**BUSINESS DATA**

**ANALYSIS REPORT**

Table of Contents

[Problem1: Clustering 5](#_Toc133780519)

[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. 5](#_Toc133780520)

[2.Treat missing values in CPC, CTR and CPM using the formula given. You have to basically create a user defined function and then call the function for imputing. 9](#_Toc133780521)

[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). 11](#_Toc133780522)

[4.Perform z-score scaling and discuss how it affects the speed of the algorithm. 12](#_Toc133780523)

[5.Perform clustering and do the following: 12](#_Toc133780524)

[Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance. 12](#_Toc133780525)

[6.Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm. 14](#_Toc133780526)

[7.Print silhouette scores for up to 10 clusters and identify optimum number of clusters. 14](#_Toc133780527)

[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.] 15](#_Toc133780528)

[9. Clustering: Conclude the project by providing summary of your learnings. 18](#_Toc133780529)

[Problem2 :PCA 19](#_Toc133780530)

[1.Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc. 20](#_Toc133780531)

[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 23](#_Toc133780532)

[3.We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary? 30](#_Toc133780533)

[4.Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment. 30](#_Toc133780534)

[5.Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector. 32](#_Toc133780535)

[6.Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot. 35](#_Toc133780536)

[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. 37](#_Toc133780537)

[8.PCA: Write linear equation for first PC. 42](#_Toc133780538)

**List of Figures**

[Figure 1: Clustering Clean Ads Dataset 5](#_Toc133786041)

[Figure 2:Top 5 rows in the dataset 6](#_Toc133786042)

[Figure 3:Bottom 5 rows in the dataset 6](#_Toc133786043)

[Figure 4:Count plot of Platform, Device Type and Format 7](#_Toc133786044)

[Figure 5: Count plot of Inventory and Ad type 7](#_Toc133786045)

[Figure 6: Plotting Boxplot of Numerical Variables 8](#_Toc133786046)

[Figure 7:Dataset after imputing the Null values for CTR,CPM and CPC 10](#_Toc133786047)

[Figure 8:Boxplot of Numerical Variables 11](#_Toc133786048)

[Figure 9: Boxplot after treating Outliers 12](#_Toc133786049)

[Figure 10: Z score scaling of the Numerical Variables 13](#_Toc133786050)

[Figure 11: Dendrogram using metric='Euclidean' and method='ward’ with all the Clusters 14](#_Toc133786051)

[Figure 12:Dendrogram using method='ward’ with all the Clusters 14](#_Toc133786052)

[Figure 13: Dendrogram using metric='Euclidean' and method='ward’ last 10 Clusters 14](#_Toc133786053)

[Figure 14: WSS Scores 15](#_Toc133786054)

[Figure 15:Elbow Plot of WSS 15](#_Toc133786055)

[Figure 16:Silhouette Scores. 16](#_Toc133786056)

[Figure 17 Silhouette Plot. 16](#_Toc133786057)

[Figure 18: Dataset with Clusters 17](#_Toc133786058)

[Figure 19 :Graphical Representation of Cluster wise trends in clicks, spend, revenue, CPM, CTR, & CPC based on Device Type 18](#_Toc133786059)

[Figure 20: Countplot of Device Types and Clusters 18](#_Toc133786060)

[Figure 21:Top 5 observations of PCA India Data Census 20](#_Toc133786061)

[Figure 22: Bottom 5 observations of PCA India Data Census 20](#_Toc133786062)

[Figure 23: Distribution Plot and Boxplot of Number of Literate Males 24](#_Toc133786063)

[Figure 24: Distribution Plot and Boxplot of Total Number of Males 24](#_Toc133786064)

[Figure 25: Distribution Plot and Boxplot of Total Number of Females 24](#_Toc133786065)

[Figure 26: Distribution Plot and Boxplot of Total Number of Male Workers 24](#_Toc133786066)

[Figure 27: Distribution Plot and Boxplot of Total Number of Female Workers 24](#_Toc133786067)

[Figure 28: Bar plot of Literate males in each state 25](#_Toc133786068)

[Figure 29: bagplot of Total Males in each state 26](#_Toc133786069)

[Figure 30: bar plot of Total Females in each state 26](#_Toc133786070)

[Figure 31: bar plot of Total Working Males in each state 26](#_Toc133786071)

[Figure 32: bar plot of Total Working Females in each state 27](#_Toc133786072)

[Figure 33:Scatterplot of Numerical Variables 27](#_Toc133786073)

[Figure 34: bar plot of Gender Ratio is Each State 28](#_Toc133786074)

[Figure 35:District wise Ratio of Gender 28](#_Toc133786075)

[Figure 36:State wise Gender Ratio of Male to Female Working Population 28](#_Toc133786076)

[Figure 37: State wise ratio of literate Male to Total Male Population 29](#_Toc133786077)

[Figure 38: State wise ratio of Working Male to Total Male Population 29](#_Toc133786078)

[Figure 39: : State wise ratio of Working Female to Total Female Population 30](#_Toc133786079)

[Figure 40 :Boxplot of Variables before scaling 31](#_Toc133786080)

[Figure 41: Boxplot of Variables after scaling 31](#_Toc133786081)

[Figure 42: Scaled Data 32](#_Toc133786082)

[Figure 43:Correlation heatmap between 57 variables 34](#_Toc133786083)

[Figure 44:Scree Plot with 57 Principal Components 35](#_Toc133786084)

[Figure 45 PCA Variance with 10 components 36](#_Toc133786085)

[Figure 46 PCA Variance Ratio with 10 components 36](#_Toc133786086)

[Figure 47:Cumulative Variance of the Components 36](#_Toc133786087)

[Figure 48: Scree Plot with 10 Principal Components 36](#_Toc133786088)

[Figure 49: PCA Loading of PC1 Figure 50: PCA Loading of PC2 39](#_Toc133786089)

[Figure 51: PCA Loading of PC3 Figure 52: PCA Loading of PC4 39](#_Toc133786090)

[Figure 53: PCA Loading of PC5 and PC6 39](#_Toc133786091)

[Figure 54:Correlation map of Principal Components and Actual Columns 40](#_Toc133786092)

[Figure 55:Heatmap of PC’s 42](#_Toc133786093)

**List of Tables**

[Table 1: Clean Ads Dataset Columns Information and Data Types 6](#_Toc133786019)

[Table 2:Data Summary of Categorical Variables 7](#_Toc133786020)

[Table 3: Data Summary of Numerical Variables 8](#_Toc133786021)

[Table 4: Count of Null Values 9](#_Toc133786022)

[Table 5: Count of Null Values in CTR, CPC and CPM 9](#_Toc133786023)

[Table 6: Data Summary Before and After Imputing Null Values 10](#_Toc133786024)

[Table 7:Data Summary after Treating Outliers 12](#_Toc133786025)

[Table 8: Data Summary after scaling using Z score method 13](#_Toc133786026)

[Table 9: Grouping data by clusters with mean to identify trends in clicks, spend, revenue, CPM, CTR, & CPC based on Device Type 17](#_Toc133786027)

[Table 10: Dataset Column Information1 21](#_Toc133786028)

[Table 11: Dataset Column Information2 21](#_Toc133786029)

[Table 12: Describing the Numerical Columns 22](#_Toc133786030)

[Table 13: : Describing the Categorical Columns 22](#_Toc133786031)

[Table 14: Null Values in the Dataset 23](#_Toc133786032)

[Table 15: Data Description of the 5 variables 25](#_Toc133786033)

[Table 16Snapshot of the Covariance Matrix of all the numerical Variables 33](#_Toc133786034)

[Table 17:Eigen Vectors 34](#_Toc133786035)

[Table 18:Eigen Values 35](#_Toc133786036)

[Table 19 Variances and Standard Deviations of each PC 37](#_Toc133786037)

[Table 20 PC scores 41](#_Toc133786038)

[Table 21:Correlation between PC’s 41](#_Toc133786039)

[Table 22: linear equation for first PC. 43](#_Toc133786040)

# Problem1: **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**](https://olympus.mygreatlearning.com/courses/88920/files/8022141/download?verifier=hhhxM0VRf4Yt8nx0EOpHSQIM7Q2rs6YyOMawRMZp&wrap=1)**Excel File.**

Perform the following in given order:

## 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: Load the sample data for number of observations and columns

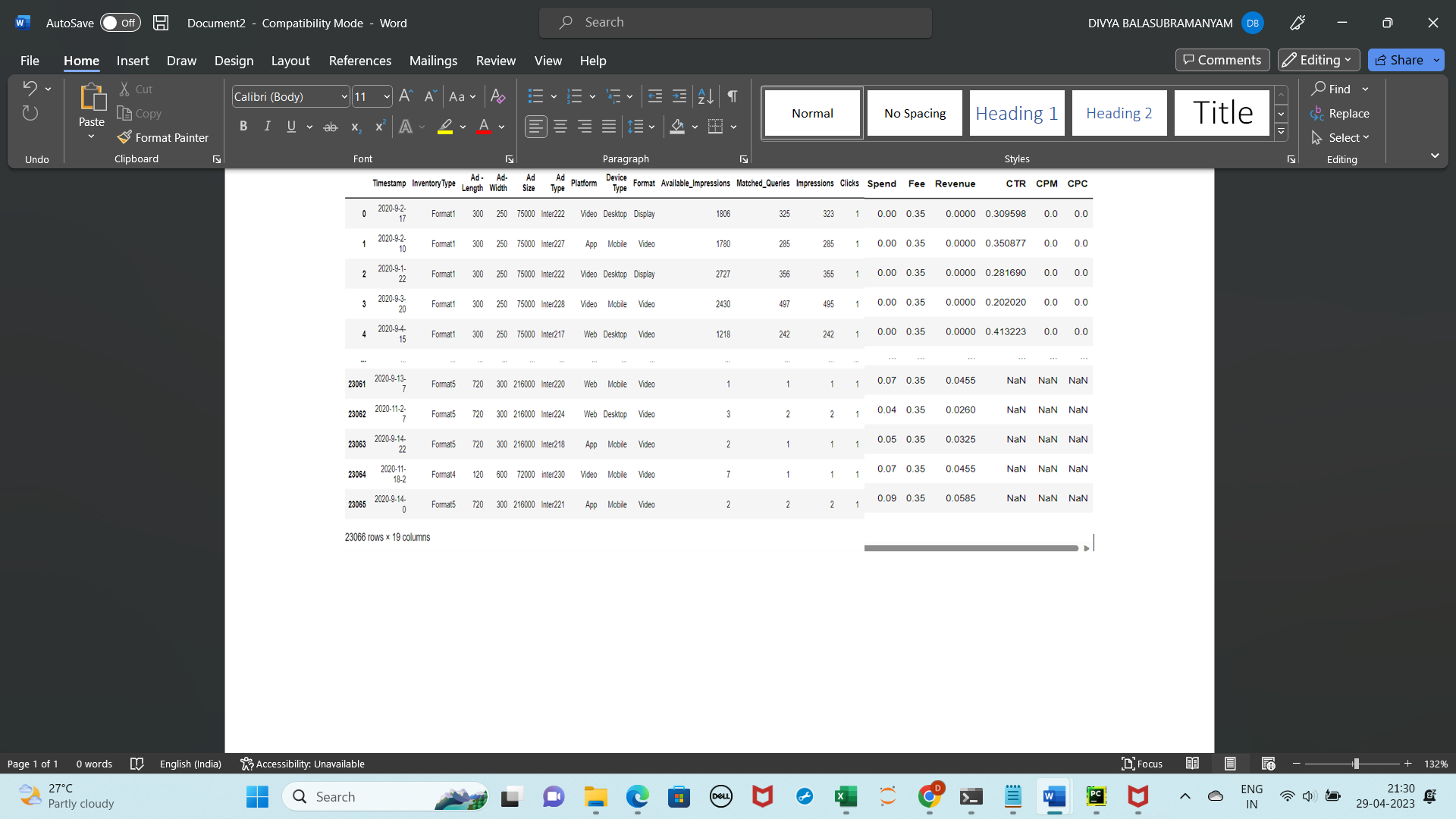


Figure 1: Clustering Clean Ads Dataset

As can be seen there are 23066 rows and 19 columns

Note: In the Original dataset , the Column CTR was expressed as % and not as per formula for CTR column. Hence Column in the excel has been modified as per the CTR formula mentioned before loading the dataset.

Let’s see the top 5 and bottom 5 records



Figure 2:Top 5 rows in the dataset

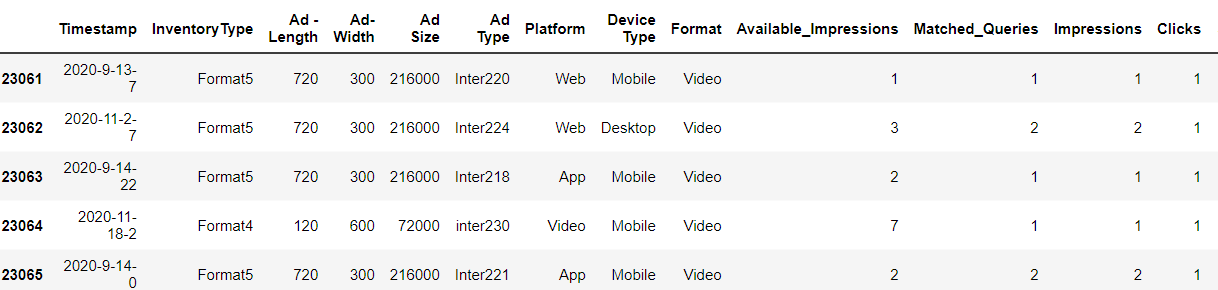


Figure 3:Bottom 5 rows in the dataset

Let’s see the columns description and their Data Types

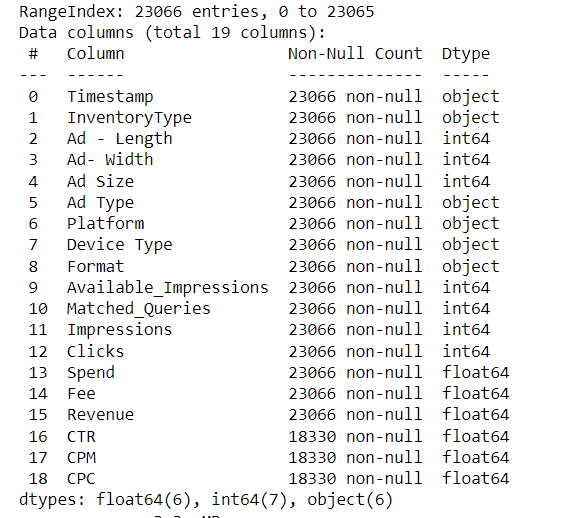


Table 1: Clean Ads Dataset Columns Information and Data Types

There are 13 Numeric type Columns of which are 6 are Float Type and 7 are Int type and 6 object type columns. Time Stamp column is marked as object type column though its Date Type Column, but timestamp column is not of any relevance to the clustering ,so we will not change the column Type.

**Let’s see the data summary of the Categorical variables**

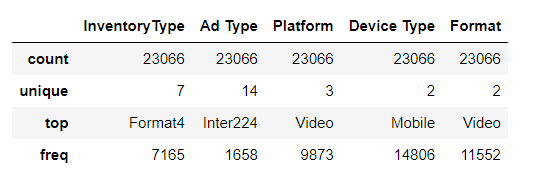


Table 2:Data Summary of Categorical Variables

Let’s see some graphical analysis on categorical variables

**Univariate Analysis of Categorical Variables**

Let’s see the graphical representation of categorical variables

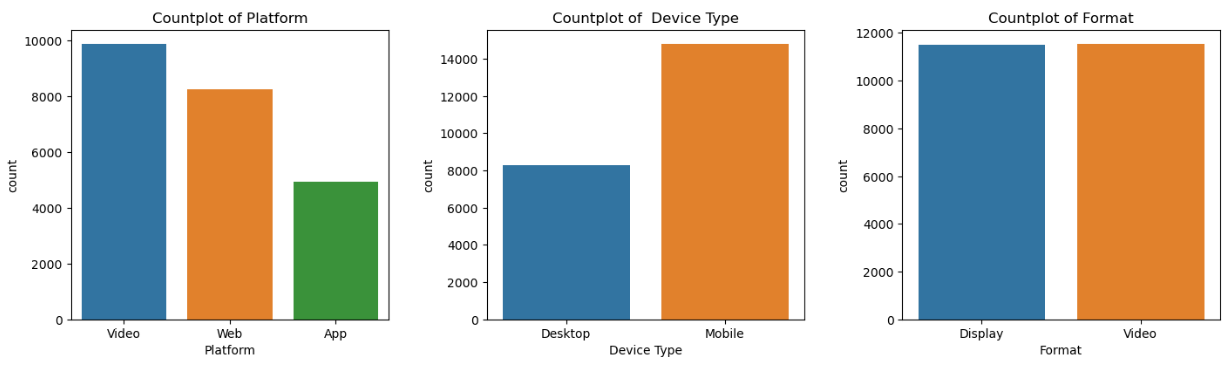


Figure 4:Count plot of Platform, Device Type and Format

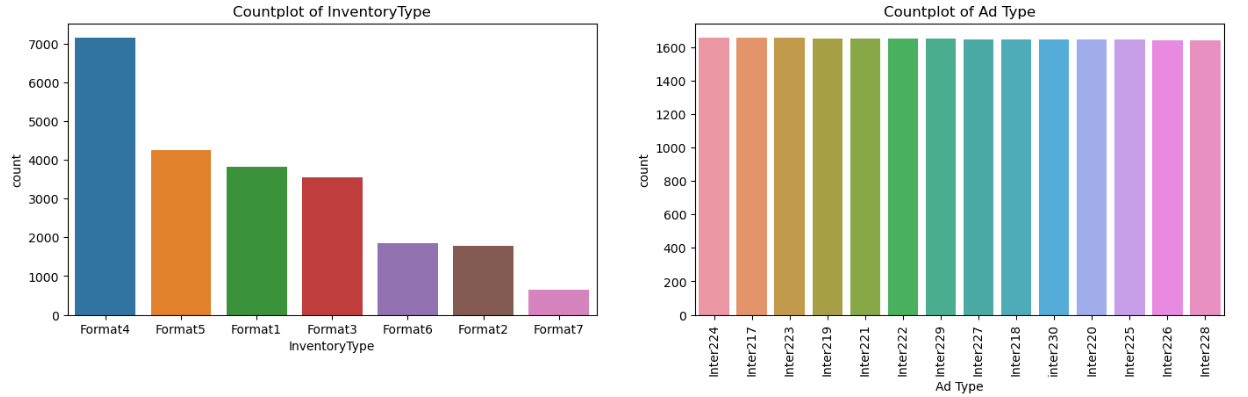


Figure 5: Count plot of Inventory and Ad type

**Inferences from the above graphical representation**:

1.There are 3 types of Platforms namely Video, Web and App. Video is the most used form of Platform for Ads and App is least used form for Ads.

2.There are 2 device types that support these Ads ,they are Mobile and Desktop. Mobile is most used Device for Ads .

3.Display and Video are 2 formats in which Ads are displayed . Video format of Ads are slightly more than Display.

4.There are 7 types of Inventories namely Format1,Format2,Format3,Format4,Format5,Format6,Format7.Format 4 is most used Inventory Type for Ads and Format 7 is least used .

