Business Report

Index-

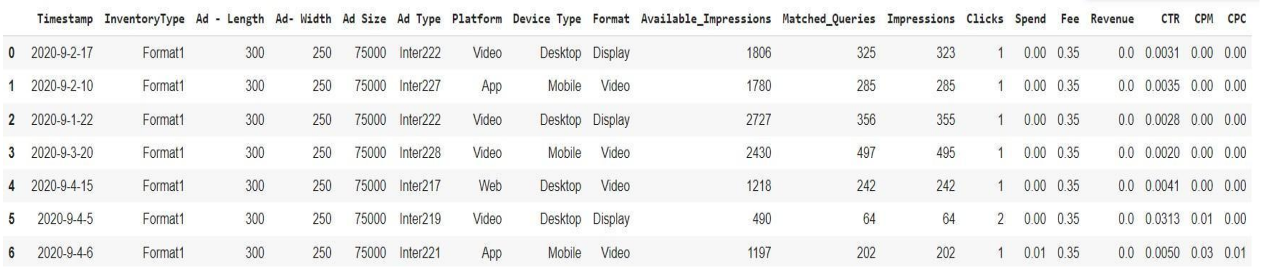
Part 1 - Clustering:

Part 1 - Clustering:

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

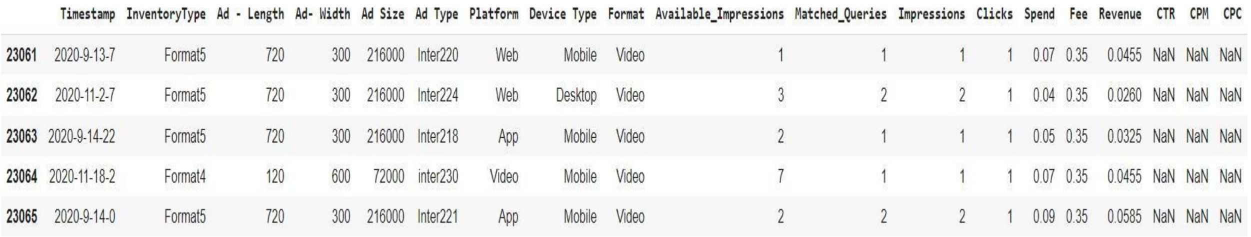
Answer-

Below are some of the rows of Data set head-



	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.00	0.35	0.0	0.0031	0.00	0.00
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.00	0.35	0.0	0.0035	0.00	0.00
2	2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.00	0.35	0.0	0.0028	0.00	0.00
3	2020-9-3-20	Format1	300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.00	0.35	0.0	0.0020	0.00	0.00
4	2020-9-4-15	Format1	300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.00	0.35	0.0	0.0041	0.00	0.00
5	2020-9-4-5	Format1	300	250	75000	Inter219	Video	Desktop	Display	490	64	64	2	0.00	0.35	0.0	0.0313	0.01	0.00
6	2020-9-4-6	Format1	300	250	75000	Inter221	App	Mobile	Video	1197	202	202	1	0.01	0.35	0.0	0.0050	0.03	0.01

Below are some of the rows of Data set tail-

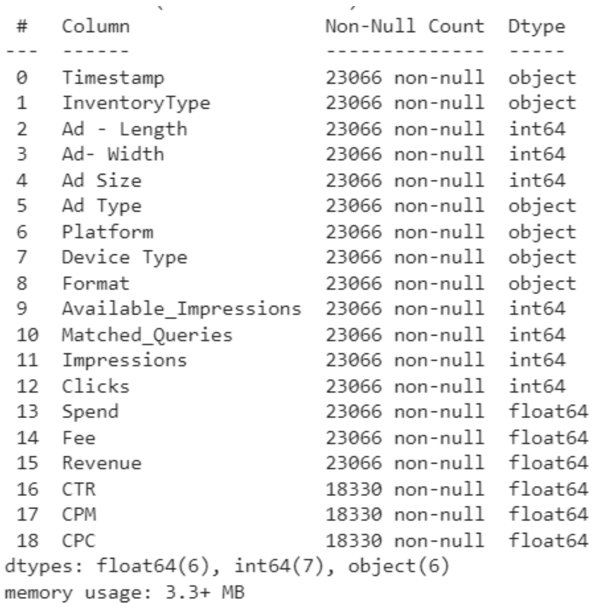


	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
23061	2020-9-13-7	Format5	720	300	216000	Inter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
23062	2020-11-2-7	Format5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

-Some of the missing values can be spotted in tail of dataset.

-Checking shape of data we found that there are 23066 rows and 19 columns in data set.

Let’s check more information about dataset.



