Clustering Clean Ads

(2022-2023)

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Course Name

Post Graduate Program in
Data Science and Business Analytics

Batch Id

(PGP-DSBA-June22C)

Submitted by

Jayant Singh

Clustering

1.1 Question: *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: Few rows (head & tail)

	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	2020-9-2-18	Format1	300	250	75000	Inter223	Web	Mobile	Display	1979	384	380	0	0.0
2	2020-9-3-16	Format6	336	250	84000	Inter217	Web	Desktop	Video	1566	298	297	0	0.0
3	2020-9-3-2	Format1	300	250	75000	Inter224	Web	Desktop	Display	643	103	102	0	0.0
4	2020-9-3-13	Format1	300	250	75000	Inter225	Video	Mobile	Display	1550	347	345	0	0.0

	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks
25852	2020-10-1-5	Format5	720	300	216000	Inter222	Video	Desktop	Video	1	1	1	0
25853	2020-11-18-2	Format4	120	600	72000	inter230	Video	Mobile	Video	7	1	1	1
25854	2020-9-14-0	Format5	720	300	216000	Inter221	App	Mobile	Video	2	2	2	1
25855	2020-9-30-4	Format7	300	600	180000	Inter228	Video	Mobile	Display	1	1	1	0
25856	2020-10-17-3	Format5	720	300	216000	Inter225	Video	Mobile	Display	1	1	1	0

Info

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25857 entries, 0 to 25856
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Timestamp             25857 non-null  object
 1   InventoryType         25857 non-null  object
 2   Ad - Length           25857 non-null  int64
 3   Ad- Width             25857 non-null  int64
 4   Ad Size               25857 non-null  int64
 5   Ad Type               25857 non-null  object
 6   Platform              25857 non-null  object
 7   Device Type           25857 non-null  object
 8   Format                25857 non-null  object
 9   Available_Impressions 25857 non-null  int64
 10  Matched_Queries       25857 non-null  int64
 11  Impressions           25857 non-null  int64
 12  Clicks                25857 non-null  int64
 13  Spend                 25857 non-null  float64
 14  Fee                   25857 non-null  float64
 15  Revenue               25857 non-null  float64
 16  CTR                   19392 non-null  float64
 17  CPM                   19392 non-null  float64
 18  CPC                   18330 non-null  float64
dtypes: float64(6), int64(7), object(6)
memory usage: 3.7+ MB
```

Data Summary

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Rev
count	25857.000000	25857.000000	25857.000000	2.585700e+04	2.585700e+04	2.585700e+04	25857.000000	25857.000000	25857.000000	25857.00
mean	390.431218	332.182774	99683.276482	2.169621e+06	1.155322e+06	1.107525e+06	9525.881386	2414.473115	0.336729	1716.54
std	230.696051	194.260924	62640.685612	4.542680e+06	2.407244e+06	2.326648e+06	16721.686071	3932.835240	0.030540	2993.02
min	120.000000	70.000000	33600.000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000	0.000000	0.210000	0.00
25%	120.000000	250.000000	72000.000000	9.133000e+03	5.451000e+03	2.558000e+03	305.000000	36.030000	0.350000	23.42
50%	300.000000	300.000000	75000.000000	3.309680e+05	1.894490e+05	1.621620e+05	3457.000000	1173.660000	0.350000	762.88
75%	720.000000	600.000000	84000.000000	2.208484e+06	1.008171e+06	9.496930e+05	10681.000000	2692.280000	0.350000	1749.98
max	728.000000	600.000000	216000.000000	2.759286e+07	1.470202e+07	1.419477e+07	143049.000000	26931.870000	0.350000	21276.18

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                   6465
CPM                   6465
CPC                   7527
dtype: int64
```

Duplicate Value Check

Null Duplicate Value found

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

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

```
2414473.1148238387
```

