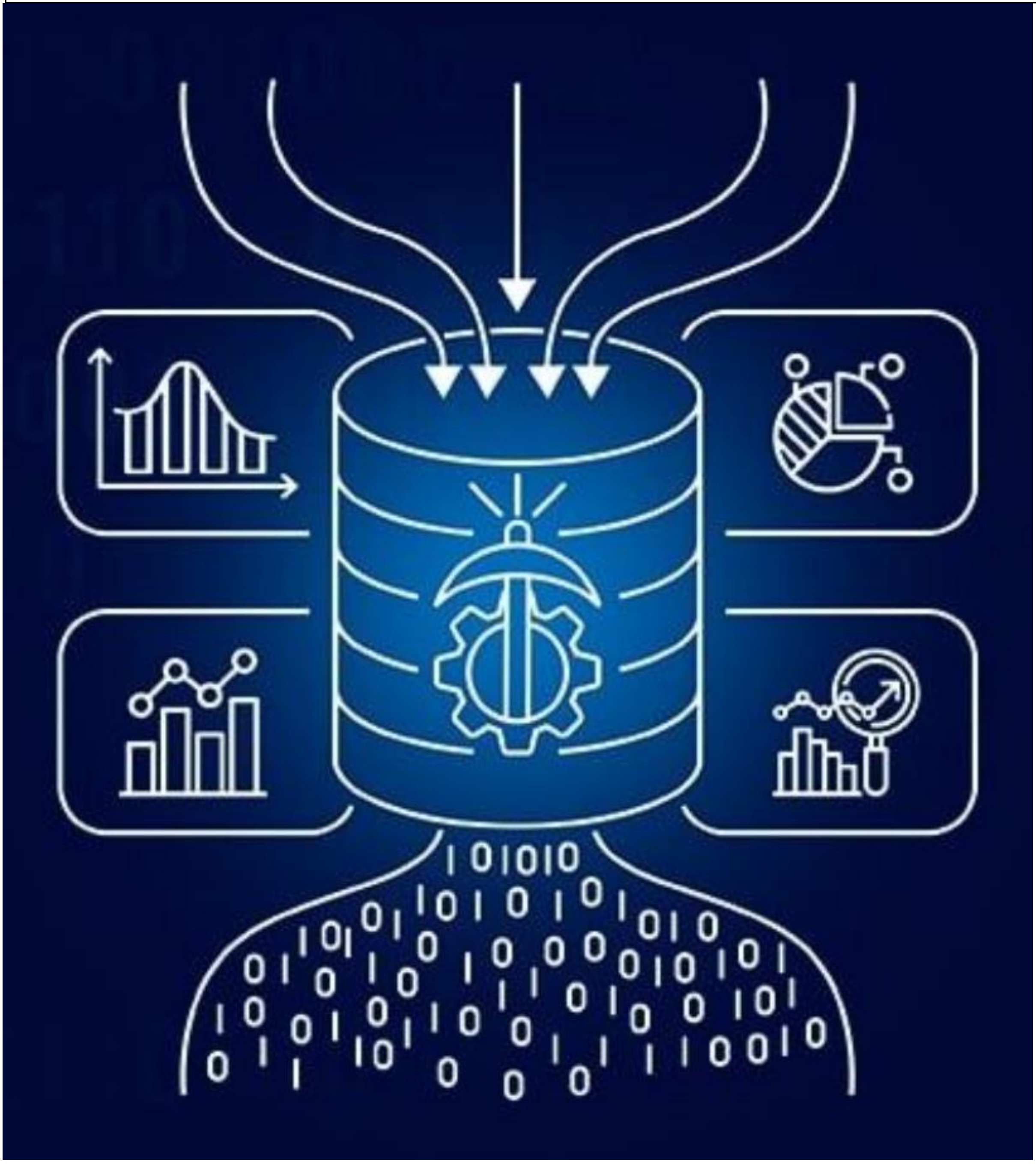
DATA MINING PROJECT ON CLUSTER & PCA



-RAHUL SHARMA

SR NO.	CONTENT	PAGE NO.
A	**CLUSTERING**	1-12
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.	1-3
1.2	Treat missing values in CPC, CTR and CPM using the formula given.	4-5
1.3	Check and treat if there are any outliers.	5-7
1.4	Perform z-score scaling and discuss how it acts the speed of the algorithm	8
1.5	Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.	8-9
1.6	Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.	9
1.7	Print silhouee scores and identify optimum number of clusters.	10

1.8	Profile the ads based on optimum number of clusters using silhouee 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]	11
1.9	Conclude the project by providing summary of your learning	12
B	**PCA**	12-21
2.1	Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.	13-14
2.2	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	14-18
2.3	We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?	18-19
2.4	Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment	18-19
2.5	Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector	19

2.6	Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.	20
2.7	Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.	21
2.8	Write linear equation for first PC	21

PART A:- 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.

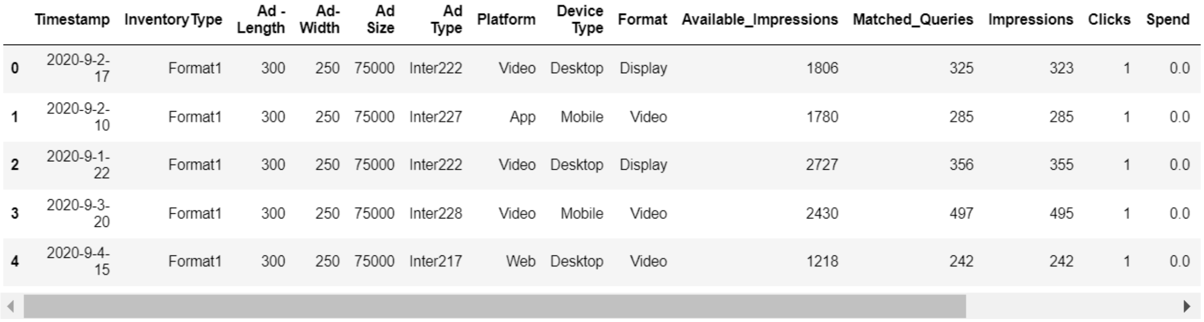
CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column in the dataset and the Number of Clicks refers to the 'Clicks' Column in the dataset.

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.

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.

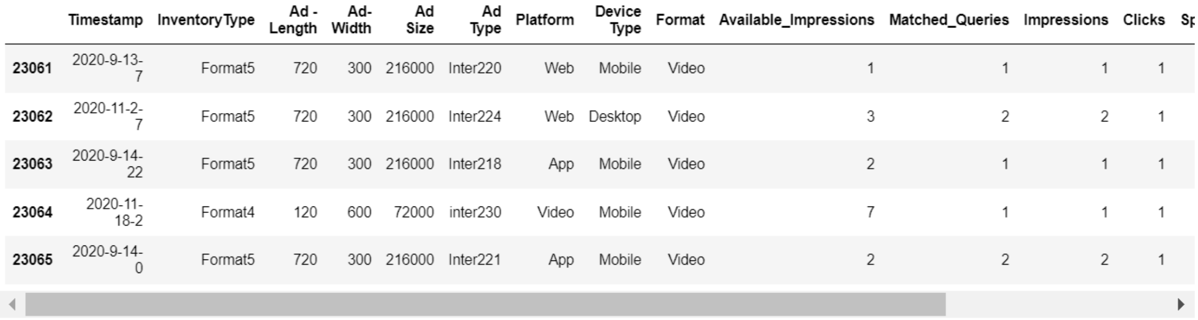
Answer:-

Top 5 rows:-



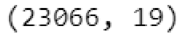
	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	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

Last 5 rows:-



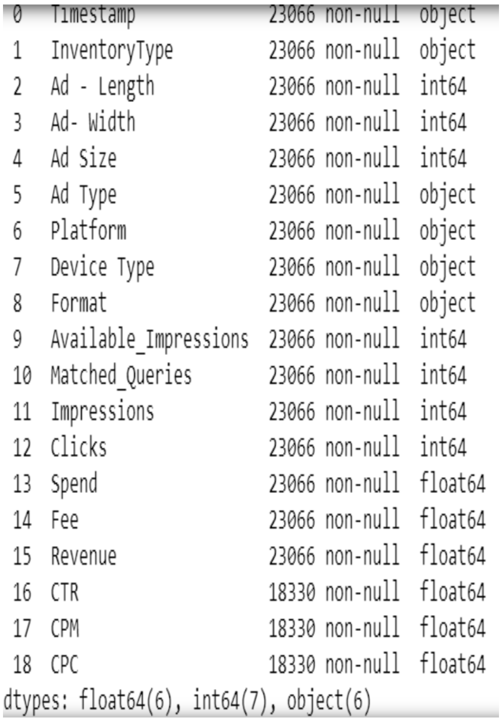
	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Sp
23061	2020-9-13-7	Format5	720	300	216000	Inter220	Web	Mobile	Video	1	1	1	1	
23062	2020-11-2-7	Format5	720	300	216000	Inter224	Web	Desktop	Video	3	2	2	1	
23063	2020-9-14-22	Format5	720	300	216000	Inter218	App	Mobile	Video	2	1	1	1	
23064	2020-11-18-2	Format4	120	600	72000	inter230	Video	Mobile	Video	7	1	1	1	
23065	2020-9-14-0	Format5	720	300	216000	Inter221	App	Mobile	Video	2	2	2	1	

Shape of the dataset:-

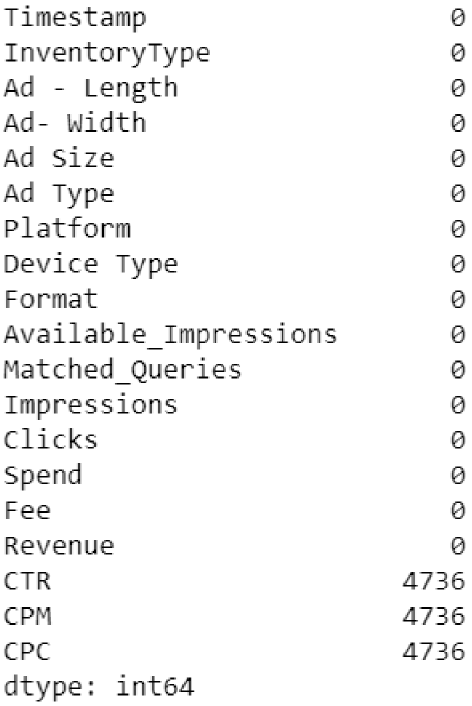


```
(23066, 19)
```

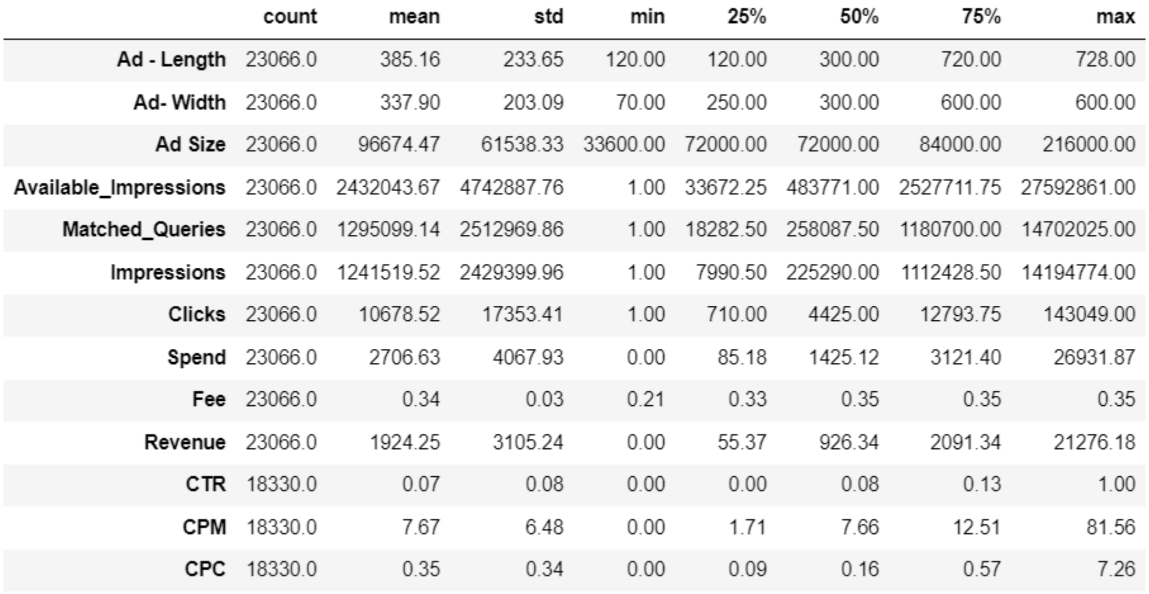
Info of the dataset:-



```
 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)
```