5.There are 14 Ad types and Inter224 is the most used Ad Type.

**Let’s see the data summary of the numerical variables:**

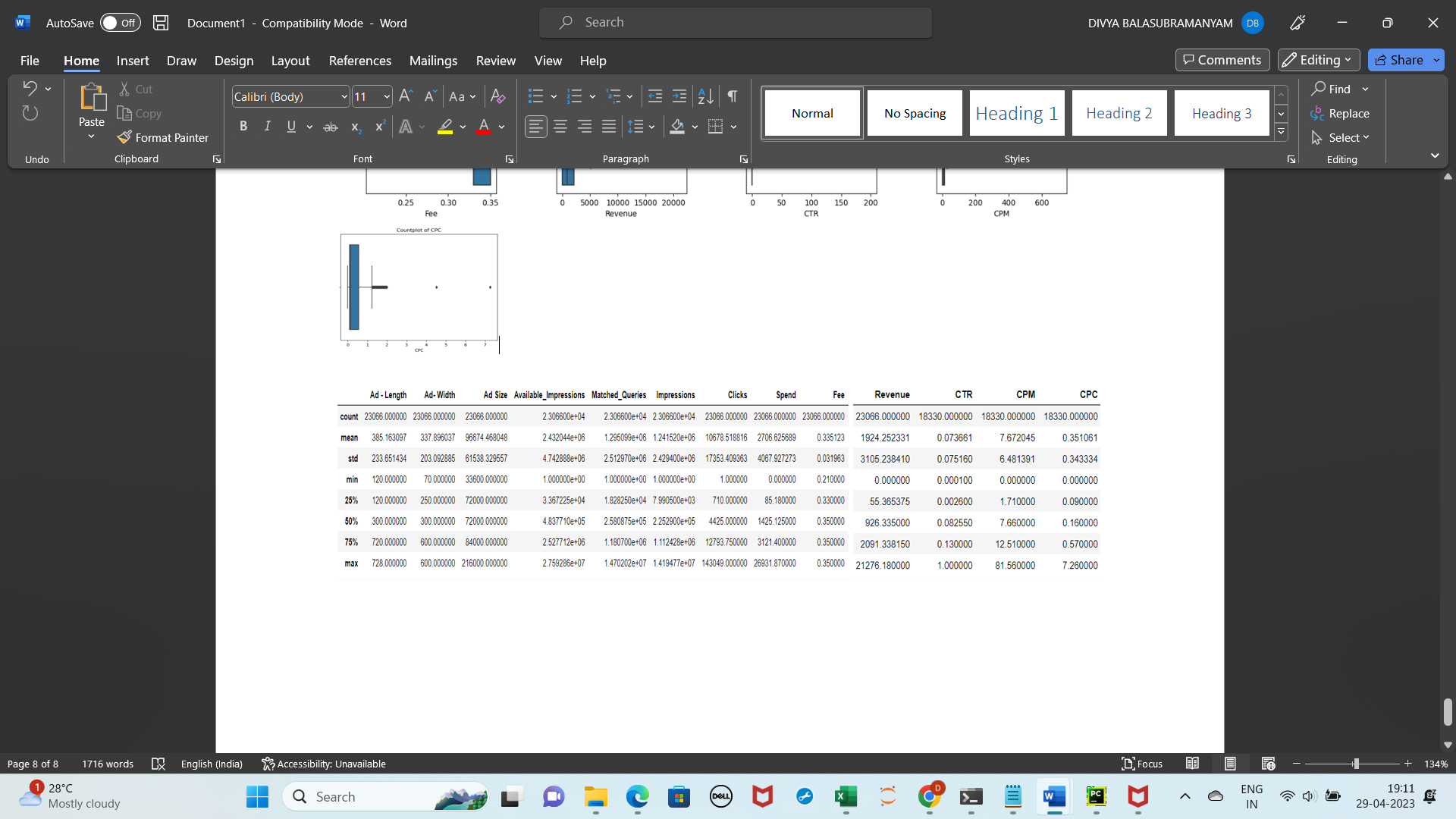


Table 3: Data Summary of Numerical Variables

**Univariate Analysis of Numerical Variables**

Let’s see the graphical representation of numerical variables

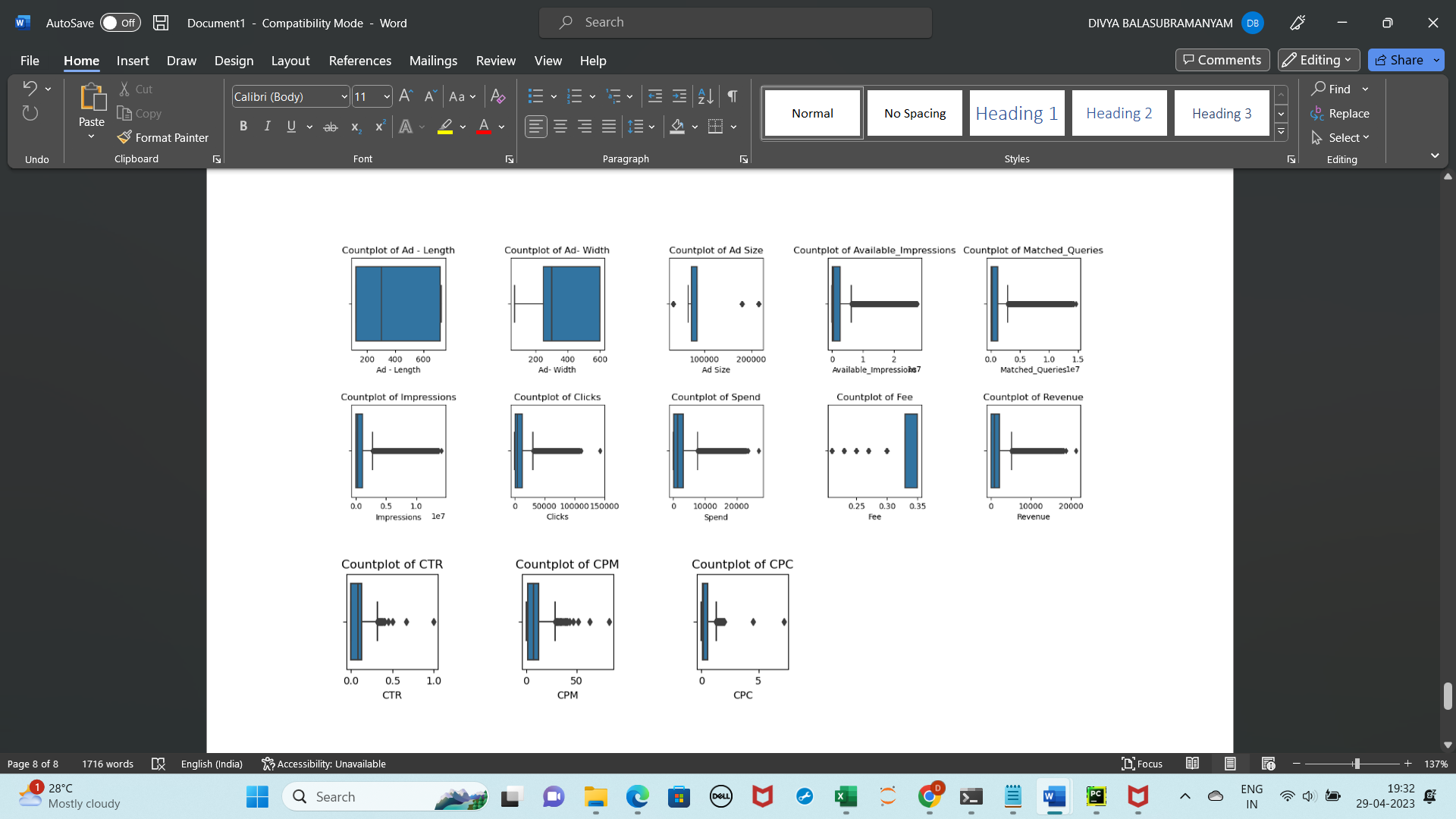


Figure 6: Plotting Boxplot of Numerical Variables

Inferences from the above graphical representation:

1.Except for Ad-width and Ad-Length, all other numerical variables have outliers .

2.Most of the variables are right skewed .Fee Variable is left skewed.

3.Range of values of each of the variables is significantly different, we need to scale them so that all features are given equal weight.

4.There are ads with 0 Revenue and 0 Spend.

Now let’s check for anomalies in the data

Let check for Null values in the columns

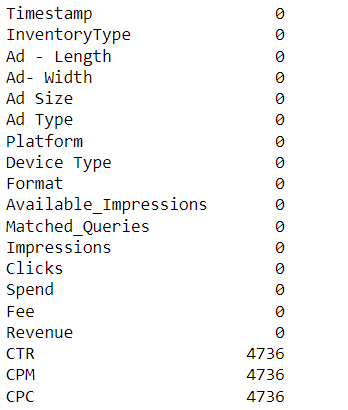
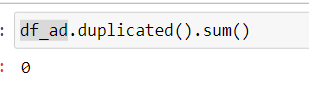


Table 4: Count of Null Values

As can be seen, columns CTR, CPM and CPC have 4736 Null values each.

There are no duplicate records in the Dataset.



## 2.Treat missing values in CPC, CTR and CPM using the formula given. You have to basically create a user defined function and then call the function for imputing.

**Using the formula for** CPC, CTR and CPM given below , we can impute the missing values

**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.**

We create a user defined Lambda function that will compute CPC, CTR and CPM and then we will apply that function on the columns CPC, CTR and CPM and thus impute the Null values.

There are some anomalies in the original dataset . There are some CPM values that are non -zero for cases where spend is 0 and values for CTR in the original dataset are expressed as percentage, we will correct those values as per the formula .

