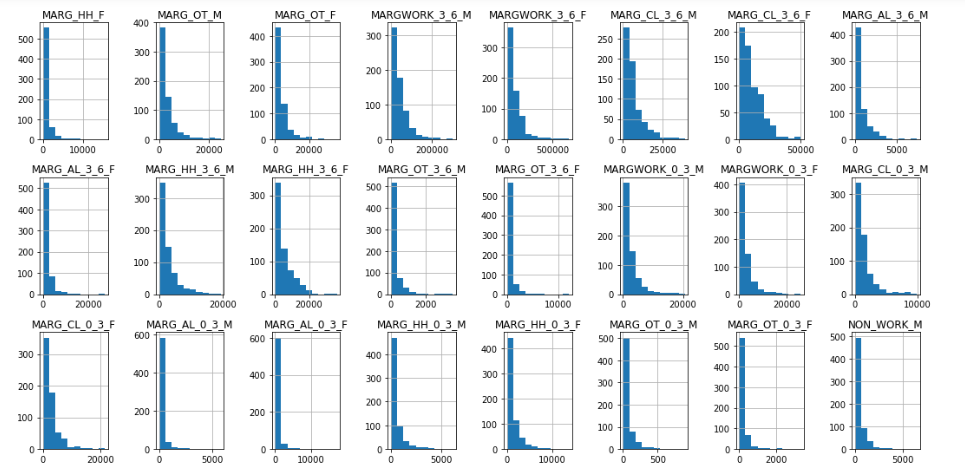
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **StateCode** | 640.0 | 17.11 | 9.43 | 1.0 | 9.00 | 18.0 | 24.00 | 35.0 | | **DistCode** | 640.0 | 320.50 | 184.90 | 1.0 | 160.75 | 320.5 | 480.25 | 640.0 | | **No\_HH** | 640.0 | 51222.87 | 48135.41 | 350.0 | 19484.00 | 35837.0 | 68892.00 | 310450.0 | | **TOT\_M** | 640.0 | 79940.58 | 73384.51 | 391.0 | 30228.00 | 58339.0 | 107918.50 | 485417.0 | | **TOT\_F** | 640.0 | 122372.08 | 113600.72 | 698.0 | 46517.75 | 87724.5 | 164251.75 | 750392.0 | | **M\_06** | 640.0 | 12309.10 | 11500.91 | 56.0 | 4733.75 | 9159.0 | 16520.25 | 96223.0 | | **F\_06** | 640.0 | 11942.30 | 11326.29 | 56.0 | 4672.25 | 8663.0 | 15902.25 | 95129.0 | | **M\_SC** | 640.0 | 13820.95 | 14426.37 | 0.0 | 3466.25 | 9591.5 | 19429.75 | 103307.0 | | **F\_SC** | 640.0 | 20778.39 | 21727.89 | 0.0 | 5603.25 | 13709.0 | 29180.00 | 156429.0 | | **M\_ST** | 640.0 | 6191.81 | 9912.67 | 0.0 | 293.75 | 2333.5 | 7658.00 | 96785.0 | | **F\_ST** | 640.0 | 10155.64 | 15875.70 | 0.0 | 429.50 | 3834.5 | 12480.25 | 130119.0 | | **M\_LIT** | 640.0 | 57967.98 | 55910.28 | 286.0 | 21298.00 | 42693.5 | 77989.50 | 403261.0 | | **F\_LIT** | 640.0 | 66359.57 | 75037.86 | 371.0 | 20932.00 | 43796.5 | 84799.75 | 571140.0 | | **M\_ILL** | 640.0 | 21972.60 | 19825.61 | 105.0 | 8590.00 | 15767.5 | 29512.50 | 105961.0 | | **F\_ILL** | 640.0 | 56012.52 | 47116.69 | 327.0 | 22367.00 | 42386.0 | 78471.00 | 254160.0 | | **TOT\_WORK\_M** | 640.0 | 37992.41 | 36419.54 | 100.0 | 13753.50 | 27936.5 | 50226.75 | 269422.0 | | **TOT\_WORK\_F** | 640.0 | 41295.76 | 37192.36 | 357.0 | 16097.75 | 30588.5 | 53234.25 | 257848.0 | | **MAINWORK\_M** | 640.0 | 30204.45 | 31480.92 | 65.0 | 9787.00 | 21250.5 | 40119.00 | 247911.0 | | **MAINWORK\_F** | 640.0 | 28198.85 | 29998.26 | 240.0 | 9502.25 | 18484.0 | 35063.25 | 226166.0 | | **MAIN\_CL\_M** | 640.0 | 5424.34 | 4739.16 | 0.0 | 2023.50 | 4160.5 | 7695.00 | 29113.0 | | **MAIN\_CL\_F** | 640.0 | 5486.04 | 5326.36 | 0.0 | 1920.25 | 3908.5 | 7286.25 | 36193.0 | | **MAIN\_AL\_M** | 640.0 | 5849.11 | 6399.51 | 0.0 | 1070.25 | 3936.5 | 8067.25 | 40843.0 | | **MAIN\_AL\_F** | 640.0 | 8926.00 | 12864.29 | 0.0 | 1408.75 | 3933.5 | 10617.50 | 87945.0 | | **MAIN\_HH\_M** | 640.0 | 883.89 | 1278.64 | 0.0 | 187.50 | 498.5 | 1099.25 | 16429.0 | | **MAIN\_HH\_F** | 640.0 | 1380.77 | 3179.41 | 0.0 | 248.75 | 540.5 | 1435.75 | 45979.0 | | **MAIN\_OT\_M** | 640.0 | 18047.10 | 26068.48 | 36.0 | 3997.50 | 9598.0 | 21249.50 | 240855.0 | | **MAIN\_OT\_F** | 640.0 | 12406.04 | 18972.20 | 153.0 | 3142.50 | 6380.5 | 14368.25 | 209355.0 | | **MARGWORK\_M** | 640.0 | 7787.96 | 7410.79 | 35.0 | 2937.50 | 5627.0 | 9800.25 | 47553.0 | | **MARGWORK\_F** | 640.0 | 13096.91 | 10996.47 | 117.0 | 5424.50 | 10175.0 | 18879.25 | 66915.0 | | **MARG\_CL\_M** | 