Clustering Clean\_Ads

**Project by Ganesh Aryan**

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.

The Data Dictionary and the detailed description of the formulas for CPM, CPC and CTR are given in the sheet 2 of the Clustering Clean ads\_data Excel File.

|  |  |  |
| --- | --- | --- |
| **Sl. No** | **Column Name** | **Column Description** |
| 1 | Timestamp | The Timestamp of the particular Advertisement. |
| 2 | Inventory Type | The Inventory Type of the particular Advertisement. Format 1 to 7. This is a Categorical Variable. |
| 3 | Ad - Length | The Length Dimension of the particular Advertisement. |
| 4 | Ad- Width | The Width Dimension of the particular Advertisement. |
| 5 | Ad Size | The Overall Size of the particular Advertisement. Length\*Width. |
| 6 | Ad Type | The type of the particular Advertisement. This is a Categorical Variable. |
| 7 | Platform | The platform in which the particular Advertisement is displayed. Web, Video or App. This is a Categorical Variable. |
| 8 | Device Type | The type of the device which supports the particular Advertisement. This is a Categorical Variable. |
| 9 | Format | The Format in which the Advertisement is displayed. This is a Categorical Variable. |
| 10 | Available\_Impressions | How often the particular Advertisement is shown. An impression is counted each time an Advertisement is shown on a search result page or other site on a Network. |
| 11 | Matched\_Queries | Matched search queries data is pulled from Advertising Platform and consists of the exact searches typed into the search Engine that generated clicks for the particular Advertisement. |
| 12 | Impressions | The impression count of the particular Advertisement out of the total available impressions. |
| 13 | Clicks | It is a marketing metric that counts the number of times users have clicked on the particular advertisement to reach an online property. |
| 14 | Spend | It is the amount of money spent on specific ad variations within a specific campaign or ad set. This metric helps regulate ad performance. |
| 15 | Fee | The percentage of the Advertising Fees payable by Franchise Entities. |
| 16 | Revenue | It is the income that has been earned from the particular advertisement. |
| 17 | CTR | CTR stands for "Click through rate". CTR is the number of clicks that your ad receives divided by the number of times your ad is shown. Formula used here is CTR = Total Measured Clicks / Total Measured Ad Impressions x 100. Note that the Total Measured Clicks refers to the 'Clicks' Column and the Total Measured Ad Impressions refers to the 'Impressions' Column. |
| 18 | CPM | CPM stands for "cost per 1000 impressions." Formula used here is CPM = (Total Campaign Spend / Number of Impressions) \* 1,000. Note that the Total Campaign Spend refers to the 'Spend' Column and the Number of Impressions refers to the 'Impressions' Column. |
| 19 | CPC | CPC stands for "Cost-per-click". Cost-per-click (CPC) bidding means that you pay for each click on your ads. The Formula used here is CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column and the Number of Clicks refers to the 'Clicks' Column. |

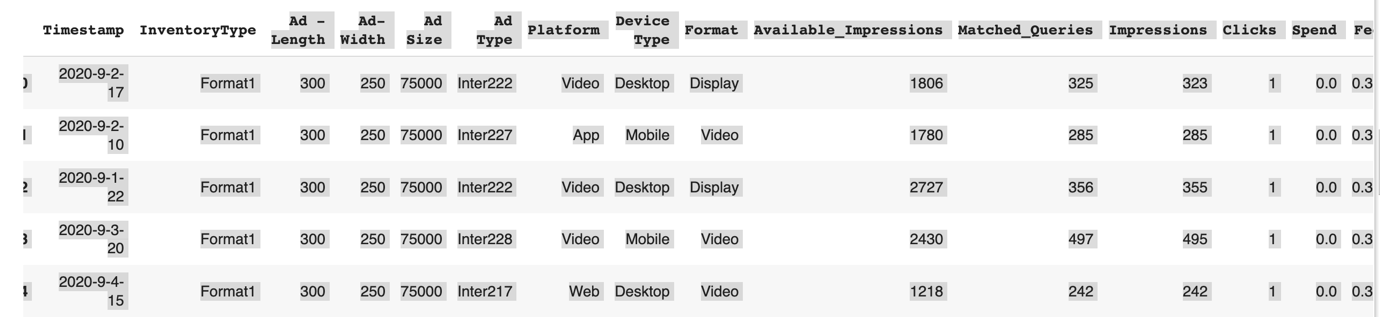
**Data Description**

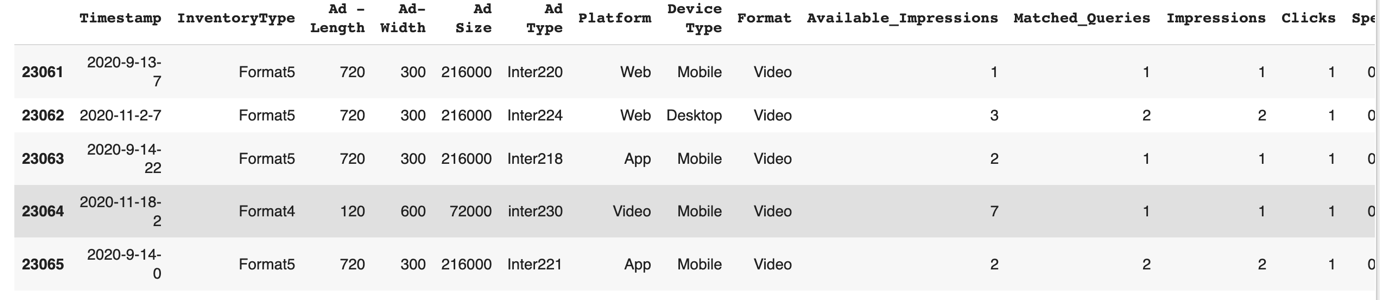
**Question1:**

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

**Basic Analysis of Data provided:**

*Top & bottom 5 rows of the data:*





**Data Info:**

Data columns (total 19 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Timestamp 23066 non-null object