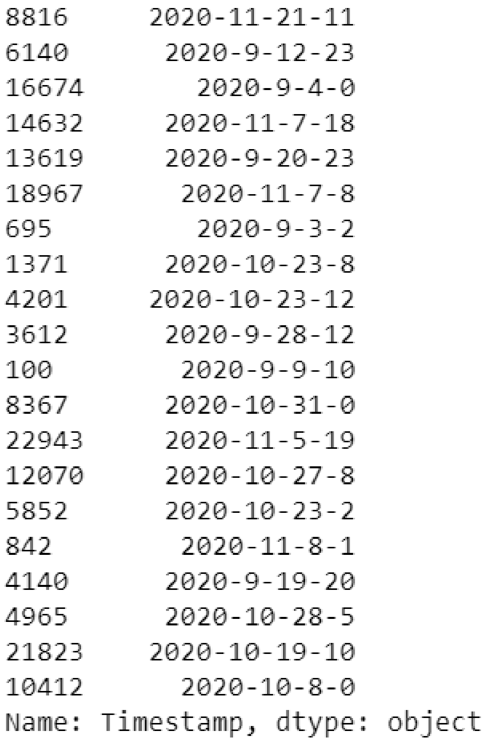
```
Timestamp                   0
InventoryType               0
Ad - Length                 0
Ad- Width                   0
Ad Size                     0
Ad Type                     0
Platform                    0
Device Type                 0
Format                      0
Available_Impressions       0
Matched_Queries             0
Impressions                 0
Clicks                      0
Spend                       0
Fee                         0
Revenue                     0
CTR                      4736
CPM                      4736
CPC                      4736
dtype: int64
```



	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	385.16	233.65	120.00	120.00	300.00	720.00	728.00
Ad- Width	23066.0	337.90	203.09	70.00	250.00	300.00	600.00	600.00
Ad Size	23066.0	96674.47	61538.33	33600.00	72000.00	72000.00	84000.00	216000.00
Available_Impressions	23066.0	2432043.67	4742887.76	1.00	33672.25	483771.00	2527711.75	27592861.00
Matched_Queries	23066.0	1295099.14	2512969.86	1.00	18282.50	258087.50	1180700.00	14702025.00
Impressions	23066.0	1241519.52	2429399.96	1.00	7990.50	225290.00	1112428.50	14194774.00
Clicks	23066.0	10678.52	17353.41	1.00	710.00	4425.00	12793.75	143049.00
Spend	23066.0	2706.63	4067.93	0.00	85.18	1425.12	3121.40	26931.87
Fee	23066.0	0.34	0.03	0.21	0.33	0.35	0.35	0.35
Revenue	23066.0	1924.25	3105.24	0.00	55.37	926.34	2091.34	21276.18
CTR	18330.0	0.07	0.08	0.00	0.00	0.08	0.13	1.00
CPM	18330.0	7.67	6.48	0.00	1.71	7.66	12.51	81.56
CPC	18330.0	0.35	0.34	0.00	0.09	0.16	0.57	7.26

Duplicates of the dataset:- zero.

Changing Datatype of Timestamp from Object to datetime64:-



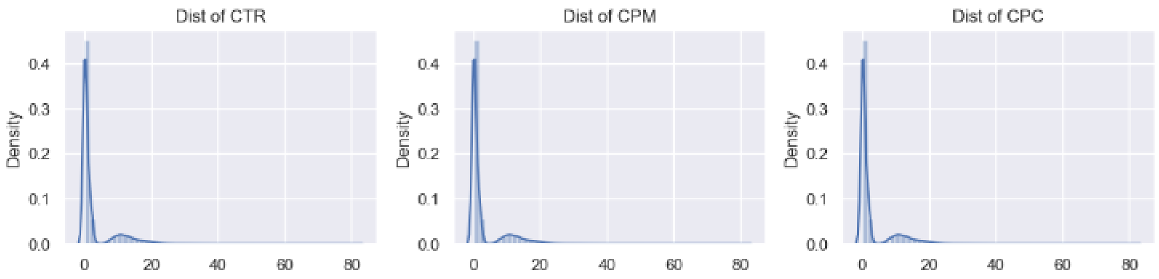
```
8816      2020-11-21-11
6140       2020-9-12-23
16674        2020-9-4-0
14632      2020-11-7-18
13619      2020-9-20-23
18967       2020-11-7-8
695          2020-9-3-2
1371       2020-10-23-8
4201      2020-10-23-12
3612       2020-9-28-12
100         2020-9-9-10
8367       2020-10-31-0
22943      2020-11-5-19
12070      2020-10-27-8
5852       2020-10-23-2
842         2020-11-8-1
4140       2020-9-19-20
4965       2020-10-28-5
21823     2020-10-19-10
10412       2020-10-8-0
Name: Timestamp, dtype: object
```

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

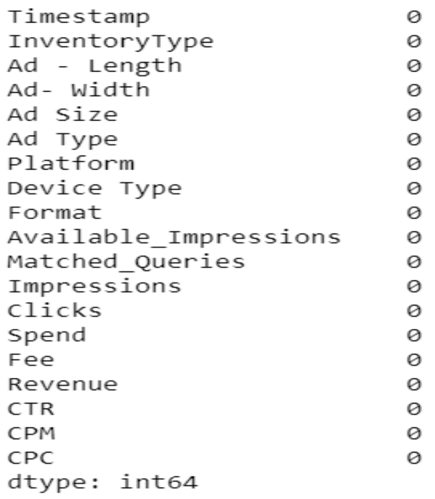
Answer:-

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

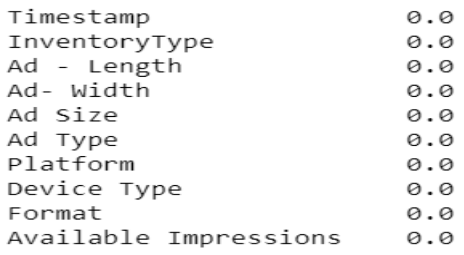
Excluding the nan values, The distribution looks normal for all 3 Features.
#To keep the data symmetric we will impute the null values with median



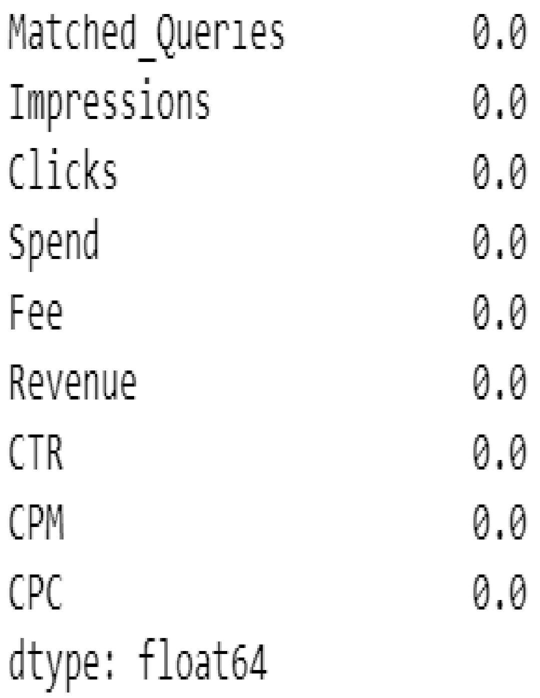
As the computation method of all 3 parameters are given, we will use the same to fill the null value



```
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
dtype: int64
```



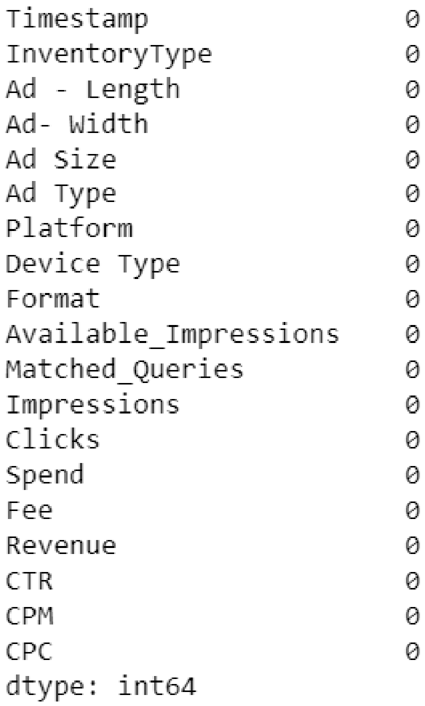
```
Timestamp                  0.0
InventoryType              0.0
Ad - Length                0.0
Ad- Width                  0.0
Ad Size                    0.0
Ad Type                    0.0
Platform                   0.0
Device Type                0.0
Format                     0.0
Available Impressions      0.0
```



```
Matched_Queries      0.0
Impressions          0.0
Clicks               0.0
Spend                0.0
Fee                  0.0
Revenue              0.0
CTR                  0.0
CPM                  0.0
CPC                  0.0
dtype: float64
```

After imputation the missing values are reduced to - CTR(0.8% nan/219), CPM(.8% nan/219) and CTC(10% nan/2586)

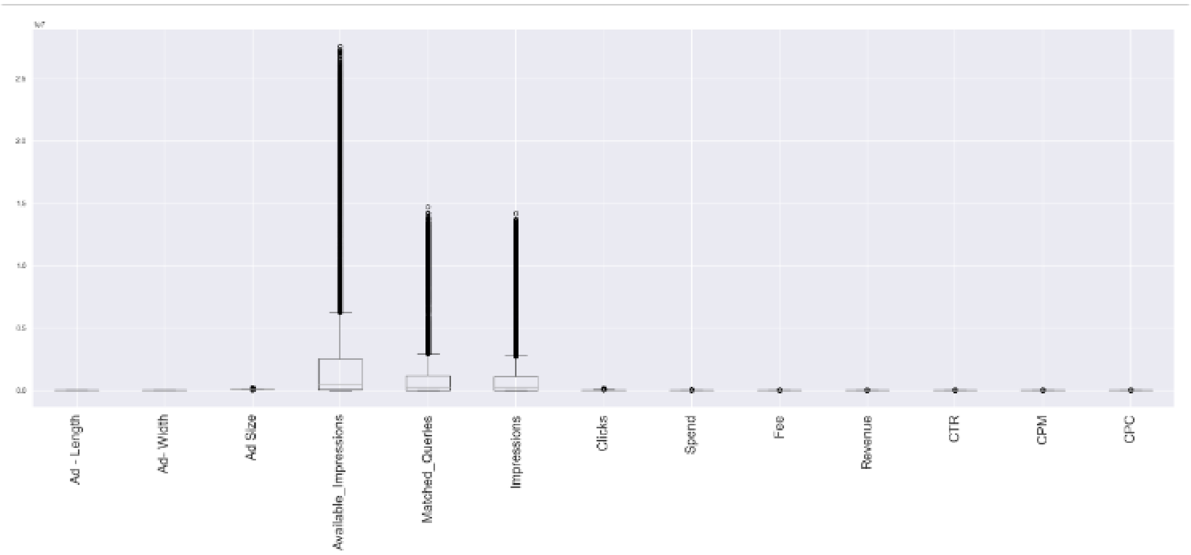
The remaining null values are present due to null value in the parameters (impressions, clicks and sales). We will remove these rows from the dataset for further analysis.