After Imputing let’s check the count of Null values in CPC, CTR and CPM .

As we can see there are no null values



Table 5: Count of Null Values in CTR, CPC and CPM

Let’s again load dataset and see if the missing values are imputed

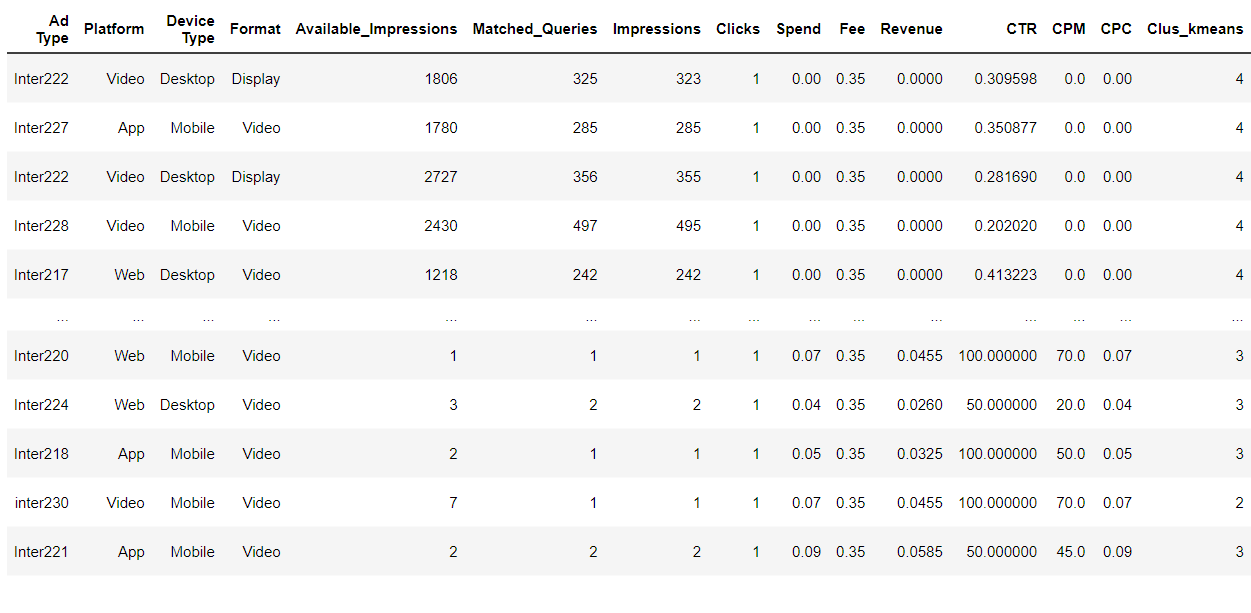


Figure 7:Dataset after imputing the Null values for CTR,CPM and CPC

As we can see now the null values are imputed as per the formulas . The Data Summary of the three columns after we have imputed the Null Values looks very different from the original data

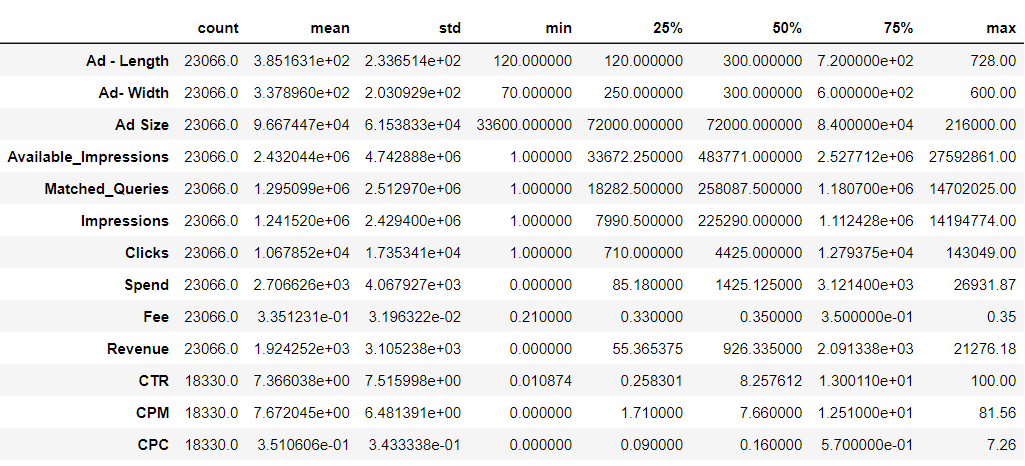
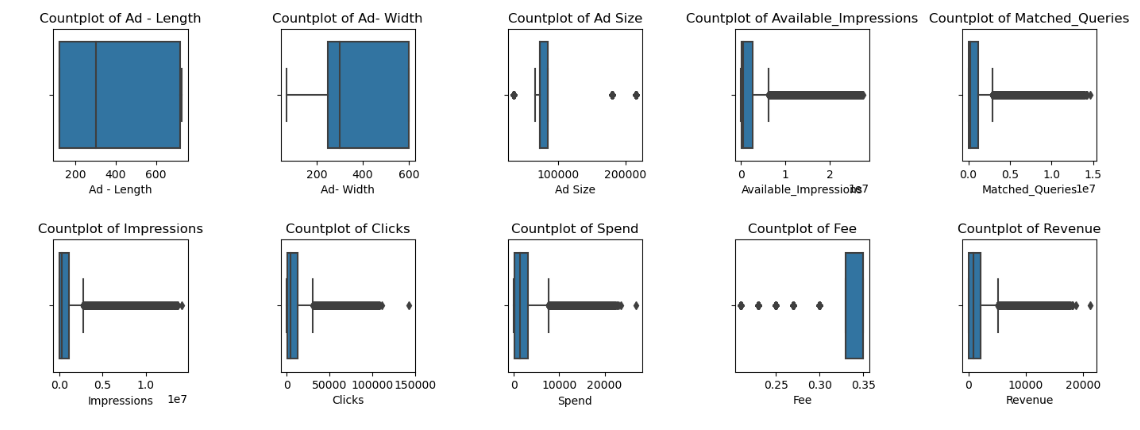
 

Table 6: Data Summary Before and After Imputing Null Values

## 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).



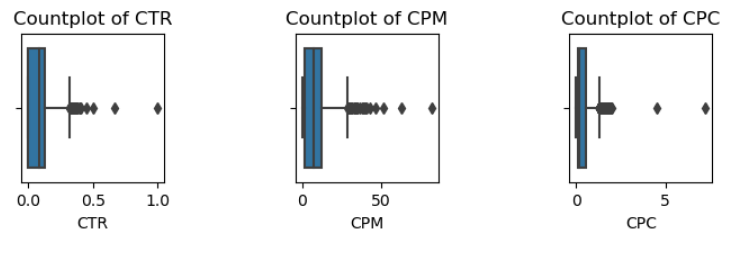


Figure 8:Boxplot of Numerical Variables

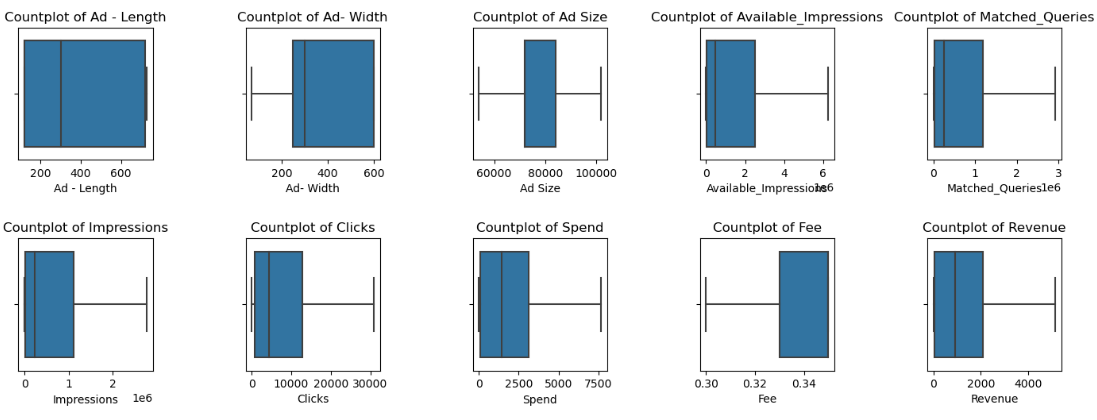
As can be seen from the boxplot above, there are outliers in many of the variables

After applying IQR method we see that there are outliers in some of the variables.

 Outlier increases the mean of data .Since K-Means algorithm is about finding mean of clusters, the algorithm is influenced by outliers. The k-means algorithm updates the cluster centres by taking the average of all the data points that are closer to each cluster centre. When all the points are packed nicely together, the average makes sense. However, when you have outliers, this can affect the average calculation of the whole cluster. As a result, this will push cluster centre closer to the outlier. The benefit of removing outliers is to enhance the accuracy and stability of statistical models by reducing their impact on results. Outliers can distort statistical analyses and skew results as they are extreme values that differ from the rest of the data. Removing outliers makes the results more robust and accurate by eliminating their influence.

To treat outlier, we use the Capping and Flooring method where in data points that are exceeding the upper whisker or Q3+1.5\*IQR ( Q3 is the 75th percentile and IQR is Interquartile range)is equated to upper whisker limit of the distribution and the values less than lower whisker or Q1 -1.5\*IQR(Q1 is the 25th percentile ) are all equated to lower whisker limit value of the distribution. Let’s apply this method and check if outliers are treated .