1 InventoryType 23066 non-null object

2 Ad - Length 23066 non-null int64

3 Ad- Width 23066 non-null int64

4 Ad Size 23066 non-null int64

5 Ad Type 23066 non-null object

6 Platform 23066 non-null object

7 Device Type 23066 non-null object

8 Format 23066 non-null object

9 Available\_Impressions 23066 non-null int64

10 Matched\_Queries 23066 non-null int64

11 Impressions 23066 non-null int64

12 Clicks 23066 non-null int64

13 Spend 23066 non-null float64

14 Fee 23066 non-null float64

15 Revenue 23066 non-null float64

16 CTR 18330 non-null float64

17 CPM 18330 non-null float64

18 CPC 18330 non-null float64

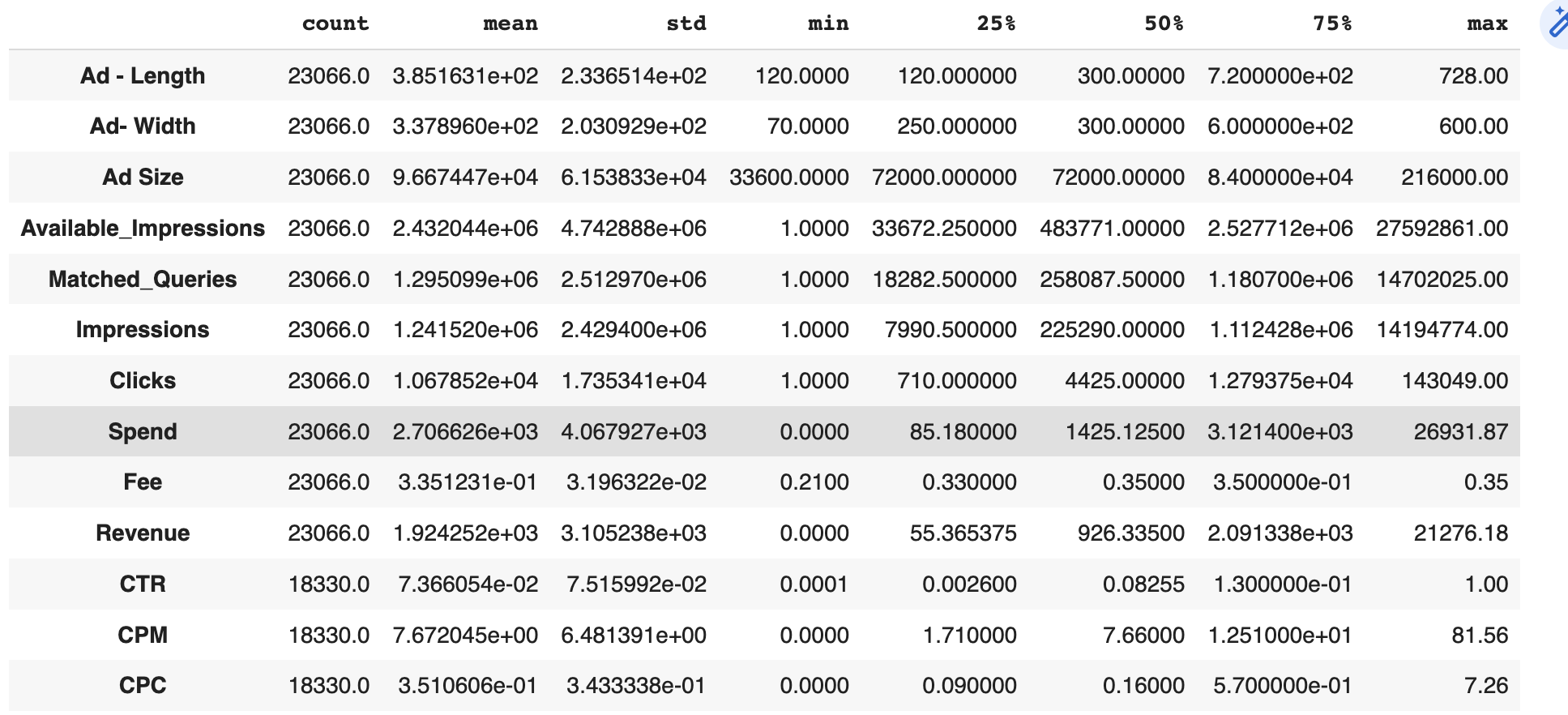
***Observations:***

*There are total* ***19*** *columns and* ***23066*** *rows in the data*

*There are null values present in columns CTR/CPM & CPC*

*6 columns are categorical and remaining are iNT and float*

**Summery:**



**Duplicate:** No Duplicate values found in dataset.

**Null Values:** There are Missing values in columns CTR, CPM and CPC

Timestamp 0

InventoryType 0

Ad - Length 0

Ad- Width 0

Ad Size 0

Ad Type 0

Platform 0

Device Type 0

Format 0

Available\_Impressions 0

Matched\_Queries 0

Impressions 0

Clicks 0

Spend 0

Fee 0

Revenue 0

CTR 4736

CPM 4736

CPC 4736

**Question2:**

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

Treated missing values in CPC, CTR and CPM by using the **“impute\_missing”** function.