CPC = Total Cost (spend) / Number of Clicks

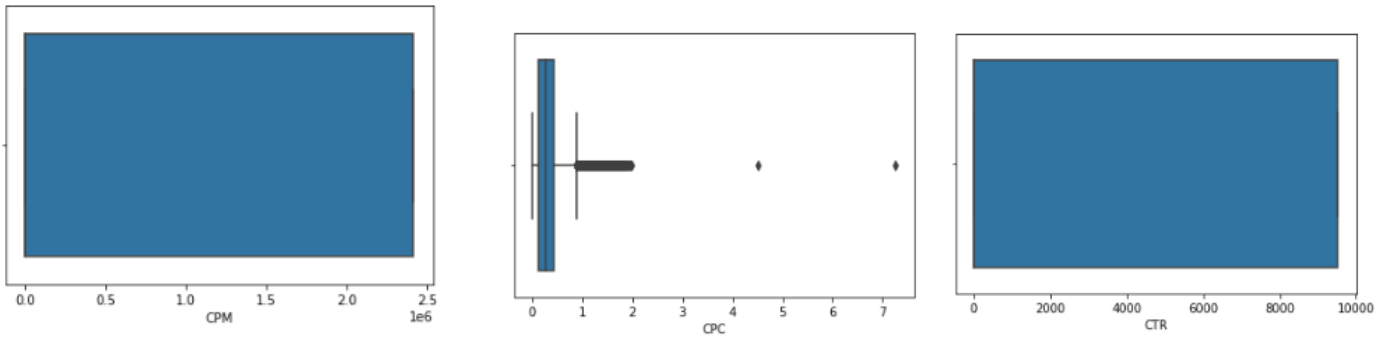
```
0.25346453697720783
```

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

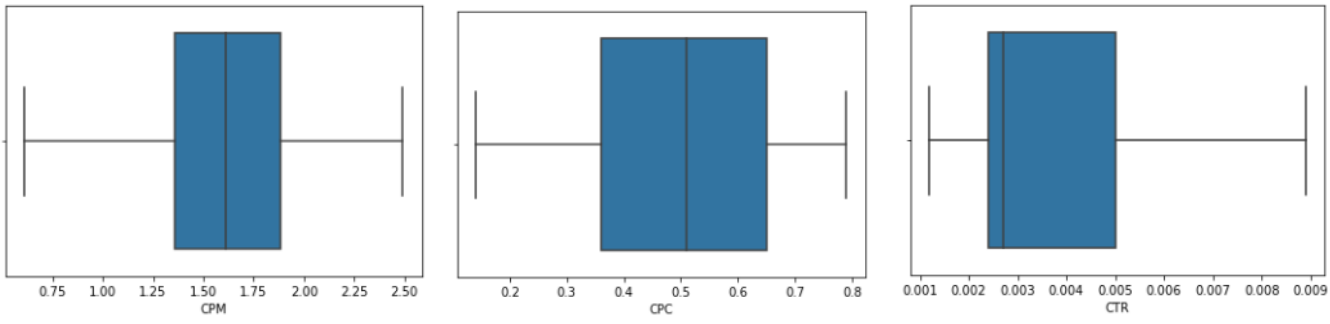
```
9525.881386085006
```

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. (As an analyst your judgement may be different from another analyst).

Yes, i think outlier treatment is necessary for K-Means as it performs on eucledian distance based algorithm. I'll use seaborn boxplot to visually detected the outlier and by slicing the array we detected the outlier using IQR based treatment.



Outlier Removal (We will Not apply outlier removal technique as there are lots of them instead, we will use standard scaler to reduce their effect)



Part 1.4

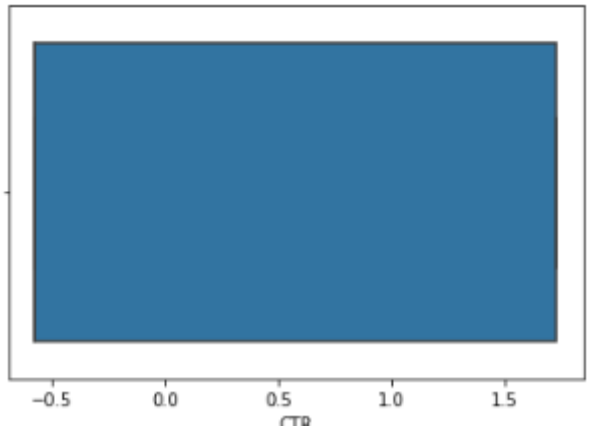
Clustering: Perform z-score scaling and discuss how it affects the speed of the algorithm.

