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.

The following three features are commonly used in digital marketing:

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

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

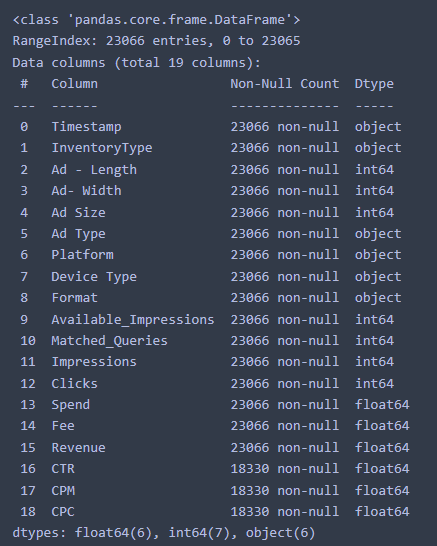
**CTR = (Total Measured Clicks / Total Measured Ad Impressions) x 100.** Note that the Total Measured Clicks refers to the 'Clicks' Column in the dataset and the Total Measured Ad Impressions refers to the 'Impressions' Column in the dataset.

* 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.**

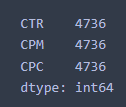
Ans.)

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

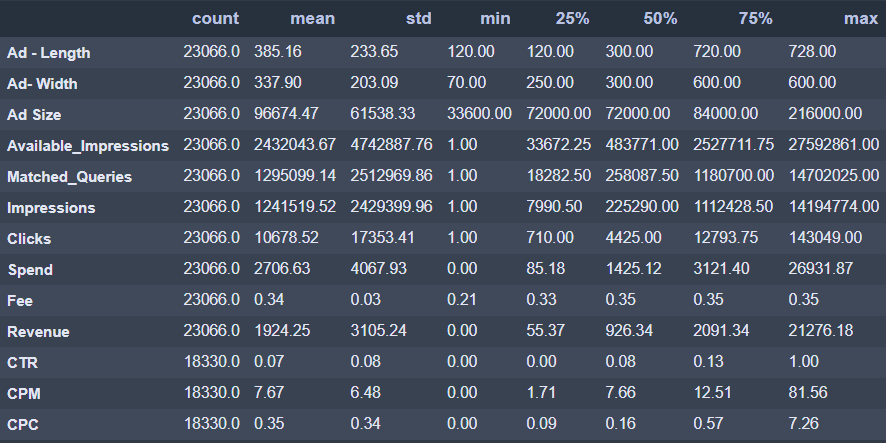


Null values,



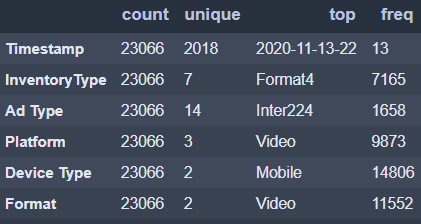
* Null values in CTR, CPM, CPC columns.
* These columns can be imputed using formulas described in the problem statement.
* There are no duplicate rows as well.

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.**

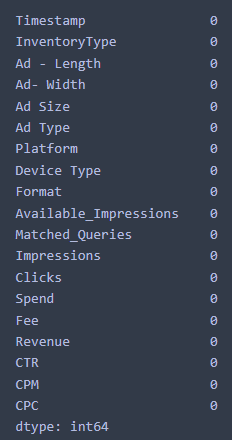
Ans.)

**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.

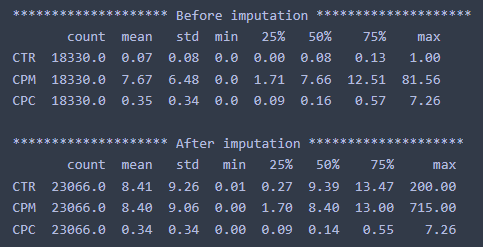
**CTR = (Total Measured Clicks / Total Measured Ad Impressions) x 100.** Note that the Total Measured Clicks refers to the 'Clicks' Column in the dataset and the Total Measured Ad Impressions refers to the 'Impressions' Column in the dataset.

