DATA MINING PROJECT REPORT

DSBA



Table of Contents

Problem 1 – Clustering

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

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

1.3 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).

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

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

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

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

1.8 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].

1.9 Conclude the project by providing summary of your learnings.

Problem 2 - PCA

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

2.2 Perform detailed Exploratory analysis by creating certain questions like

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

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

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

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

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

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

2.8 Write linear equation for first PC.

**Problem 1**

**Clustering:**

**Digital Ads Data:**

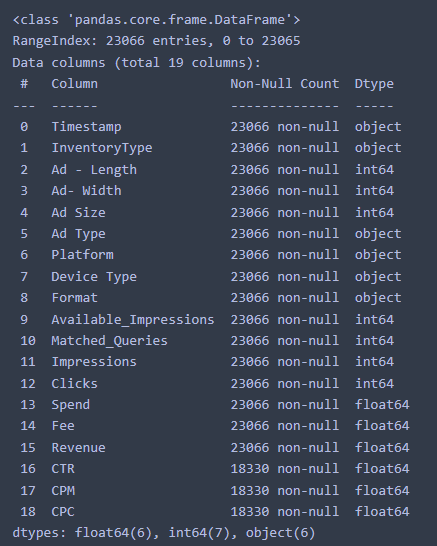
The ads24x7 is a Digital Marketing company which has now got seed funding of $10 Million. They are expanding their wings in Marketing Analytics. They collected data from their Marketing Intelligence team and now wants you (their newly appointed data analyst) to segment type of ads based on the features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

* 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.**

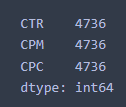
Ans.)

High-level glimpse at the data,

* There are 23066 rows and 19 columns.
* Data-Type variables,
  + Float Datatype - 6
  + Int Datatype – 7
  + Object Datatype - 6

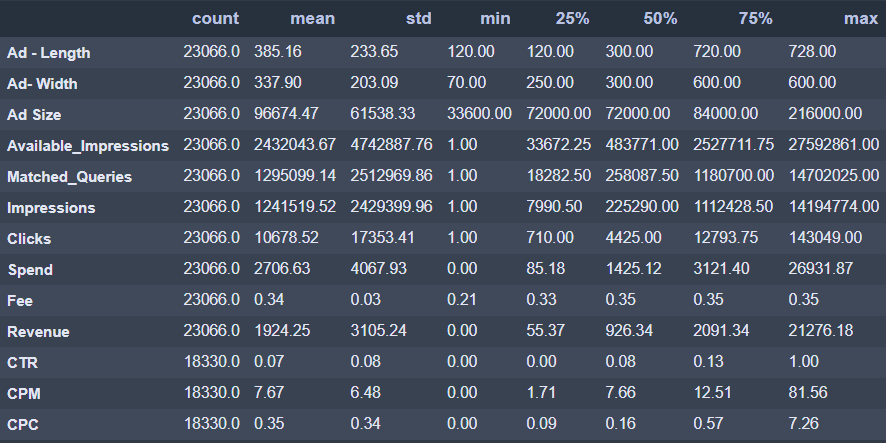


Null values,



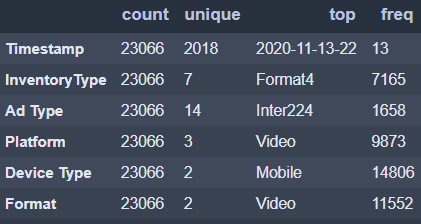
* Null values in CTR, CPM, CPC columns.
* These columns can be imputed using formulas described in the problem statement.
* There are no duplicate rows as well.

Numerical summary of the data,



* Values are in different scale, so we must do scaling of the data. Either standard scaler or Min-Max-scaling.
* There is difference between mean and median values of few columns, hence we can expect outliers and/or skewness in the data.

Categorical summary of the data,



* Timestamp column is not required for this Clustering analysis dataset.
* There are no ordinal related columns. So, it is suggested to implement One-Hot encoding.
* There are is no incorrect/bad data in categorical columns.
* InventoryType has 7 unique values and out of which 'Format4' is telecasted frequently.
* Ad-Type is almost equally shared for all categories.
* There is a pattern in Platform column, where Ads are advertised in ratio Video > Web > App.
* Mobiles are highly supported device for Ads, followed by Desktop devices.
* Almost all Ads are showcased in Video and Display formats only.
  1. **Treat missing values in CPC, CTR and CPM using the formula given.**

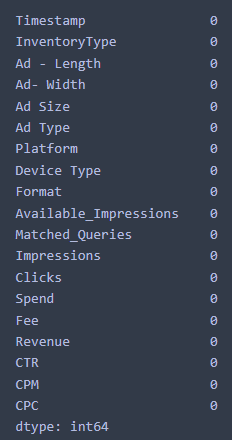
Ans.)

**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.

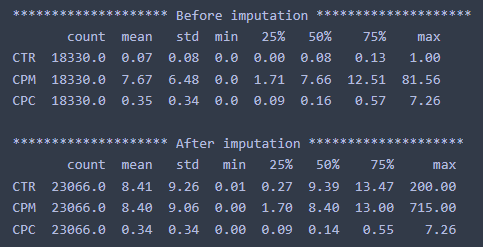
**CTR = (Total Measured Clicks / Total Measured Ad Impressions) x 100.** Note that the Total Measured Clicks refers to the 'Clicks' Column in the dataset and the Total Measured Ad Impressions refers to the 'Impressions' Column in the dataset.

