7/30/2023

Cluster Analysis

PGP-DSBA

Karthick Raj S

Table of Contents

Clustering 4

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. 8

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). 9

Part 1 - Clustering: Perform z-score scaling and discuss how it affects the speed of the algorithm. 11

Part 1 - Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance. 12

Part 1 - Clustering: Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm. 14

Part 1 - Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters. 15

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]. 16

Part 1 - Clustering: Conclude the project by providing summary of your learnings. 21

List of Tables

Table 1 First Five Rows of Cluster Data 4
Table 2 Last Five Rows of Cluster Data 5
Table 3 Summary of Cluster Data 7
Table 4 Null values of Cluster Data 7
Table 5 Null Values of Cluster Data(After) 8
Table 6 WSS Values 14
Table 7 Silhouette Score 15
Table 8 Revenue For Clusters 19
Table 9 Revenue, Spend & Fee For Clusters 19
Table 10 Clicks, CTR, CPM & CPC For Clusters 20

List of Figures

Figure A Boxplot of Cluster Data 9
Figure B Boxplot of Cluster Data(After) 10
Figure C Dendrogram 12
Figure D Dendrogram (truncated) 13
Figure E Scree plot for K-Means 14
Figure F Silhouette Score Plot 15
Figure G Count Plot for Clusters 16
Figure H Count Plot for Platform 16
Figure I Count Plot for Inventory Type 17

Clustering:

Digital Ads Data:

The ads24x7 is a Digital Marketing company which has now got seed funding of $10 Million. They are expanding their wings in Marketing Analytics. They collected data from their Marketing Intelligence team and now wants you (their newly appointed data analyst) to segment type of ads based on the features provided. Use Clustering procedure to segment ads into homogeneous groups.

The following three features are commonly used in digital marketing:

CPM = (Total Campaign Spend / Number of Impressions) * 1,000. Note that the Total Campaign Spend refers to the 'Spend' Column in the dataset and the Number of Impressions refers to the 'Impressions' Column in the dataset.

CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column in the dataset and the Number of Clicks refers to the 'Clicks' Column in the dataset.

CTR = Total Measured Clicks / Total Measured Ad Impressions x 100. Note that the Total Measured Clicks refers to the 'Clicks' Column in the dataset and the Total Measured Ad Impressions refers to the 'Impressions' Column in the dataset.

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

Perform the following in given order:

- Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.
- Treat missing values in CPC, CTR and CPM using the formula given. You may refer to the Bank_KMeans Solution File to understand the coding behind treating the missing values using a specific formula. You have to basically create an user defined function and then call the function for imputing.
- 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).
- Perform z-score scaling and discuss how it affects the speed of the algorithm.
- Perform clustering and do the following:

- Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.
- Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.
- Print silhouette scores for up to 10 clusters and identify optimum number of clusters.
- 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.]
- Conclude the project by providing summary of your learnings.

Clustering

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.

Shape:

The Shape of the dataset is (23066,19).

There are **23066** Rows and **19** columns in the dataset.

First Five (Head):

The First Five rows of the dataset (The rows and columns has been transposed for easier view). *Refer Clustering jupyter workings for the output*.

	0	**1**	**2**	**3**	**4**
Timestamp	2020-9-2-17	2020-9-2-10	2020-9-1-22	2020-9-3-20	2020-9-4-15
InventoryType	Format1	Format1	Format1	Format1	Format1
Ad - Length	300	300	300	300	300
Ad- Width	250	250	250	250	250
Ad Size	75000	75000	75000	75000	75000
Ad Type	Inter222	Inter227	Inter222	Inter228	Inter217
Platform	Video	App	Video	Video	Web
Device Type	Desktop	Mobile	Desktop	Mobile	Desktop
Format	Display	Video	Display	Video	Video
Available_Impressions	1806	1780	2727	2430	1218
Matched_Queries	325	285	356	497	242
Impressions	323	285	355	495	242
Clicks	1	1	1	1	1
Spend	0.0	0.0	0.0	0.0	0.0
Fee	0.35	0.35	0.35	0.35	0.35
Revenue	0.0	0.0	0.0	0.0	0.0
CTR	0.0031	0.0035	0.0028	0.0020	0.0041
CPM	0.0	0.0	0.0	0.0	0.0
CPC	0.0	0.0	0.0	0.0	0.0

Table 1 First Five Rows of Cluster Data

Last Five (Tail):

The Last Five rows of the dataset (The rows and columns has been transposed for easier view). *Refer Clustering jupyter workings for the output*.