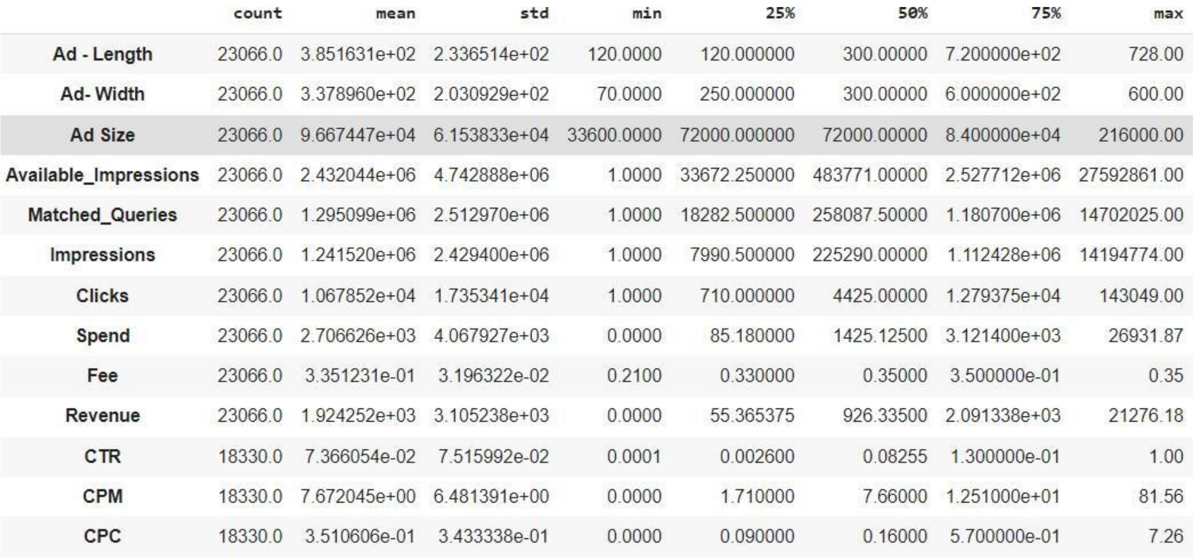
```
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Timestamp              23066 non-null  object
 1   InventoryType          23066 non-null  object
 2   Ad - Length            23066 non-null  int64
 3   Ad- Width              23066 non-null  int64
 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
```

-we can see that there are 6 float type , 7 integer type and 6 object type features are there.

-or there are 13 continuous and 6 categorical features are present.

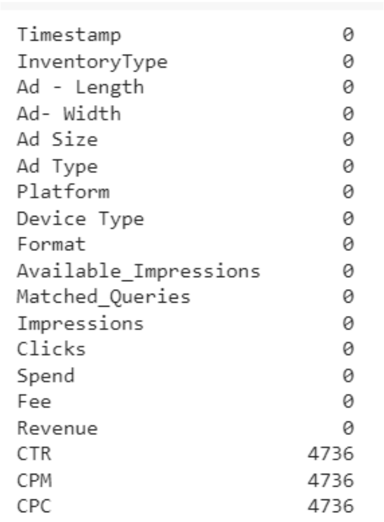
- Noticeably there are only 18330 entries in CTR, CPM and CPC features as compared to 23066 entries in remaining feature. Thus a lot of missing values are there.

Let’s check how continuous variables are distributed.



	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	3.851631e+02	2.336514e+02	120.0000	120.000000	300.00000	7.200000e+02	728.00
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.0000	250.000000	300.00000	6.000000e+02	600.00
Ad Size	23066.0	9.667447e+04	6.153833e+04	33600.0000	72000.000000	72000.00000	8.400000e+04	216000.00
Available_Impressions	23066.0	2.432044e+06	4.742888e+06	1.0000	33672.250000	483771.00000	2.527712e+06	27592861.00
Matched_Queries	23066.0	1.295099e+06	2.512970e+06	1.0000	18282.500000	258087.50000	1.180700e+06	14702025.00
Impressions	23066.0	1.241520e+06	2.429400e+06	1.0000	7990.500000	225290.00000	1.112428e+06	14194774.00
Clicks	23066.0	1.067852e+04	1.735341e+04	1.0000	710.000000	4425.00000	1.279375e+04	143049.00
Spend	23066.0	2.706626e+03	4.067927e+03	0.0000	85.180000	1425.12500	3.121400e+03	26931.87
Fee	23066.0	3.351231e-01	3.196322e-02	0.2100	0.330000	0.35000	3.500000e-01	0.35
Revenue	23066.0	1.924252e+03	3.105238e+03	0.0000	55.365375	926.33500	2.091338e+03	21276.18
CTR	18330.0	7.366054e-02	7.515992e-02	0.0001	0.002600	0.08255	1.300000e-01	1.00
CPM	18330.0	7.672045e+00	6.481391e+00	0.0000	1.710000	7.66000	1.251000e+01	81.56
CPC	18330.0	3.510606e-01	3.433338e-01	0.0000	0.090000	0.16000	5.700000e-01	7.26

-No negative value or as such anomaly could be spotted in above table.



```
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
```

-There are almost 4736 values missing in all three of CTR, CPM and CPC features.

