DATA MINING PROJECT REPORT

DSBA

I. Contents

Problem Statement 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.
The Data Dictionary and the detailed description of the formulas for CPM, CPC and CTR are given in the sheet 2 of the Clustering clean ads_data Excel file.

Part1- Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.

The dataset that is provided is Clustering Clean_ads is being loaded in the dataframe.

- Printing the first 5 rows of the data-frame .

Note- (Two screenshots are there because in single screenshots all columns are not visible)

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

1 to 5 of 5 entries Filter

- Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0	0.35	0.0	0.0031	0.0	0.0
300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.0	0.35	0.0	0.0035	0.0	0.0
300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0	0.35	0.0	0.0028	0.0	0.0
300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.0	0.35	0.0	0.002	0.0	0.0
300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.0	0.35	0.0	0.0041	0.0	0.0

Show 25 per page

- Printing the last 5 rows of the data-frame .

Note- (Two screenshots are there because in single screenshots all columns are not visible)

	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks
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

Ad- dth	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
300	216000	Inter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
300	216000	Inter224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.0260	NaN	NaN	NaN
300	216000	Inter218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	NaN	NaN	NaN
600	72000	inter230	Video	Mobile	Video	7	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
300	216000	Inter221	App	Mobile	Video	2	2	2	1	0.09	0.35	0.0585	NaN	NaN	NaN

- Information of the dataset provided:

```
<bound method DataFrame.info of          Timestamp InventoryType  Ad - Length  Ad- Width  Ad Size   Ad Type  \
0       2020-9-2-17       Format1          300        250    75000  Inter222
1       2020-9-2-10       Format1          300        250    75000  Inter227
2       2020-9-1-22       Format1          300        250    75000  Inter222
3       2020-9-3-20       Format1          300        250    75000  Inter228
4       2020-9-4-15       Format1          300        250    75000  Inter217
...             ...           ...          ...        ...      ...       ...
23061   2020-9-13-7       Format5          720        300   216000  Inter220
23062   2020-11-2-7       Format5          720        300   216000  Inter224
23063  2020-9-14-22       Format5          720        300   216000  Inter218
23064  2020-11-18-2       Format4          120        600    72000  inter230
23065   2020-9-14-0       Format5          720        300   216000  Inter221

      Platform Device Type   Format  Available_Impressions  Matched_Queries  \
0        Video     Desktop  Display                   1806              325
1          App      Mobile    Video                   1780              285
2        Video     Desktop  Display                   2727              356
3        Video      Mobile    Video                   2430              497
4          Web     Desktop    Video                   1218              242
...        ...         ...      ...                    ...              ...
23061      Web      Mobile    Video                      1                1
23062      Web     Desktop    Video                      3                2
23063      App      Mobile    Video                      2                1
23064    Video      Mobile    Video                      7                1
23065      App      Mobile    Video                      2                2
```

```
        Impressions  Clicks  Spend   Fee  Revenue     CTR  CPM  CPC
0               323       1   0.00  0.35   0.0000  0.0031  0.0  0.0
1               285       1   0.00  0.35   0.0000  0.0035  0.0  0.0
2               355       1   0.00  0.35   0.0000  0.0028  0.0  0.0
3               495       1   0.00  0.35   0.0000  0.0020  0.0  0.0
4               242       1   0.00  0.35   0.0000  0.0041  0.0  0.0
...             ...     ...    ...   ...      ...     ...  ...  ...
23061             1       1   0.07  0.35   0.0455     NaN  NaN  NaN
23062             2       1   0.04  0.35   0.0260     NaN  NaN  NaN
23063             1       1   0.05  0.35   0.0325     NaN  NaN  NaN
23064             1       1   0.07  0.35   0.0455     NaN  NaN  NaN
23065             2       1   0.09  0.35   0.0585     NaN  NaN  NaN

[23066 rows x 19 columns]>
```

- Shape of the dataset(that is rows and columns)

```
(23066, 19)
```

- Describe function to find out mean , std,count etc.