	23061	23062	23063	23064	23065
Timestamp	2020-9-13-7	2020-11-2-7	2020-9-14-22	2020-11-18-2	2020-9-14-0
InventoryType	Format5	Format5	Format5	Format4	Format5
Ad - Length	720	720	720	120	720
Ad- Width	300	300	300	600	300
Ad Size	216000	216000	216000	72000	216000
Ad Type	Inter220	Inter224	Inter218	inter230	Inter221
Platform	Web	Web	App	Video	App
Device Type	Mobile	Desktop	Mobile	Mobile	Mobile
Format	Video	Video	Video	Video	Video
Available_Impressions	1	3	2	7	2
Matched_Queries	1	2	1	1	2
Impressions	1	2	1	1	2
Clicks	1	1	1	1	1
Spend	0.07	0.04	0.05	0.07	0.09
Fee	0.35	0.35	0.35	0.35	0.35
Revenue	0.0455	0.0260	0.0325	0.0455	0.0585
CTR	NaN	NaN	NaN	NaN	NaN
CPM	NaN	NaN	NaN	NaN	NaN
CPC	NaN	NaN	NaN	NaN	NaN

Table 2 Last Five Rows of Cluster Data

Info:

The Info of the dataset is

```
<class 'pandas.core.frame.DataFrame'>
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
```

Summary:

The Summary of the dataset is

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066	3.85E+02	2.34E+02	120	120	300	7.20E+02	728
Ad- Width	23066	3.38E+02	2.03E+02	70	250	300	6.00E+02	600
Ad Size	23066	9.67E+04	6.15E+04	33600	72000	72000	8.40E+04	216000
Available_Impressions	23066	2.43E+06	4.74E+06	1	33672.25	483771	2.53E+06	27592861
Matched_Queries	23066	1.30E+06	2.51E+06	1	18282.5	258087.5	1.18E+06	14702025
Impressions	23066	1.24E+06	2.43E+06	1	7990.5	225290	1.11E+06	14194774
Clicks	23066	1.07E+04	1.74E+04	1	710	4425	1.28E+04	143049
Spend	23066	2.71E+03	4.07E+03	0	85.18	1425.125	3.12E+03	26931.87
Fee	23066	3.35E-01	3.20E-02	0.21	0.33	0.35	3.50E-01	0.35
Revenue	23066	1.92E+03	3.11E+03	0	55.36538	926.335	2.09E+03	21276.18
CTR	18330	7.37E-02	7.52E-02	0.0001	0.0026	0.08255	1.30E-01	1
CPM	18330	7.67E+00	6.48E+00	0	1.71	7.66	1.25E+01	81.56
CPC	18330	3.51E-01	3.43E-01	0	0.09	0.16	5.70E-01	7.26

Table 3 Summary of Cluster Data

Duplicates and Null Values:

There is no Duplicates in the dataset.

There are some null values in the CPC, CTR and CPM columns.

Columns	No. of Null Values
CTR	4736
CPM	4736
CPC	4736

Table 4 Null values of Cluster Data

Part 1- Clustering: Treat missing values in CPC, CTR and CPM using the formula given.

The Missing Values in CPC, CTR and CPM are treated using the formula:

CPM = (Total Campaign Spend / Number of Impressions) x1,000.

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

CTR = (Total Measured Clicks / Total Measured Ad Impressions) x 100.

The Null Values after treating the columns are:

Columns	No. of Null Values
CTR	0
CPM	0
CPC	0

Table 5 Null Values of Cluster Data(After)

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).

There are outliers in the dataset. The outliers are visible from the below boxplots

