**CLUSTERING CLEAN\_ADS**

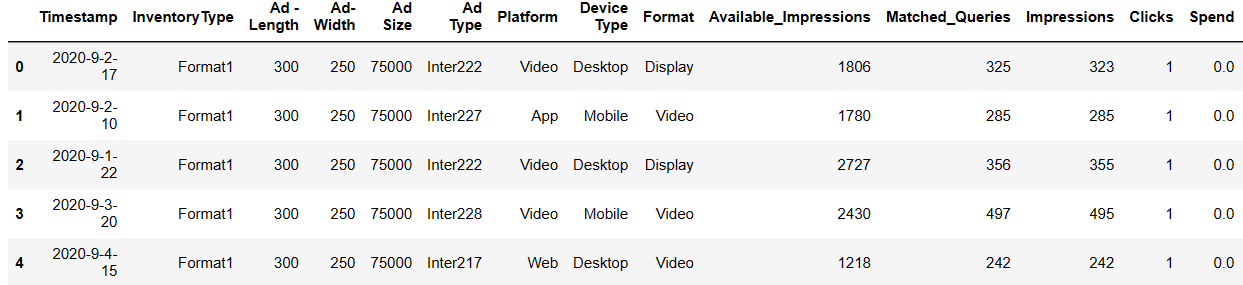
|  |  |
| --- | --- |
| Part 1 - Clustering: Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc. | 4 |
| Part 1 - Clustering: Treat missing values in CPC, CTR and CPM using the formula given. | 4 |
| Part 1 - Clustering: 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). | 3 |
| Part 1 - Clustering: Perform z-score scaling and discuss how it affects the speed of the algorithm. | 3 |
| Part 1 - Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance. | 4 |
| Part 1 - Clustering: Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm. | 4 |
| Part 1 - Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters. | 4 |
| Part 1 - Clustering: 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]. | 4 |
| Part 1 - Clustering: Conclude the project by providing summary of your learnings. | 3 |

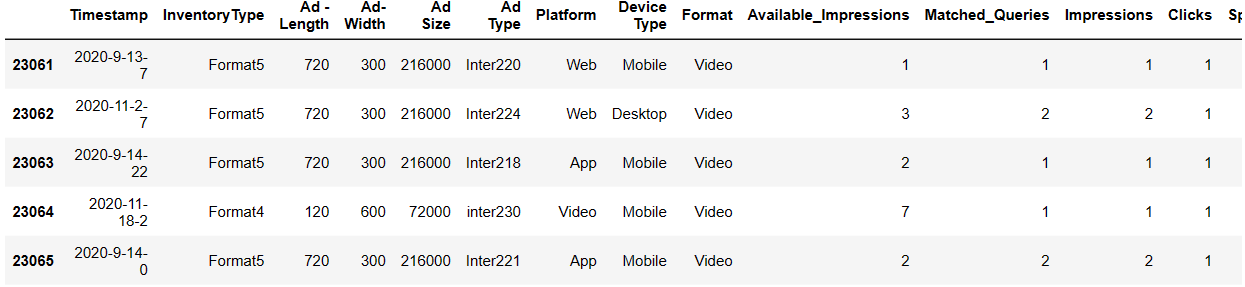
**Part 1 Clustering: Read the data and Perform basic analysis such as printing a few rows(head and tail),info,data summary,null values, duplicate values etc. (4 Marks)**

**Ans:-** We have loaded the all the required packages and loaded Clustering clean Ads Data file using Pandas.

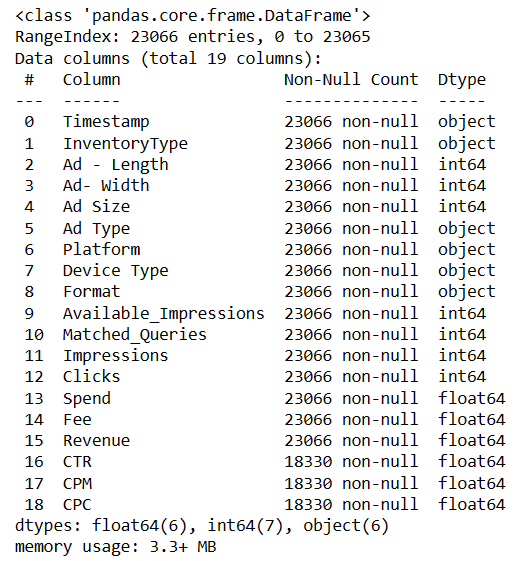
Dataset has 23066 rows and 19 columns.

We have viewed first and last few rows using head() and tail() functions respectively.





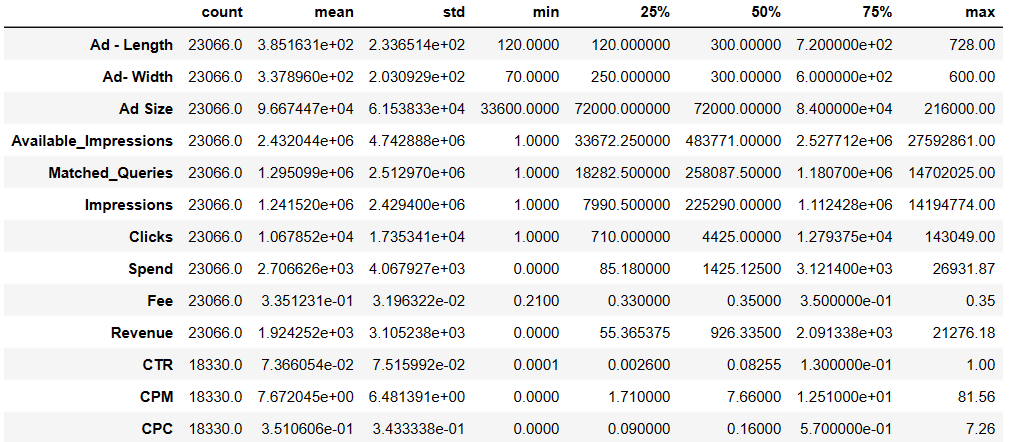
We can view dataset information using info()



We can see that there are 2 objects data types , 6 floats and 7 integers. There are null values in CTR, CPM, CPC columns.

There are no duplicate values.

Dataset can be described using describe() function.



Mean,standard deviation, count etc can be seen using describe() function for all the numerical fields.

1. Ad length ranges from minimum 120 to 728 max.The average Ad length is 3.8.Ad width ranges from 70 to 600 and has an average of 2.03.
2. CTR, CPM and CPC has 0 minimum.
3. Available Impression is counted each time ad is shown so impression is counted at least 1 time and at max it is counted 14194774 times. It has an average of 1.24 times.
4. Clicks had minimum of 1 times so user has clicked on an ad at least 1 times and at max 143049 times.
5. Minimum revenue earned by advertisement is 0 and maximum is 21276. On an average Revenue earned is 1.92.

**Part 1: Clustering: Treat Missing values in CPC, CTR,CPM using the formula given .(4Marks)**

**Ans**:

**Null values:**

There are null values in CPC, CTR and CPM columns

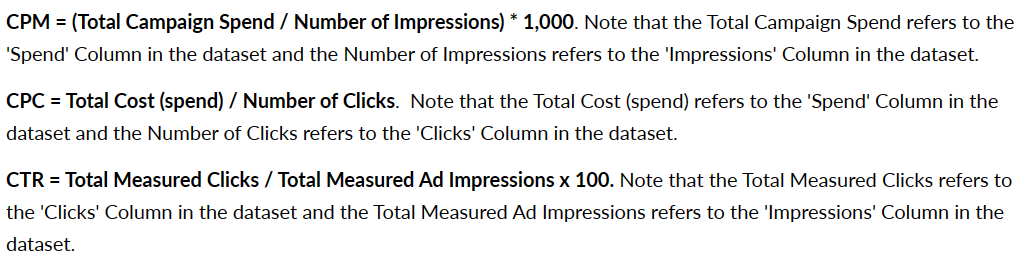
CPC :- Total 4736 null values

CTR:- Total 4736 null values

CPM:- Total 4736 null values.

**Treating Null values:**

Formula to calculate CPC, CTR and CPM are given the description.



There are many rows missing in the CTR,CPM and CPC columns so we cannot drop the rows.

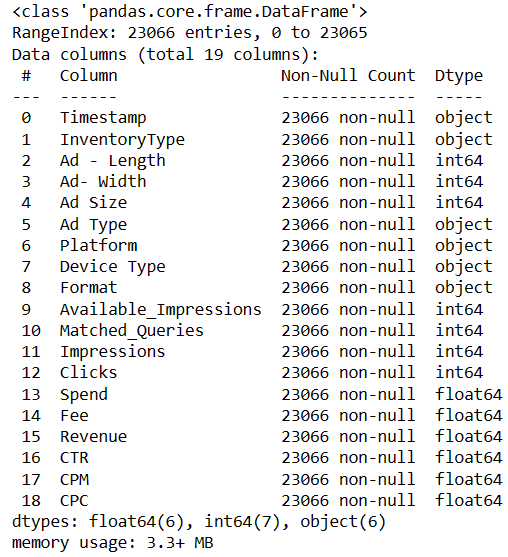
Data field with less value missing can be dropped.

These 3 fields has float data type.

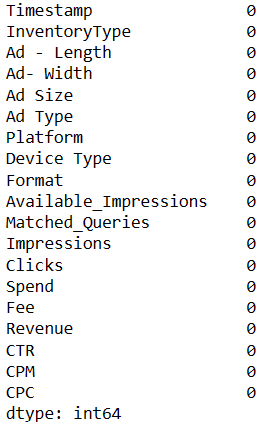
Function was created using given formulas to calculate values of CTR,CPM and CPC.

Using this function all values are calculated. Null values or NaN values in these 3 columns are replaced by values which are calculated using the function.

Dataset information can be viewed to check if there are any null values



Dataset has no values.

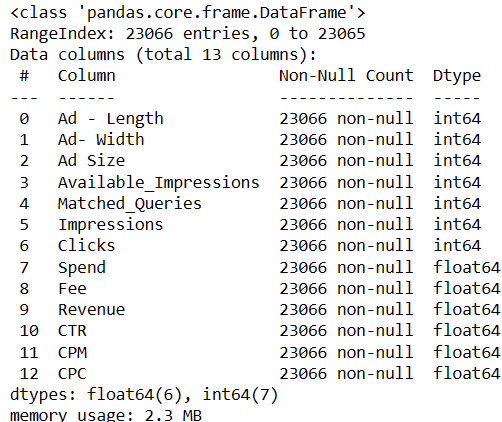


**Part 1: Clustering Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an analyst your judgement may be different from another analyst) (3 Marks)**

**Ans:**

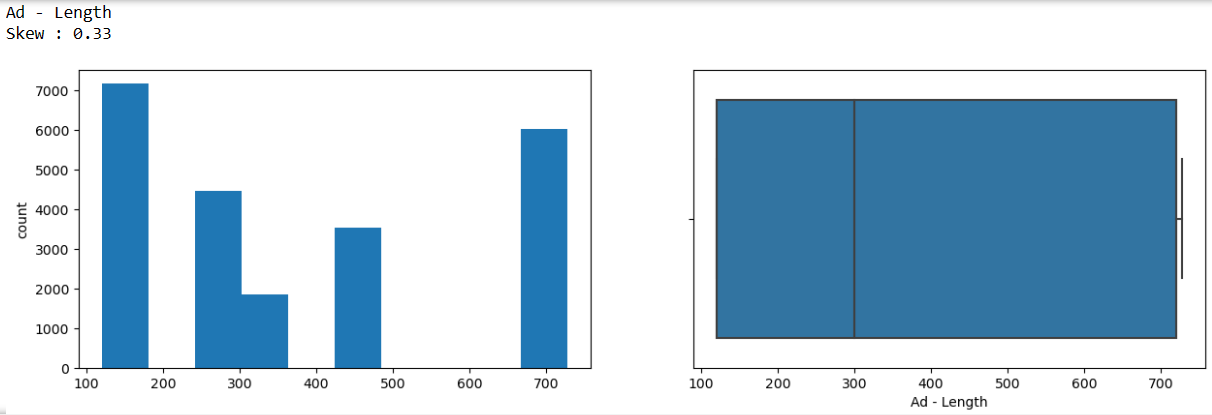
Lets check the distribution and outliers for each column in the data.