We take only the variables that need to be treated and create a user defined function to treat the outliers for each of those variables and apply that function of each of those variables



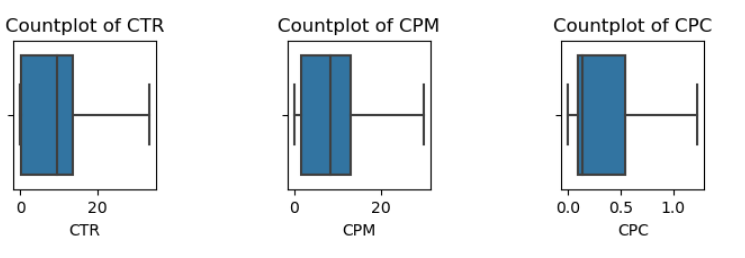


Figure 9: Boxplot after treating Outliers

After applying the function, we can see from the plots above , that the outliers have been treated and there are no outliers in the variables now .Lets see the data summary after treating outliers.The maximum and minimum values are now capped and floored

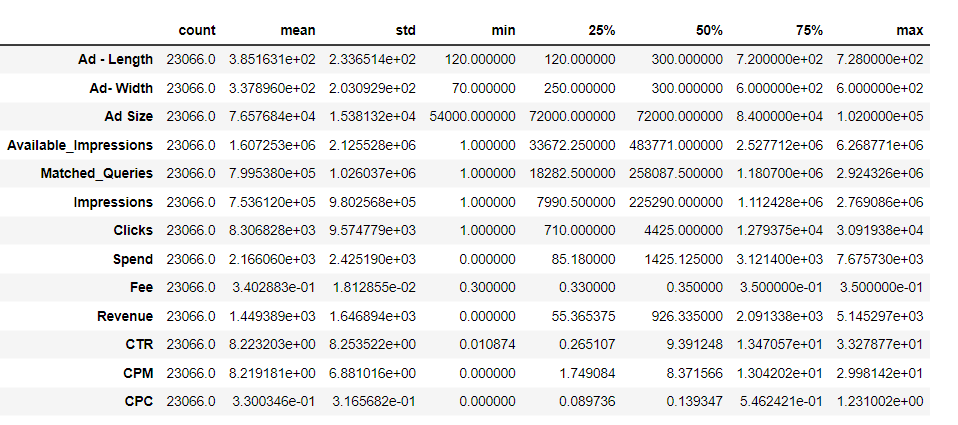


Table 7:Data Summary after Treating Outliers

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

In the machine learning algorithms if the values of the features are closer to each other there are chances for the algorithm to get trained well and faster instead of the data set where the data points or features values have high differences with each other .It will take more time to understand the data and the accuracy will be lower. So, if the data in any conditions has data points far from each other, scaling is a technique to make them closer to each other. Scaling is used for making data points generalized so that the distance between them will be lower. Scaling features restricts modules from being biased towards features having lower /higher magnitude.

Therefore its important we scale the features so that they all are on same scale which in turn helps model to assign equal importance to all features and make predictions without bias

Now let’s scale the numerical variables using Z score

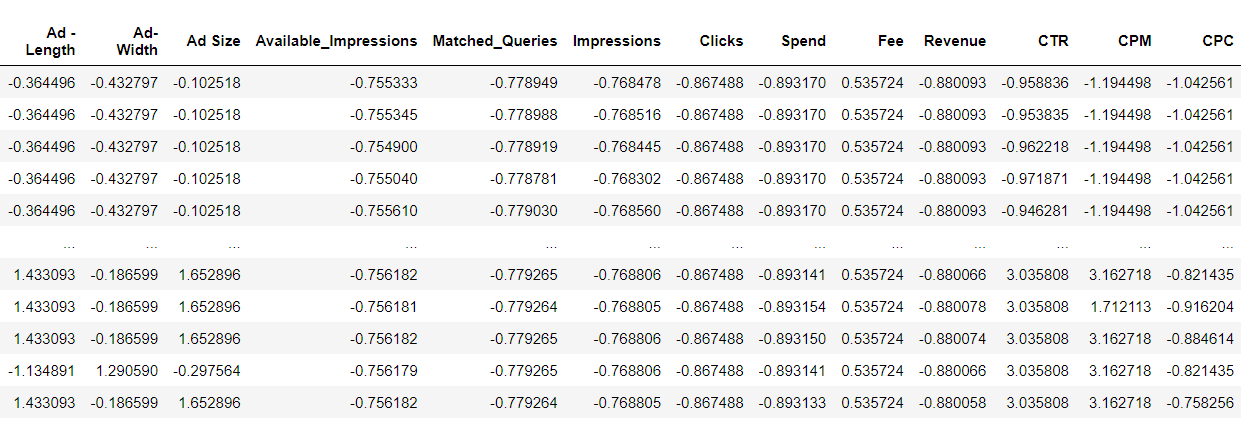


Figure 10: Z score scaling of the Numerical Variables

As we can observe now all the variables are now scaled with mean close to 0 and standard deviation close to 1



Table 8: Data Summary after scaling using Z score method

## 5.Perform clustering and do the following:

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

A dendrogram is a visual representation of cluster-making. On the x-axis are the cluster names or numbers and the y-axis is the distance or height. The vertical straight lines denote the height where two clusters combine. The higher the level of combining, the distant the individual items or clusters are. In hierarchical clustering, all items must combine to make one cluster. Each cluster is a representative of a different population. After constructing the dendrogram, we decide the level where the resultant tree needs to be cut. If the number of clusters is large, the cluster size is small and the clusters homogeneous. If the number of clusters is small, each contains more item sand hence clusters are more heterogeneous. Depending on the distance measure and linkage used, the number of clusters and their composition maybe different. Now let us proceed to the clustering using Euclidean distance and ward linkage method. Using scipy.cluster.hierarchy library we create dendrograms for ward and euclidean disatnce

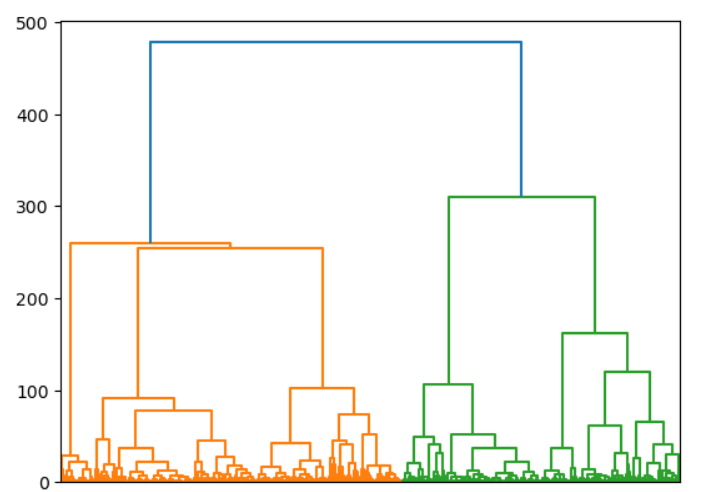


Figure 11: Dendrogram using metric='Euclidean' and method='ward’ with all the Clusters

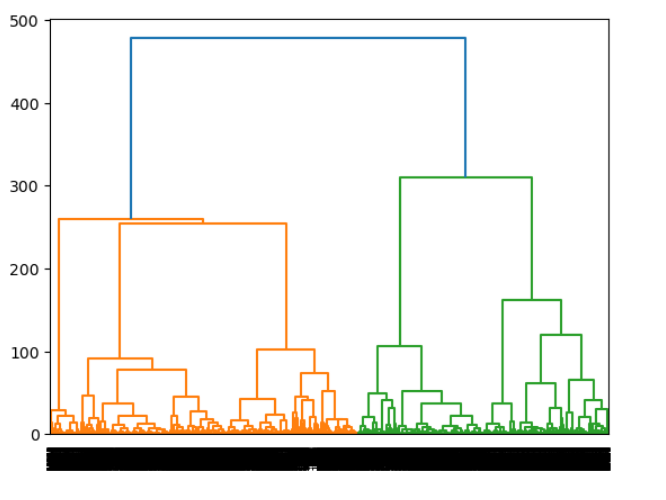


Figure 12:Dendrogram using method='ward’ with all the Clusters

Dendrogram of Figure 11 includes all clusters . Going by color\_threshold, there are 2 clusters. It is difficult for a business to understand segments from just 2 clusters. After multiple iterations, we can identify the suitable number of clusters.

Let’s view the Last 10 clusters on the dendrogram

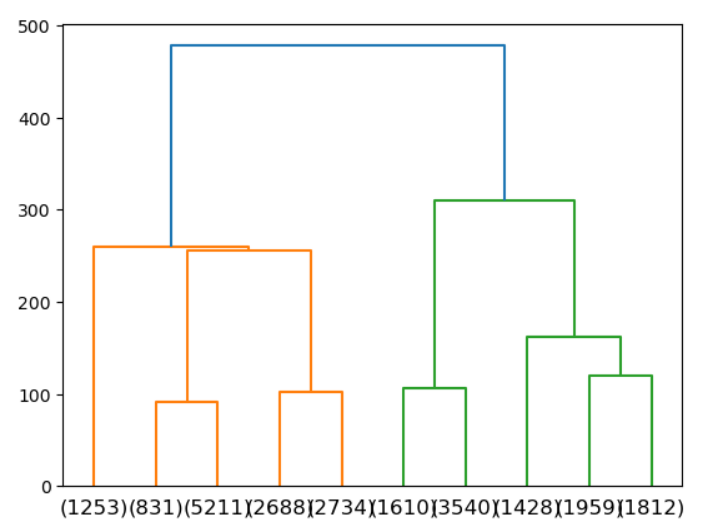


Figure 13: Dendrogram using metric='Euclidean' and method='ward’ last 10 Clusters

Figure 13 does not show the full-grown tree but just the last 10 clusters. Cluster size is not equal for all 10 clusters, as indicated in the above diagram(X-axis give the cluster size). In the dendrogram we locate the largest vertical difference between nodes, and in the middle pass a horizontal line. The number of vertical lines intersecting it is the optimal number of clusters (when affinity is calculated using the method set in linkage). It is important to note that based on the method selected for linkage and affinity, cluster membership and cluster size could vary.

