Clustering (Hierarchical and K-means)

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

The primary step is check a data in Head and tail to find the how many features, rows and columns present in dataset.

	Timestamp	InventoryType	Ad - Length	Ad-Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Sp
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	
2	2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	
3	2020-9-3-20	Format1	300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	
4	2020-9-4-15	Format1	300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	

	Timestamp	InventoryType	Ad - Length	Ad-Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Click
23061	2020-9-13-7	Format5	720	300	216000	Inter220	Web	Mobile	Video	1	1	1	
23062	2020-11-2-7	Format5	720	300	216000	Inter224	Web	Desktop	Video	3	2	2	
23063	2020-9-14-22	Format5	720	300	216000	Inter218	App	Mobile	Video	2	1	1	
23064	2020-11-18-2	Format4	120	600	72000	inter230	Video	Mobile	Video	7	1	1	
23065	2020-9-14-0	Format5	720	300	216000	Inter221	App	Mobile	Video	2	2	2	

There is no null in the dataset and 6 float, 7 integers and 6 object are presented

```
RangeIndex: 23066 entries, 0 to 23065
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
dtypes: float64(6), int64(7), object(6)
memory usage: 3.3+ MB
```

In our Dataset have 4736 zero values in CPR and CTR and CPM.

```
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
dtype: int64
```

Treat the missing values in CPC, CTR and CPM through formulas

CTR=Clicks/Impression*100, CPM=Spend/Impression*1000, CPC=Spend/clicks

CTR:

We treat the missing values through the lambda function.

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	300	250	75000	1806	325	323	1	0.00	0.35	0.0000	0.309598	0.0	0.0
1	300	250	75000	1780	285	285	1	0.00	0.35	0.0000	0.350877	0.0	0.0
2	300	250	75000	2727	356	355	1	0.00	0.35	0.0000	0.281690	0.0	0.0
3	300	250	75000	2430	497	495	1	0.00	0.35	0.0000	0.202020	0.0	0.0
4	300	250	75000	1218	242	242	1	0.00	0.35	0.0000	0.413223	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
23061	720	300	216000	1	1	1	1	0.07	0.35	0.0455	100.000000	NaN	NaN
23062	720	300	216000	3	2	2	1	0.04	0.35	0.0260	50.000000	NaN	NaN
23063	720	300	216000	2	1	1	1	0.05	0.35	0.0325	100.000000	NaN	NaN
23064	120	600	72000	7	1	1	1	0.07	0.35	0.0455	100.000000	NaN	NaN
23065	720	300	216000	2	2	2	1	0.09	0.35	0.0585	50.000000	NaN	NaN

23066 rows × 13 columns

CPM

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
73	300	250	75000	1789	306	300	1	0.00	0.35	0.00	0.333333	0.000000	NaN
74	300	250	75000	1784	274	274	1	0.00	0.35	0.00	0.364964	0.000000	NaN
75	300	250	75000	2095	379	375	1	0.00	0.35	0.00	0.266667	0.000000	NaN
76	300	250	75000	1800	308	307	1	0.00	0.35	0.00	0.325733	0.000000	NaN
77	300	250	75000	1732	275	273	1	0.00	0.35	0.00	0.366300	0.000000	NaN
78	300	250	75000	1594	250	247	1	0.00	0.35	0.00	0.404858	0.000000	NaN
79	300	250	75000	1605	215	213	1	0.00	0.35	0.00	0.469484	0.000000	NaN
80	300	250	75000	631	67	65	1	0.00	0.35	0.00	1.538462	0.000000	NaN
81	300	250	75000	1671	373	372	1	0.00	0.35	0.00	0.268817	0.000000	NaN
82	300	250	75000	1424	274	269	1	0.00	0.35	0.00	0.371747	0.000000	NaN
83	300	250	75000	1619	214	211	1	0.00	0.35	0.00	0.473934	0.000000	NaN
84	300	250	75000	1775	349	348	1	0.00	0.35	0.00	0.287356	0.000000	NaN
85	300	250	75000	1888	376	368	1	0.00	0.35	0.00	0.271739	0.000000	NaN
86	300	250	75000	1842	286	285	1	0.00	0.35	0.00	0.350877	0.000000	NaN
87	300	250	75000	1970	359	355	1	0.00	0.35	0.00	0.281690	0.000000	NaN
88	300	250	75000	1914	472	465	1	0.00	0.35	0.00	0.215054	0.000000	NaN

CPC:

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
73	300	250	75000	1789	306	300	1	0.00	0.35	0.00	0.333333	0.000000	0.00
74	300	250	75000	1784	274	274	1	0.00	0.35	0.00	0.364964	0.000000	0.00
75	300	250	75000	2095	379	375	1	0.00	0.35	0.00	0.266667	0.000000	0.00
76	300	250	75000	1800	308	307	1	0.00	0.35	0.00	0.325733	0.000000	0.00
77	300	250	75000	1732	275	273	1	0.00	0.35	0.00	0.366300	0.000000	0.00
78	300	250	75000	1594	250	247	1	0.00	0.35	0.00	0.404858	0.000000	0.00
79	300	250	75000	1605	215	213	1	0.00	0.35	0.00	0.469484	0.000000	0.00
80	300	250	75000	631	67	65	1	0.00	0.35	0.00	1.538462	0.000000	0.00
81	300	250	75000	1671	373	372	1	0.00	0.35	0.00	0.268817	0.000000	0.00
82	300	250	75000	1424	274	269	1	0.00	0.35	0.00	0.371747	0.000000	0.00
83	300	250	75000	1619	214	211	1	0.00	0.35	0.00	0.473934	0.000000	0.00
84	300	250	75000	1775	349	348	1	0.00	0.35	0.00	0.287356	0.000000	0.00
85	300	250	75000	1888	376	368	1	0.00	0.35	0.00	0.271739	0.000000	0.00
86	300	250	75000	1842	286	285	1	0.00	0.35	0.00	0.350877	0.000000	0.00

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).

Treating outliers is necessary for Kmeans clustering because the main reasons are **sensitive to outliers**. It uses the mean to calculate cluster centroids. outliers can pull the centroids away from the true centre of cluster. which might be inaccurate results.

Outliers can form the own cluster or distort the shape and size of the existing cluster. It's challenging to interpret the cluster

Outliers can affect your Scaling of your data. Hence we need to normalize your data.

This is the main reason to treat the Outliers.

But this project we don’t treat the outlier because it affect my scaling.

	Ad - Length	Ad-Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	-0.3645	-0.4328	-0.3522	-0.5124	-0.5152	-0.5109	-0.6153	-0.6654	0.4654	-0.6197	-0.8746	-0.9271	-0.9866
1	-0.3645	-0.4328	-0.3522	-0.5124	-0.5153	-0.5109	-0.6153	-0.6654	0.4654	-0.6197	-0.8701	-0.9271	-0.9866
2	-0.3645	-0.4328	-0.3522	-0.5122	-0.5152	-0.5109	-0.6153	-0.6654	0.4654	-0.6197	-0.8776	-0.9271	-0.9866
3	-0.3645	-0.4328	-0.3522	-0.5123	-0.5152	-0.5108	-0.6153	-0.6654	0.4654	-0.6197	-0.8862	-0.9271	-0.9866
4	-0.3645	-0.4328	-0.3522	-0.5125	-0.5153	-0.5110	-0.6153	-0.6654	0.4654	-0.6197	-0.8634	-0.9271	-0.9866
...	...	...	...	...	...	...	...	...	...	...	...	...	...
23061	1.4331	-0.1866	1.9391	-0.5128	-0.5154	-0.5111	-0.6153	-0.6654	0.4654	-0.6197	9.8890	6.8013	-0.7815
23062	1.4331	-0.1866	1.9391	-0.5128	-0.5154	-0.5110	-0.6153	-0.6654	0.4654	-0.6197	4.4905	1.2810	-0.8694
23063	1.4331	-0.1866	1.9391	-0.5128	-0.5154	-0.5111	-0.6153	-0.6654	0.4654	-0.6197	9.8890	4.5932	-0.8401
23064	-1.1349	1.2906	-0.4010	-0.5128	-0.5154	-0.5111	-0.6153	-0.6654	0.4654	-0.6197	9.8890	6.8013	-0.7815
23065	1.4331	-0.1866	1.9391	-0.5128	-0.5154	-0.5110	-0.6153	-0.6653	0.4654	-0.6197	4.4905	4.0412	-0.7229

23066 rows × 13 columns

Ad- length, Ad-Width and Ad-size:

- This metrics have been assessed the campaigns impact and reach.

Other metrics is to understand the user engagement and financial performance.

CTR (Click through Rate)

- The CTR standard deviations are above 2 refers that the audience segment is highly responsive to the ads.

CPM: Standard Deviations are above at 1 so this is also high responsive to ads

CPC: This metrics is a almost normal because the Standard deviation are -0.78 to -0.98(It's Normal)

Lower side: In **Impressions** and **Clicks** are column are low because the standard deviations are below -0.5. so it as lower response from ad-side.

Matched queries, Ad-Width is lower response (Below -0.4 to -0.1).

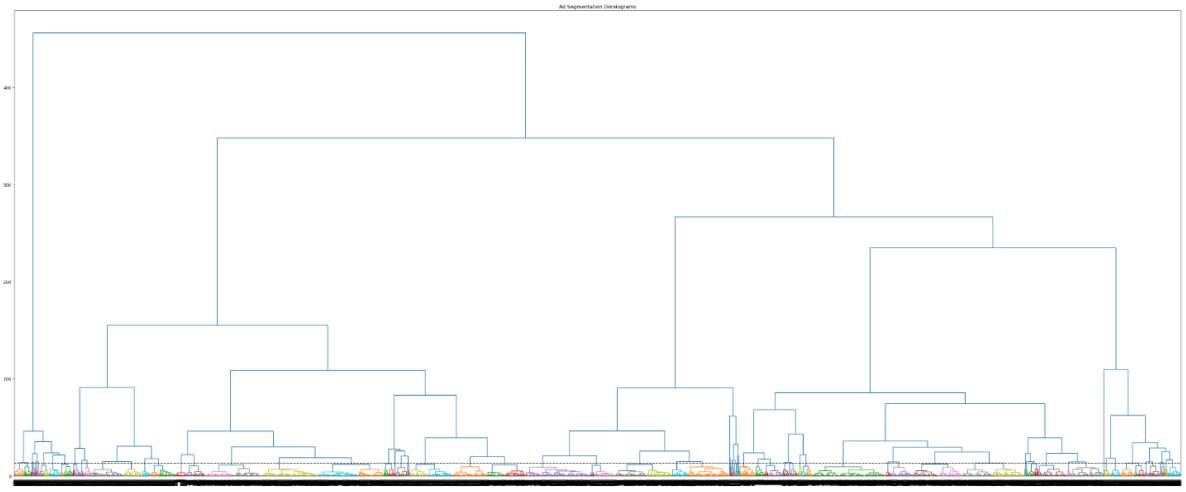
Spend: It's also the -0.66 STD so it's a lower side.