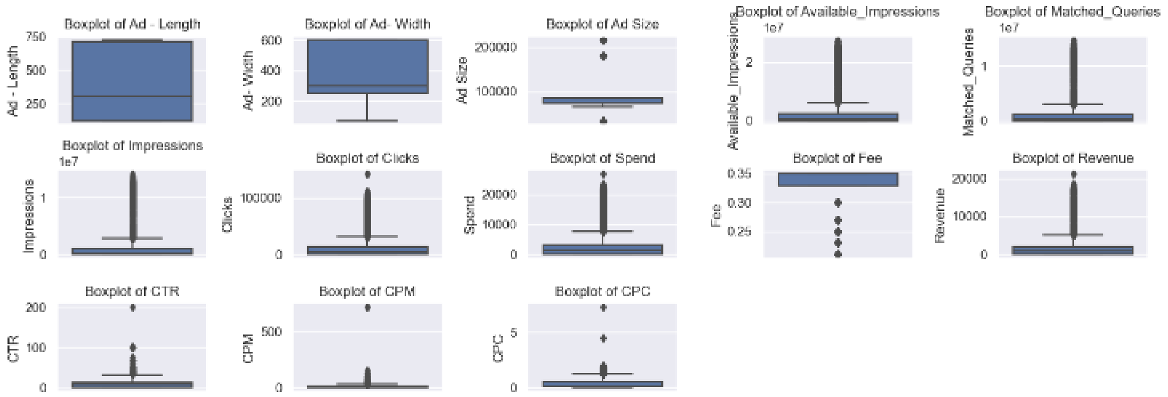
```
Timestamp                0
InventoryType            0
Ad - Length              0
Ad- Width                0
Ad Size                  0
Ad Type                  0
Platform                 0
Device Type              0
Format                   0
Available_Impressions    0
Matched_Queries          0
Impressions              0
Clicks                   0
Spend                    0
Fee                      0
Revenue                  0
CTR                      0
CPM                      0
CPC                      0
dtype: int64
```

1.3 Check and treat if there are any outliers.

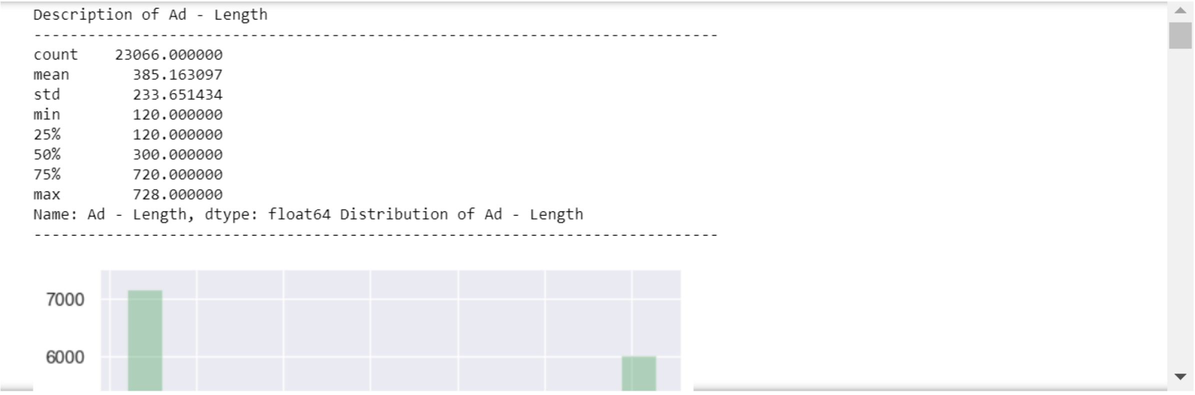
Answer:- Method1-



Method 2-



OBS (outliers) : From the above set of box plots, its evident that Outliers are present in all numeric Features except for Ad-length and Ad-width

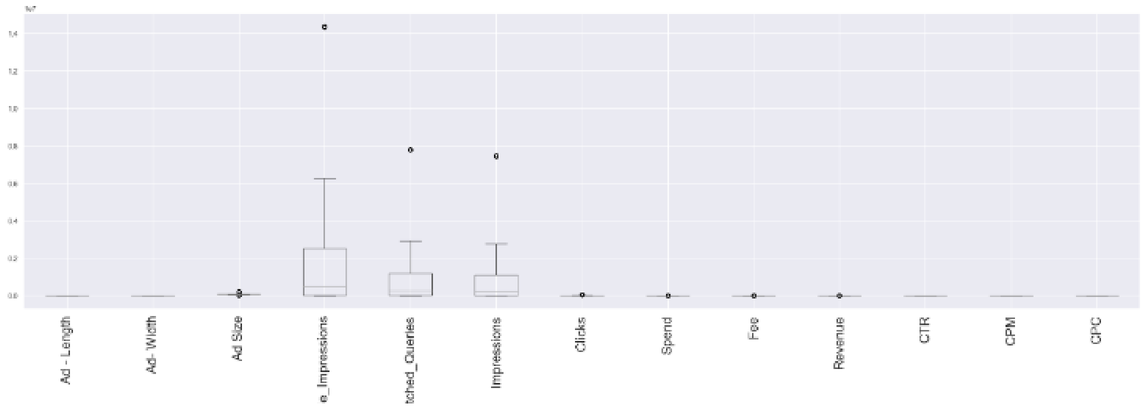


```
Description of Ad - Length
----------------------------------------------------------------------------
count    23066.000000
mean       385.163097
std        233.651434
min        120.000000
25%        120.000000
50%        300.000000
75%        720.000000
max        728.000000
Name: Ad - Length, dtype: float64 Distribution of Ad - Length
----------------------------------------------------------------------------
```

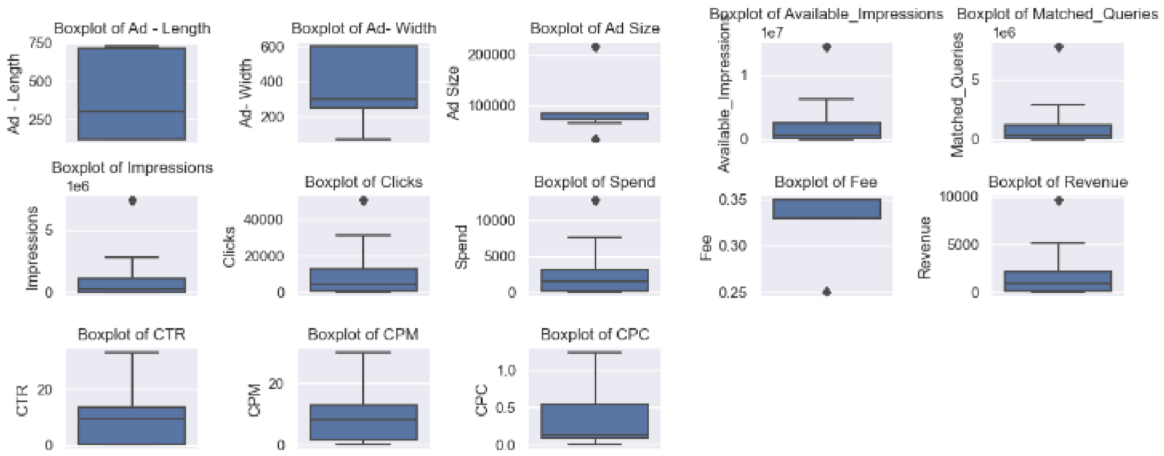
Data doesn’t display completely here, please go through my jupiter notebook file.

