Data MINING (pca & clusturing) Project

By Vinish Vincent

Great Learnings

PGP-DSBA - 23

# Table of content

Table of Content

**Problem 1**: 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

* 1. 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
  2. Treat missing values in CPC, CTR and CPM using the formula given. You may refer to the Bank\_KMeans Solution File to understand the coding behind treating the missing values using a specific formula. You have to basically create a user defined function and then call the function for imputing…………………………………………………………………………………………………………………….………….4
  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)…..……….……….4
  4. Perform z-score scaling and discuss how it affects the speed of the algorithm………………………..….6
  5. Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance………….6
  6. Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm..7
  7. Print silhouette scores for up to 10 clusters and identify optimum number of clusters……………….8
  8. Profile the ads based on optimum number of clusters using silhouette score and your domain understanding……………………………………………………………………………………………………………………………..8
  9. Conclude the project by providing summary of your learnings.………………………………………………………………………………………………………….…………………..…..10

**Problem 2**: 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…………………………………………10

* Note: The 24 variables given in the Rubric is just for performing EDA. You will have to consider the entire dataset, including all the variables for performing PCA……………….……….13

# Problem 1

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:

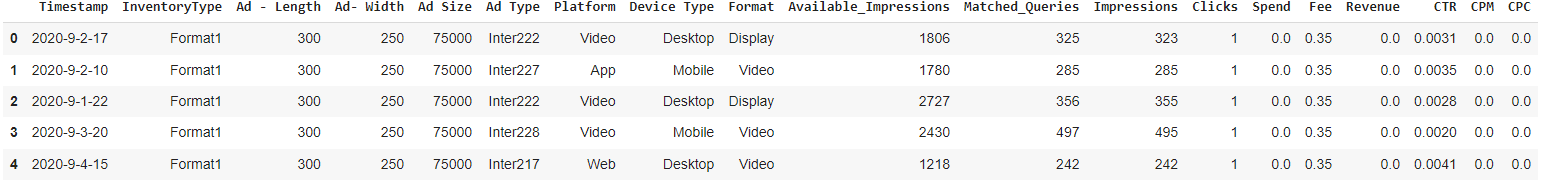
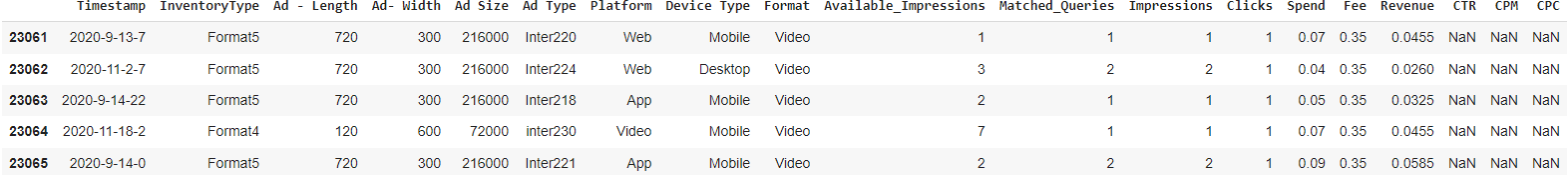
CPM = (Total Campaign Spend / Number of Impressions) \* 1,000

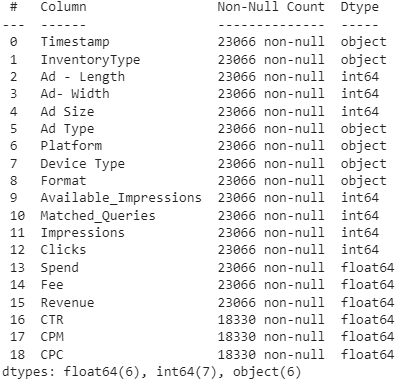
CPC = Total Cost (spend) / Number of Clicks

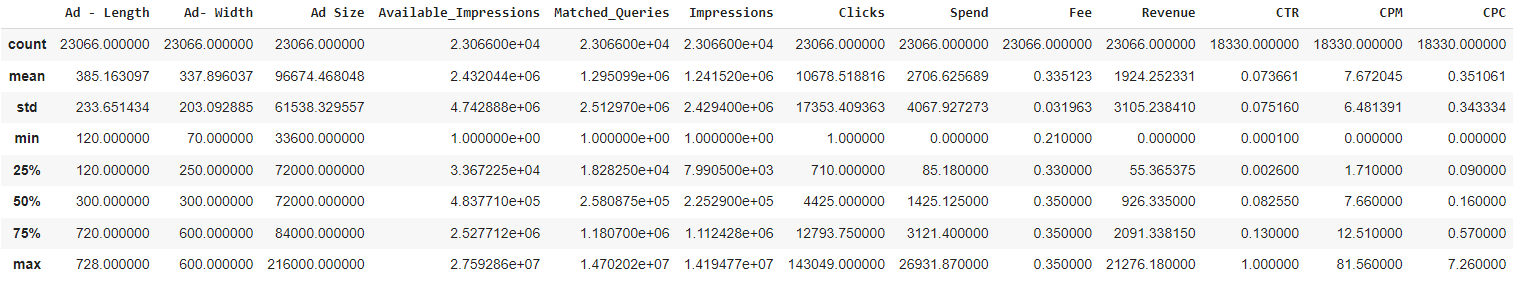
CTR = Total Measured Clicks / Total Measured Ad Impressions x 100

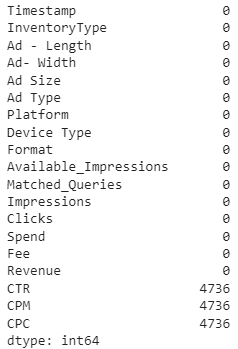
* 1. **Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.**

**Solution**:

* ****Data top 5 records:
* Data last 5 records:
* There are 23066 rows and 19 columns in this dataset
* Data info:



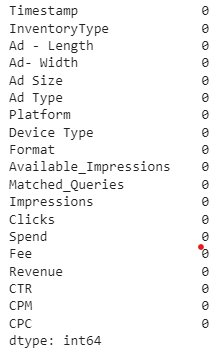
* Data summary:
* There are no duplicates in data set
* There are 4736 values missing in CTR, CPM and CPC respectively.



* 1. **Treat missing values in CPC, CTR and CPM using the formula given**

**Solution**: we have calculated using def calculate\_ function for all the 3 missing records, i.e., CTR, CPM and CPC. We have replaced the null value by using .apply(lambda x: …) function.

