DATA MINING PROJECT BUISNESS REPORT

Shubham Nagargoje july2022 DSBA

Content Table

Part 2 –

1.PCA: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.

2.PCA: Perform detailed Exploratory analysis by creating certain questions like (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio? (Example Questions). Pick 5 variables out of the given 24 variables below for EDA: No_HH, TOT_M, TOT_F, M_06, F_06, M_SC, F_SC, M_ST, F_ST, M_LIT, F_LIT, M_ILL, F_ILL, TOT_WORK_M, TOT_WORK_F, MAINWORK_M, MAINWORK_F, MAIN_CL_M, MAIN_CL_F, MAIN_AL_M, MAIN_AL_F, MAIN_HH_M, MAIN_HH_F, MAIN_OT_M, MAIN_OT_F

3 - PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?

4. - PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.

5 - PCA: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.

6 - PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.

7 - PCA: Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.

8 - PCA: Write linear equation for first PC.

PART 1

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.

Reading the basic data set given

	Timestamp	Inventory Type	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.0
2	2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0
3	2020-9-3-20	Format1	300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.0
4	2020-9-4-15	Format1	300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.0

Applying basic EDA to explore the data , info,summary,null values and duplicates

```
<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
```

There are 6 float, 7 integer and 6 object data types

```
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
```

So there are 4736 null values in CTR CPM CPC.

There are no duplicate records present in thr dataframe
Summary

- There are 6 categorical type and 13 numerical type of datatypes.
- No null values except CPC CPM CTR
- There are no duplicates found
- Data is not scaled and has a great difference range of datapoints.

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

For CPM , CPC , CPM

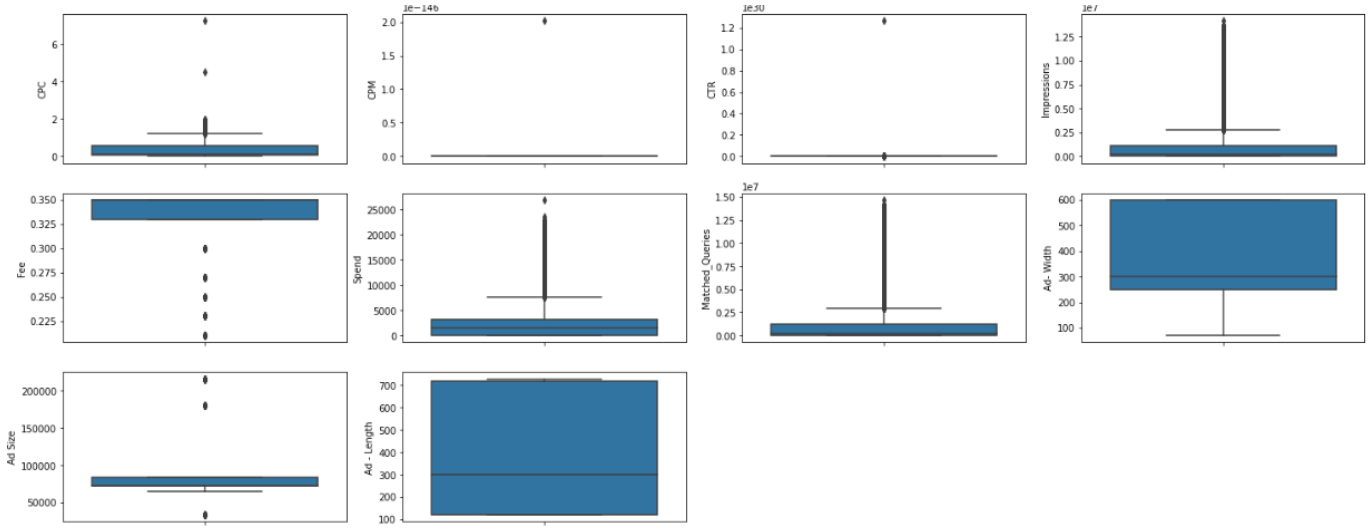
CPM = (Total Campaign Spend / Number of Impressions) * 1,000.

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

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

Almost all the values are filled by using the givem formula.

3 - 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).