[10]

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	3.851631e+02	2.336514e+02	120.0000	120.000000	300.00000	7.200000e+02	728.00
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.0000	250.000000	300.00000	6.000000e+02	600.00
Ad Size	23066.0	9.667447e+04	6.153833e+04	33600.0000	72000.000000	72000.00000	8.400000e+04	216000.00
Available_Impressions	23066.0	2.432044e+06	4.742888e+06	1.0000	33672.250000	483771.00000	2.527712e+06	27592861.00
Matched_Queries	23066.0	1.295099e+06	2.512970e+06	1.0000	18282.500000	258087.50000	1.180700e+06	14702025.00
Impressions	23066.0	1.241520e+06	2.429400e+06	1.0000	7990.500000	225290.00000	1.112428e+06	14194774.00
Clicks	23066.0	1.067852e+04	1.735341e+04	1.0000	710.000000	4425.00000	1.279375e+04	143049.00
Spend	23066.0	2.706626e+03	4.067927e+03	0.0000	85.180000	1425.12500	3.121400e+03	26931.87
Fee	23066.0	3.351231e-01	3.196322e-02	0.2100	0.330000	0.35000	3.500000e-01	0.35
Revenue	23066.0	1.924252e+03	3.105238e+03	0.0000	55.365375	926.33500	2.091338e+03	21276.18
CTR	18330.0	7.366054e-02	7.515992e-02	0.0001	0.002600	0.08255	1.300000e-01	1.00
CPM	18330.0	7.672045e+00	6.481391e+00	0.0000	1.710000	7.66000	1.251000e+01	81.56
CPC	18330.0	3.510606e-01	3.433338e-01	0.0000	0.090000	0.16000	5.700000e-01	7.26

- To check the duplicate rows/ values present in dataset.
  We found that there is no duplicate

```
0
```

- Finding null values present in dataset.

By below screenshot we can see that there are null values that is present in CTR, CPM,CPC and we need to treat the null values as to move further.

```
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
```

Part 1 - Clustering: Treat missing values in CPC, CTR and CPM using the formula given

We found that there are many missing/null values present in CPC,CTR,CPM as shown in below screenshot.

```
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
```

We need to treat these missing values as we cannot proceed with missing values present in dataset. So we are treating one by one all the 3 parameters null values .

Formula of each of CTR,CPM, CPC is described below so that we can relate how each parameter is calculated.

1. CTR stands for "Click through rate". CTR is the number of clicks that your ad receives divided by the number of times your ad is shown. Formula used here is CTR = Total Measured Clicks / Total Measured Ad Impressions x 100. Note that the Total Measured Clicks refers to the 'Clicks' Column and the Total Measured Ad Impressions refers to the 'Impressions' Column.
2. CPM stands for "cost per 1000 impressions." Formula used here is CPM = (Total Campaign Spend / Number of Impressions) * 1,000. Note that the Total Campaign Spend refers to the 'Spend' Column and the Number of Impressions refers to the 'Impressions' Column.
3. CPC stands for "Cost-per-click". Cost-per-click (CPC) bidding means that you pay for each click on your ads. The Formula used here is CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column and the Number of Clicks refers to the 'Clicks' Column.

After applying those functions shown above to respective columns we get the below output

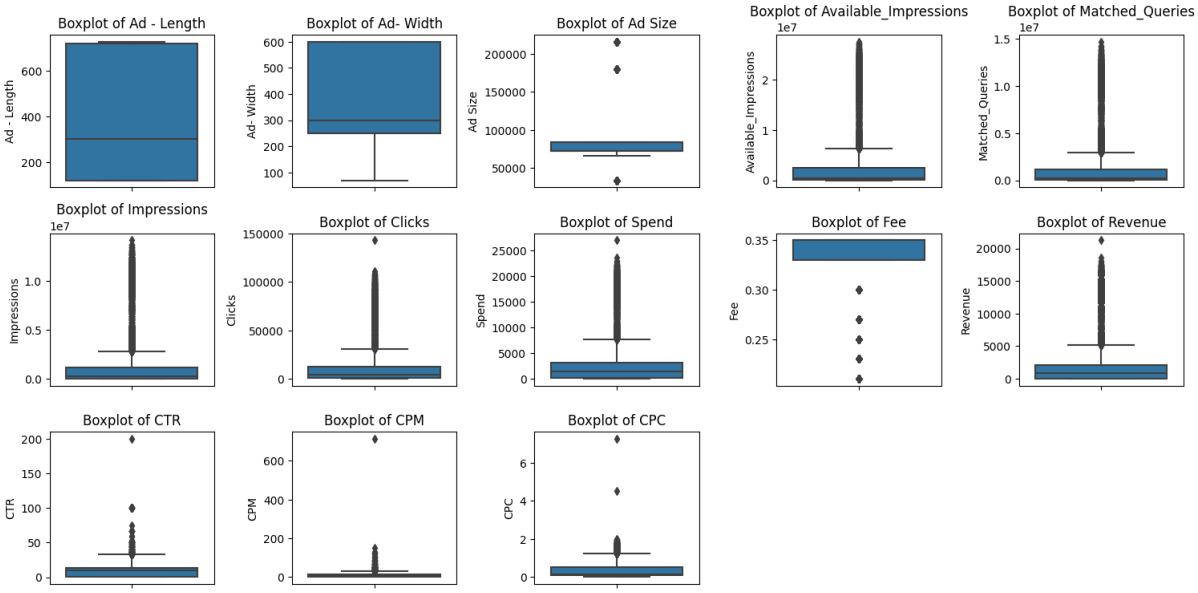
```
Timestamp 0
InventoryType 0
Ad - Length 0
Ad- Width 0
Ad Size 0
Ad Type 0
Platform 0
Device Type 0
Format 0
Available_Impressions 0
Matched_Queries 0
Impressions 0
Clicks 0
Spend 0
Fee 0
Revenue 0
CTR 0
CPM 0
CPC 0
```