-There are no duplicate records present in data.

Part 1 - Clustering:

1.2- Treat missing values in CPC, CTR and CPM using the formula given.

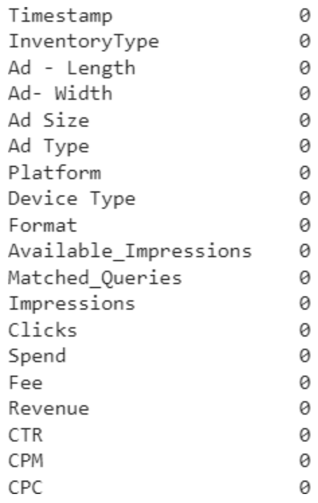
Answer-

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

CPC = Total Cost (spend) / Number of Clicks

CTR = (Total Measured Clicks / Total Measured Ad Impressions)* 100

We have made a user defined function for each formula and imputed the missing values accordingly.



```
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                      0
CPM                      0
CPC                      0
```

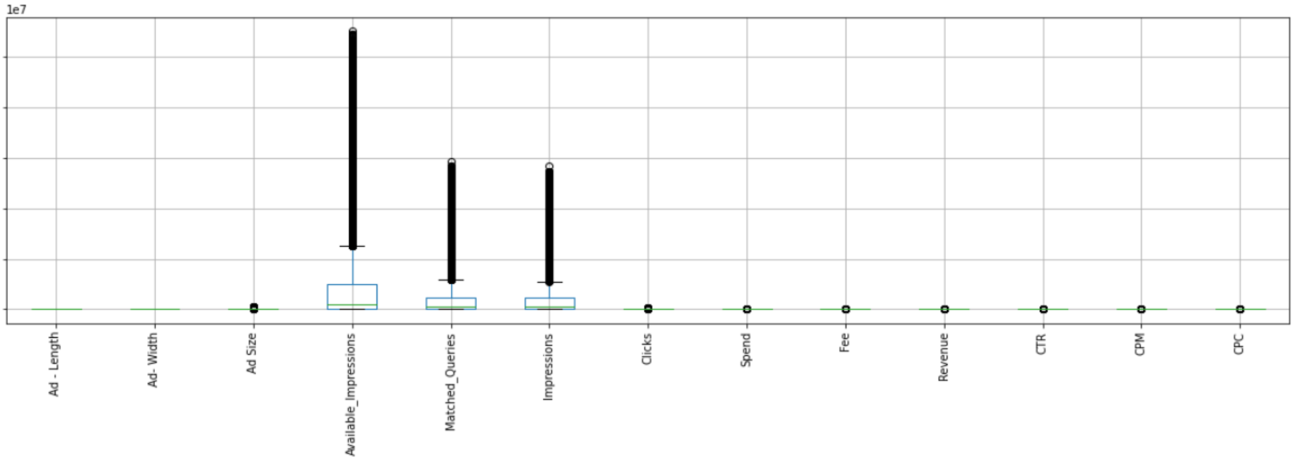
There are currently no missing values present.

Part 1 - Clustering:

1.3-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.

Answer-

Let’s check if there are outliers in dataset –



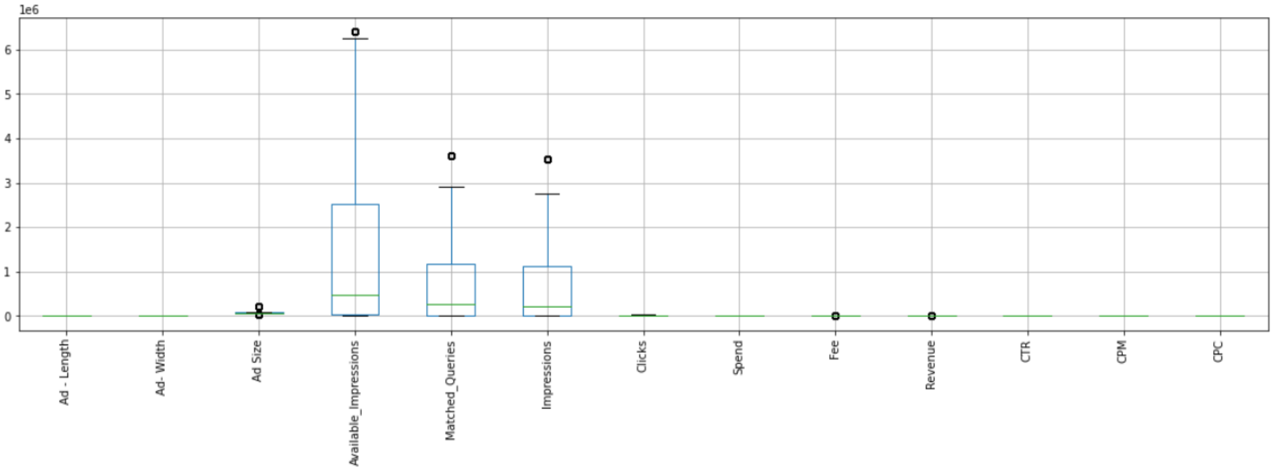
There are outliers present in almost all the continuous features except Ad-Length and Ad-width.