640.0 | 1040.74 | 1311.55 | 0.0 | 311.75 | 606.5 | 1281.00 | 13201.0 | | **MARG\_CL\_F** | 640.0 | 2307.68 | 3564.63 | 0.0 | 630.25 | 1226.0 | 2659.25 | 44324.0 | | **MARG\_AL\_M** | 640.0 | 3304.33 | 3781.56 | 0.0 | 873.50 | 2062.0 | 4300.75 | 23719.0 | | **MARG\_AL\_F** | 640.0 | 6463.28 | 6773.88 | 0.0 | 1402.50 | 4020.5 | 9089.25 | 45301.0 | | **MARG\_HH\_M** | 640.0 | 316.74 | 462.66 | 0.0 | 71.75 | 166.0 | 356.50 | 4298.0 | | **MARG\_HH\_F** | 640.0 | 786.63 | 1198.72 | 0.0 | 171.75 | 429.0 | 962.50 | 15448.0 | | **MARG\_OT\_M** | 640.0 | 3126.15 | 3609.39 | 7.0 | 935.50 | 2036.0 | 3985.25 | 24728.0 | | **MARG\_OT\_F** | 640.0 | 3539.32 | 4115.19 | 19.0 | 1071.75 | 2349.5 | 4400.50 | 36377.0 | | **MARGWORK\_3\_6\_M** | 640.0 | 41948.17 | 39045.32 | 291.0 | 16208.25 | 30315.0 | 57218.75 | 300937.0 | | **MARGWORK\_3\_6\_F** | 640.0 | 81076.32 | 82970.41 | 341.0 | 26619.50 | 56793.0 | 107924.00 | 676450.0 | | **MARG\_CL\_3\_6\_M** | 640.0 | 6394.99 | 6019.81 | 27.0 | 2372.00 | 4630.0 | 8167.00 | 39106.0 | | **MARG\_CL\_3\_6\_F** | 640.0 | 10339.86 | 8467.47 | 85.0 | 4351.50 | 8295.0 | 15102.00 | 50065.0 | | **MARG\_AL\_3\_6\_M** | 640.0 | 789.85 | 905.64 | 0.0 | 235.50 | 480.5 | 986.00 | 7426.0 | | **MARG\_AL\_3\_6\_F** | 640.0 | 1749.58 | 2496.54 | 0.0 | 497.25 | 985.5 | 2059.00 | 27171.0 | | **MARG\_HH\_3\_6\_M** | 640.0 | 2743.64 | 3059.59 | 0.0 | 718.75 | 1714.5 | 3702.25 | 19343.0 | | **MARG\_HH\_3\_6\_F** | 640.0 | 5169.85 | 5335.64 | 0.0 | 1113.75 | 3294.0 | 7502.25 | 36253.0 | | **MARG\_OT\_3\_6\_M** | 640.0 | 245.36 | 358.73 | 0.0 | 58.00 | 129.5 | 276.00 | 3535.0 | | **MARG\_OT\_3\_6\_F** | 640.0 | 585.88 | 900.03 | 0.0 | 127.75 | 320.5 | 719.25 | 12094.0 | | **MARGWORK\_0\_3\_M** | 640.0 | 2616.14 | 3036.96 | 7.0 | 755.00 | 1681.5 | 3320.25 | 20648.0 | | **MARGWORK\_0\_3\_F** | 640.0 | 2834.55 | 3327.84 | 14.0 | 833.50 | 1834.5 | 3610.50 | 25844.0 | | **MARG\_CL\_0\_3\_M** | 640.0 | 1392.97 | 1489.71 | 4.0 | 489.50 | 949.0 | 1714.00 | 9875.0 | | **MARG\_CL\_0\_3\_F** | 640.0 | 2757.05 | 2788.78 | 30.0 | 957.25 | 1928.0 | 3599.75 | 21611.0 | | **MARG\_AL\_0\_3\_M** | 640.0 | 250.89 | 453.34 | 0.0 | 47.00 | 114.5 | 270.75 | 5775.0 | | **MARG\_AL\_0\_3\_F** | 640.0 | 558.10 | 1117.64 | 0.0 | 109.00 | 247.5 | 568.75 | 17153.0 | | **MARG\_HH\_0\_3\_M** | 640.0 | 560.69 | 762.58 | 0.0 | 136.50 | 308.0 | 642.00 | 6116.0 | | **MARG\_HH\_0\_3\_F** | 640.0 | 1293.43 | 1585.38 | 0.0 | 298.00 | 717.0 | 1710.75 | 13714.0 | | **MARG\_OT\_0\_3\_M** | 640.0 | 71.38 | 107.90 | 0.0 | 14.00 | 35.0 | 79.00 | 895.0 | | **MARG\_OT\_0\_3\_F** | 640.0 | 200.74 | 309.74 | 0.0 | 43.00 | 113.0 | 240.00 | 3354.0 | | **NON\_WORK\_M** | 640.0 | 510.01 | 610.60 | 0.0 | 161.00 | 326.0 | 604.50 | 6456.0 | | **NON\_WORK\_F** | 640.0 | 704.78 | 910.21 | 5.0 | 220.50 | 464.5 | 853.50 | 10533.0 | |