As we see the number of vertical lines intersecting is 5, if we draw a horizontal line at the largest vertical distance , so let’s assume the cluster size as 5.We can further analyse using K-means clustering to conclude on the number of clusters .

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

k-means clustering is the most used non-hierarchical clustering technique. It aims to partition n observations into k clusters in which each observation belongs to the cluster whose mean (centroid) is nearest to it, serving as a prototype of the cluster. It minimizes within-cluster variances (squared Euclidean distances).For a given number of clusters, the total within-cluster sum of squares (WSS) is computed. That value of k is chosen to be optimum, where addition of one more cluster does not lower the value of total WSS appreciably. Following are WSS scores for each value of K starting from k=1 up to k=10.

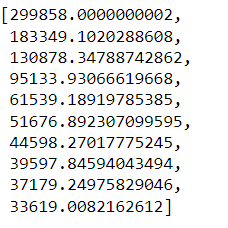


Figure 14: WSS Scores

The Elbow method looks at the total WSS as a function of the number of clusters. Let’s plot the WSS scores for each cluster

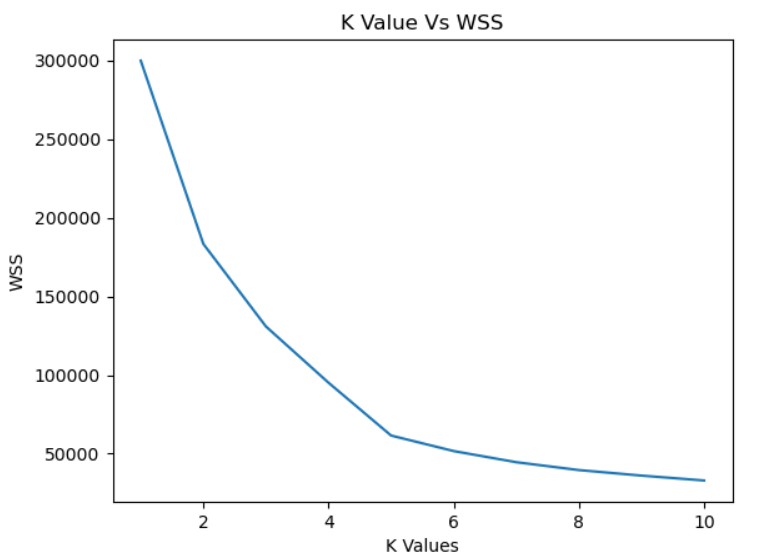


Figure 15:Elbow Plot of WSS

In this Elbow method, we are actually varying the number of clusters (K) from 1 – 10. For each value of K, we are calculating WCSS (Within-Cluster Sum of Square). WCSS is the sum of the squared distance between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot looks like an Elbow. As the number of clusters increases, the WCSS value will start to decrease. WCSS value is largest when K = 1. When we analyse the graph, we can see that the graph will rapidly change at a point and thus creating an elbow shape. From this point, the graph moves almost parallel to the X-axis. The K value corresponding to this point is the optimal value of K or an optimal number of clusters. As can be seen from the plot above , the point at which the elbow shape is created is 5; that is, our K value or an optimal number of clusters is 5.

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

Silhouette method measures how tightly the observations are clustered and the average distance between clusters. For each observation a silhouette score is constructed which is a function of the average distance between the point and all other points in the cluster to which it belongs, and the distance between the point and all other points in all other clusters, that it does not belong to. The maximum value of the statistic indicates the optimum value of k.

Following are the Silhouette scores up to 10 clusters

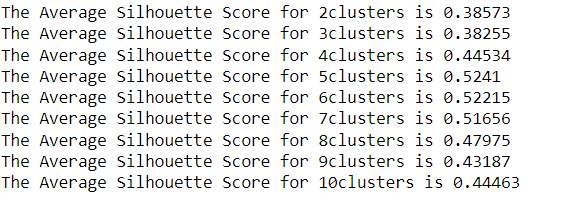


Figure 16:Silhouette Scores.

As can be seen from the Silhouette scores, the maximum value of the statistic is at k value 5 after which scores are decreasing , therefore we can say that optimum value of k is 5.

Let’s plot the Silhouette scores against each cluster and verify the same .

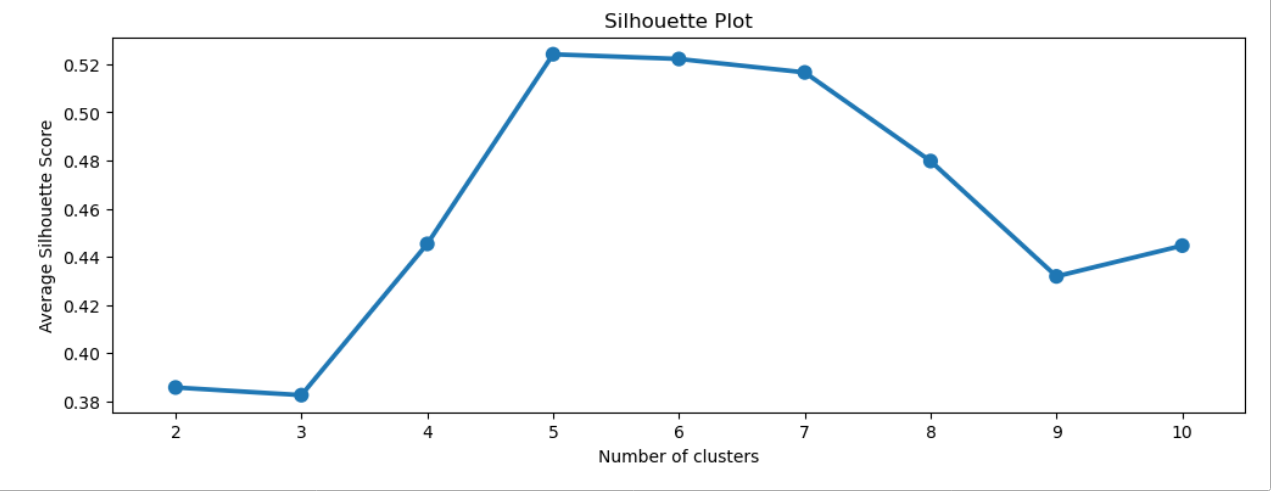


Figure 17 Silhouette Plot.

It is clear from Fig 17 that the maximum value of average silhouette score is achieved for k= 5, which, therefore, is considered to be the optimum number of clusters for this data.

## 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.]

We found that optimum number of clusters is 5 , so we can add a new column Clus\_Kmeans to the Original Dataset which can now be profiled into 5 different clusters .The Figure below depicts the dataset assigned with 5 different clusters .There are 5 clusters 1,2,3,4,5 and each observation is grouped into one of these 5 clusters .

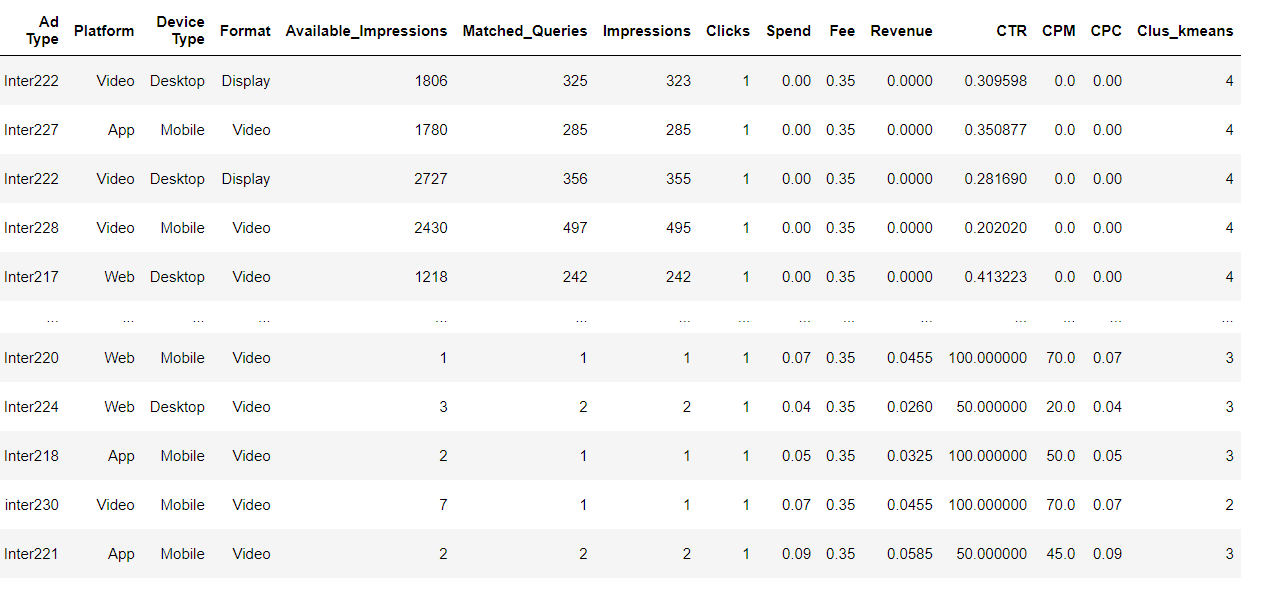


Figure 18: Dataset with Clusters

Let’s group the data into clusters and observe the behaviour of the important variables in each of the clusters by taking the means of these variables to identify trends in clicks, spend, revenue, CPM, CTR, & CPC.



