Data Mining Project:

-Prapthi Pandian

Table of Contents

2.4 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.

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

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.

2.8 Write linear equation for first PC.

Problem 1: 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.

Printing 1st 5 rows:

	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	0.0	0.35	0.0	0.0031	0.0	0.0
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.0	0.35	0.0	0.0035	0.0	0.0
2	2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0	0.35	0.0	0.0028	0.0	0.0
3	2020-9-3-20	Format1	300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.0	0.35	0.0	0.0020	0.0	0.0
4	2020-9-4-15	Format1	300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.0	0.35	0.0	0.0041	0.0	0.0

Printing last 5 rows:

	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

Checking data type of columns:

```
<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
```

Shape:

```
(23066, 19)
```

There are 23066 observations and 19 columns in the data

6 variables of Object data type and 13 variables of numeric data type (int and float)

Basic description of data –

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
count	23066.000000	23066.000000	23066.000000	2.306600e+04	2.306600e+04	2.306600e+04	23066.000000	23066.000000	23066.000000	23066.000000	18330.000000	18330.000000	18330.000000
mean	385.163097	337.896037	96674.468048	2.432044e+06	1.295099e+06	1.241520e+06	10678.518816	2706.625689	0.335123	1924.252331	0.073661	7.672045	0.351061
std	233.651434	203.092885	61538.329557	4.742888e+06	2.512970e+06	2.429400e+06	17353.409363	4067.927273	0.031963	3105.238410	0.075160	6.481391	0.343334
min	120.000000	70.000000	33600.000000	1.000000e+00	1.000000e+00	1.000000e+00	1.000000	0.000000	0.210000	0.000000	0.000100	0.000000	0.000000
25%	120.000000	250.000000	72000.000000	3.367225e+04	1.828250e+04	7.990500e+03	710.000000	85.180000	0.330000	55.365375	0.002600	1.710000	0.090000
50%	300.000000	300.000000	72000.000000	4.837710e+05	2.580875e+05	2.252900e+05	4425.000000	1425.125000	0.350000	926.335000	0.082550	7.660000	0.160000
75%	720.000000	600.000000	84000.000000	2.527712e+06	1.180700e+06	1.112428e+06	12793.750000	3121.400000	0.350000	2091.338150	0.130000	12.510000	0.570000
max	728.000000	600.000000	216000.000000	2.759286e+07	1.470202e+07	1.419477e+07	143049.000000	26931.870000	0.350000	21276.180000	1.000000	81.560000	7.260000

We have also observed that there are no duplicate values in the dataframe.

Checking for null values-

```
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
```

We can observe that the fields CTR, CPM and CPC contain null values i.e. 4736 entries each which we will treat in the next step.

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

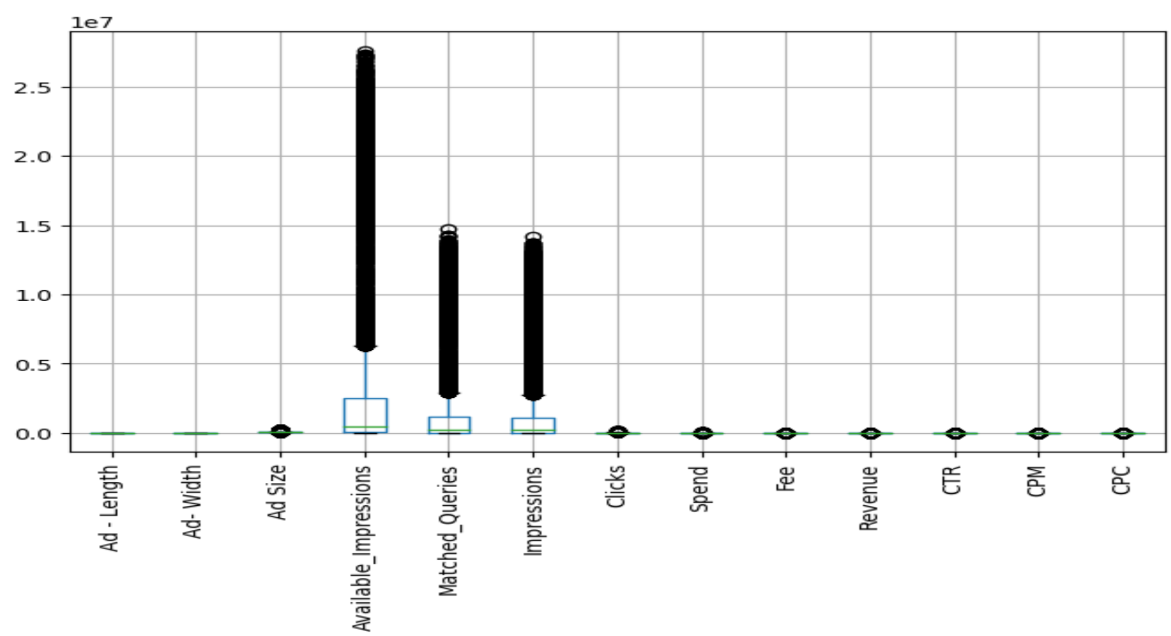
Treating null values using the given formula-

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.

```
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
```

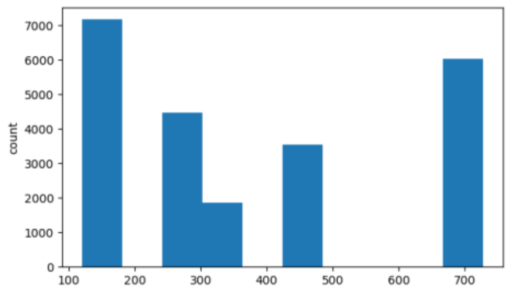
Post treating the null values, the result is as above. We can now observe that there are no null values in the data.

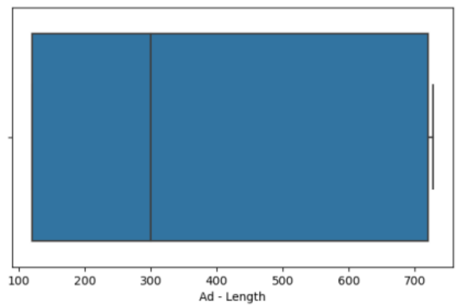
1.3 Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ.



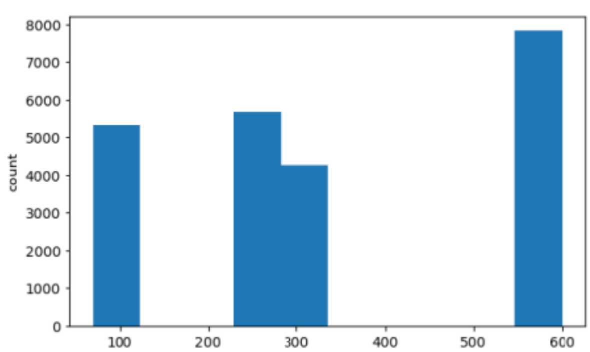
Univariate analysis-

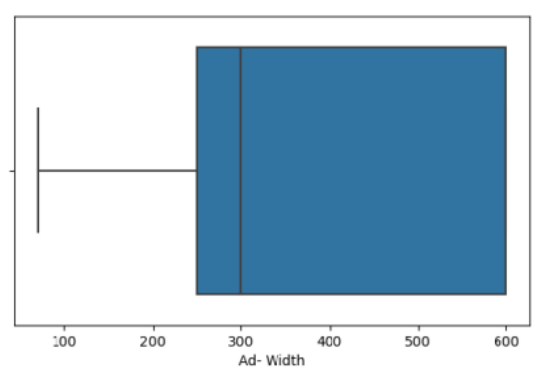
```
Description of Ad - Length
---------------------------------------------------------------------------
count    23066.000000
mean       385.163097
std        233.651434
min        120.000000
25%        120.000000
50%        300.000000
75%        720.000000
max        728.000000
Name: Ad - Length,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

Skew : 0.33

Distribution of Ad - Length
---------------------------------------------------------------------------
```



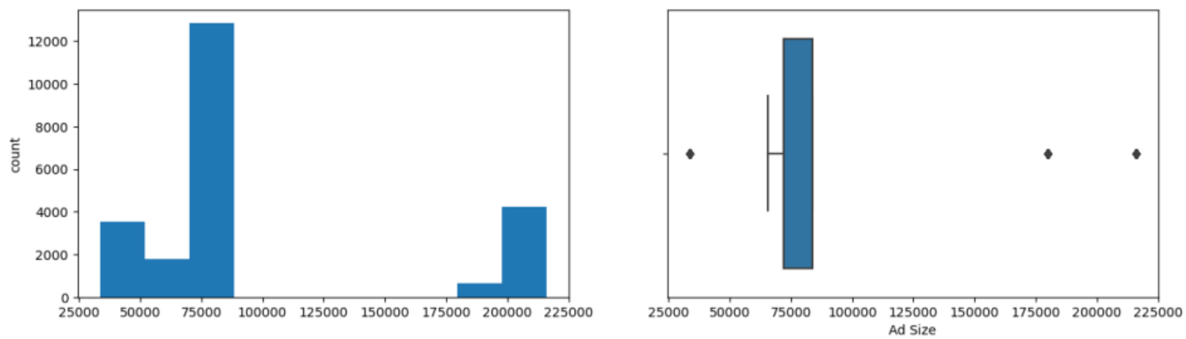



```
Description of Ad- Width
---------------------------------------------------------------------------
count    23066.000000
mean       337.896037
std        203.092885
min         70.000000
25%        250.000000
50%        300.000000
75%        600.000000
max        600.000000
Name: Ad- Width, dtype: float64

Skew : 0.21

Distribution of Ad- Width
---------------------------------------------------------------------------
```

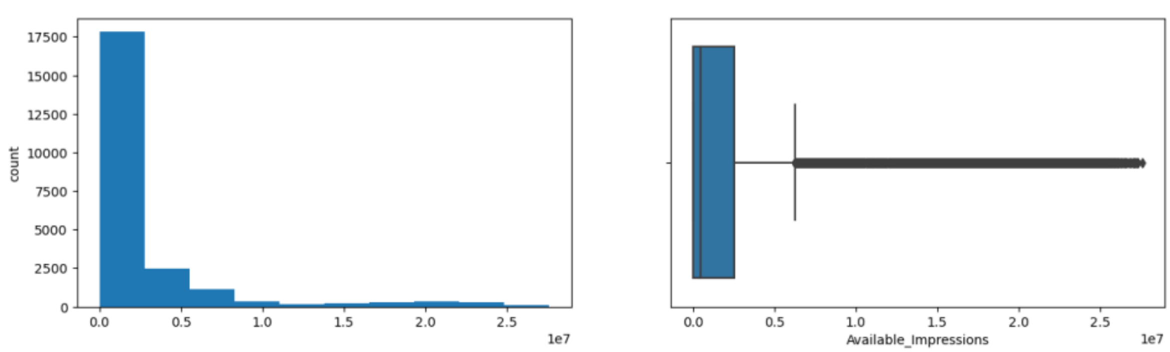




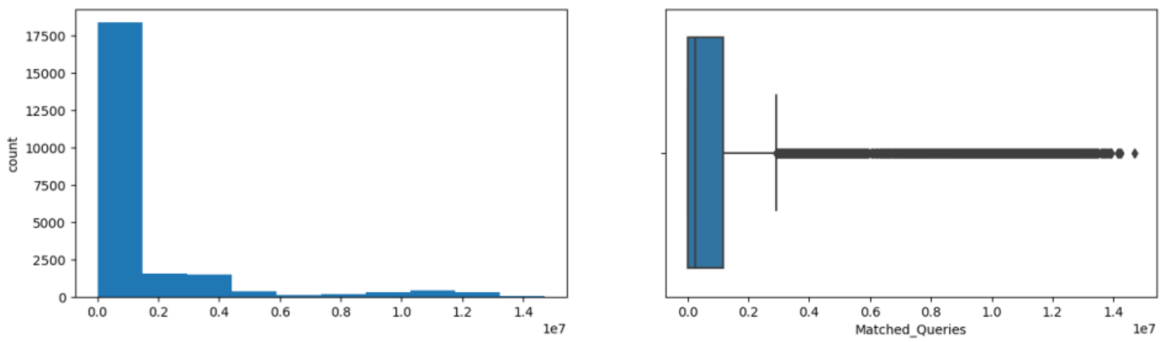
```
Description of Ad Size
----------------------------------------------------------------------------
count      23066.000000
mean       96674.468048
std        61538.329557
min        33600.000000
25%        72000.000000
50%        72000.000000
75%        84000.000000
max       216000.000000
Name: Ad Size, dtype: float64

Skew : 1.21


Distribution of Ad Size
----------------------------------------------------------------------------
```