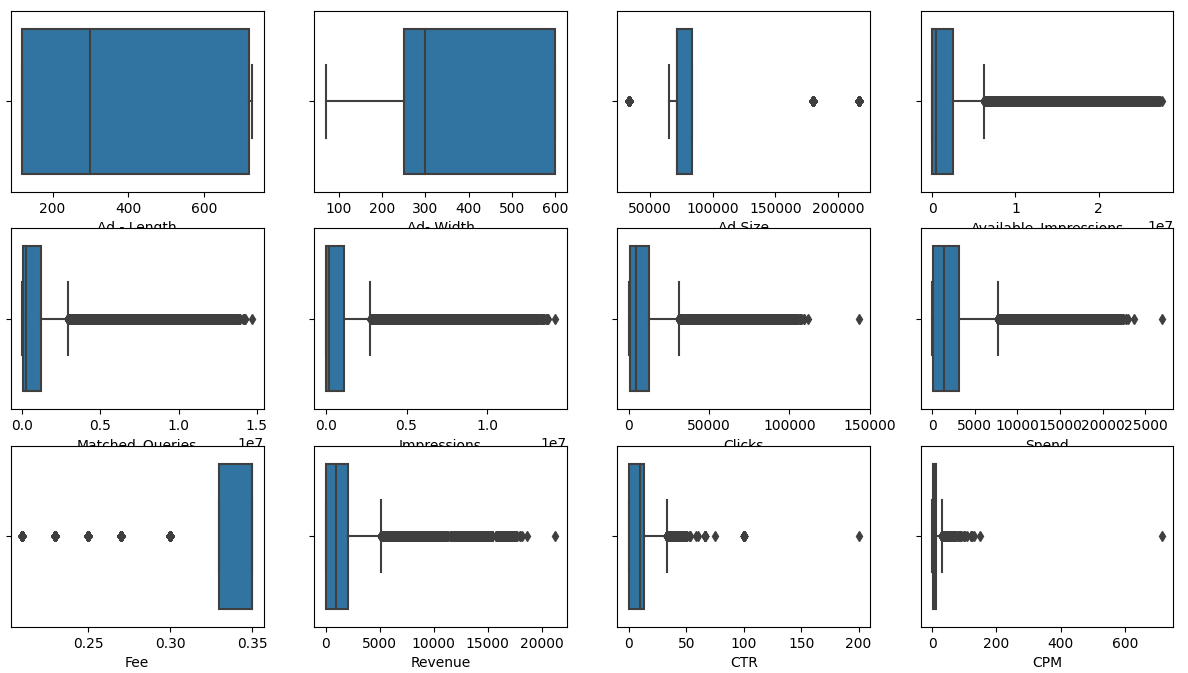
After treating null values

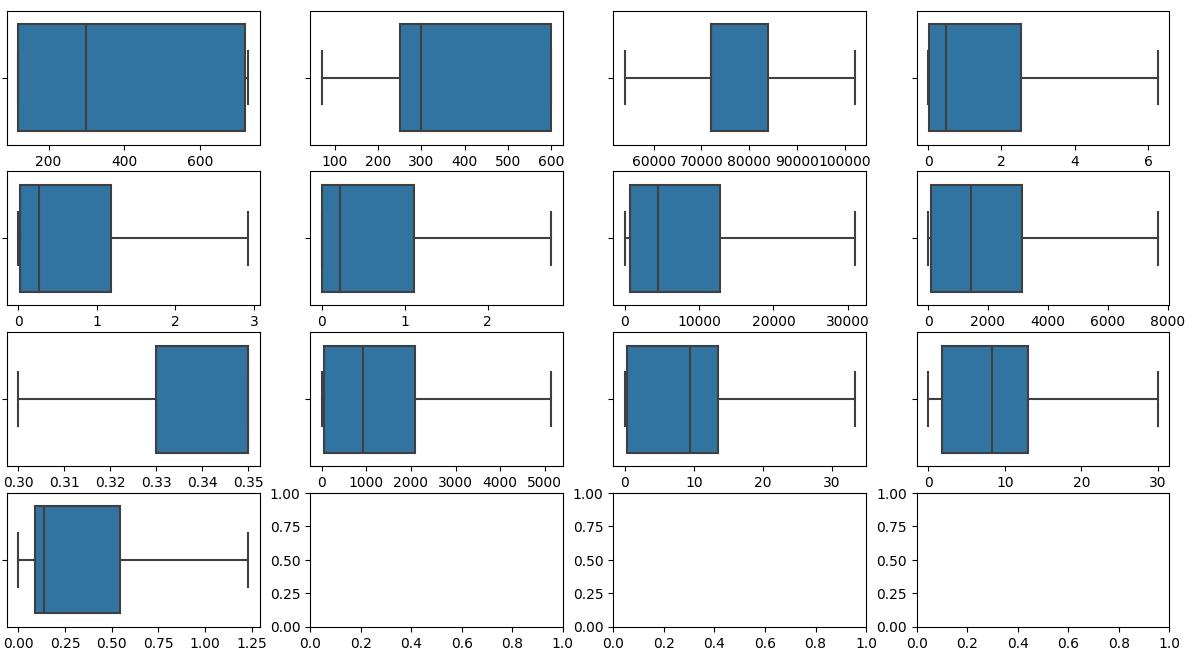
****

* 1. **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.**

**Solution**: There are multiple variables as outliers including: Ad size, Available Impressions, Matched Queries, Impressions, Clicks, Speed, Fee, Revenue, CTR, CPM and CPC.