It can affect the speed of the algorithm

The factors are faster convergence, Efficient Distance calculation.

Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance

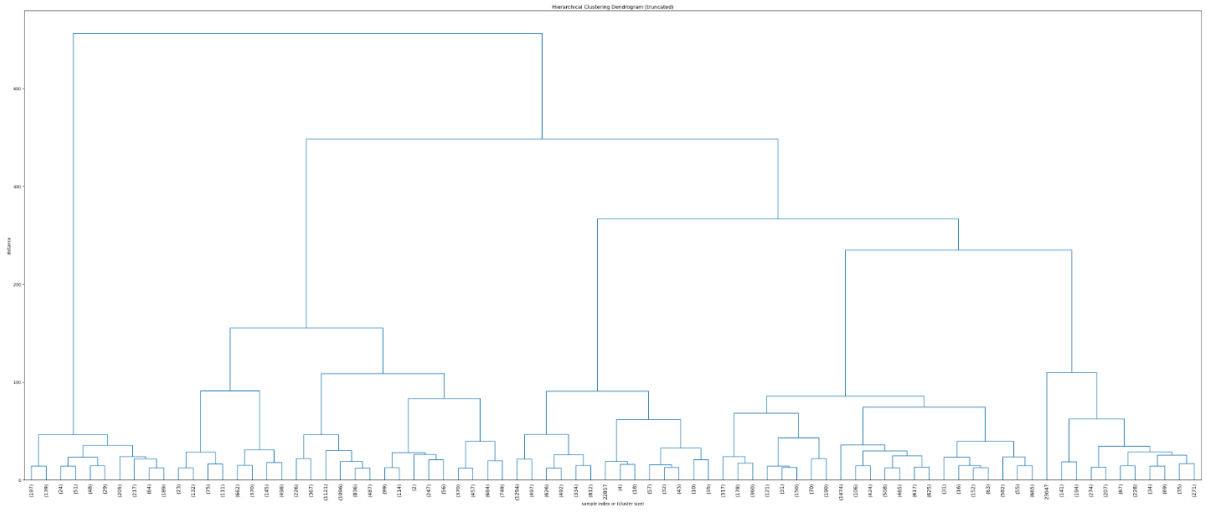
In ward,



This is a entire dataset which I ran through **ward method.**

I plotted the Euclidean method. I selected the last 80 in truncated mode in x and y label.
The intra is minimized and the inter is maximized.

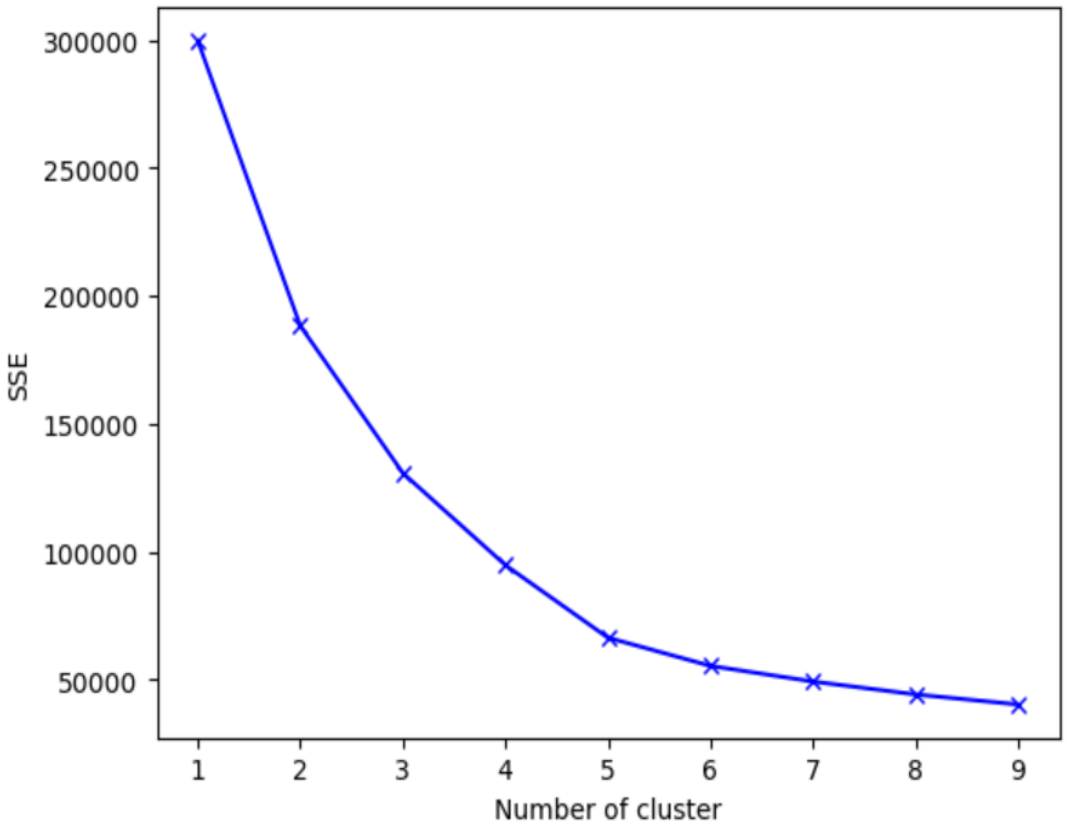
Euclidean method is a familiar rather than manhattan and chebyshev because the penality is higher, if I take modulus (manhattan) function which might not give a greater penality that ‘s why chose the Euclidean distance (main thing is squared the function)