Timestamp 0

InventoryType 0

Ad - Length 0

Ad- Width 0

Ad Size 0

Ad Type 0

Platform 0

Device Type 0

Format 0

Available\_Impressions 0

Matched\_Queries 0

Impressions 0

Clicks 0

Spend 0

Fee 0

Revenue 0

CTR 0

CPM 0

CPC 0

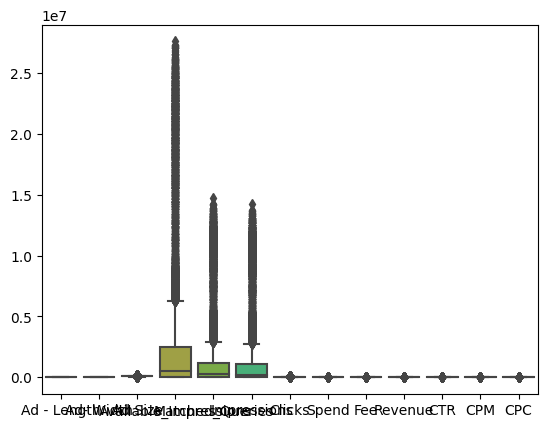
dtype: int64

**Question3:**

**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.**

**Outliers.**

There are outliers and considering the fact that removing the outliers from the this data might affect the quality of clustering given the fact that data available in these outliers might be useful. Therefore continuing with as it is.