Dataset after doing imputation,



No Null values

Summary on columns containing null values before and after imputation,

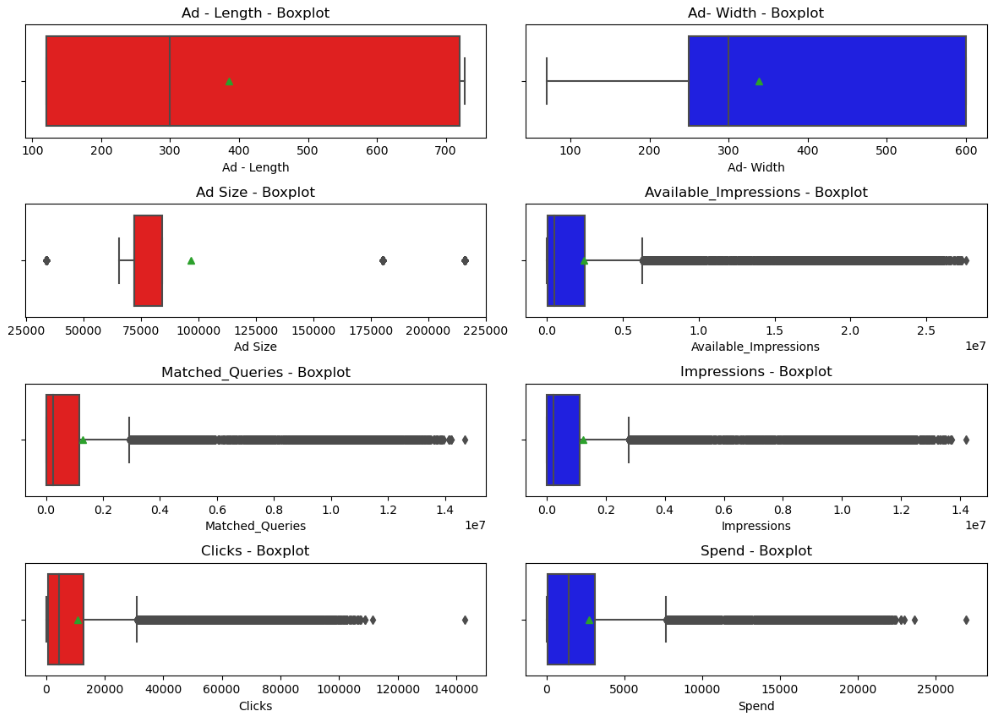


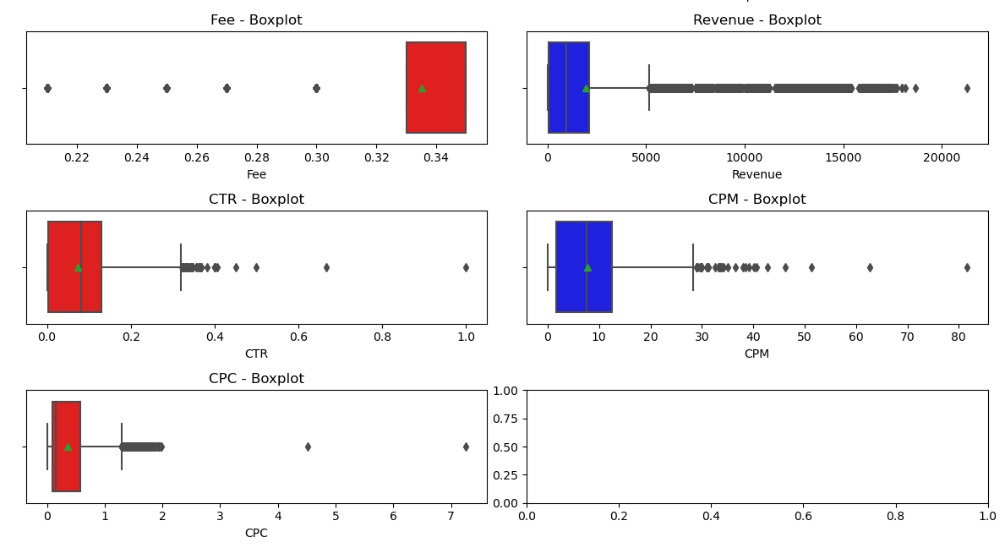
**1.3 Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ.**

Ans.)

* Yes, treating outliers is important as well as necessary for K-Means algorithm.
* K-Means algorithm designed on the basis on distances calculated between 2 data points or observations.
* When data is having outliers, then K-Means algorithm will try to get biased over outliers and grasp more noise and variance will be very high.
* As K-Means is sensitive to outliers, It is recommended to treat outliers using IQR or any other method and then do scaling to implement Clustering techniques.
* In the dataset, we use IQR rather than z-score because z-score assume data is normally distributed. But in IQR, we are not modifying the data within the box and whiskers so it is more robust to use.
* And yes, there are lots of outliers from almost all columns in the dataset.

Visualization of Boxplot to detect outliers,



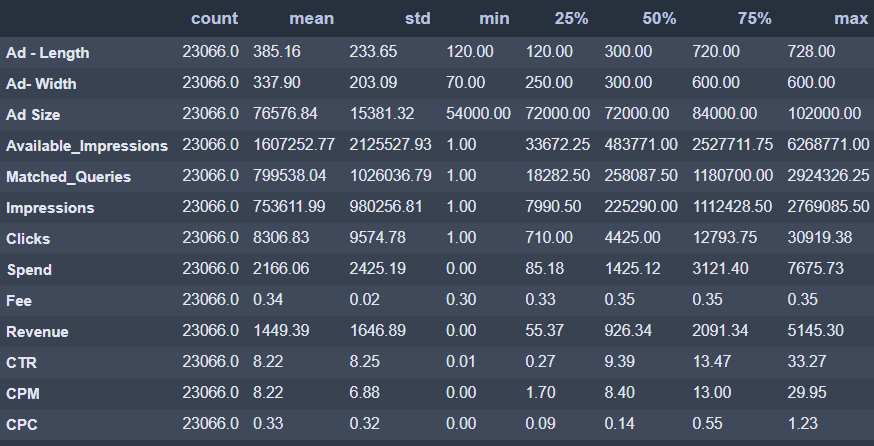


* Lots of discrepancy in the distributions of Ad-Length, Ad-Width, Ad-Size.
* Very few outliers in Ad-Size. Available\_impressions column is highly right skewed with lots of outliers.
* Columns Matched\_Queries, Impressions, Clicks, Spend are fully right skewed with lots of outliers.
* Fee column has inconsistency data distribution. Few outliers on left skewed.
* Columns Revenue, CTR, CPM, CPC are completely right skewed with lots of outliers.

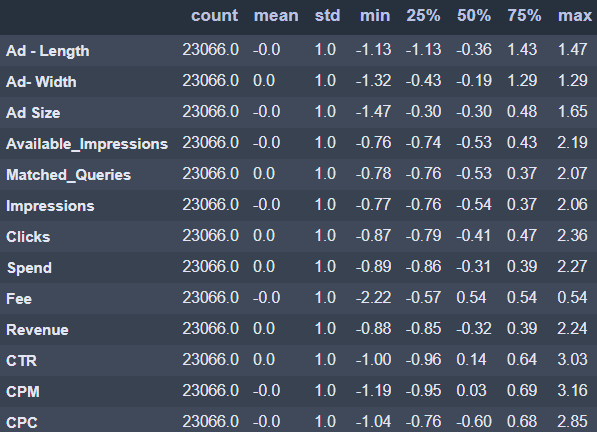
**1.4 Perform z-score scaling and discuss how it affects the speed of the algorithm.**

Ans.)

Data before scaling,



Data after scaling,



How scaling affects the speed of the Algorithm?

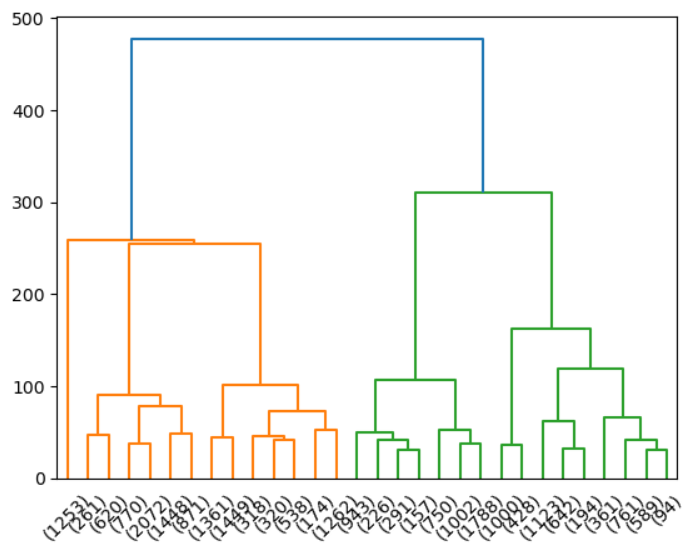
* In **Distance calculation**, when we are using distance-based algos like K-Means, Hierarchical, etc, the sole process depends on the calculation between 2 data points.
* We do use different types of metrics when to calculate between 2 data points like Euclidean, Manhattan, Minkowski, Cosine direction, etc. These distance metrics rely on calculation of distances between data points.
* Hence, scaling the variables allows us to do proper calculation because variables on different scales leads to bias and causes wrong interpretation over algos.
* When all variables are on same scale, then clustering algos like K-Means will have to do re-calculation every time to find right centroids of the clusters. Hence, scaling will **enhance faster calculations** and not much delay in finding optimum centroid.
* Outliers are completely on different scales in the dataset. These outliers will have huge impact on distance based algos. Hence, scaling the dataset will **scale outliers** to be in range and algos will perform faster and better.

