PART 1

List of Figures

List of Tables

Data Description

We have a data on The ads24*7, Digital Marketing company which has now got seed funding $10M.They are expanding wings in Marketing Analytics. There is the data description

```
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
```

Where we can see Timestamp, Inventory, Ad Type, Platform, Device Type, Format is a object data type and some of int64 and float64 datatypes are there.

TEST FOR MISSING VALUES AND TREAT THE MISSING VALUES

We can clearly see that in CTR,CPM & CPC has some missing values. Next I use to calculate the missing values using

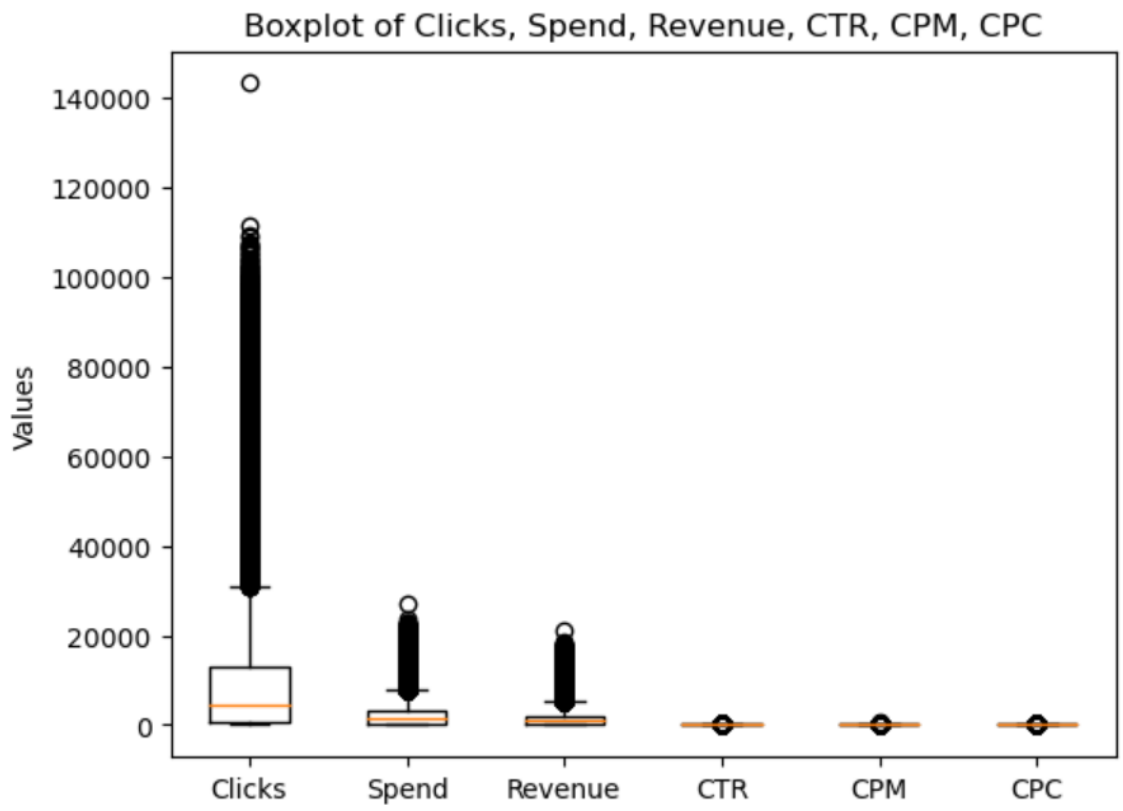
CPM = (Total Campaign Spend / Number of Impressions) * 1,000

CPC = Total Cost (spend) / Number of Clicks

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

These formulas.

CHEKING IF THERE ANY OUTLIERS



In this boxplot we can clearly see there is outliers available.

I do not think treating outliers is necessary. Because if we treat the outliers it will impact on the data set. It will change the entire data.