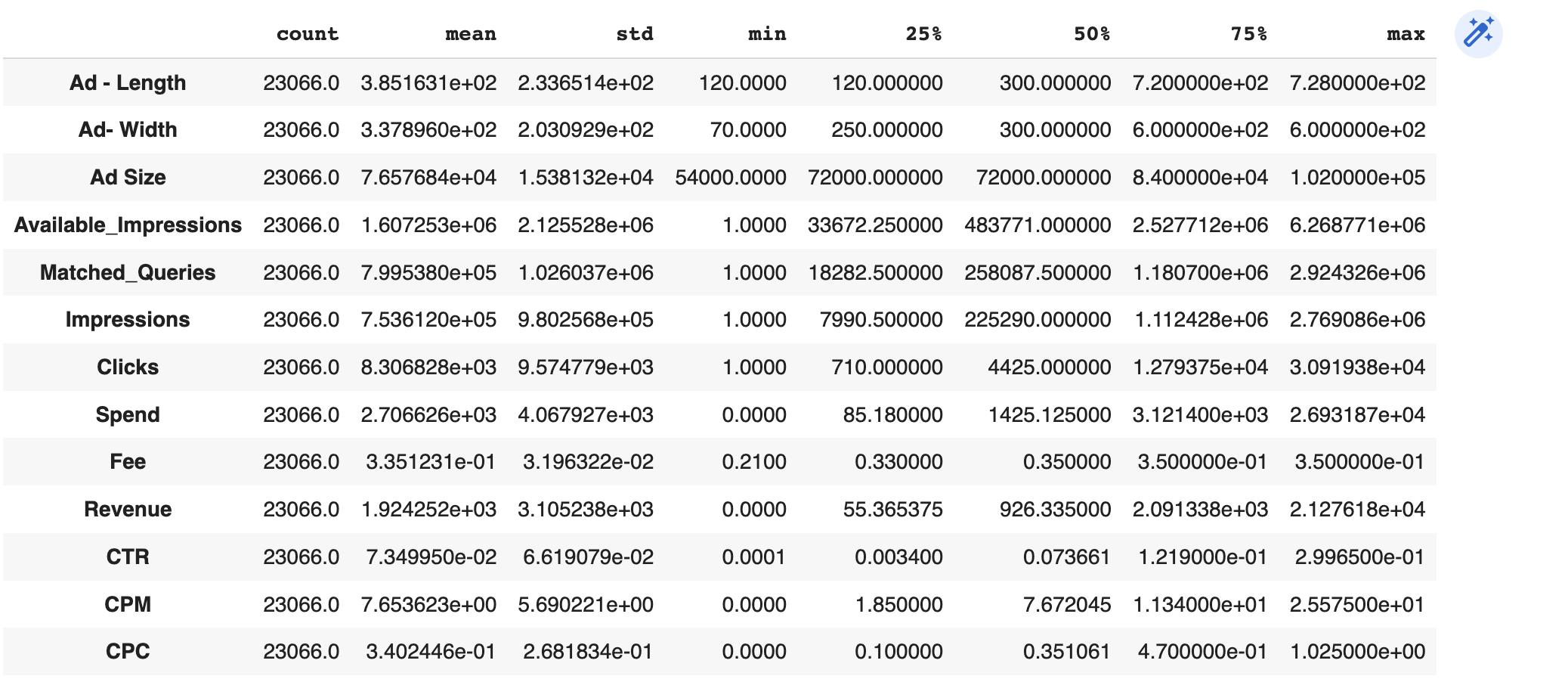


**Question5:**

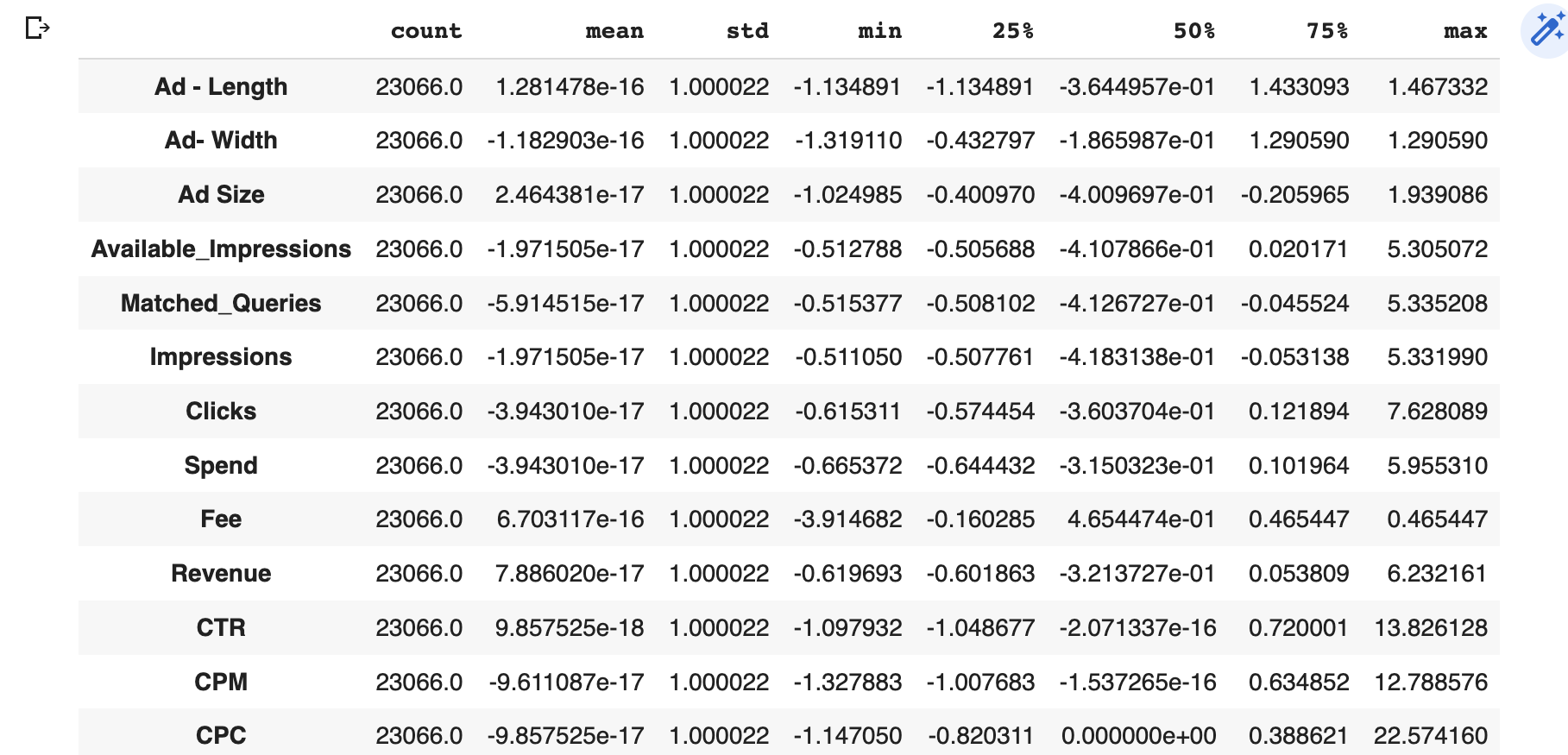
**Perform z-score scaling and discuss how it affects the speed of the algorithm.**

*Have performed standard scaling instead and below the snap shot of the data before and after.*

**Data before performing scaling:**



**New data after Performing scaling:**

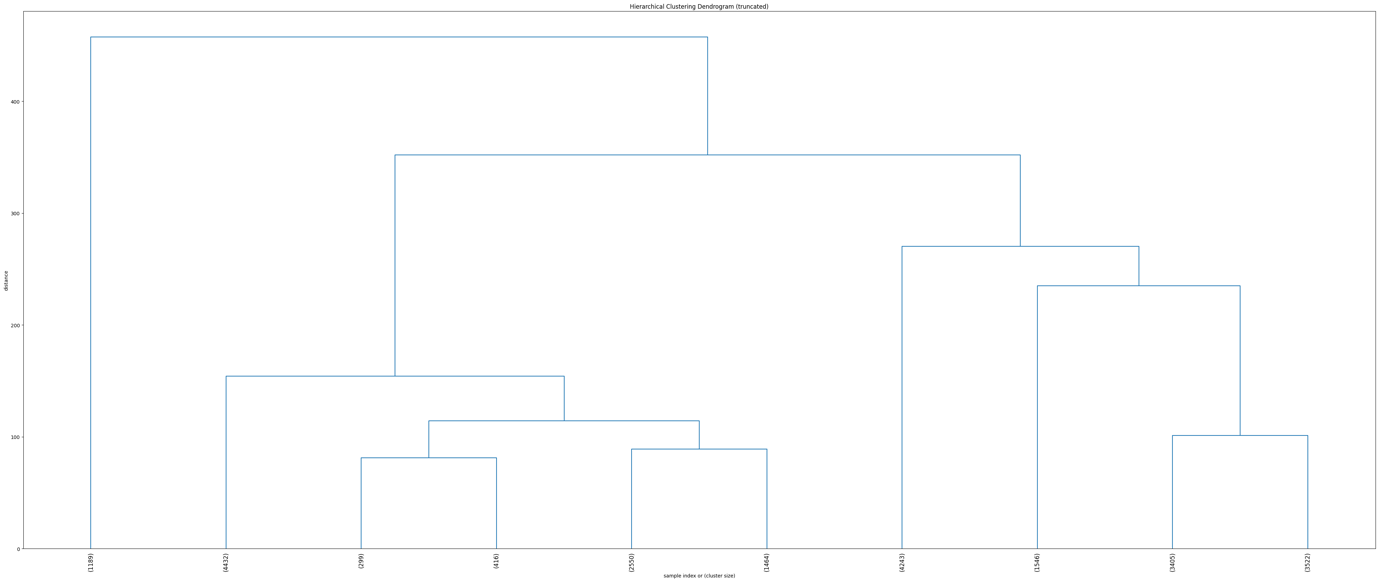


**Question6:**

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

**Performing clustering and the following actions are been taken:**

Please find below Dendrogram performed for Hierarchical using WARD and Euclidean Distance on the Scaled Data such as “data1\_scaled”.In this Dendrogram, value of P = 10, which means that only the last 10 merged clusters are shown



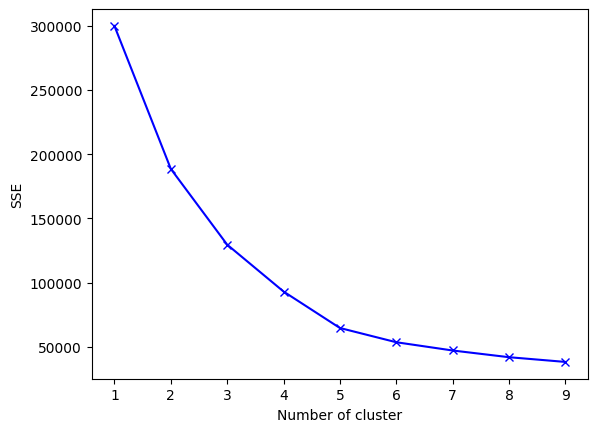
**Question7:**

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

For checking the Optimal number of clusters we use ESS (Sum Of Square of error)

As per the check.