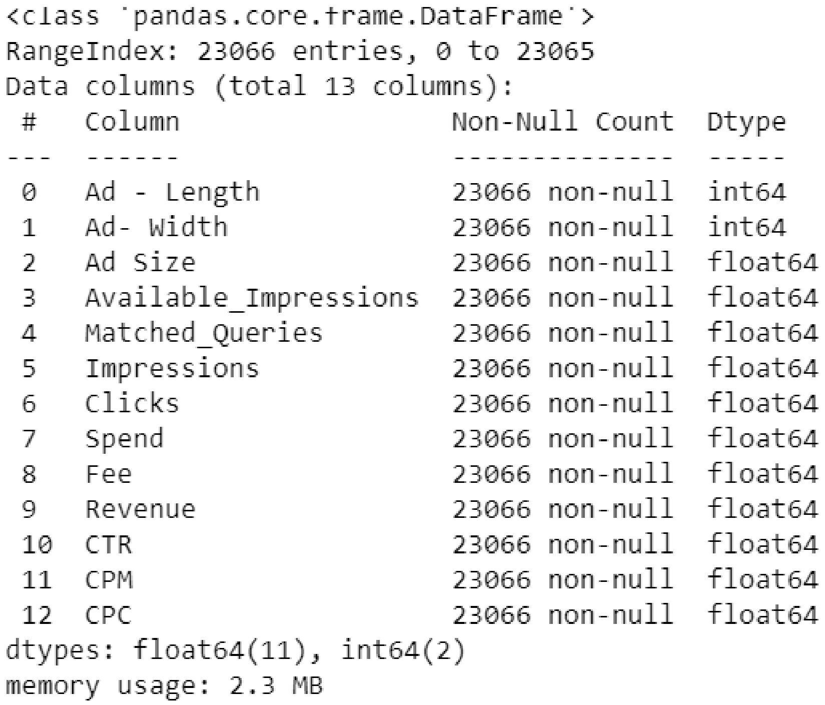
OUTLIER TREATMENT

Method 1-

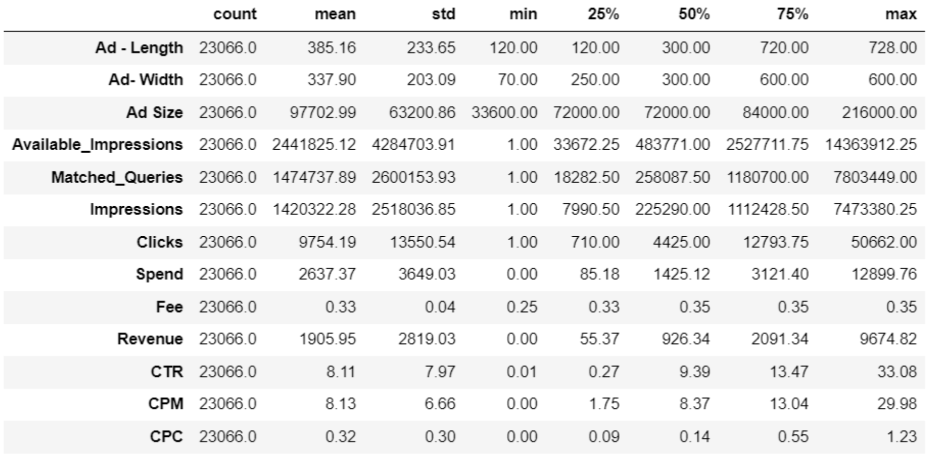


Method 2-





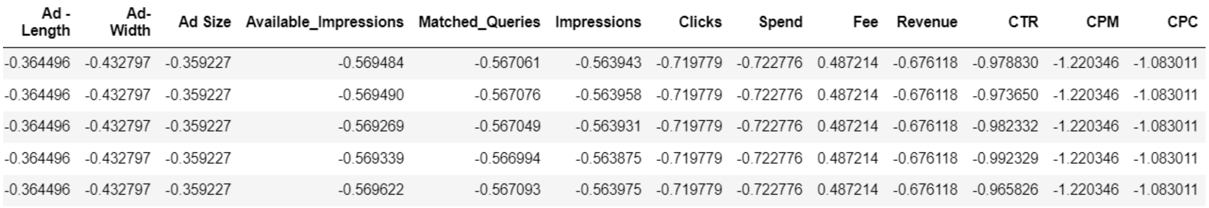
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23066 entries, 0 to 23065
Data columns (total 13 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Ad - Length            23066 non-null  int64
 1   Ad- Width              23066 non-null  int64
 2   Ad Size                23066 non-null  float64
 3   Available_Impressions  23066 non-null  float64
 4   Matched_Queries        23066 non-null  float64
 5   Impressions            23066 non-null  float64
 6   Clicks                 23066 non-null  float64
 7   Spend                  23066 non-null  float64
 8   Fee                    23066 non-null  float64
 9   Revenue                23066 non-null  float64
 10  CTR                    23066 non-null  float64
 11  CPM                    23066 non-null  float64
 12  CPC                    23066 non-null  float64
dtypes: float64(11), int64(2)
memory usage: 2.3 MB
```



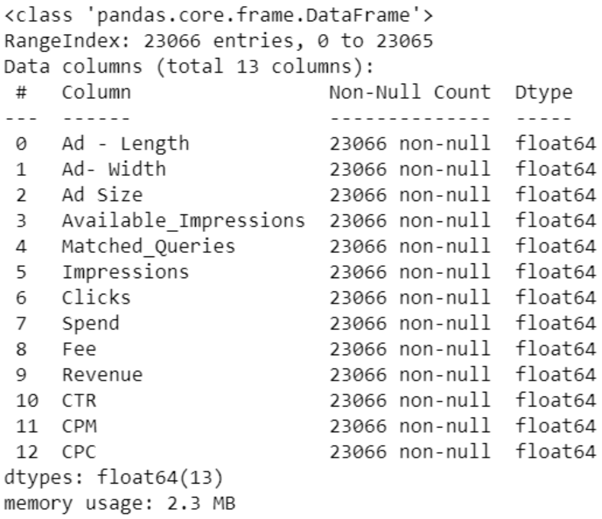
	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	385.16	233.65	120.00	120.00	300.00	720.00	728.00
Ad- Width	23066.0	337.90	203.09	70.00	250.00	300.00	600.00	600.00
Ad Size	23066.0	97702.99	63200.86	33600.00	72000.00	72000.00	84000.00	216000.00
Available_Impressions	23066.0	2441825.12	4284703.91	1.00	33672.25	483771.00	2527711.75	14363912.25
Matched_Queries	23066.0	1474737.89	2600153.93	1.00	18282.50	258087.50	1180700.00	7803449.00
Impressions	23066.0	1420322.28	2518036.85	1.00	7990.50	225290.00	1112428.50	7473380.25
Clicks	23066.0	9754.19	13550.54	1.00	710.00	4425.00	12793.75	50662.00
Spend	23066.0	2637.37	3649.03	0.00	85.18	1425.12	3121.40	12899.76
Fee	23066.0	0.33	0.04	0.25	0.33	0.35	0.35	0.35
Revenue	23066.0	1905.95	2819.03	0.00	55.37	926.34	2091.34	9674.82
CTR	23066.0	8.11	7.97	0.01	0.27	9.39	13.47	33.08
CPM	23066.0	8.13	6.66	0.00	1.75	8.37	13.04	29.98
CPC	23066.0	0.32	0.30	0.00	0.09	0.14	0.55	1.23

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

Answer:-



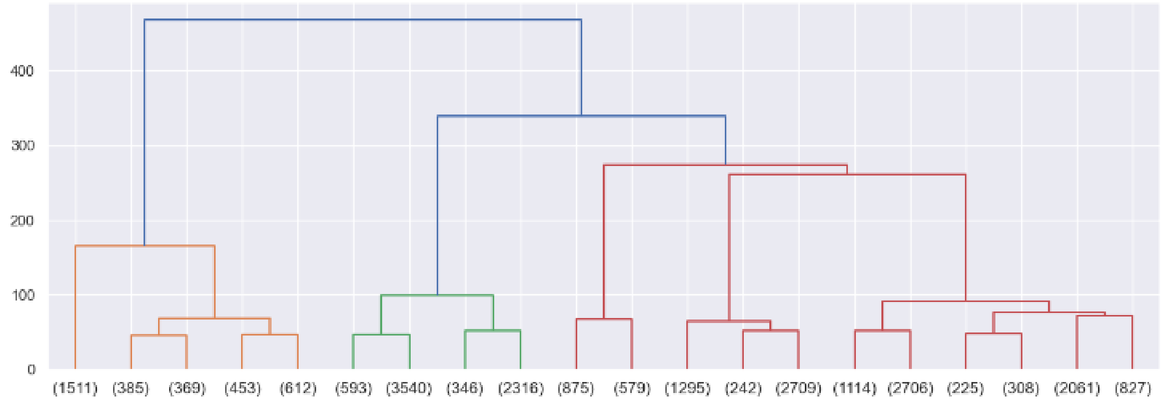
Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
-0.364496	-0.432797	-0.359227	-0.569484	-0.567061	-0.563943	-0.719779	-0.722776	0.487214	-0.676118	-0.978830	-1.220346	-1.083011
-0.364496	-0.432797	-0.359227	-0.569490	-0.567076	-0.563958	-0.719779	-0.722776	0.487214	-0.676118	-0.973650	-1.220346	-1.083011
-0.364496	-0.432797	-0.359227	-0.569269	-0.567049	-0.563931	-0.719779	-0.722776	0.487214	-0.676118	-0.982332	-1.220346	-1.083011
-0.364496	-0.432797	-0.359227	-0.569339	-0.566994	-0.563875	-0.719779	-0.722776	0.487214	-0.676118	-0.992329	-1.220346	-1.083011
-0.364496	-0.432797	-0.359227	-0.569622	-0.567093	-0.563975	-0.719779	-0.722776	0.487214	-0.676118	-0.965826	-1.220346	-1.083011



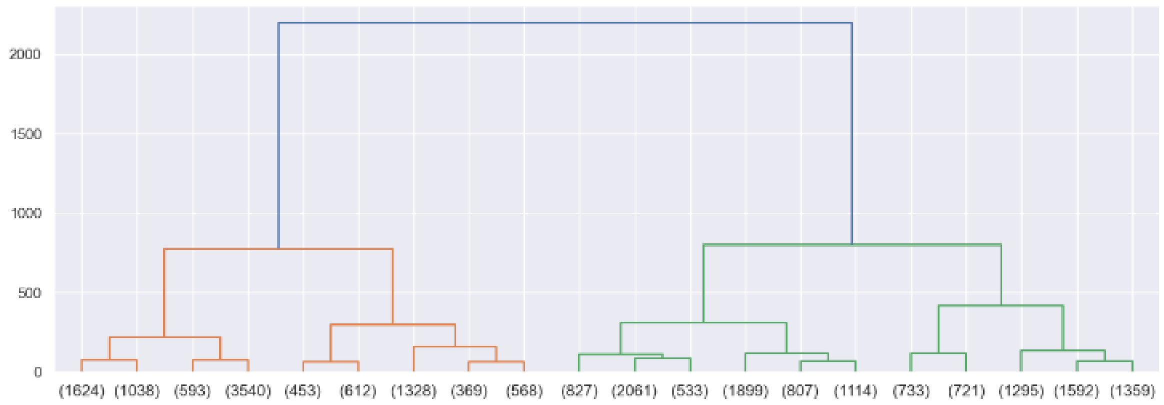
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23066 entries, 0 to 23065
Data columns (total 13 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   Ad - Length            23066 non-null  float64
 1   Ad- Width              23066 non-null  float64
 2   Ad Size                23066 non-null  float64
 3   Available_Impressions  23066 non-null  float64
 4   Matched_Queries        23066 non-null  float64
 5   Impressions            23066 non-null  float64
 6   Clicks                 23066 non-null  float64
 7   Spend                  23066 non-null  float64
 8   Fee                    23066 non-null  float64
 9   Revenue                23066 non-null  float64
 10  CTR                    23066 non-null  float64
 11  CPM                    23066 non-null  float64
 12  CPC                    23066 non-null  float64
dtypes: float64(13)
memory usage: 2.3 MB
```

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

Answer:-

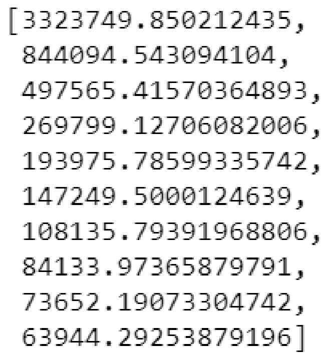


DENDOGRAM USING EUCLIDEAN DISTANCES

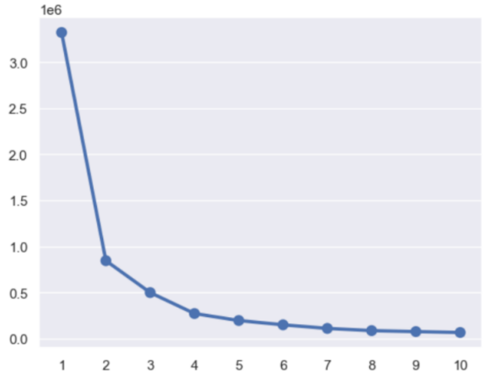


1.6 Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.

Answer:- k-mean inertia= 63944.29253879197



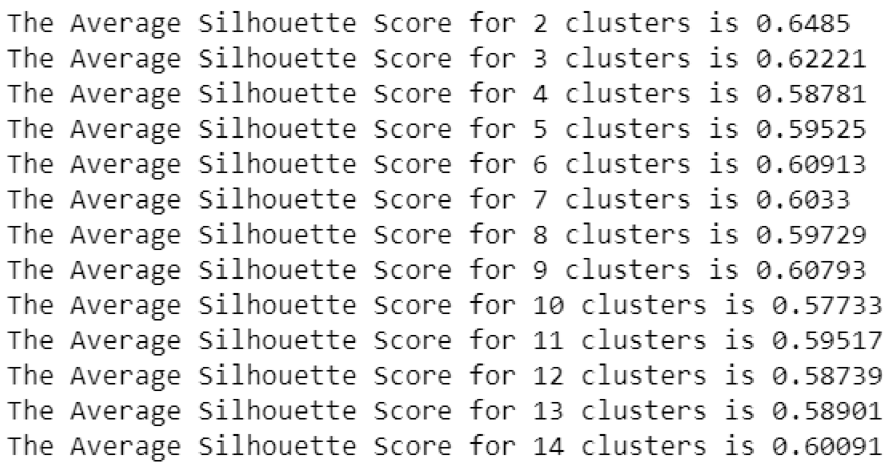
```
[3323749.850212435,
 844094.543094104,
 497565.41570364893,
 269799.12706082006,
 193975.78599335742,
 147249.5000124639,
 108135.79391968806,
 84133.97365879791,
 73652.19073304742,
 63944.29253879196]
```



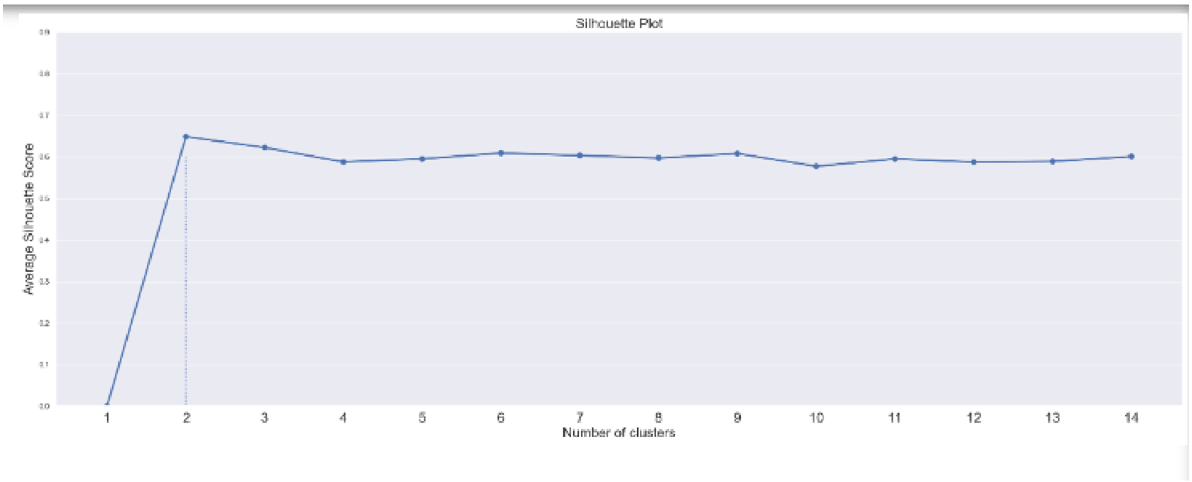
When we move from k=1 to k=2 , we see that there is a significant drop in the value , also when we move from k=2 to k=3,k=3 to k=4 there is a significant drop as well.But from k=4 to k=5 , k=5 to k=6 , the drop in values reduces significantly.