Table 9: Grouping data by clusters with mean to identify trends in clicks, spend, revenue, CPM, CTR, & CPC based on Device Type

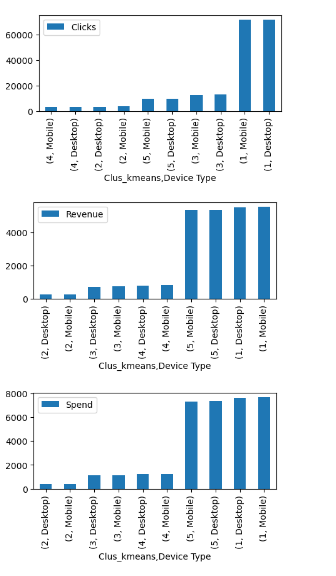
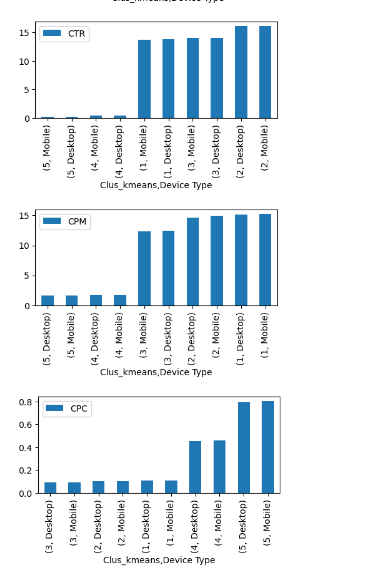
 

Figure 19 :Graphical Representation of Cluster wise trends in clicks, spend, revenue, CPM, CTR, & CPC based on Device Type

**Some observations based on the data summary and Visualizations**:

Cluster 1 has most Clicks and Cluster 4 has lowest Clicks

Cluster 1 has highest Revenue and Spend and Cluster 2 has lowest Revenue and Spend

Cluster 2 has highest CTR and Cluster 5 has lowest CTR

Cluster 1 has highest CPM and Cluster 5 has lowest CPM

Cluster 5 has highest CPC and Cluster 3 has lowest CPC

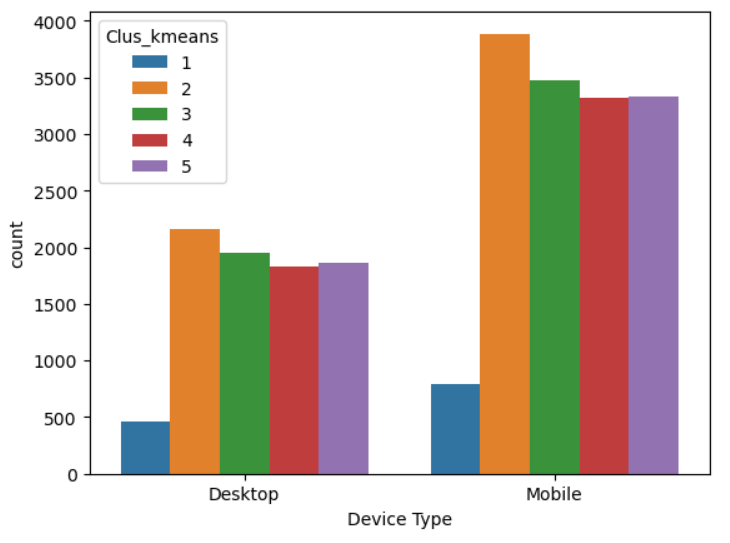


Figure 20: Countplot of Device Types and Clusters

There are more Ads played on Mobile device than on Desktop for all clusters

## 9. Clustering: Conclude the project by providing summary of your learnings.

1.Cluster 1 has the highest Revenue among all clusters, also highest spends. But the Fee is lowest among all clusters . They also have the highest Clicks and highest CPM.Such ads should be preferred as they are having low fee but highest revenue generating Ads.

2.Ads in Cluster2 are the lowest on spending and revenue is also very low . Their Fee is also quite high compared to another Ad’s. They have highest CTR (click to Impression ratio) and CPM (Spend per 1000 Impressions) is also high. CPC(spend per Click) is low but it’s not increasing their revenue. They have second lowest clicks probably because Fee is quite high, so to increase revenue they can lower the Fees.

3.Ads in Cluster3 has the highest Ad size, with second least Revenue and High Fee and lowest CPC(spend per Click), CPM(Spend per 1000 Impressions) is second highest . Maybe they can reduce their Ad size as its not guaranteeing more clicks or more revenue which could reduce some cost and spend more on creating better Ads.

4.Ads in Cluster4 has highest Fees among clusters and lowest Clicks and not a high revenue generating cluster . CPM(Spend per 1000 Impressions) is low . CPC (spend per Click) is second highest and CTR (click to Impression ratio) is very low. They have highest number of Impressions , but lowest clicks .May be reducing the Fees can increase number of clicks thereby increasing Revenue.

5.Ads in Cluster 5 has highest number of Impressions . Their Fees is also not very high with second highest Revenue and Spend. Highest CPC (spend per Click),Lowest CPM(Spend per 1000 Impressions). and lowest CTR(click to Impression ratio). They are not getting many clicks though the spend per click is high. Such Ads are generating good revenue amount but they have highest CPC so cost should be reduced to increase more revenue