All the null values of these three parameters have been treated as shown in above screenshot.

Part 1 - Clustering: Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an analyst your judgement may be different from another analyst).

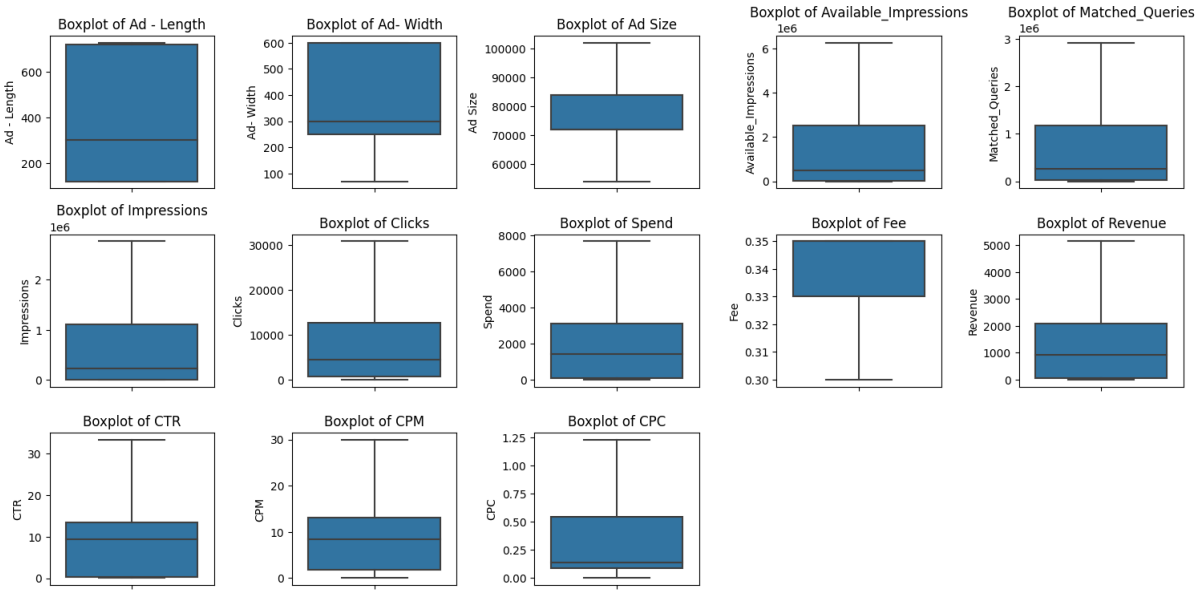
- Check whether there is outliers present in the dataset provided.

We found that there is outliers present in the dataset as shown below-



We decided to treat the outliers that are present in the dataset as Clustering is sensitive to outliers.

After treating the outliers with IQR method we have removed the outliers as shown below



Part 1 - Clustering: Perform z-score scaling and discuss how it affects the speed of the algorithm.

- Dataset that is present before performing the z-score scaling.