Between K1 & K2, there is a significant drop. Also, moving k2 to k5 there is a significant drop as well. As It reduces after k5, hence the K 5 would be the optimal number of clusters

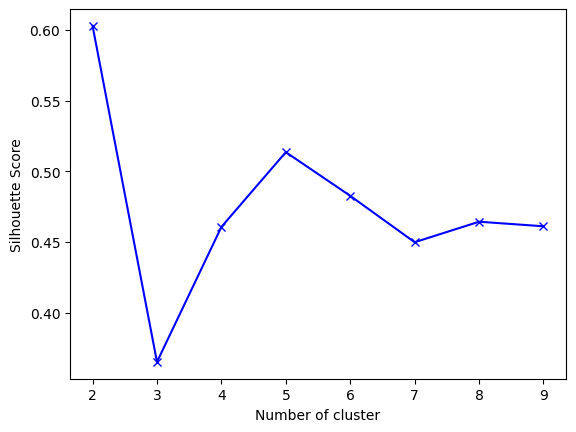


**Question8:**

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

The Silhouette Score is a measure of how similar an object is to its own cluster compared to other clusters, and it ranges from -1 to 1, with higher values indicating better clustering.

As per both the observations it is safe to conclude that the **optimal number of clusters should be 5** as 2 would be very less to consider



For n\_clusters=2, The Silhouette Coefficient is 0.602856419557812

For n\_clusters=3, The Silhouette Coefficient is 0.3652575679239419

For n\_clusters=4, The Silhouette Coefficient is 0.46072044314349486

For n\_clusters=5, The Silhouette Coefficient is 0.5135883146481809

For n\_clusters=6, The Silhouette Coefficient is 0.4826959081616031

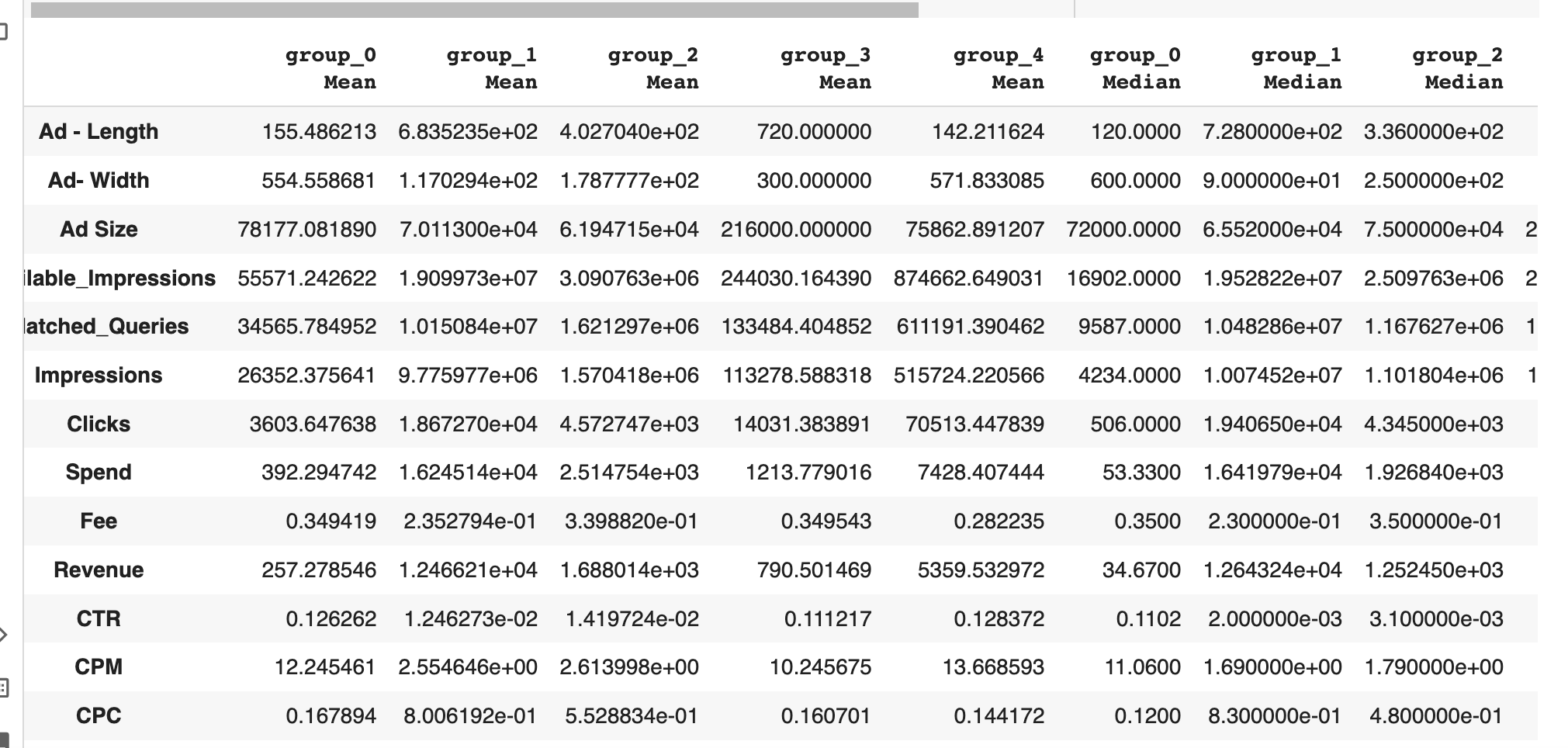
For n\_clusters=7, The Silhouette Coefficient is 0.44949580299210745

