Report

Part-1:

Solution-1:

* Data Summary:

• There are 23066 Rows and 19 columns.

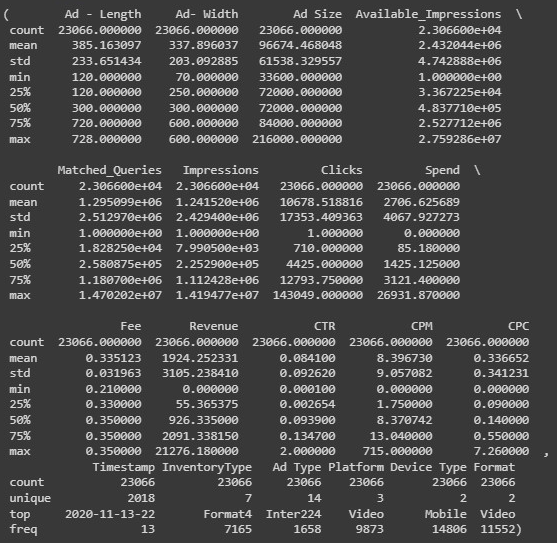
• There are Object, Float and Integer Datatypes.

• There are missing values under CPM (4736), CPC (4736) and CTR (4736) which we will fill using the given formulas.

• There are no Duplicate values in the data.

• Divided the columns as numerical and categorical for further analysis.

• Basic statistical summary:



Solution-2:

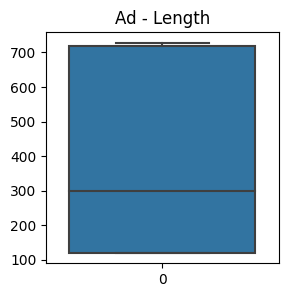
• Using below formulas, missing values for CPC, CTR and CPM were imputed

• CPM = (Total Campaign Spend / Number of Impressions) \* 1,000

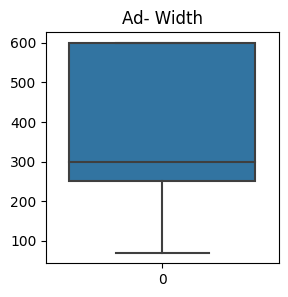
• CPC = Total Cost (spend) / Number of Clicks

• CTR = Total Measured Clicks / Total Measured Ad Impressions

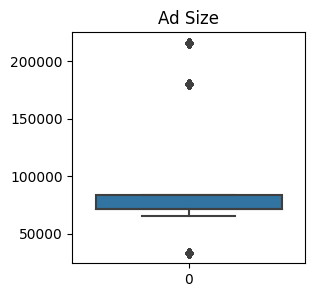
Solution-3:



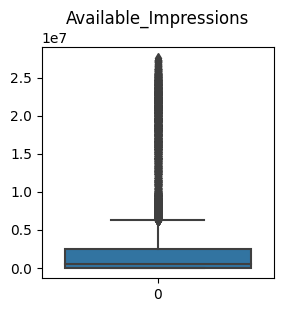
total number of outliers in Ad - Length is 0



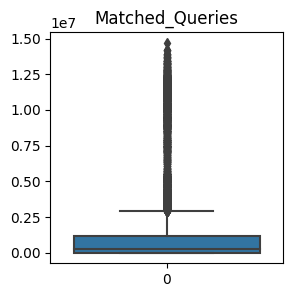
total number of outliers in Ad- Width is 0



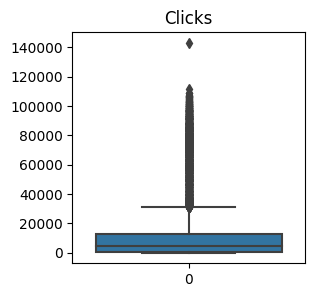
total number of outliers in Ad Size is 8448