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

Ans.)

Dendrogram using,

* Method – Ward
* Metric – Euclidean
* truncate\_mode – lastp
* p - 30



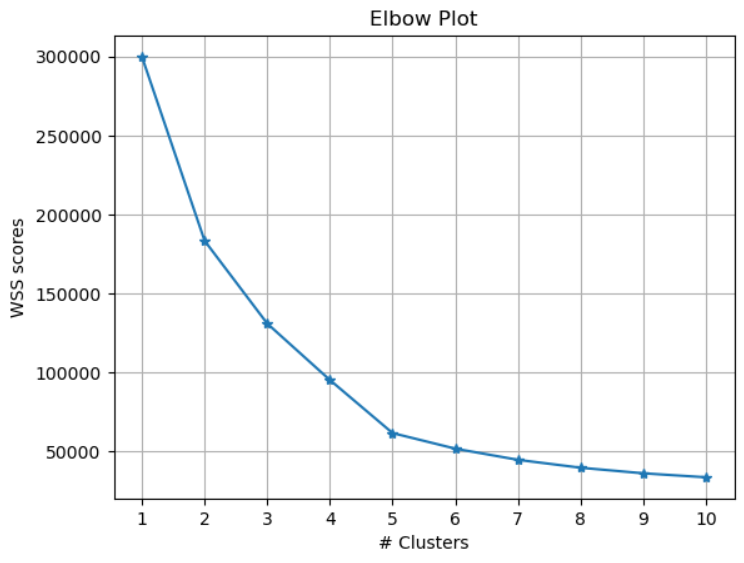
* It appears that drawing 5 clusters from above dendrogram makes sense.
* There is lots of variance ~200 margin, when we draw a horizontal line ~200, then it intersects the vertical lines at 5 points, which is an indication to use 5 clusters.
* Method-1 using **criterion as Maxclust=5** gives - [4 4 4 4 4 4 4 4 4 4]
* Method-2 using **criterion as distance=200** gives - [4 4 4 4 4 4 4 4 4 4]

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

Ans.)

Table 1- WSS using K-Means Clustering

|  |  |
| --- | --- |
| **# Clusters** | **WSS (Within Sum of Scores)** |
| 1 | 299858.0 |
| 2 | 183346.54 |
| 3 | 130874.90 |
| 4 | 95130.65 |
| 5 | 61535.91 |
| 6 | 51673.66 |
| 7 | 44594.82 |
| 8 | 39594.62 |
| 9 | 36058.68 |
| 10 | 33540.10 |



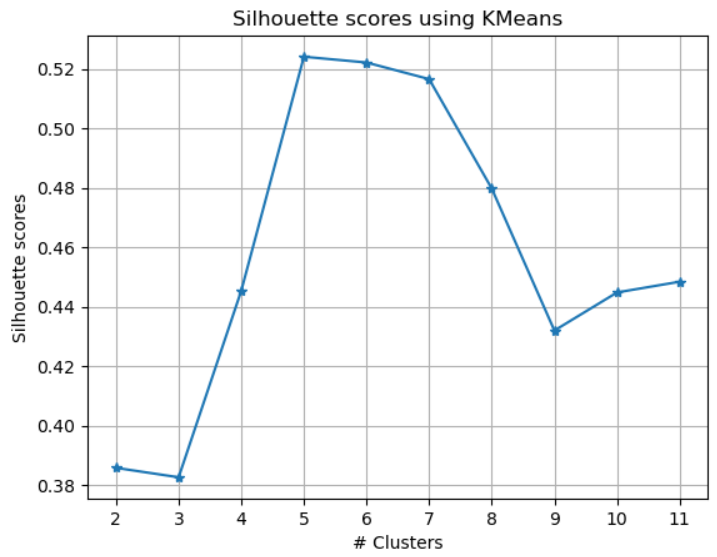
* From elbow plot, there is constant decrease in WSS score from cluster 1 to 5.
* From cluster 5 to 10, there is very less significant decrease in variance.
* Hence, we can conclude that **optimum number of clusters is 5** but we should also verify using silhouette scores**.**

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

Ans.)

Table 2- Silhouette scores using K-Means Clustering

|  |  |
| --- | --- |
| **# Clusters** | **Silhouette Scores** |
| 2 | 0.39 |
| 3 | 0.38 |
| 4 | 0.45 |
| 5 | 0.53 |
| 6 | 0.52 |
| 7 | 0.51 |
| 8 | 0.48 |
| 9 | 0.43 |
| 10 | 0.44 |
| 11 | 0.45 |



* From Silhouette scores, we can conclude that optimum number of clusters required is 5.
* **Silhouette score is high when we have 5 clusters** in above table -2.

**1.8. 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].**

Ans.)

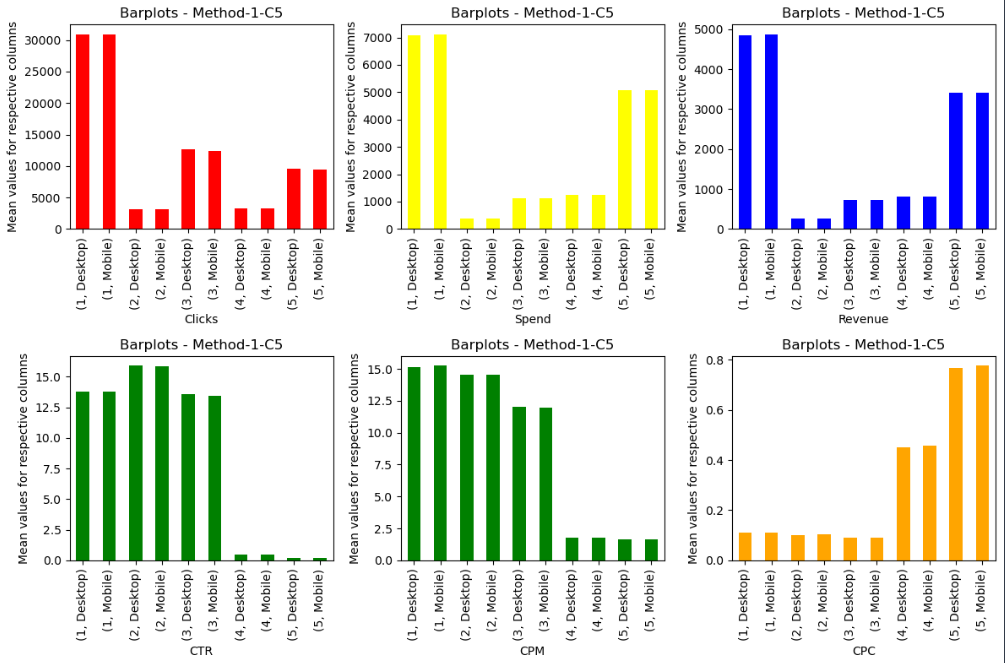
3 Cluster labels were recorded,

1. Method-1-C5 – From dendrogram using criterion Maxclust=5
2. Method-1-d200 – From dendrogram using criterion distance=200
3. K-Means clustering for k=5

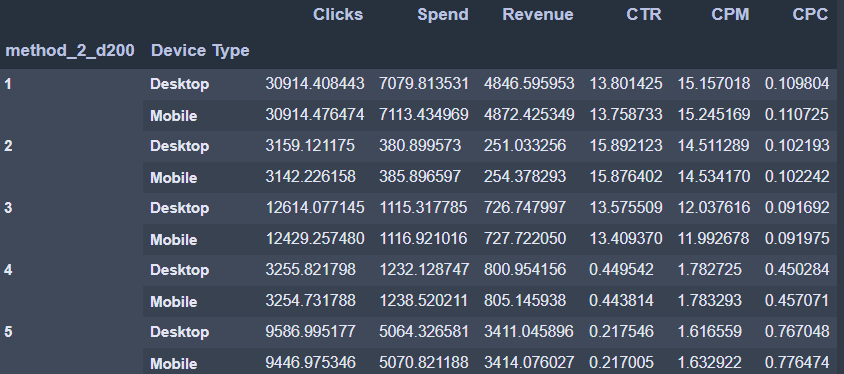
In the all 3 methods, Have grouped by mean/average.