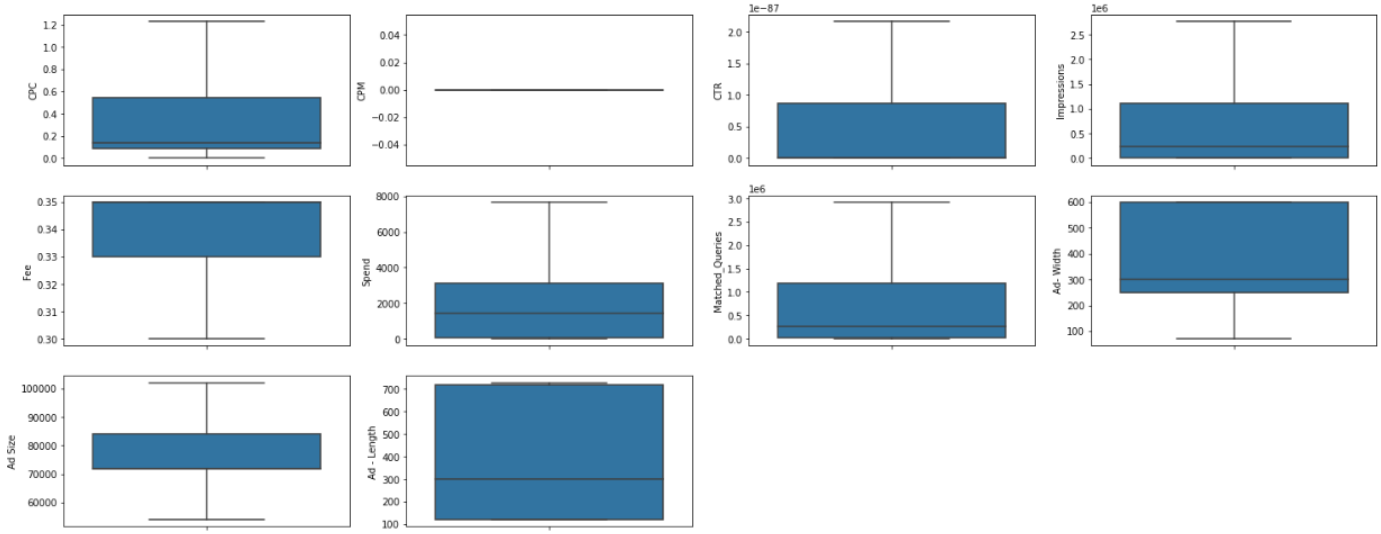


- There are outliers present in the data and i think that for k means clustering , we should treat the outliers.

There are two ways to treate outliers,

- Z_score method normalising data with z score, mean=0 , std=1
- MIn-Max method assinging the maximum value to the outliers greater than upper range=Q3+(1.5 * IQR) and lower_range= Q1-(1.5 * IQR) to the minimum value.

Here we will use min max method to treat the outliers.



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

	Timestamp	Inventory Type	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.0
2	2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0
3	2020-9-3-20	Format1	300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.0
4	2020-9-4-15	Format1	300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.0

after appling z score the data is converted to.

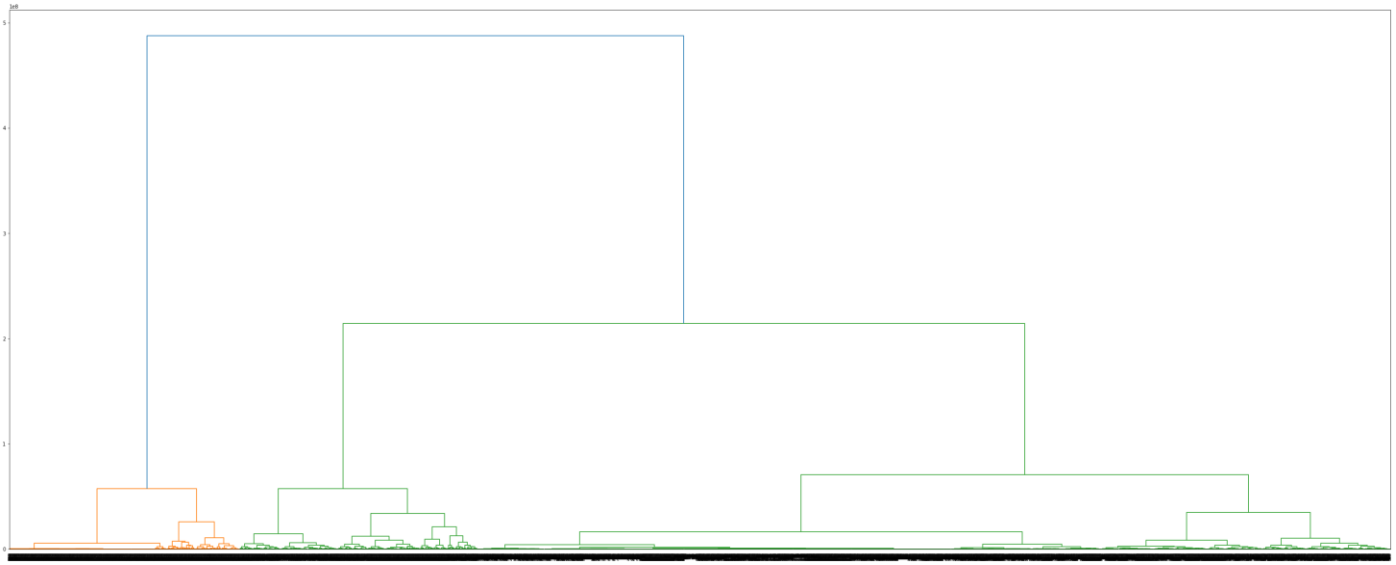
	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	-0.364496	-0.432797	-0.102518	-0.755333	-0.778949	-0.768478	-0.867488	-0.89317	0.535724	-0.880093	-0.586626	NaN	-1.042561
1	-0.364496	-0.432797	-0.102518	-0.755345	-0.778988	-0.768516	-0.867488	-0.89317	0.535724	-0.880093	-0.586626	NaN	-1.042561
2	-0.364496	-0.432797	-0.102518	-0.754900	-0.778919	-0.768445	-0.867488	-0.89317	0.535724	-0.880093	-0.586626	NaN	-1.042561
3	-0.364496	-0.432797	-0.102518	-0.755040	-0.778781	-0.768302	-0.867488	-0.89317	0.535724	-0.880093	-0.586626	NaN	-1.042561
4	-0.364496	-0.432797	-0.102518	-0.755610	-0.779030	-0.768560	-0.867488	-0.89317	0.535724	-0.880093	-0.586626	NaN	-1.042561

And has been normalised to a range.