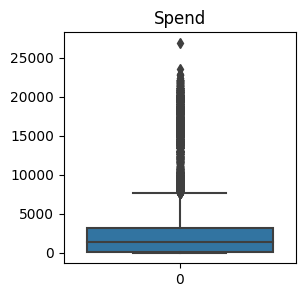
total number of outliers in Available\_Impressions is 3622



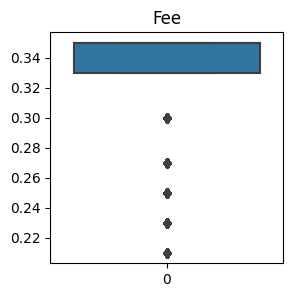
total number of outliers in Impressions is 4038



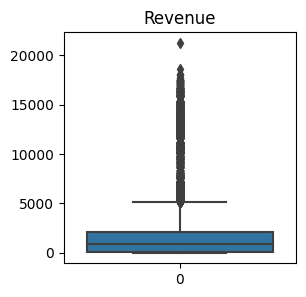
total number of outliers in Clicks is 2953



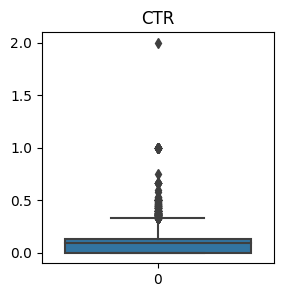
total number of outliers in Spend is 3092



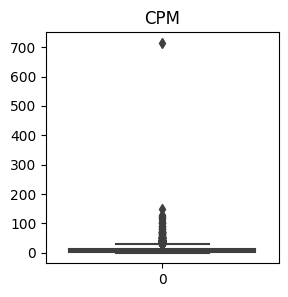
total number of outliers in Fee is 3517



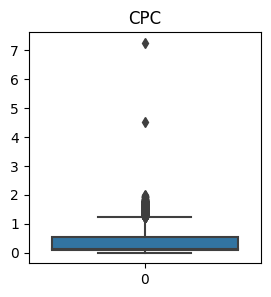
total number of outliers in Revenue is 3320



total number of outliers in CTR is 1229



total number of outliers in CPM is 814

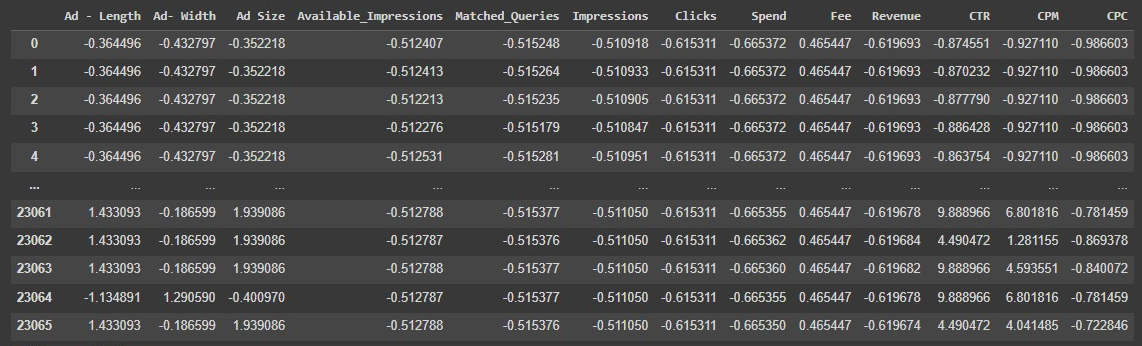


total number of outliers in CPC is 1353.

* As we can observe from above visualization there are outliers in every numerical column except 'Ad-width' and 'Ad-lenght'.
* And there are 8448 outliers in 'Ad-size' column.
* We can perform K-Means clustering both with and without outliers, and compare the results. If outliers are causing cluster centroids to be pulled away from the main cluster structure, it might be a good idea to address them and using clustering algorithms that are less sensitive to outliers, such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) or hierarchical clustering.