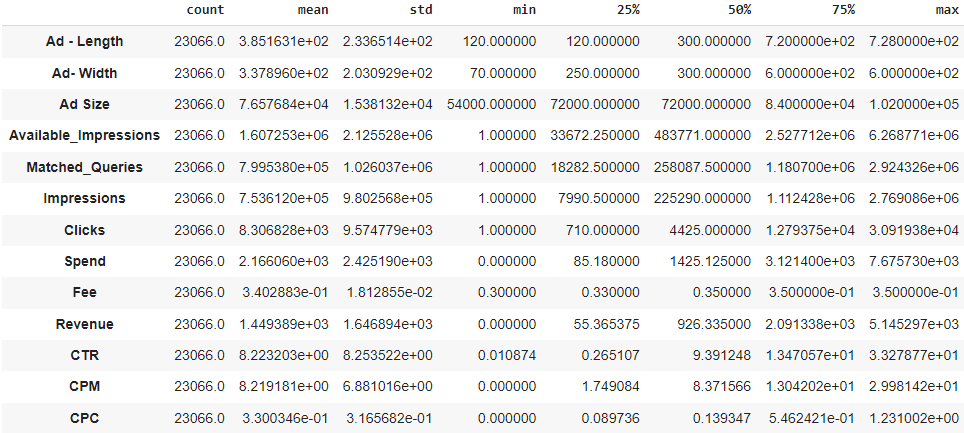
Ad-Length and Ad-width has no outliers



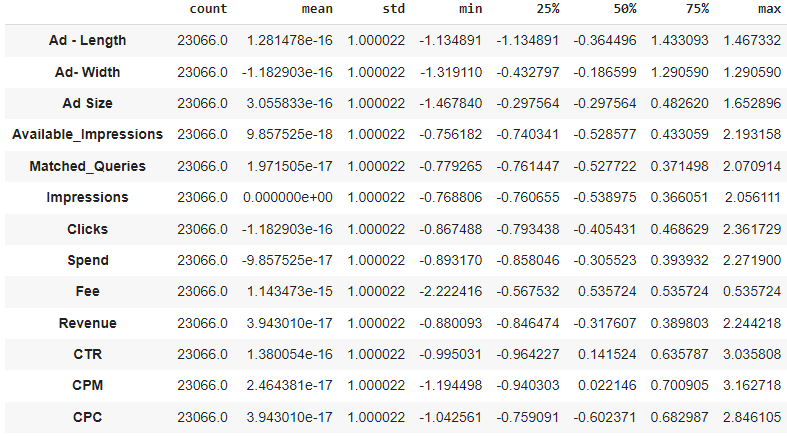
**Treating outliers with IQR method**

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

**Solution**: We imported z-score from scipy.stats

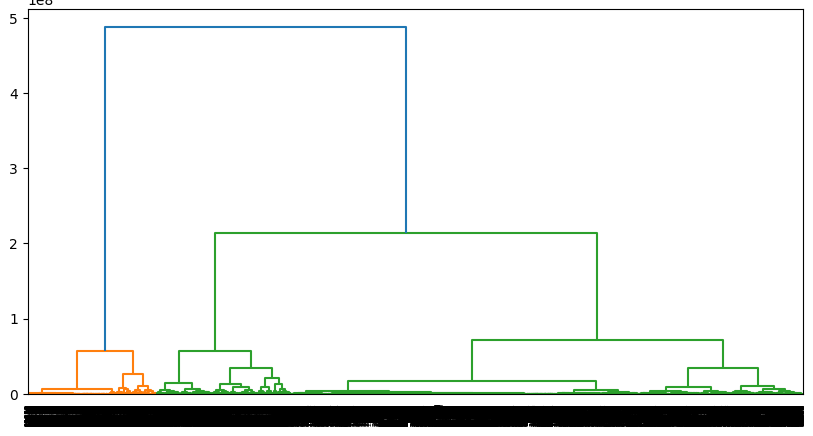
****Before scaling:

After scaling:

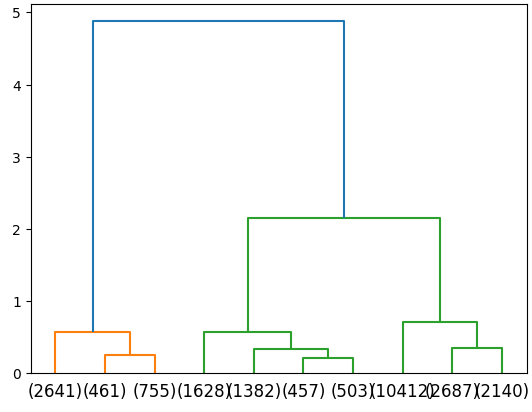


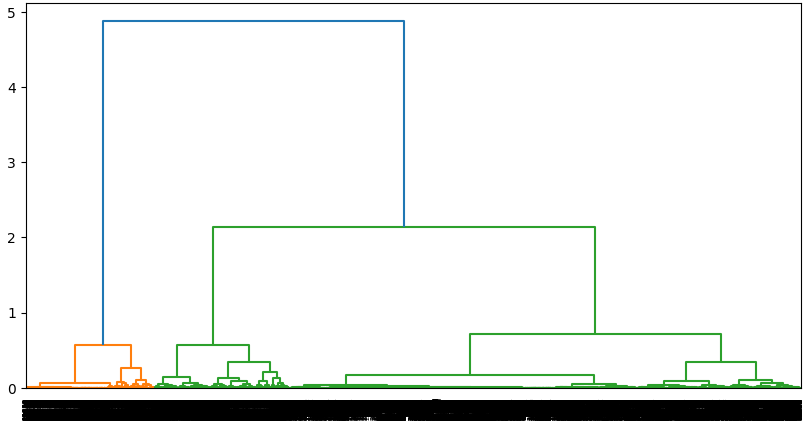
* 1. **Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance**

**Solution:** Dendrogram using ward linkage

****

**Dendrogram with last 10 clusters:**

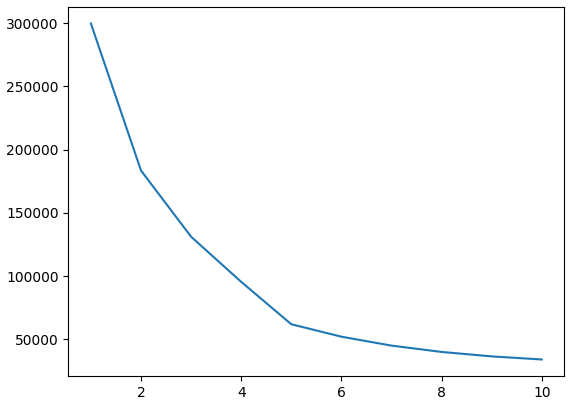
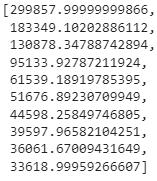
****

**Dendrogram using Euclidean distance method**

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

**Solution**: Elbow plot

**WSS**

****

The optimum no. of clusters or K has been identified as **5, i.e., k=5**

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

**Solution:**

no\_of\_cluster\_3 - 0.3825486036570082

no\_of\_cluster\_4 - 0.45324270552598256

no\_of\_cluster\_5 - 0.5240956940501831

We can see silhouette score decreasing after 5 clusters, so optimum no. of clusters as per score is **5**

no\_of\_cluster\_6 - 0.5221533662938636

no\_of\_cluster\_7 - 0.5165635029478517

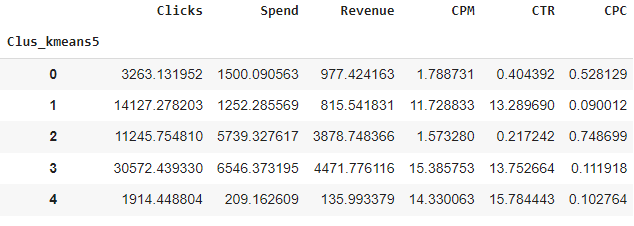
no\_of\_cluster\_8 - 0.47972249893837277

no\_of\_cluster\_9 - 0.4320636564025043

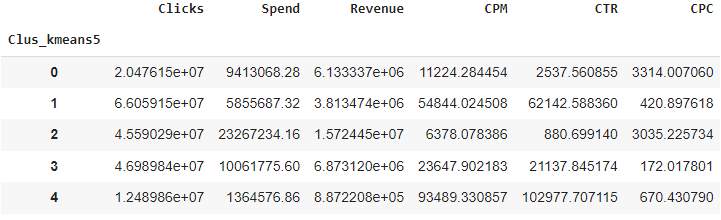
no\_of\_cluster\_10 - 0.43124854581084165

* 1. **Profile the ads based on optimum number of clusters using silhouette score and your domain understanding**

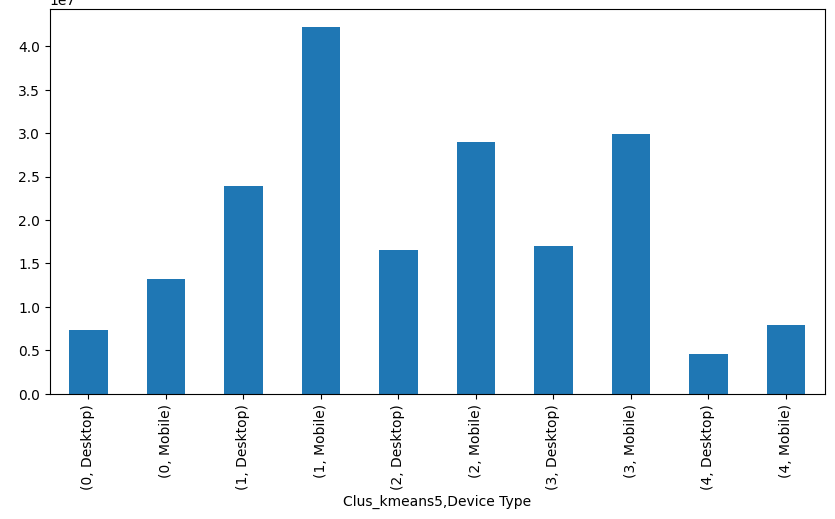
**Solution:** Cluster grouped by mean:

****

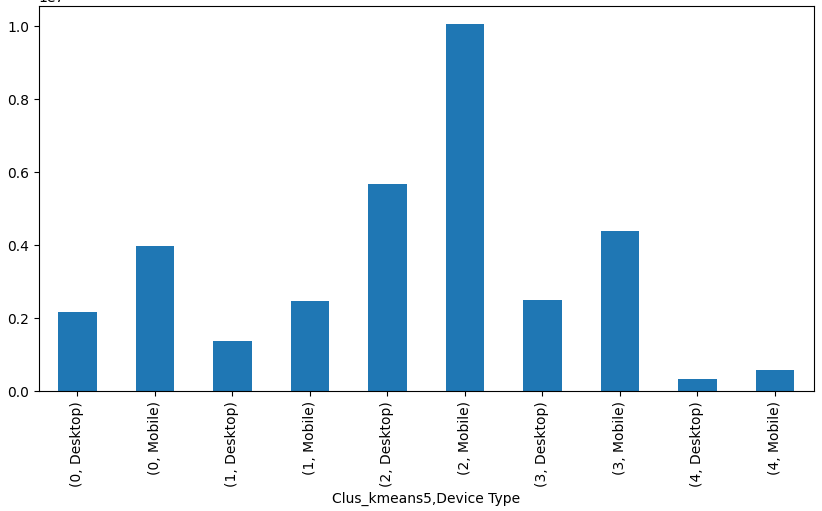
Cluster grouped by sum:

****

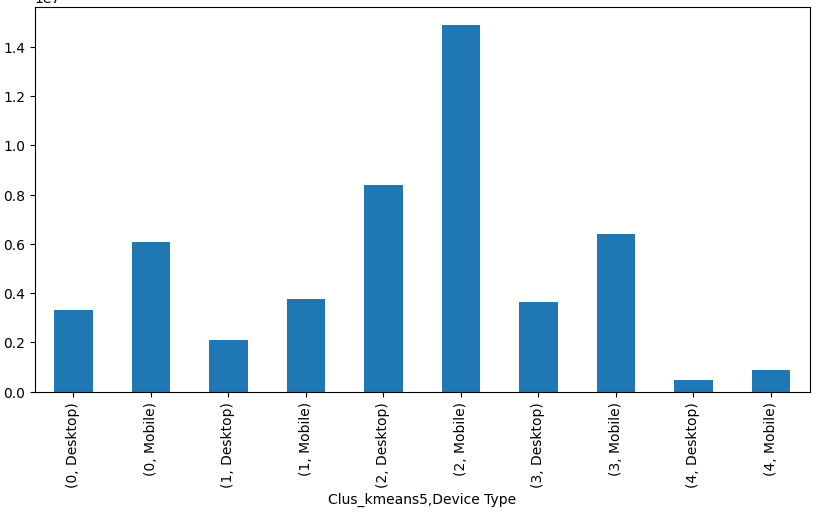
Sum cluster by clicks with Device Type:



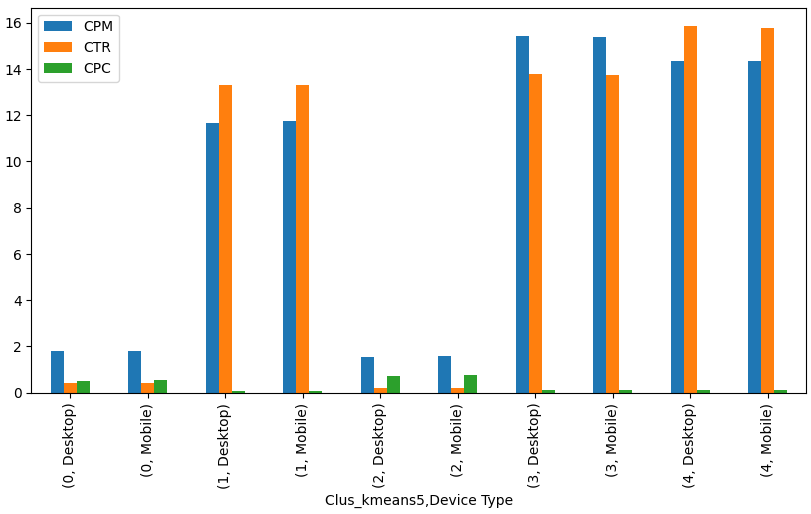
Sum cluster by Revenue with Device Type:



Sum cluster by Spend with Device Type:



Sum cluster by 'CPM', 'CTR', 'CPC' with Device Type:



* 1. **Conclude the project by providing summary of your learnings**

**Solution**:

* The dataset has 23066 rows and 19 columns
* There is no duplicate values. There are missing values in CPC, CTR and CPM. They are treated by using the formula provided in the data sheet
* There are outliers in integer (numeric) variables i.e., Ad-Length, Ad-width, Ad size, Avaliable impressions, Matched Queries, Impressions, Clicks, Spend, Fee, Revenue, CTR, CPM and CPC.
* Outliers are treated using IQR method and scaled the data using Z-Score.
* Dendrogram is the visualization and linkage are for computing the distances and merging the clusters from n to 1, the output of Linkage is visualized by Dendrogram
* We will create linkage using Ward's method and also Euclidean distance method.
* Using fit - transform function and viewing the output - The dataframe is now stored in an array and K-means was performed
* We have found WSS values and mapped the elbow plot
* We have calculated the sil-score and preformed this function for cluster no. 2 till cluster no. 10.
* Considering all the sil-scores, we identified 5 is optimal number of clusters

# Problem 2

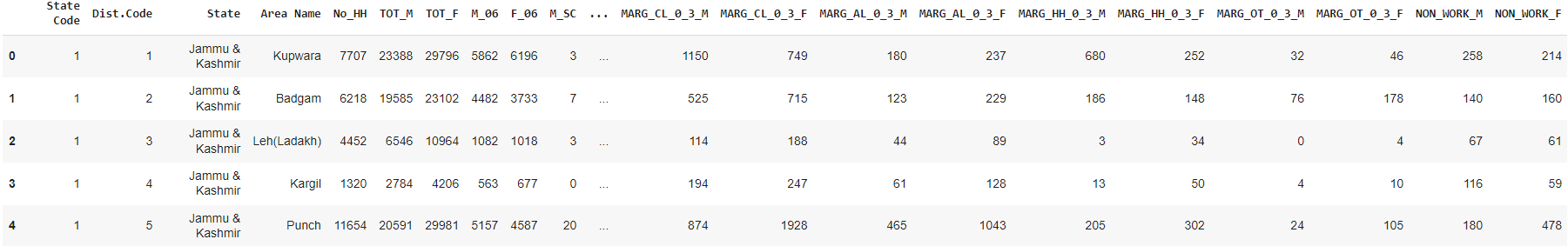
Primary census abstract for female headed households excluding institutional households (India & States/UTs - District Level), Scheduled tribes - 2011 PCA for Female Headed Household Excluding Institutional Household. The Indian Census has the reputation of being one of the best in the world. The first 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,