1.7 Print silhouette scores and identify optimum number of clusters.

Answer:-



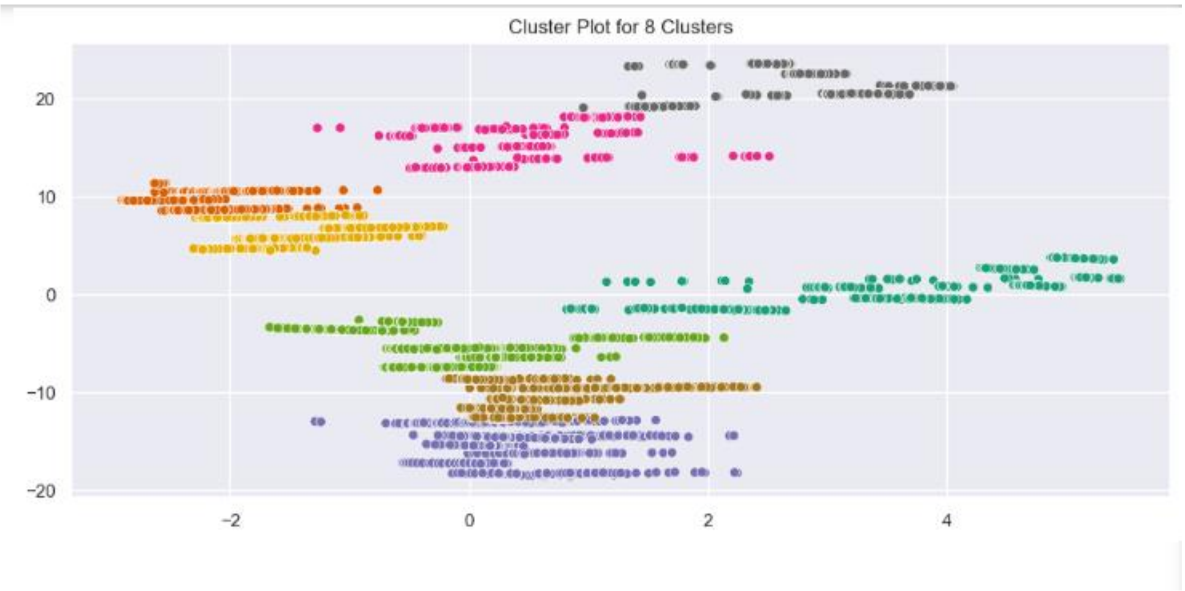
```
The Average Silhouette Score for 2 clusters is 0.6485
The Average Silhouette Score for 3 clusters is 0.62221
The Average Silhouette Score for 4 clusters is 0.58781
The Average Silhouette Score for 5 clusters is 0.59525
The Average Silhouette Score for 6 clusters is 0.60913
The Average Silhouette Score for 7 clusters is 0.6033
The Average Silhouette Score for 8 clusters is 0.59729
The Average Silhouette Score for 9 clusters is 0.60793
The Average Silhouette Score for 10 clusters is 0.57733
The Average Silhouette Score for 11 clusters is 0.59517
The Average Silhouette Score for 12 clusters is 0.58739
The Average Silhouette Score for 13 clusters is 0.58901
The Average Silhouette Score for 14 clusters is 0.60091
```

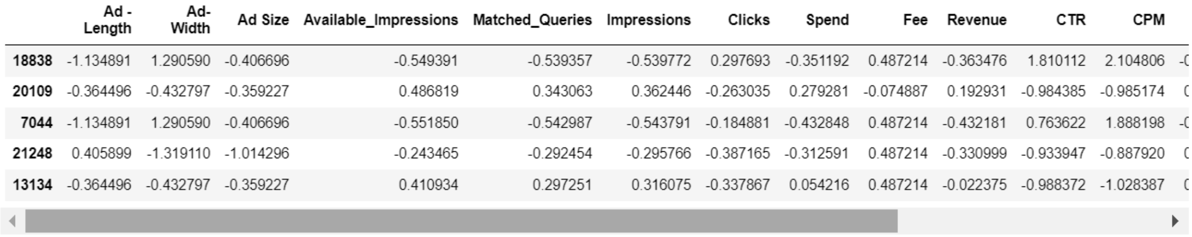


It is clear from above plot that the maximum value of average silhouette score is achieved for k = 8, which, therefore, is considered to be the optimum number of clusters for this data.

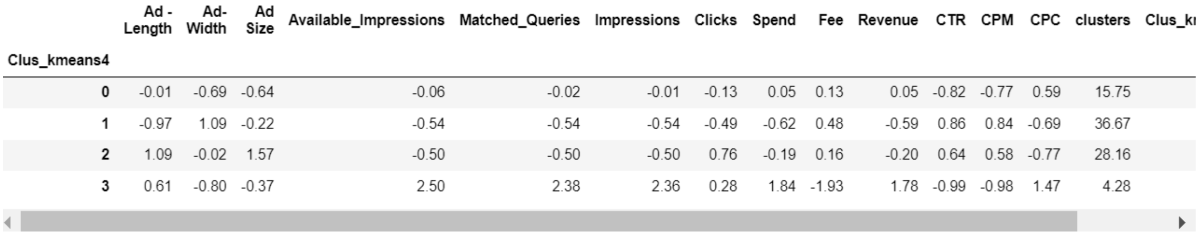
1.8 Profile the ads based on optimum number of clusters using silhouee score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots]

Answer:-





	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM
18838	-1.134891	1.290590	-0.406696	-0.549391	-0.539357	-0.539772	0.297693	-0.351192	0.487214	-0.363476	1.810112	2.104806
20109	-0.364496	-0.432797	-0.359227	0.486819	0.343063	0.362446	-0.263035	0.279281	-0.074887	0.192931	-0.984385	-0.985174
7044	-1.134891	1.290590	-0.406696	-0.551850	-0.542987	-0.543791	-0.184881	-0.432848	0.487214	-0.432181	0.763622	1.888198
21248	0.405899	-1.319110	-1.014296	-0.243465	-0.292454	-0.295766	-0.387165	-0.312591	0.487214	-0.330999	-0.933947	-0.887920
13134	-0.364496	-0.432797	-0.359227	0.410934	0.297251	0.316075	-0.337867	0.054216	0.487214	-0.022375	-0.988372	-1.028387



Clus_kmeans4	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	clusters	Clus_kı
0	-0.01	-0.69	-0.64	-0.06	-0.02	-0.01	-0.13	0.05	0.13	0.05	-0.82	-0.77	0.59	15.75	
1	-0.97	1.09	-0.22	-0.54	-0.54	-0.54	-0.49	-0.62	0.48	-0.59	0.86	0.84	-0.69	36.67	
2	1.09	-0.02	1.57	-0.50	-0.50	-0.50	0.76	-0.19	0.16	-0.20	0.64	0.58	-0.77	28.16	
3	0.61	-0.80	-0.37	2.50	2.38	2.36	0.28	1.84	-1.93	1.78	-0.99	-0.98	1.47	4.28	

1.9 Conclude the project by providing summary of your learning

Answer:-

- The dataset has 25857 rows and 19 columns.
- The missing values in CPC, CTR and CPM are treated by using the formulae given and writing a user-defined function, and calling it.
- We check for outliers, we can see there are outliers in the variables.
- Dendogram is the visualization and linkage is for computing the distances and merging the clusters from n to 1.
- The output of Linkage is visualized by Dendogram.
- We will create linkage using Ward's method and run linkage function on the usable columns of the data.
- The linkage now stores the various distance at which the n clusters are sequentially merged into a single cluster.
- Using FIt – transform function and viewing the output -
The dataframe is now stored in an array.
- Using this array we can now perform k-means
- The one requirement before we run the k-means algorithm, is to know how many clusters we require as output
- From the plot we have following observations:
- When we move from k=1 to k=2 , we see that there is a significant drop in the value ,also when we move from k=2 to k=3,k=3 to k=4 there is a significant drop as well.
- But from k=4 to k=5 , k=5 to k=6 , the drop in values reduces significantly
- So 4 is optimal number of clusters.

PART B:- PCA

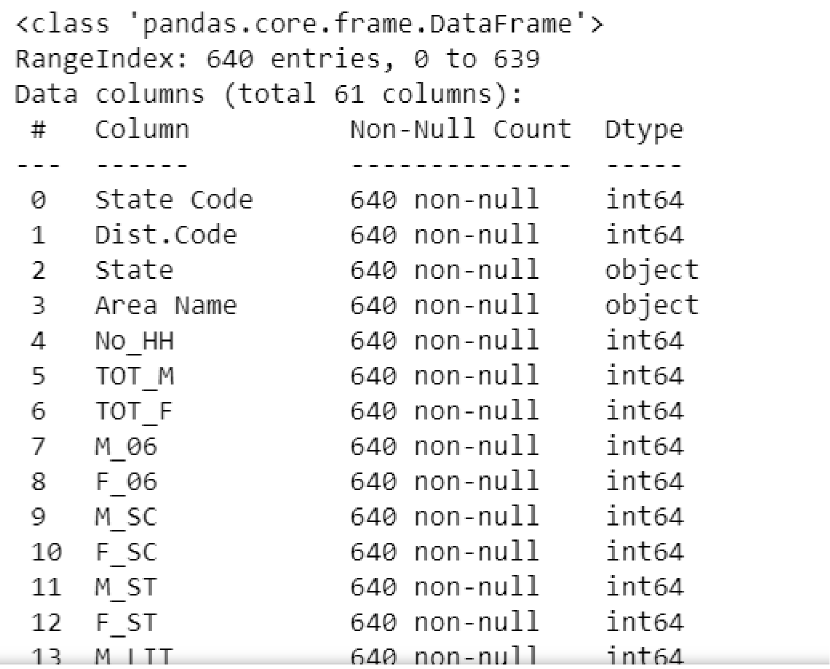
PCA FH (FT): Primary census abstract for female headed households excluding institutional households (India & States/UTs - District Level), Scheduled tribes - 2011 PCA for Female Headed Household Excluding Institutional Household. The Indian Census has the reputation of being one of the best in the world. The first Census in India was conducted in the year 1872. This was conducted at different points of time in different parts of the country. In 1881 a Census was taken for the entire country simultaneously. Since then, Census has been conducted every ten years, without a break. Thus, the Census of India 2011 was the fifteenth in this unbroken series since 1872, the seventh after independence and the second census of the third millennium and twenty first century. The census has been uninterruptedly continued despite of several adversities like wars, epidemics, natural calamities, political unrest, etc. The Census of India is conducted under the provisions of the Census Act 1948 and the Census Rules, 1990. The Primary Census Abstract which is important publication of 2011 Census gives basic information on Area, Total Number of Households, Total Population, Scheduled Castes, Scheduled Tribes Population, Population in the age group 0-6, Literates, Main Workers and Marginal Workers classified by the four broad industrial categories, namely, (i) Cultivators, (ii) Agricultural Laborers, (iii) Household Industry Workers, and (iv) Other Workers and also Non-Workers. The characteristics of the Total Population include Scheduled Castes, Scheduled Tribes, Institutional and House-less Population and are presented by sex and rural-urban residence. Census 2011 covered 35 States/Union Territories, 640 districts, 5,924 sub-districts, 7,935 Towns and 6,40,867 Villages.