Solution-4:

* Scaled data:



Scaling the data can have a significant impact on the speed of clustering algorithms. The effect of scaling depends on the specific algorithm, the scaling method used

1. **Convergence Speed:** Clustering algorithms often rely on distance calculations between data points or cluster centroids. Scaling the data ensures that features are on similar scales, which can lead to faster convergence. Algorithms like K-Means converge faster when the distances between data points are comparable, resulting in quicker updates of cluster centroids.

2. **Distance Calculations:** Many clustering algorithms involve distance-based calculations. Scaling the data can make these calculations more efficient. When features are on the same scale, distance computations become simpler and require fewer floating-point operations, which can speed up the algorithm's execution.

3. **Numerical Stability:** Scaling the data can improve numerical stability by reducing the risk of numerical precision issues. Algorithms might behave more consistently and converge faster when the data is scaled, particularly in cases where the algorithm involves iterative updates.

4. **Outlier Impact:** Scaling can reduce the impact of outliers on clustering results. Outliers can significantly affect the distances between data points, leading to slower convergence or the formation of distorted clusters. Scaling makes the algorithm less sensitive to extreme values.

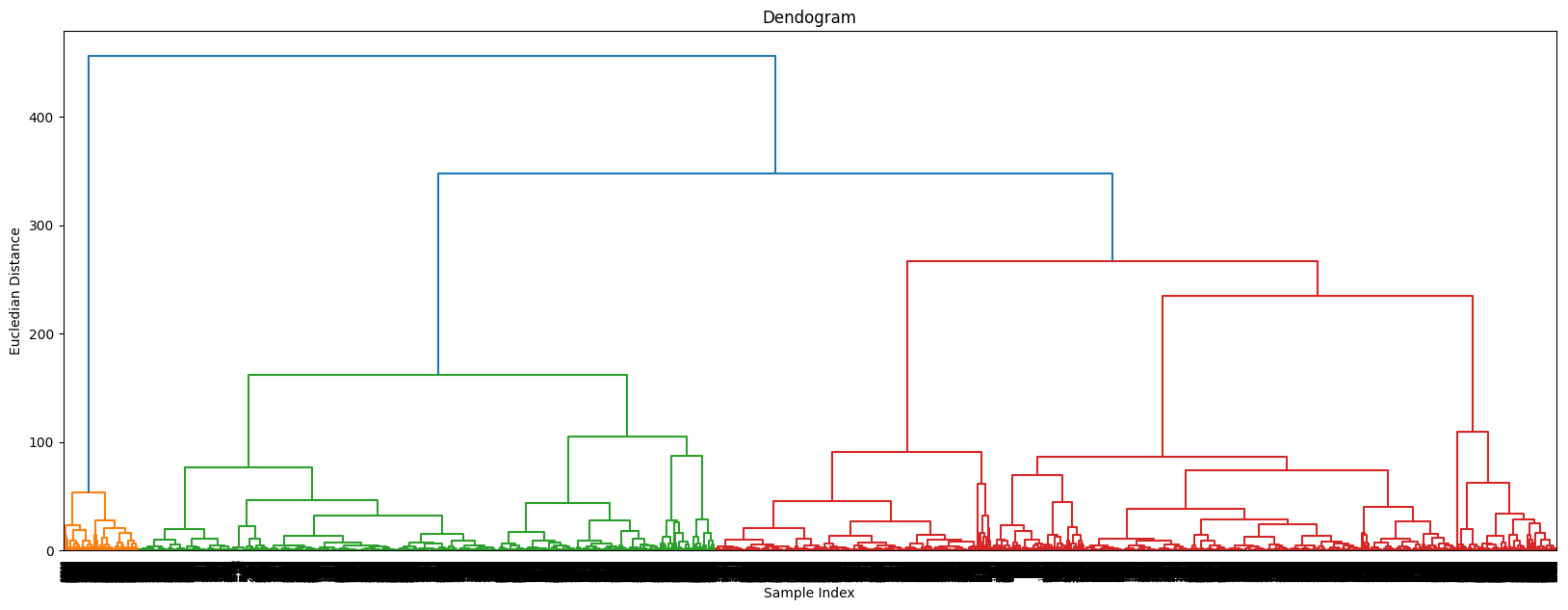
5. **Efficient Memory Usage:** Algorithms that involve large distance matrices or similarity matrices (e.g., hierarchical clustering) can benefit from scaling. Smaller feature ranges lead to more efficient memory usage and faster matrix computations.