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	3.851631e+02	2.336514e+02	120.000000	120.000000	300.000000	7.200000e+02	7.280000e+02
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.000000	250.000000	300.000000	6.000000e+02	6.000000e+02
Ad Size	23066.0	9.667447e+04	6.153833e+04	33600.000000	72000.000000	72000.000000	8.400000e+04	2.160000e+05
Available_Impressions	23066.0	2.432044e+06	4.742888e+06	1.000000	33672.250000	483771.000000	2.527712e+06	2.759286e+07
Matched_Queries	23066.0	1.295099e+06	2.512970e+06	1.000000	18282.500000	258087.500000	1.180700e+06	1.470202e+07
Impressions	23066.0	1.241520e+06	2.429400e+06	1.000000	7990.500000	225290.000000	1.112428e+06	1.419477e+07
Clicks	23066.0	1.067852e+04	1.735341e+04	1.000000	710.000000	4425.000000	1.279375e+04	1.430490e+05
Spend	23066.0	2.706626e+03	4.067927e+03	0.000000	85.180000	1425.125000	3.121400e+03	2.693187e+04
Fee	23066.0	3.351231e-01	3.196322e-02	0.210000	0.330000	0.350000	3.500000e-01	3.500000e-01
Revenue	23066.0	1.924252e+03	3.105238e+03	0.000000	55.365375	926.335000	2.091338e+03	2.127618e+04
CTR	23066.0	8.409941e+00	9.262048e+00	0.010874	0.265107	9.391248	1.347057e+01	2.000000e+02
CPM	23066.0	8.396849e+00	9.057760e+00	0.000000	1.749084	8.371566	1.304202e+01	7.150000e+02
CPC	23066.0	3.366776e-01	3.412527e-01	0.000000	0.089736	0.139347	5.462421e-01	7.264000e+00

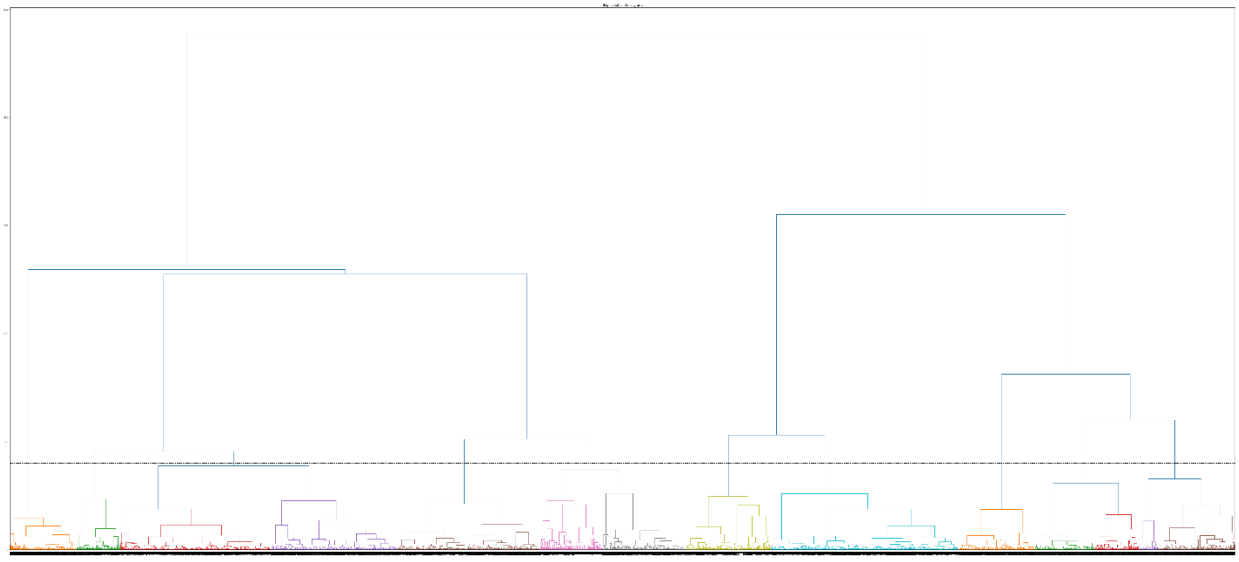
- Dataset that is present after performing z-score scaling

Scaling can increase the computational complexity of algorithms, as it involves additional computations to transform the data.