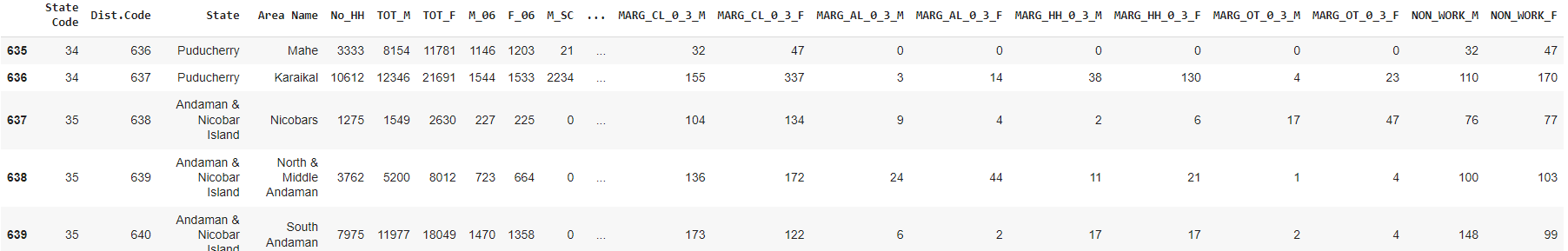
1. Cultivators
2. Agricultural Laborers
3. Household Industry Workers
4. Other Workers and also Non-Workers.

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.

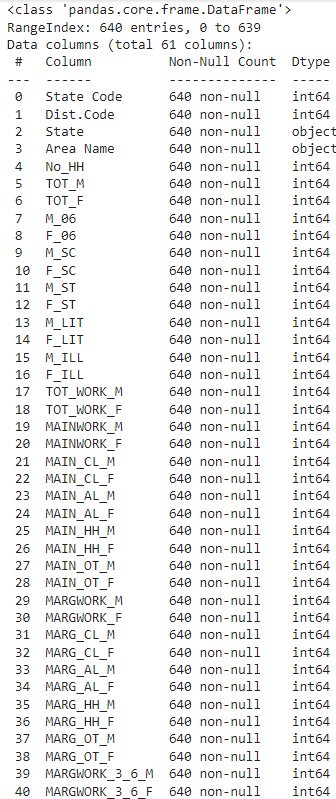
* 1. **Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.**

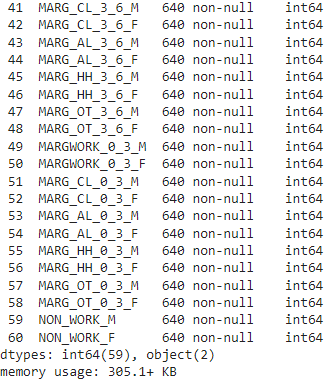
**Solution**:

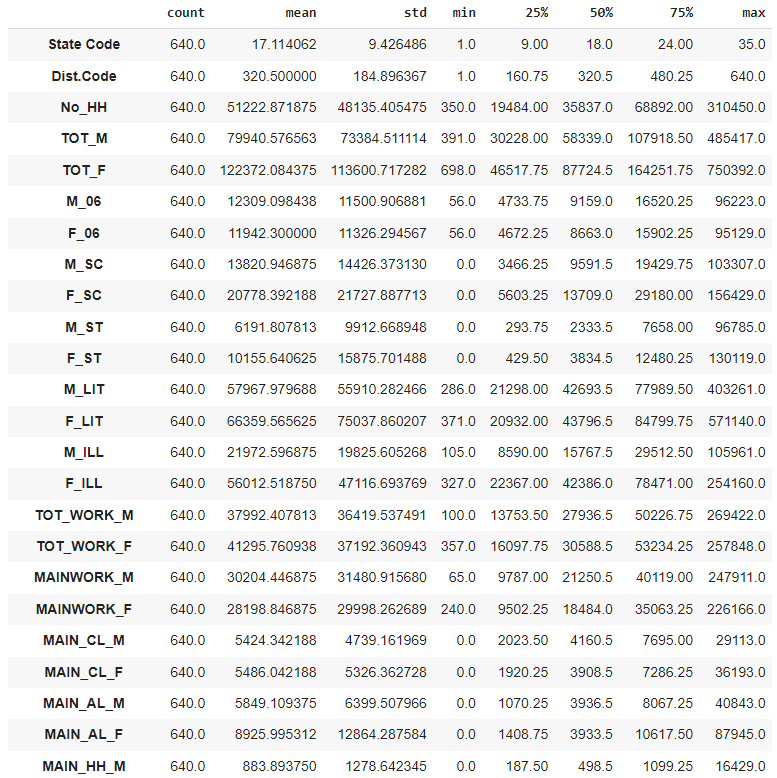
* Data top 5 records:
* Data last 5 records:

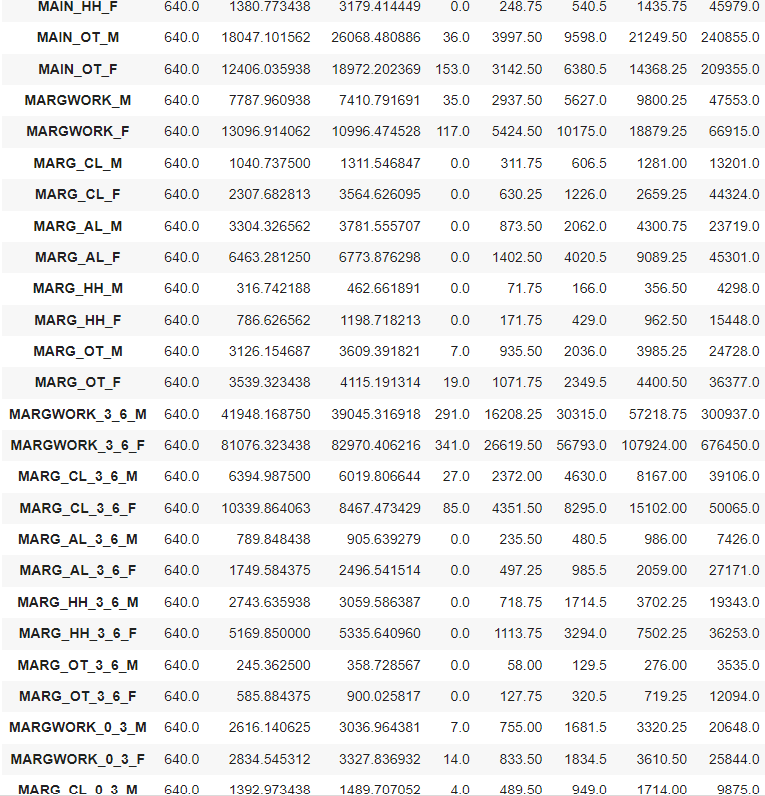


* There are 640 rows and 61 columns in this dataset
* Data info:



****

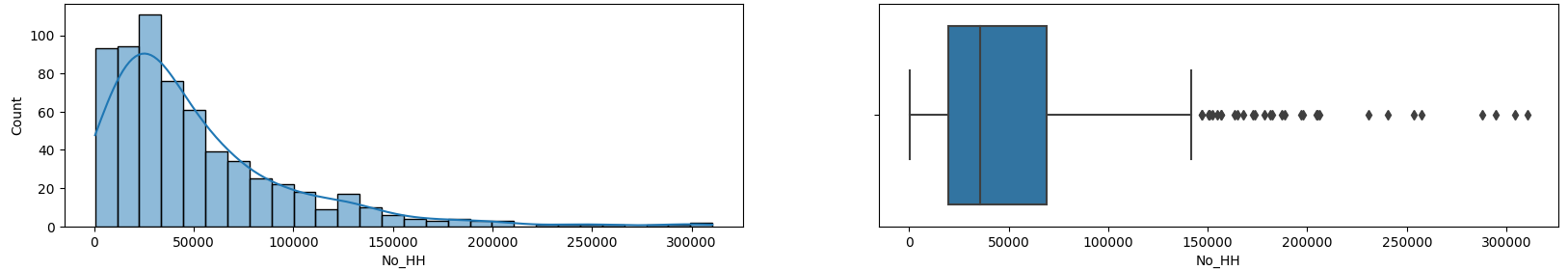
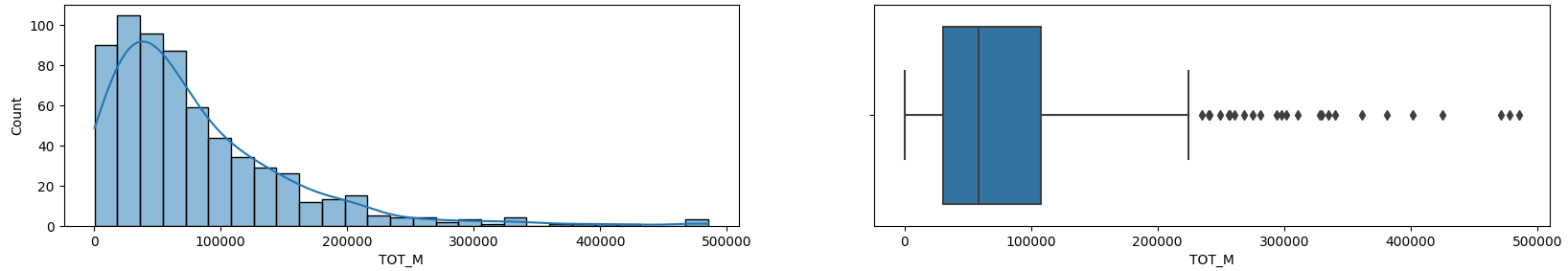
* Data summary:

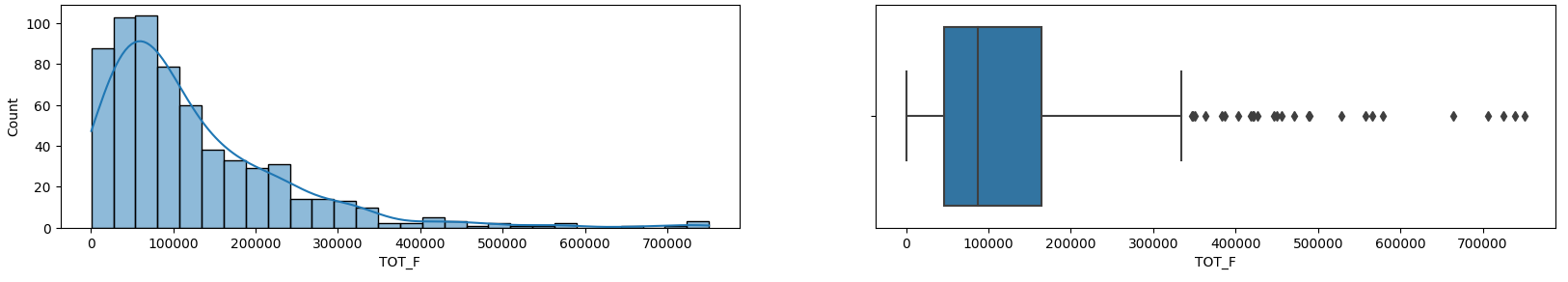
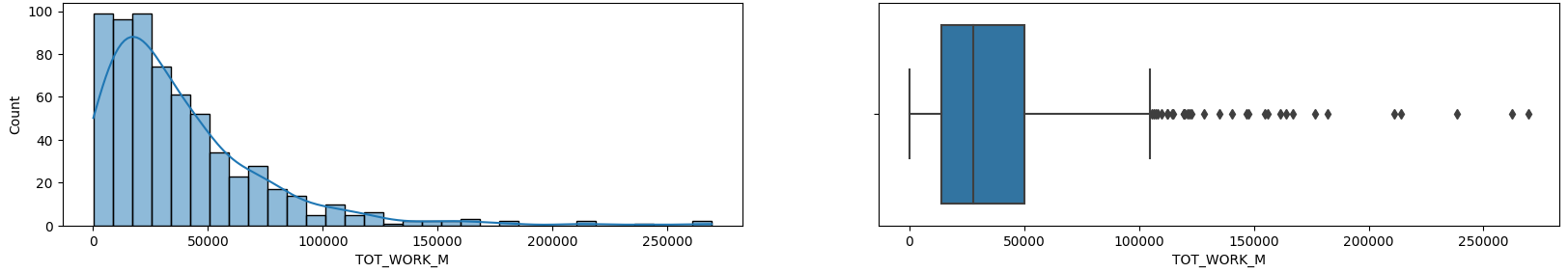
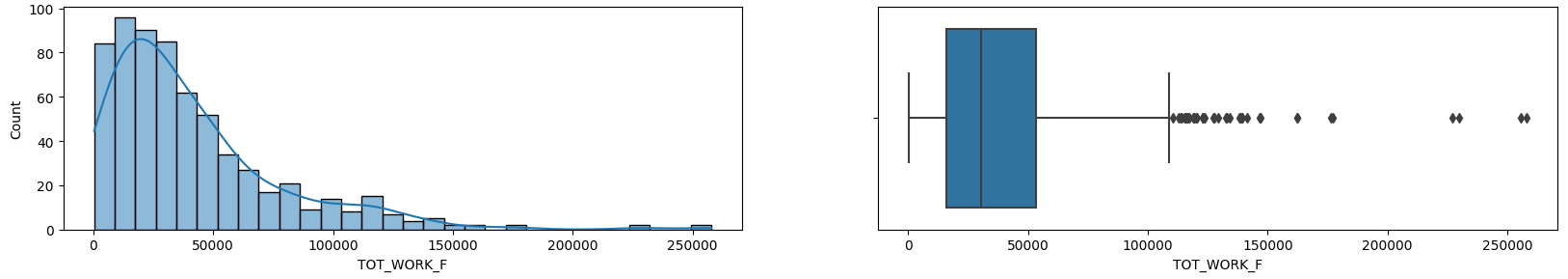