Applying z-score scaling

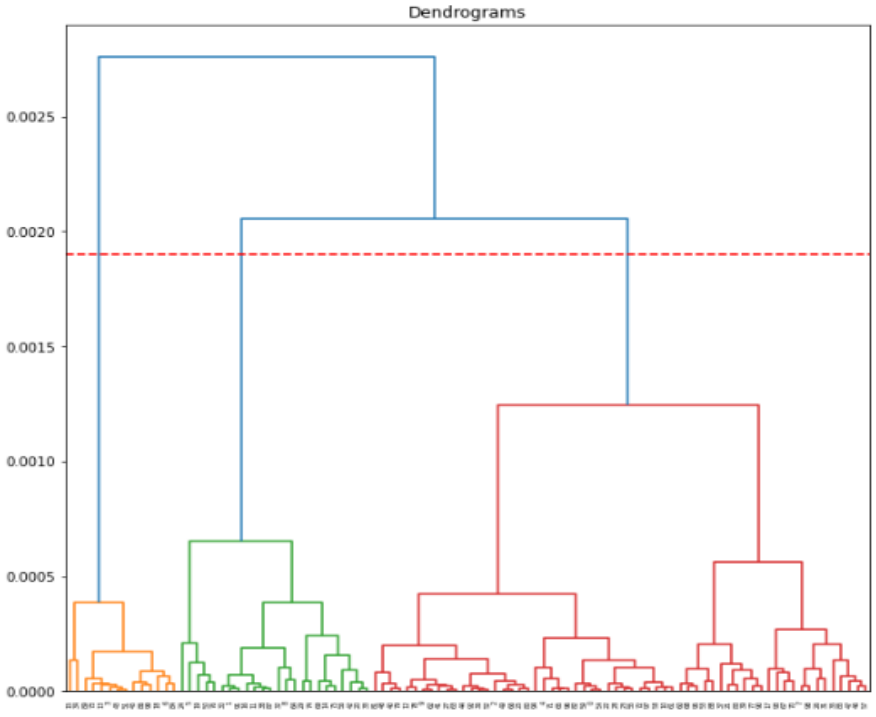
	Timestamp	Ad - Length	Ad-Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	...	Inter220	Inter221
0	-1.189701	-0.392000	-0.423062	-0.394053	-0.477220	-0.479810	-0.475888	-0.569624	-0.613939	0.434559	...	-0.277437	-0.277437
1	-1.189701	-0.392000	-0.423062	-0.394053	-0.477182	-0.479785	-0.475864	-0.569683	-0.613939	0.434559	...	-0.277437	-0.277437
2	-1.189701	-0.235948	-0.423062	-0.250374	-0.477273	-0.479821	-0.475899	-0.569683	-0.613939	0.434559	...	-0.277437	-0.277437
3	-1.189701	-0.392000	-0.423062	-0.394053	-0.477476	-0.479902	-0.475983	-0.569683	-0.613939	0.434559	...	-0.277437	-0.277437
4	-1.189701	-0.392000	-0.423062	-0.394053	-0.477276	-0.479801	-0.475879	-0.569683	-0.613939	0.434559	...	-0.277437	-0.277437

5 rows × 37 columns

Checking the effect of outlier after applying Standard scaler

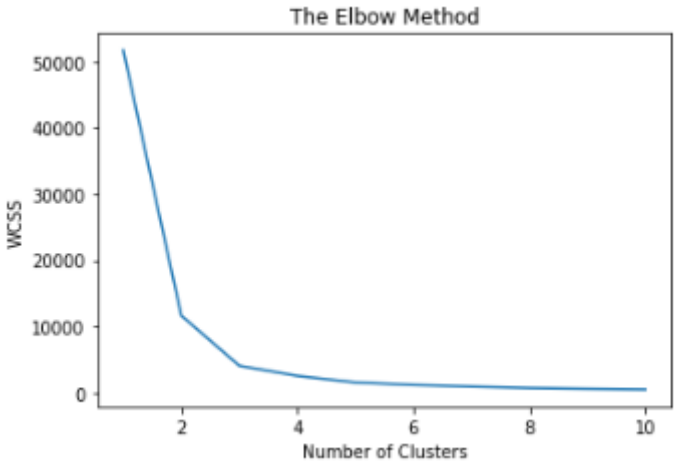


Part 1.5 - Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.