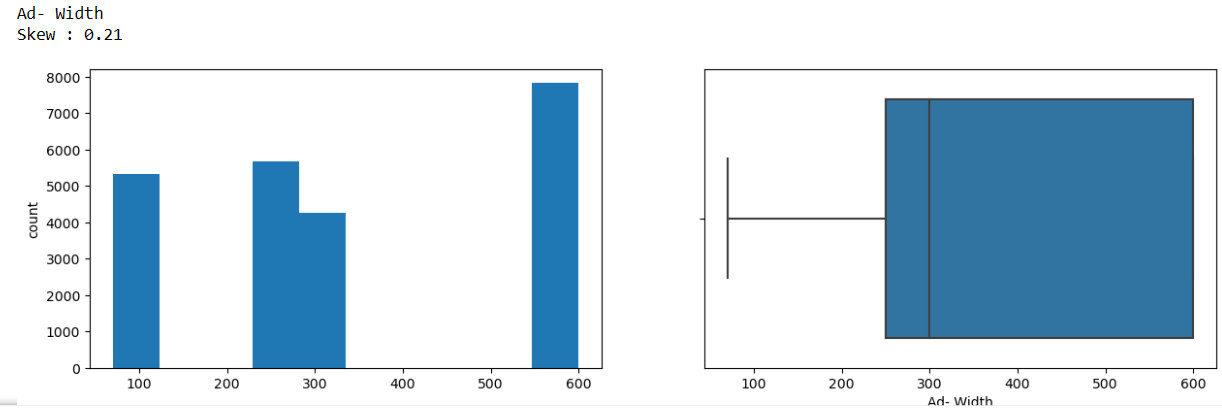
Before checking for outliers lets drop all the categorical column( 'Timestamp','InventoryType','Ad Type','Platform','Device Type','Format') and create new dataframe with only numerical columns.

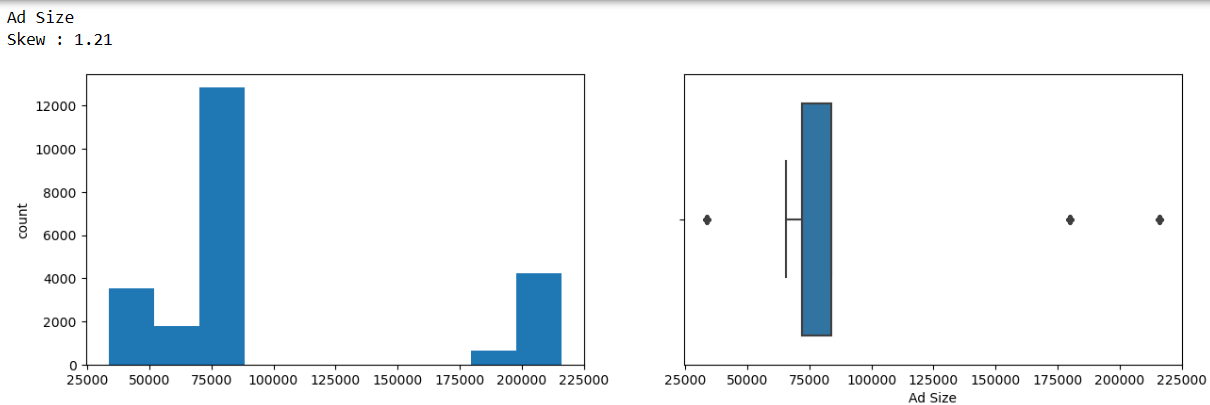


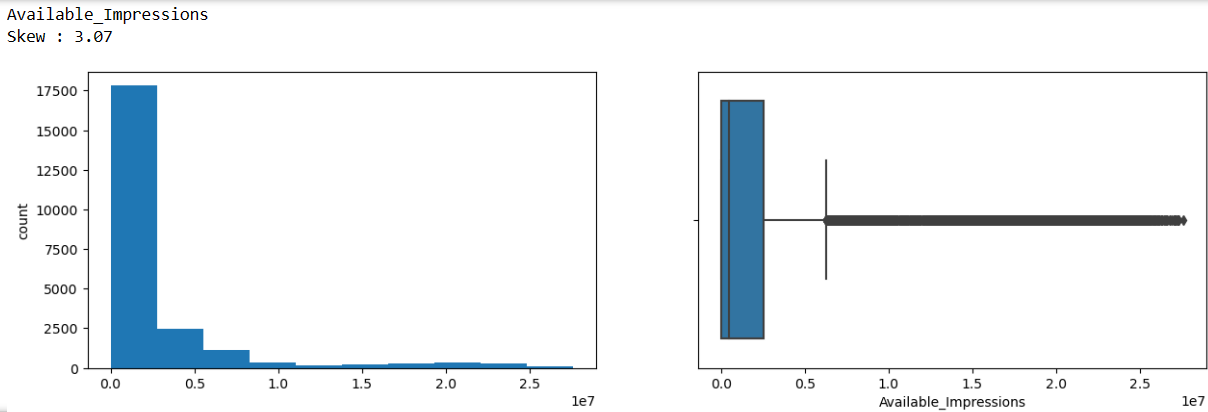
Lets check the outliers by plotting the boxplot for all the numerical fields.

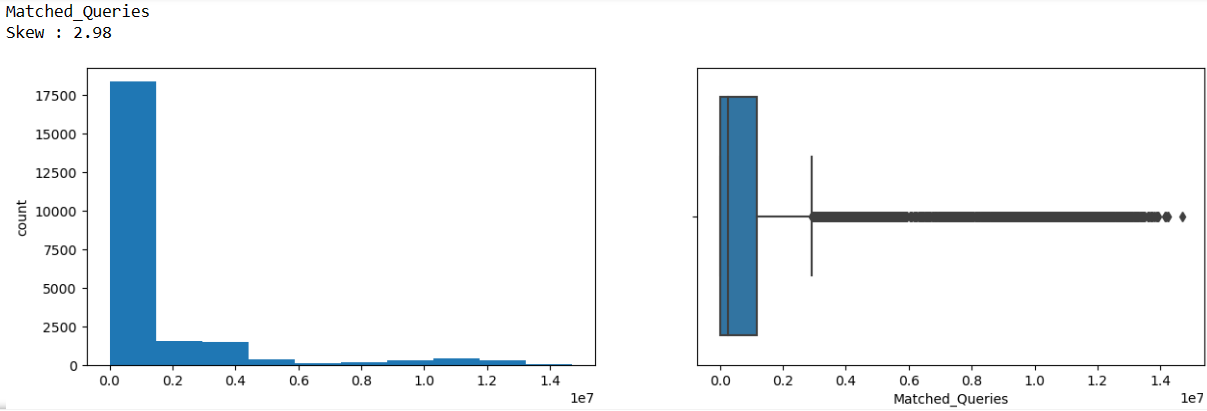
This also shows plotting graphs for All Numeric variables which is known as Univariate Ananlysis.