For K-means clustering treating outliers is very important as it is a distance based algorithm, so outliers will have a significant impact on clustering. Outliers will impact the way clusters are formed. The shifting of centroid in each iteration in K-means clustering is influenced by outliers, resulting in not so good clusters. Within cluster variance would be very high in these clusters.

Technique we used is –

We have initially calculated the 10th percentile and 90th percentile value of each column where outliers are present.

The data points that are lesser than the 10th percentile for each column are replaced with the 10th percentile value of that column and the data points that are greater than the 90th percentile are replaced with 90th percentile value of that column.



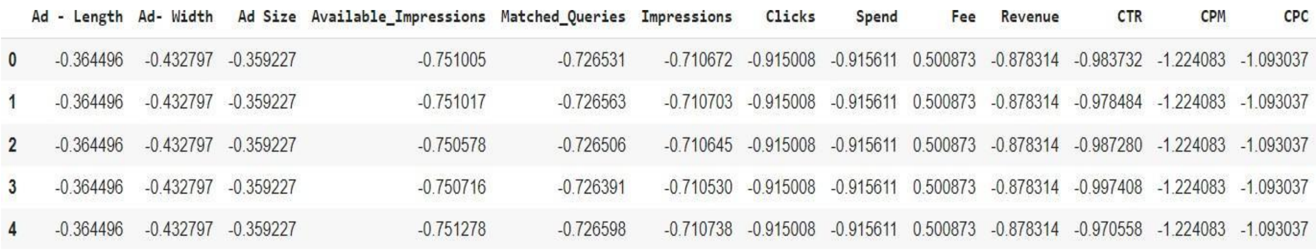
-Features after applying above mentioned capping. Outliers are capped.

Part 1 - Clustering:

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

Answer-

-Z-score scaling is performed, below is the resulting dataset.

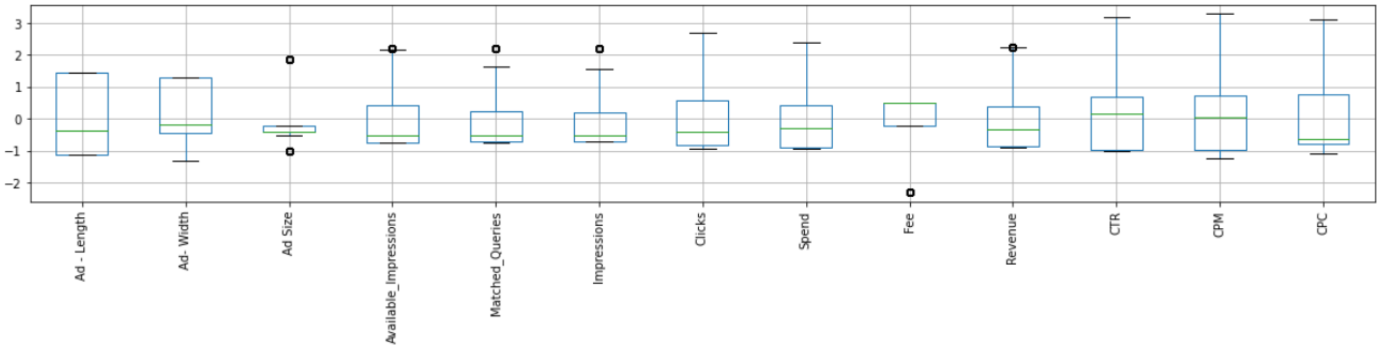


	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	-0.364496	-0.432797	-0.359227	-0.751005	-0.726531	-0.710672	-0.915008	-0.915611	0.500873	-0.878314	-0.983732	-1.224083	-1.093037
1	-0.364496	-0.432797	-0.359227	-0.751017	-0.726563	-0.710703	-0.915008	-0.915611	0.500873	-0.878314	-0.978484	-1.224083	-1.093037
2	-0.364496	-0.432797	-0.359227	-0.750578	-0.726506	-0.710645	-0.915008	-0.915611	0.500873	-0.878314	-0.987280	-1.224083	-1.093037
3	-0.364496	-0.432797	-0.359227	-0.750716	-0.726391	-0.710530	-0.915008	-0.915611	0.500873	-0.878314	-0.997408	-1.224083	-1.093037
4	-0.364496	-0.432797	-0.359227	-0.751278	-0.726598	-0.710738	-0.915008	-0.915611	0.500873	-0.878314	-0.970558	-1.224083	-1.093037

Scaling does not have any affect on speed of the algorithm. As it has to calculate same number of distances each time, weather data is scaled or not.