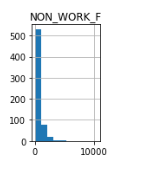
We drop the columns Statecode, Distcode, state and area name since these are all categorical variables and we analyze only the numerical columns stored in data set d2.

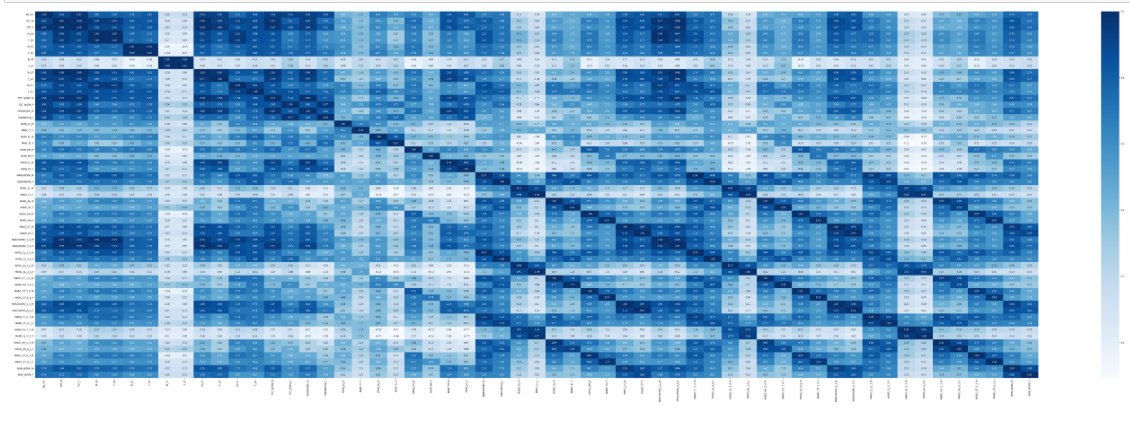
###### Q2-  Perform detailed Exploratory analysis by creating certain questions like (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio? (Example Questions). Pick 5 variables out of the given 24 variables below for EDA: No\_HH, 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