2.1 Read the data and perform basic checks shape, data types, statistical summary.

Answer:-

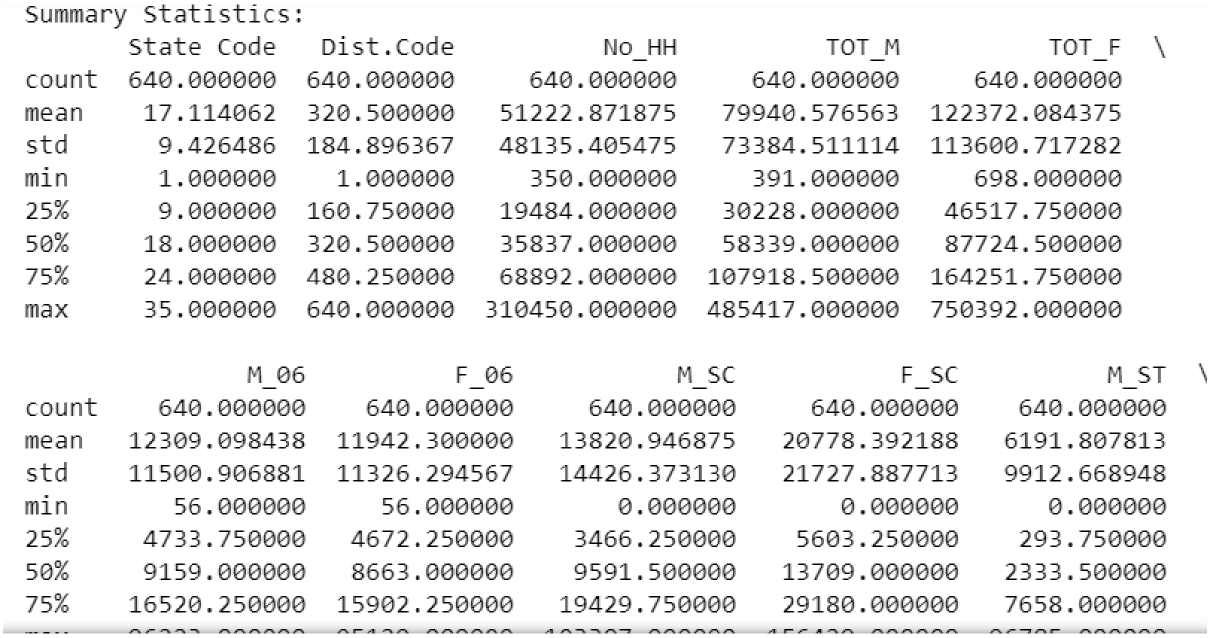
Shape of the dataset:- (5*61)

Data type:-



```
<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
```

Statistical Summary:-



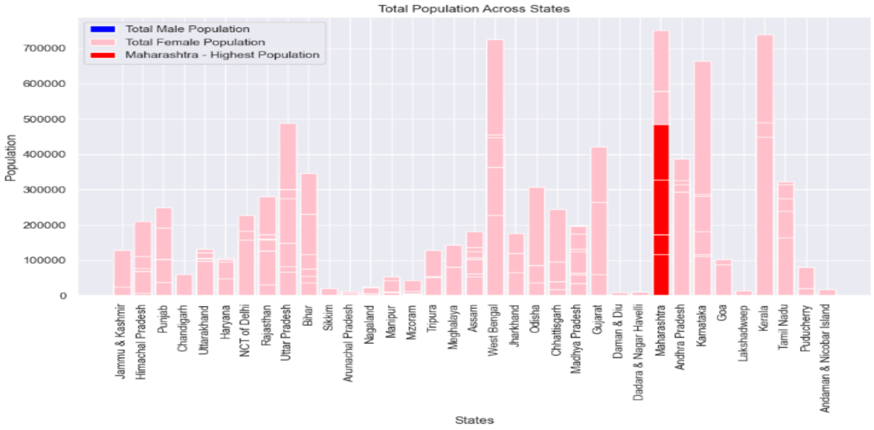
```
Summary Statistics:
        State Code   Dist.Code          No_HH          TOT_M          TOT_F  \
count   640.000000  640.000000     640.000000     640.000000     640.000000
mean     17.114062  320.500000   51222.871875   79940.576563  122372.084375
std       9.426486  184.896367   48135.405475   73384.511114  113600.717282
min       1.000000    1.000000     350.000000     391.000000     698.000000
25%       9.000000  160.750000   19484.000000   30228.000000   46517.750000
50%      18.000000  320.500000   35837.000000   58339.000000   87724.500000
75%      24.000000  480.250000   68892.000000  107918.500000  164251.750000
max      35.000000  640.000000  310450.000000  485417.000000  750392.000000

                M_06           F_06           M_SC           F_SC          M_ST  \
count     640.000000     640.000000     640.000000     640.000000    640.000000
mean    12309.098438   11942.300000   13820.946875   20778.392188   6191.807813
std     11500.906881   11326.294567   14426.373130   21727.887713   9912.668948
min        56.000000      56.000000       0.000000       0.000000      0.000000
25%      4733.750000    4672.250000    3466.250000    5603.250000    293.750000
50%      9159.000000    8663.000000    9591.500000   13709.000000   2333.500000
75%     16520.250000   15902.250000   19429.750000   29180.000000   7658.000000
```

All information is incomplete please go through my ipynb file.

2.2 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

(i) Which state has the highest & lowest population?

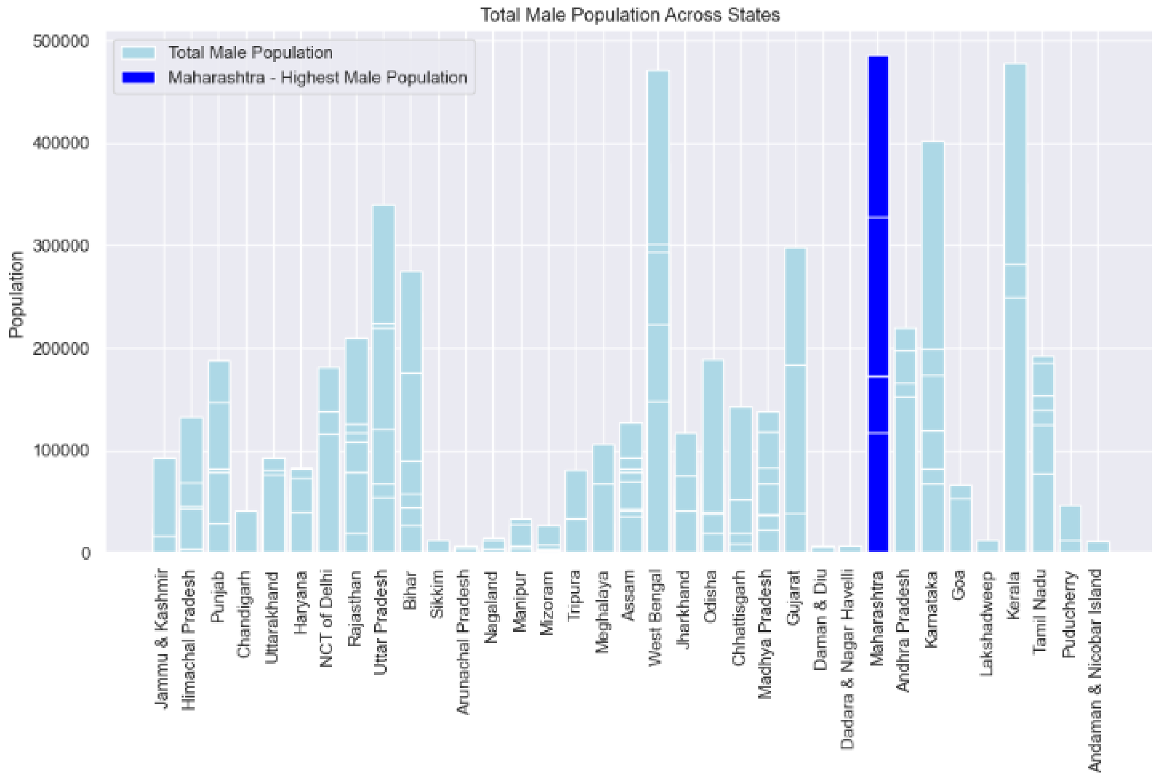
Answer:-



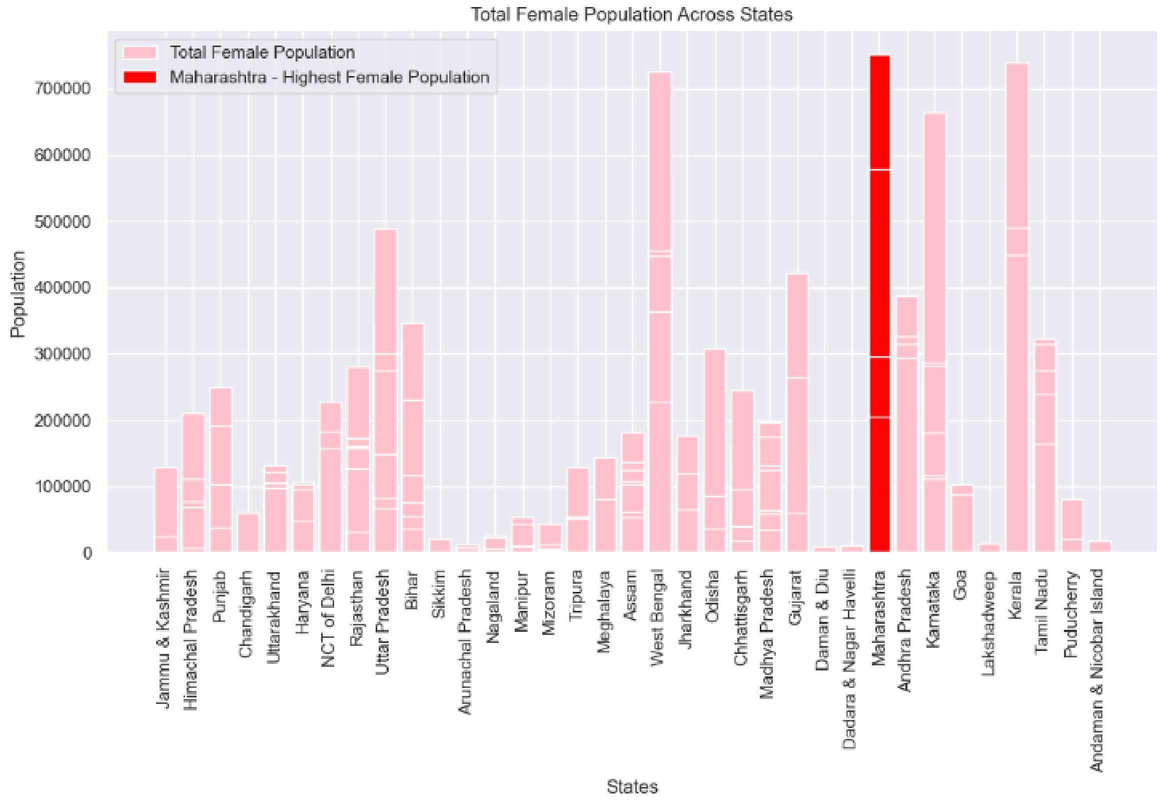
Maharashtra has highest population. & Daman &Diu has lowest.

(ii) Which state has the highest & lowest gender ratio?