Time-lapse of 1-2 secs could be noticed, total time of execution of algorithm is around 27 secs. There is no significant difference whether we apply algorithm on scaled data or unscaled data, time taken by algorithm is same.

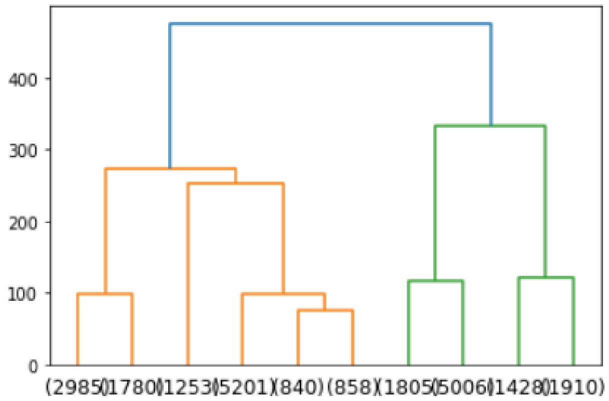
Distribution of Data after scaling looks like this.



Part 1 - Clustering:

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

Answer



This is the Dendrogram we got after performing Hierarchical Clustering using Ward Linkage and Euclidean distance.

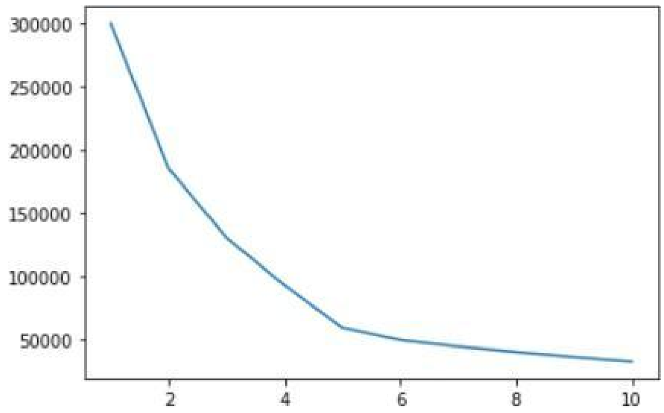
We can go with either 2 or 3 or 5 clusters according to above pattern. We will identify the right number of clusters using Business intelligence and silhouette scores in below answers.

Part 1 - Clustering:

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

We use k-means algorithm and below is the Elbow plot.

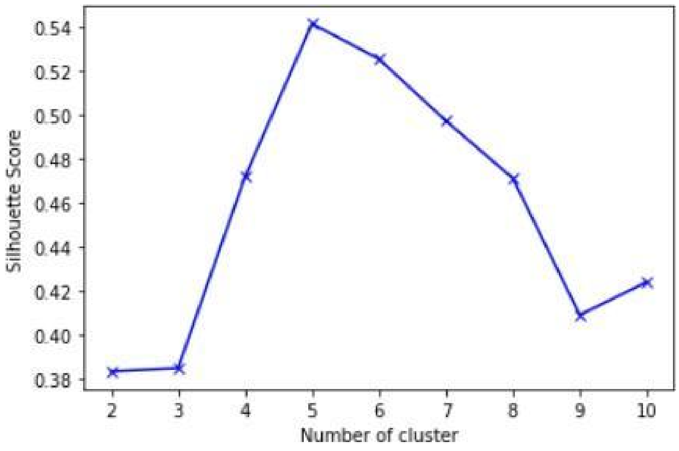
- Elbow plot tells us within-cluster sum of squared distances as a function of number of clusters.

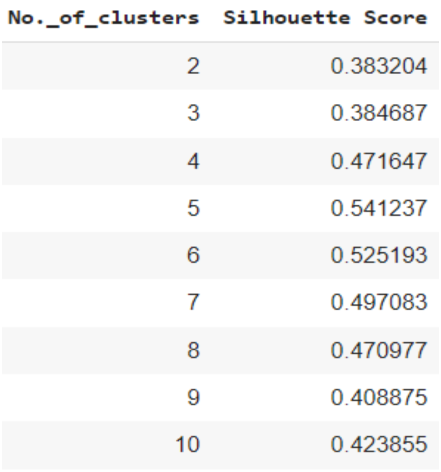


It could be noticed that we have an elbow @ 5 after which the curve flattens.

Part 1 - Clustering:

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





No._of_clusters	Silhouette Score
2	0.383204
3	0.384687
4	0.471647
5	0.541237
6	0.525193
7	0.497083
8	0.470977
9	0.408875
10	0.423855

- Silhouette score is max at 5 clusters.