Univariante anlysis





 Multivariant analysis

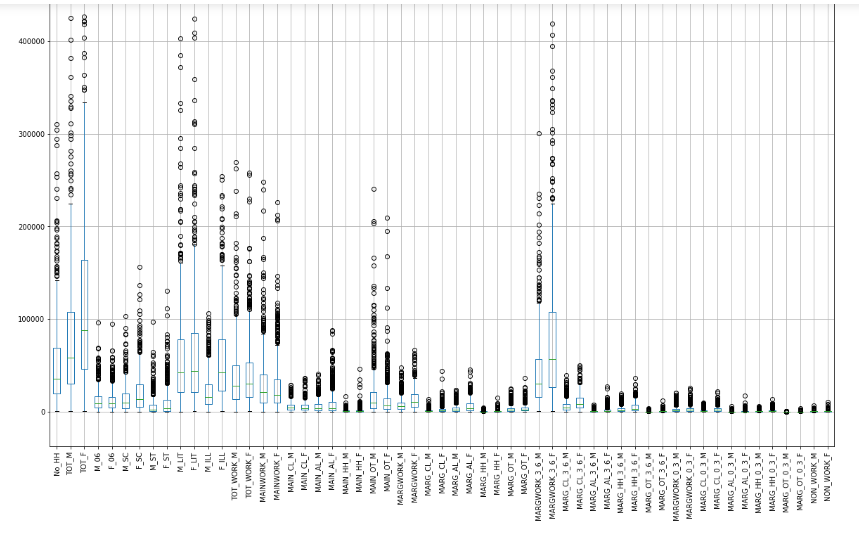


Which state has highest gender ratio and which has the lowest?