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

Using the elbow method to find the optimal number of clusters

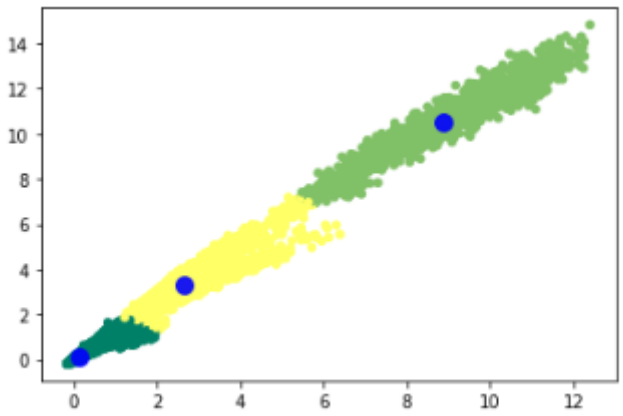


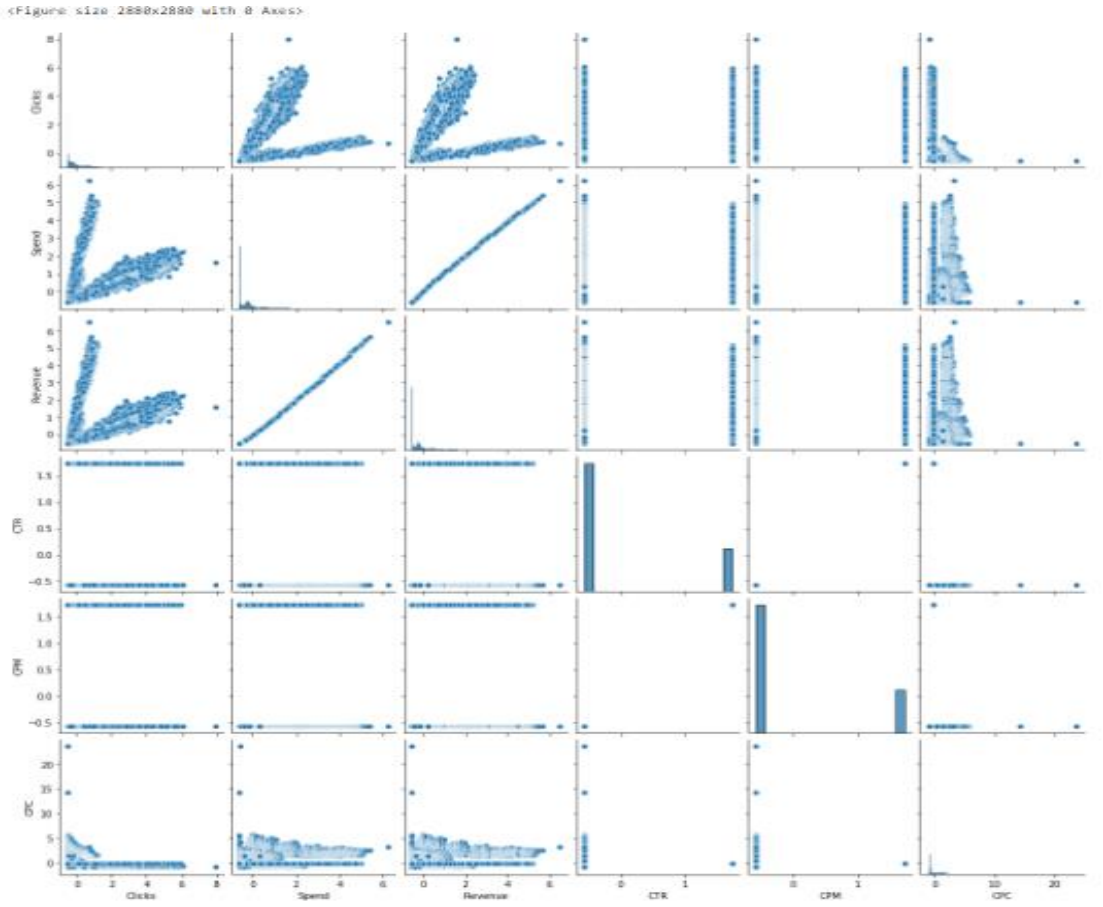
Part 1.7 - Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters

```
[0.6837816953500804]
```