PERFORM Z-SCORE

In the data set there are 23066 rows and 21 columns which are unscaled data. Here is the sample data

	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	

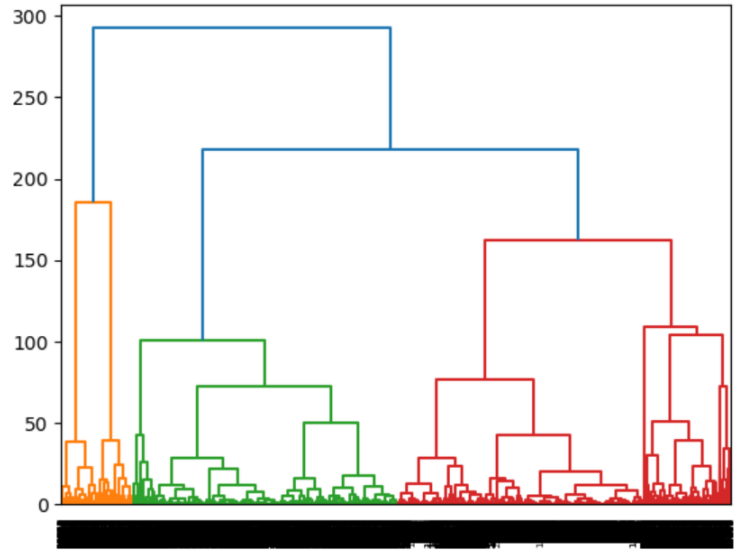
We cannot go through the clustering process with uncalled data so here we apply z-score.

Here is the sample

	Clicks	Spend	Revenue	CTR	CPM	CPC
0	-0.615311	-0.665372	-0.619693	-0.332572	-0.92711	-0.986603
1	-0.615311	-0.665372	-0.619693	-0.332521	-0.92711	-0.986603
2	-0.615311	-0.665372	-0.619693	-0.332610	-0.92711	-0.986603
3	-0.615311	-0.665372	-0.619693	-0.332712	-0.92711	-0.986603
4	-0.615311	-0.665372	-0.619693	-0.332444	-0.92711	-0.986603

HIERARCHICAL CLUSTERING AND DENDROGRAM

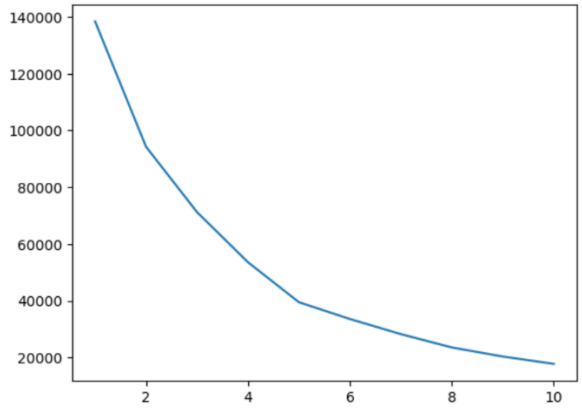
In hierarchical clustering we take all the rows as a cluster and we check the dendrogram.



Dendrogram suggests for 3 cluster.

ELBOW PLOT

For k-means algorithm we take hole data as 1cluster and end with each row as a cluster. Here we observe the WSS plot to take the number of cluster.



In this plot we can see that 5 is the significant optimum number taking as cluster number.

SILHOUETTE SCORES

the silhouette score for a clustering solution is greater than 0.5, it indicates that the clusters are relatively well-separated and the data points within each cluster are more similar to each other than to data points in other clusters.

the clustering results have a moderate to good level of separation and coherence based on the silhouette score

SAMMARY

Clustering technique is used to make some group on different behaviours of customers. It was used to identified them and put them in a group for taking a decision.

In this project I use the clustering technique, make some cluster and find there sum, average and see each clusters different characteristics.

PART 2

LIST OF FIGURES

LIST OF TABLES

Data Description

In this data set we have 640 rows and 61 columns. Here we use PCA for dimensionality reduction . we only take the independent variables for this technique. We consider each variable as a pc and reduce its dimensionality before machine learning technique. Here is the sample data

	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_A
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180	
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123	
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	...	114	188	44	
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	...	194	247	61	
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	...	874	1928	465	

STATE AND DISTRICT HAS HIGHEST AND LOWEST GENDER RATIO

```
State with the highest gender ratio: Andhra Pradesh
State with the lowest gender ratio: Lakshadweep
District with the highest gender ratio: 547
District with the lowest gender ratio: 587
```

OUTLIERS

Here we were asked not to choose outliers. But PCA is a technique which depends on the means. As we know means are highly sensitive of outliers. So treating outliers in PCA is necessary.