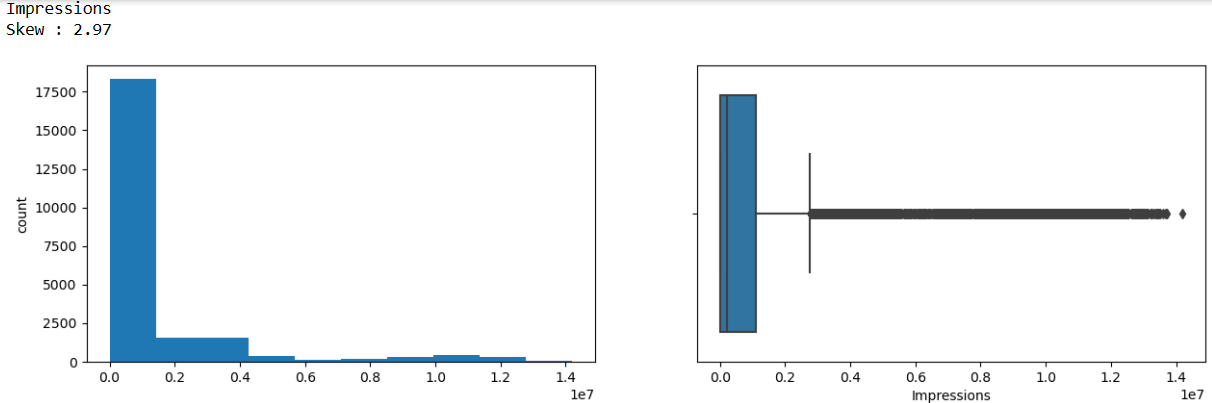


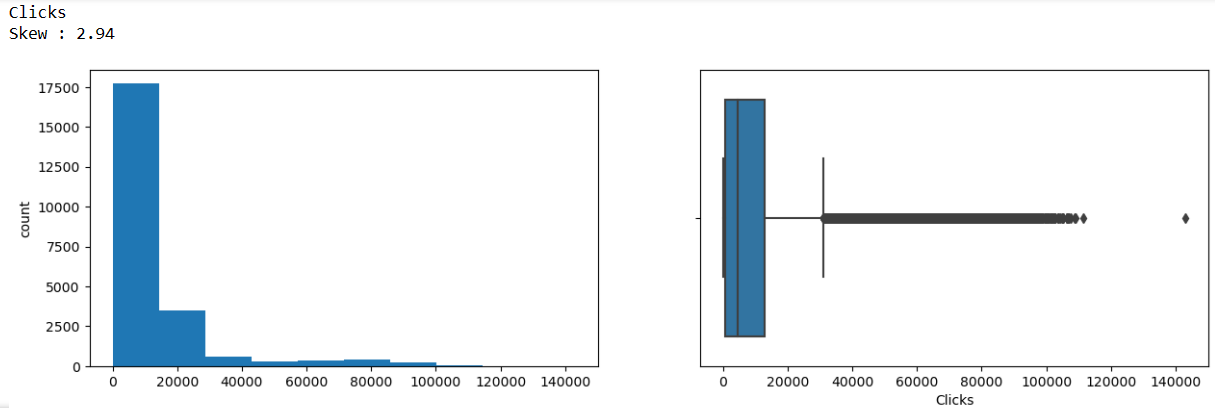


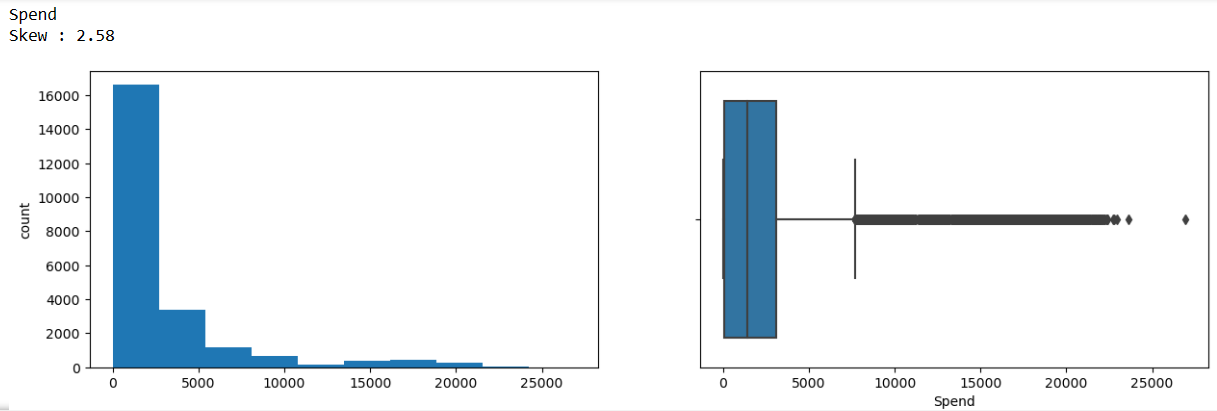


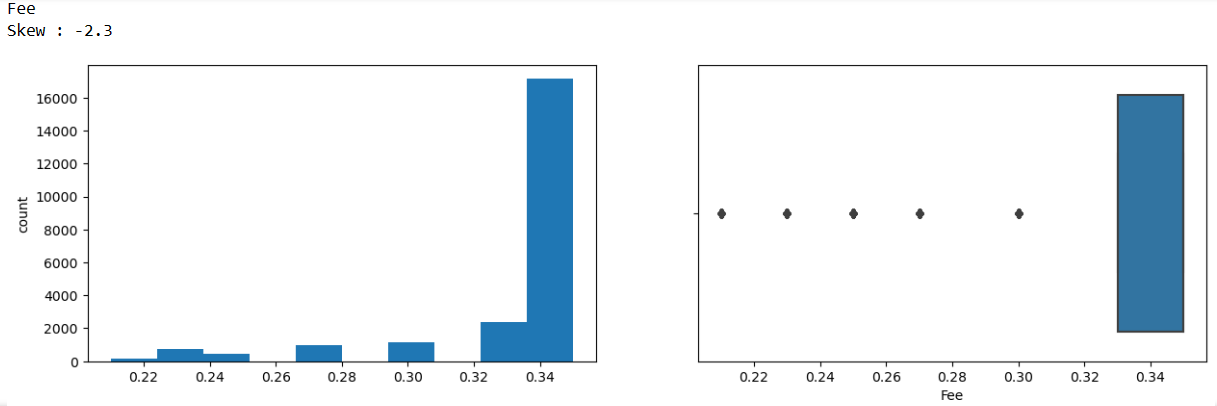


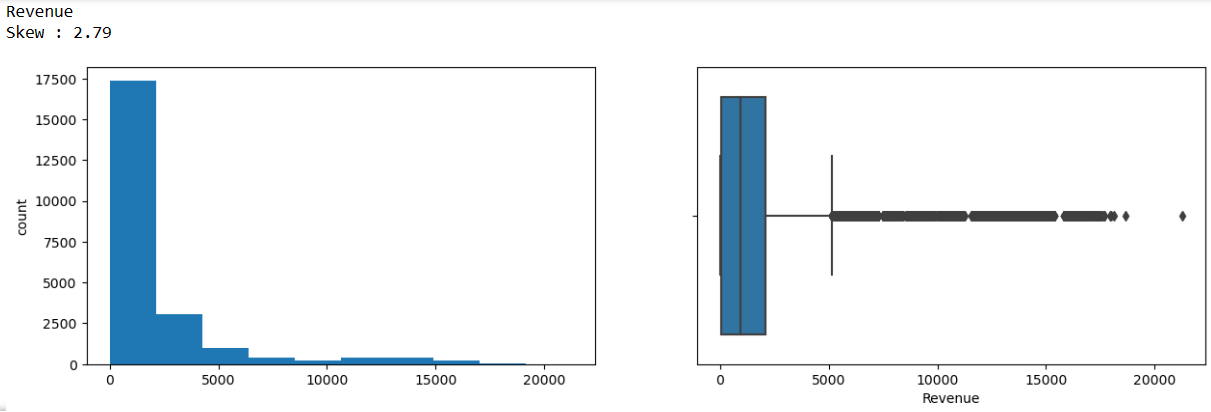


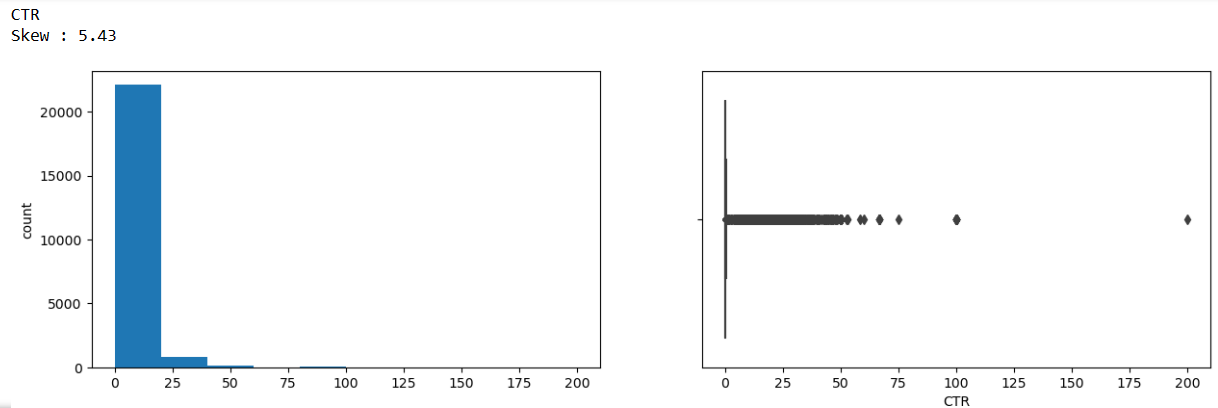


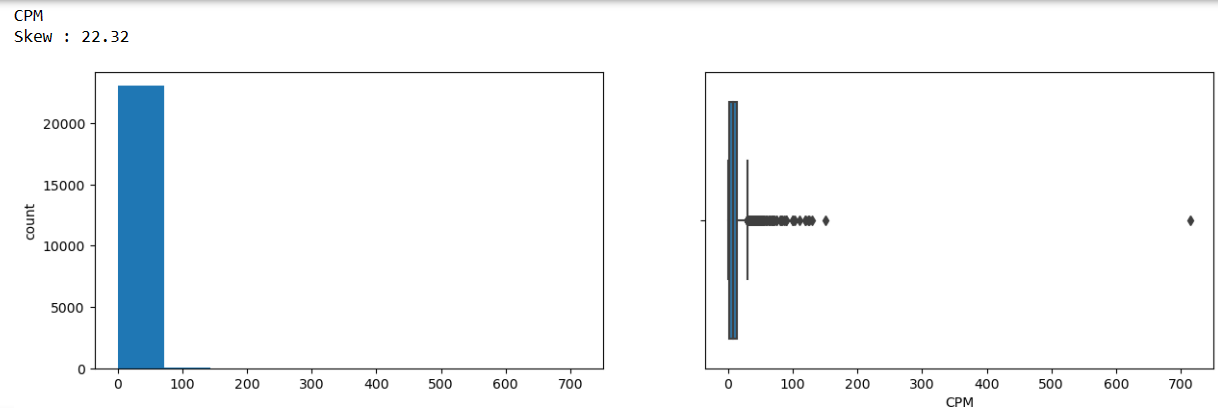


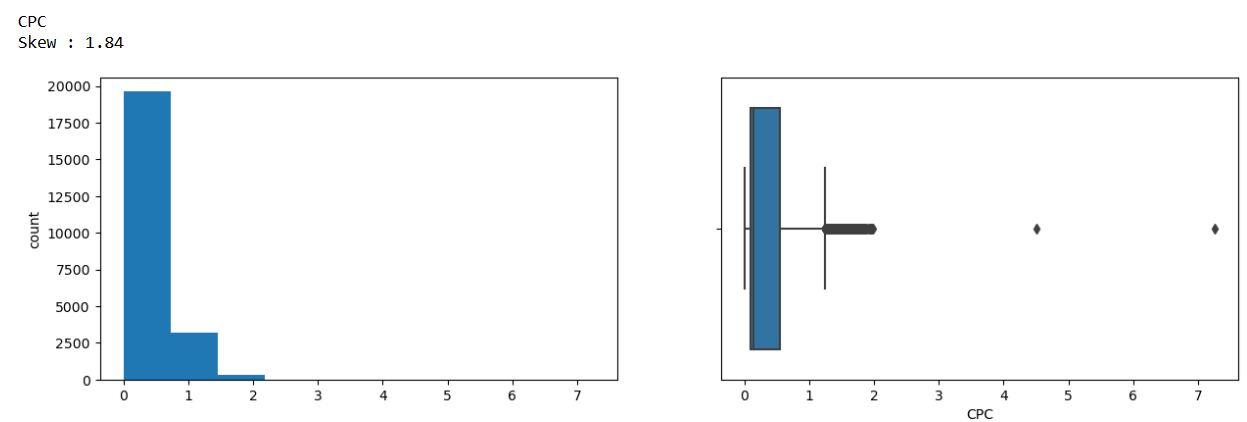












Insights:

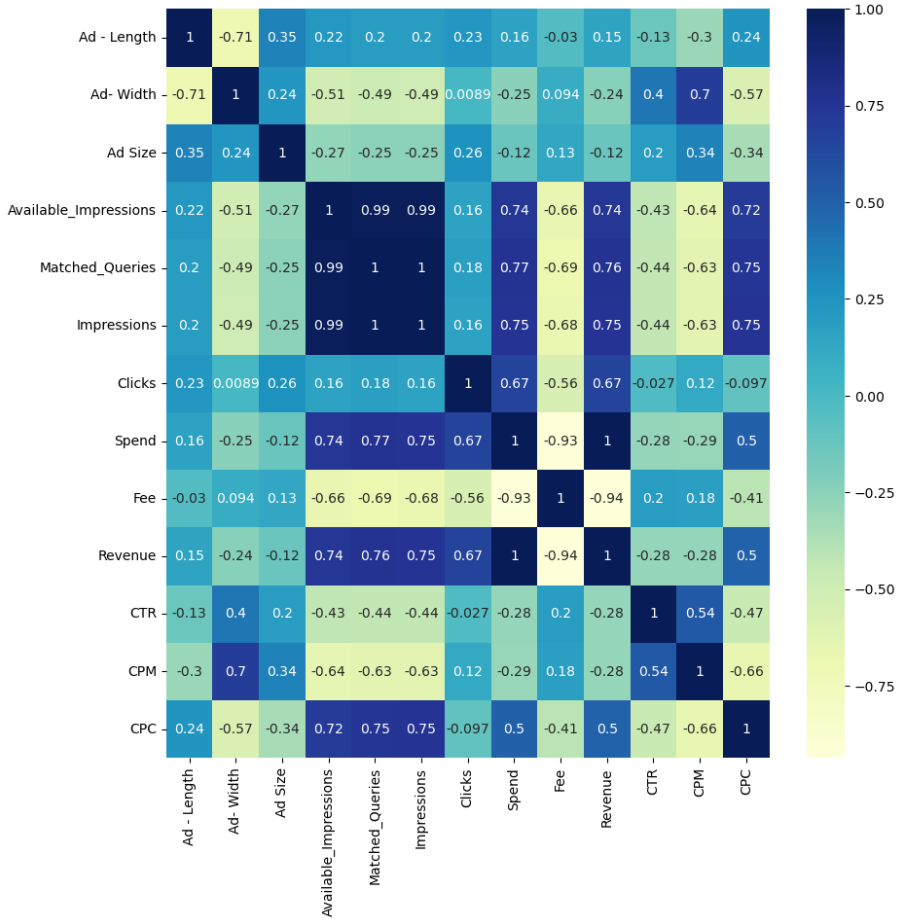
1. As observed, Most of the variables have skewed distributions.
2. Ad Length and Ad Width are slightly skewed with no outliers.
3. Fee is the only variable which is skewed to the left meaning high percentage of the Advertising Fees payable by Franchise Entities.
4. The distribution for all other variables is skewed to the right. All these variables have some outliers to the right end.