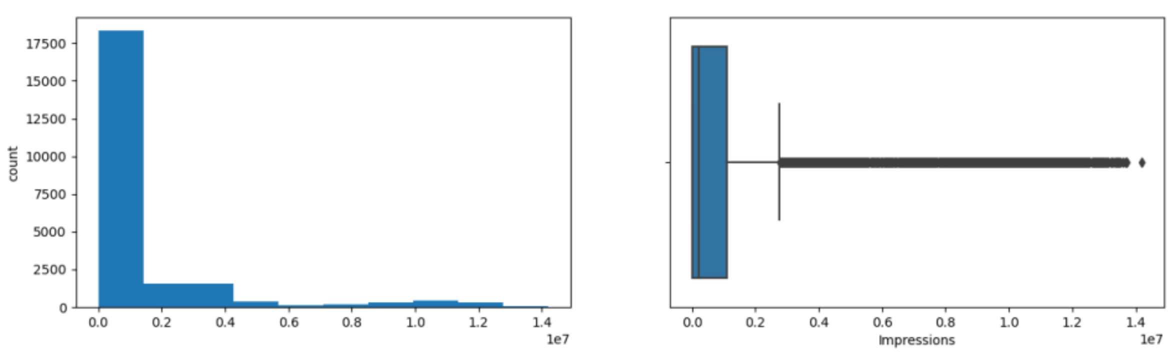

```
Description of Available_Impressions
----------------------------------------------------------------------------
count    2.306600e+04
mean     2.432044e+06
std      4.742888e+06
min      1.000000e+00
25%      3.367225e+04
50%      4.837710e+05
75%      2.527712e+06
max      2.759286e+07
Name: Available_Impressions, dtype: float64

Skew : 3.07


Distribution of Available_Impressions
----------------------------------------------------------------------------
```



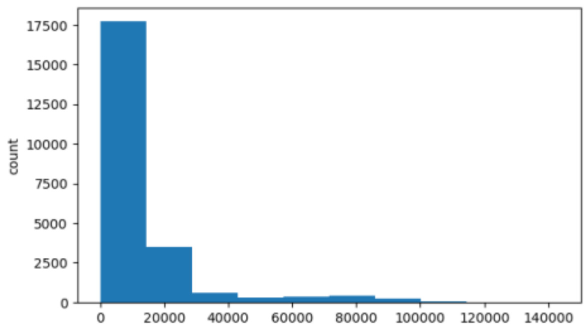
```
Description of Matched_Queries
----------------------------------------------------------------------------
count    2.306600e+04
mean     1.295099e+06
std      2.512970e+06
min      1.000000e+00
25%      1.828250e+04
50%      2.580875e+05
75%      1.180700e+06
max      1.470202e+07
Name: Matched_Queries, dtype: float64

Skew : 2.98

Distribution of Matched_Queries
----------------------------------------------------------------------------
```

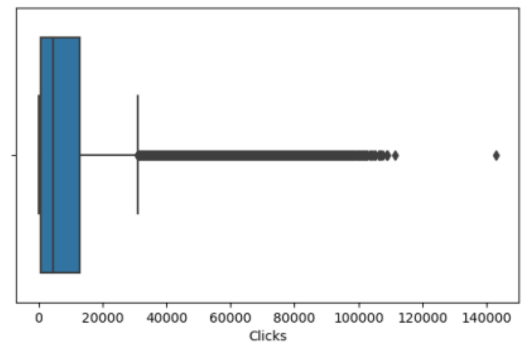



```
Description of Impressions
----------------------------------------------------------------------------
count    2.306600e+04
mean     1.241520e+06
std      2.429400e+06
min      1.000000e+00
25%      7.990500e+03
50%      2.252900e+05
75%      1.112428e+06
max      1.419477e+07
Name: Impressions, dtype: float64

Skew : 2.97

Distribution of Impressions
----------------------------------------------------------------------------
```

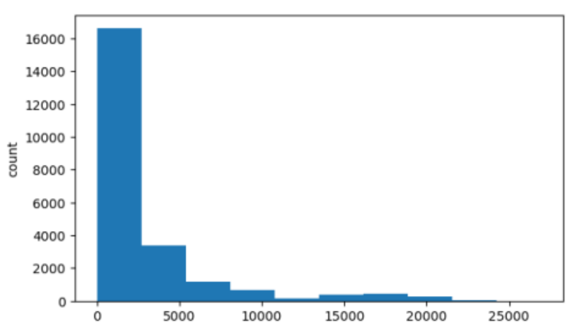


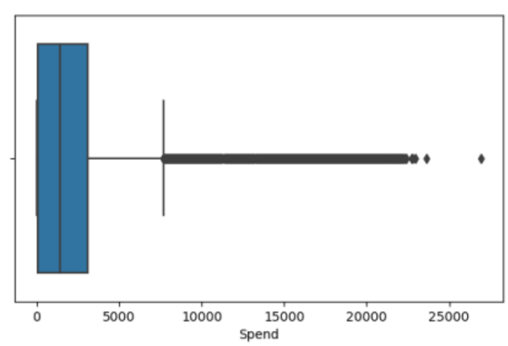
```
Description of Clicks
--------------------------------------------------------------------------------
count     23066.000000
mean      10678.518816
std       17353.409363
min           1.000000
25%         710.000000
50%        4425.000000
75%       12793.750000
max      143049.000000
Name: Clicks, dtype: float64

Skew : 2.94


Distribution of Clicks
--------------------------------------------------------------------------------
```





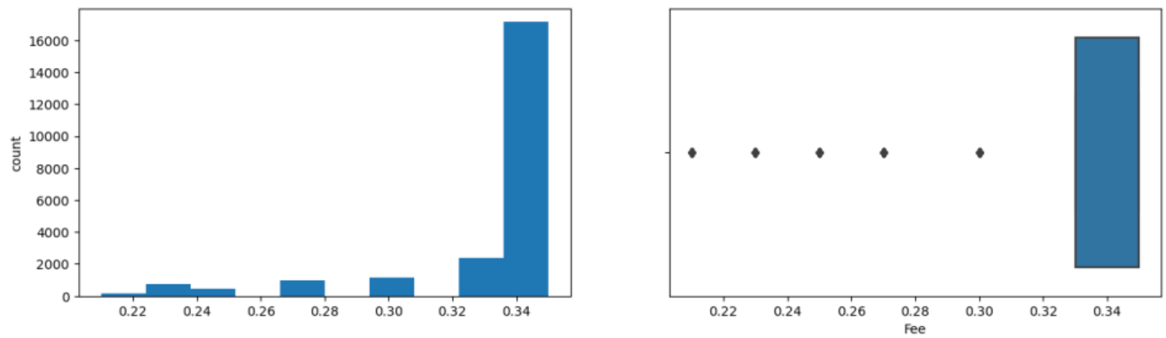

```
Description of Spend
--------------------------------------------------------------------------------
count    23066.000000
mean      2706.625689
std       4067.927273
min          0.000000
25%         85.180000
50%       1425.125000
75%       3121.400000
max      26931.870000
Name: Spend, dtype: float64

Skew : 2.58


Distribution of Spend
--------------------------------------------------------------------------------
```





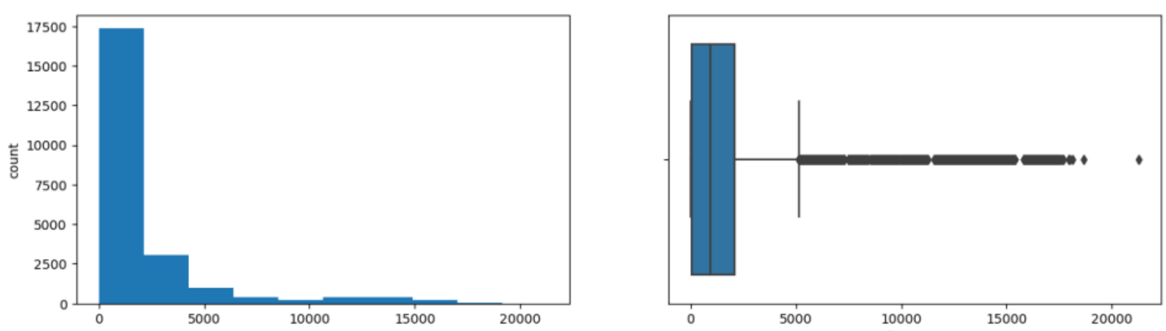
```
Description of Fee
---------------------------------------------------------------------------
count    23066.000000
mean         0.335123
std          0.031963
min          0.210000
25%          0.330000
50%          0.350000
75%          0.350000
max          0.350000
Name: Fee, dtype: float64

Skew : -2.3
```

```
Distribution of Fee
---------------------------------------------------------------------------
```

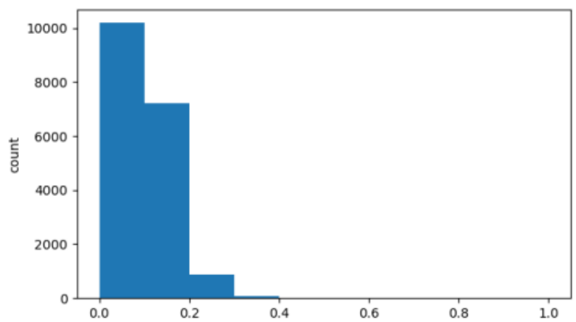


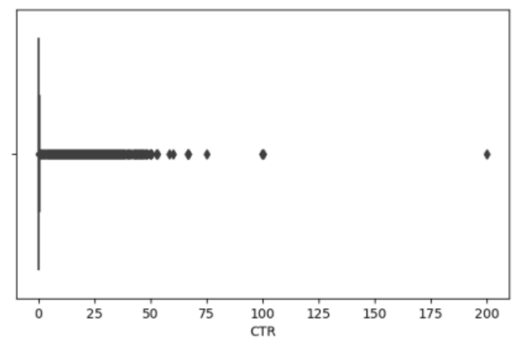

```
Description of Revenue
---------------------------------------------------------------------------
count    23066.000000
mean      1924.252331
std       3105.238410
min          0.000000
25%         55.365375
50%        926.335000
75%       2091.338150
max      21276.180000
Name: Revenue, dtype: float64

Skew : 2.79
```

```
Distribution of Revenue
---------------------------------------------------------------------------
```

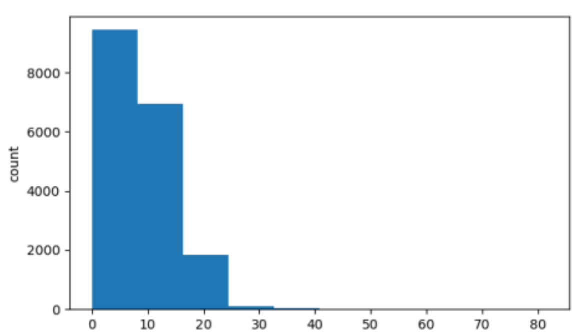


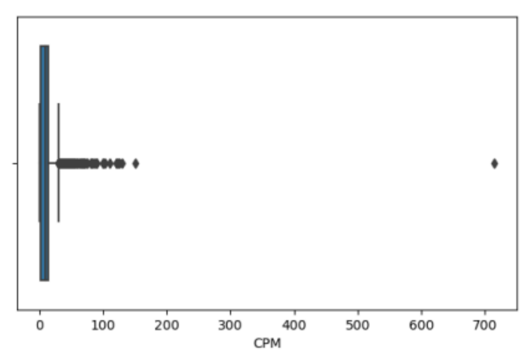
```
Description of CTR
----------------------------------------------------------------------------
count    23066.000000
mean         2.614863
std          7.853405
min          0.000100
25%          0.003400
50%          0.112650
75%          0.183778
max        200.000000
Name: CTR, dtype: float64

Skew : 5.43


Distribution of CTR
----------------------------------------------------------------------------
```



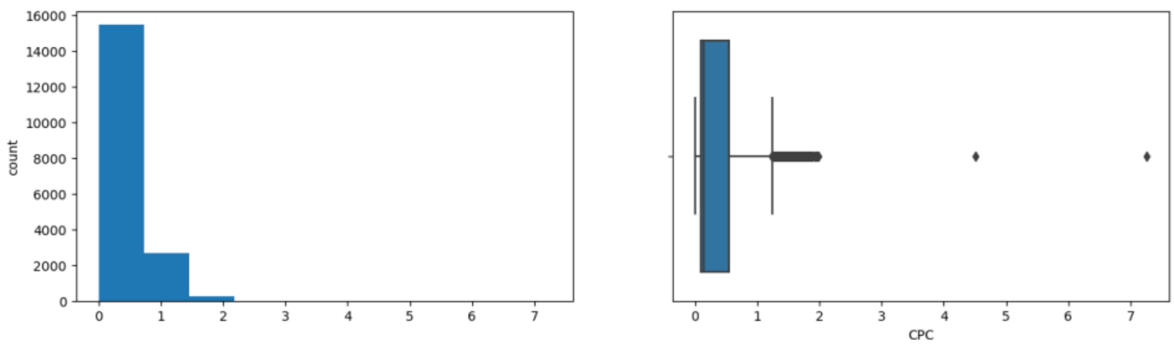



```
Description of CPM
----------------------------------------------------------------------------
count    23066.000000
mean         8.396730
std          9.057082
min          0.000000
25%          1.750000
50%          8.370742
75%         13.040000
max        715.000000
Name: CPM, dtype: float64

Skew : 22.32


Distribution of CPM
----------------------------------------------------------------------------
```





```
Description of CPC
--------------------------------------------------------------------------------
count    23066.000000
mean         0.336652
std          0.341231
min          0.000000
25%          0.090000
50%          0.140000
75%          0.550000
max          7.260000
Name: CPC, dtype: float64

Skew : 1.84


Distribution of CPC
--------------------------------------------------------------------------------
```



We can observe that there are outliers present in all except for the Ad Length and Ad Width column.
Fee is left skewed wherein mean <median
CPM is highly skewed.