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Rev
count	2.306600e+04	2.306600e+04	2.306600e+04	2.306600e+04	2.306600e+04	2.306600e+04	2.306600e+04	2.306600e+04	2.306600e+04	2.30660(
mean	-4.030447e-15	5.390161e-15	-4.156304e-15	-3.617510e-15	1.341008e-15	-1.224345e-15	1.960656e-15	1.250852e-15	-2.322121e-14	3.13622
std	1.000022e+00	1.000022e+00	1.000022e+00	1.000022e+00	1.000022e+00	1.000022e+00	1.000022e+00	1.000022e+00	1.000022e+00	1.00002
min	-1.134891e+00	-1.319110e+00	-1.467840e+00	-7.561823e-01	-7.792648e-01	-7.688060e-01	-8.674882e-01	-8.931702e-01	-2.222416e+00	-8.800
25%	-1.134891e+00	-4.327968e-01	-2.975645e-01	-7.403406e-01	-7.614468e-01	-7.606554e-01	-7.934379e-01	-8.580464e-01	-5.675316e-01	-8.464
50%	-3.644957e-01	-1.865987e-01	-2.975645e-01	-5.285774e-01	-5.277221e-01	-5.389745e-01	-4.054310e-01	-3.055230e-01	5.357244e-01	-3.176
75%	1.433093e+00	1.290590e+00	4.826195e-01	4.330590e-01	3.714976e-01	3.660513e-01	4.686290e-01	3.939323e-01	5.357244e-01	3.89802
max	1.467332e+00	1.290590e+00	1.652896e+00	2.193158e+00	2.070914e+00	2.056111e+00	2.361729e+00	2.271900e+00	5.357244e-01	2.24421

By appling z-score , we have reduced the range of the data , which increases the speed of the alogrithim

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

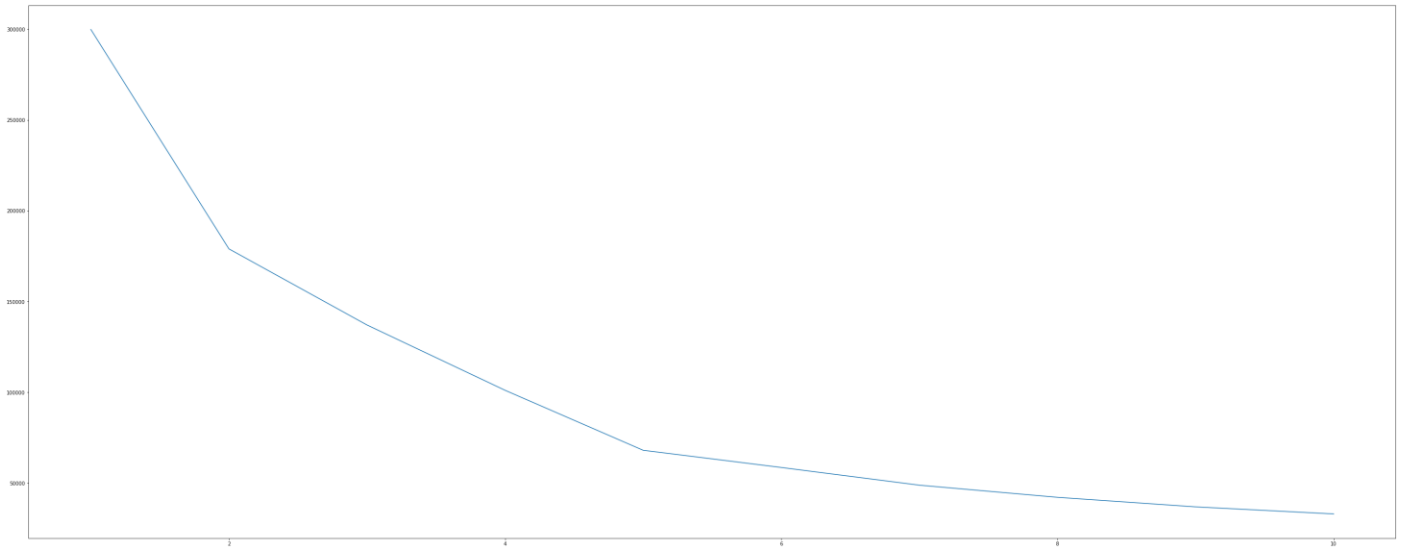


Here, 4 is the optimum number of clusters that should be made.