There is a need to treat Outliers in our case. In some cases Outliers provide useful information and in some cases Outliers can affect the result of Analysis.

Outliers needs to treated while doing clustering as it affects the result of scaled Data. As we are here using Zscore method for scaling so Outlier treatment is must as it affects value of mean.

Here we are using (1.5 \* IQR) method to treat outliers.

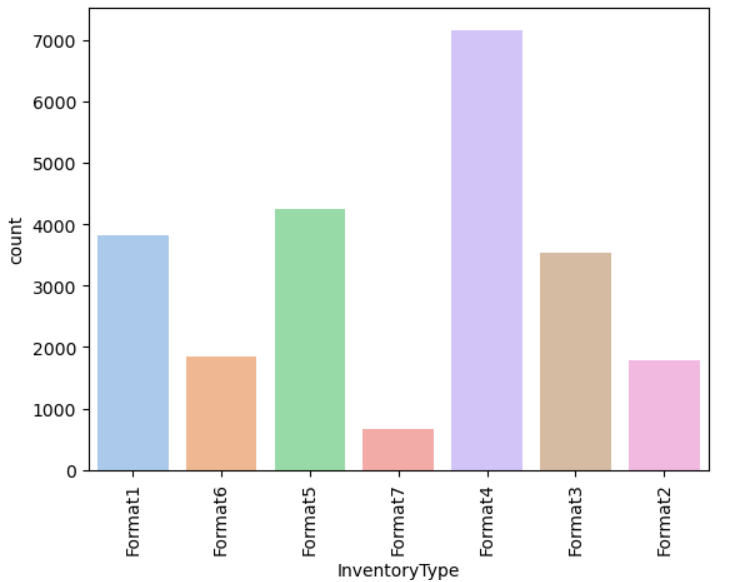
**Let's check the correlation among the variables**

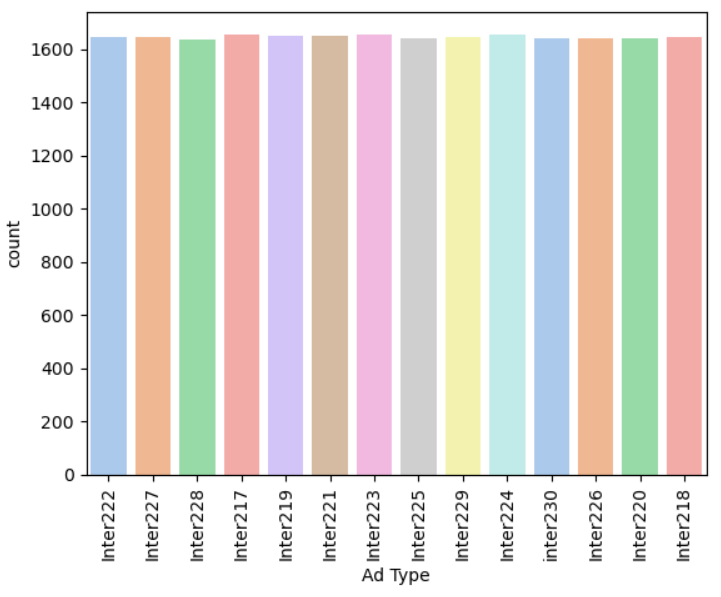


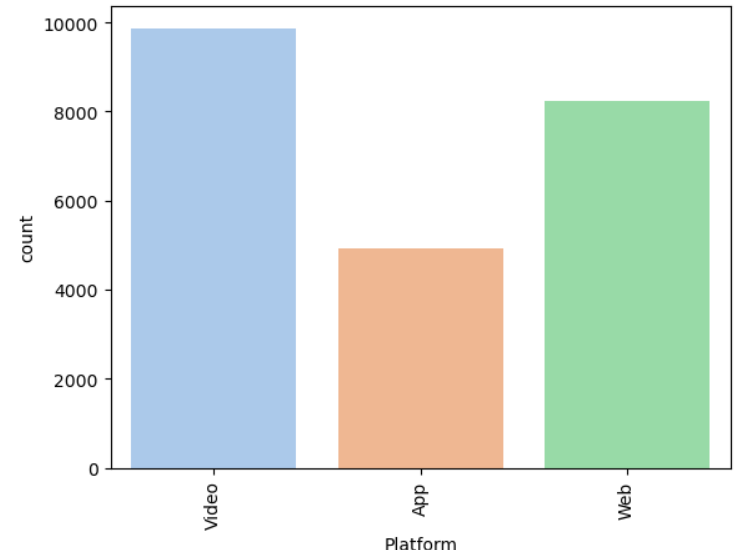
Insights:

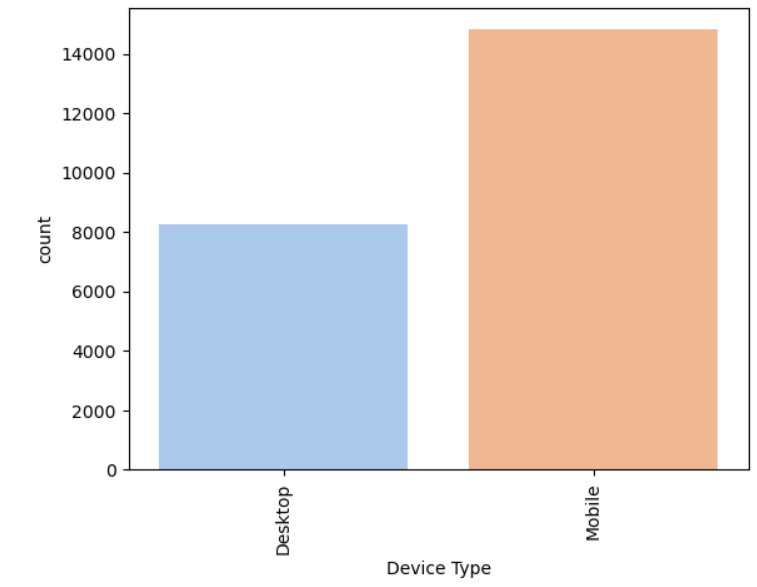
1. There is strong Positive correlation of Available\_Impressions with Impressions and Matched\_Queries. Available\_Impressions is moderately correlated with the revenue.
2. There is correlation between Ad length and Ad size which means that as the Ad length increases Ad size also increases.
3. Spend and Fee are highly negatively correlated.
4. CPC is positively correlated with Available Impressions,Matched\_Queries and Impressions. Cost per click increases as Impressions,Matched\_Queries and Impressions increases.

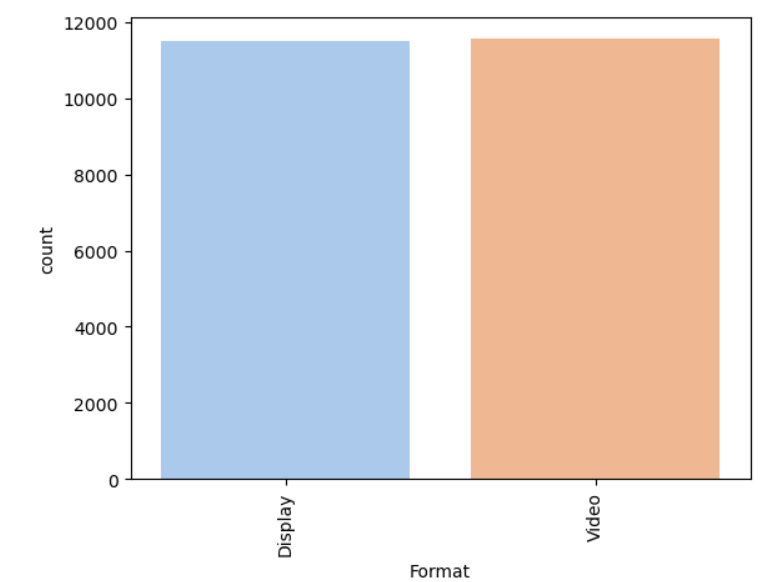
Univariate Analysis for all the Categorical columns.







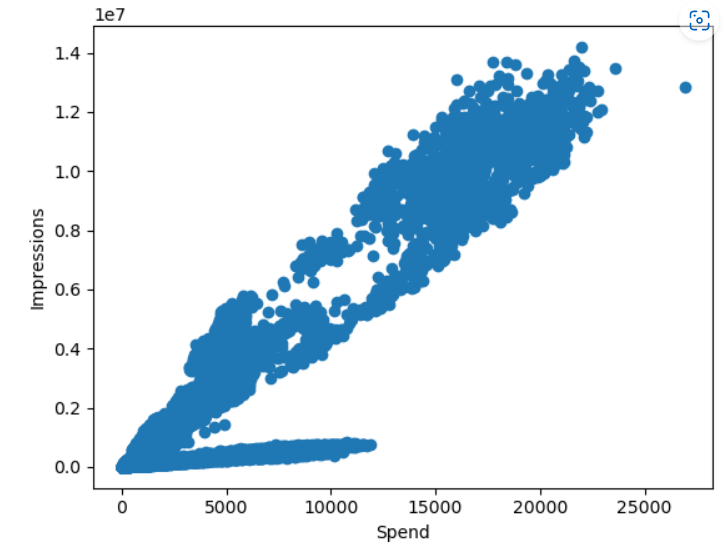




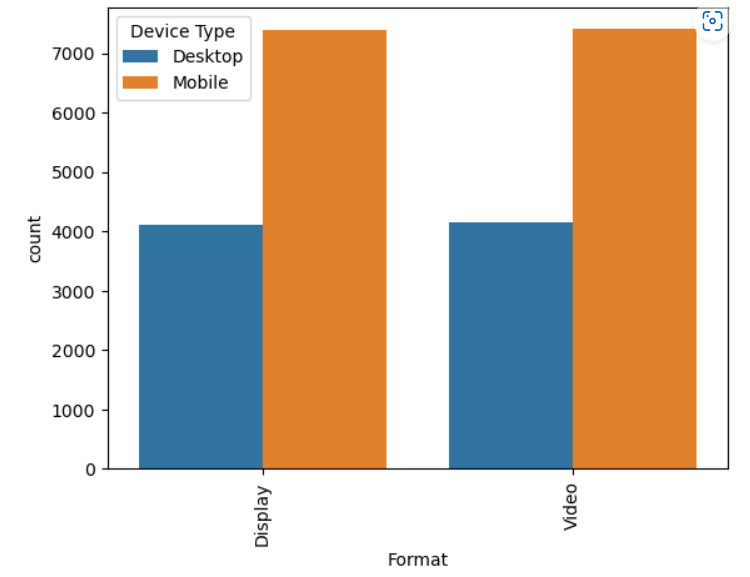
Observations:

1. Inventory type format4 is larger in number followed by format5 and others.
2. Video Platform is used more to display Ads followed by Web and App.
3. Mobile is more supportive to most of the Ads as compared to Desktop
4. Format is displayed equally in Video and Display.