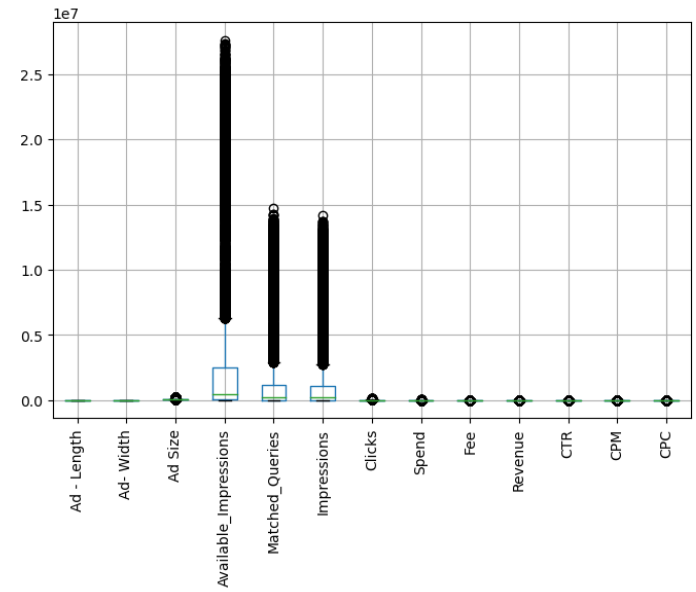
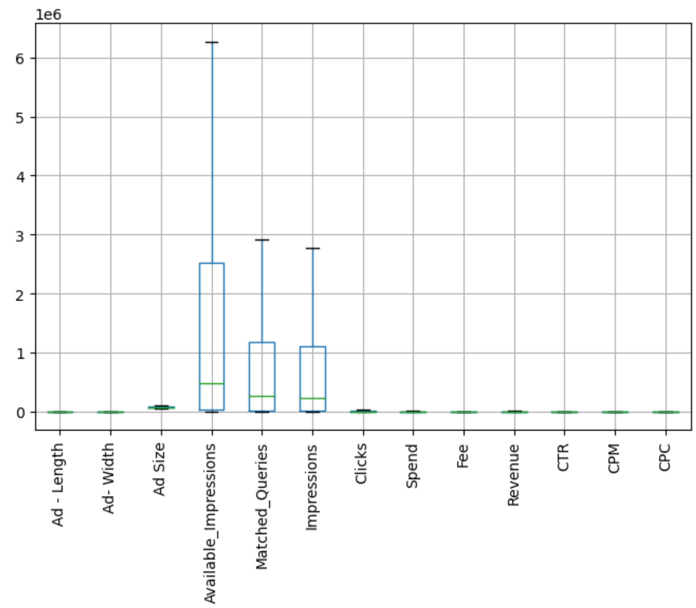
Since K-Means is sensitive to outliers, in order to perform clustering, we will consider outlier treatment to reduce their impact.

And here we are treating outliers using IQR (Interquartile Range).

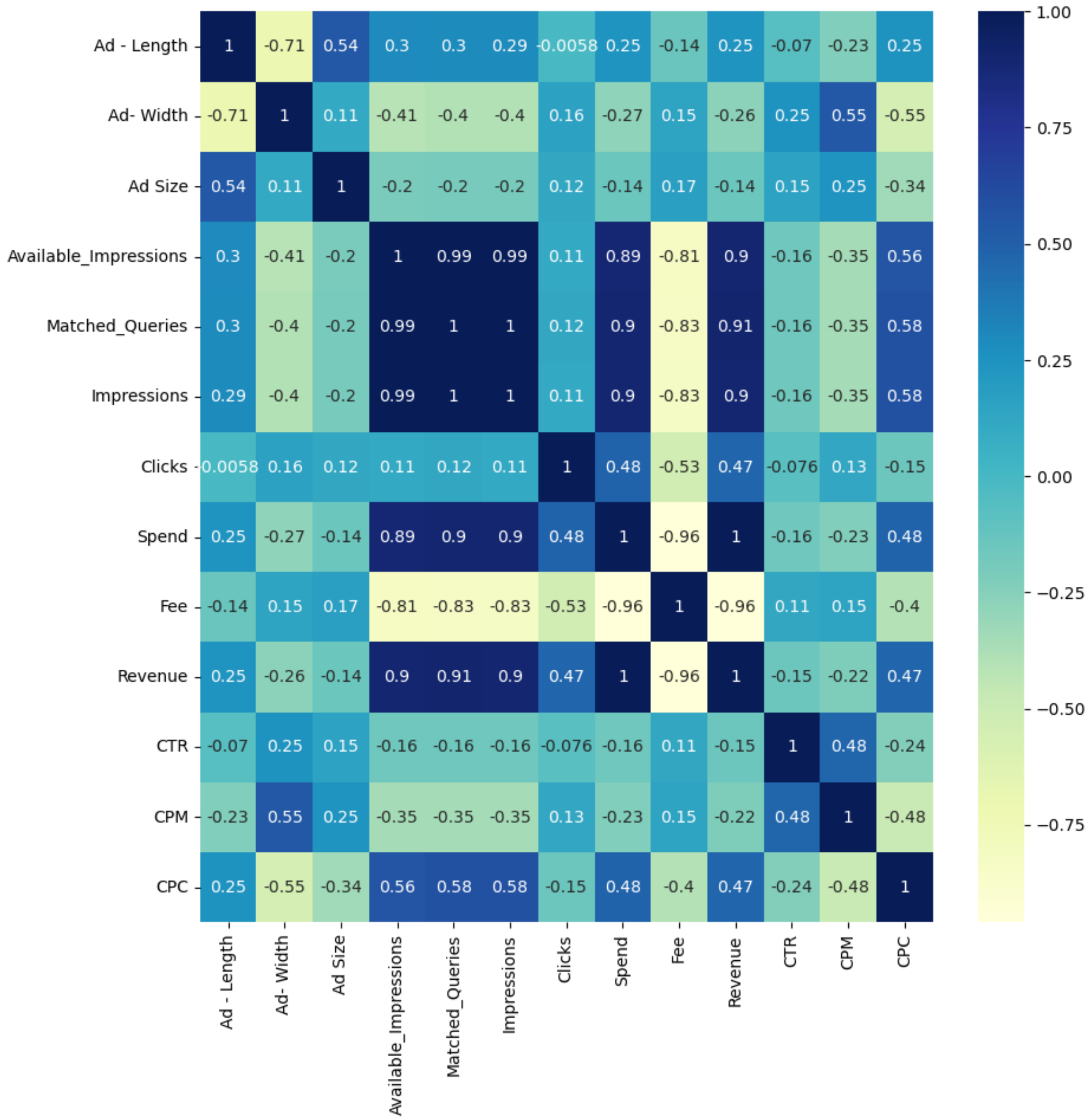
Number of outliers present in the data-

```
Ad - Length                0
Ad- Width                  0
Ad Size                 8448
Available_Impressions   2378
Matched_Queries         3192
Impressions             3269
Clicks                  1691
Spend                   2081
Fee                     3517
Revenue                 2325
CTR                     3487
CPM                      208
CPC                      568
dtype: int64
```

For the higher outliers we will treat it to get it at 95 percentile value and for Lower-level outliers we will treat it to get it at 5 percentile value.

Before Outlier	**After Outlier Treatment**
	

Bivariate analysis-

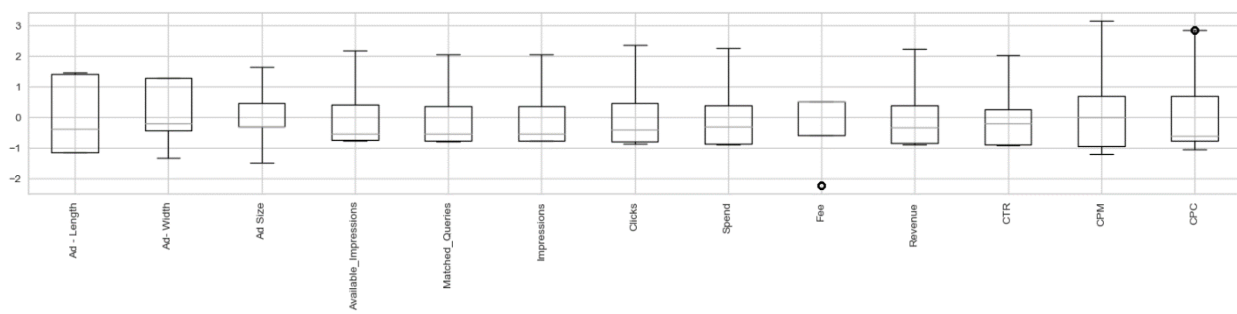


1.4 Perform z-score scaling and discuss how it affects the speed of the algorithm.

First 5 rows of the scaled data

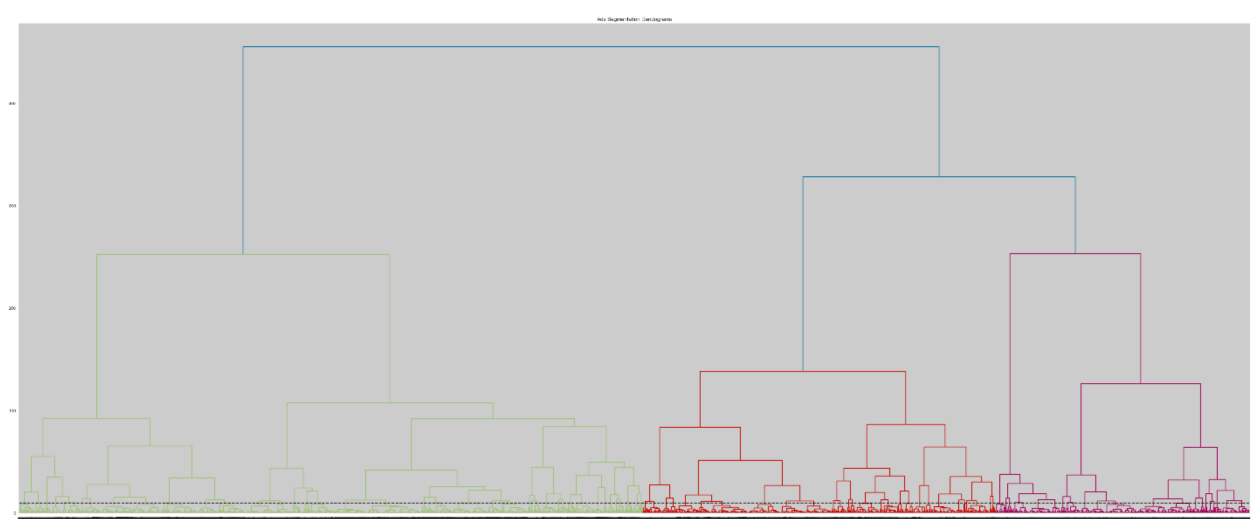
	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	-0.364496	-0.432797	-0.102518	-0.755333	-0.778949	-0.768478	-0.867488	-0.89317	0.535724	-0.880093	-0.891201	-1.194562	-1.04114
1	-0.364496	-0.432797	-0.102518	-0.755345	-0.778988	-0.768516	-0.867488	-0.89317	0.535724	-0.880093	-0.888615	-1.194562	-1.04114
2	-0.364496	-0.432797	-0.102518	-0.754900	-0.778919	-0.768445	-0.867488	-0.89317	0.535724	-0.880093	-0.893142	-1.194562	-1.04114
3	-0.364496	-0.432797	-0.102518	-0.755040	-0.778781	-0.768302	-0.867488	-0.89317	0.535724	-0.880093	-0.898315	-1.194562	-1.04114
4	-0.364496	-0.432797	-0.102518	-0.755610	-0.779030	-0.768560	-0.867488	-0.89317	0.535724	-0.880093	-0.884734	-1.194562	-1.04114

```
<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
```

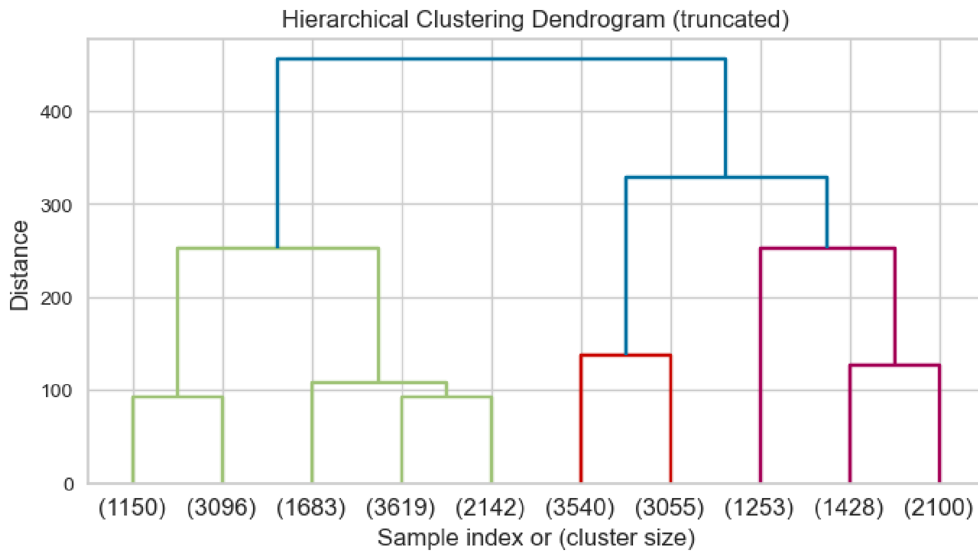


In the scaled data, the mean tends to 0 and Standard deviation to 1.
It is a preprocessing step in data analysis which helps in reducing the influence of features with larger scales or high variances so that we can obtain clusters more efficiently.