* There are no duplicates in data set
* There are no missing values

* 1. **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.**

******Solution**:

**Selected 5 variables:** No\_HH, TOT\_M, TOT\_F, TOT\_WORK\_M, and TOT\_WORK\_F

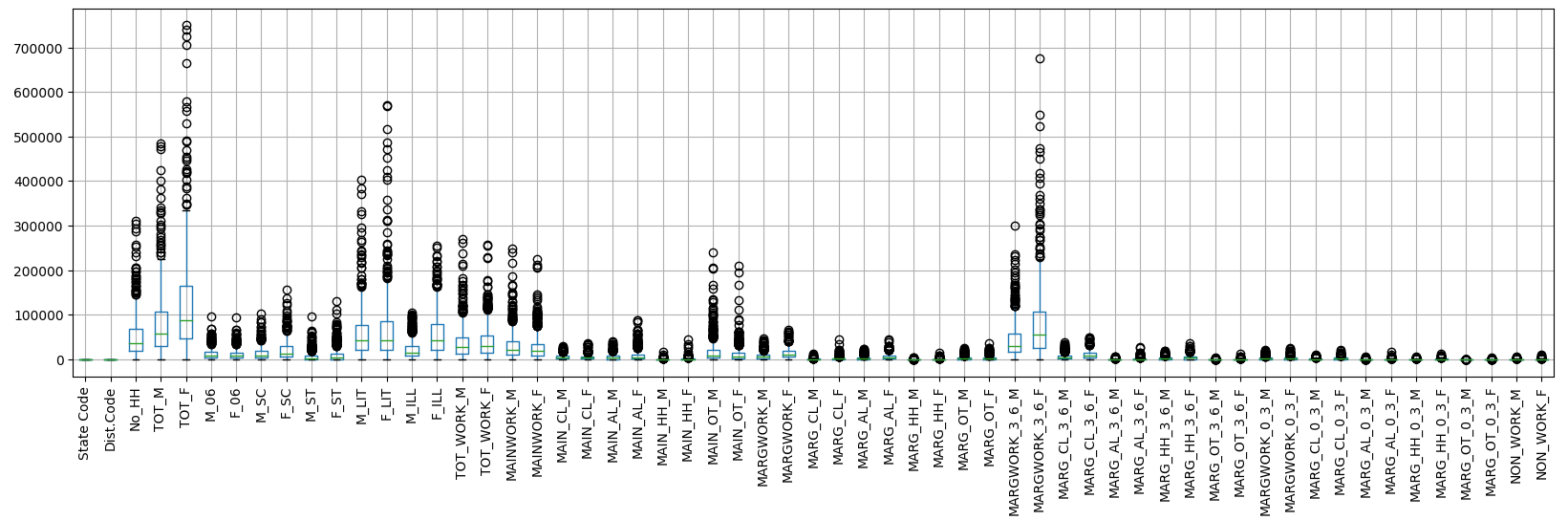


* 1. **We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?**

**Solution**: No, treating outliers is not necessary because if you remove the outliers it may lead to high bias. It is recommended to use PCA to handle the outliers.

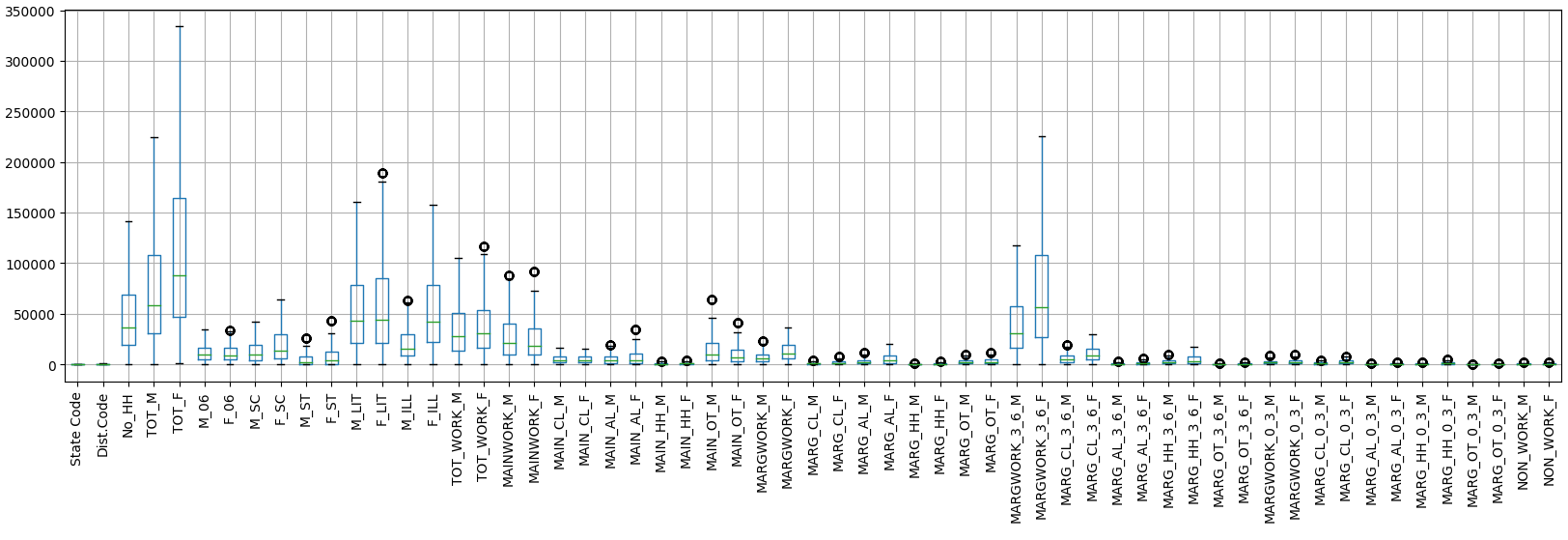
* 1. **Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment**

**Solution**:

Before scaling:

Treating outliers using IQR method (Z-score)

After scaling:

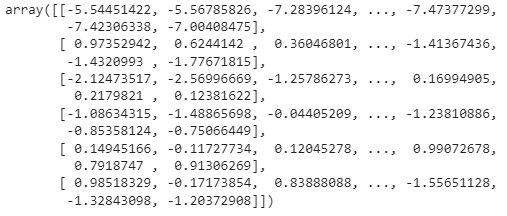
No, Scaling didn't have high impact on outliers