Clusters	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	
1	468.551724	198.216230	72900.388903	5.811260e+06	2.846720e+06	2.707899e+06	11599.900441	5831.979704	0.312681	3941.835894	5.27
2	429.870025	153.506297	64778.962217	2.569059e+06	1.218455e+06	1.165571e+06	4484.104471	2090.528877	0.347572	1367.305327	1.09
3	326.689973	466.006531	81939.012678	7.434439e+04	4.606658e+04	3.764695e+04	4420.519641	453.424741	0.349848	295.252722	9.75
4	407.890615	324.820800	77651.336234	7.635432e+05	4.444652e+05	3.975864e+05	17202.431712	2993.150706	0.335736	2014.825412	4.95

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

```
The WSS value for 1 clusters is 299857.9999999998
The WSS value for 2 clusters is 178957.67749378303
The WSS value for 3 clusters is 137024.77550794432
The WSS value for 4 clusters is 101036.63569196296
The WSS value for 5 clusters is 67986.59137885594
The WSS value for 6 clusters is 58515.880559607926
The WSS value for 7 clusters is 48763.82989556023
The WSS value for 8 clusters is 42076.58046464563
The WSS value for 9 clusters is 36816.622632624065
The WSS value for 10 clusters is 32928.91438710698
```

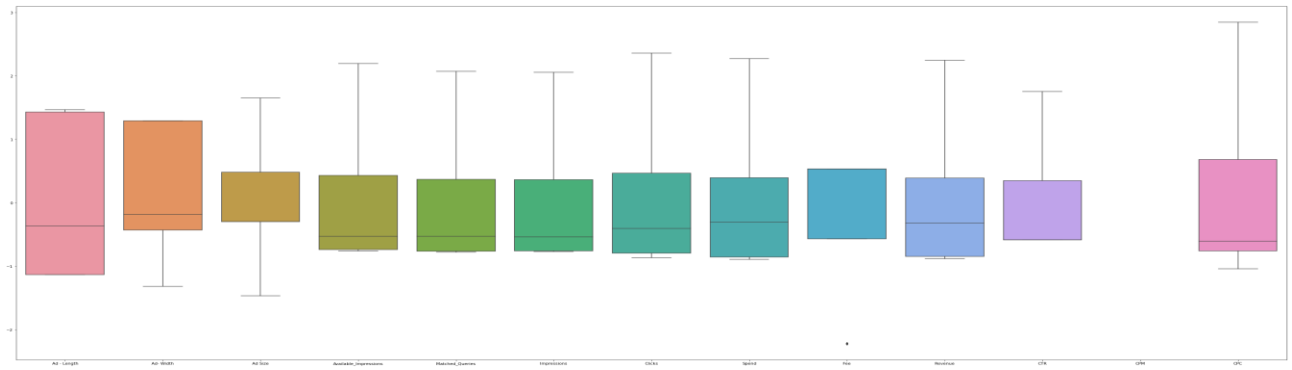


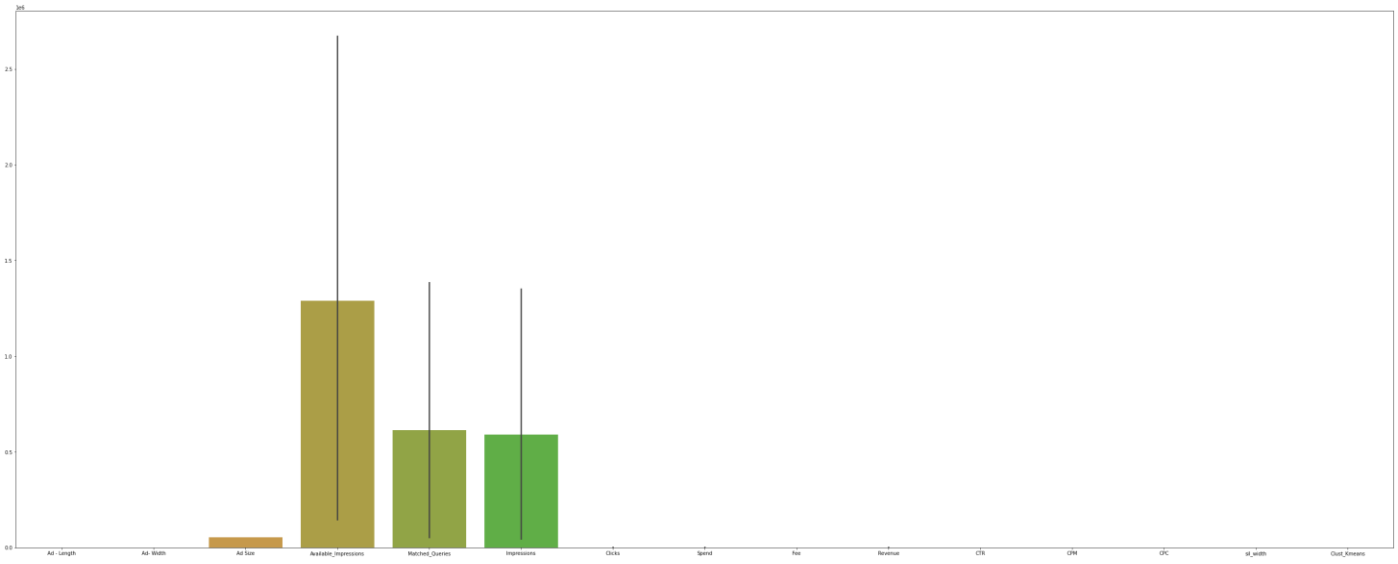
THe OPtimum no. of clusters are 5 for k mean