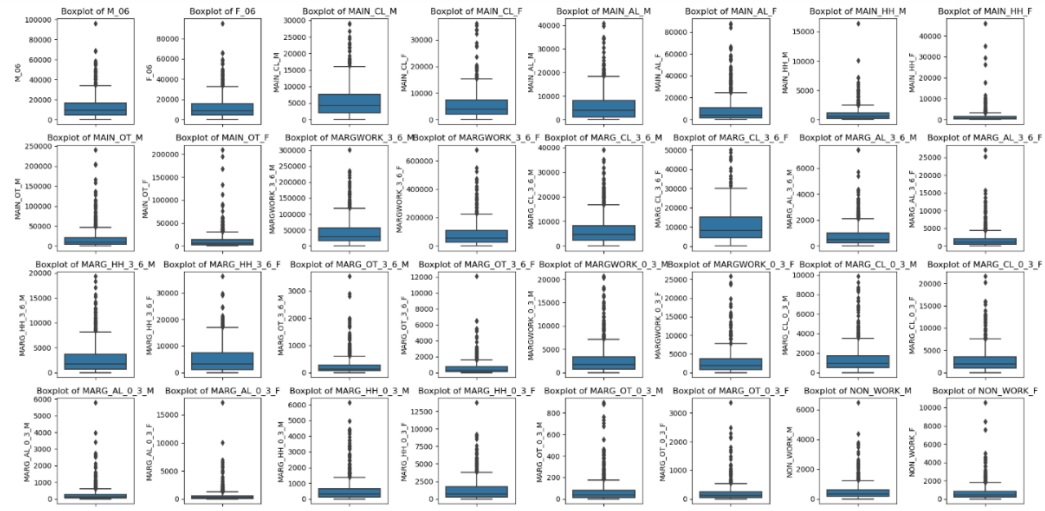
SCALING THE DATA

we use z-score for scaling the data.

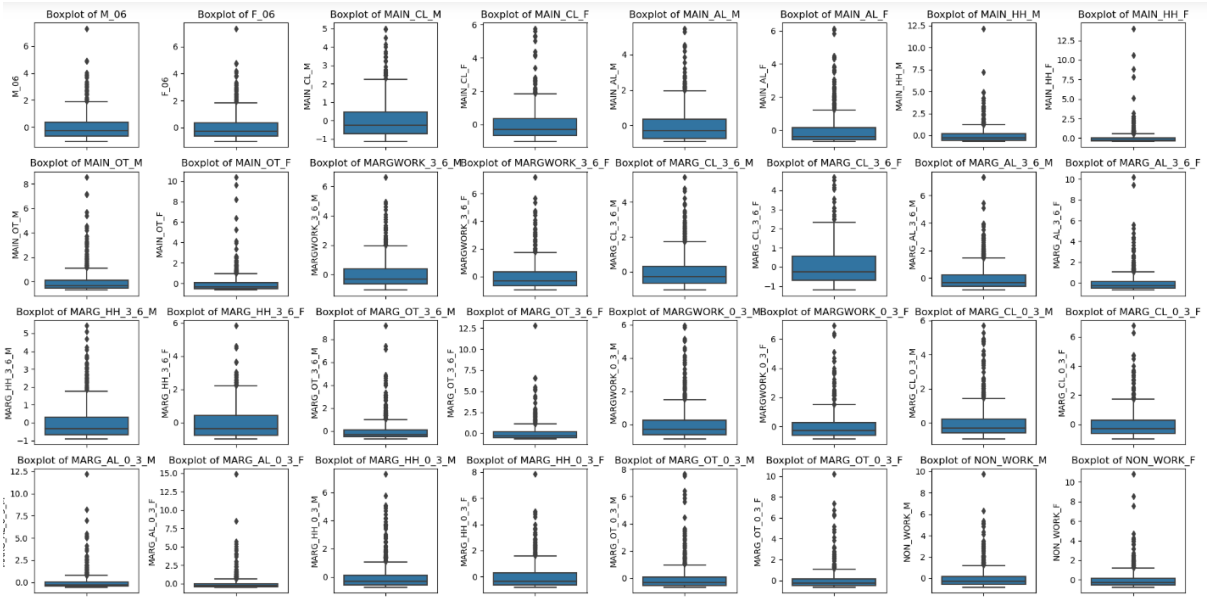
z-score does not impact outliers.