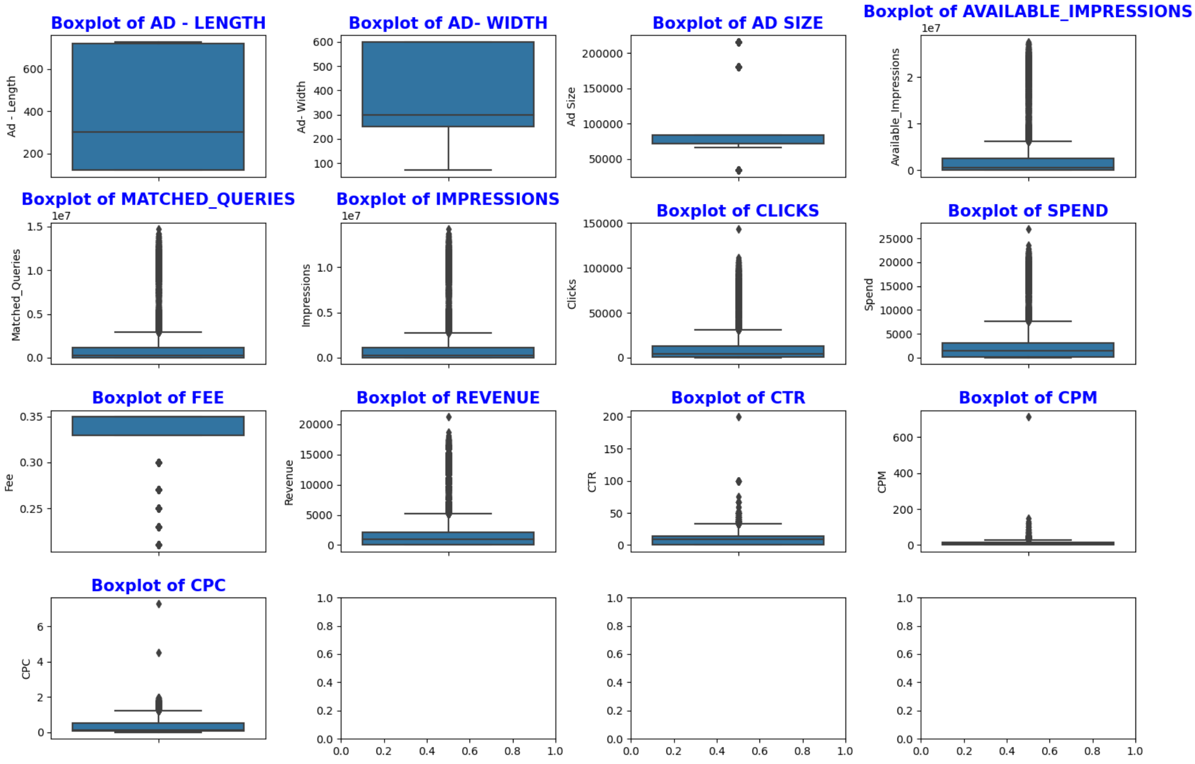

Figure A Boxplot of Cluster Data

Treating outliers is a necessary step for K-Means clustering as it influences the cluster. K-Means algorithm finds the mean or centroid of the clusters that can be affected by the outliers. So, the outliers has to be treated.

The most used and efficient method in treating outlier is using InterQuartile range for calculating the Lower and Upper range.

After Outlier Treatment:

The Boxplot after treating outliers

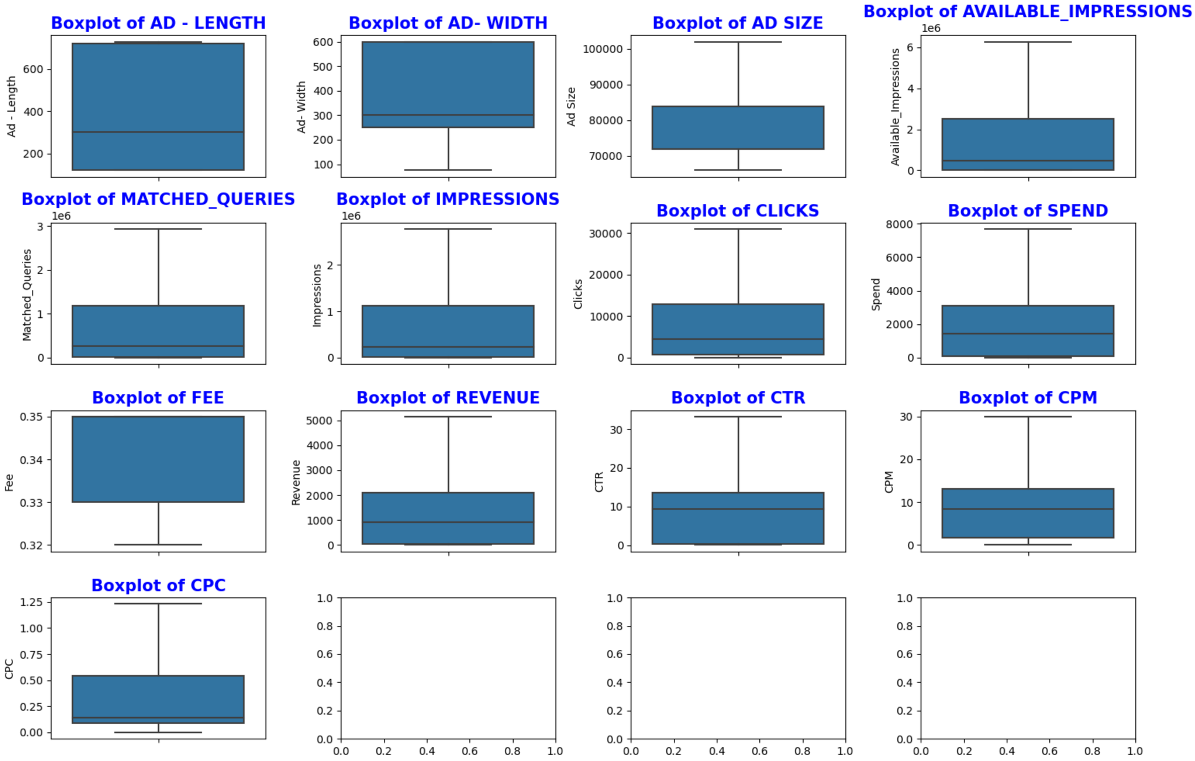

Figure B Boxplot of Cluster Data(After)

The dataset has been treated for both missing values and outliers.

Part 1- Clustering: Perform z-score scaling and discuss how it affects the speed of the algorithm.

Before scaling the dataset, the categorical variables in the dataset are labelled through one hot encoding.

The new dataset with labelled categorical variables has 23066 Rows and 36 Columns.

The dataset has been scaled through StandardScaler. Now the dataset has mean of 0 and standard deviation of 1. *(Refer jupyter working for the summary after scaling)*.

Without scaling, the cluster is affected by the variable unit which is highest. Through Scaling all the columns are normalized.

Part 1- Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.

The Hierarchical cluster has been performed with Ward linkage and Euclidean distance for the dataset and the dendrogram is constructed.