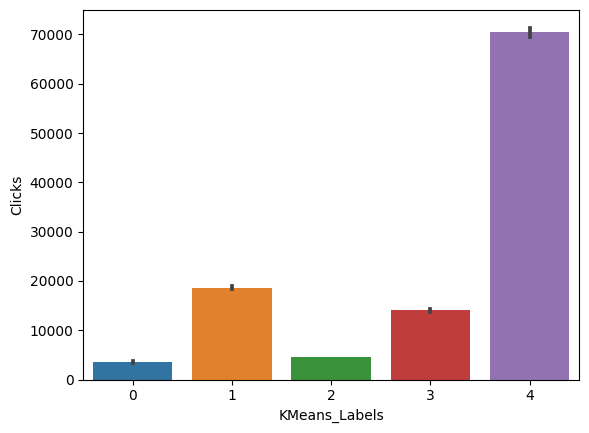
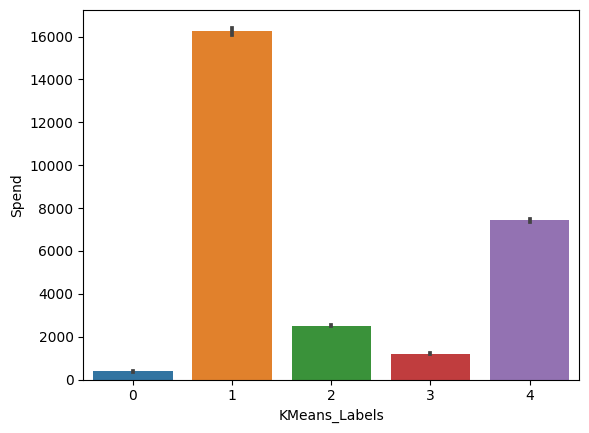
For n\_clusters=8, The Silhouette Coefficient is 0.46442872466744733

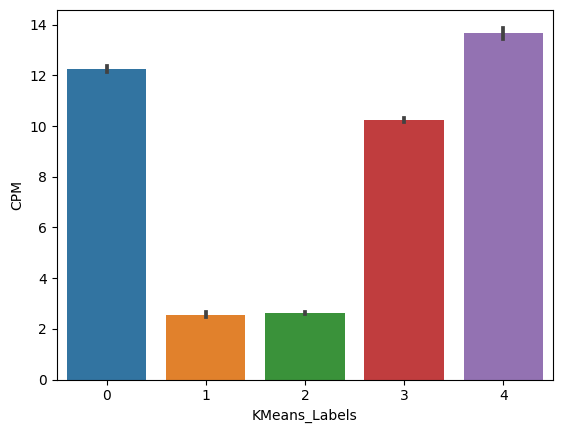
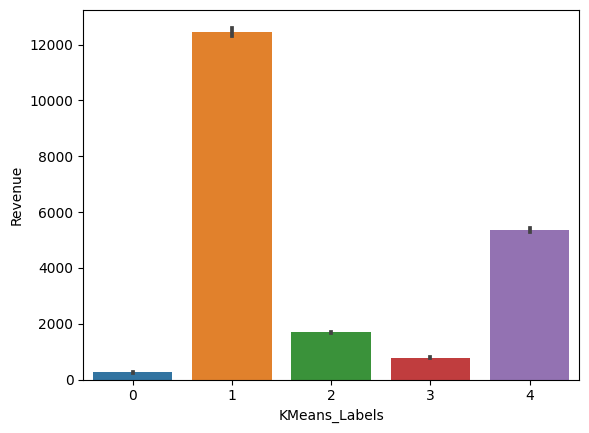
For n\_clusters=9, The Silhouette Coefficient is 0.46194271486474986

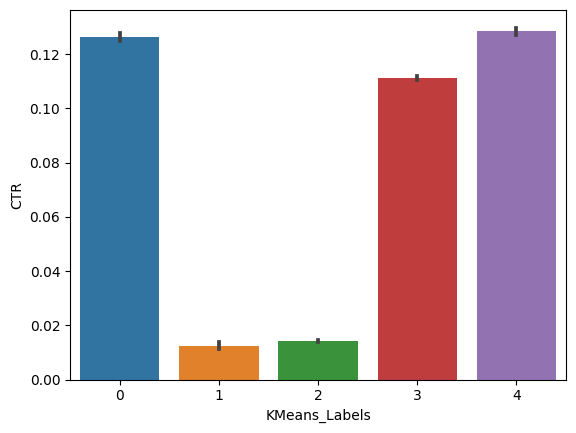
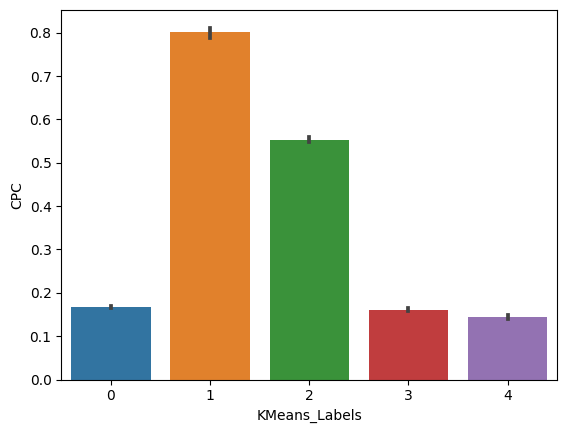
**Question9:**

**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].**

****

****

****

** **

* I have performed KMeans Clustering on scaled data, and then added the predicted cluster
* labels to two different data sets: scaled\_data\_copy and df.
* The KMeans function from scikit-learn is used to create a KMeans object with n\_clusters=5
* (i.e., 5 clusters).
* Created clusters for the Ads based on optimum number of clusters using silhouette score.