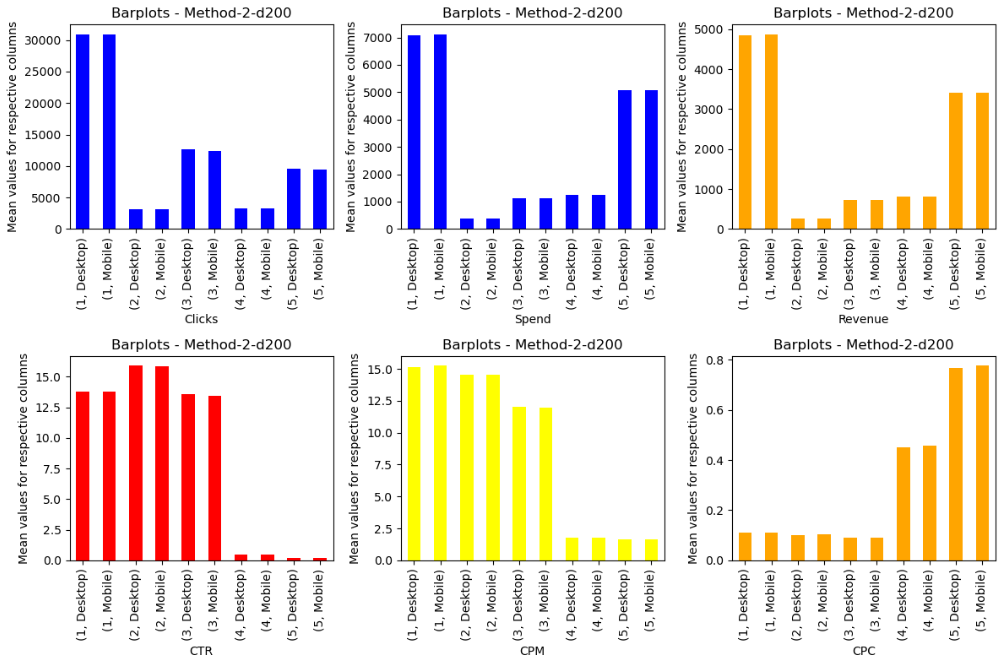
**Method-1-C5**





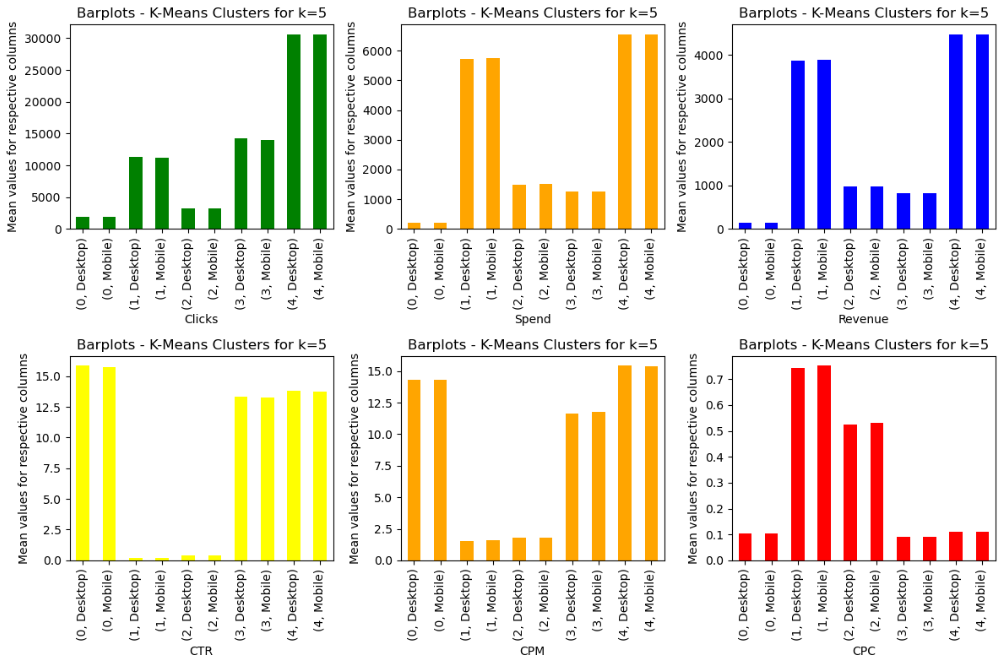
**Method-2-d200**





**K-Means Clustering for k=5**





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

Ans.)

**In method-1-C5,**

* **Clicks**: The average number of clicks varies across the clusters and device types. Cluster 1 (Desktop) has the highest average clicks, while Cluster 2 (Mobile) has the lowest average clicks.
* **Spend**: The average spending on advertising also varies across the clusters and device types. Cluster 1 (Mobile) has the highest average spend, while Cluster 2 (Desktop) has the lowest average spend.
* **Revenue**: The average revenue generated from the advertising campaigns shows variations across the clusters and device types. Cluster 1 (Mobile) has the highest average revenue, while Cluster 2 (Desktop) has the lowest average revenue.

**In method-2-d200,**

* **Cluster 1 (Desktop):** This cluster shows the highest values for Clicks, Spend, Revenue, CTR, CPM, and CPC in both Desktop and Mobile device types. It indicates that users in this cluster are highly engaged, leading to higher clicks, spend, and revenue. Focus on optimizing advertising strategies and targeting this cluster to maximize returns.
* **Cluster 2 (Desktop):** This cluster has relatively lower values for Clicks, Spend, Revenue, CTR, CPM, and CPC compared to Cluster 1. However, it still shows better performance than Clusters 3, 4, and 5. Targeting this cluster with specific advertising campaigns and optimizations can help improve the overall performance.
* **Clusters 3, 4, and 5:** These clusters generally exhibit lower values for Clicks, Spend, Revenue, CTR, CPM, and CPC compared to Clusters 1 and 2. They represent segments with lower engagement and conversion rates. It is recommended to analyze the characteristics and preferences of users in these clusters to identify opportunities for improvement and develop targeted strategies to increase their engagement.