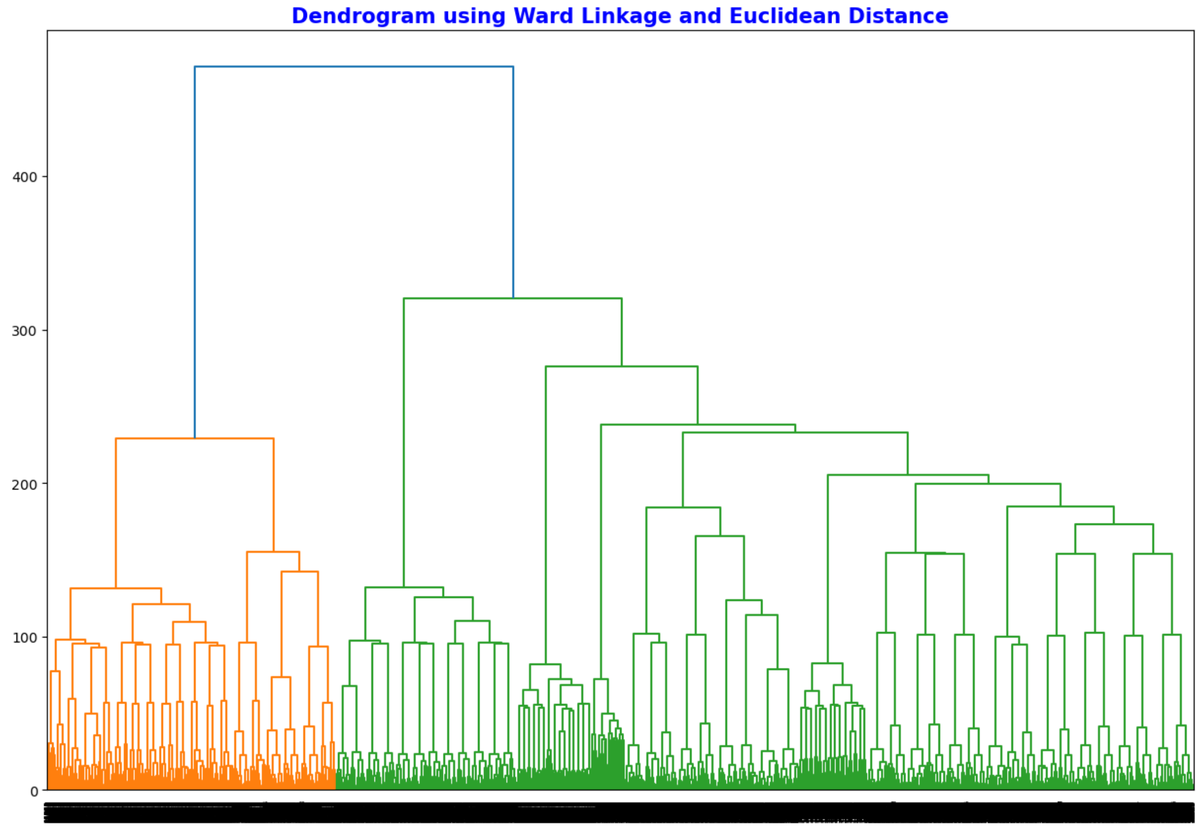

Figure C Dendrogram

For Easier representation of the dendrogram, it has been truncated to 12 clusters.