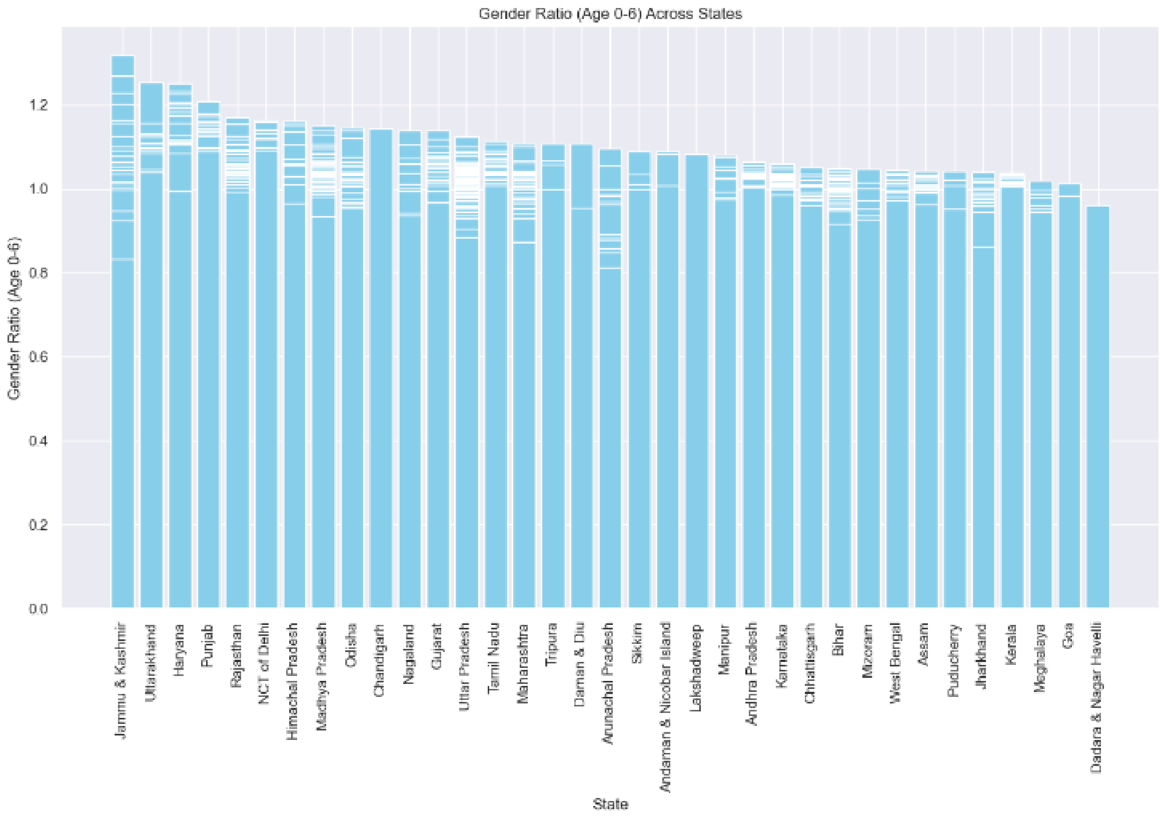
Answer:-



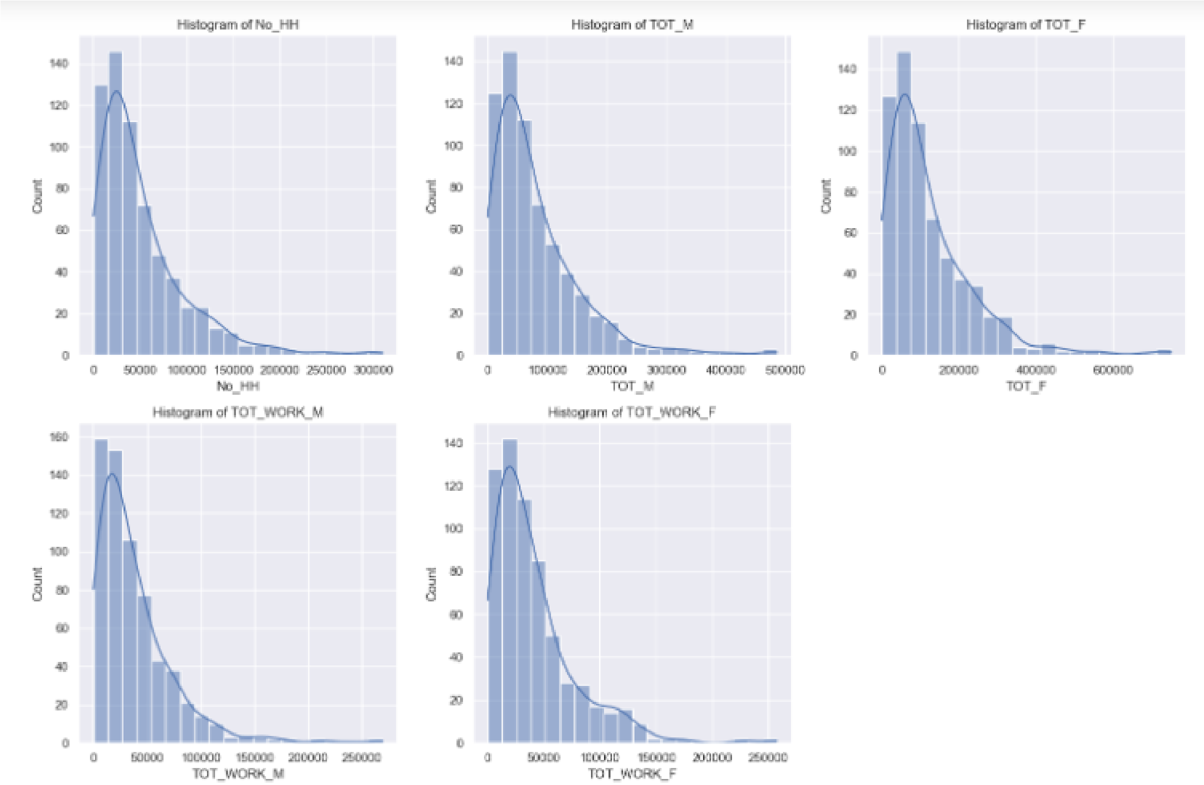
Total Male Population Across States

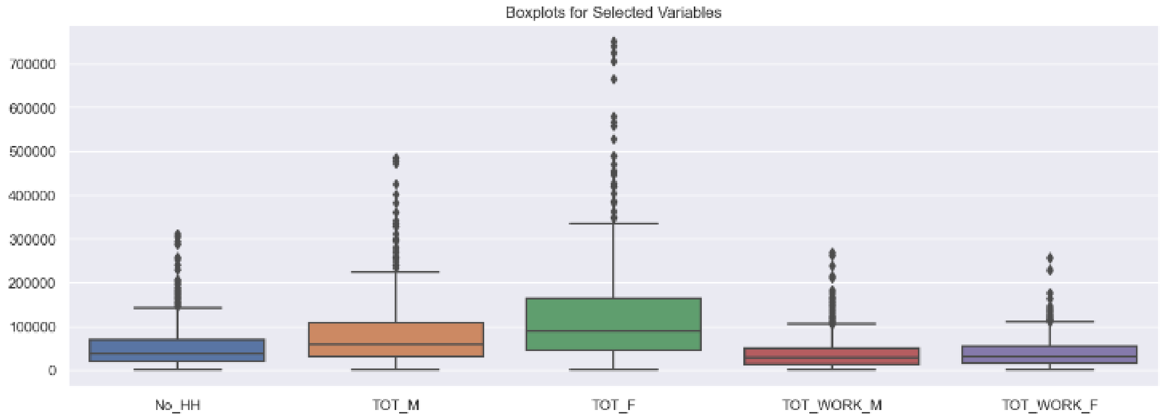


Total Female Population Across States

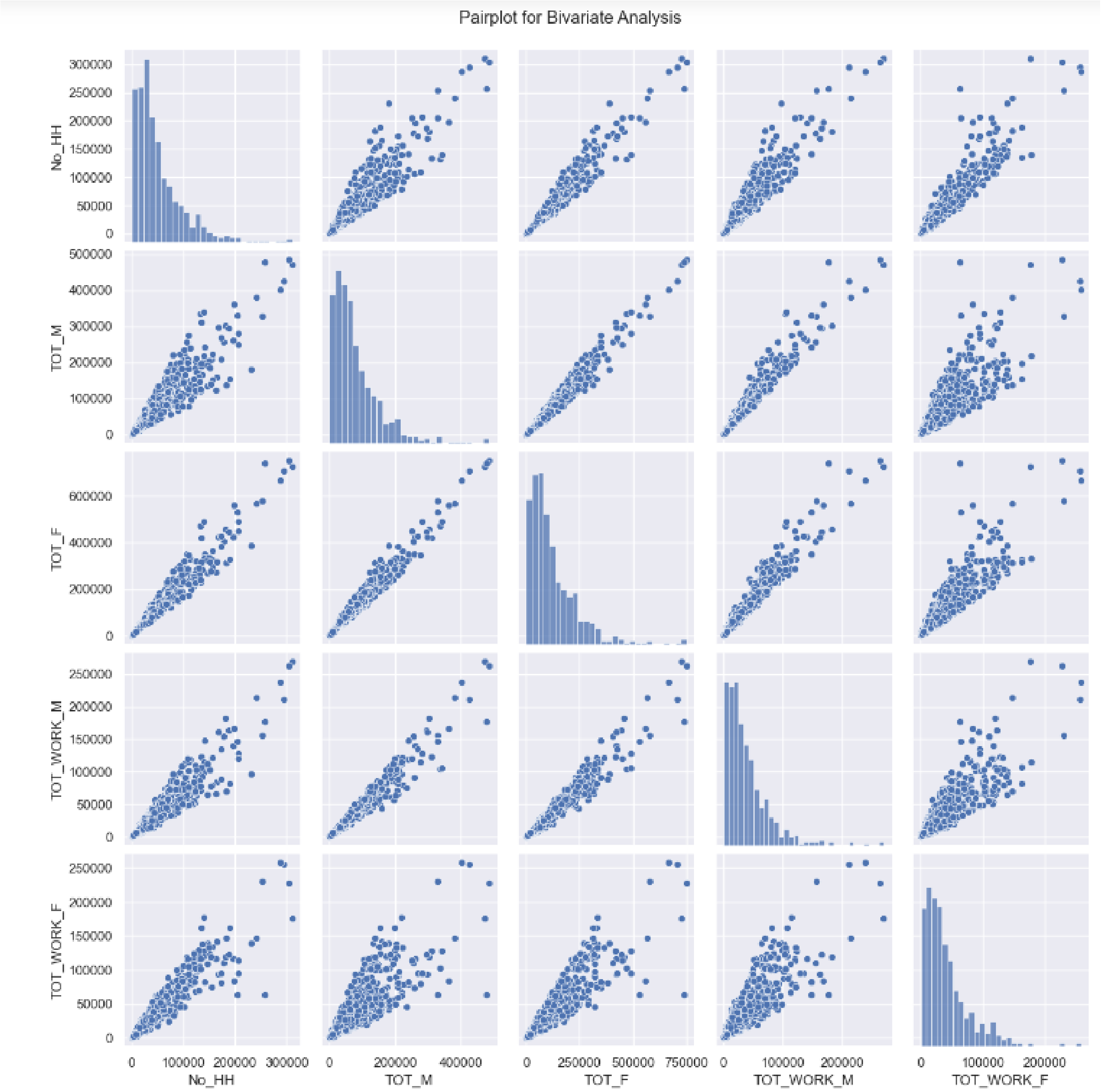


For EDA - Variables considered:No_HH TOT_M TOT_F TOT_WORK_M TOT_WORK_FNo of HouseholdTotal population MaleTotal population FemaleTotal Worker Population MaleTotal Worker Population FemaleUnivariate Analysis:Plotting histogram and boxplots for the above variables:-