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

```
sil score for 2  is 0.43796685070315555
sil score for 3  is 0.3380282471161297
sil score for 4  is 0.40200501676284645
sil score for 5  is 0.49502489534287814
sil score for 6  is 0.49190880475307874
sil score for 7  is 0.5042344609844847
sil score for 8  is 0.4895488238624933
sil score for 9  is 0.5230820335171237
sil score for 10  is 0.536145934891218
```

8 - 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].




1e6
2.5
2.0
1.5
1.0
0.5
0.0
Ad - Length
Ad- Width
Ad Size
Available_Impressions
Matched_Queries
Impressions
Clicks
Spend
Fee
Revenue
CTR
CPM
CPC
id_width
Clust_Kmeans

Part 2.

PCA: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.

MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
749	180	237	680	252	32	46	258	214
715	123	229	186	148	76	178	140	160
188	44	89	3	34	0	4	67	61
247	61	128	13	50	4	10	116	59
1928	465	1043	205	302	24	105	180	478

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 640 entries, 0 to 639
Data columns (total 61 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   State Code      640 non-null    int64
 1   Dist.Code       640 non-null    int64
 2   State           640 non-null    object
 3   Area Name       640 non-null    object
 4   No_HH           640 non-null    int64
 5   TOT_M           640 non-null    int64
 6   TOT_F           640 non-null    int64
 7   M_06            640 non-null    int64
 8   F_06            640 non-null    int64
 9   M_SC            640 non-null    int64
 10  F_SC            640 non-null    int64
 11  M_ST            640 non-null    int64
 12  F_ST            640 non-null    int64
 13  M_LIT           640 non-null    int64
 14  F_LIT           640 non-null    int64
 15  M_ILL           640 non-null    int64
 16  F_ILL           640 non-null    int64
 17  TOT_WORK_M      640 non-null    int64
 18  TOT_WORK_F      640 non-null    int64
 19  MAINWORK_M      640 non-null    int64
 20  MAINWORK_F      640 non-null    int64
 21  MAIN_CL_M       640 non-null    int64
 22  MAIN_CL_F       640 non-null    int64
 23  MAIN_AL_M       640 non-null    int64
 24  MAIN_AL_F       640 non-null    int64
 25  MAIN_HH_M       640 non-null    int64
 26  MAIN_HH_F       640 non-null    int64
```

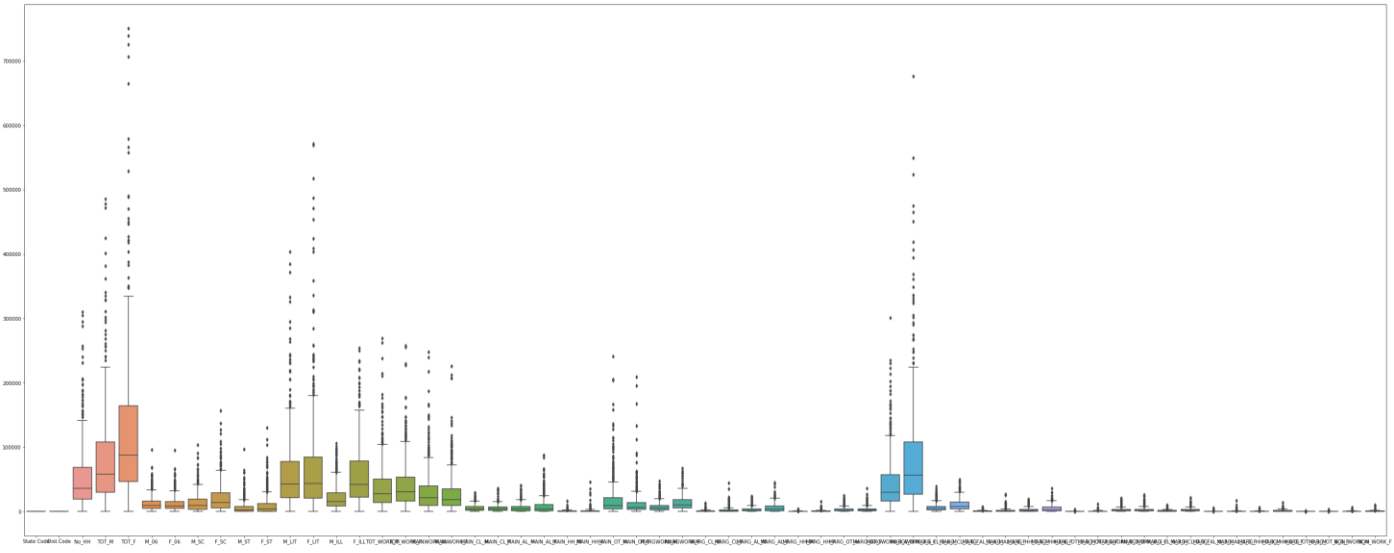
```
State Code       0
Dist.Code        0
State            0
Area Name        0
No_HH            0
                ..
MARG_HH_0_3_F    0
MARG_OT_0_3_M    0
MARG_OT_0_3_F    0
NON_WORK_M       0
NON_WORK_F       0
Length: 61, dtype: int64
```