**K-Means Clustering for k=5,**

* **Cluster 0 (Desktop):** This cluster shows relatively lower values for Clicks, Spend, Revenue, CTR, CPM, and CPC compared to other clusters. It represents a segment with lower engagement and conversion rates. Targeting this cluster with specific advertising campaigns and optimizations can help improve its performance and increase user engagement.
* **Cluster 0 (Mobile):** Similar to the desktop segment in Cluster 0, the mobile segment also exhibits lower values for Clicks, Spend, Revenue, CTR, CPM, and CPC. It is recommended to focus on mobile optimization strategies, such as improving mobile website or app experiences and mobile-specific advertising creatives, to enhance user engagement and conversion rates within this cluster.
* **Cluster 1 (Desktop):** This cluster has the highest values for Clicks, Spend, Revenue, CTR, CPM, and CPC among all the clusters. It represents a segment with high engagement and conversion rates. Businesses can capitalize on this segment by allocating more resources and developing targeted advertising strategies to maximize returns.
* **Cluster 1 (Mobile):** Similar to the desktop segment in Cluster 1, the mobile segment also shows high values for Clicks, Spend, Revenue, CTR, CPM, and CPC. It indicates that users in this cluster are highly engaged on mobile devices. Businesses should prioritize mobile advertising efforts and optimize mobile campaigns to leverage the potential of this segment.
* **Cluster 2, Cluster 3, and Cluster 4:** These clusters exhibit intermediate values for Clicks, Spend, Revenue, CTR, CPM, and CPC. They represent segments with moderate engagement and conversion rates. It is recommended to analyze the characteristics and preferences of users in these clusters to identify opportunities for improvement and develop targeted strategies to enhance their performance.

**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.

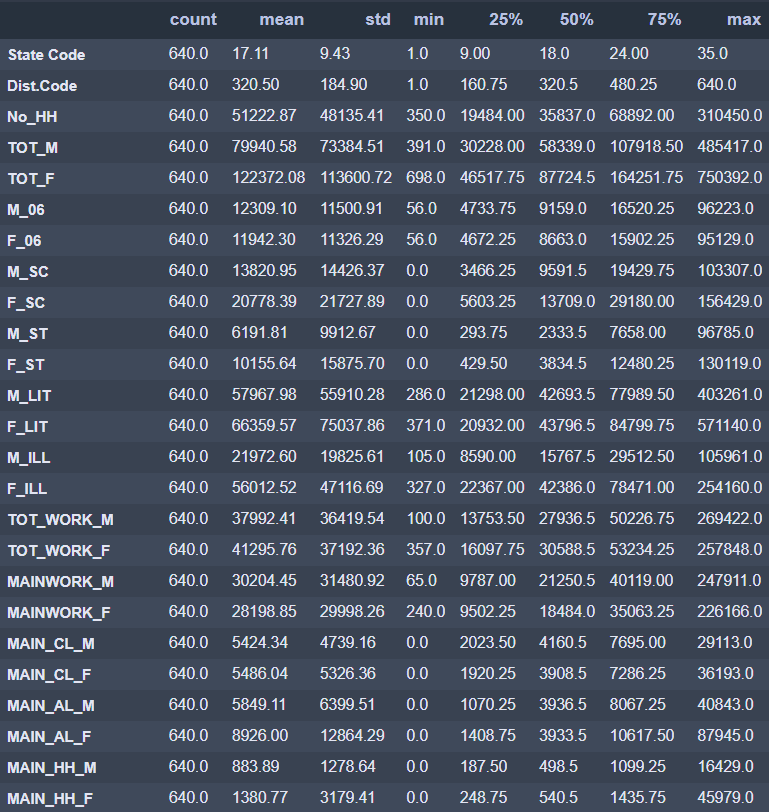
**2.1 Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.**

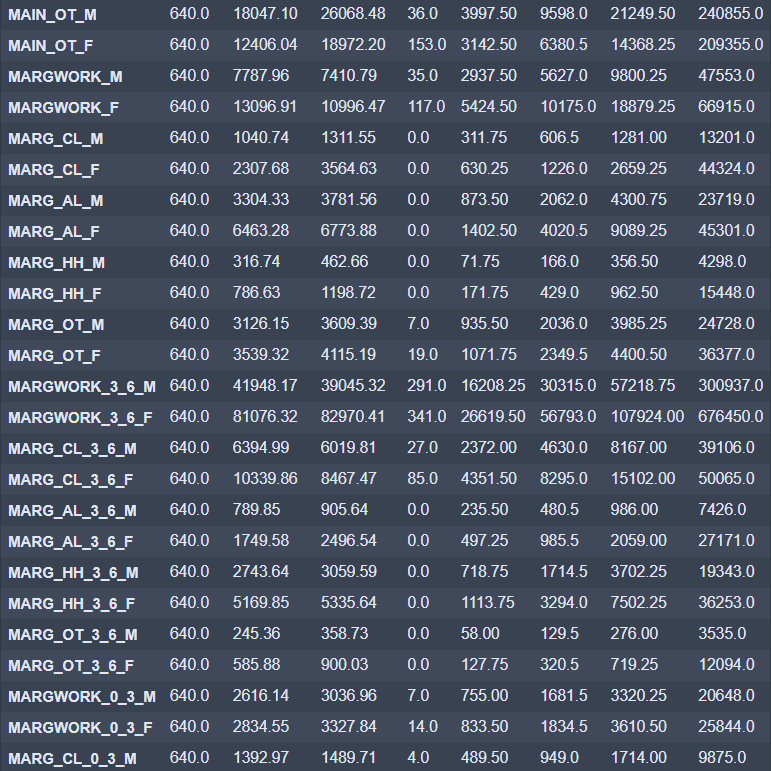
Ans.)

High-level glimpse at the data,

* There are 640 rows and 61 columns.
* Data-Type variables,
  + Int Datatype – 59
  + Object Datatype – 2
* No null values in the data.
* No duplicate entries/rows.

Numerical summary of the data,

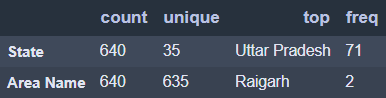






* Values are in different scale, so we must do scaling of the data. Either standard scaler or Min-Max-scaling.
* There is difference between mean and median values of few columns, hence we can expect outliers and/or skewness in the data.

Categorical summary of the data,



* Area Name column has 635 unique values out of 640 rows. Which is high likely that whole column is unique.
* Both State and Area Name do not add value to the further analysis. We can drop them before scaling or PCA analysis.

**2.2 Perform detailed Exploratory analysis by creating certain questions like**

**(i) Which state has highest gender ratio and which has the lowest?**

**(ii) Which district has the highest & lowest gender ratio? (Example Questions).**

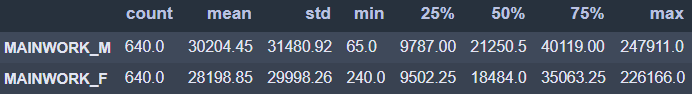
**Pick 5 variables out of the given 24 variables below for EDA: No\_HH, TOT\_M, TOT\_F, M\_06, F\_06, M\_SC, F\_SC, M\_ST, F\_ST, M\_LIT, F\_LIT, M\_ILL, F\_ILL, TOT\_WORK\_M, TOT\_WORK\_F, MAINWORK\_M, MAINWORK\_F, MAIN\_CL\_M, MAIN\_CL\_F, MAIN\_AL\_M, MAIN\_AL\_F, MAIN\_HH\_M, MAIN\_HH\_F, MAIN\_OT\_M, MAIN\_OT\_F**

Ans.)

Let us do EDA about Main Workers from the dataset.

1) In total, who are working more from all states combined? Males or Females?