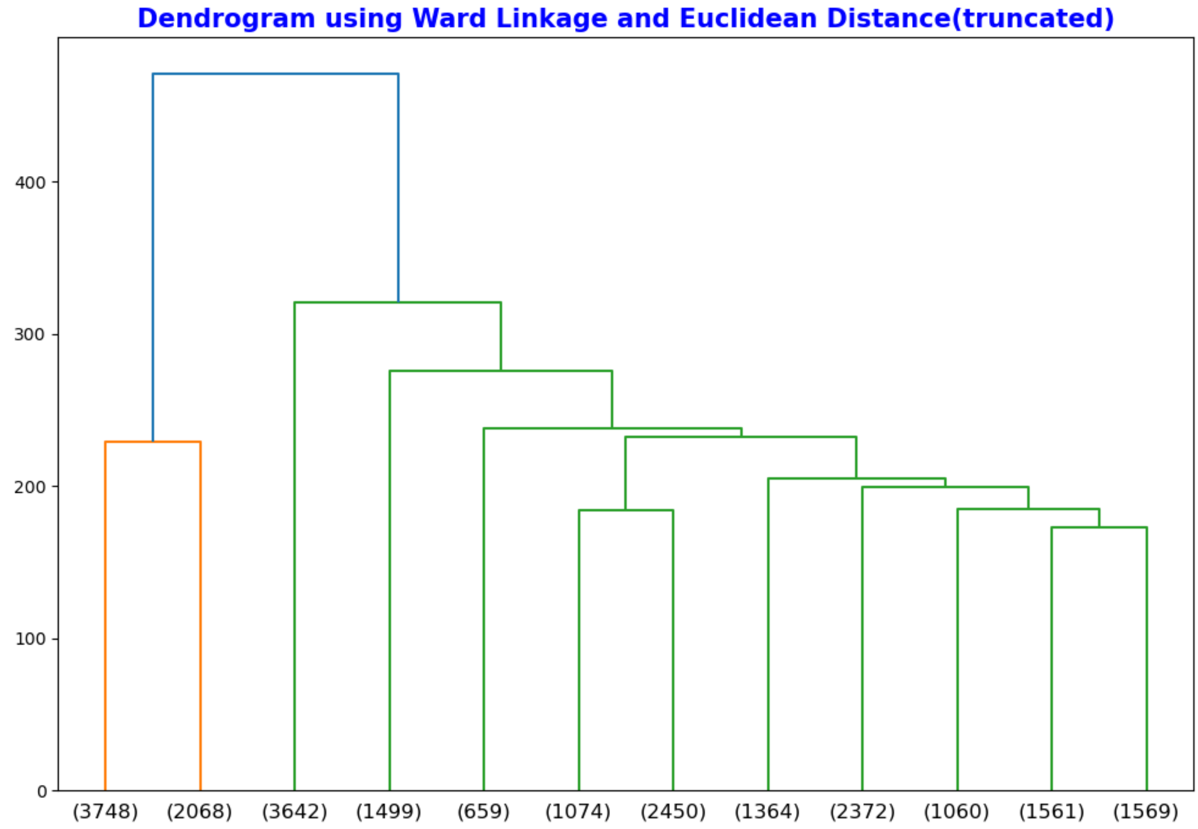

Figure D Dendrogram (truncated)

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

Within Sum of Square for each cluster

Number of clusters	WSS	difference
1	830376.00	NaN
2	695004.98	-135371.02
3	626983.77	-68021.21
4	565851.69	-61132.08
5	530418.98	-35432.71
6	506446.04	-23972.94
7	490590.12	-15855.92
8	466912.27	-23677.85
9	455951.28	-10960.99
10	439558.48	-16392.80

Table 6 WSS Values

From 5 clusters to 6 clusters and so on, the difference between the clusters has been reduced drastically.

From the Elbow plot, it can be seen the elbow is formed at 5.

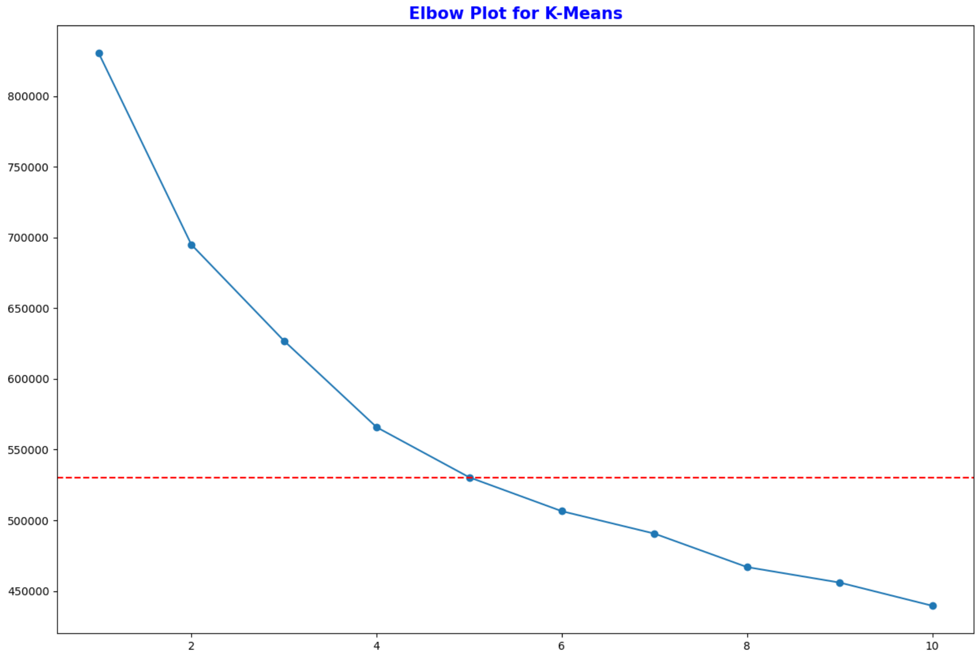

Figure E Scree plot for K-Means

To confirm accurately, silhouette score can be used.

Part 1- Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters.

The Silhouette score is highest for the 5 clusters. It is also clearly seen from the plot also

Number of clusters	Silhouette score
2	0.160321
3	0.147746
4	0.173651
5	***0.189136***
6	0.188508
7	0.172752
8	0.167041
9	0.171424
10	0.181049