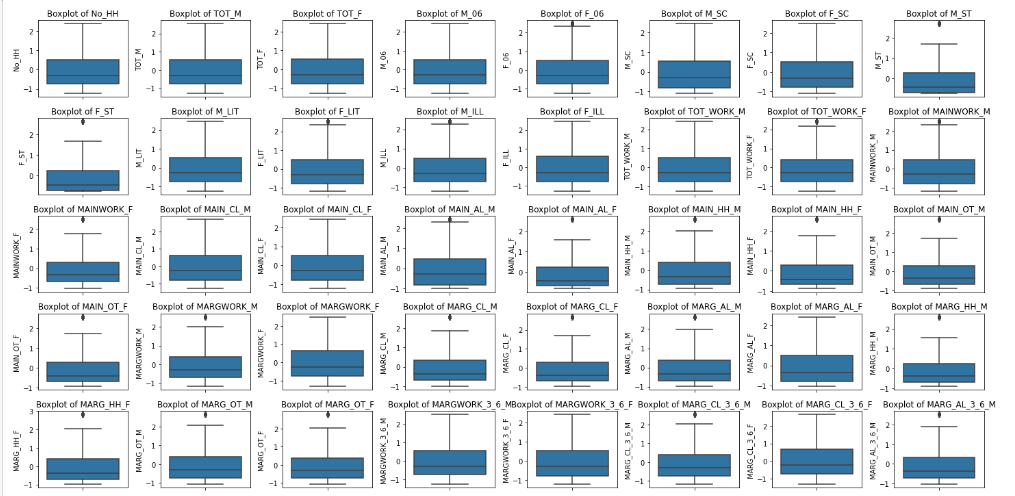
**Q3- We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?**

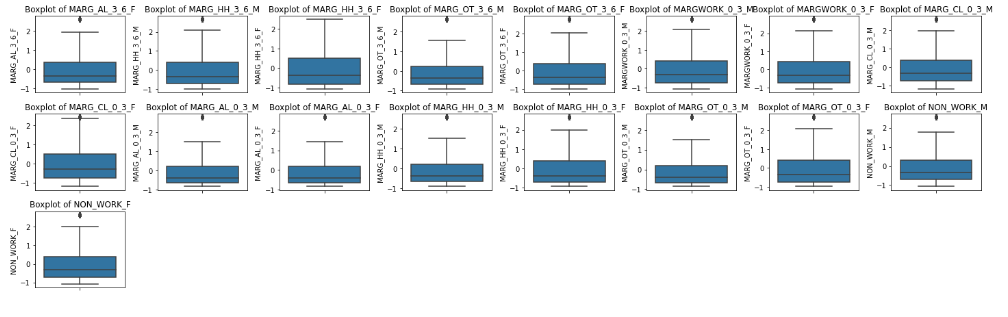
Yes we will be removing outliers, since PCA is sensitive to outliers it might lead to inefficient results. Also the dataset has a lot of outliers which needs to be removed.

Boxplot for dataset



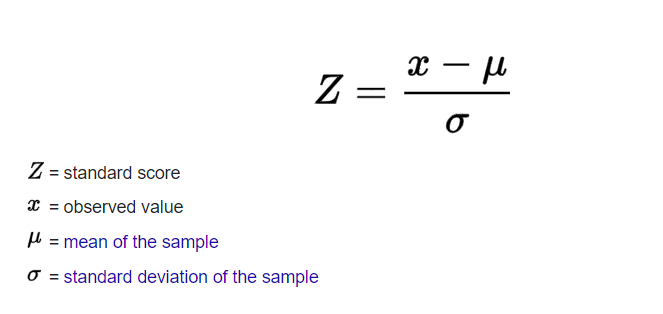
Boxplot with removed outliers





Q3- **Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.**

Scaling does not have an effect on outliers, scaling only reduces the magnitude of outliers but does not minimize them.

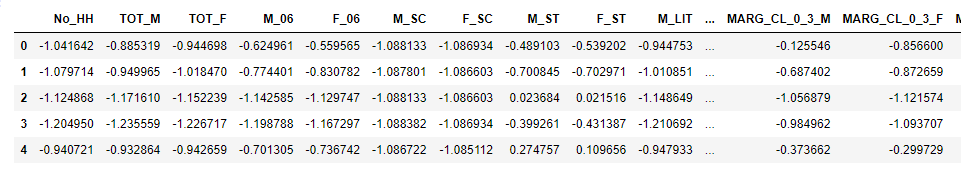
****

Q4- **Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.**

We have dropped statecode, Distcode, state and areaname column since all these are categorical variables and are not needed for performing PCA, since PCA is performed only on continuous and numerical variables.