Dataset after doing imputation,



No Null values

Summary on columns containing null values,

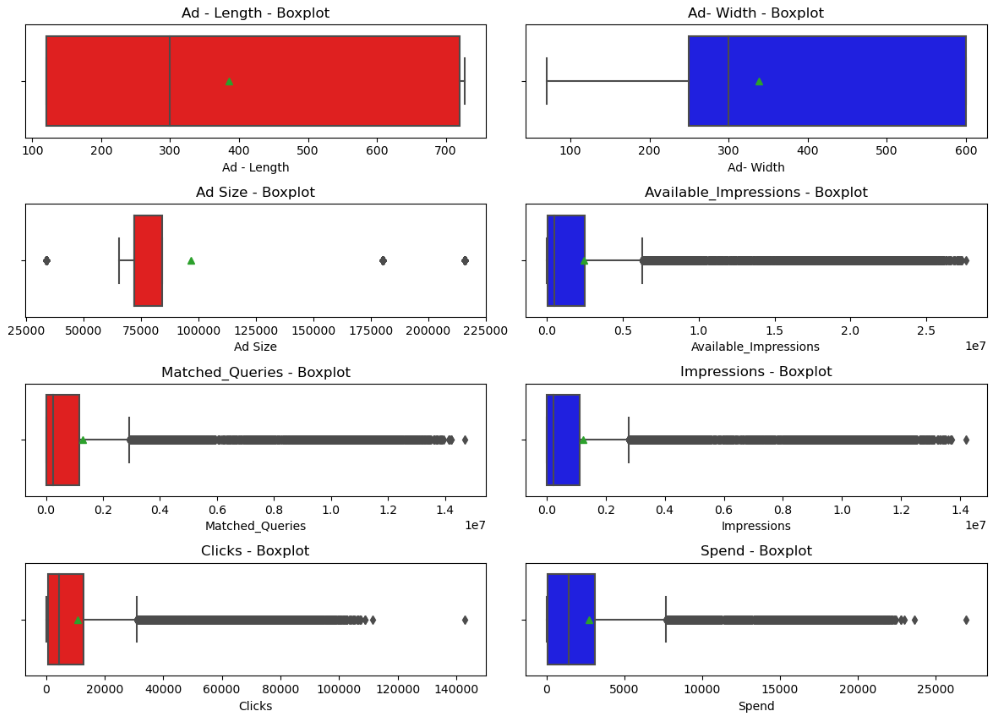


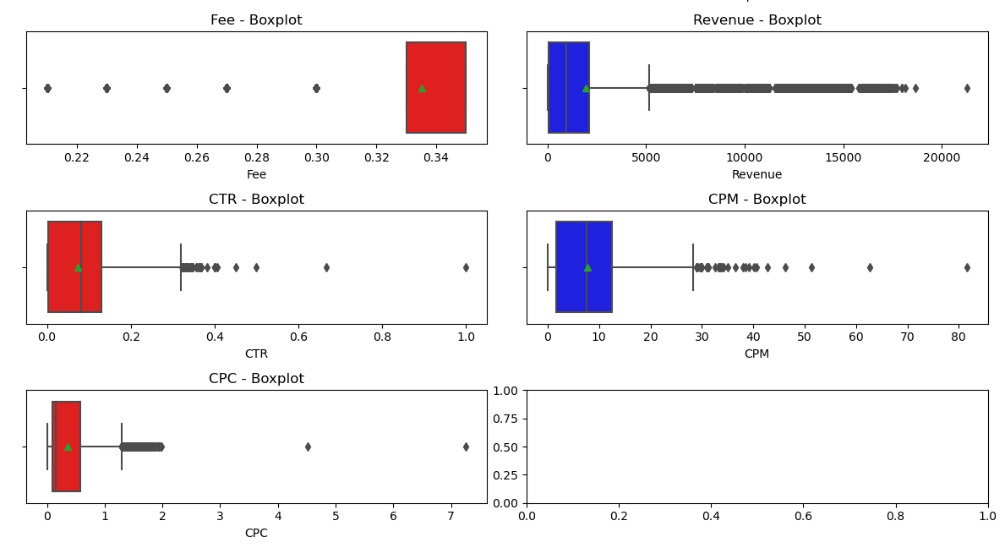
**1.3 Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ.**