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



Below is the truncated dendrogram displaying the last 10 clusters (p=10).

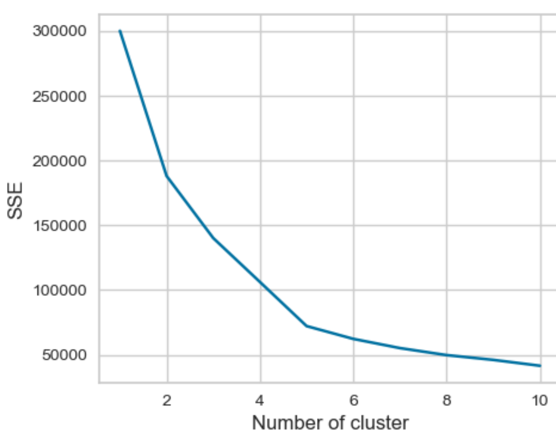


1.6 Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.

Within sum of squares-

```
{1: 299858.0,
 2: 187902.64770993276,
 3: 139992.87426412938,
 4: 106152.74229789544,
 5: 72133.6934158383,
 6: 62259.98939794785,
 7: 55151.52147681743,
 8: 49733.040051637116,
 9: 46049.7390221088,
 10: 41531.157690828666}
```

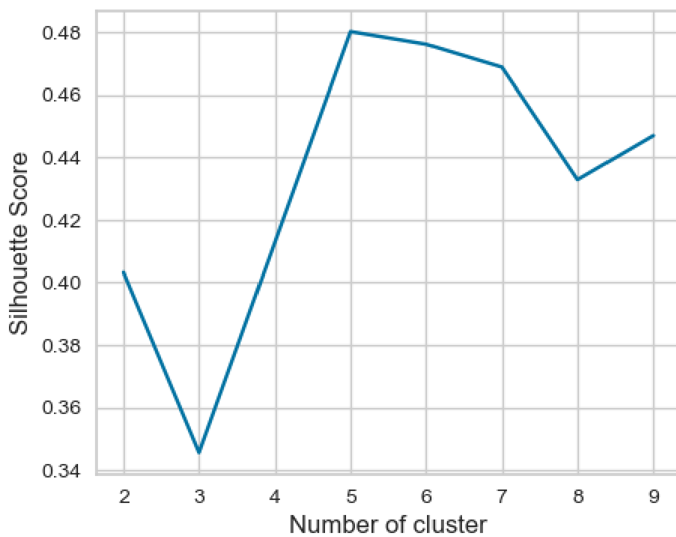
Elbow Plot:



We have used WSS to check the optimal number of clusters. And can observe that the WSS reduces as K keeps increasing.
There is a significant wss value drop as we move from k= 2 to k=5.
We can choose any from 2 to 5 as our number of clusters. Since there is not a significant reduction at k=6, we will choose 5 as the optimal number of clusters.

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

```
{2: 0.40318725804432765,
 3: 0.34547066630442486,
 4: 0.41284225649057377,
 5: 0.48020321346347616,
 6: 0.47613989974053916,
 7: 0.46883074857917595,
 8: 0.43286664054059454,
 9: 0.4470009074272004}
```



We can observe that the silhouette score is highest for k=5. So we can consider the optimal number of clusters as 5.

1.8 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.

We have grouped data based on 5 clusters. Clus_kmeans column denotes the cluster number.

	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	Clus_kmeans
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0	0.35	0.0	0.0031	0.0	0.0	3
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.0	0.35	0.0	0.0035	0.0	0.0	3
2	2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0	0.35	0.0	0.0028	0.0	0.0	3
3	2020-9-3-20	Format1	300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.0	0.35	0.0	0.0020	0.0	0.0	3
4	2020-9-4-15	Format1	300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.0	0.35	0.0	0.0041	0.0	0.0	3

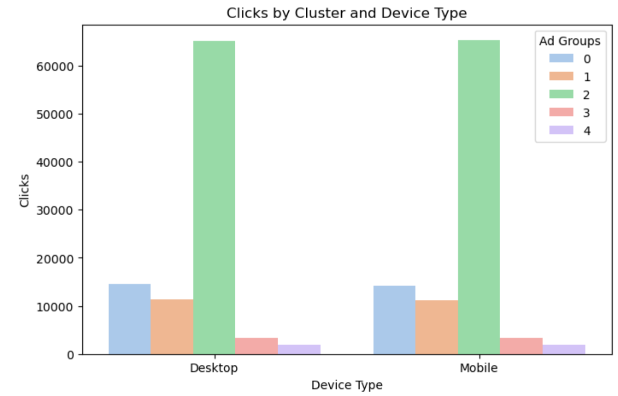
Frequency of the clusters-

```
0    4699
1    4049
2    1539
3    6139
4    6640
Name: Clus_kmeans, dtype: int64
```

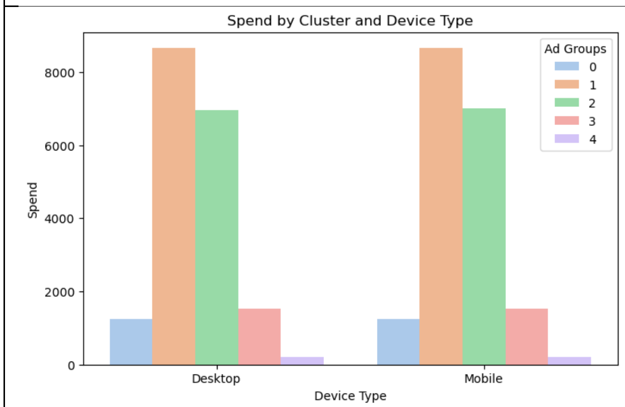
We have calculated the mean of the original data for each label and grouped them accordingly.

	group_0 Mean	group_1 Mean	group_2 Mean	group_3 Mean	group_4 Mean
Ad - Length	681.939136	4.658810e+02	141.543860	4.244913e+02	146.024096
Ad- Width	305.309640	1.992122e+02	572.482131	1.462127e+02	568.373494
Ad Size	206053.202809	7.520506e+04	75680.311891	5.350448e+04	77139.759036
Available_Impressions	263137.601405	1.039627e+07	805593.964263	1.838153e+06	36489.929217
Matched_Queries	141872.516706	5.630305e+06	566390.274854	8.783969e+05	21813.339006
Impressions	120873.592041	5.451651e+06	477750.160494	8.398562e+05	15667.703916
Clicks	14361.254097	1.125400e+04	65260.276803	3.304247e+03	1888.464759
Spend	1254.077176	8.653044e+03	6985.407472	1.524133e+03	210.052837
Fee	0.349545	2.903853e-01	0.288356	3.492344e-01	0.349991
Revenue	816.684693	6.378677e+03	5013.785448	9.931507e+02	136.562174
CTR	4.473420	3.417136e-02	2.208029	6.252699e-02	5.327328
CPM	12.046037	1.572871e+00	15.390007	1.805811e+00	14.448043
CPC	0.091051	7.607038e-01	0.111935	5.524783e-01	0.104421

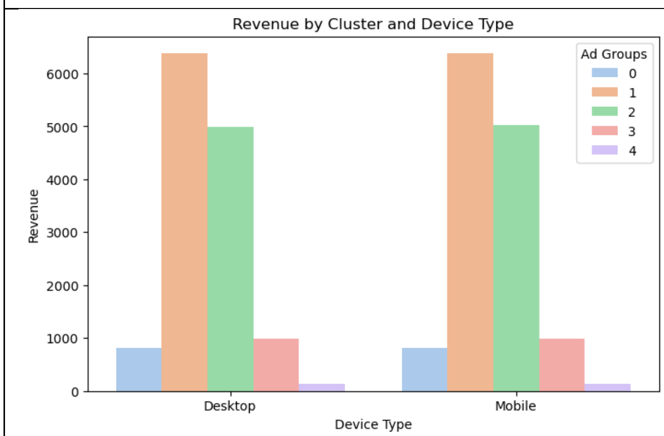
	Clus_kmeans	Device Type	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	freq
0	0	Desktop	14501.612634	1253.277342	0.349547	816.130957	4.443102	12.022119	0.090368	4699.0
1	0	Mobile	14283.292618	1254.521440	0.349543	816.992263	4.490260	12.059322	0.091430	4049.0
2	1	Desktop	11327.692886	8647.606334	0.290773	6374.965943	0.034836	1.560723	0.754071	1539.0
3	1	Mobile	11212.350599	8656.117445	0.290166	6380.773762	0.033795	1.579737	0.764452	6139.0
4	2	Desktop	65203.080645	6965.583154	0.288548	4997.376742	2.287759	15.440287	0.111979	6640.0
5	2	Mobile	65292.810398	6996.683690	0.288247	5023.118841	2.162678	15.361408	0.111910	NaN
6	3	Desktop	3311.025688	1522.011083	0.349257	991.712318	0.061328	1.810714	0.549474	NaN
7	3	Mobile	3300.514271	1525.300859	0.349222	993.942780	0.063187	1.803111	0.554132	NaN
8	4	Desktop	1918.356003	208.883552	0.349992	135.799904	5.314257	14.262153	0.104181	NaN
9	4	Mobile	1871.743072	210.706956	0.349991	136.988601	5.334641	14.552033	0.104555	NaN



We can observe that there is maximum click on 3rd cluster (Ad Group 2) in both desktop and mobile device.

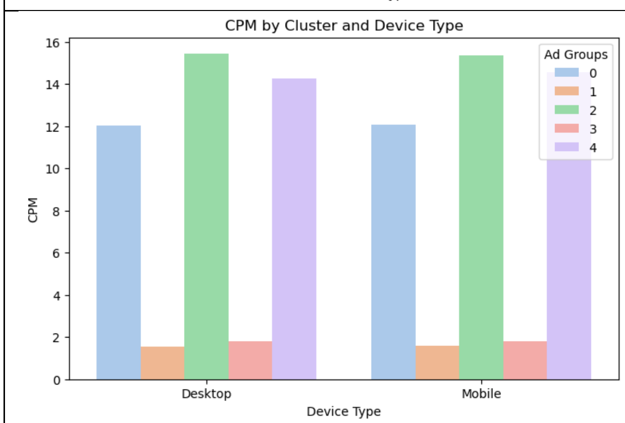


We can observe that there is maximum amount of money spent on specific ad variations for a campaign by 2nd cluster (Ad Group 1) followed by 3rd cluster (Ad Group 2) in both desktop and mobile device.
And least by 5th Cluster (Ad Group 4).



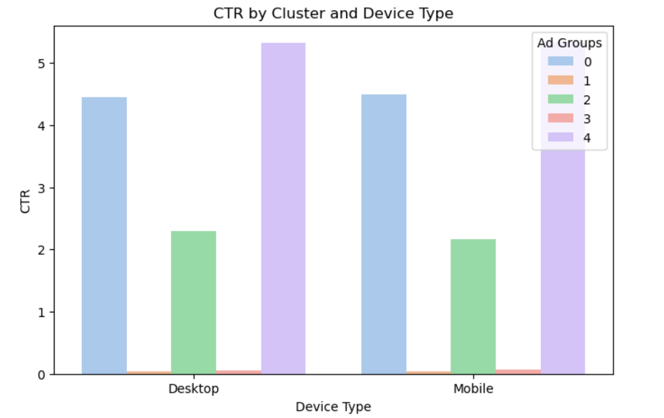
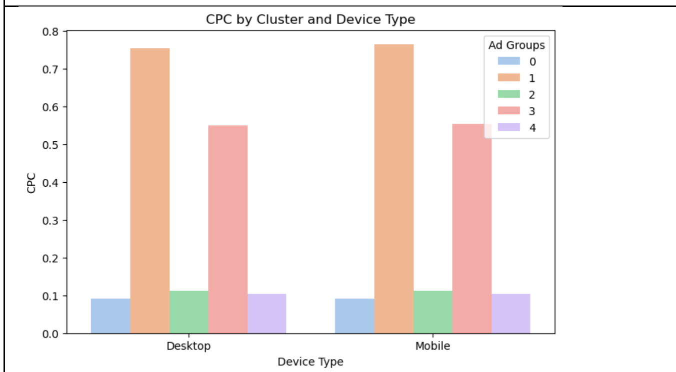
We can observe that maximum income has been earned by 2nd cluster (Ad Group 1) followed by 3rd cluster (Ad Group 2) and least by 5th Cluster (Ad Group 4).

And there is not much variations amongst the device types.



We can observe that the cost per 1000 impressions here is maximum for 3rd cluster (Ad Group 2) with a slight or very little variation with the 5th cluster (Ad Group 4).

And there is very less on 2nd and 4th cluster i.e. Ad Group 1, Ad Group 3.

<table>
<tr>
<td>
</td>
<td>We can observe that the reach created for an ad has been high by 5th cluster (Ad Group 4) followed by 1st cluster (Ad Group 0) and least created by 2nd and 4th cluster i.e. Ad Group 1 and Ad Group 3.</td>
</tr>
<tr>
<td>
</td>
<td>We can observe that the Cluster 2 (Ad Group 1) has paid more amount comparatively for each of the clicks on their ads followed by Cluster 4 (Ad Group 3).</td>
</tr>
</table>

1.9 Conclude the project by providing summary of your learnings.

Below is the final clustered dataset.