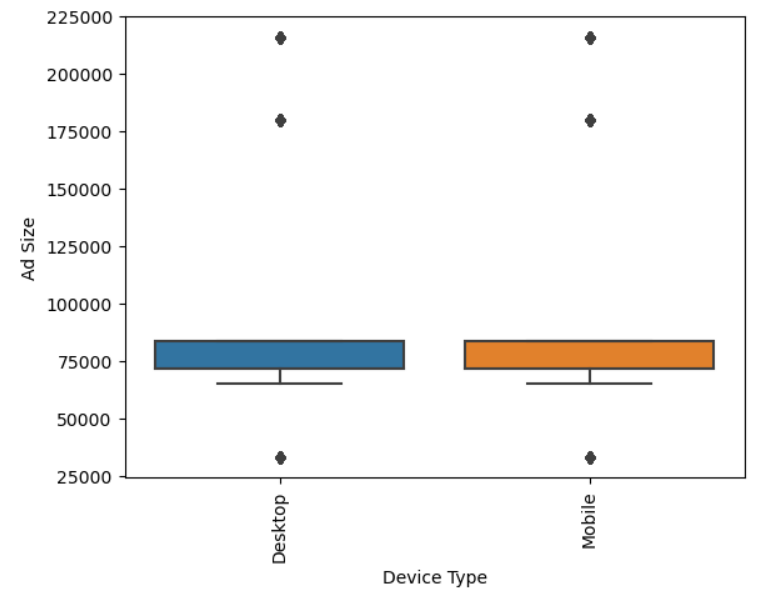
Numeric vs Numeric

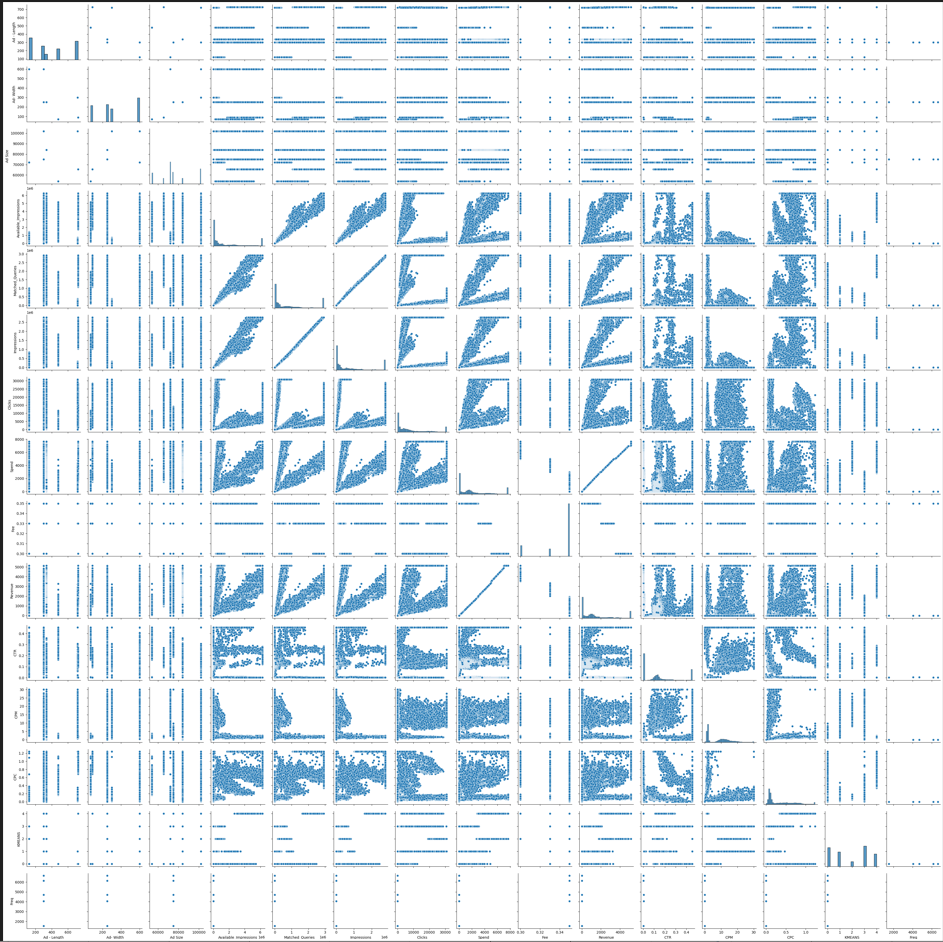
1. 
2. Spend and Impressions are Positively correlated.

Categorical vs Categorical



Categorical vs Numeric





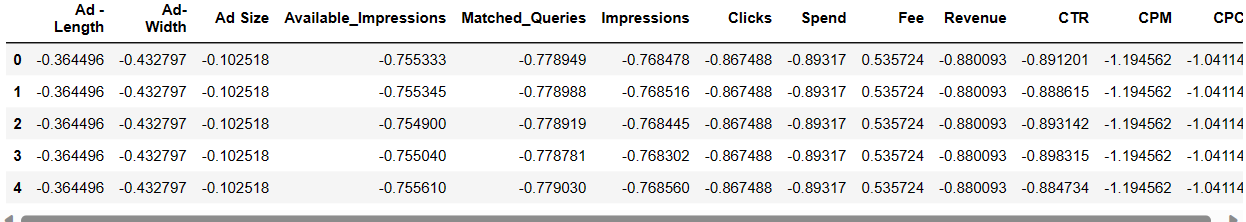
**Part 1: Clustering Perform z-score scaling and discuss how it affects the speed of the algorithm.(3 Marks)**

**Ans:**

**Scaling the Data**

1. Categorical data was removed from the original DataFrame and formed a new DataFrame data\_df with only numerical Values.
2. Zscore imported from scipy.stats.
3. Zscore method is used for scaling the data.

Lets have a look at few rows of scaled data.



Now all the values from all the columns are scaled. There are no extreme values.

Clustering Algorithms are distance based algorithm.

Both in Hierarchical and Non-Hierarchical techniques, clusters are formed based on distance calculation.

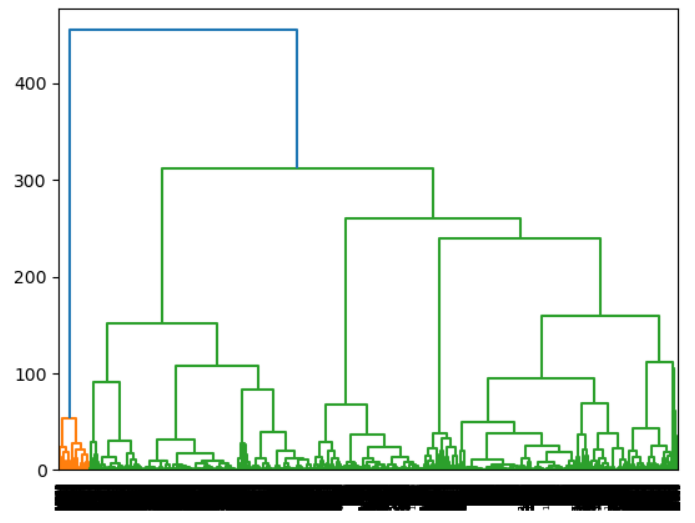
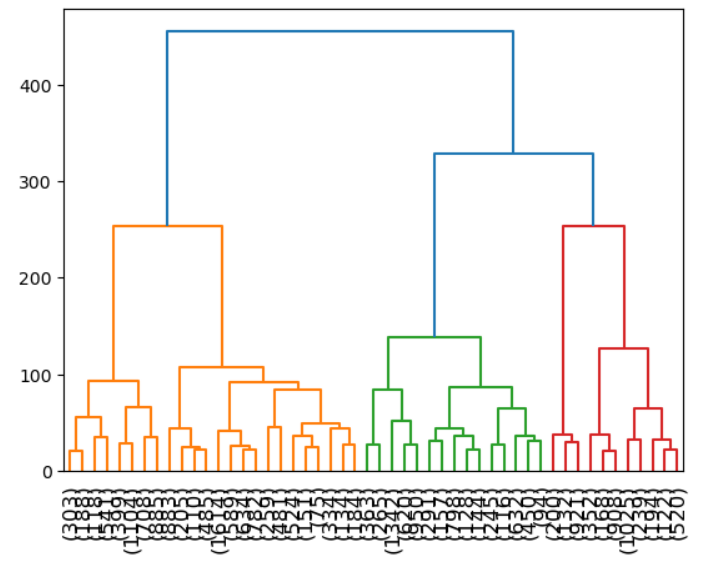
So if some features have much larger values than the other then they might affect the analysis of the result.

Scaling helps in standardization of the data. Clustering is very sensitive to the scaled variables.Once the scaling is done all the variables will be at the same scale which in turn increases the speed of the algorithm.as Scaling is Distance based algorithm.