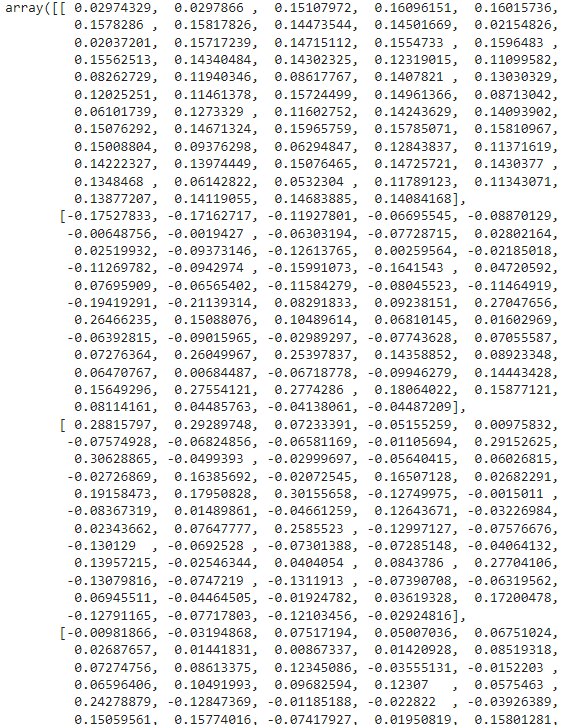
* 1. **Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.**

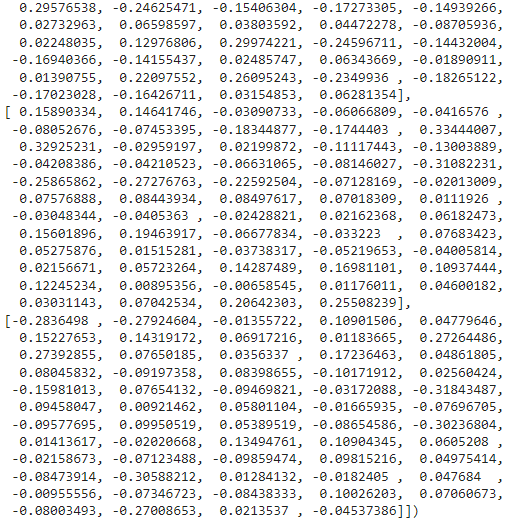
**Solution**: PCA is performed after using kmo model to check the adequacy of the data.

KMO Model = 0.9449981830784535

We imported PCA from sklearn.decomposition to find values.



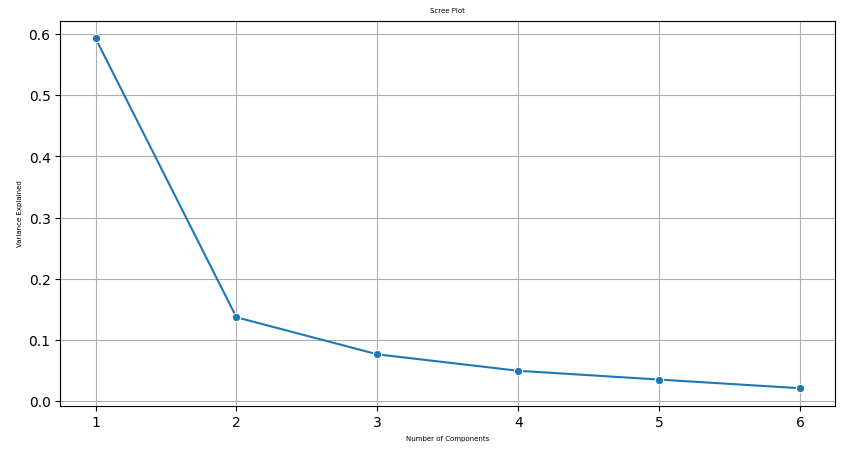
PCA components:



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

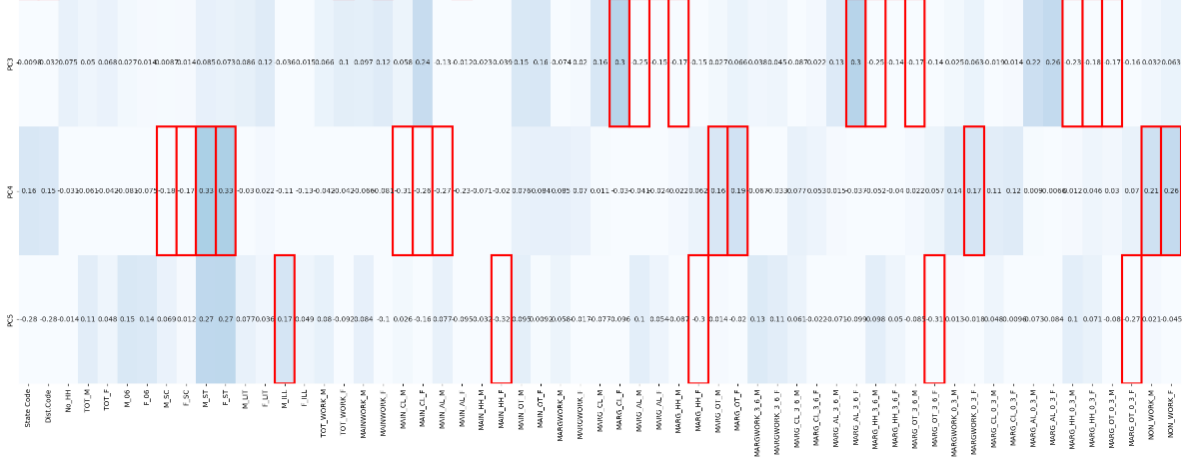


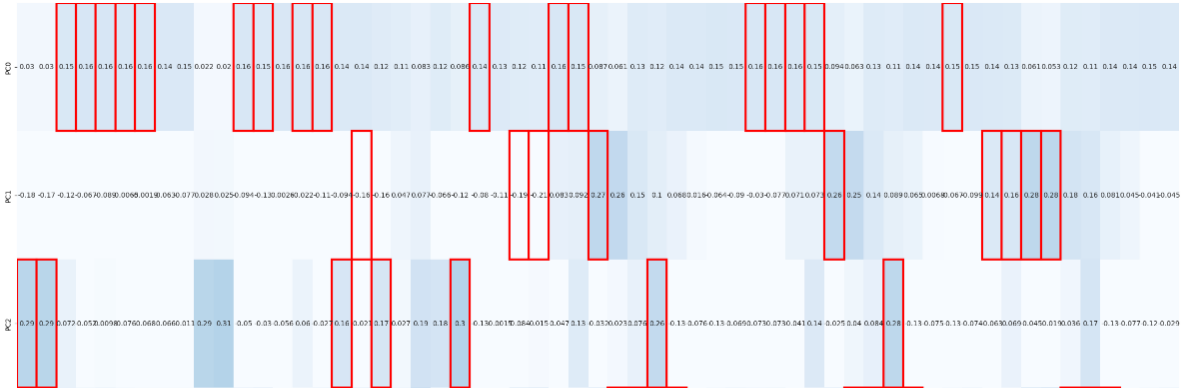
* 1. **Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.**



**Solution**: The optimum no of PC is 6, taking at least 90% explained variance

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

**Solution:**

****

* 1. **Write linear equation for first PC**

**Solution:** PC0 = [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] + [ 0.13 \* MARG\_CL\_0\_3\_F] + [0.06 \* MARG\_AL\_0\_3\_M] + [0.05 \* MARG\_AL\_0\_3\_F] + [0.12 \* 55 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]

**Thank you**