FinalClustering.csv

- There are 23066 observations and 19 columns in the data
- 6 variables of Object data type and 13 variables of numeric data type (int and float)
- We have also observed that there are no duplicate values in the dataframe.
- The columns CTR, CPM and CPC contain null values i.e. 4736 entries each
- We treated null values using the given formula-

  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.

  Post which there were no null values in the data.
- There were outliers present in many columns.

- Fee was left skewed wherein mean <median
- CPM was highly skewed.
- Since K-Means is sensitive to outliers, in order to perform clustering, we considered outlier treatment to reduce their impact using IQR (Interquartile Range).
- We scaled the data, the mean tended to 0 and Standard deviation to 1.
  It is a preprocessing step in data analysis which helps in reducing the influence of features with larger scales or high variances so that we can obtain clusters more efficiently.
- Performed truncated dendrogram displaying the last 10 clusters (p=10).
- We used WSS to check the optimal number of clusters. And observed that the WSS reduces as K increased.
- There was a significant wss value drop as we move from k= 2 to k=5.
  Since there is not a significant reduction at
  k=6, we chose 5 as the optimal number of clusters.
- Silhouette score was highest for k=5. So we considered the optimal number of clusters as 5.
- We grouped data based on 5 clusters.
- Cluster 5 had high frequency and 3rd cluster the least.
- The device type did not have much impact on the Clicks, Spend, Fee, Revenue, CTR, CPM, CPC.
- Based on the analysis, we found that Ad Group 4 is more preferrable followed by Ad group 3 and the Ad Group 0 the least preferrable.

Problem 2: 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 Houseless 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 like checking head, info, summary, nulls, and duplicates, etc.

First 5 rows of dataframe

	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_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180	237	680	252	32	4
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123	229	186	148	76	17
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	...	114	188	44	89	3	34	0	
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	...	194	247	61	128	13	50	4	1
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	...	874	1928	465	1043	205	302	24	1(

Last 5 rows of dataframe

	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_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_
635	34	636	Puducherry	Mahe	3333	8154	11781	1146	1203	21	...	32	47	0	0	0	0	0	
636	34	637	Puducherry	Karaikal	10612	12346	21691	1544	1533	2234	...	155	337	3	14	38	130	4	
637	35	638	Andaman & Nicobar Island	Nicobars	1275	1549	2630	227	225	0	...	104	134	9	4	2	6	17	
638	35	639	Andaman & Nicobar Island	North & Middle Andaman	3762	5200	8012	723	664	0	...	136	172	24	44	11	21	1	
639	35	640	Andaman & Nicobar Island	South Andaman	7975	11977	18049	1470	1358	0	...	173	122	6	2	17	17	2	

Basic information about the dataframe-

```
<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
 27  MAIN_OT_M       640 non-null    int64
 28  MAIN_OT_F       640 non-null    int64
 29  MARGWORK_M      640 non-null    int64
 30  MARGWORK_F      640 non-null    int64

 31  MARG_CL_M       640 non-null    int64
 32  MARG_CL_F       640 non-null    int64
 33  MARG_AL_M       640 non-null    int64
 34  MARG_AL_F       640 non-null    int64
 35  MARG_HH_M       640 non-null    int64
 36  MARG_HH_F       640 non-null    int64
 37  MARG_OT_M       640 non-null    int64
 38  MARG_OT_F       640 non-null    int64
 39  MARGWORK_3_6_M  640 non-null    int64
 40  MARGWORK_3_6_F  640 non-null    int64
 41  MARG_CL_3_6_M   640 non-null    int64
 42  MARG_CL_3_6_F   640 non-null    int64
 43  MARG_AL_3_6_M   640 non-null    int64
 44  MARG_AL_3_6_F   640 non-null    int64
 45  MARG_HH_3_6_M   640 non-null    int64
 46  MARG_HH_3_6_F   640 non-null    int64
 47  MARG_OT_3_6_M   640 non-null    int64
 48  MARG_OT_3_6_F   640 non-null    int64
 49  MARGWORK_0_3_M  640 non-null    int64
 50  MARGWORK_0_3_F  640 non-null    int64
 51  MARG_CL_0_3_M   640 non-null    int64

 52  MARG_CL_0_3_F   640 non-null    int64
 53  MARG_AL_0_3_M   640 non-null    int64
 54  MARG_AL_0_3_F   640 non-null    int64
 55  MARG_HH_0_3_M   640 non-null    int64
 56  MARG_HH_0_3_F   640 non-null    int64
 57  MARG_OT_0_3_M   640 non-null    int64
 58  MARG_OT_0_3_F   640 non-null    int64
 59  NON_WORK_M      640 non-null    int64
 60  NON_WORK_F      640 non-null    int64
dtypes: int64(59), object(2)
memory usage: 305.1+ KB
```

It contains 640 entries.

Two of object data type and 59 of numeric data type

State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F
count	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	...	640.000000	640.000000	640.000000
mean	17.114062	320.500000	51222.871875	79940.576563	122372.084375	12309.098438	11942.300000	13820.946875	20778.392188	6191.807813	...	1392.973438	2757.050000	250.889062
std	9.426486	184.896367	48135.405475	73384.511114	113600.717282	11500.906881	11326.294567	14426.373130	21727.887713	9912.668948	...	1489.707052	2788.776676	453.336594
min	1.000000	1.000000	350.000000	391.000000	698.000000	56.000000	56.000000	0.000000	0.000000	0.000000	...	4.000000	30.000000	0.000000
25%	9.000000	160.750000	19484.000000	30228.000000	46517.750000	4733.750000	4672.250000	3466.250000	5603.250000	293.750000	...	489.500000	957.250000	47.000000
50%	18.000000	320.500000	35837.000000	58339.000000	87724.500000	9159.000000	8663.000000	9591.500000	13709.000000	2333.500000	...	949.000000	1928.000000	114.500000
75%	24.000000	480.250000	68892.000000	107918.500000	164251.750000	16520.250000	15902.250000	19429.750000	29180.000000	7658.000000	...	1714.000000	3599.750000	270.750000
max	35.000000	640.000000	310450.000000	485417.000000	750392.000000	96223.000000	95129.000000	103307.000000	156429.000000	96785.000000	...	9875.000000	21611.000000	5775.000000

Shape: `(640, 61)`

There are no duplicates found in the dataframe.

Null value check:

```
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
```

No null values found.

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 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

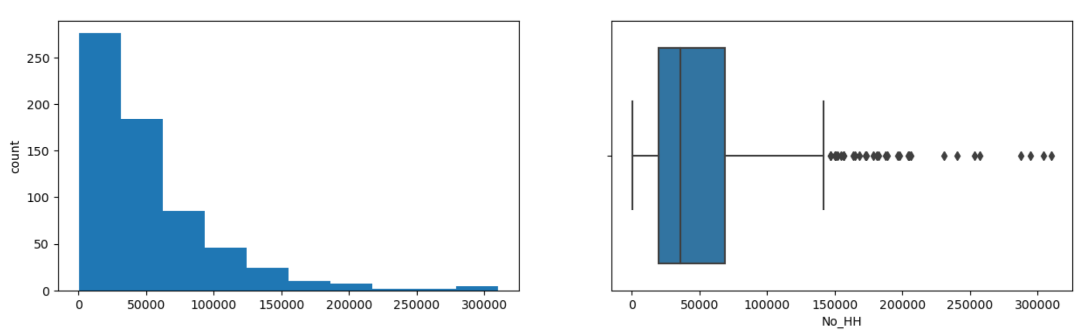
We are choosing following variables for analysis-

No_HH	No of Household
TOT_M	Total population Male

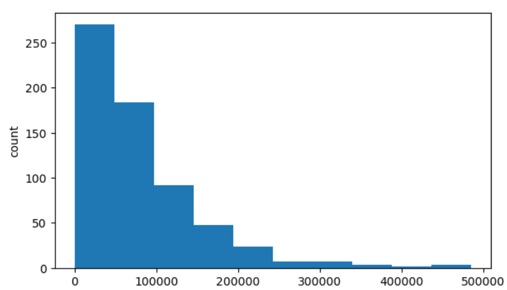
TOT_F	Total population Female
TOT_WORK_M	Total Worker Population Male
TOT_WORK_F	Total Worker Population Female

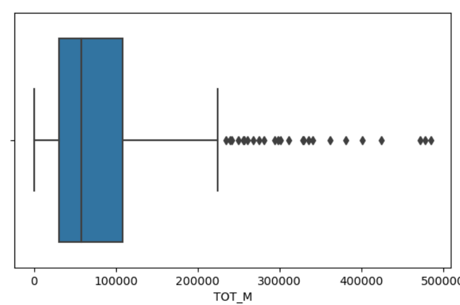
Univariate analysis-

```
Description of No_HH
----------------------------------------------------------------------------
count        640.000000
mean       51222.871875
std        48135.405475
min          350.000000
25%        19484.000000
50%        35837.000000
75%        68892.000000
max       310450.000000
Name: No_HH, dtype: float64

Skew : 2.02


Distribution of No_HH
----------------------------------------------------------------------------
```

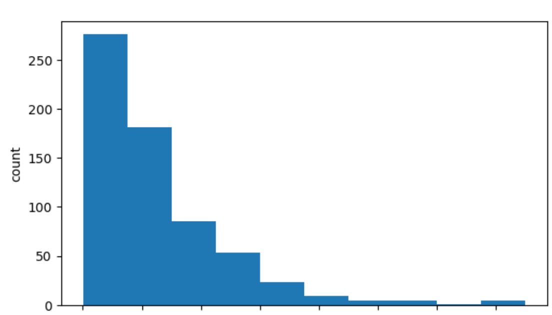


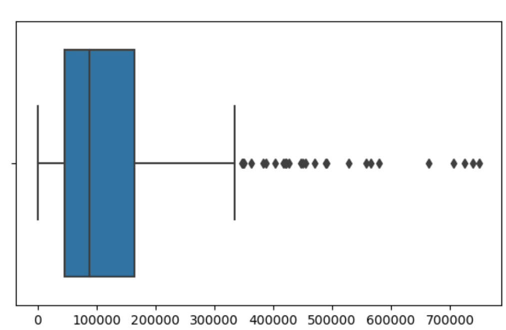
```
Description of TOT_M
----------------------------------------------------------------------------
count       640.000000
mean      79940.576563
std       73384.511114
min         391.000000
25%       30228.000000
50%       58339.000000
75%      107918.500000
max      485417.000000
Name: TOT_M, dtype: float64

Skew : 2.03


Distribution of TOT_M
----------------------------------------------------------------------------
```



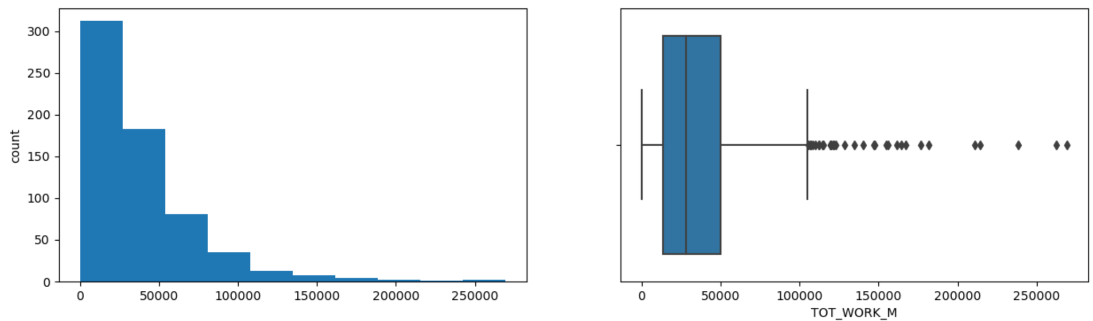



```
Description of TOT_F
----------------------------------------------------------------------------
count       640.000000
mean     122372.084375
std      113600.717282
min         698.000000
25%       46517.750000
50%       87724.500000
75%      164251.750000
max      750392.000000
Name: TOT_F, dtype: float64

Skew : 2.11


Distribution of TOT_F
----------------------------------------------------------------------------
```