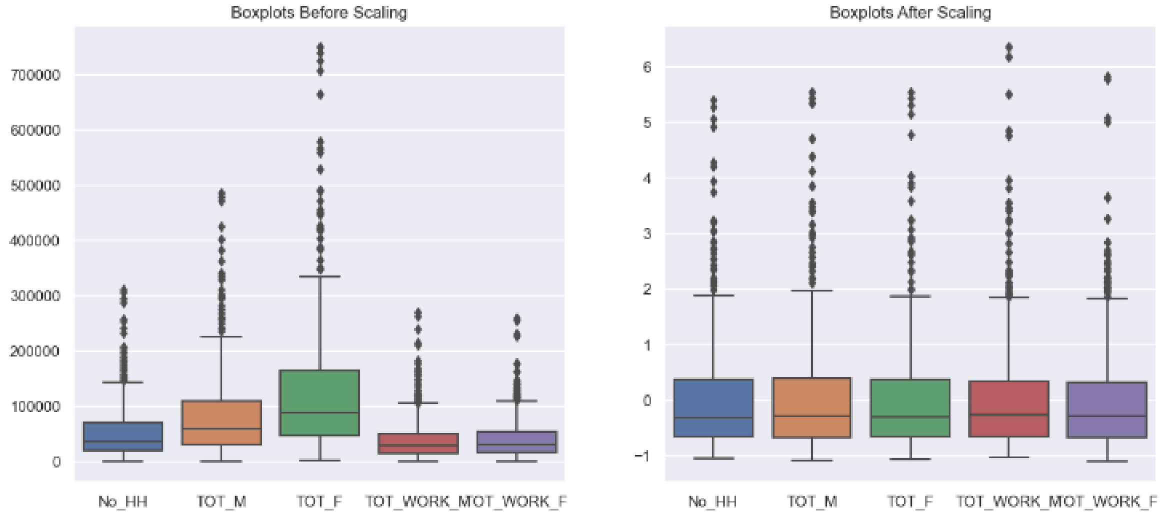




for bivariate analysis:-

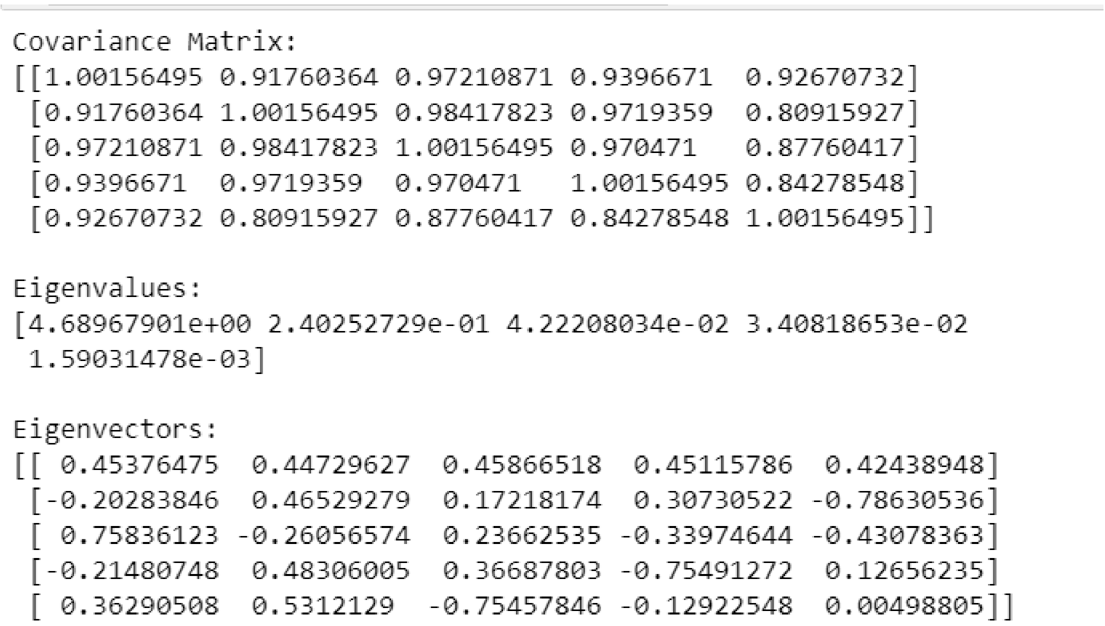


2.3 &2.4 :- We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary? Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.



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

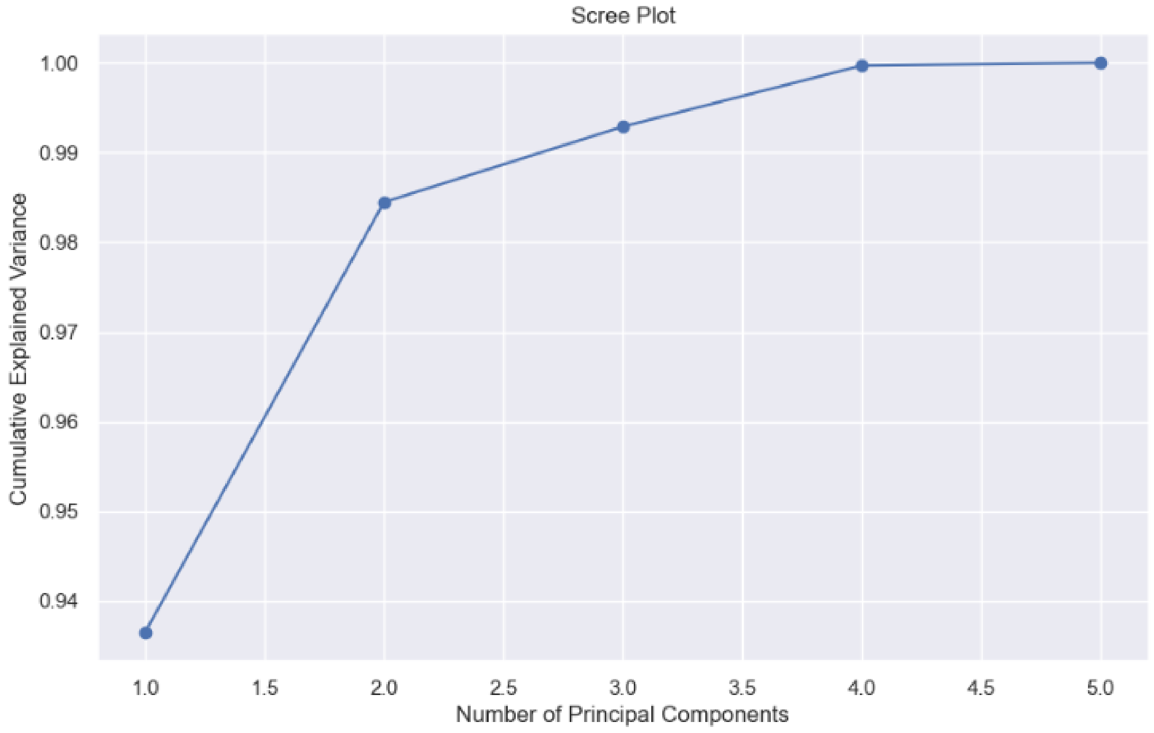
Answer:-



```
Covariance Matrix:
[[1.00156495 0.91760364 0.97210871 0.9396671  0.92670732]
 [0.91760364 1.00156495 0.98417823 0.9719359  0.80915927]
 [0.97210871 0.98417823 1.00156495 0.970471   0.87760417]
 [0.9396671  0.9719359  0.970471   1.00156495 0.84278548]
 [0.92670732 0.80915927 0.87760417 0.84278548 1.00156495]]

Eigenvalues:
[4.68967901e+00 2.40252729e-01 4.22208034e-02 3.40818653e-02
 1.59031478e-03]

Eigenvectors:
[[ 0.45376475  0.44729627  0.45866518  0.45115786  0.42438948]
 [-0.20283846  0.46529279  0.17218174  0.30730522 -0.78630536]
 [ 0.75836123 -0.26056574  0.23662535 -0.33974644 -0.43078363]
 [-0.21480748  0.48306005  0.36687803 -0.75491272  0.12656235]
 [ 0.36290508  0.5312129  -0.75457846 -0.12922548  0.00498805]]
```

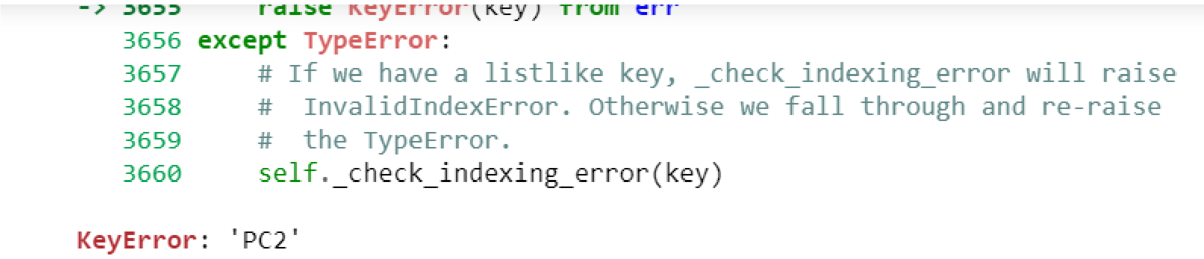
2.6:- Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.

Answer:-

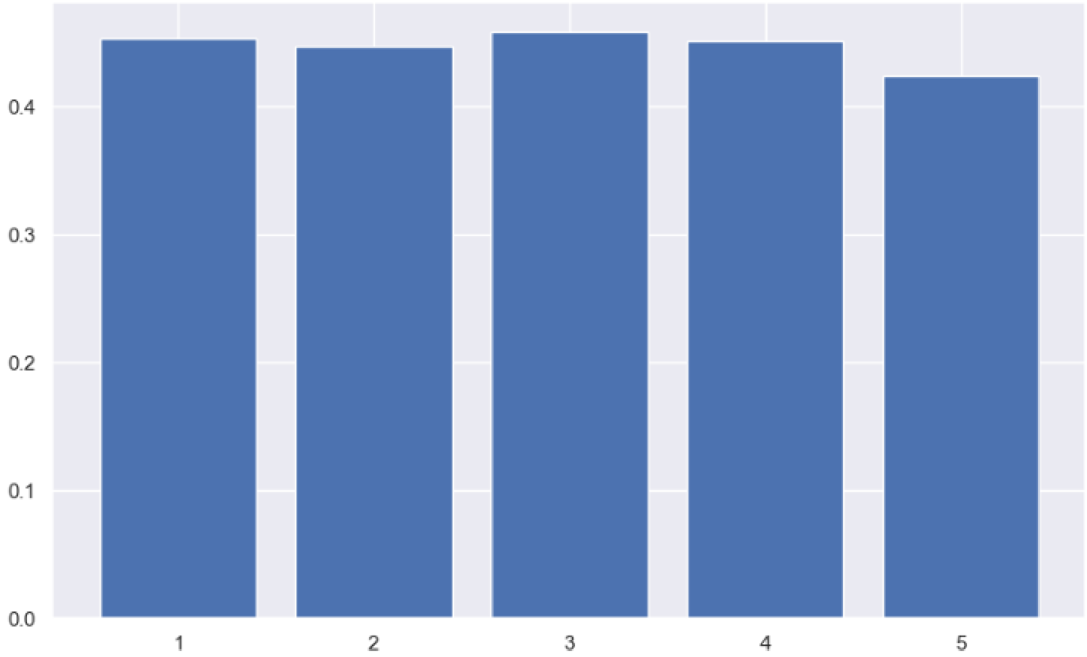


2.7:- Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.

Answer:-



```
-> 3655         raise KeyError(key) from err
   3656     except TypeError:
   3657         # If we have a listlike key, _check_indexing_error will raise
   3658         #  InvalidIndexError. Otherwise we fall through and re-raise
   3659         #  the TypeError.
   3660         self._check_indexing_error(key)

KeyError: 'PC2'
```



2.8:- Write linear equation for first PC.

Answer:-

PC 1 = a1x1 + a2x2 + a3X3 +a4X4 + …….+ a57x5724