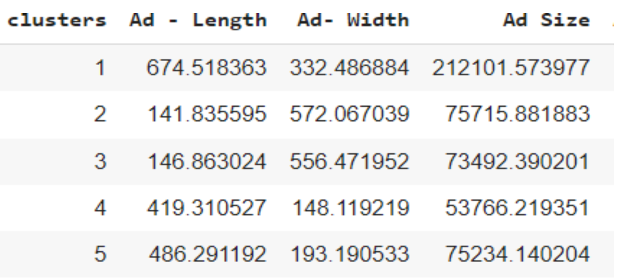
- If we choose to go with 5 clusters we would have clusters that are well apart from each other and clearly distinguished.

-Optimum number of clusters = 5.

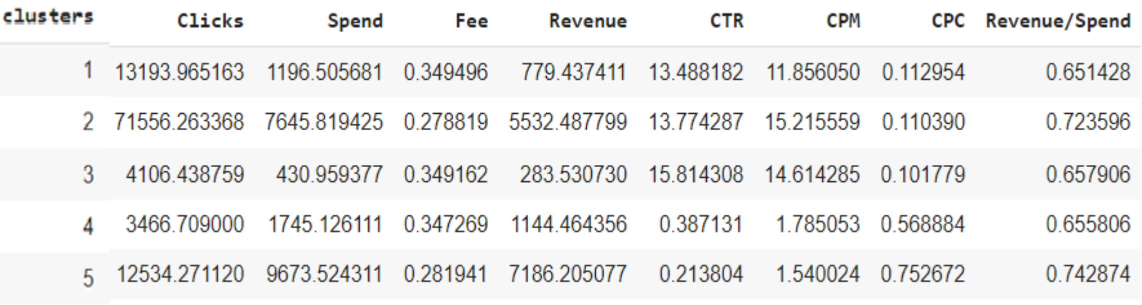
Part 1 - Clustering:

1.8-Profile the ads based on optimum number of clusters using silhouette score and your domain understanding.

Let’s check out the mean values of different features of all 5 clusters.



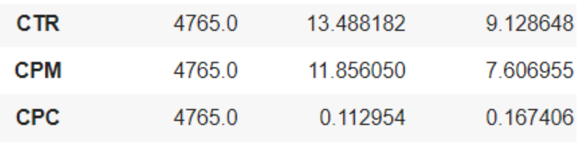
clusters	Ad - Length	Ad- Width	Ad Size
1	674.518363	332.486884	212101.573977
2	141.835595	572.067039	75715.881883
3	146.863024	556.471952	73492.390201
4	419.310527	148.119219	53766.219351
5	486.291192	193.190533	75234.140204



clusters	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	Revenue/Spend
1	13193.965163	1196.505681	0.349496	779.437411	13.488182	11.856050	0.112954	0.651428
2	71556.263368	7645.819425	0.278819	5532.487799	13.774287	15.215559	0.110390	0.723596
3	4106.438759	430.959377	0.349162	283.530730	15.814308	14.614285	0.101779	0.657906
4	3466.709000	1745.126111	0.347269	1144.464356	0.387131	1.785053	0.568884	0.655806
5	12534.271120	9673.524311	0.281941	7186.205077	0.213804	1.540024	0.752672	0.742874

We have kept a track of Ad-sizes covering 10th and 90th percentile values, for each cluster. We will also cover the mean and standard deviation of CTR, CPM & CPC. We will also see the impact of all these mentioned parameters on Revenue/Spend.

Cluster 1- Ad-size- Range [180000-216000] | Mean- 212101



	Count	Mean	Std.
CTR	4765.0	13.488182	9.128648
CPM	4765.0	11.856050	7.606955
CPC	4765.0	0.112954	0.167406

This cluster covers the ads where Ad Size is very large compared to other clusters.

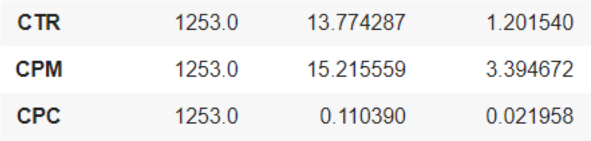
Ad length > Ad- width

Its CTR is around 13.48; we are almost getting 13 clicks when ad is shown 100 times. This type of ad has of CPM of 11.85 and spend/click that is CPC = 0.11.

Standard deviation of CTR, CPM and CPC is 9.1, 7.6 and 0.16 respectively . This is quite high Standard deviation.

This type of Ad is capable of generating almost 0.65% of the revenue of total spends.

Cluster 2- Ad-size- Range [72000-84000] | Mean- 75716



	Count	Mean	Std.
CTR	1253.0	13.774287	1.201540
CPM	1253.0	15.215559	3.394672
CPC	1253.0	0.110390	0.021958

This cluster covers the ads where Ad Size is medium compared to other clusters.