**Part 1: Clustering Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance. (4 Marks)**

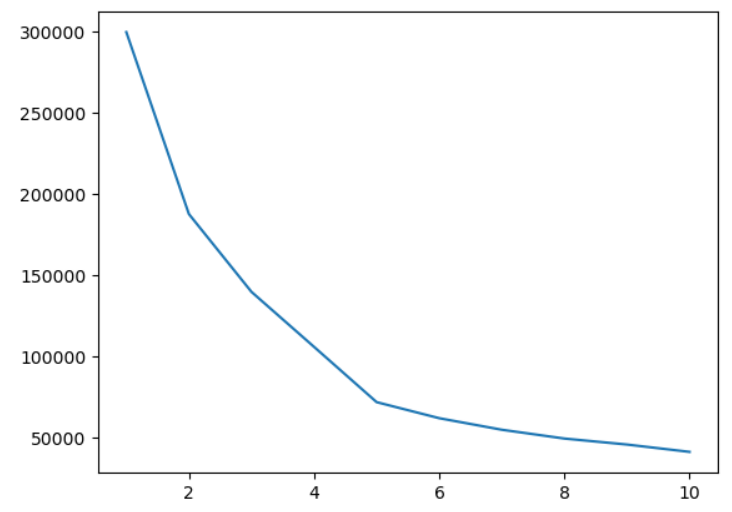
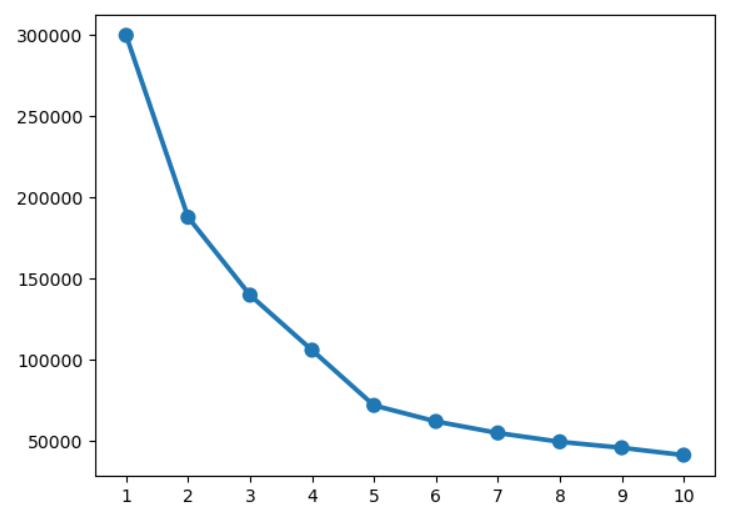
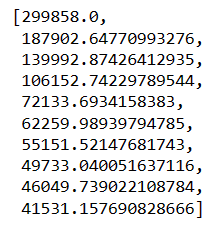
**Ans:**

**Creating a Dendrogram**

1. Importing Dendrogram and linkage module.
2. Choosing Ward and Euclidean distance Method to form clusters.
3. 
4. Cutting the Dendrogram with suitable clusters
5. 
6. We are able to get more information with 5 clusters. Hence let’s consider 5 clusters and plot the clusters to confirm if the derived clusters are providing the required segmentation details.
7. Importing fcluster modules to create clusters.
8. With Method 1 and Method 2 using Criterion as Maxclust and Distance respectively we got below result
9. 
10. So forming 5 clusters for Cleaning Ad Data will provide us with good insights.

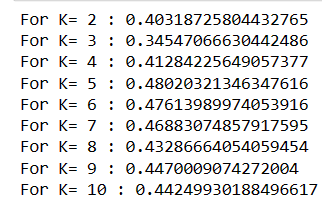
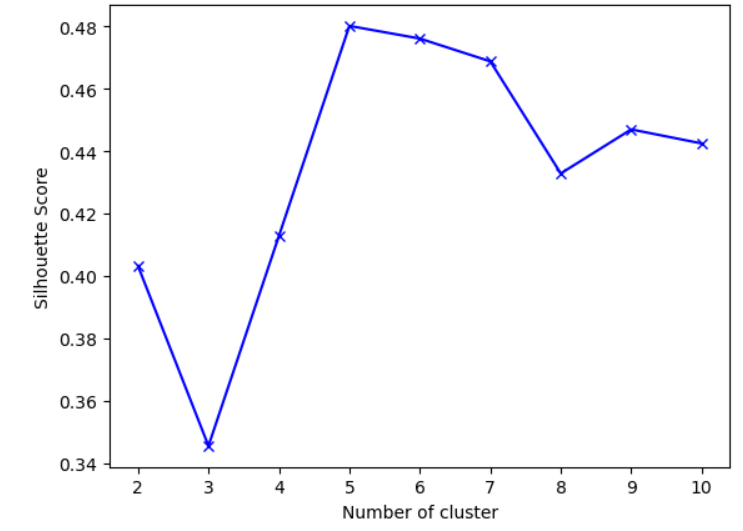
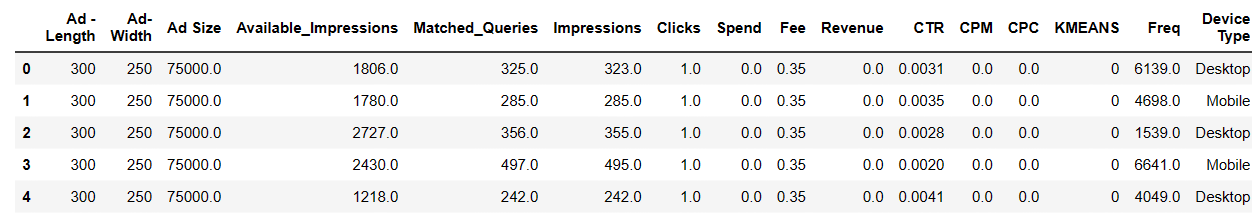
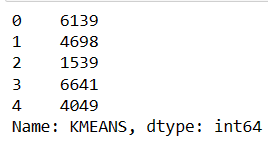
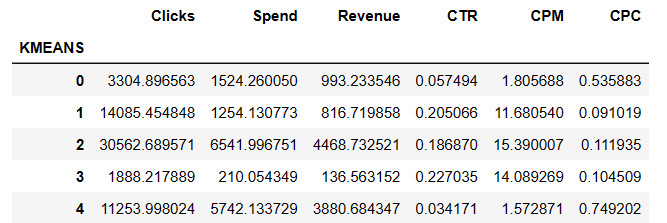
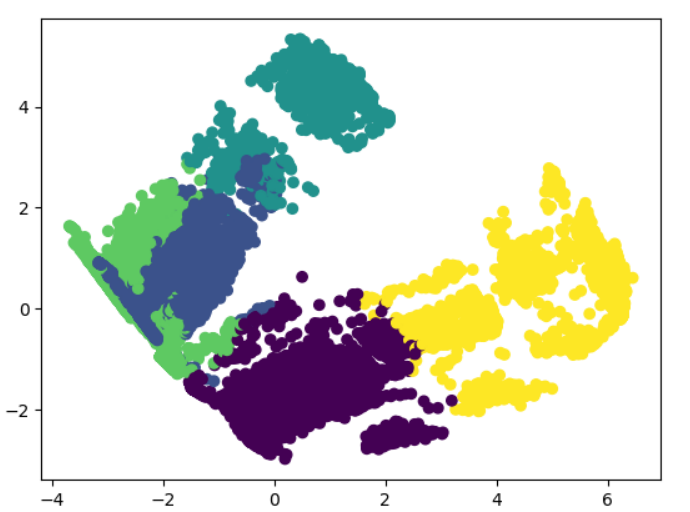
**Part 1 Clustering Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.(4 Marks)**

**Ans:**

1. Identifying Number of clusters using KMeans Algorithm.
2. Lets have a look at within sum of squres using Elbow plot for 10 clusters.
3. 
4. 
5. There is significant drop from 2 to 5. There are two elbows at 2 and 5.
6. From the graph it is clear that 5 is the optimal no of K value as after 5 there is no significant drop in the graph.
7. Lets use silhouette\_score to verify correct number of cluster.
8. WSS values also shows that after 5 the decrease in the value in not very visible.
9. 

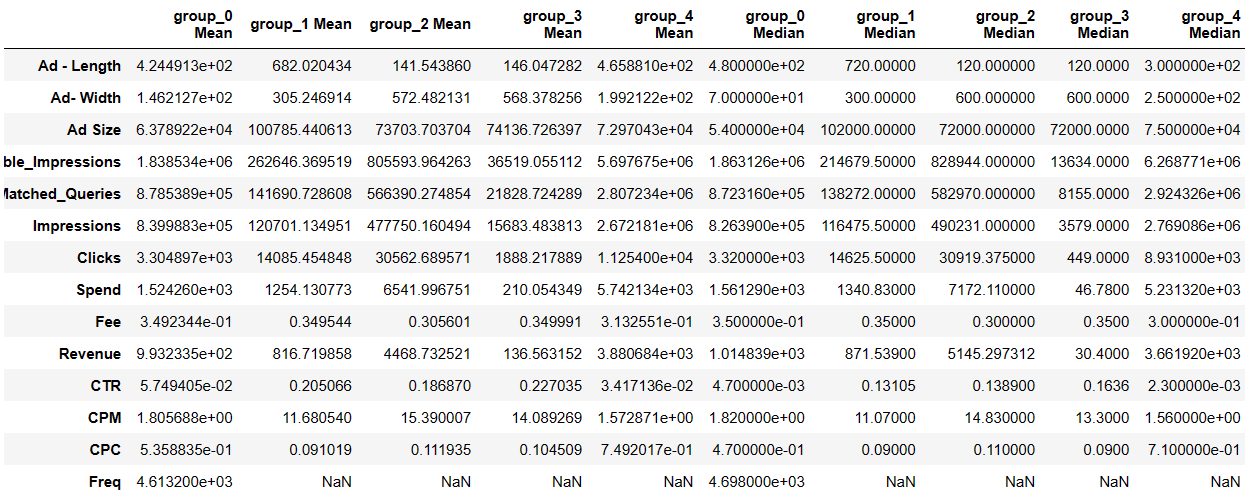
**Part 1: Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters.(4 Marks)**

**Ans:**

1. Importing silhouette\_samples and silhouette\_score
2. Lets get the silhouette\_score for upto 10 clusters
3. 
4. 
5. We can see from the graph that 5 has highest number of silhouette\_score.
6. Also Considering WSS we consider no of clusters from 2 to 6.
7. As above graph shows 5 with silhouette\_score so lets first consider 5 as number of cluster.
8. After trying 2,4 and 5 as number of cluster we have come to conclusion that 5 is the perfect number of cluster.
9. Lets add Labels to the original Dataframe and take mean of all the columns with respect to label column to check the overlapping.
10. 
11. 
12. There uniform distribution of data within each cluster. Cluster 2 has lowest distribution and Cluster 3 has highest distribution.
13. Clusters do not overlap each other to a greater extend.
14. 
15. 

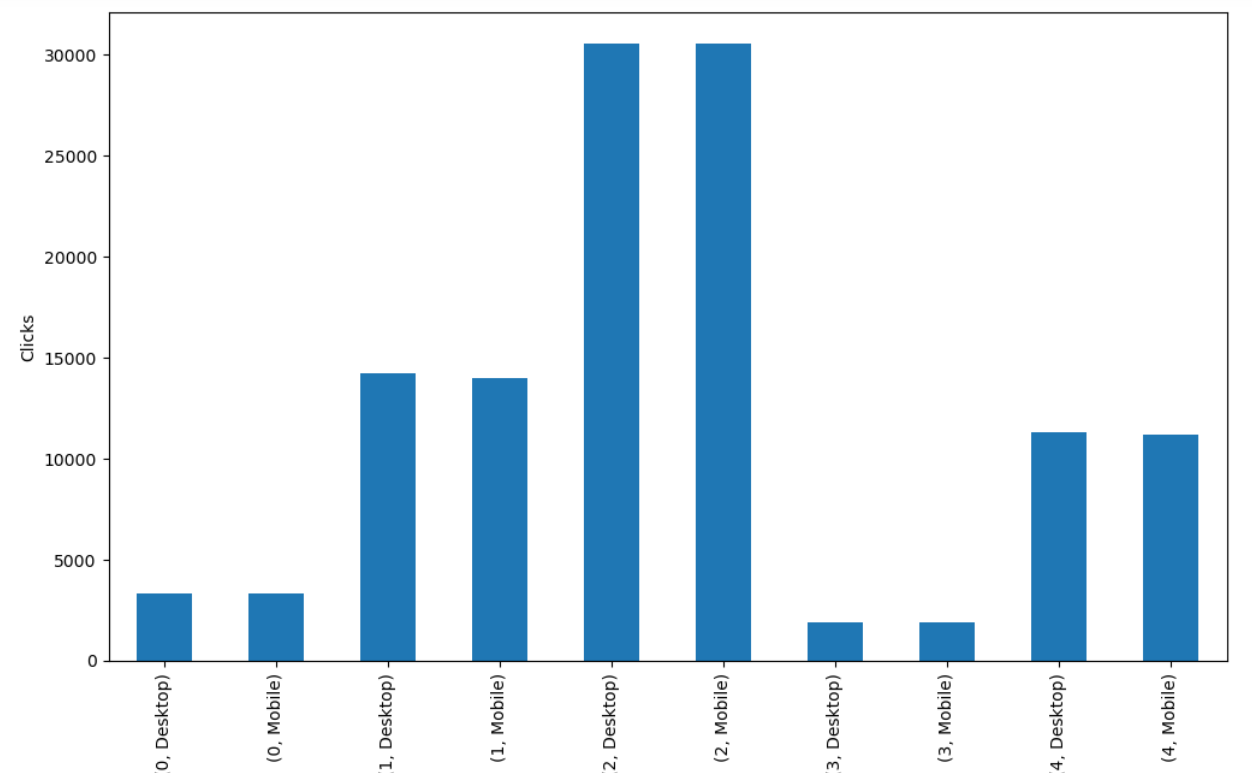
**Part : Clustering Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots].**