# Problem2 :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.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | 1.Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc. Ans: Lets load the PCA India Data Census    There are 640 observations and 61 columns  Let’s Analyse the Dataset provided PCA India Data Census.xlsx and print the top 5 observations.    Figure 21:Top 5 observations of PCA India Data Census    Figure 22: Bottom 5 observations of PCA India Data Census  Printing the information about 61 columns    Table 10: Dataset Column Information1      Table 11: Dataset Column Information2  As can be seen from the figures above there are 59 Integer type and 2 Object types Columns  State code and Dist Code are numerical values , but they are actually categorical values as they are not  continuous variables  Printing the summary of the numerical columns(State code and Dist code are excluded)        Table 12: Describing the Numerical Columns    Table 13: : Describing the Categorical Columns  As we can see all the Numerical variables are have a right skewed distribution.  As we can among categorical variables ,Uttar Pradesh has the greatest number of observations and Raigarh is appearing twice because there both states Chhattisgarh and Maharashtra have area names called Raigarh  There are no Null values in the Dataset.  There are no duplicated observations in the dataset    Table 14: Null Values in the Dataset |  | | 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 Let’s pick 5 variables M\_LIT, TOT\_M, TOT\_F, TOT\_WORK\_M, TOT\_WORK\_F for our analysis  M\_LIT-Literates population Male  TOT\_M- Total population Male  TOT\_F- Total population Female  TOT\_WORK\_M- Total Worker Population Male  TOT\_WORK\_F- Total Worker Population Female  **Univariate Analysis**    Figure 23: Distribution Plot and Boxplot of Number of Literate Males    Figure 24: Distribution Plot and Boxplot of Total Number of Males    Figure 25: Distribution Plot and Boxplot of Total Number of Females    Figure 26: Distribution Plot and Boxplot of Total Number of Male Workers    Figure 27: Distribution Plot and Boxplot of Total Number of Female Workers    Table 15: Data Description of the 5 variables  **Inferences from above representations of the 5 Numerical Variables**  1.All the distributions are right skewed with outliers  2.The average number of females are more than the males.  3.There average number of females working population is also slightly more than male working population  4.Standard deviation of total females is very high compared to total males indicating the distribution is more spread out .  5.There are a minimum of 286 literate males in each district  **Bivariate Analysis of Numerical and Categorical Variables**  Let’s Analyse the state wise numbers of these variables  Observation from the EDA:  When we plot the Average Number of Literate males against states, we see that Uttar Pradesh seems to be having highest whereas Dadra and Nagar Haveli the least number of Literate Males    Figure 28: Bar plot of Literate males in each state  When we plot the Number of Males against states, we see that Uttar Pradesh seems to be having highest number of males whereas Dadar and Nagar Haveli has the least number of males    Figure 29: bagplot of Total Males in each state  When we plot the Number of Females against states, we see that Uttar Pradesh seems to be having highest number of females whereas Dadar and Nagar Haveli has the least number of females    Figure 30: bar plot of Total Females in each state  When we plot the number of working males against states, we see that Uttar Pradesh seems to be having highest number of working males whereas Dadar and Nagar Haveli has the least number of working males    Figure 31: bar plot of Total Working Males in each state  When we plot the number of working females against states, we see that Uttar Pradesh seems to be having highest number of working females whereas Dadar and Nagar Haveli has the least number of working females    Figure 32: bar plot of Total Working Females in each state  **Scatterplot of Numerical Variables to see the correlation among the numerical variables**    Figure 33:Scatterplot of Numerical Variables  We can see that there is a positive correlation among the numerical variables as seen in the scatterplot    When we plot the Gender Ratio against each state, we can observe that Lakshadweep seems to have a very high Male to Female Ratio whereas Andhra Pradesh seems to have least Male to Female Ratio, whereas when we compare the districts, Lakshadweep seems to have a high gender ratio of .87 and Krishna District in Andhra Pradesh seems to have the lowest ratio with .437 (Plotting will not possible because there are 640 districts ).    Figure 34: bar plot of Gender Ratio is Each State    Figure 35:District wise Ratio of Gender  When we plot the Gender Ratio of working population against each state, we can observe that Lakshadweep seems to have a very high Male to Female working population Ratio whereas Arunachal Pradesh seems to have least Male to Female working population Ratio    Figure 36:State wise Gender Ratio of Male to Female Working Population  When we plot the Ratio of Literate population of Males against total Male population in each state , we can observe that Goa seems to have a very high Ratio of Literate population of Males against total Male population whereas Bihar seems to have least Ratio of Literate population of Males against total Male population    Figure 37: State wise ratio of literate Male to Total Male Population  When we plot the Ratio of working population of Males against total Male population in each state , we can observe that Karnataka seems to have a very high Ratio of working population of Males against total Male population whereas Arunachal Pradesh seems to have least Ratio of working population of Males against total Male population    Figure 38: State wise ratio of Working Male to Total Male Population  When we plot the Ratio of working population of females against total females’ population in each state , we can observe that Nagaland seems to have a very high Ratio of working population of females against total females’ population whereas Lakshadweep seems to have least Ratio of working population of females against total females’ population    Figure 39: : State wise ratio of Working Female to Total Female Population |  | | 3.We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary? In a dataset, outliers are data points that strongly deviate from the norm and can impact model accuracy. Normally , treating outliers from a dataset before model training can improve the performance of the resulting model.  But some reasons why we should not remove outliers in a dataset:   * When there are a lot of observations in the dataset as it could mean something about the data that need to be analysed further * When the model results are critical and can easily pose risks, for example, if dealing with sensitive use cases in the health or self-driving sectors * When the outliers are natural to the data, in this case population of the entire country. They could have hidden patterns which otherwise wouldn't be unearthed if they were to be removed.   So, in this case treating outliers is not a good option |  | | 4.Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment. From the data description we see that variances of the variables are so widely different; it is not a good idea to perform PCA on the unscaled variables. PCA works on the total variance which is the sum of the variances in the data. If one variance or more variance(s) is very high compared to the rest, it will dominate the construction of the PCs and all variables will not have proper representation. When sample variances of the original variables show differences by large magnitude, variables need to be normalized.  Let’s Visualize data before scaling    Figure 40 :Boxplot of Variables before scaling    Figure 41: Boxplot of Variables after scaling  We used the Z score method to scale the data .  As we can see from the boxplot before and after scaling , there is no impact of scaling on the outliers.  Printing the Dataset after scaling    Figure 42: Scaled Data |  | | 5.Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector. PCA is a method that: Measures how each variable is associated with one another using a **Covariance matrix.** Understands the directions of the spread of our data using **Eigenvectors.** Brings out the relative importance of these directions using **Eigenvalues.**  **PCA uses this concept of eigen decomposition.** We center the n predictors to their respective means and then get an n x n covariance matrix. This covariance matrix is then decomposed into eigenvalues and eigenvectors. So, a covariance matrix has variances (covariance of a predictor with itself) and covariances (between predictors).Eigenvectors are unit vectors with length or magnitude equal to 1. They are often referred to as right vectors, which simply means a column vector.Eigenvalues are coefficients applied to eigenvectors that give the vectors their length or magnitude  Here a snapshot of Covariance Matrix for the all 57 variables    Table 16Snapshot of the Covariance Matrix of all the numerical Variables    Figure 43:Correlation heatmap between 57 variables  The correlation matrix and heatmap indicate high correlation among some variables such as no of Households and No of Males./Females, No of working Males/Females. Moderate correlation may be detected between several pairs of variables, such as; between number of Males/Females and Main Work population etc. Existence of such pairs of high and moderate correlations indicate that dimension reduction must be considered for the data.  There are 57 variables , hence we get 57 Principal components  Following is a snapshot of the Eigen Vectors for all 57 Principal components .Since it’s not possible to list all the 57 eigen vectors for 57 PCA, below is just a snapshot    Table 17:Eigen Vectors  Following are the 57 Eigen values for all the 57 principal components .    Table 18:Eigen Values  In the above output, eigenvectors give the PCA components and eigenvalues give the explained variances of the components. The first few PC components explain the maximum variance .If there are eigenvalues close to zero, they represent components that may be discarded because they don’t explain much of the data.  Let’s draw the Scree Plot with PCA =57    Figure 44:Scree Plot with 57 Principal Components  Above Scree plot is a graphic that shows the explained variance per principal component. The measure of the plot can be the percentage or the absolute value of the explained variance (eigenvalues). The first few principal components explain the major amount of variance. |  | | 6.Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot. The scree plot is a useful visual tool to select optimum number of PC. On the X-axis are shown the indices of the PCs and on the Y-axis are shown the variances. If there is a distinct break point in the line joining the variances (elbow point) beyond which the line becomes approximately horizontal, then that point may be taken as the value of 𝑘, provided other conditions are also satisfied.  Since its not feasible to have 57 PC’s and first few components explain the maximum variance let’s take number of PC’s as 10 and get the variances of each of them . Let’s get the variances ratio of each component and the cumulative variance . As we can see PC1(first value in the array) explains maximum variance with 31.8 as variance    Figure 45 PCA Variance with 10 components  Let’s get the explained variance ratios. Explained Variance Ratio is explained variance of component / (total of all explained variances).Here PC1 has ratio of 55.7    Figure 46 PCA Variance Ratio with 10 components  Let’s get the cumulative variance ratios.PC1 has maximum ratio of 55.7. PC 1 and 2 together explain 69.5% variance .PC1 ,2 and 3 explain 76.7% variance and so on.    Figure 47:Cumulative Variance of the Components  Let plot a scree plot of PC components equal to 10    Figure 48: Scree Plot with 10 Principal Components  In Fig 48, there is a distinct break at 2. However, 𝑘 cannot be taken to be 2 since the first two PCs explain only 69.5% of total variance. The PCs must be taken so as to explain at least 90% of the total variance. If 𝑘=6, then the first 6 PCs explain 90.4% of the total variance. One choice of 𝑘 could be 6. Also, from the scree plot we can see that after 6 Principal components , the Variance explained isn’t changing much , so we can select 6 PCs to represent the entire dataset    Table 19 Variances and Standard Deviations of each PC  The principal components are constructed in decreasing order of magnitude of their standard deviations, which is equivalent to decreasing order of magnitude of their variances. In Table 19 the variances of the constructed principal components and their sum total is given. |  | | 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. Since we have selected 6 as optimum number of principal components , the new dataset will now be represented by these 6 variables instead of 57 variables . Principal components are linear combinations of the original variables or scaled variables, as the case may be. It is possible that some of the coefficients are very small numbers or close to 0. We present the linear combinations that make up the first 6 PC’s  Let’s have a look at the PCA loadings of each of these PC components with each of the columns      Here is a Graphical representation of Principal components with the corresponding loadings for each column of the dataset    Figure 49: PCA Loading of PC1 Figure 50: PCA Loading of PC2    Figure 51: PCA Loading of PC3 Figure 52: PCA Loading of PC4      Figure 53: PCA Loading of PC5 and PC6  Heatmap of PC loading shows the correlation of PC with actual columns    Figure 54:Correlation map of Principal Components and Actual Columns    PC1 is showing the positive correlation with many variables  PC2 is showing the very high correlation with variables MAIN\_OT\_M,MAIN\_OT\_F,MARG\_CL\_M,MARG\_CL\_F,MARG\_AL\_3\_6\_M,MARG\_AL\_3\_6\_F,MARG\_AL\_0\_3\_M,MARG\_AL\_0\_3\_F representing Main Other Workers Population Male and Female, Marginal Cultivator Population Male and Female, Marginal Agriculture Labourers Population 3-6 Male and Female, Marginal Agriculture Labourers Population 0-3 Male and Female  PC3 is showing the maximum correlation with variables MAIN\_AL\_M,MAIN\_AL\_F,MARG\_CL\_F,MARG\_AL\_3\_6\_F,MARG\_HH\_3\_6\_F,MARG\_AL\_0\_3\_M,MARG\_AL\_0\_3\_F indicating columns representing Main Agricultural Labourers Population Male and Female and  Marginal Agriculture Labourers Population 0-3 Male and Female, Marginal Household Industries Population 3-6 Female, Marginal Agriculture Labourers Population 3-6 Female, Marginal Cultivator Population Female  PC4 is showing the maximum correlation with variables MAIN\_AL\_F,MAIN\_CL\_F, MARG\_OT\_3\_6\_M , MARG\_OT\_0\_3\_M , MARG\_HH\_M,MAIN\_WORK\_F,TOT\_WORK\_F indicating columns representing Main Agricultural Labourers Population Female and Main Cultivator Population Female, Main Working Population Female, Total Worker Population Female  PC5 is showing the maximum correlation with variables M\_ST, F\_ST , MAIN\_CL\_M, MAIN\_CL\_F , M\_SC, F\_SC indicating a column representing Scheduled Tribes Population , Scheduled Caste Population and Cultivator Population.  PC6 is showing the maximum correlation with variables MAIN\_HH\_F,MARG\_OT\_3\_6\_F,MARG\_HH\_F,MARG\_OT\_0\_3\_F indicating a column representing Marginal Working Population  Principal components are linear combinations of the original variables. Each PC is a linear combination of all variables, or scaled variables, as the case may be .Once the original variables are replaced by the PCs, the latter are used for any further analysis. Just as each observed unit has a particular value of each variable, similarly each observation has a particular value for each PC. These values are called PC scores. These scores are obtained by putting scaled values of the variables in the expression of PCs as shown below    Table 20 PC scores    To check that the PCs are orthogonal, correlation matrix is computed. As can be seen the correlation between the PC’s is 0, indicating there is no correlation between them.    Table 21:Correlation between PC’s    Figure 55:Heatmap of PC’s |  | | 8.PCA: Write linear equation for first PC. Formula for Linear equation for PC1 = a1x1 + a2x2 + a3x3 + ............+ anXn , where a1,a2....aN are the coefficients or loadings and x1,x2,x3.....xn are the observed data.  For each PC, these are the coefficients with which the corresponding variables need to be multiplied to get the PC. Note that the weights can be positive or negative    Using the above weights w arrive at the linear equation for PC1 as below . Note the coefficient have been rounded off to 2 decimal places    Table 22: linear equation for first PC. |  | |  |  | |