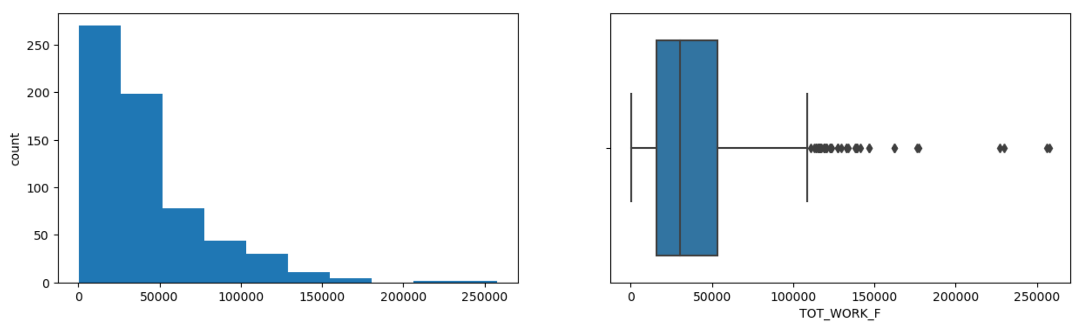




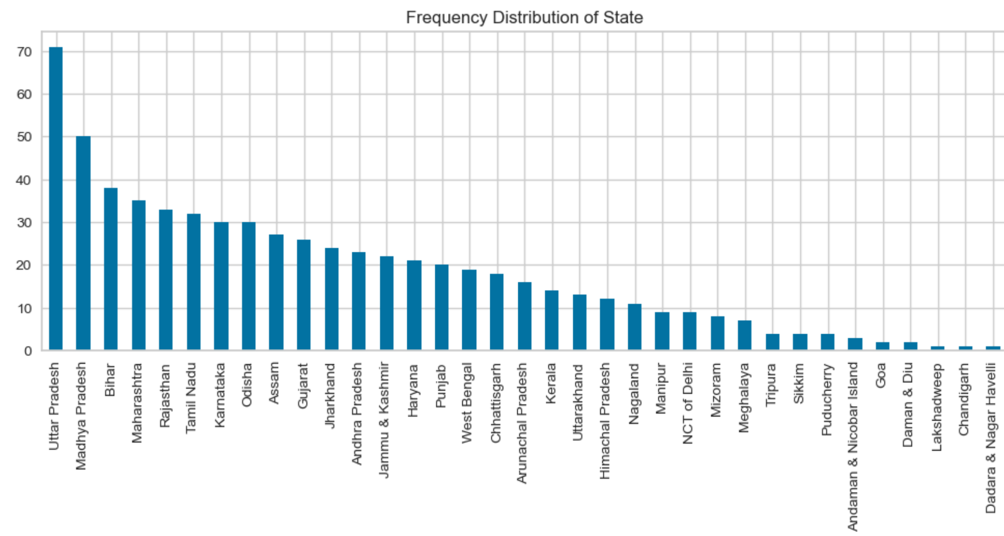
```
Description of TOT_WORK_M
--------------------------------------------------------------------------------
count       640.000000
mean      37992.407813
std       36419.537491
min         100.000000
25%       13753.500000
50%       27936.500000
75%       50226.750000
max      269422.000000
Name: TOT_WORK_M, dtype: float64

Skew : 2.3


Distribution of TOT_WORK_M
```



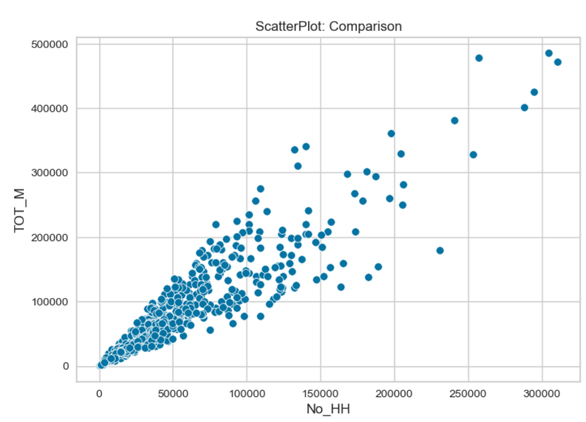

```
Description of TOT_WORK_F
----------------------------------------------------------------------------
count       640.000000
mean      41295.760938
std       37192.360943
min         357.000000
25%       16097.750000
50%       30588.500000
75%       53234.250000
max      257848.000000
Name: TOT_WORK_F, dtype: float64

Skew : 1.93


Distribution of TOT_WORK_F
----------------------------------------------------------------------------
```

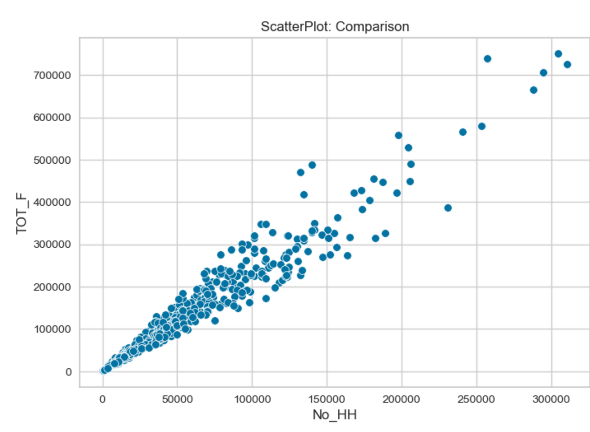


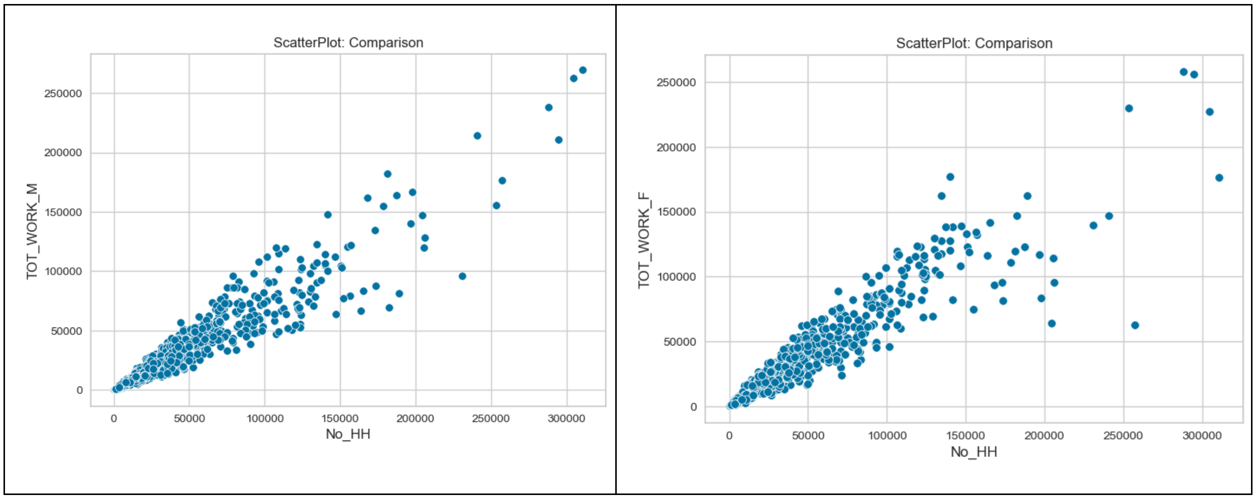
All the entries are positively skewed and contains outliers.



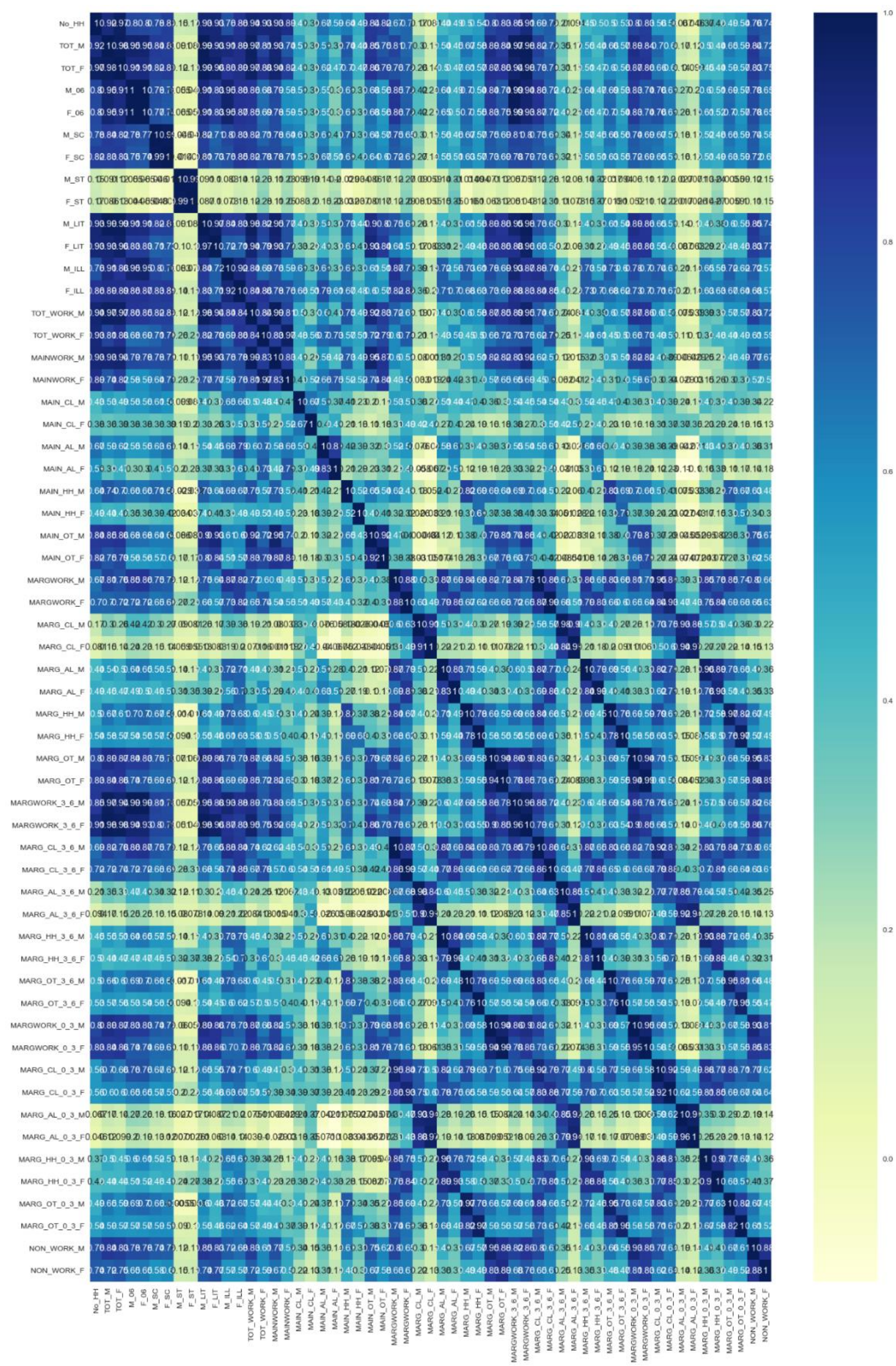
Bivariate analysis-







ScatterPlot: Comparison
TOT_WORK_M
No_HH
0
50000
100000
150000
200000
250000
300000
ScatterPlot: Comparison
TOT_WORK_F
No_HH



All the variables are positively correlated to each other

We can obtain Gender ratio using following formula-

Gender ratio= Female pop *1000 / male pop

```
State
Uttar Pradesh               54132.937245
Madhya Pradesh              31775.695694
Bihar                       28359.179844
Rajasthan                   22689.095737
Maharashtra                 20137.830181
Karnataka                   19094.178717
Assam                       18512.947635
Tamil Nadu                  17342.585846
Gujarat                     17221.226828
Odisha                      16659.085662
Haryana                     16332.590075
Jharkhand                   16263.942064
Jammu & Kashmir             16016.296979
Punjab                      14956.868233
West Bengal                 12337.016850
Andhra Pradesh              12303.635809
Chhattisgarh                 9673.837123
Arunachal Pradesh            9148.075615
Kerala                       8284.307828
Uttarakhand                  8045.597434
Himachal Pradesh             7545.342854
NCT of Delhi                 6897.924454
Nagaland                     6467.275656
Manipur                      5773.922018
Meghalaya                    5152.737661
Mizoram                      5064.937805
Sikkim                       2628.330933
Tripura                      2489.246453
Puducherry                   2407.197208
Andaman & Nicobar Island     1901.582314
Daman & Diu                  1404.381804
Goa                          1240.316274
Lakshadweep                   868.061197
Chandigarh                    700.036886
Dadara & Nagar Havelli        644.631151
Name: Gender ratio, dtype: float64
```

```
Area Name
Aurangabad       1380.559600
Hamirpur         1343.379253
Bilaspur         1247.883501
Bijapur          1158.490941
Raigarh          1040.772462
                     ...
Baudh             451.455047
West Godavari     450.075676
Virudhunagar      449.351612
Koraput           440.768731
Krishna           437.972258
Name: Gender ratio, Length: 635, dtype: float64
```

We can conclude that Uttar Pradesh has highest gender ratio and Dadara & Nagar Havelli with the least.

In terms of area, Aurangabad has highest gender ratio while Krishna the least.

Considering around 50,000 households, we have total of around 79,000 male population and 1,00,000 female population.

In which, total working population of male is around 38,000 and 41,000 female.

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

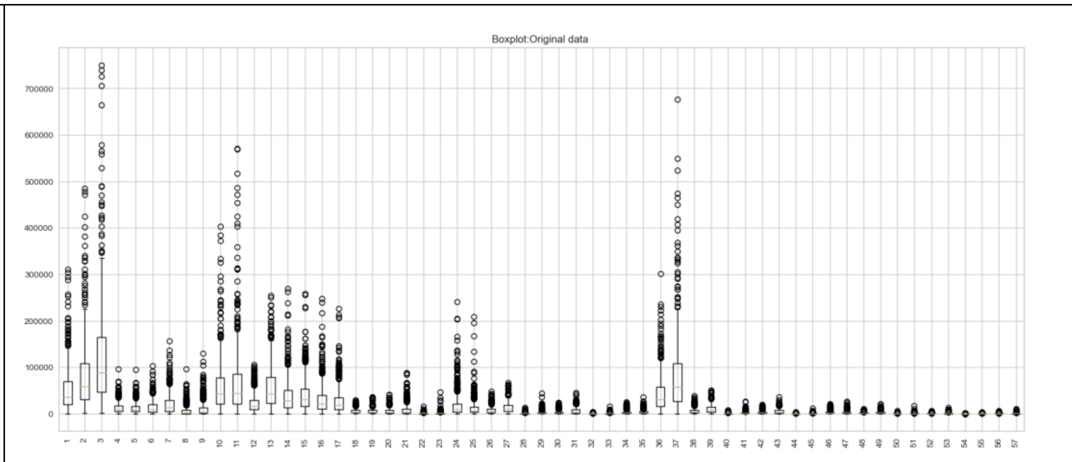
Treating outliers depends on the nature of data. Since we have scaled the data to make sure the high variance data does not impact, outlier treatment is not necessary in this case.