Before performing PCA we apply zscore method to scale the data since the numerical variables are highly different in magnitude.

Few rows of scaled data(Not all the columns are visible)



We check the correlation of variables, and looking at the pairplot we see that there is some correlation in the given variables.

Barlett sphericity Test

To confirm the statistical signinficance of correlations we peform Bartlett sphericity test and we get p values as 0 which means we reject the null hypothesis(H0=Correlations are not significant). Hence there are significant correlations in the dataset provided.

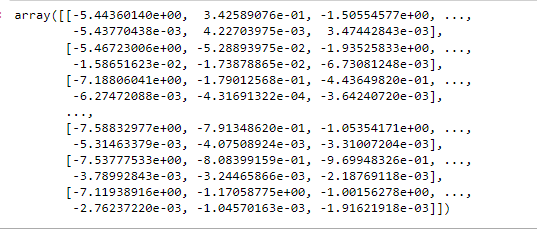
Calculating kmao

Since we got kmo value=0.95, which is greater than 0.7 this shows us that we have adequate sample size to perform PCA.

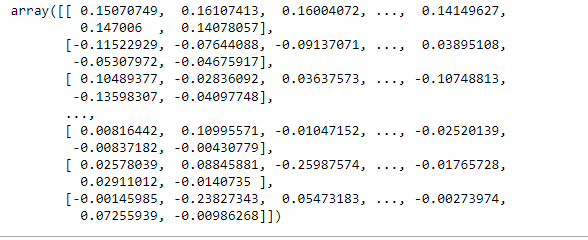
After this we apply PCA taking all the 57 features we have got out of 61 features( dropped a few features).

For this we import PCA from sklearn.decomposition and we put random state=123 so that we get the same results everytime we perform the algorithms. And then we performed pca.fit\_transform to the scaled data set.

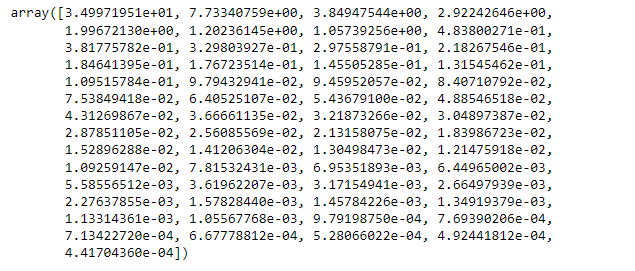
Covariance Matrix



PCA Components(Eign vectors)

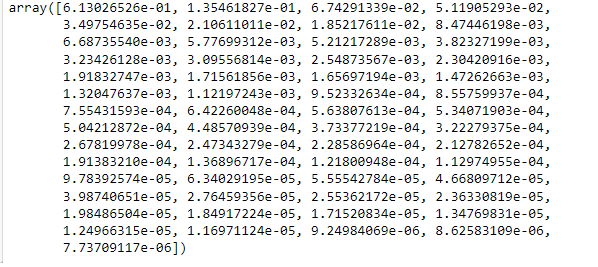


Eigen values



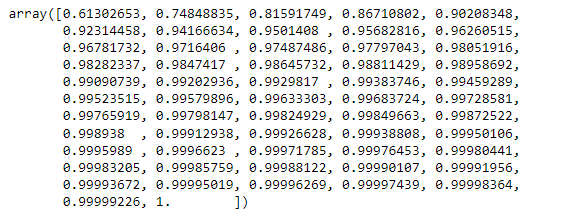
Explained variance ratio

(eigen value of each PC)/(sum of eigen values of all PCs)