* The groupby method from Pandas is used to group the data by the KMeans cluster labels and the mean and median methods are used to compute the mean and median values of each feature within each cluster. The resulting data frames are stored in the variables mean and median.
* Then, the pd.concat method from Pandas is used to concatenate the mean and median data frames along the vertical axis (axis=0) to create a single data frame df\_kmeans. The index attribute of the resulting data frame is set to a list of strings representing the names of each group followed by either 'Mean' or 'Median'.

**Question10:**

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

* When Click on Ads gets increases then Revenue is also increases.
* When amount of money spent on specific ad variations within a specific campaign or ad set.
* When impression count of the particular Advertisement increases then Revenue is also increases

PCA:

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.

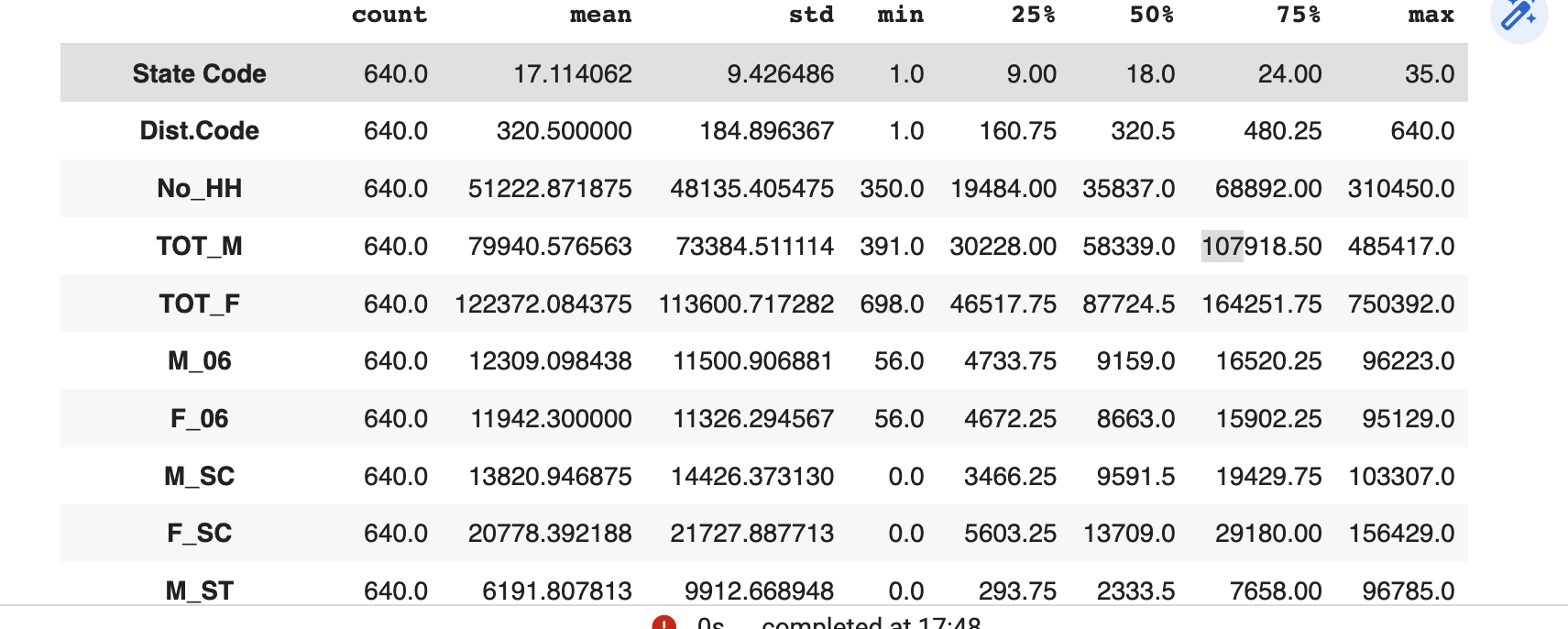
Note: The 24 variables given in the Rubric is just for performing EDA.

You will have to consider the entire dataset, including all the variables for performing PCA.