Table 7 Silhouette Score

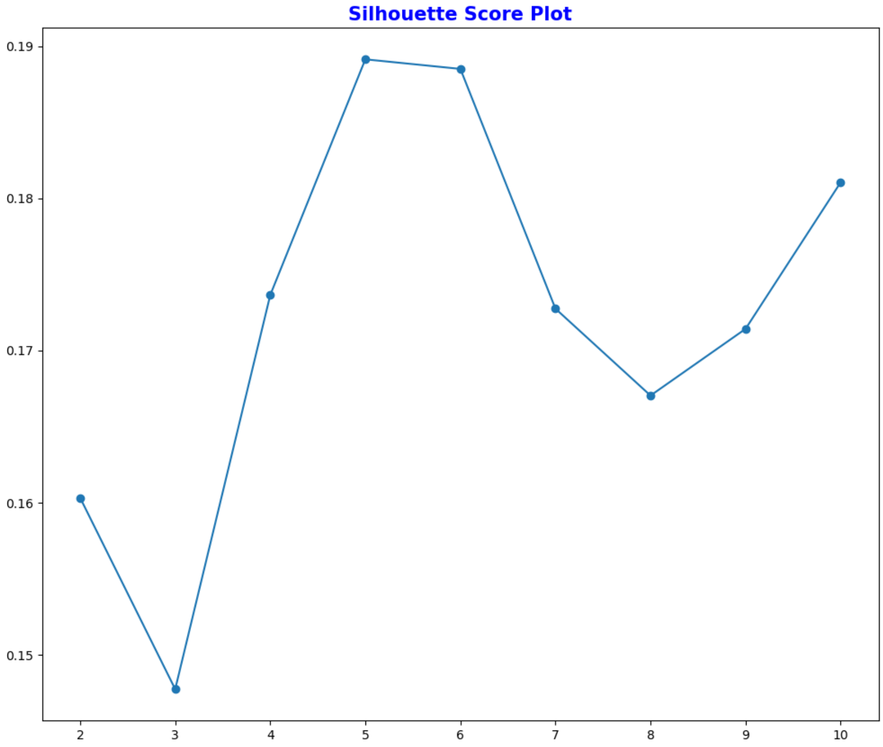

Figure F Silhouette Score Plot

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].

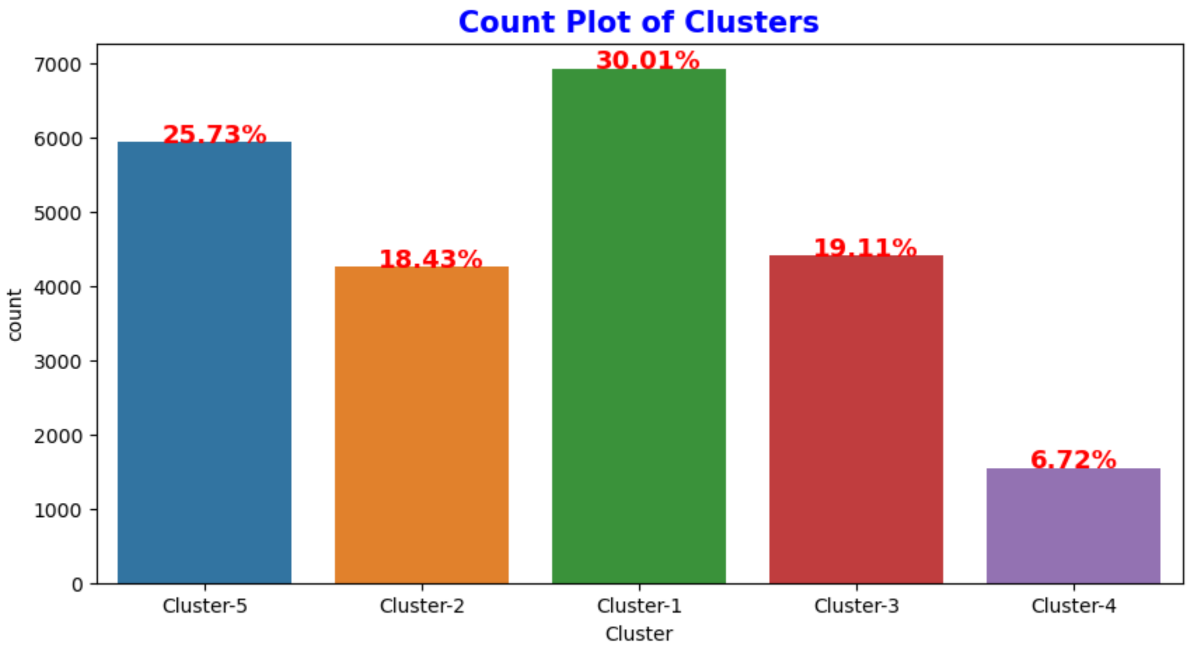

Figure G Count Plot for Clusters

Cluster -1 has the highest percentage of the dataset. Cluster 1 and 5 Covers about half of the dataset.