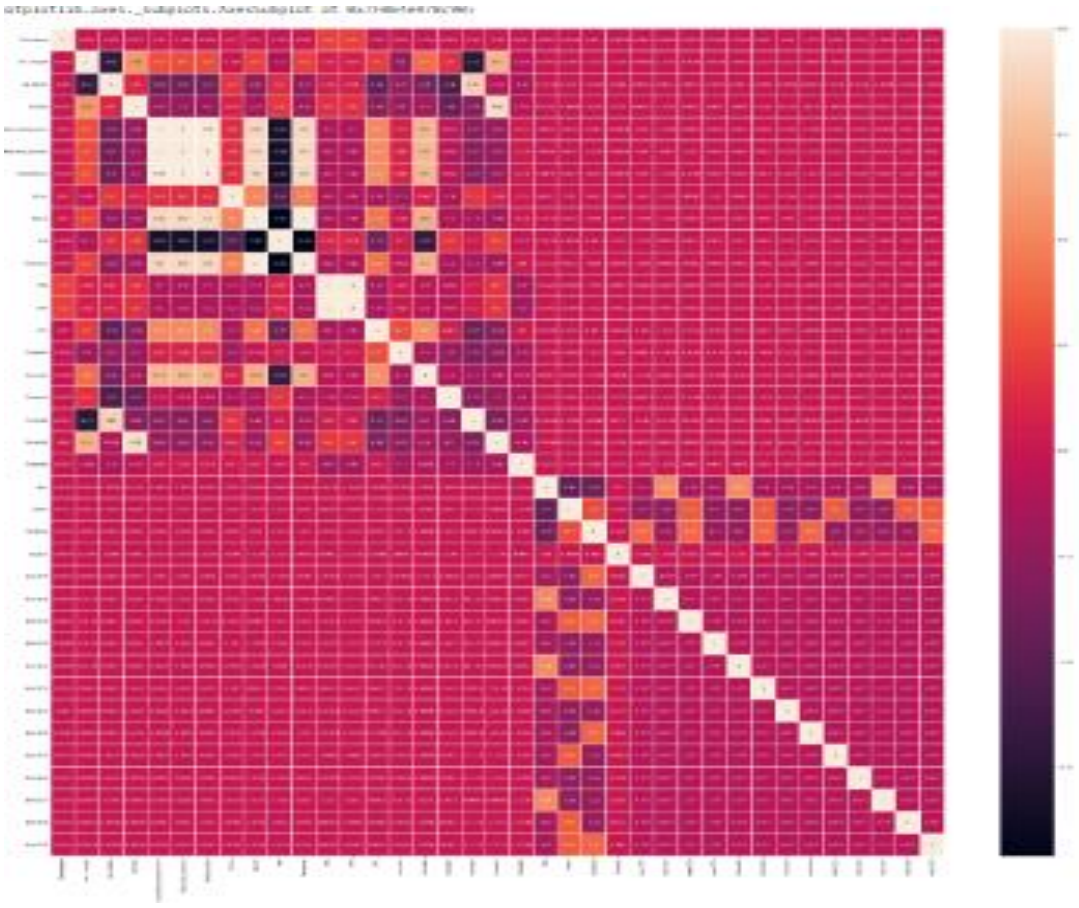
Part 1.8 - Clustering: Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots]

Bar Plot





Heatmap to check correlations between features



1.9 Conclude the project by providing summary of your learnings

Part 2

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

Head : 5 rows, 61 Columns

	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_
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180	
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123	
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	...	114	188	44	
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	...	194	247	61	
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	...	874	1928	465	

5 rows × 61 columns

Info

```
<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
```

Summary:

	State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MAI
count	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	...	
std	9.426486	184.896367	48135.405475	73384.511114	113600.717282	11500.906881	11326.294567	14426.373130	21727.887713	9912.668948	...	
min	1.000000	1.000000	350.000000	391.000000	698.000000	56.000000	56.000000	0.000000	0.000000	0.000000	...	
25%	9.000000	160.750000	19484.000000	30228.000000	46517.750000	4733.750000	4672.250000	3466.250000	5603.250000	293.750000	...	
50%	18.000000	320.500000	35837.000000	58339.000000	87724.500000	9159.000000	8663.000000	9591.500000	13709.000000	2333.500000	...	
75%	24.000000	480.250000	68892.000000	107918.500000	164251.750000	16520.250000	15902.250000	19429.750000	29180.000000	7658.000000	...	
max	35.000000	640.000000	310450.000000	485417.000000	750392.000000	96223.000000	95129.000000	103307.000000	156429.000000	96785.000000	...	

8 rows × 59 columns

Null Values

```
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
```

Duplicate Value Check

0

Part 2.2 - PCA: Perform detailed Exploratory analysis by creating certain questions like (i) Which state has highest gender ratio and which has the lowest?

(ii) Which district has the highest & lowest gender ratio? (Example Questions).

Pick 5 variables out of the given 24 variables below for EDA: No_HH, TOT_M, TOT_F, M_06, F_06, M_SC, F_SC, M_ST, F_ST, M_LIT, F_LIT, M_ILL, F_ILL, TOT_WORK_M, TOT_WORK_F, MAINWORK_M, MAINWORK_F, MAIN_CL_M, MAIN_CL_F, MAIN_AL_M, MAIN_AL_F, MAIN_HH_M, MAIN_HH_F, MAIN_OT_M, MAIN_OT_F

(i) Highest Gender ratio from the code below is of `Lakshadweep`

```
State Code      35.000000
Gender ratio     0.868061
dtype: float64
```

Lowest Gender ratio from the code below is of Andhra Pradesh

```
State Code      1.000000
Gender ratio    0.437972
dtype: float64
```