```
State Code       0
Dist.Code        0
State            0
Area Name        0
No_HH            0
                ..
MARG_HH_0_3_F    0
MARG_OT_0_3_M    0
MARG_OT_0_3_F    0
NON_WORK_M       0
NON_WORK_F       0
Length: 61, dtype: int64
```

Summary,

- There are no null values , and duplicates
- data is highly ranged
- There are many outliers

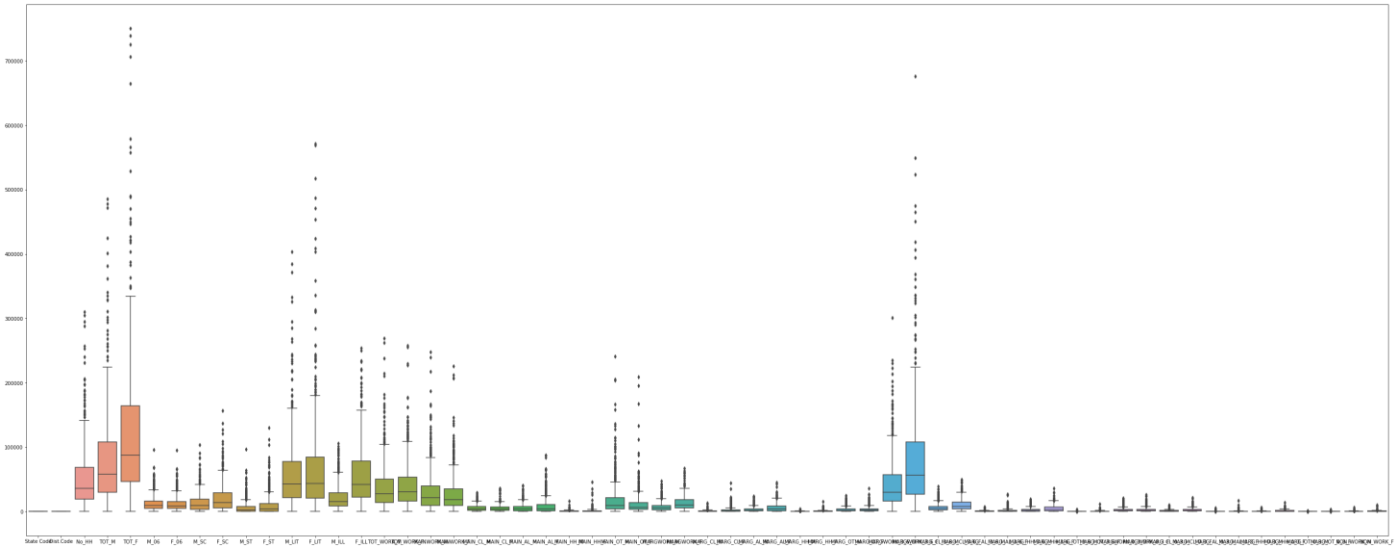
PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?



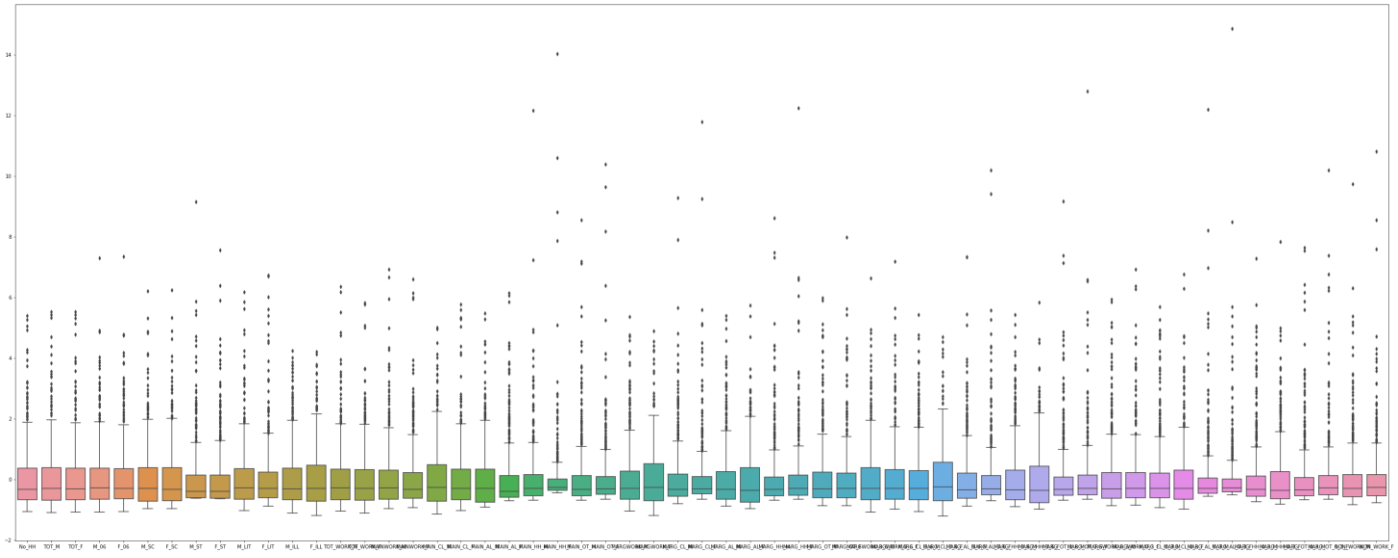
As we can see a large no. of outliers , not treating the outliers will have no affect on the pca as the data will be scaled and it will reduce the effect of outliers.

Additionally , after scaling , it will increase the speed of the operations too.

PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.



Before scaling



After scaling

AS we compare both the boxplots, that is before and after scaling.

- There are still outliers in the data
- The data is scaled to 0 to 15 range.
- After scaling , algorithem speeds up.

PCA: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector

Bartletts Test of Sphericity Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.

H0: All variables in the data are uncorrelated Ha: At least one pair of variables in the data are correlated If the null hypothesis cannot be rejected, then PCA is not advisable.

If the p-value is small, then we can reject the null hypothesis and agree that there is atleast one pair of vairbales in the data wihich are correlated hence PCA is recommended.

from factor_analyzer.factor_