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,



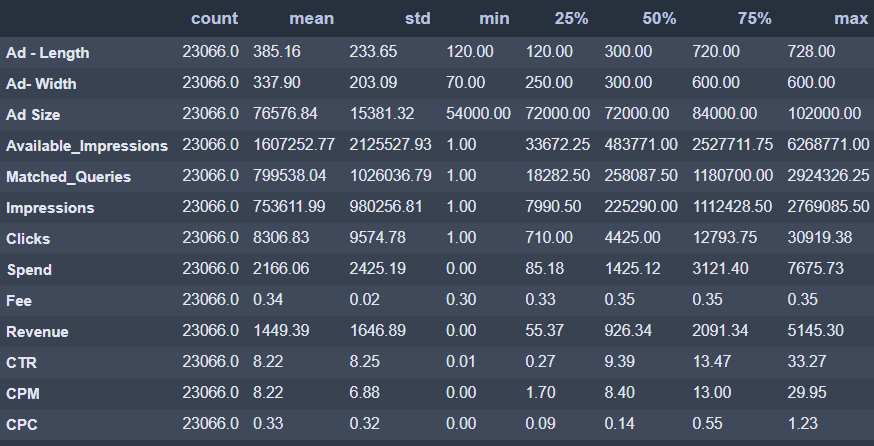


* Lots of discrepancy in the distributions of Ad-Length, Ad-Width, Ad-Size.
* Very few outliers in Ad-Size. Available\_impressions column is highly right skewed with lots of outliers.
* Columns Matched\_Queries, Impressions, Clicks, Spend are fully right skewed with lots of outliers.
* Fee column has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

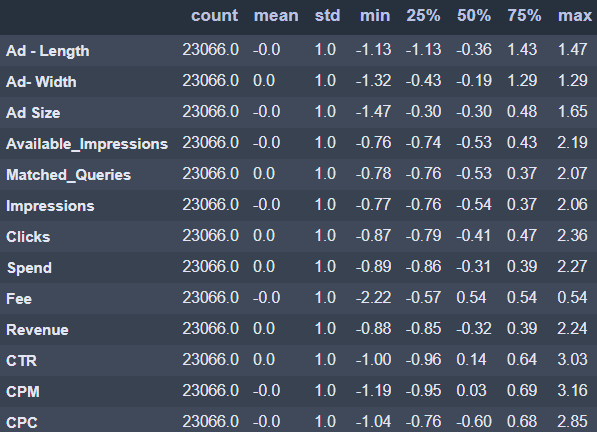
**1.4 Perform z-score scaling and discuss how it affects the speed of the algorithm.**

Ans.)

Data before scaling,



Data after scaling,

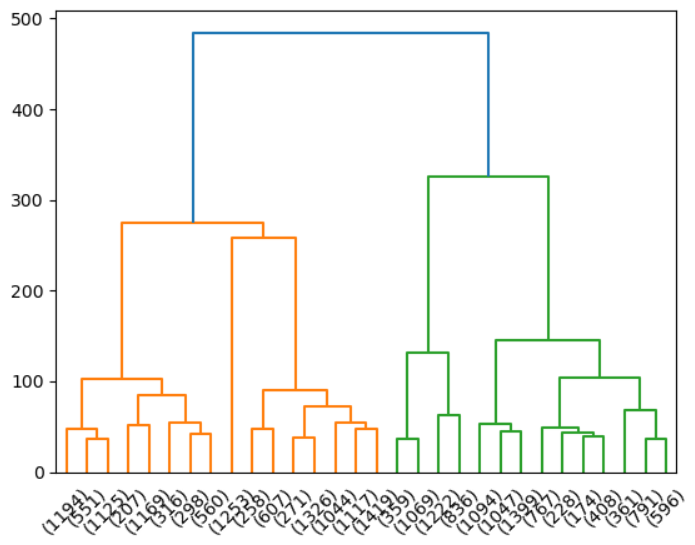


**1.5 Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.**

Ans.)