Which district has the highest gender ratio

District Number: 587

Which district has the Lowest gender ratio

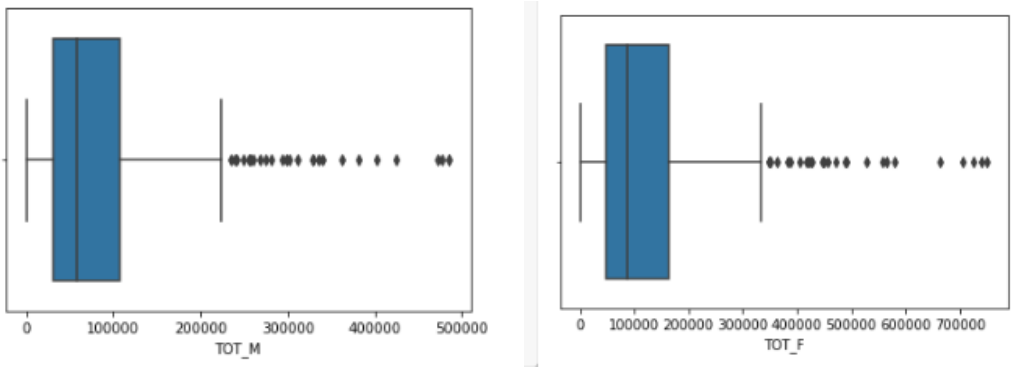
District Number: 547

Pick 5 variables out of the given 24 variables below for EDA

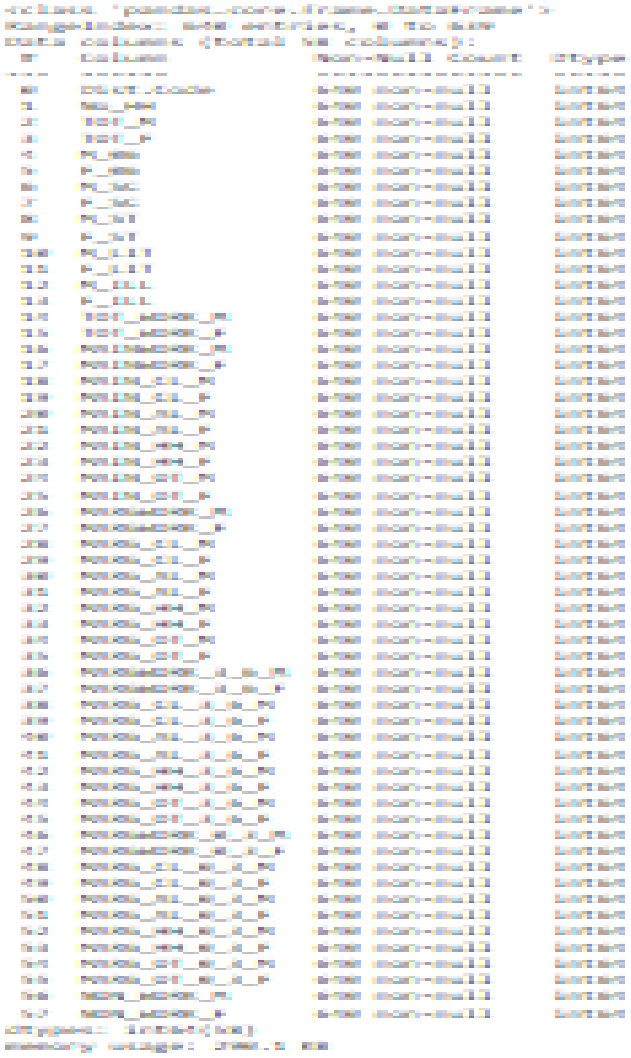
	TOT_M	TOT_F	M_LIT	MARGWORK_3_6_F	TOT_WORK_M	State
0	23388	29796	13381	26044	6723	Jammu & Kashmir
1	19585	23102	10513	18902	6982	Jammu & Kashmir
2	6546	10964	4534	6164	2775	Jammu & Kashmir
3	2784	4206	1842	3088	1002	Jammu & Kashmir
4	20591	29981	13243	22289	5717	Jammu & Kashmir