* Total Male workers are 1,93,30,846
* Total Female workers are 1,80,47,262



* We can see there are more main male workers compared to female workers.
* But the difference is very huge, which is 12,83,584

2) From which state and area does the highest and lowest number of workers coming from?

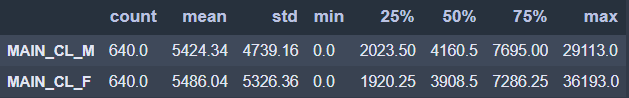
2a) States

* *Uttar Pradesh and Maharashtra* states has more number of Main workers among all states i.e. 4,00,000+
* *Lakshadweep* has very less Main workers i.e., 3711

2b) Areas

* Areas *Mumbai Suburban and Bangalore* has 4,00,000+ Main workers.
* There are less than 1000 Main workers from *Kargil, Nicobars, and Dibang Valley areas.*

3) In Cultivators sector, does males or females participate more?



* In cultivator sector, Total Male workers are 34,71,579
* In cultivator sector, Total Female workers are 35,11,067
* In cultivator sector, Female workers are more by 39,488

4) In Cultivators sector, from which state and area does the highest and lowest number of workers coming from?

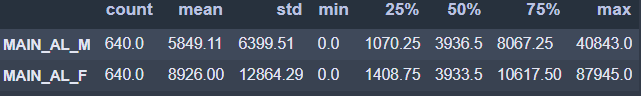
4a) States

* *Uttar Pradesh* has more number of Cultivator workers among all states i.e. 12+ lakhs.
* *Lakshadweep* has NO cultivator workers.

4b) Areas

* Area *Jaunpur* has 60,000+ Cultivator workers.
* There are <10 Cultivator workers from *New Delhi and Yanam* areas.

5) In Agricultural sector, does males or females participate more?



* In Agricultural sector, Total Male workers are 37,43,430
* In Agricultural sector, Total Female workers are 57,12,637
* In Agricultural sector, Female workers are more by 19,69,207

6) In Agricultural sector, from which state and area does the highest and lowest number of workers coming from?

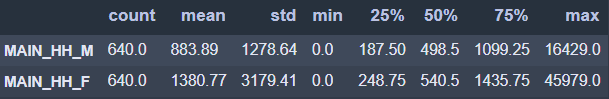
6a) States

* *Andhra Pradesh* has more number of Agricultural workers among all states i.e. 15+ lakhs.
* *Lakshadweep* has NO Agricultural workers.

6b) Areas

* Area *Kurnool* has 1,20,000+ Agricultural workers.
* There are <10 Agricultural workers from *Nicobars and Dibang Valley* areas.

7) In Household industry, does males or females participate more?



* In Household industry, Total Male workers are 5,65,692
* In Household industry, Total Female workers are 8,83,695
* In Household industry, Female workers are more by 3,18,003

8) In Household industry, from which state and area does the highest and lowest number of workers coming from?

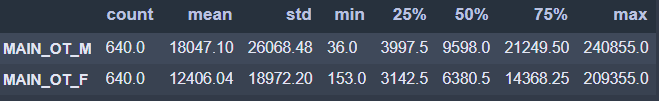
8a) States

* *Uttar Pradesh* has more number of Household workers among all states i.e. 2.5+ lakhs.
* *Lakshadweep and Daman & Diu* has <100 Household workers.

8b) Areas

* Area *Dakshina Kannada and Murshidabad* has 40,000+ Household workers.
* There are <10 Household workers from *Anjaw and Dibang Valley* areas.

9) In other workers/ Non-workers category, does males or females participate more?



* In other workers/Non-workers category, Total Male workers are 1,15,50,145
* In other workers/Non-workers category, Total Female workers are 79,39,863
* In other workers/Non-workers category, Male workers are more by 36,10,282

10) In other workers/ Non-workers category, from which state and area does the highest and lowest number of workers coming from?

10a) States

* *Uttar Pradesh* has more number of other workers/Non-workers category workers among all states i.e. 20+ lakhs.
* *Lakshadweep, Dadara & Nagar Havelli, and Daman & Diu* has <10000 other workers/Non-workers category workers.

10b) Areas

* Areas *Mumbai Suburban and Bangalore* has 4,00,000+ other workers/Non-workers category workers.
* There are <300 other workers/Non-workers category workers from *Anjaw, Longleng, and Dibang Valley* areas.

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

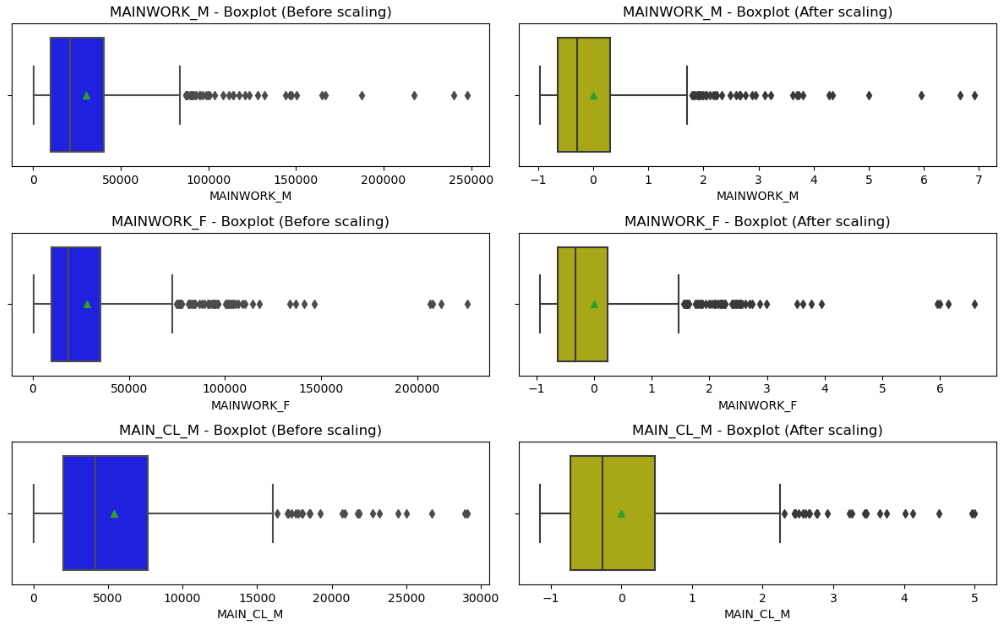
Ans.)