Dendrogram using,

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



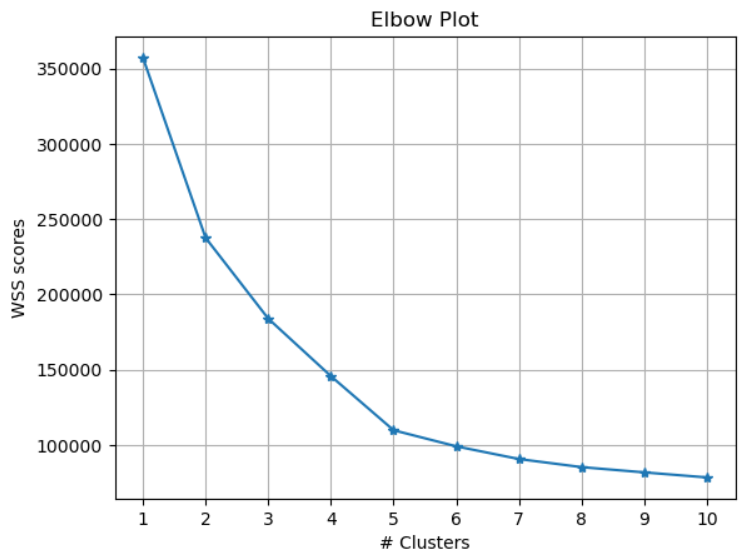
* It appears that drawing 5 clusters from above dendrogram makes.
* There is lots of variance ~200 margin, when we draw a horizontal line ~200, then it intersects the vertical lines at 5 points, which is an indication to use 5 clusters.

**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 | 357149.63 |
| 2 | 237475.97 |
| 3 | 183827.39 |
| 4 | 145618.51 |
| 5 | 109697.25 |
| 6 | 99011.03 |
| 7 | 90578.60 |
| 8 | 85204.98 |
| 9 | 81712.29 |
| 10 | 78356.23 |



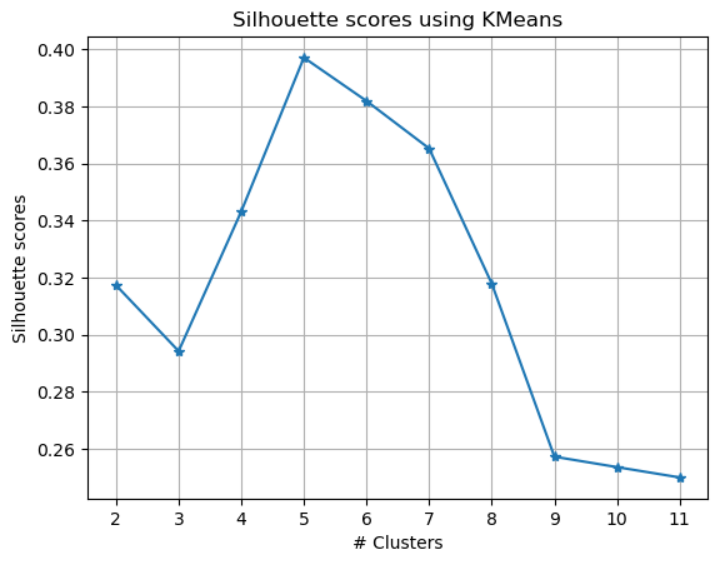
* From elbow plot, there is constant decrease in WSS score from cluster 1 to 5.
* From cluster 5 to 10, there is very less significant decrease in variance.
* Hence, we could conclude that **optimum number of clusters is 5.**

**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 | 0.32 |
| 3 | 0.29 |
| 4 | 0.34 |
| 5 | 0.40 |
| 6 | 0.38 |
| 7 | 0.36 |
| 8 | 0.32 |
| 9 | 0.26 |
| 10 | 0.25 |
| 11 | 0.25 |



* 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.)

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

Ans.)