```
0.0
```

KMO Test The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.

Generally, if MSA is less than 0.5, PCA is not recommended, since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

```
0.8039889932781528
```

This confirms the statistical significance of corelations.

```
#eigen vector
pca.components_
```

```
array([[ 1.56020579e-01,  1.67117635e-01,  1.65553179e-01,
         1.62192948e-01,  1.62566396e-01,  1.51357849e-01,
         1.51566500e-01,  2.72341946e-02,  2.81833150e-02,
         1.61992837e-01,  1.46872680e-01,  1.61749445e-01,
         1.65248187e-01,  1.59871988e-01,  1.45935804e-01,
         1.46200730e-01,  1.23970284e-01,  1.03127159e-01,
         7.45397856e-02,  1.13355712e-01,  7.38821590e-02,
         1.31572584e-01,  8.33826397e-02,  1.23526242e-01,
         1.11021264e-01,  1.64615479e-01,  1.55395618e-01,
         8.23885414e-02,  4.91953957e-02,  1.28598563e-01,
         1.14305073e-01,  1.40853227e-01,  1.27669598e-01,
         1.55262872e-01,  1.47286584e-01,  1.64971950e-01,
         1.61253433e-01,  1.65501611e-01,  1.55647049e-01,
         9.30142064e-02,  5.15358640e-02,  1.28576116e-01,
         1.10645843e-01,  1.39592763e-01,  1.24545909e-01,
         1.54293786e-01,  1.46285654e-01,  1.50125706e-01,
         1.40157047e-01,  5.25417829e-02,  4.17859530e-02,
         1.21840354e-01,  1.16011410e-01,  1.39868774e-01,
         1.32192245e-01,  1.50375578e-01,  1.31066203e-01],
```