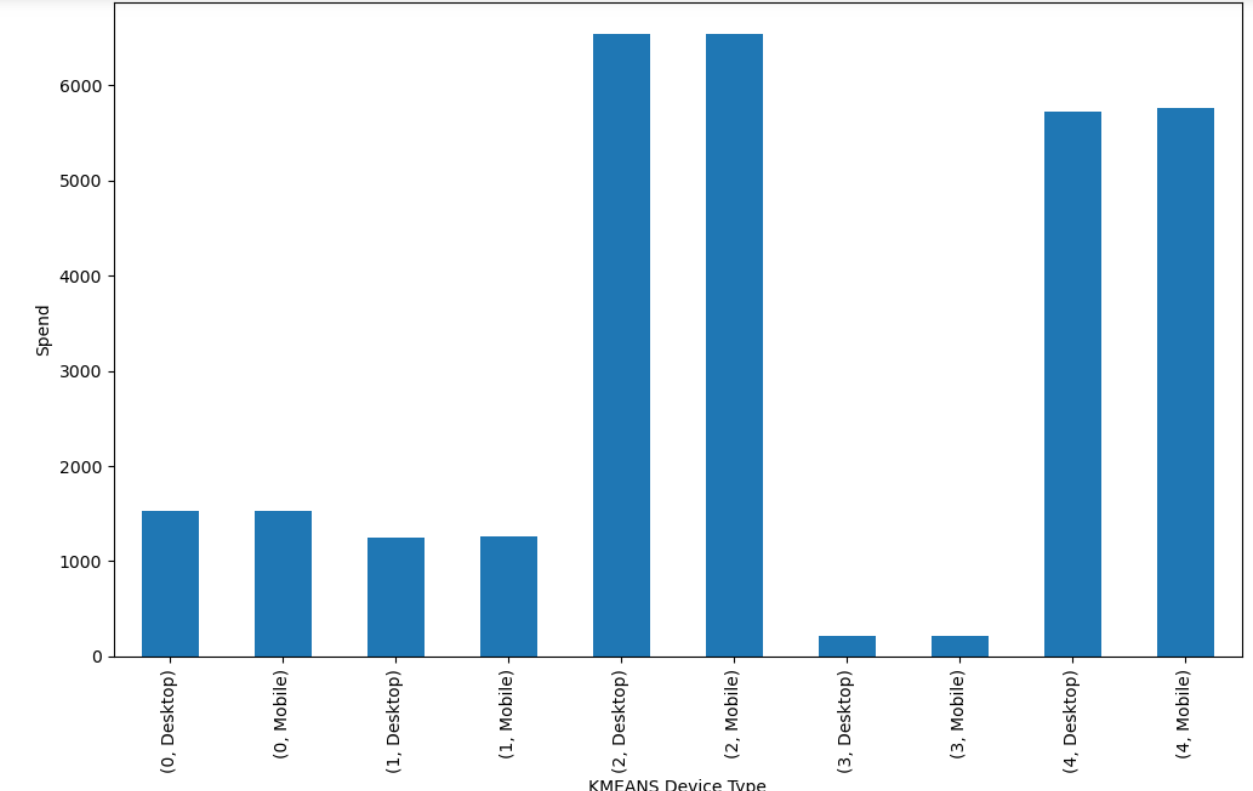
**Ans:**

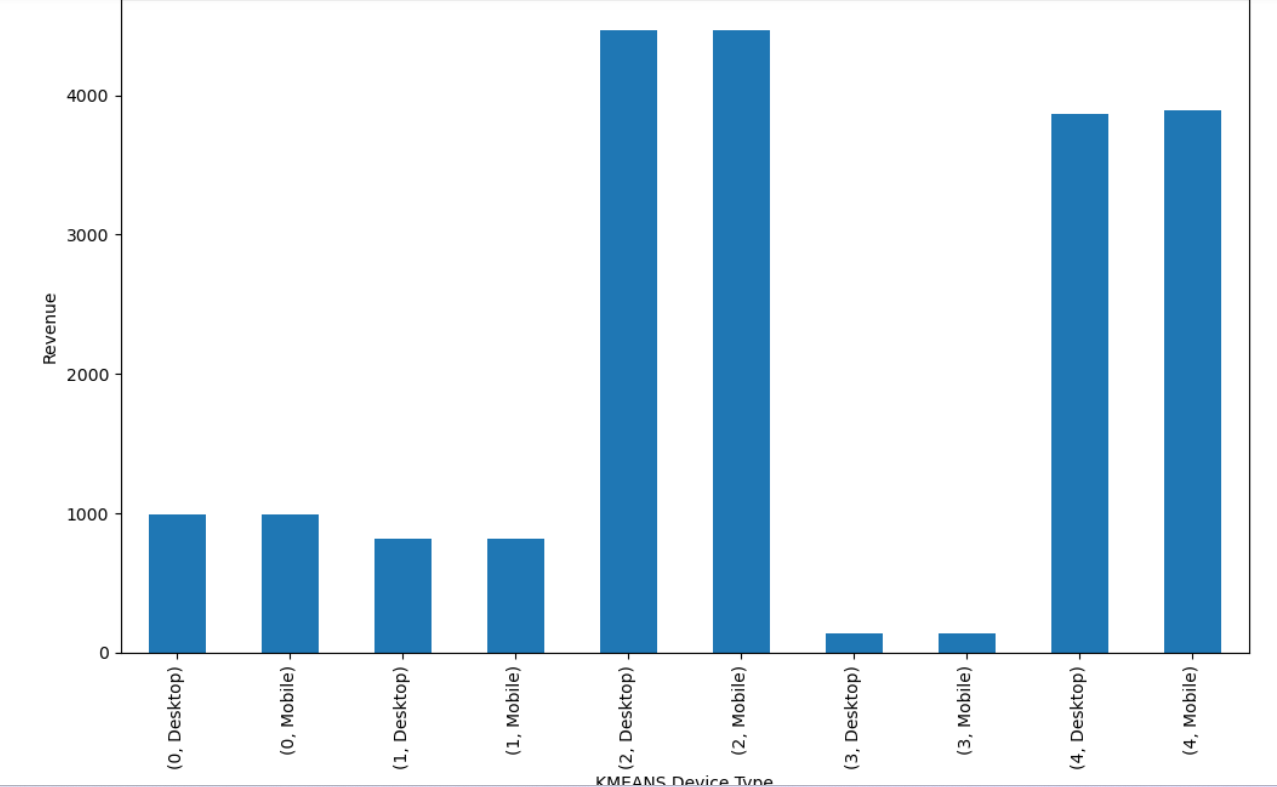
****

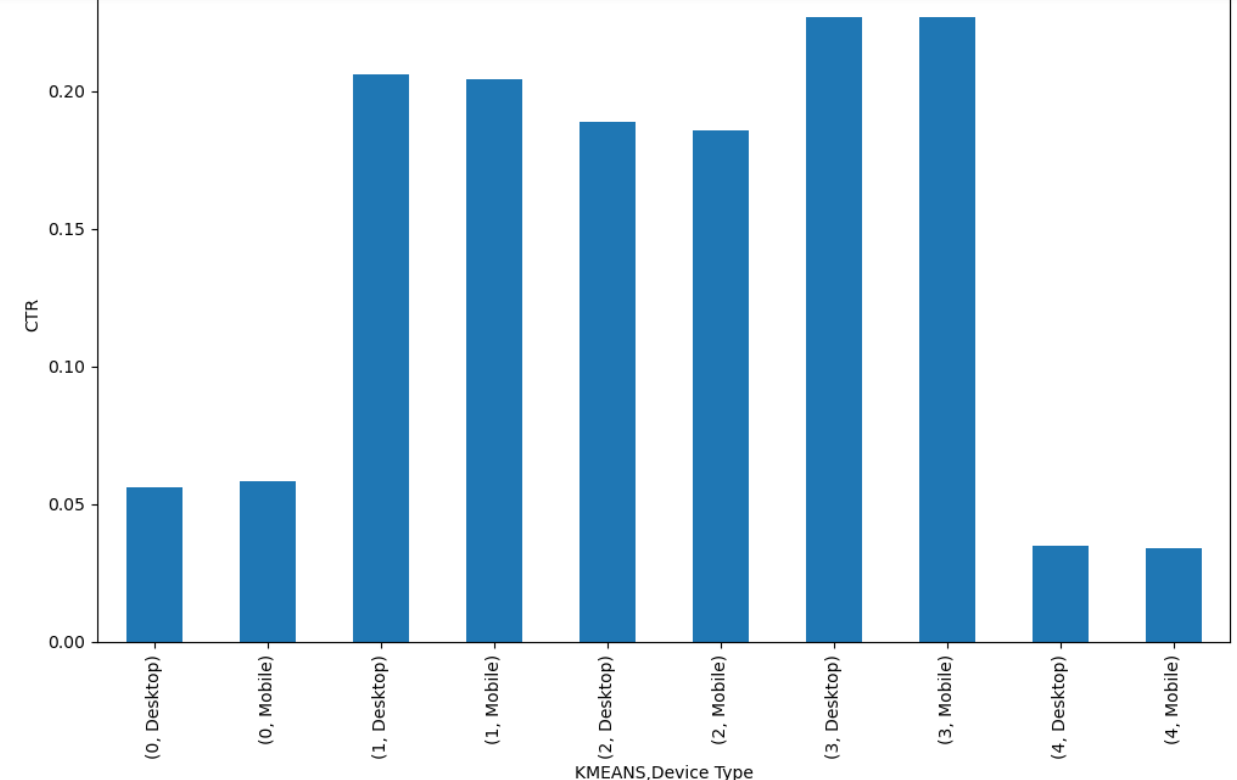
Cluster 2 has high revenue whereas cluster 4 has lowest revenue.