**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.

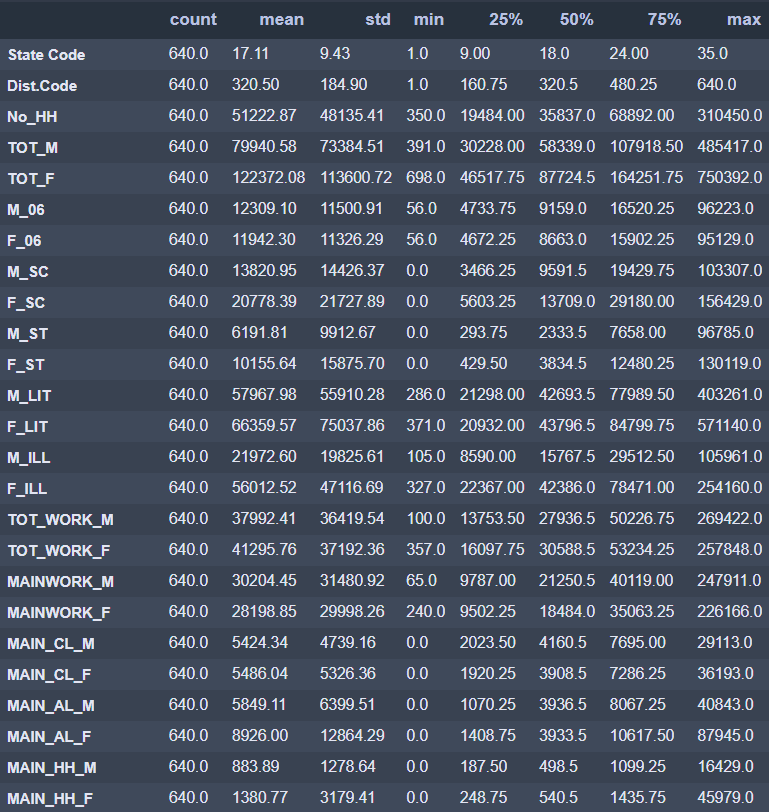
**2.1 Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.**

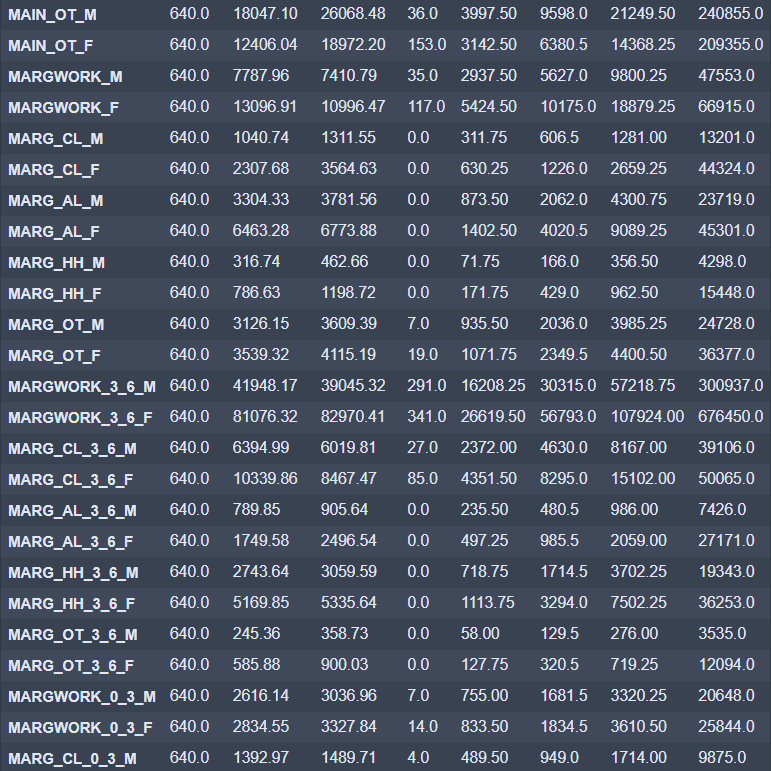
Ans.)