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

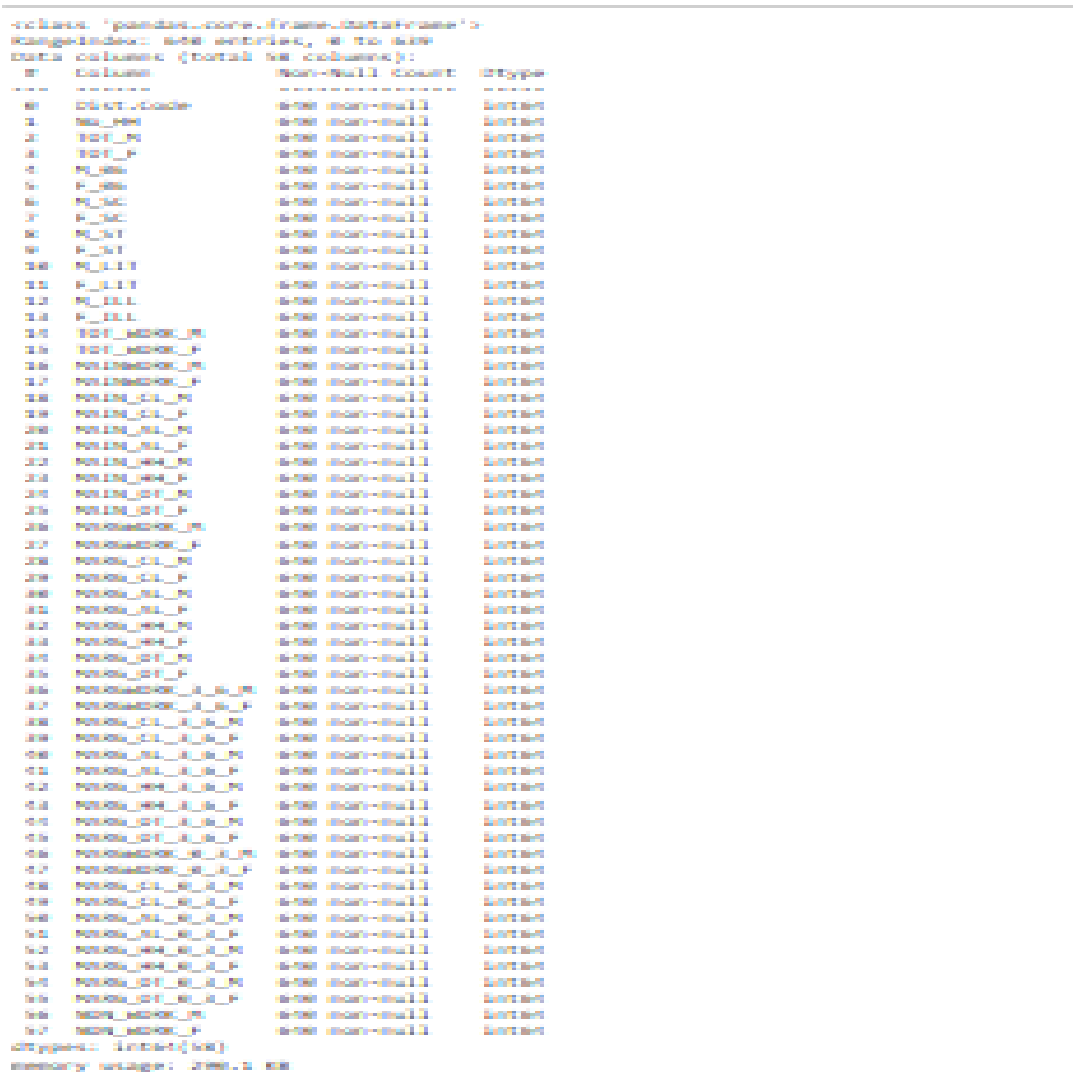
Yes I think it is necessary to treat outlier in PCA as it is sensitive to outliers. In our scenario we can see from boxplots below that feature does contain a lot of outliers that need to be treated.



Dropping State Variables

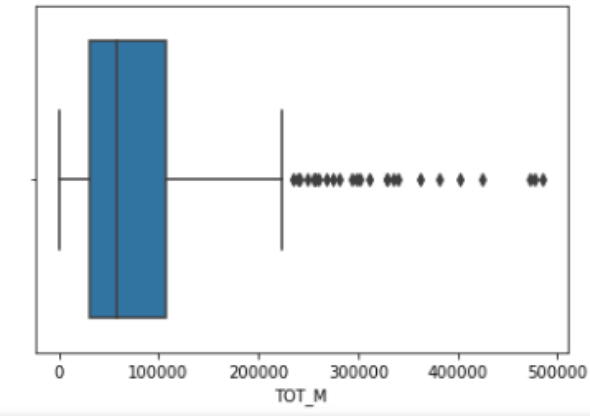


Dropping Area Name

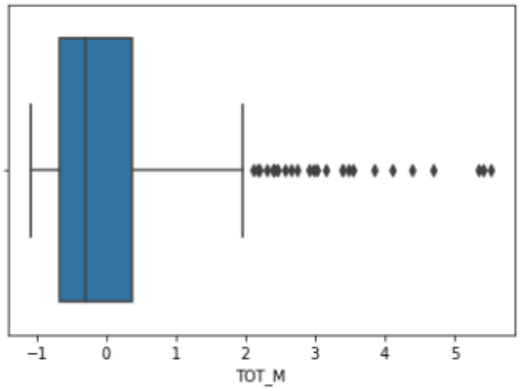


Part 2.4 - PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.