Ad Width > Ad Length

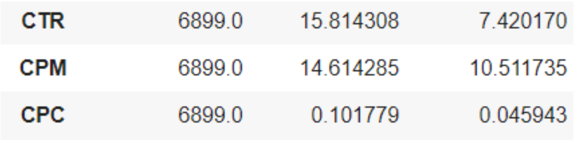
This cluster has lowest Fee=0.27 Its CTR is around 13.77, we are almost getting 13 clicks when ad is shown 100 times. This type of ad has highest CPM of 15.85 that is spend / 1000 times ad is shown is 15.85. And spend/click that is CPC = 0.11.

Standard deviation of CTR, CRM and CPC is 1.2, 3.39 and 0.02 respectively. This is quite low Standard deviation. Variability in CTR, CPM and CPC is low.

This type of Ad is capable of generating second highest almost 0.72% of the revenue of total spends.

Cluster 3- Ad-size- Range [72000-84000] | Mean- 73492.3

This cluster covers the ads of same range as cluster 2, but generates less Revenue/spent compared to cluster 2.



	Count	Mean	Std.
CTR	6899.0	15.814308	7.420170
CPM	6899.0	14.614285	10.511735
CPC	6899.0	0.101779	0.045943

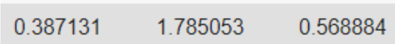
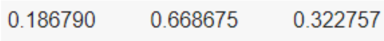
Ad Width > Ad Length

This cluster has highest Fee=0.34.Its CTR is around 15.81, we are almost getting 16 clicks when ad is shown 100 times. This type of ad has highest CPM of 14.61 that is spend / 1000 times ad is shown is 14.61. And spend/click is lowest that is CPC = 0.10.

Standard deviation of CTR, CRM and CPC is 7.4, 10.51 and 0.045 respectively. This is quite high Standard deviation. Variability in CTR, CPM and CPC is high.

This type of Ad is capable of generating almost 0.65% of the revenue of total spends.

Cluster 4- Ad-size- Range [33600-75000] | Mean- 53766.2

	CTR	CPM	CPC
Mean	0.387131	1.785053	0.568884
Std	0.186790	0.668675	0.322757

This cluster covers the ads where Ad Size is lowest compared to other clusters.

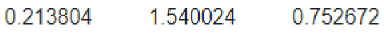
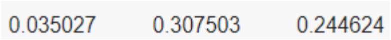
Ad length > Ad- width

Its CTR is around 0.38; we are almost getting 0.38 clicks when ad is shown 100 times. This type of ad has of second lowest CPM of 1.78 and spend/click that is second highest CPC = 0.56.

Standard deviation of CTR, CPM and CPC is 0.18, 0.66 and 0.32 respectively. This is quite high Standard deviation. Variability in CTR, CRM and CPC is high compared to mean values.

This type of Ad is capable of generating almost 0.65% of the revenue of total spends.

Cluster 5- Ad-size- Range [65520- 84000] | Mean- 75234

	CTR	CPM	CPC
Mean	0.213804	1.540024	0.752672
Std	0.035027	0.307503	0.244624

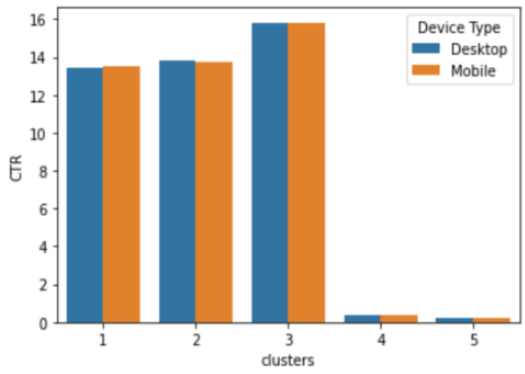
This cluster covers the ads where Ad Size is medium compared to other clusters.

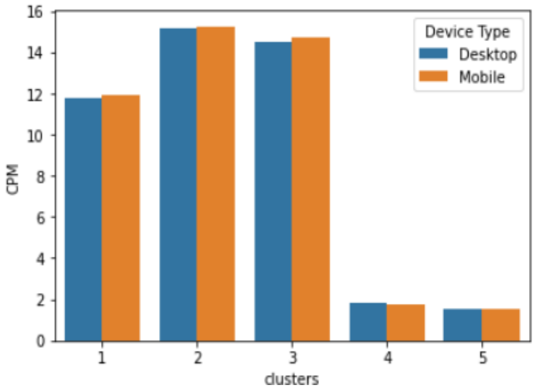
Ad length > Ad- width

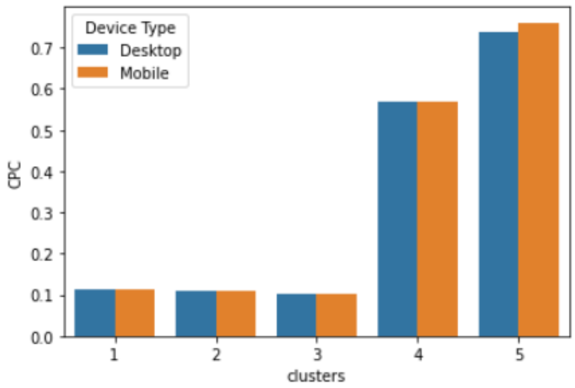
Its CTR is around 0.21(Lowest); we are almost getting 0.21 clicks when ad is shown 100 times. This type of ad has of lowest CPM of 1.54 and spend/click that is highest CPC = 0.75.