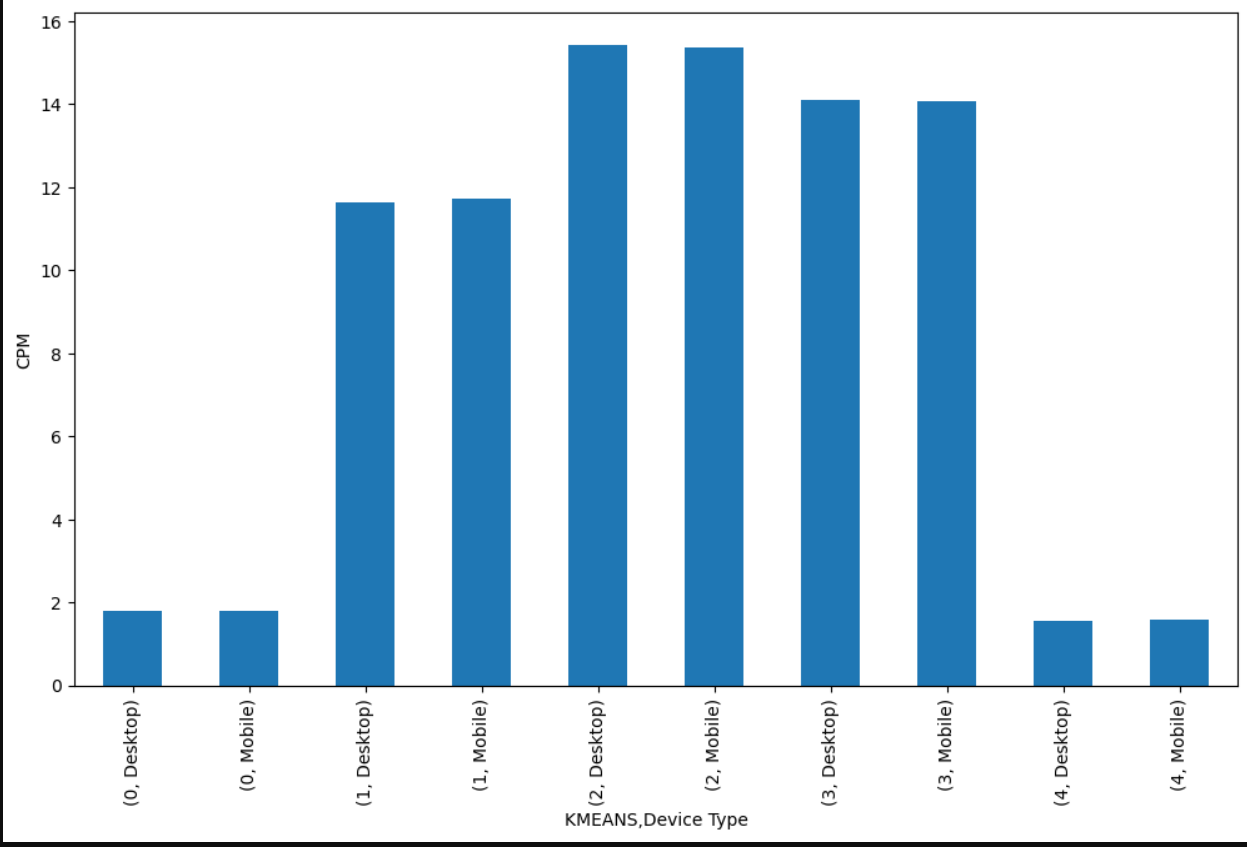
Cluster 3 is the biggest cluster as well.

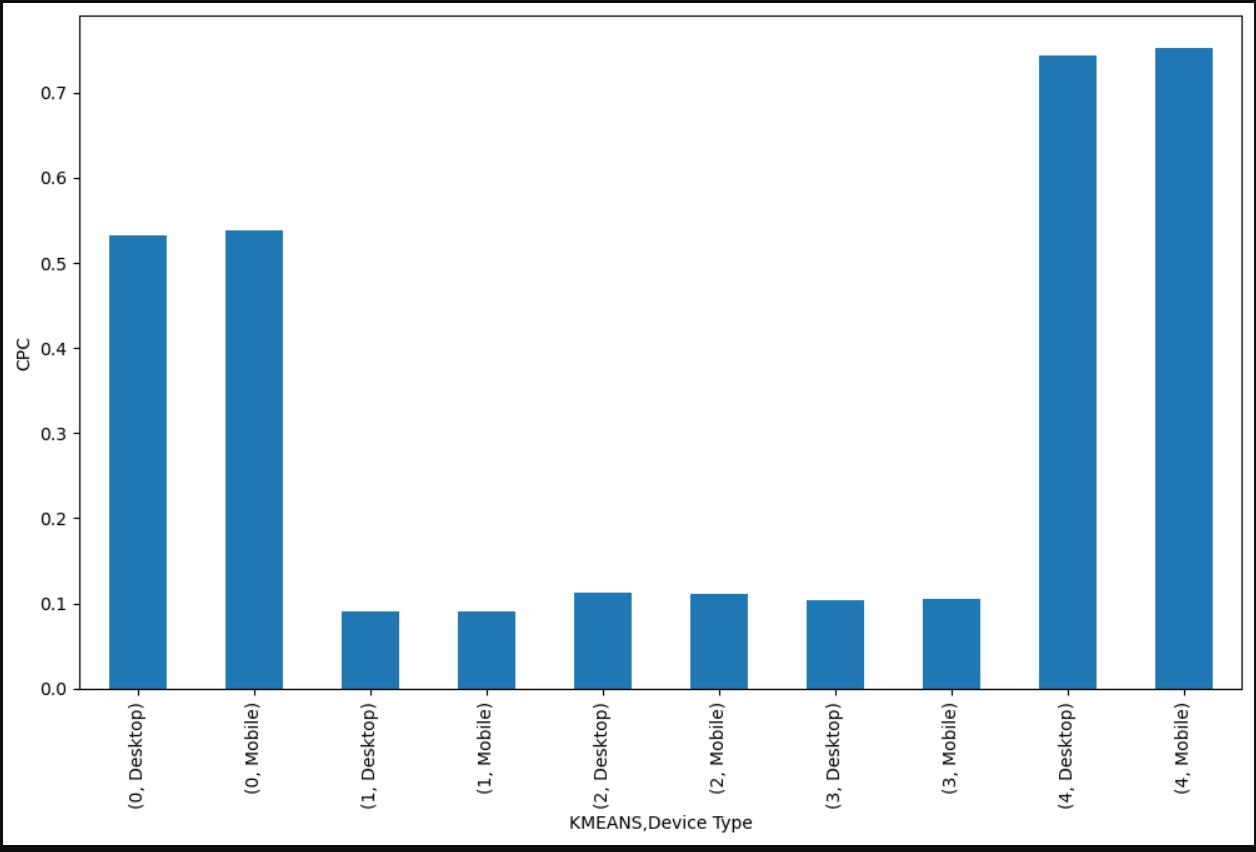




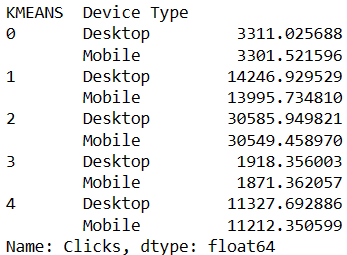
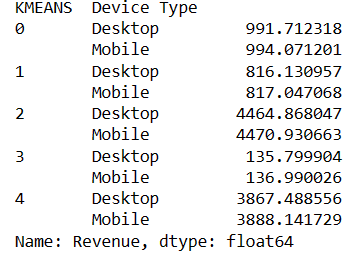
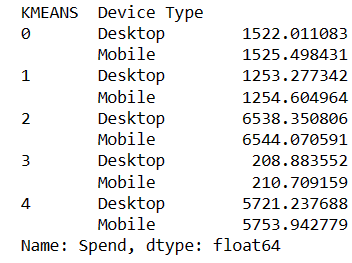
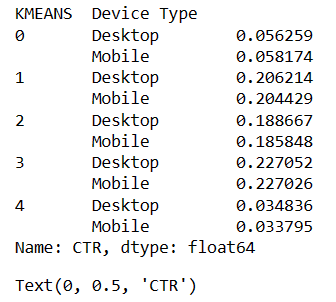
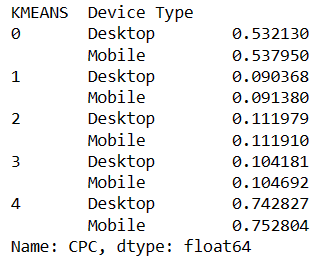
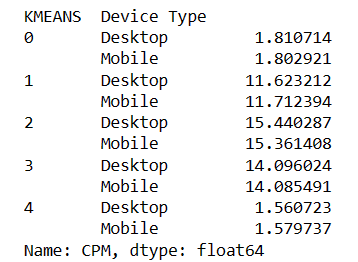








**Cluster Profiles:**

1. ****
2. **Cluster 2 has with device type Desktop and Mobile has highest number of clicks.This means that number of times user clicks on particular advertisement is more for Cluster 2 has is lowest for 3.**
3. ****
4. **Clusters 2 has highest number of Revenue whereas Cluster 3 has Lowest Revenue. Income Earned from Distribution of Cluster 2 is always highest than any other Cluster.**
5. ****
6. **There is same pattern going on with Spend similar to Clicks and Revenue. So We can say that Clicks, Revenue and Spend are positively correlated to each other.**
7. ****
8. **In this Cluster 3 has highest number of CTR followed by Cluster 1 and Cluster 2. Cluster 4 and 0 have very less number of CTR.**
9. ****
10. **In this case Cluster 4 has highest CPC values followed b Cluster 0 respectively.**
11. ****
12. **Cluster 3 and 2 have highest number of CPM values.Cluster 0 and Cluster covers almost similar values of CPM.**

**Part 1 - Clustering: Conclude the project by providing summary of your learnings(3 Marks)**

**Ans:-**

1. **There are 23066 rows and 19 columns in the dataset including Categorical columns.**
2. **There were few missing values in the column CTR,CPM and CPC which were replaced by using formulas given in the description.**
3. **Catergorical fields were removed before performing clustering on the data. So now we are left only 16 columns.**
4. **Outlier treatment was necessary in this case as we did scaling using Zscore technique. Zscore requires mean value which is affected by outliers.**
5. **Clustering is performed on the scaled data.**
6. **There are two types of clustering.**
7. **Hierarchichal clustering was done using ward linkage and Euclidean method.\**
8. **Non Hierarchichal clustering was done using KMeans Method.**
9. **For Hierarchichal Clustering the output is visualized by using Dendrogram.**
10. **By drawing the line at 220 on dendrogram we got 5 clusters which was again verified using Maxclust and distance method.**
11. **For now we can say that 5 is number of cluster.**
12. **Using KMeans Hierarchichal clustering was performed.**
13. **Within sum of square was calculated. We can see from the graph that WSS decreases as the number of clusters increased.**
14. **Using WSS we have got a lineplot using Elbow plot.**
15. **We can see that there are elbows at 2 and 5.**
16. **For now we have considered clusters from 2 to 5.**
17. **silhouette\_score is a method to find out the number of cluster.**
18. **We have plotted a graph using silhouette\_score which shows that 5 has highest number of silhouette\_score.**
19. **This concludes that 5 is the number of cluster we need.**
20. **We have also plotted graph showing relation between Kmeans labels value,Device Type and Spend,Revenue,Clicks,CTR,CPM and CPC.**
21. **Overall, this clustering solution does give us good insights into potential clusters.**
22. **Cluster 3 has highest Clicks , Revenue and Spend.**
23. **Cluster 1 has Lowest Clicks , Revenue and Spend..**
24. **CPR,CPM and CPC shows different trend along the cluster.**
25. **In other WSS is not significantly dropping beyond 5 to we say that 5 is number of cluster for this case.**