Before scaling



After scaling



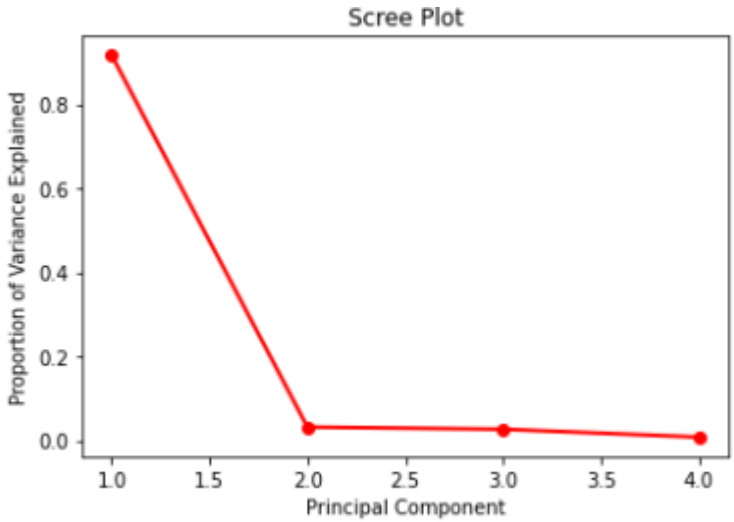
As we can see after scaling the variance has been decreased due to which the outliers are closely located to other data points thus reducing their effect.

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

The eigen values and eigen vectors are :

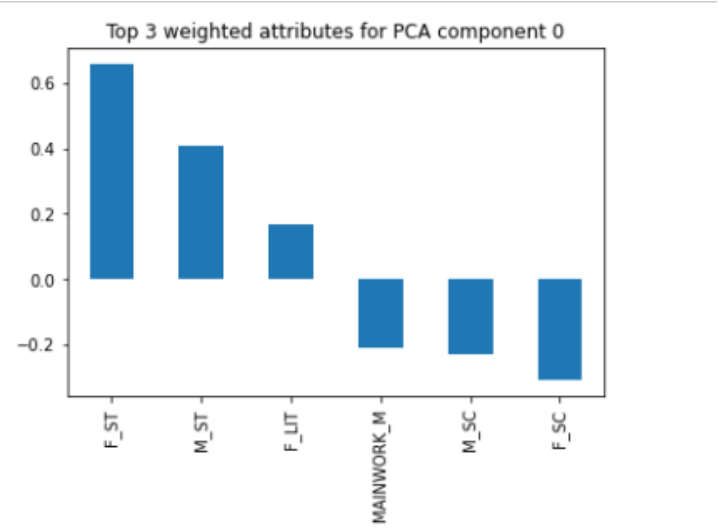
```
[4.40513827e+10 1.51948727e+09 3.67892434e+08
1.26894306e+09] [[ 1.00000000e+00 -8.98306322e-18  1.09328119e-18 -2.69434504e-33]
 [ 0.00000000e+00 -1.00000000e+00 -5.85873983e-17 -2.91146100e-16]
 [ 0.00000000e+00 -5.50824873e-17 -4.40659898e-16  1.00000000e+00]
 [ 0.00000000e+00 -1.26099935e-16  1.00000000e+00 -8.44548393e-17]]
```

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



The graph tells the first principal component captures about 95% of the information present in the original mathematical space i.e. explains ~ 95% of the variation in the data. On taking the first two components together, the total explained variation is close to 98% and when taking the first three components, then the cumulative explained variation is ~ 98% and ~ 99% of the variation is captured by the first four principal components.

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




```
F_ST           0.66
M_ST           0.41
F_LIT          0.17
MAINWORK_M    -0.21
M_SC          -0.23
F_SC          -0.31
Name: 3, dtype: float64
```

Part 2.7 - PCA: Write linear equation for first PC

Let us assume the base measures ‘x’ meter. Hence, each of the legs measure y = (x + 4) meters.

The Perimeter of a triangle is the sum of the three sides. The equations are formed and solved as follows:

Solve for x, 2x – 4 = 0

Solution:

Add 4 both sides

2x – 4 + 4 = 0 + 4

2x = 4

Divide each side by 2, we get

2x/2 = 4/2

x = 4/2 = 2

So, x = 2 is the answer.