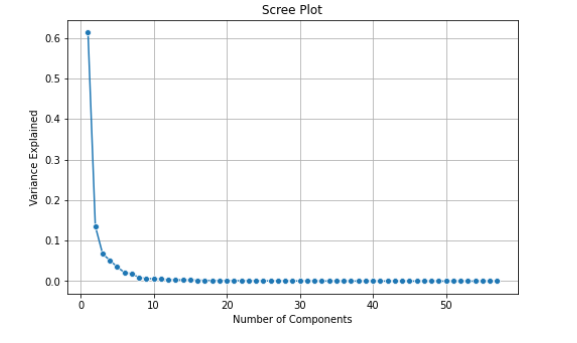


**Q5- Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.**

Cumulative explained variance ratio

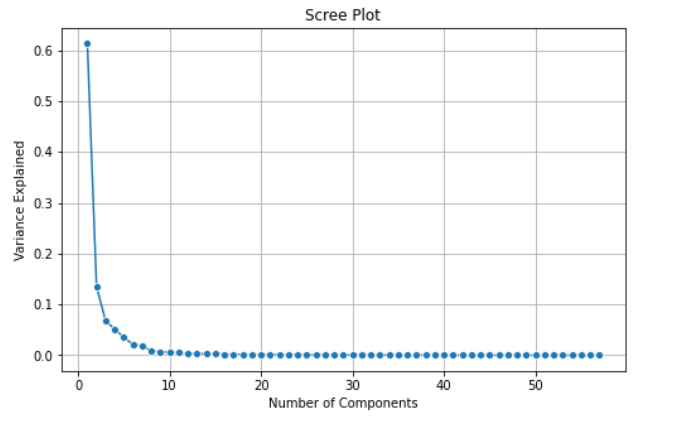
****

**Scree plot**

****

**Although there are 57 observed variables the first 5 principal components can explain more than 90% of total variation, hence the optimum number of PC’s that we take is 5 PCs instead of original 57, thereby reducing the dimensions to a significant number.**

**We also look at the scree plot which has indices of the PCs on the X-axis and variances on the Y-axsis. We can see that at point 5 there is a distinct break point and the line joining the variances (elbow pint) becomes approximetly horizontal.**