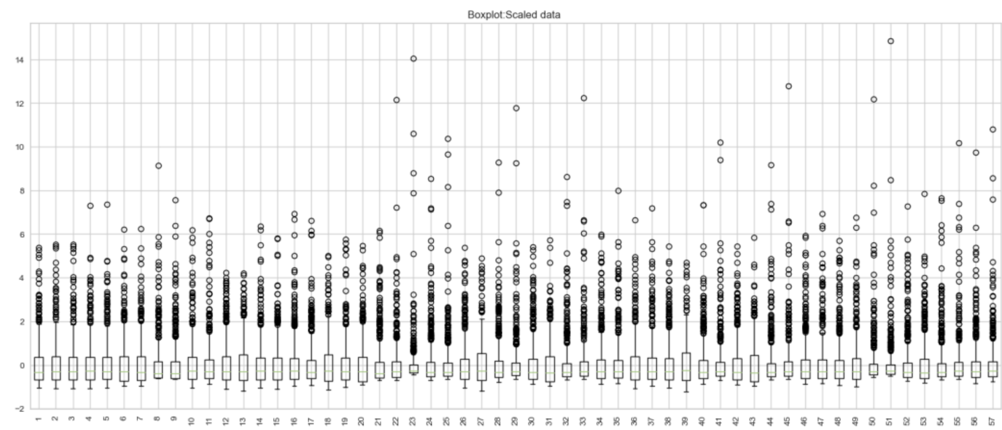
2.4 Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.

Printing first few rows of scaled data:

	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_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
0	-0.904738	-0.771236	-0.815563	-0.561012	-0.507738	-0.958575	-0.957049	-0.423306	-0.476423	-0.798097	...	-0.163229	-0.720610	-0.156494	-0.287524	0.156577	-0.657412	-0.36
1	-0.935695	-0.823100	-0.874534	-0.681096	-0.725367	-0.958297	-0.956772	-0.582014	-0.607607	-0.849434	...	-0.583103	-0.732811	-0.282327	-0.294688	-0.491731	-0.723062	0.04
2	-0.972412	-1.000919	-0.981466	-0.976956	-0.965262	-0.958575	-0.956772	-0.038951	-0.027273	-0.956457	...	-0.859212	-0.921931	-0.456727	-0.420050	-0.731894	-0.795026	-0.66
3	-1.037530	-1.052224	-1.041001	-1.022118	-0.995393	-0.958783	-0.957049	-0.355965	-0.390060	-1.004643	...	-0.805468	-0.900758	-0.419198	-0.385127	-0.718770	-0.784926	-0.62
4	-0.822676	-0.809381	-0.813933	-0.622359	-0.649908	-0.957395	-0.955529	0.149238	0.043330	-0.800568	...	-0.348645	-0.297513	0.472670	0.434200	-0.466796	-0.625849	-0.43

Before



After	

Hence, we can observe that scaling does not have much impact on outliers.

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

Bartletts Test of Sphericity to test 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

P value= 0.0

p value < 0.05, hence alternate hypothesis is true and we can agree that there is atleast one pair of variables in the data wihich are correlated. Hence PCA is recommended.

KMO Test to check whether we have adequate no. of observations to perform PCA.

MSA:
0.8039889932781528

Since msa > 0.7 , we are expected to provide a considerable reduction is the dimension and extraction of meaningful components. Hence, can perform PCA.

Co-variance matrix-

```
[[1.   0.92 0.97 ... 0.54 0.76 0.74]
 [0.92 1.   0.98 ... 0.59 0.85 0.72]
 [0.97 0.98 1.   ... 0.57 0.83 0.75]
 ...
 [0.54 0.59 0.57 ... 1.   0.61 0.52]
 [0.76 0.85 0.83 ... 0.61 1.   0.88]
 [0.74 0.72 0.75 ... 0.52 0.88 1.  ]]
```

Since we performed z scaling, co-variance and correlation matrix are same.

These are 57 PC scores-

```
array([[-4.62,  0.14,  0.33, ..., -0.  ,  0.  ,  0.  ],
       [-4.77, -0.11,  0.24, ...,  0.  ,  0.  ,  0.  ],
       [-5.96, -0.29,  0.37, ...,  0.  , -0.  ,  0.  ],
       ...,
       [-6.29, -0.64,  0.11, ...,  0.  ,  0.  , -0.  ],
       [-6.22, -0.67,  0.27, ..., -0.  , -0.  ,  0.  ],
       [-5.9 , -0.94,  0.35, ..., -0.  ,  0.  , -0.  ]])
```

```
Eigen Vectors:
 [[ 0.16  0.17  0.17 ...  0.13  0.15  0.13]
 [-0.13 -0.09 -0.1  ...  0.05 -0.07 -0.07]
 [-0.    0.06  0.04 ... -0.08  0.11  0.1 ]
 ...
 [ 0.    0.21  0.25 ... -0.07  0.   -0.07]
 [ 0.    0.29 -0.21 ...  0.04 -0.03  0.01]
 [-0.    0.19  0.03 ... -0.03 -0.14 -0.02]]
```

```
Eigen Values:
 [3.181e+01 7.870e+00 4.150e+00 3.670e+00 2.210e+00 1.940e+00 1.180e+00
 7.500e-01 6.200e-01 5.300e-01 4.300e-01 3.500e-01 3.000e-01 2.800e-01
 1.900e-01 1.400e-01 1.100e-01 1.100e-01 1.000e-01 8.000e-02 6.000e-02
 4.000e-02 4.000e-02 3.000e-02 3.000e-02 2.000e-02 1.000e-02 1.000e-02
 1.000e-02 1.000e-02 1.000e-02 1.000e-02 0.000e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 0.000e+00]
```

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

Obtaining % of variability explained by each PC

explained variance = (eigen value of each pc)/(sum of eigen values of all pc's)

```
[0.557 0.138 0.073 0.064 0.039 0.034 0.021 0.013 0.011 0.009 0.008 0.006
 0.005 0.005 0.003 0.002 0.002 0.002 0.002 0.001 0.001 0.001 0.001 0.001
 0.    0.    0.    0.    0.    0.    0.    0.    0.    0.    0.    0.
 0.    0.    0.    0.    0.    0.    0.    0.    0.    0.    0.    0.
 0.    0.    0.    0.    0.    0.    0.    0.    0.   ]
```

In %

```
[56. 14.  7.  6.  4.  3.  2.  1.  1.  1.  1.  1.  1.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.]
```

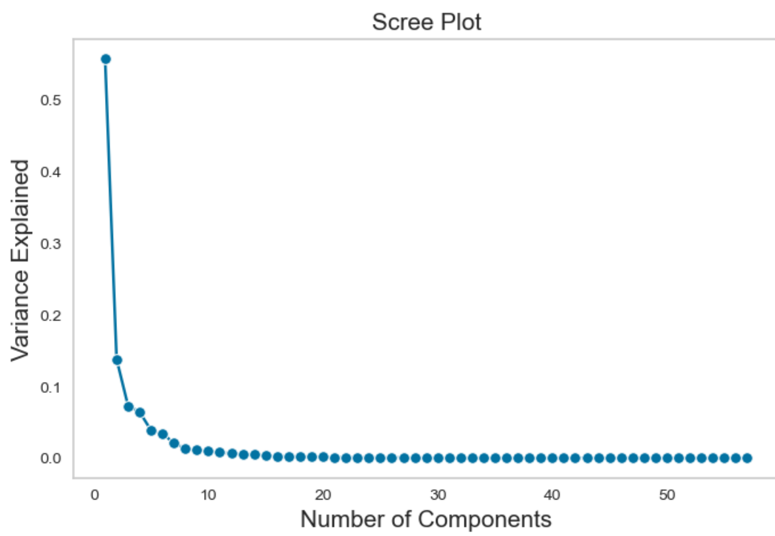
Obtaining the Cumulative Sum of the Explained Variance

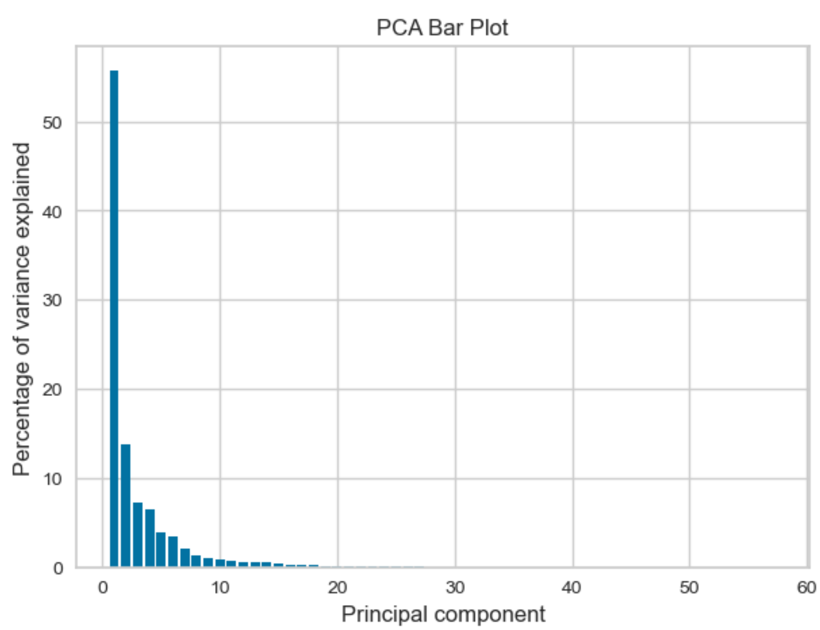
```
Cumulative Variance Explained in Percentage:
 [ 55.73  69.51  76.79  83.21  87.08  90.47  92.53  93.85  94.93  95.85
  96.61  97.23  97.75  98.24  98.57  98.81  99.01  99.2   99.37  99.51
  99.61  99.69  99.75  99.81  99.85  99.89  99.92  99.94  99.96  99.97
  99.98  99.99 100.   100.   100.   100.   100.   100.   100.   100.
 100.   100.   100.   100.   100.   100.   100.   100.   100.   100.
 100.   100.   100.   100.   100.   100.   100.  ]
```

While adding the variance of PCs, and stopping at 90% variance, 6 PCs contribute 90% variance

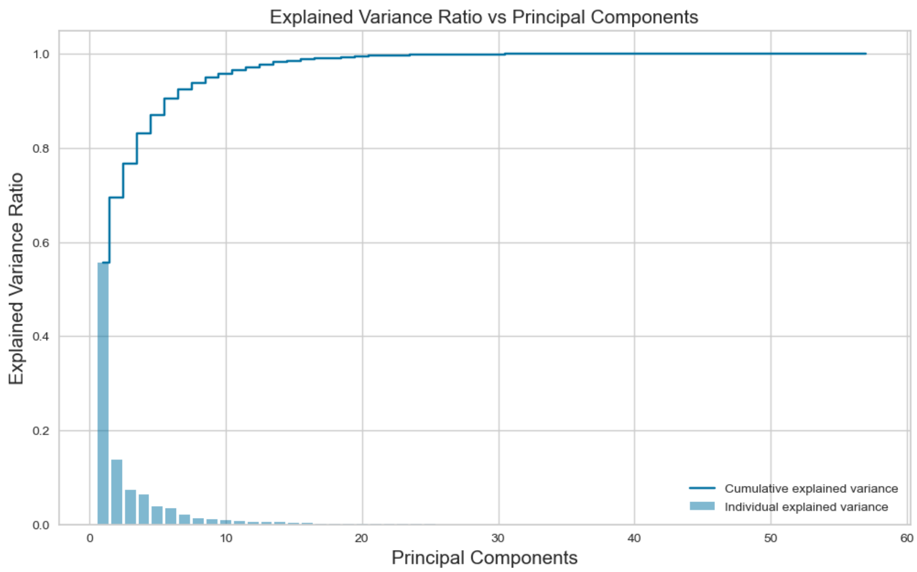
Hence, 90% of variability is explained by 6 PCs.

Scree Plot to identify the number of components to be built.





Plotting Cumulative explained variance and individual explained variance vs Principal Components



We can also find the least number of components that can explain more than 90% variance using enumerate function which has provided below result-

```
Number of PCs that explain at least 90% variance:  6
```

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.

We have 6 set of eigen vectors and each vector has 57 coefficients.