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

The SSE Vs Number of cluster

The No of Cluster is increases, SSE is reduced that is no error.
We can see the diagram for each and every cluster like 1,2,3,4,5,6,7,8



The above diagram will reduce the SSE also the cluster will increase but it doesn't clear to tell about which **cluster is high**.

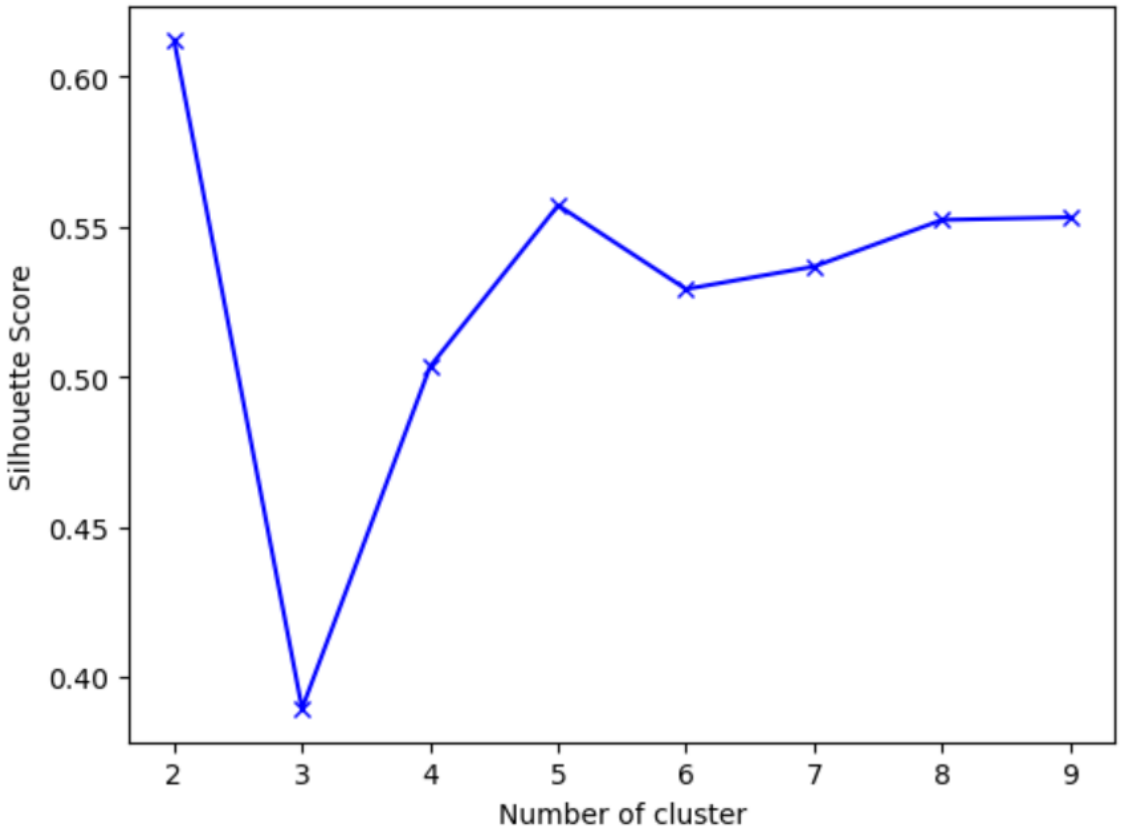
Print silhouette scores for up to 10 clusters and identify optimum number of clusters.

In **Silhouette** method, we can say which is high.
The cluster is the high value when compare to the clusters in the below silhouette graph,
So I can take the Cluster for
The silhouette formula is

S=b-a/max(a, b)

Why we need to put Silhouette's?