Here is the example

Before z-score



after z-score

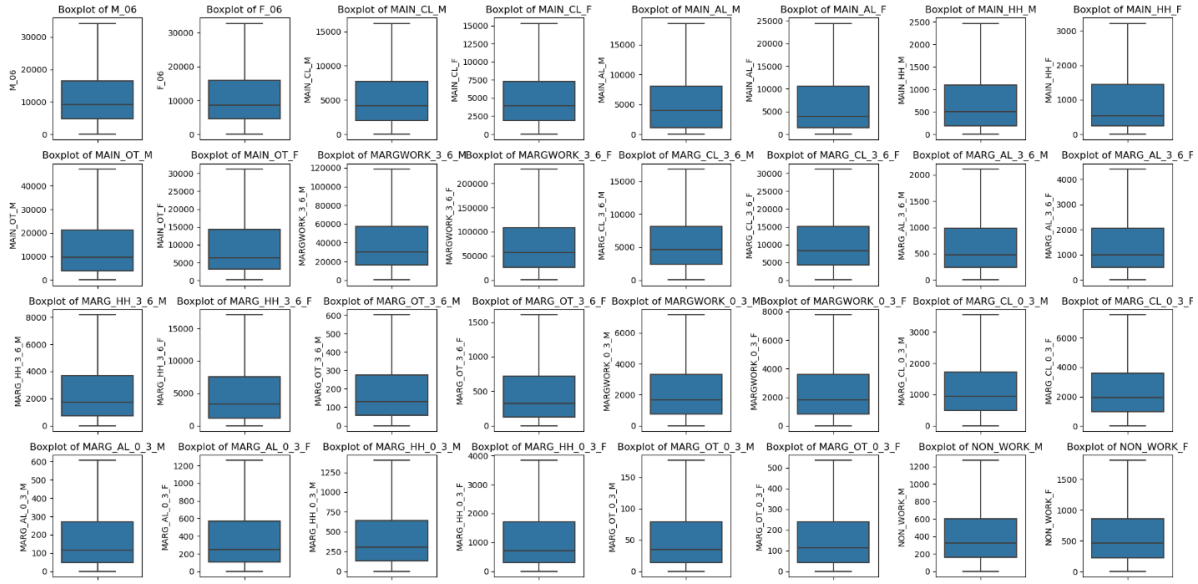


z-score use for scaling the data set, it does not effect the outliers. In PCA treating outliers and z-score both are important.

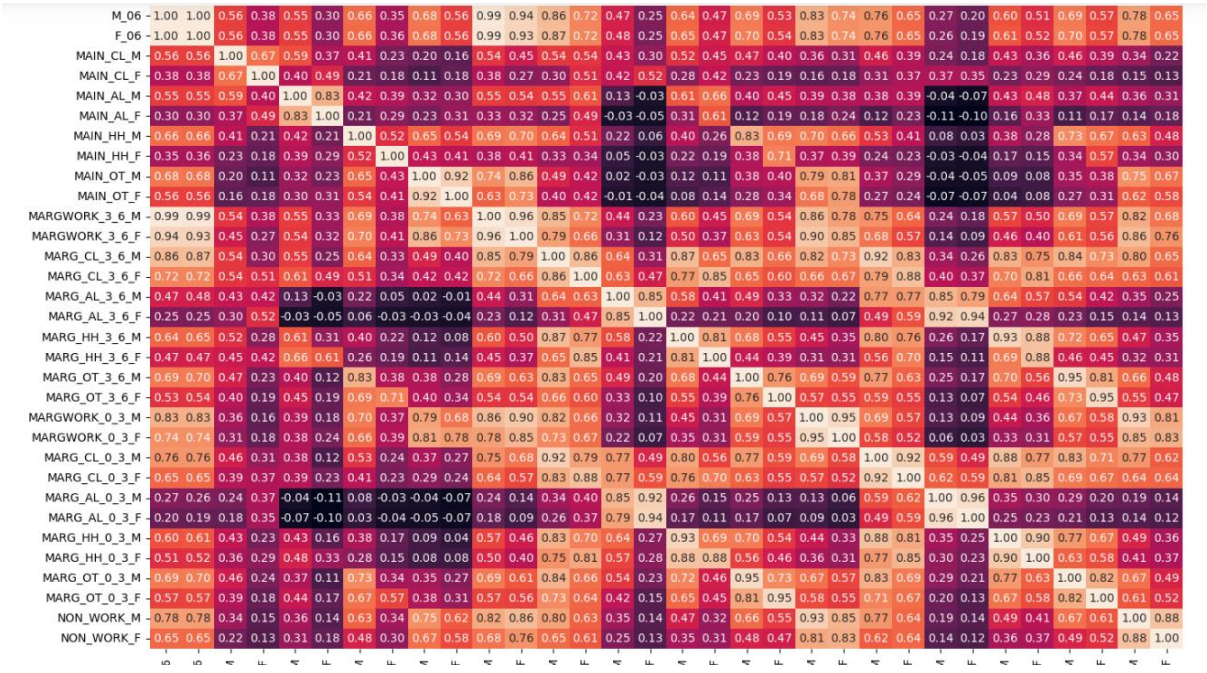
PERFORMING PCA

PCA cannot go with the outliers and its works on only independent variables.