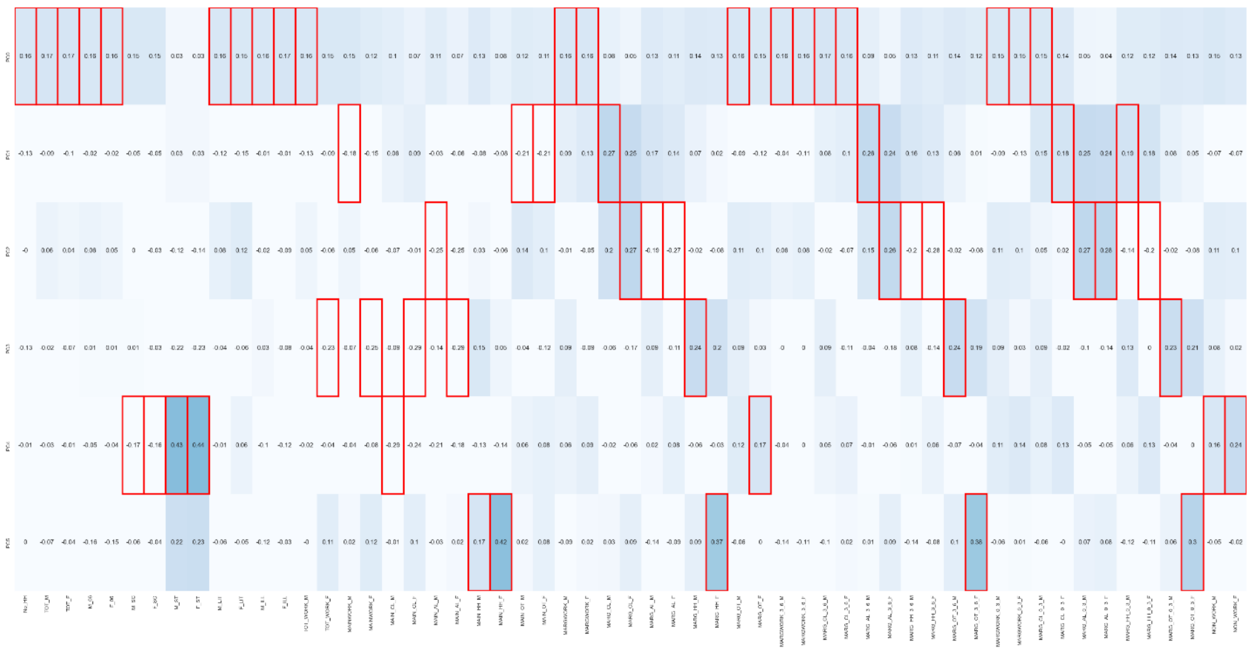
The below are the eigen vectors-

	PC1	PC2	PC3	PC4	PC5	PC6
No_HH	0.16	-0.13	-0.00	-0.13	-0.01	0.00
TOT_M	0.17	-0.09	0.06	-0.02	-0.03	-0.07
TOT_F	0.17	-0.10	0.04	-0.07	-0.01	-0.04
M_06	0.16	-0.02	0.06	0.01	-0.05	-0.16
F_06	0.16	-0.02	0.05	0.01	-0.04	-0.15
M_SC	0.15	-0.05	0.00	0.01	-0.17	-0.06
F_SC	0.15	-0.05	-0.03	-0.03	-0.16	-0.04
M_ST	0.03	0.03	-0.12	-0.22	0.43	0.22
F_ST	0.03	0.03	-0.14	-0.23	0.44	0.23
M_LIT	0.16	-0.12	0.08	-0.04	-0.01	-0.06
F_LIT	0.15	-0.15	0.12	-0.06	0.06	-0.05
M_ILL	0.16	-0.01	-0.02	0.03	-0.10	-0.12
F_ILL	0.17	-0.01	-0.09	-0.08	-0.12	-0.03
TOT_WORK_M	0.16	-0.13	0.05	-0.04	-0.02	-0.00
TOT_WORK_F	0.15	-0.09	-0.06	-0.23	-0.04	0.11
MAINWORK_M	0.15	-0.18	0.05	-0.07	-0.04	0.02
MAINWORK_F	0.12	-0.15	-0.06	-0.25	-0.08	0.12
MAIN_CL_M	0.10	0.06	-0.07	-0.09	-0.29	-0.01
MAIN_CL_F	0.07	0.09	-0.01	-0.29	-0.24	0.10
MAIN_AL_M	0.11	-0.03	-0.25	-0.14	-0.21	-0.03
MAIN_AL_F	0.07	-0.06	-0.25	-0.29	-0.18	0.02
MAIN_HH_M	0.13	-0.08	0.03	0.15	-0.13	0.17
MAIN_HH_F	0.08	-0.08	-0.06	0.05	-0.14	0.42
MAIN_OT_M	0.12	-0.21	0.14	-0.04	0.06	0.02
MAIN_OT_F	0.11	-0.21	0.10	-0.12	0.08	0.08
MARGWORK_M	0.16	0.09	-0.01	0.09	0.06	-0.09
MARGWORK_F	0.16	0.13	-0.05	-0.09	0.09	0.02
MARG_CL_M	0.08	0.27	0.20	-0.06	-0.02	0.03
MARG_CL_F	0.05	0.25	0.27	-0.17	-0.06	0.09
MARG_AL_M	0.13	0.17	-0.19	0.09	0.02	-0.14
MARG_AL_F	0.11	0.14	-0.27	-0.11	0.08	-0.09
MARG_HH_M	0.14	0.07	-0.02	0.24	-0.06	0.09
MARG_HH_F	0.13	0.02	-0.08	0.20	-0.03	0.37
MARG_OT_M	0.16	-0.09	0.11	0.09	0.12	-0.06
MARG_OT_F	0.15	-0.12	0.10	0.03	0.17	0.00
MARGWORK_3_6_M	0.16	-0.04	0.06	-0.00	-0.04	-0.14
MARGWORK_3_6_F	0.16	-0.11	0.08	0.00	0.00	-0.11
MARG_CL_3_6_M	0.17	0.08	-0.02	0.09	0.05	-0.10
MARG_CL_3_6_F	0.16	0.10	-0.07	-0.11	0.07	0.02
MARG_AL_3_6_M	0.09	0.26	0.15	-0.04	-0.01	0.01
MARG_AL_3_6_F	0.05	0.24	0.26	-0.18	-0.06	0.09
MARG_HH_3_6_M	0.13	0.16	-0.20	0.08	0.01	-0.14
MARG_HH_3_6_F	0.11	0.13	-0.28	-0.14	0.06	-0.08
MARG_OT_3_6_M	0.14	0.06	-0.02	0.24	-0.07	0.10
MARG_OT_3_6_F	0.12	0.01	-0.08	0.19	-0.04	0.38
MARGWORK_0_3_M	0.15	-0.09	0.11	0.09	0.11	-0.06
MARGWORK_0_3_F	0.15	-0.13	0.10	0.03	0.14	0.01
MARG_CL_0_3_M	0.15	0.15	0.05	0.09	0.08	-0.06
MARG_CL_0_3_F	0.14	0.18	0.02	-0.02	0.13	-0.00
MARG_AL_0_3_M	0.05	0.25	0.27	-0.10	-0.05	0.07
MARG_AL_0_3_F	0.04	0.24	0.28	-0.14	-0.05	0.08
MARG_HH_0_3_M	0.12	0.19	-0.14	0.13	0.06	-0.12
MARG_HH_0_3_F	0.12	0.18	-0.20	0.00	0.13	-0.11
MARG_OT_0_3_M	0.14	0.08	-0.02	0.23	-0.04	0.06
MARG_OT_0_3_F	0.13	0.05	-0.08	0.21	0.00	0.30
NON_WORK_M	0.15	-0.07	0.11	0.08	0.16	-0.05
NON_WORK_F	0.13	-0.07	0.10	0.02	0.24	-0.02

To identify which features have maximum loading across the components, we will first plot the component loading on a heatmap.

For each feature, we have found the maximum loading value across the components and marked the same with help of rectangular box.

Features marked with rectangular red box are the one having maximum loading on the respective component. We will consider these marked features to decide the context that the component represents.



PC2 provides more information about the Marginal Cultivator population, Marginal Agriculture Labourers.

PC3 on Marginal Agriculture Labourers, Marginal Cultivator Population.

PC4 on Marginal Household Industries Population, Marginal Other Workers Population, Marginal Household Industries.

PC5 on Scheduled Tribes population, Non Working Population.

Highly coefficient columns are Scheduled Tribes population Male, Scheduled Tribes population Female, Main Household Industries Population Female, Marginal Household Industries Population Female, Marginal Other Workers Population Person 3-6 Female, Marginal Other Workers Population 0-3 Female which are very important for our data analysis.

2.8 Write linear equation for first PC.

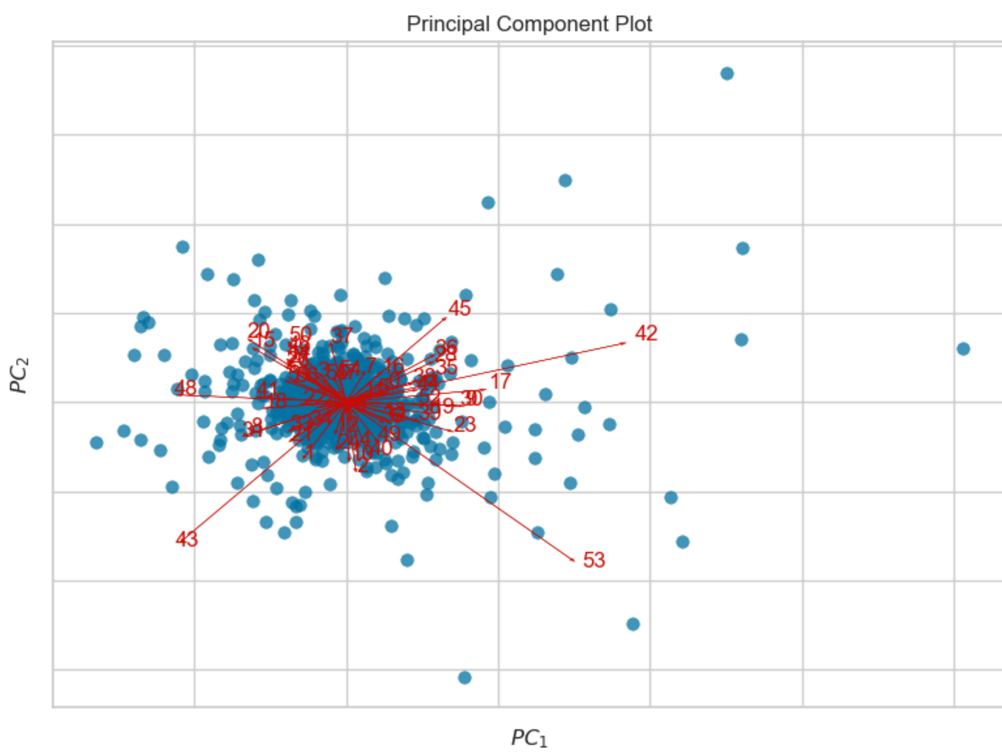
Liner equation for 1st PC:

PC1= a1x1 + a2x2 + a3x3 + a4x4 + a5x5 + a6x6

a1...- coefficients/ eigen vectors

x1... - original data

```
The Linear equation of 1st component:
0.16 * No_HH + 0.17 * TOT_M + 0.17 * TOT_F + 0.16 * M_06 + 0.16 * F_06 + 0.15 * M_SC + 0.15 * F_SC + 0.03 * M_ST + 0.03 * F_S
T + 0.16 * M_LIT + 0.15 * F_LIT + 0.16 * M_ILL + 0.17 * F_ILL + 0.16 * TOT_WORK_M + 0.15 * TOT_WORK_F + 0.15 * MAINWORK_M +
0.12 * MAINWORK_F + 0.1 * MAIN_CL_M + 0.07 * MAIN_CL_F + 0.11 * MAIN_AL_M + 0.07 * MAIN_AL_F + 0.13 * MAIN_HH_M + 0.08 * MAIN
_HH_F + 0.12 * MAIN_OT_M + 0.11 * MAIN_OT_F + 0.16 * MARGWORK_M + 0.16 * MARGWORK_F + 0.08 * MARG_CL_M + 0.05 * MARG_CL_F +
0.13 * MARG_AL_M + 0.11 * MARG_AL_F + 0.14 * MARG_HH_M + 0.13 * MARG_HH_F + 0.16 * MARG_OT_M + 0.15 * MARG_OT_F + 0.16 * MARG
WORK_3_6_M + 0.16 * MARGWORK_3_6_F + 0.17 * MARG_CL_3_6_M + 0.16 * MARG_CL_3_6_F + 0.09 * MARG_AL_3_6_M + 0.05 * MARG_AL_3_6_
F + 0.13 * MARG_HH_3_6_M + 0.11 * MARG_HH_3_6_F + 0.14 * MARG_OT_3_6_M + 0.12 * MARG_OT_3_6_F + 0.15 * MARGWORK_0_3_M + 0.15
* MARGWORK_0_3_F + 0.15 * MARG_CL_0_3_M + 0.14 * MARG_CL_0_3_F + 0.05 * MARG_AL_0_3_M + 0.04 * MARG_AL_0_3_F + 0.12 * MARG_HH
_0_3_M + 0.12 * MARG_HH_0_3_F + 0.14 * MARG_OT_0_3_M + 0.13 * MARG_OT_0_3_F + 0.15 * NON_WORK_M + 0.13 * NON_WORK_F +
```



Thankyou!
The end.