****

**Q6- Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.**

The names we give to PC’s are-

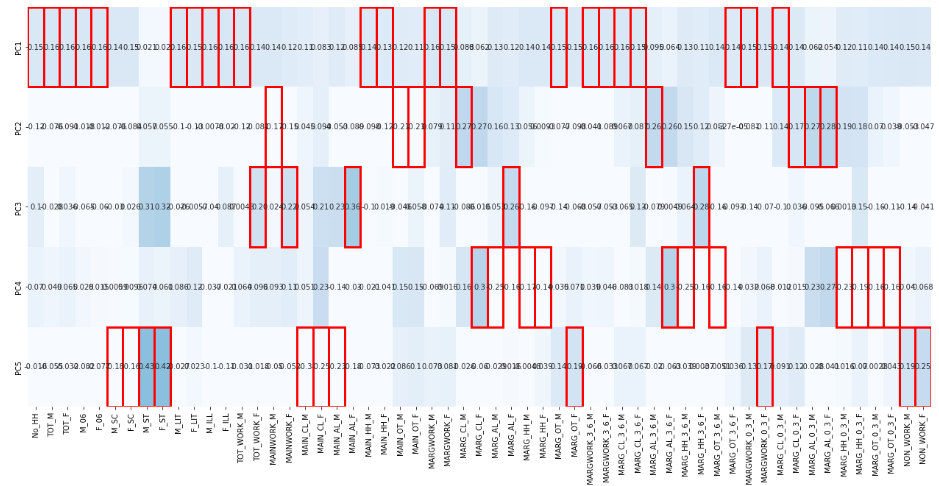
PC1=Skilled\_unskilled\_workforce

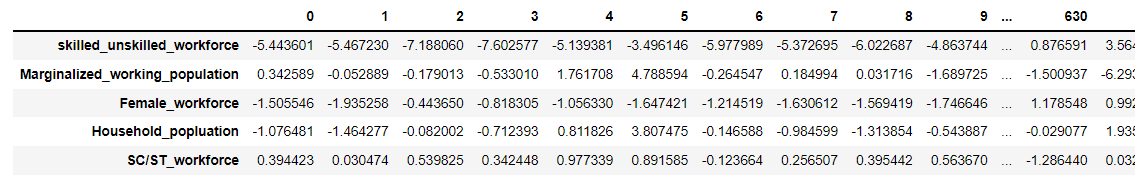
PC2=Marginalized\_working\_population

PC3=Female\_workforce

PC4=Household\_population

PC5=SC/ST Workforce

****

****

**Q7- Write linear equation for first PC.**

( 0.15 ) \* No\_HH + ( 0.16 ) \* TOT\_M + ( 0.16 ) \* TOT\_F + ( 0.16 ) \* M\_06 + ( 0.16 ) \* F\_06 + ( 0.14 ) \* M\_SC + ( 0.15 ) \* F\_SC + ( 0.02 ) \* M\_ST + ( 0.02 ) \* F\_ST + ( 0.16 ) \* M\_LIT + ( 0.15 ) \* F\_LIT + ( 0.16 ) \* M\_ILL + ( 0.16 ) \* F\_ILL + ( 0.16 ) \* TOT\_WORK\_M + ( 0.14 ) \* TOT\_WORK\_F + ( 0.14 ) \* MAINWORK\_M + ( 0.12 ) \* MAINWORK\_F + ( 0.11 ) \* MAIN\_CL\_M + ( 0.08 ) \* MAIN\_CL\_F + ( 0.12 ) \* MAIN\_AL\_M + ( 0.09 ) \* MAIN\_AL\_F + ( 0.14 ) \* MAIN\_HH\_M + ( 0.13 ) \* MAIN\_HH\_F + ( 0.12 ) \* MAIN\_OT\_M + ( 0.11 ) \* MAIN\_OT\_F + ( 0.16 ) \* MARGWORK\_M + ( 0.15 ) \* MARGWORK\_F + ( 0.09 ) \* MARG\_CL\_M + ( 0.06 ) \* MARG\_CL\_F + ( 0.13 ) \* MARG\_AL\_M + ( 0.12 ) \* MARG\_AL\_F + ( 0.14 ) \* MARG\_HH\_M + ( 0.14 ) \* MARG\_HH\_F + ( 0.15 ) \* MARG\_OT\_M + ( 0.15 ) \* MARG\_OT\_F + ( 0.16 ) \* MARGWORK\_3\_6\_M + ( 0.16 ) \* MARGWORK\_3\_6\_F + ( 0.16 ) \* MARG\_CL\_3\_6\_M + ( 0.15 ) \* MARG\_CL\_3\_6\_F + ( 0.09 ) \* MARG\_AL\_3\_6\_M + ( 0.06 ) \* 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 + ( 0.14 ) \* MARG\_OT\_3\_6\_F + ( 0.15 ) \* MARGWORK\_0\_3\_M + ( 0.15 ) \* MARGWORK\_0\_3\_F + ( 0.14 ) \* MARG\_CL\_0\_3\_M + ( 0.14 ) \* MARG\_CL\_0\_3\_F + ( 0.06 ) \* MARG\_AL\_0\_3\_M + ( 0.05 ) \* MARG\_AL\_0\_3\_F + ( 0.12 ) \* MARG\_HH\_0\_3\_M + ( 0.11 ) \* MARG\_HH\_0\_3\_F + ( 0.14 ) \* MARG\_OT\_0\_3\_M + ( 0.14 ) \* MARG\_OT\_0\_3\_F + ( 0.15 ) \* NON\_WORK\_M + ( 0.14 ) \* NON\_WORK\_F +