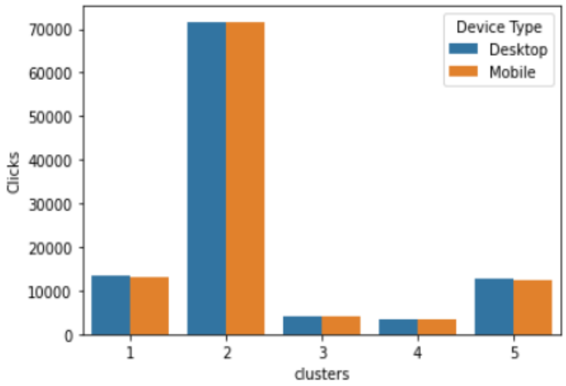
Standard deviation of CTR, CRM and CPC is 0.03, 0.30 and 0.24 respectively. This is quite low Standard deviation. Variability in CTR, CPM and CPC is low compared to cluster 4.

This type of Ad is capable of generating highest of all almost 0.74% of the revenue of total spends.Let’s compare if Device types makes any difference-







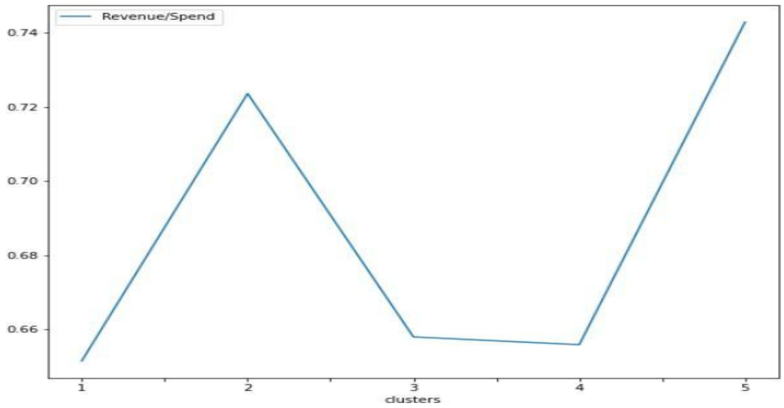


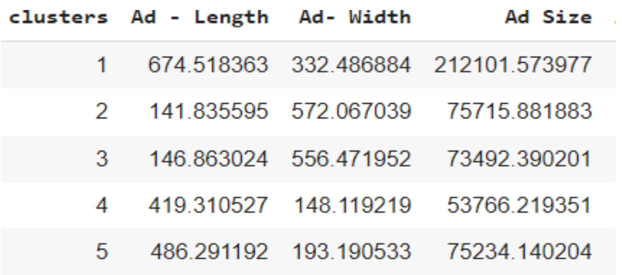
-There is as such difference that Device type is creating. Properties are same for both Mobile and Desktop compared around all the clusters.

Part 1 - Clustering:

1.9- Conclude the project by providing summary of your learnings.

We have noticed that cluster 2 and cluster 5 is generating a greater revenue/spend that is 0.72 and 0.74 respectively compared to other clusters





clusters	Ad - Length	Ad- Width	Ad Size
1	674.518363	332.486884	212101.573977
2	141.835595	572.067039	75715.881883
3	146.863024	556.471952	73492.390201
4	419.310527	148.119219	53766.219351
5	486.291192	193.190533	75234.140204

To spend the current seed funding in a better way we should use the patterns of cluster 2 and cluster 5.

Properties-

Standard deviation is quite low as compared to other clusters.

Mean of Fee (The percentage of the Advertising Fees payable by Franchise Entities.) is around 0.27 for both the clusters.

Inference1- If your ad is bound to have Ad length < Ad width.

Keep ad length and width ratio as 1: 5 and ad area around 72000.

Then

Its CTR is should be around 13.77, maintaining standard deviation of around 1.2 . This type of should have CPM of 15.85 that is spend / 1000 times ad is shown is 15.85 with a SD of around 3.39. And spend/click should be constant around CPC = 0.11, with a standard deviation of around 0.02.

Inference 2-If your ad is bound to have Ad length > Ad width.

Keep ad length and width ratio as 3:1 and ad area around 75000.

Then

Its CTR is should be around 0.21, maintaining standard deviation of around 0.03. This type of should have CPM of 1.54 that is spend / 1000 times ad is shown is 1.54 with a SD of around 0.30. And spend/click should be constant around CPC = 0.75, with a standard deviation of around 0.24.

END