In below graph, Cluster 2 is the high that means well-clustered object. Usually the high average silhouette's score are indicated that object is well- clustered.



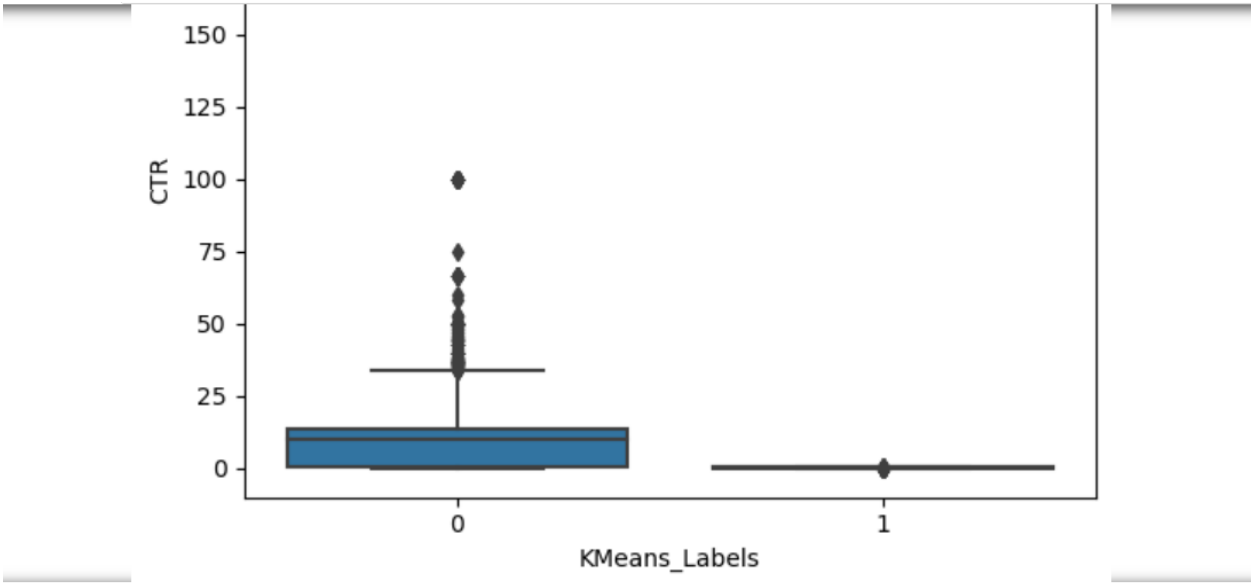
Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots].

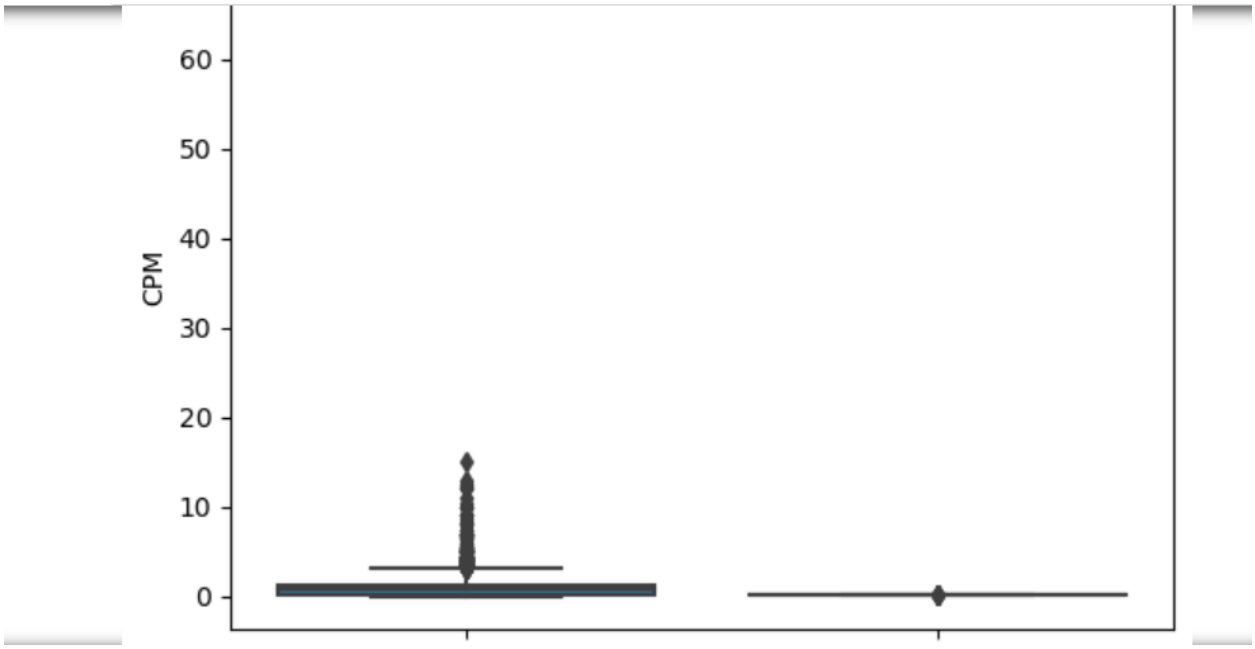
Next we predict the label and attach to the original data. In this original data, some skewed cluster is few observation, 21580 has present in cluster 0 out of 23066 entries.