```
#covarience matrix
cov_matrix=data_scaled.cov()
cov_matrix
```

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_0_3_M	MARG_CL_0_
No_HH	1.001565	0.917604	0.972109	0.798807	0.797619	0.776522	0.825137	0.149861	0.165361	0.933397	...	0.557813	0.556
TOT_M	0.917604	1.001565	0.984178	0.952313	0.949275	0.841240	0.827592	0.091565	0.086315	0.990860	...	0.699403	0.596
TOT_F	0.972109	0.984178	1.001565	0.909396	0.907975	0.818238	0.834059	0.123819	0.128848	0.986983	...	0.656373	0.599
M_06	0.798807	0.952313	0.909396	1.001565	0.999713	0.782342	0.748700	0.055361	0.044017	0.914186	...	0.761800	0.648
F_06	0.797619	0.949275	0.907975	0.999713	1.001565	0.774345	0.742846	0.065240	0.054748	0.909062	...	0.764809	0.650
M_SC	0.776522	0.841240	0.818238	0.782342	0.774345	1.001565	0.986612	-0.045738	-0.047900	0.819765	...	0.674687	0.570
F_SC	0.825137	0.827592	0.834059	0.748700	0.742846	0.986612	1.001565	-0.014144	-0.009204	0.815424	...	0.651473	0.586
M_ST	0.149861	0.091565	0.123819	0.055361	0.065240	-0.045738	-0.014144	1.001565	0.989593	0.090682	...	0.123160	0.196
F_ST	0.165361	0.086315	0.128848	0.044017	0.054748	-0.047900	-0.009204	0.989593	1.001565	0.087512	...	0.121601	0.217
M_LIT	0.933397	0.990860	0.986983	0.914186	0.909062	0.819765	0.815424	0.090682	0.087512	1.001565	...	0.653528	0.560
F_LIT	0.929539	0.933166	0.958510	0.833812	0.830425	0.715056	0.729895	0.100645	0.101050	0.969471	...	0.548153	0.485
M_ILL	0.764235	0.912965	0.859542	0.946889	0.950093	0.802028	0.763753	0.083193	0.072703	0.843152	...	0.745823	0.626
F_ILL	0.863423	0.886746	0.888305	0.864675	0.866643	0.834017	0.848529	0.138247	0.149727	0.835690	...	0.709562	0.673
TOT_WORK_M	0.939667	0.971936	0.970471	0.857112	0.854127	0.826064	0.824978	0.122835	0.119357	0.978480	...	0.601813	0.514
TOT_WORK_F	0.926707	0.809159	0.877604	0.684563	0.686421	0.714087	0.778146	0.265163	0.285420	0.816644	...	0.493600	0.549
MAINWORK_M	0.928079	0.934292	0.942489	0.790929	0.786017	0.779710	0.783570	0.113785	0.109484	0.954792	...	0.473037	0.393
MAINWORK_F	0.892700	0.745533	0.824110	0.585895	0.586599	0.645150	0.713990	0.231171	0.246706	0.769242	...	0.303333	0.340
MAIN_CL_M	0.432077	0.532567	0.488420	0.562042	0.562478	0.609108	0.579425	0.099419	0.083256	0.469297	...	0.465574	0.392

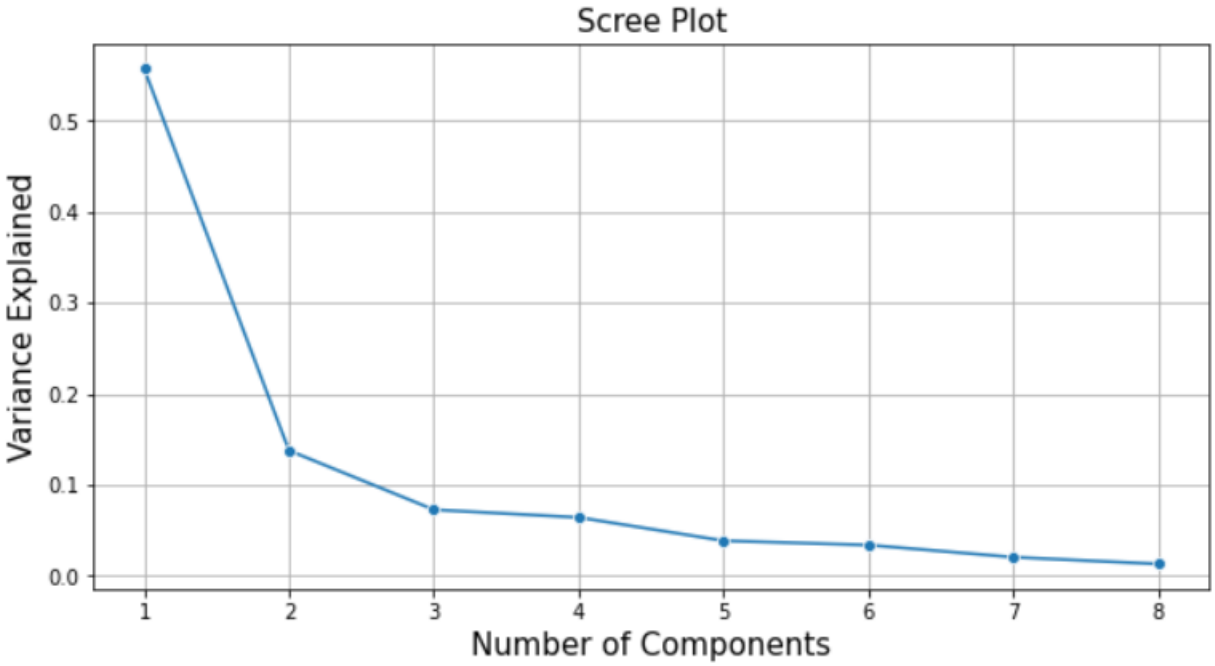
Eigen values ,

```
array([0.55726063, 0.13784435, 0.07275295, 0.06426418, 0.03865049,
       0.03395169, 0.02060239, 0.01315764])
```

PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.¶

```
#eigen values
var_exp=pca.explained_variance_ratio_
var_exp
```

```
array([0.55726063, 0.13784435, 0.07275295, 0.06426418, 0.03865049,
       0.03395169, 0.02060239, 0.01315764])
```




```
Cumulative Variance Explained [0.55726063 0.69510499 0.76785794 0.83212212 0.87077261 0
.9047243
 0.92532669 0.93848433]
```

The cumlative explained variance ratio is over almost 92.5% in 7 pca's.