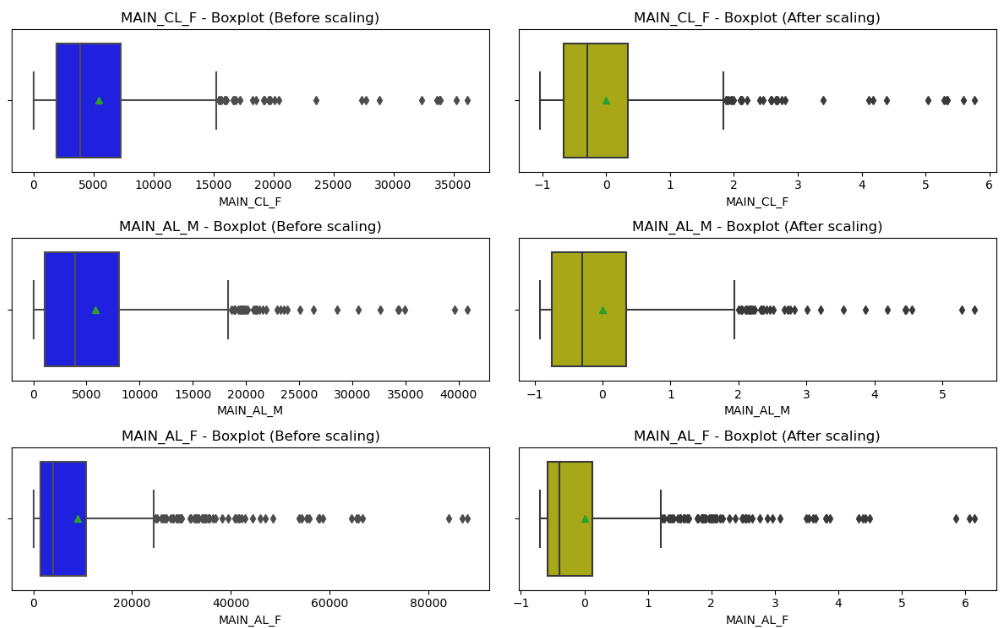
* No, we should not treat outliers in this dataset.
* Reason because, we are handling with **Population census data** where every outlier represents different **demographical or geographical** representations of the population.
* It is not recommended to treat outliers when dealing with Population like census data. Hence, robust analysis required before taking any action on the outliers in the dataset.

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

Ans.)

Glimpse at few columns – Before scaling and After scaling





* After looking at above plots, we don’t find much differences between outliers from the data.
* Outliers are showing same pattern before scaling and after scaling data.
* Because scaling does not guarantee on outliers. Scaling will only bring the scales of the data on to similar line/scale for easy computation.
* Hence, to conclude that we don’t find any impact on outliers after scaling data. Scaling will only focus on changing the scales of the data on same scale.

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

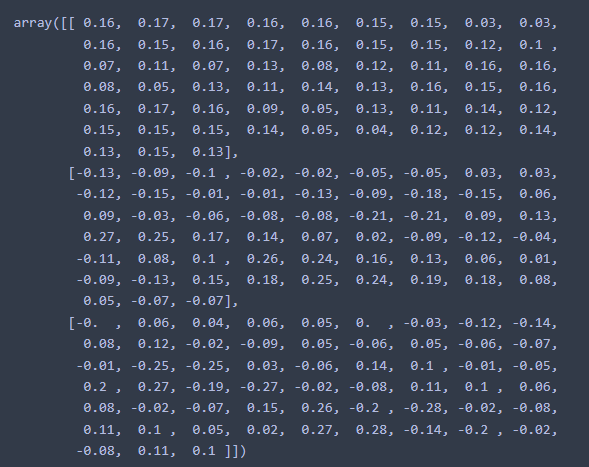
Ans.)

Pre-checks before applying PCA,

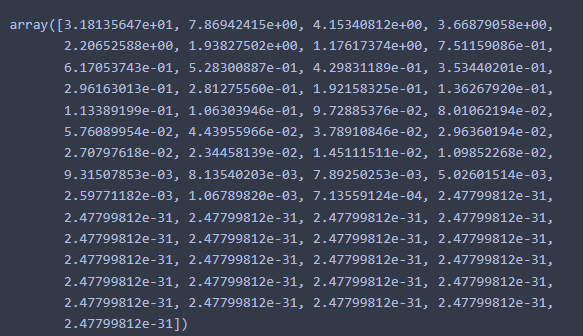
* **Bartletts Test of Sphericity**
  + Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.
    - H0: All variables in the data are uncorrelated
    - Ha: At least one pair of variables in the data are correlated
  + If the null hypothesis cannot be rejected, then PCA is not advisable.
  + If the p-value is small, then we can reject the null hypothesis and agree that there is atleast one pair of variables in the data which are correlated hence PCA is recommended.
  + When implemented on our data, we received **P-value close to 0 (<0.5).** Hence, we can reject null hypothesis by saying atleast 1 pair of variables in the data are correlated.
* **KMO Test**
  + The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.
  + Generally, if MSA is less than 0.5, PCA is not recommended, since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.
  + From output, we have received that **MSA value as 0.8 (>0.7).** Hence, we have strong evidence that reduction is required on the dimensions.

From the above two tests, we can conclude that we can perform PCA on our dataset.

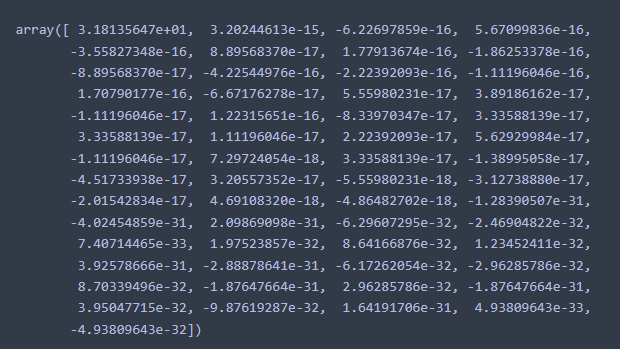
Glimpse of few Eigen vectors,



Glimpse of 57 Eigen values,



Glimpse at Covariance matrix’s 1st value out of (57,57) matrix data,



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

Ans.)