6. **Algorithm Sensitivity:** Some clustering algorithms are more sensitive to feature scaling than others. For example, K-Means and hierarchical clustering often benefit from scaling, while density-based algorithms like DBSCAN are less sensitive to scaling due to their local density approach.

7. **Preprocessing Overhead:** While scaling can improve algorithm speed, there might be a slight overhead associated with data preprocessing. However, this preprocessing overhead is usually small compared to the potential speed improvements gained during algorithm execution.

Solution-5:

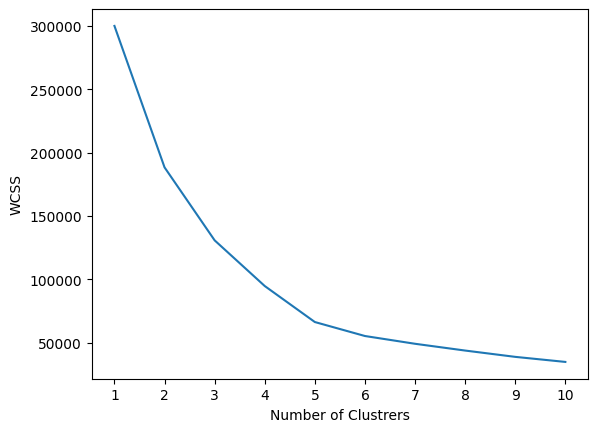
* Dendogram:



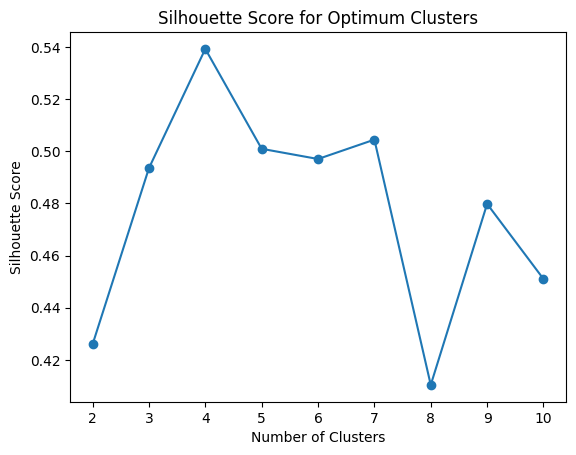
* From dendogram optimal number of clusters is found to be: 5

Solution-6:

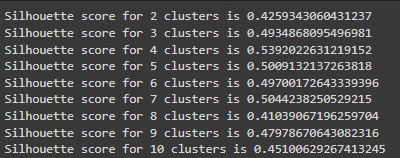
* Elbow plot:



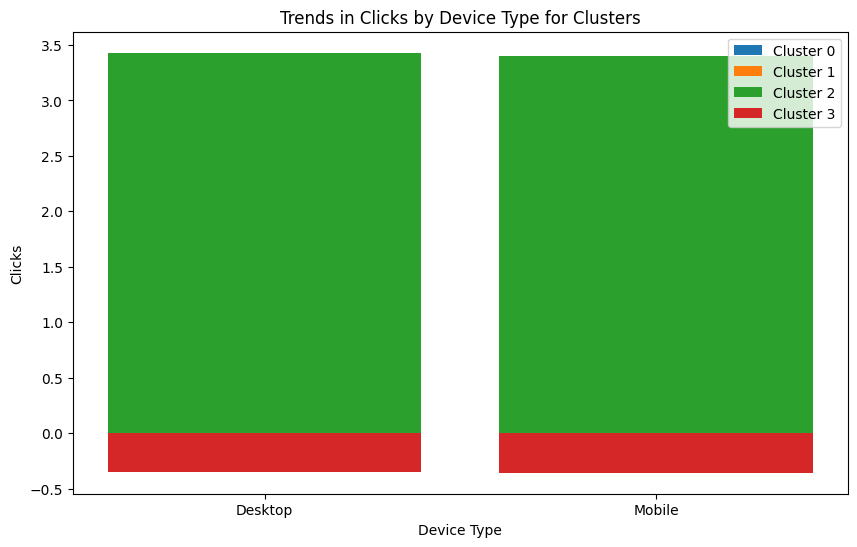
* Optimal number of clusters from K-mean clustering method is 4.

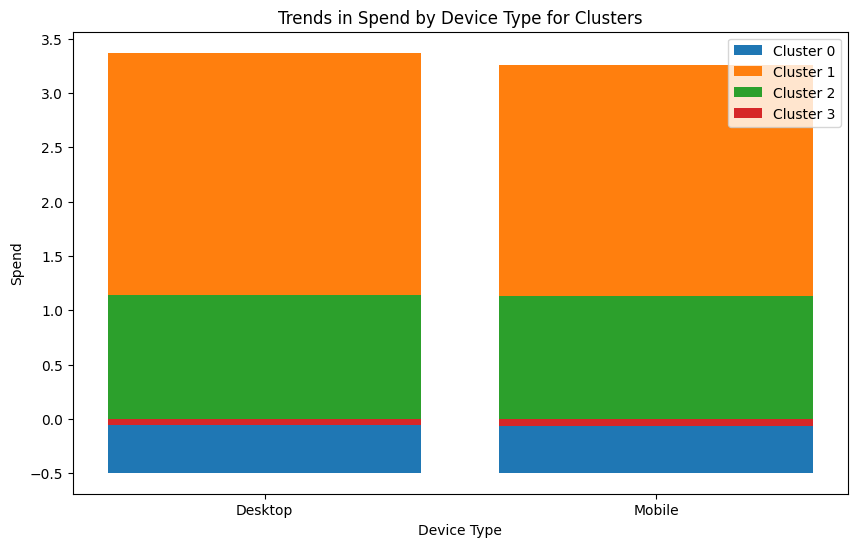


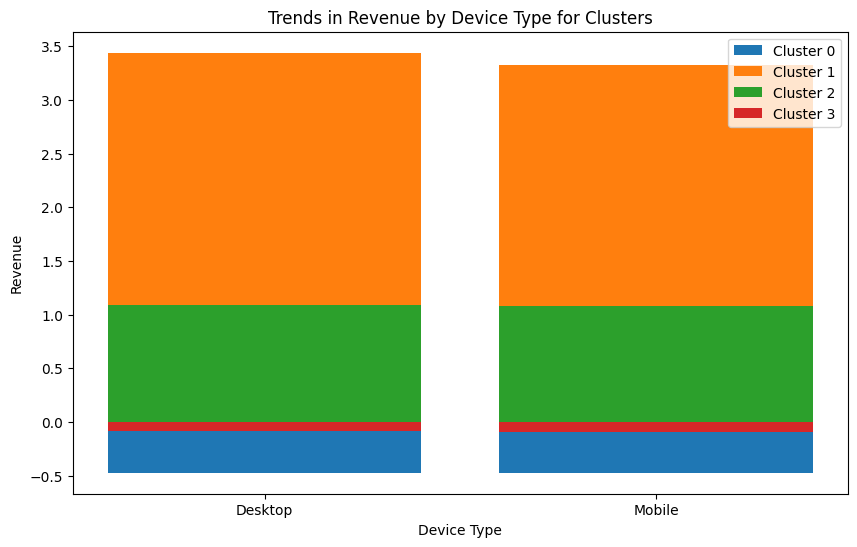
Solution-7:

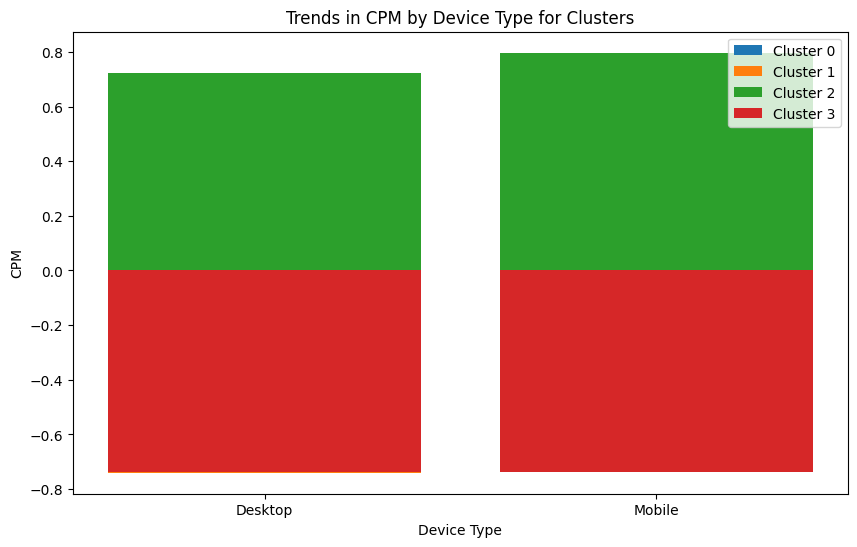


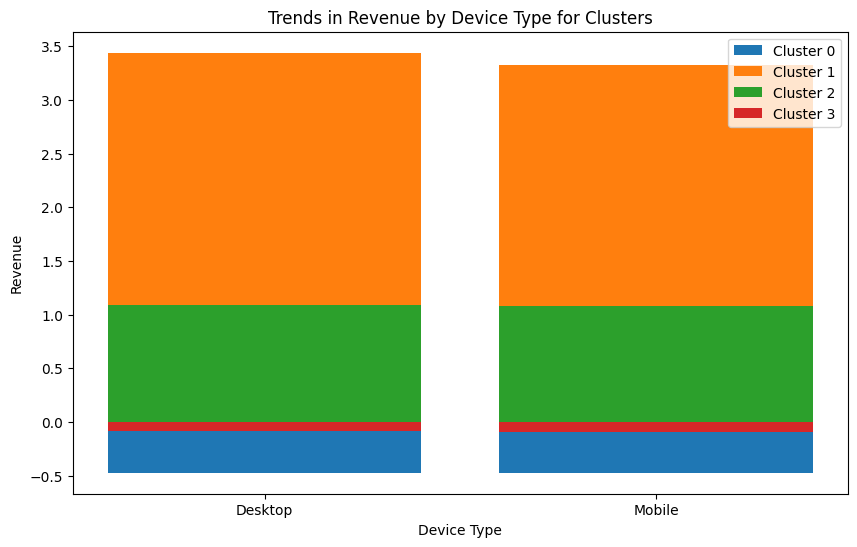
Solution-8:











1. Average clicks in Cluster 2 is highest for both Desktop and Mobile device type
2. Average Spend in Cluster 1 is highest for both Desktop and Mobile device type
3. Average Revenue in Cluster 1 is highest for both Desktop and Mobile device type
4. Average CTR in Cluster 3 is highest for both Desktop and Mobile device type
5. Average CPM in Cluster 2 and 3 is highest for both Desktop and Mobile device type
6. Average CPC in Cluster 1 is highest for both Desktop and Mobile device type