So fast I use to treat the outliers.



In the boxplot we can see there are only independent variables and containing no outliers.

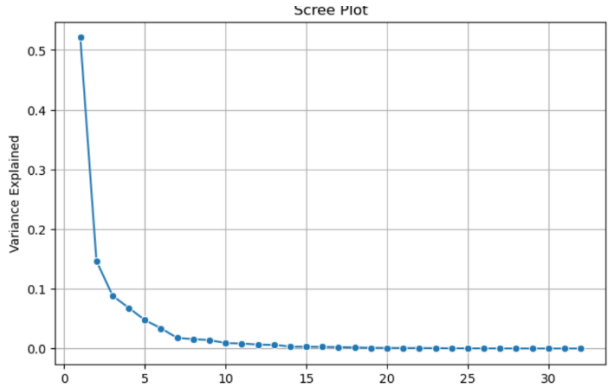


Here is the covariance matrix.

IDENTIFING OPTIMUM NUMBER OF PCs

To identify the optimal number of pcs I use the see the scree plot and explained variance ratio

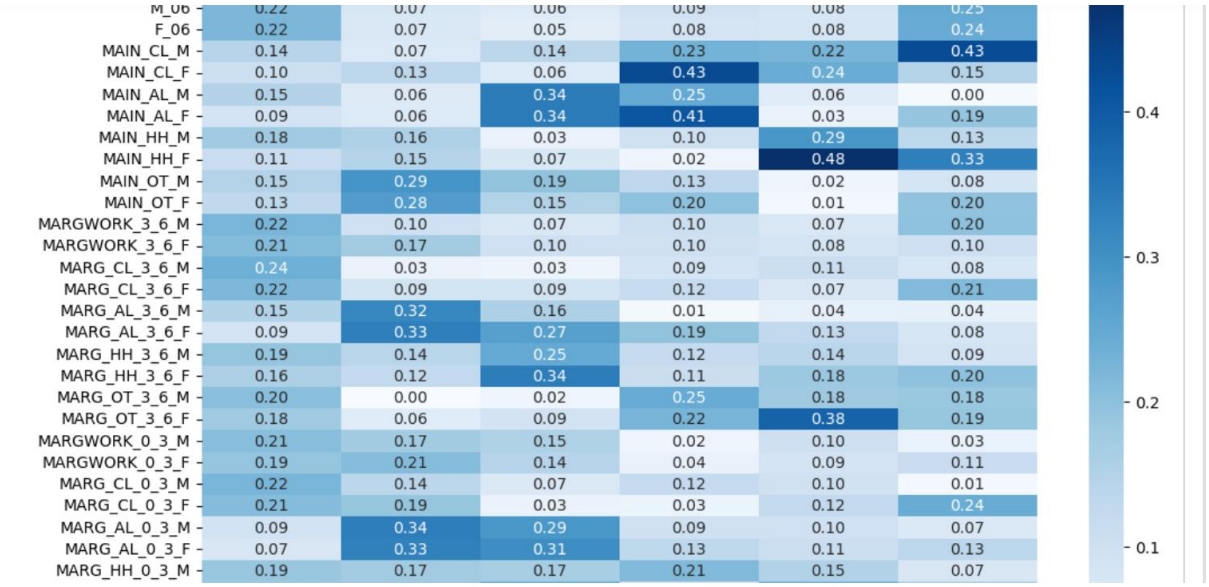
```
[0.52120054, 0.6677232 , 0.75618259, 0.82417856, 0.87170677,
 0.90546198, 0.92310679, 0.93879236, 0.952818  , 0.96188109,
 0.97010962, 0.97658872, 0.98265229, 0.9857252 , 0.98868869,
 0.99133356, 0.9935503 , 0.99511636, 0.99615724, 0.99704083,
 0.99780206, 0.99849627, 0.99901355, 0.99936005, 0.99965068,
 0.99984142, 0.99996858, 1.        , 1.        , 1.        ,
 1.        , 1.        ])
```



Lets take 90% of explained variance ratio.

PC1 is the most important among all the PCs. Only pc1 contents 52% of the data.

If we take care only pc1 components huge benefits we will get.



LINEAR EQUATION FOR FIRST PC

PC1 = w1 * X1 + w2 * X2 + w3 * X3 + ... + wn * Xn

Where X1, X2, X3, ..., Xn are the original variables, and w1, w2, w3, ..., wn are the coefficients or loadings of the first PC.