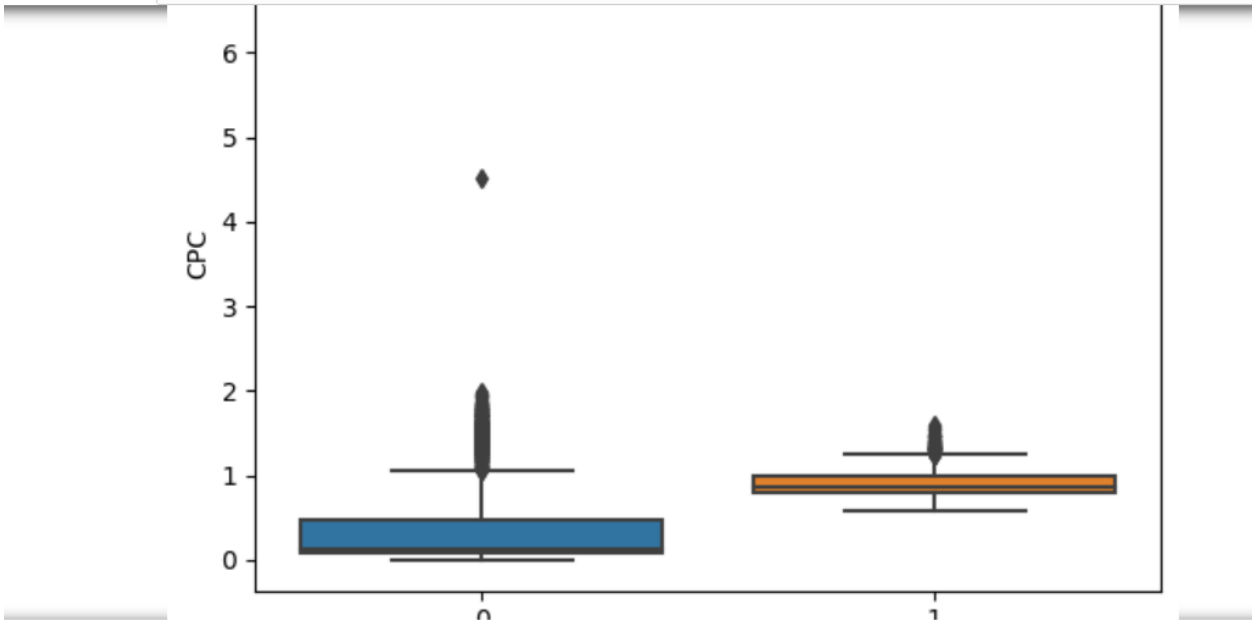
Visualization part : Box Plot

In this below chart the CTR plot is high visualization in below box plot.
In CPM also good to see in cluster 1
I digital marketing

A High CTR refers to that large percentage of users who see the Ad clicks. It is Positive sign the ad is engaging and relevant to the audience. The result is engaged in relevant traffic for high CTR.