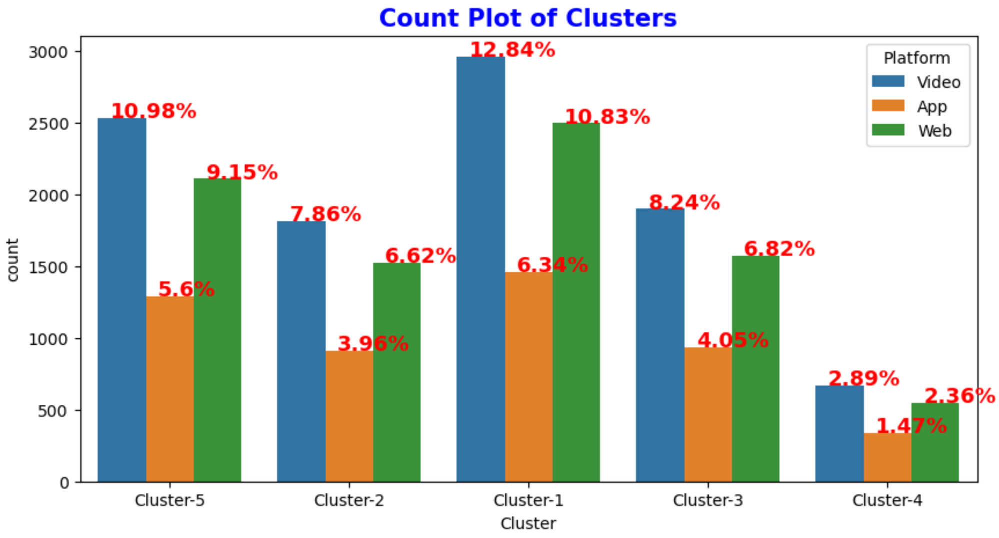

Figure H Count Plot for Platform

In all the clusters, video and web platform has the most of the dataset.

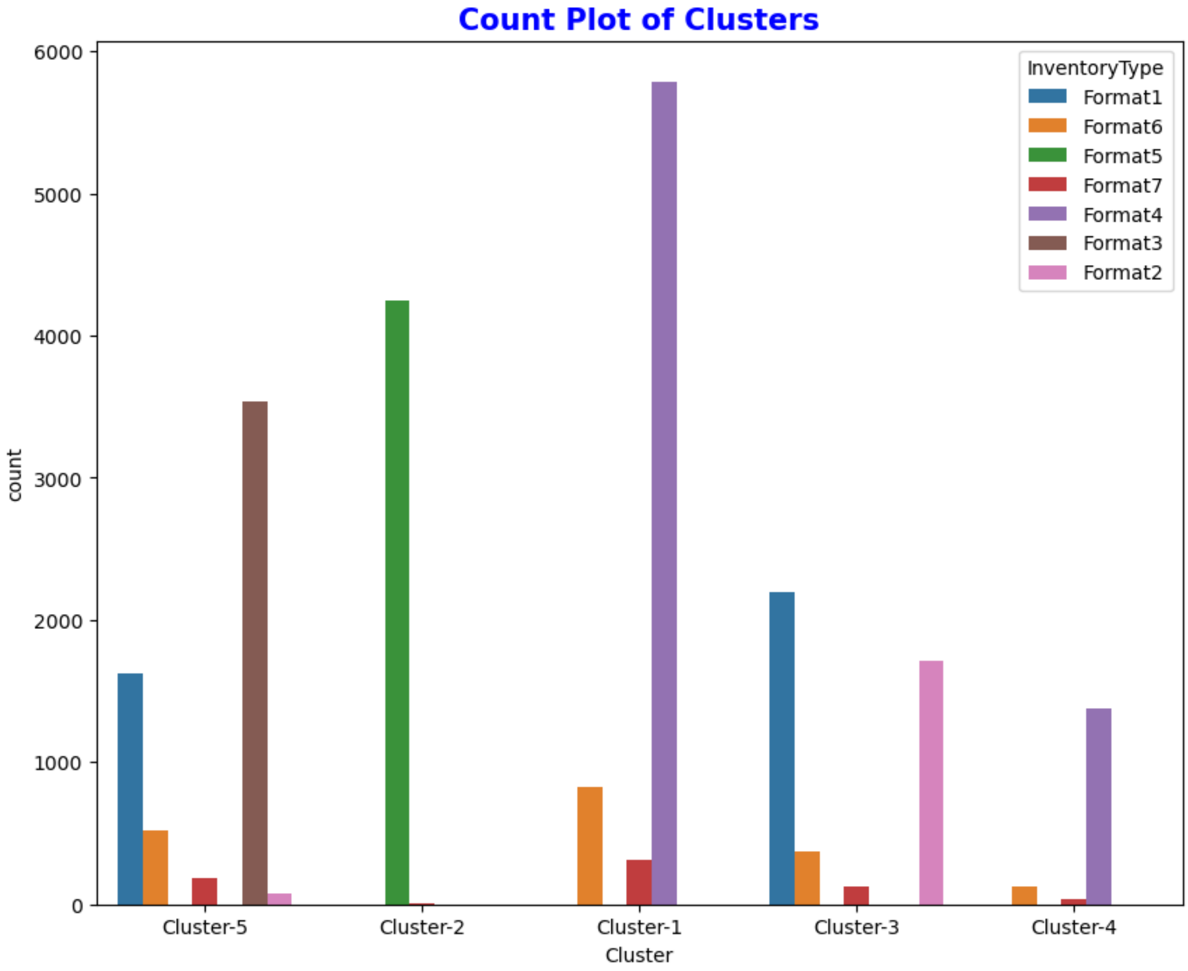

Figure I Count Plot for Inventory Type

Revenue:

	Platform	App		Video		Web		All
	Format	Display	Video	Display	Video	Display	Video	
Cluster	*Device Type*							
Cluster-1	Desktop	0.00	0.00	141015.04	132513.31	**93717.62**	95078.85	**462324.82**
	Mobile	152210.15	134920.09	128300.86	148352.20	154326.47	122606.49	840716.25
Cluster-2	Desktop	0.00	0.00	355549.06	367425.28	246322.26	233941.37	1203238
	Mobile	380301.03	339730.93	348965.29	366536.36	344375.96	378794.13	2158704
Cluster-3	Desktop	0.00	0.00	1763520.22	1784148.09	1177927.84	1184219.56	5909816
	Mobile	1788734.52	1692483.00	**1808456.26**	1709370.30	1704699.47	1780361.52	**10484105**
Cluster-4	Desktop	0.00	0.00	779442.73	681213.37	538678.90	500900.66	2500236

	Mobile	729480.34	788229.58	729986.27	792477.03	744713.80	618472.66	4403360
Cluster-5	Desktop	0.00	0.00	587207.87	580989.12	397132.19	359597.26	1924926
	Mobile	617018.37	558956.24	547778.48	614670.86	562556.56	643190.98	3544171
All		3667744.4	3667744.40	3514319.84	**7190222.08**	**7177695.92**	5964451.08	5917163.47

Table 8 Revenue For Clusters

The Highest Revenue is for cluster 3 for Device type – mobile.

The highest revenue yielding platform is Video and both the format are nearly having same amount of revenue.

The lowest revenue yielding cluster is cluster 1.

	Revenue		Spend		Fee in (%)
Cluster	**mean**	**sum**	**mean**	**sum**	**mean**
Cluster-1	188.25	**1303041.07**	289.61	**2004679**	35%
Cluster-2	790.67	3361941.69	1214.05	5162125	35%
Cluster-3	3719.97	**16393920.78**	5509.45	**24280153**	33%
Cluster-4	4451.06	6903595.33	6516.56	10107180	32%
Cluster-5	921.65	5469097.93	1416.95	8408205	35%

Table 9 Revenue, Spend & Fee For Clusters

The Highest Mean revenue is Cluster 4.

In Spend, the highest is cluster 3 and the lowest is cluster 1. For Mean also it is the same.

The Payable Fee % pending from the franchise is mostly the same for all clusters.

Clicks, CTR, CPM & CPC:

	Clicks		CTR		CPM		CPC	
Cluster	**mean**	**sum**	**mean**	**sum**	**mean**	**sum**	**mean**	**sum**
Cluster-1	2854.08	19755922	**15.63**	**108189.1**	14.18	98159.23	0.1	704.16
Cluster-2	13744.09	**58439887.88**	13.34	56737.44	11.73	49892.51	0.09	383.89
Cluster-3	10585.49	46650271	0.21	946.16	1.6	7029.66	**0.77**	3385.95
Cluster-4	**30506.55**	**47315658.25**	13.75	21320.86	**15.42**	23910.07	0.11	174.02
Cluster-5	3276.64	19443560	0.42	2482.87	1.78	10592.15	0.5	2964.56

Table 10 Clicks, CTR, CPM & CPC For Clusters

Cluster 2 and 4 has highest clicks as a whole. While the mean highest clicks are for cluster 4.

The Cost per click is highest for cluster 3 which is good when comparing with the revenue it makes. But when taking the spend also into account, none of the clusters reaches the breakeven point.

The Click through rate is highest for Cluster 1.

Part 1- Clustering: Conclude the project by providing summary of your learnings.

None of the clusters make the breakeven point.

Cluster 1:

The cluster 1 has the most ads, least revenue and spend.

It also has the highest CTR, CPM and less loss compared to others.

Cluster 1 is the most profitable and majority of ads is in that cluster. The type of ads in cluster 1 are Format 4 and the platform they use are Web and video. The Most of the Ads are supported by mobile devices.

It has the largest Ad-Size.

This cluster has the major business franchise clients for the company.

Cluster 2:

Cluster 2 has the highest Ad Length and highest Clicks. It has the Format 4 Ads.

The Revenue and Spend in this cluster is moderate compared to other clusters.

Cluster 3:

The Cluster 3 has the highest revenue but it has the second lowest number of ads in the cluster. It is because of the payable from the franchise are low compared to others. It has a poor CTR. If the Ads are used efficiently and the CTR's improved. This cluster will yield more revenue.

Format 1 and Format 2 are the dominant ads in this cluster.

This Cluster has the highest profit for the company.

Cluster 4:

The Cluster 4 has the lowest Ad size, lowest number of Ads and highest Clicks and CTR. This Cluster of Ad segment can be dropped.

It is mostly similar to cluster 1 in the Ad format and Ad Length. But the Spend for this cluster is very high when compared with the number of Ads it has.

Cluster 5:

Cluster 5 is the second highest business for the company. With cluster 1 and 5 together, Half of the company's business is based on the Ads from these clusters.

Most of the Ads in this cluster is of Format 3.