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,

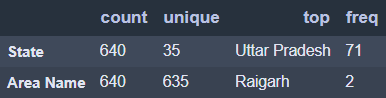






* Values are in different scale, so we must do scaling of the data. Either standard scaler or Min-Max-scaling.
* There is difference between mean and median values of few columns, hence we can expect outliers and/or skewness in the data.

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).**

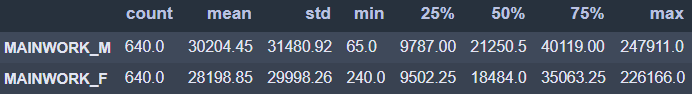
**Pick 5 variables out of the given 24 variables below for EDA: No\_HH, TOT\_M, TOT\_F, M\_06, F\_06, M\_SC, F\_SC, M\_ST, F\_ST, M\_LIT, F\_LIT, M\_ILL, F\_ILL, TOT\_WORK\_M, TOT\_WORK\_F, MAINWORK\_M, MAINWORK\_F, MAIN\_CL\_M, MAIN\_CL\_F, MAIN\_AL\_M, MAIN\_AL\_F, MAIN\_HH\_M, MAIN\_HH\_F, MAIN\_OT\_M,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

Ans.)

Let us do EDA about Main Workers from the dataset.

1) On an average, 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 not very huge, which is 12,83,584

2) From where does the highest and lowest number of workers coming from?

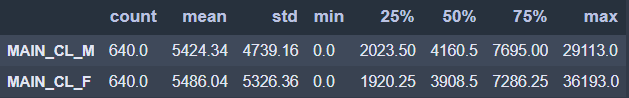
2a) States

* 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 3471579
* In cultivator sector, Total Female workers are 3511067
* In cultivator sector, Female workers are more by 39488

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

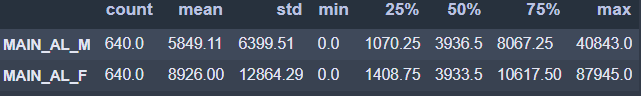
4a) States

* 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 3743430
* In Agricultural sector, Total Female workers are 5712637
* In Agricultural sector, Female workers are more by 1969207

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

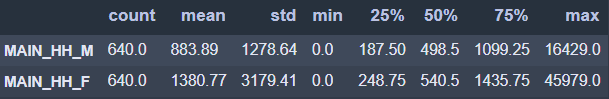
6a) States

* 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 where does the highest and lowest number of workers coming from?

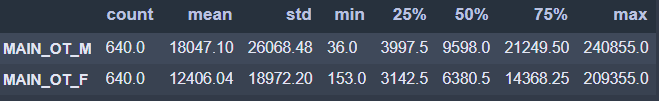
8a) States

* 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 where 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

* Area 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

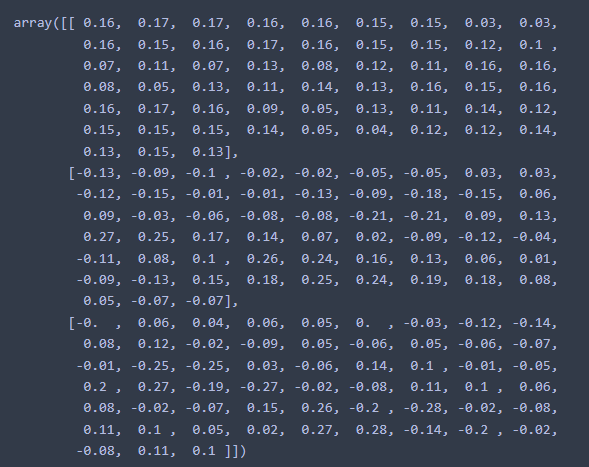
**2.5. Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.**

Ans.)

Pre-checks before applying PCA,

* calculate\_bartlett\_sphericity – 0.0 (< 0.5)
  + That means, atleast one-pair of independent variables are showing multi-collinearity between them.
* calculate\_kmo – 0.8 (> 0.5)
  + Va

Glimpse of few eigen-vectors,



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

Ans.)

**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.)

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

Ans.)