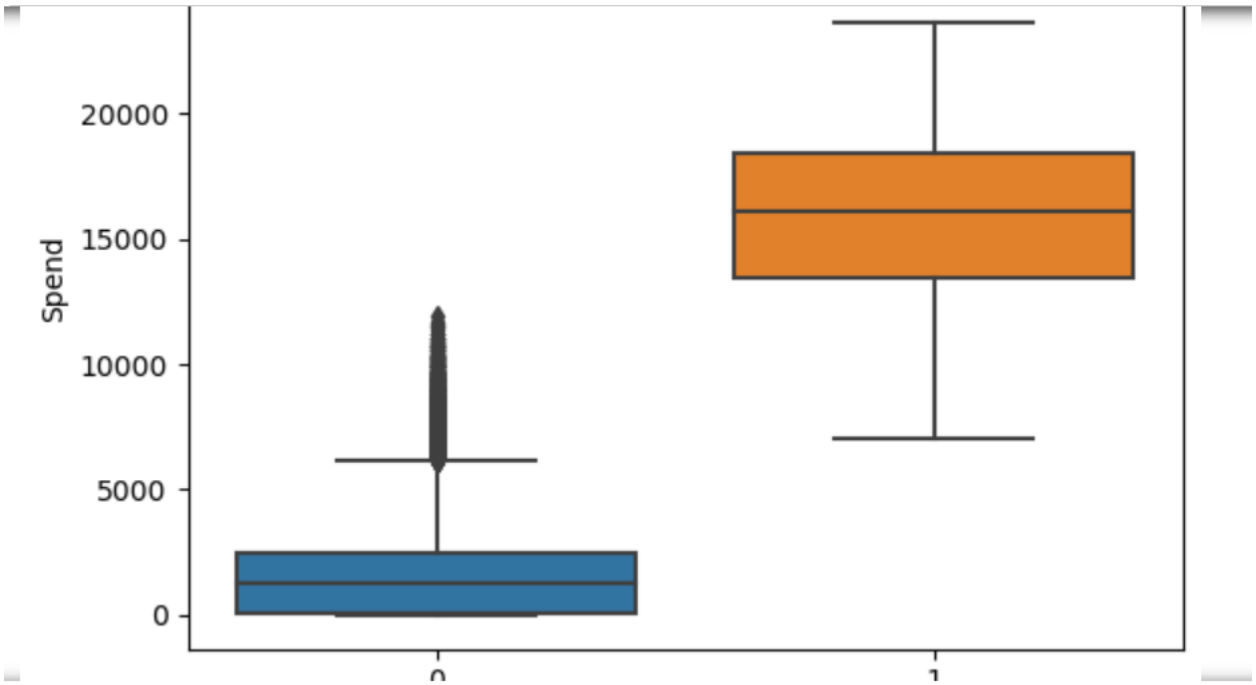


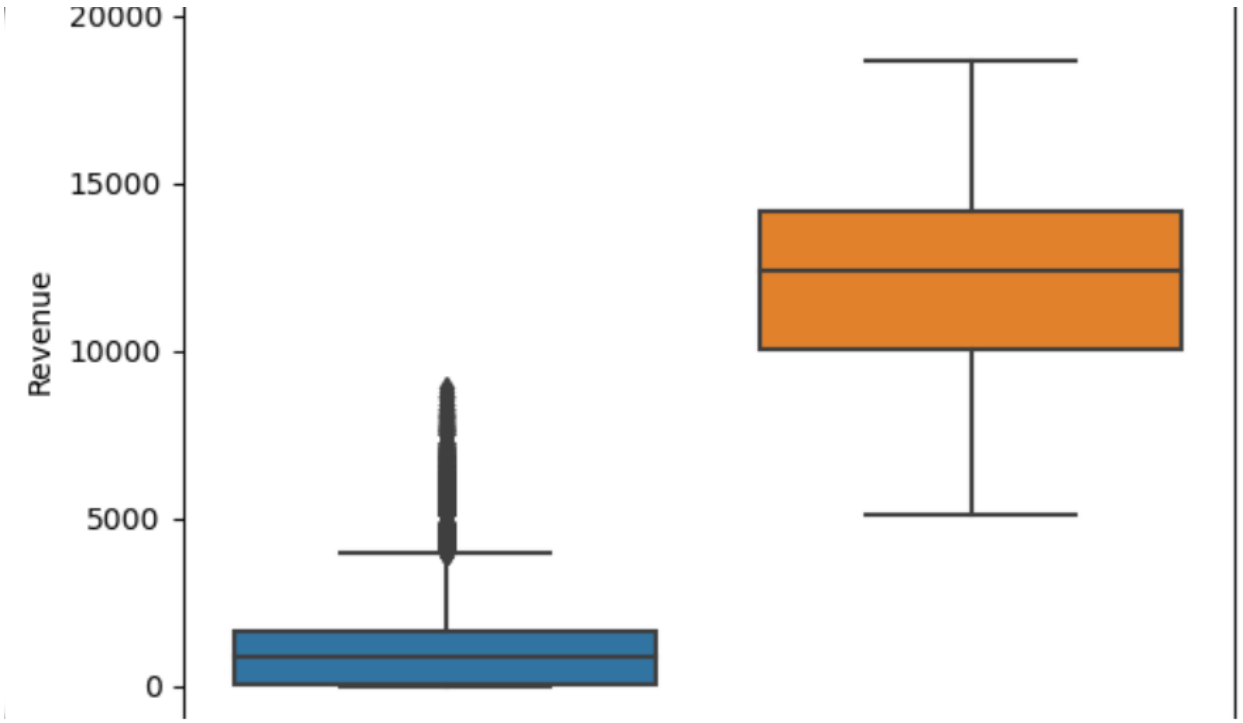

CPM
60
50
40
30
20
10
0


CPC
6
5
4
3
2
1
0
0
1

Spend and Revenue:

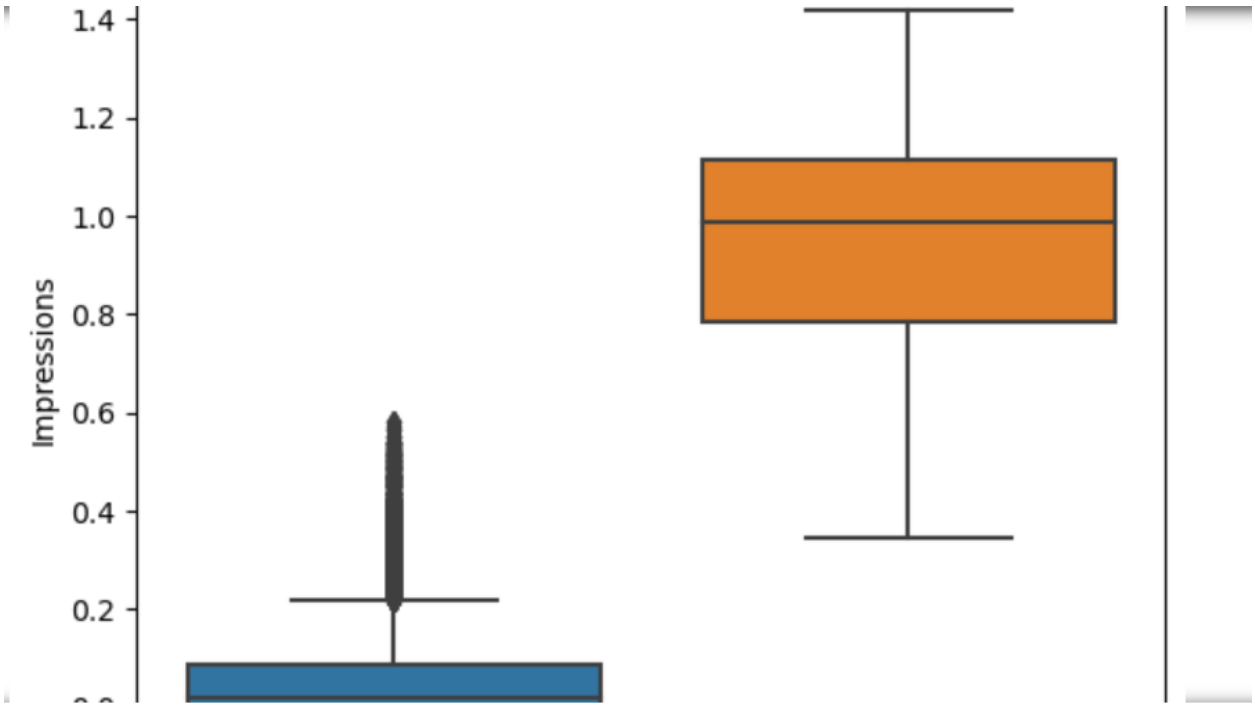
In this Ad content, the company spent so much of amount and it got high revenue.

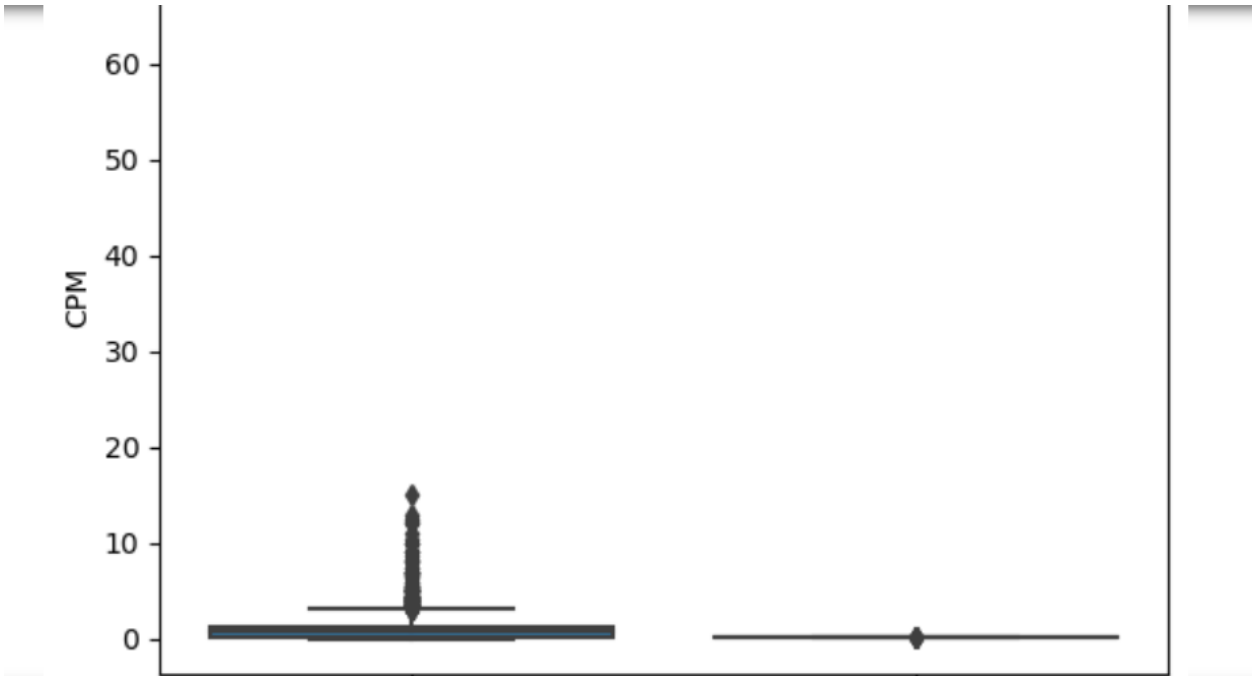




CPM and Impressions:

We can See the Impressions are low.
In this concept, the limited reach can affect brand awareness and reduce the chances of attracting potential customers.





Conclusion:

In summary, low impressions can impact **CPM** and have significant business implications, including reduced brand reach, limited data for advertisers, lower revenue, and decreased competition among advertisers. 24*7 often aim to increase impressions to provide more value to advertisers and improve their own financial viability.