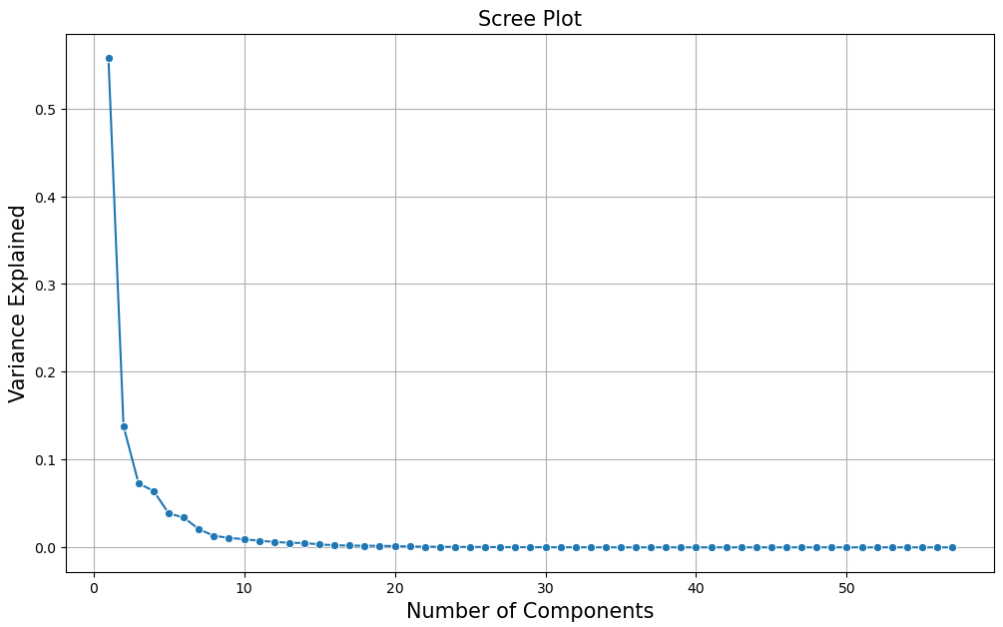


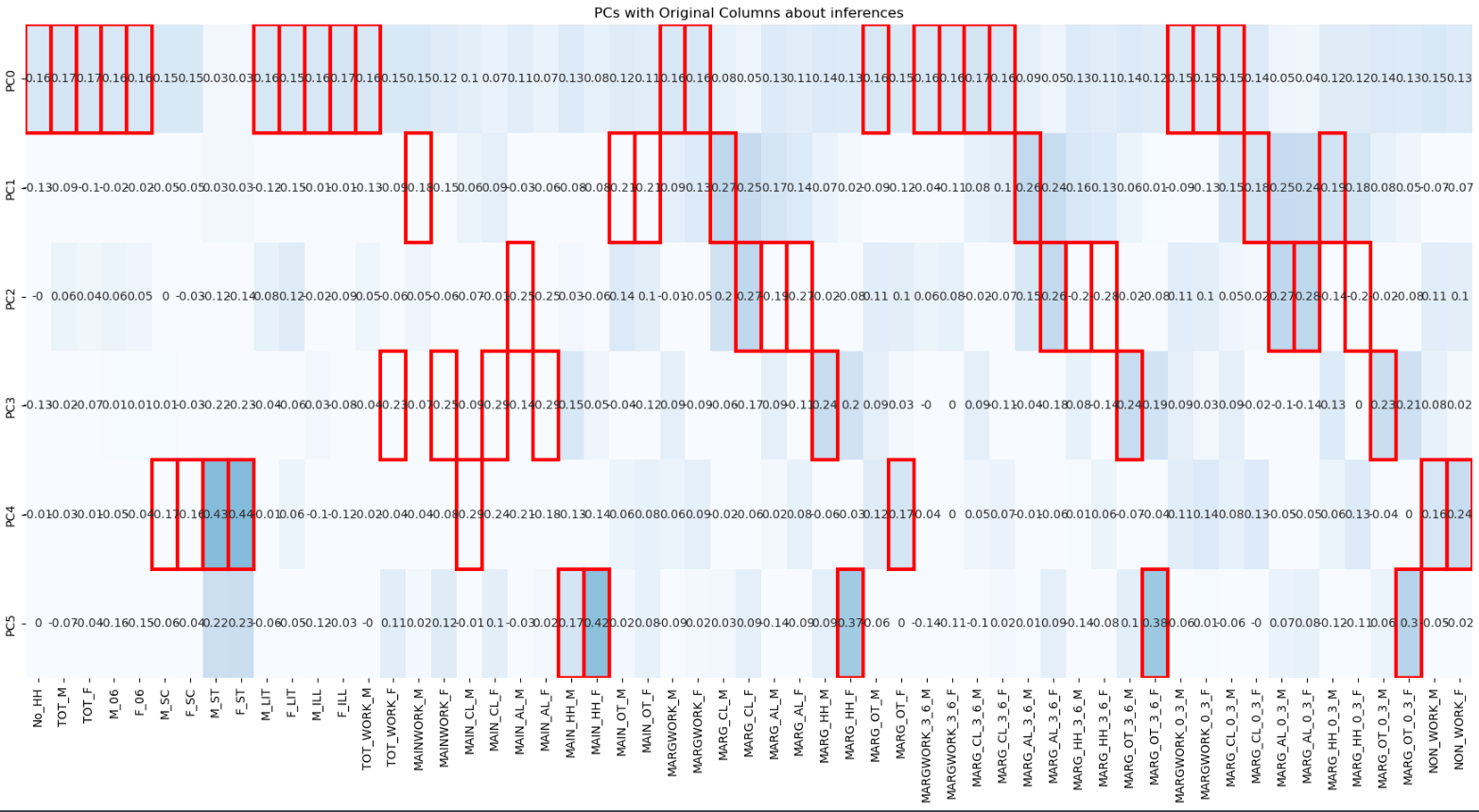
Table 3 – PCs with Explained Variance Ratio

|  |  |  |
| --- | --- | --- |
| **PCs** | **Explained\_Variance\_Ratio** | **Cumulative\_EVR\_in\_%** |
| PC-1 | 0.56 | 55.73 |
| PC-2 | 0.14 | 69.51 |
| PC-3 | 0.07 | 76.79 |
| PC-4 | 0.06 | 83.21 |
| PC-5 | 0.04 | 87.08 |
| PC-6 | 0.03 | 90.47 |
| PC-7 | 0.02 | 92.53 |
| PC-8 | 0.01 | 93.85 |
| PC-9 | 0.01 | 94.93 |
| PC-10 | 0.01 | 95.85 |
| PC-11 | 0.01 | 96.61 |
| PC-12 | 0.01 | 97.23 |
| PC-13 | 0.01 | 97.75 |
| PC-14 | 0 | 98.24 |
| PC-15 | 0 | 98.57 |
| PC-16 | 0 | 98.81 |
| PC-17 | 0 | 99.01 |
| PC-18 | 0 | 99.2 |
| PC-19 | 0 | 99.37 |
| PC-20 | 0 | 99.51 |
| PC-21 | 0 | 99.61 |
| PC-22 | 0 | 99.69 |
| PC-23 | 0 | 99.75 |
| PC-24 | 0 | 99.81 |
| PC-25 | 0 | 99.85 |
| PC-26 | 0 | 99.89 |
| PC-27 | 0 | 99.92 |
| PC-28 | 0 | 99.94 |
| PC-29 | 0 | 99.96 |
| PC-30 | 0 | 99.97 |
| PC-31 | 0 | 99.98 |
| PC-32 | 0 | 99.99 |
| PC-33 | 0 | 100 |
| PC-34 | 0 | 100 |
| PC-35 | 0 | 100 |
| PC-36 | 0 | 100 |
| PC-37 | 0 | 100 |
| PC-38 | 0 | 100 |
| PC-39 | 0 | 100 |
| PC-40 | 0 | 100 |
| PC-41 | 0 | 100 |
| PC-42 | 0 | 100 |
| PC-43 | 0 | 100 |
| PC-44 | 0 | 100 |
| PC-45 | 0 | 100 |
| PC-46 | 0 | 100 |
| PC-47 | 0 | 100 |
| PC-48 | 0 | 100 |
| PC-49 | 0 | 100 |
| PC-50 | 0 | 100 |
| PC-51 | 0 | 100 |
| PC-52 | 0 | 100 |
| PC-53 | 0 | 100 |
| PC-54 | 0 | 100 |
| PC-55 | 0 | 100 |
| PC-56 | 0 | 100 |
| PC-57 | 0 | 100 |

* Highlighted 6 PCs from table 3 are sufficient for further analysis.
* Hence, as per dimensionality reduction algorithm, we have **reduced dataset columns from 57 to 6 by retaining 90%** of the variance or information explained by the dataset.

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

Ans.)



* From above plot, PC0 which is in first row explains 57% of the variance or information from the data.
* The variances which are not explained by PC0 will be targeted by next PCs in line.
* We can see that M\_SC, F\_SC, M\_ST, F\_ST are 4 variables which are not explained by any top 1 or 2 PCs. Hence, PC4 was able to provide info about these 4 variables. By using domain knowledge, we could rephrase PC4 as Total\_SC\_ST variable.
* Similarly, PC2 is talking about Marginal workers data. Hence, we can rephrase it as Marg\_Workers variable.
* Usually the top PCs i.e., PC0, PC1, PC2 tries to explain more variance about the data, which is as follows PC0 > PC1 > PC2 and so on.
* From Table 3, we can find out that our top 3 PCs contribute around 75% of the variance from the dataset.

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

Ans.)

Linear Equation for first PC i.e., PC-1 is as combination of all 57 columns,

PC-1 = (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