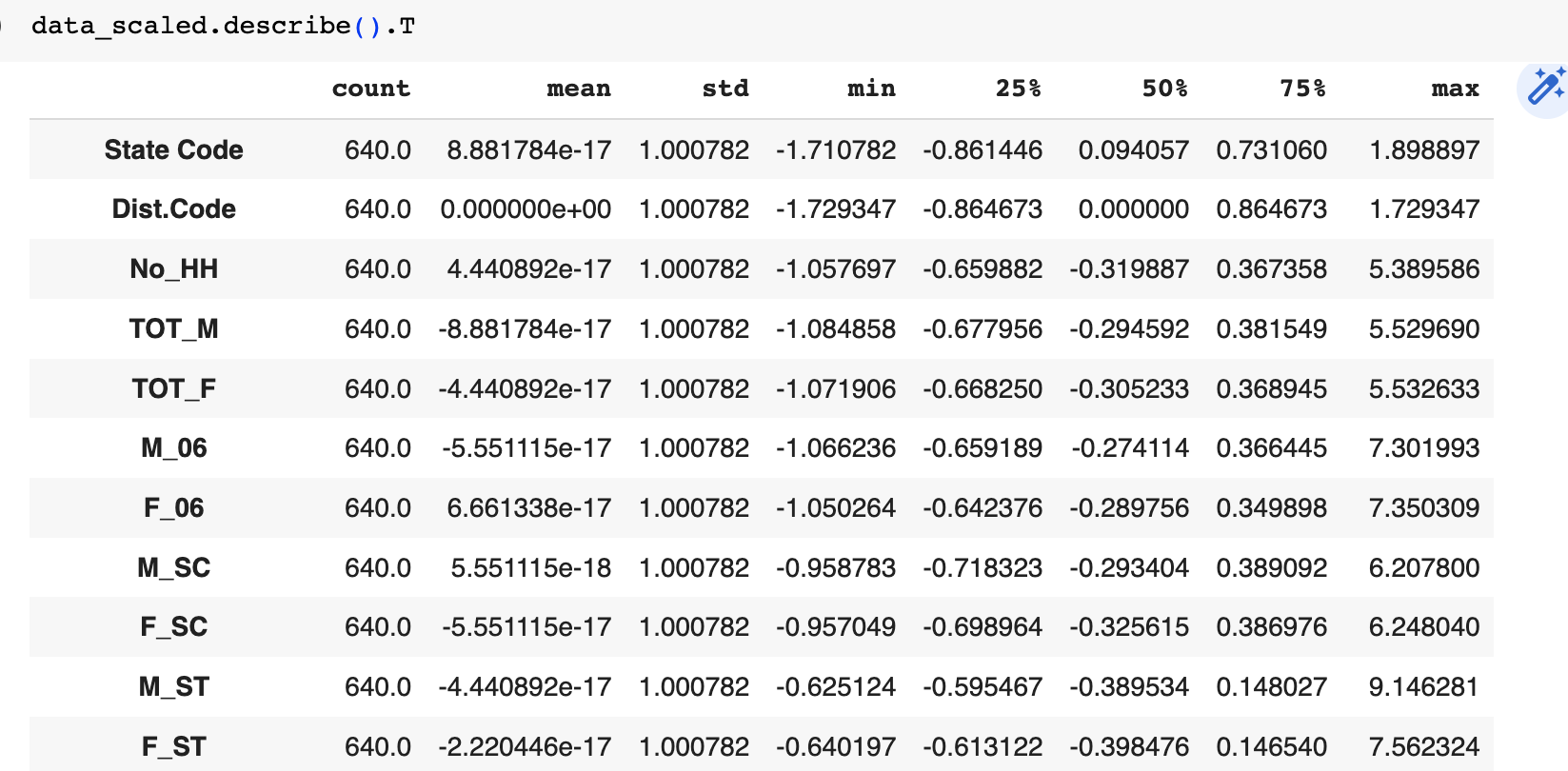
**Solution:**

*Observation:*

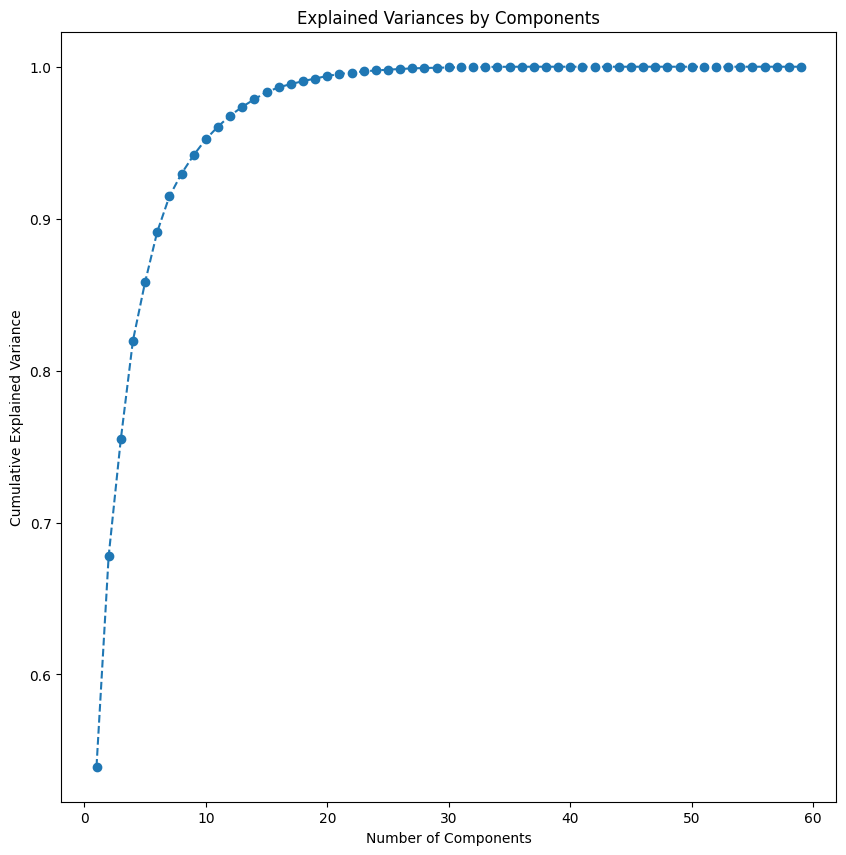
* There are 640 observations and 61 columns in the data.
* All variables State and area name are of numeric data type.
* New dataset is made and dropped categorical columns from it
* NO duplicate data found
* 5 points summery of the data



* Except State Code and District code all variables are having outliers and it is skewed towards right.
* Outliers are present in to the data.
* However After analysing the data it will be appropriate to continue with the current data and do the scaling instead .
* After the scaling the data



**Visualization of the explained variance by individual components**



* Finding the least number of components that can explain more than 90% variance
* Found 7 PCs after performing PCA.
* We can see that out of the 60 original features, we reduced the number of features through principal components to 7, these components explain more than 90% of the original variance.
* So that is about 90% reduction in the dimensionality with a loss of 10% in variance.
* Let us now look at these principal components as a linear combination of original features.