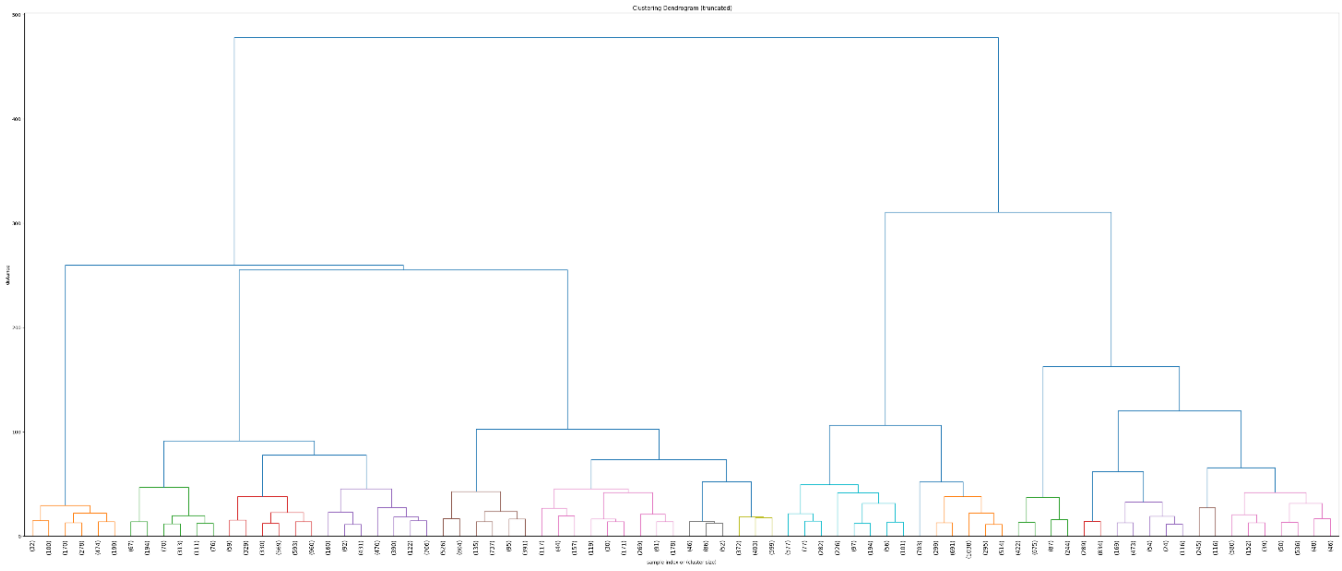
	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	1.281478e-16	1.000022	-1.134891	-1.134891	-0.364496	1.433093	1.467332
Ad- Width	23066.0	-1.182903e-16	1.000022	-1.319110	-0.432797	-0.186599	1.290590	1.290590
Ad Size	23066.0	3.055833e-16	1.000022	-1.467840	-0.297564	-0.297564	0.482620	1.652896
Available_Impressions	23066.0	9.857525e-18	1.000022	-0.756182	-0.740341	-0.528577	0.433059	2.193158
Matched_Queries	23066.0	1.971505e-17	1.000022	-0.779265	-0.761447	-0.527722	0.371498	2.070914
Impressions	23066.0	0.000000e+00	1.000022	-0.768806	-0.760655	-0.538975	0.366051	2.056111
Clicks	23066.0	-1.182903e-16	1.000022	-0.867488	-0.793438	-0.405431	0.468629	2.361729
Spend	23066.0	-9.857525e-17	1.000022	-0.893170	-0.858046	-0.305523	0.393932	2.271900
Fee	23066.0	1.143473e-15	1.000022	-2.222416	-0.567532	0.535724	0.535724	0.535724
Revenue	23066.0	3.943010e-17	1.000022	-0.880093	-0.846474	-0.317607	0.389803	2.244218
CTR	23066.0	1.380054e-16	1.000022	-0.995031	-0.964227	0.141524	0.635787	3.035808
CPM	23066.0	2.464381e-17	1.000022	-1.194498	-0.940303	0.022146	0.700905	3.162718
CPC	23066.0	3.943010e-17	1.000022	-1.042561	-0.759091	-0.602371	0.682987	2.846105

Part 1 - Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distances

The below Dendrogram is performed for WARD and Euclidean distance on the scale data that is below



The value of p=10 that means that the last 10 clusters are merged and is shown below



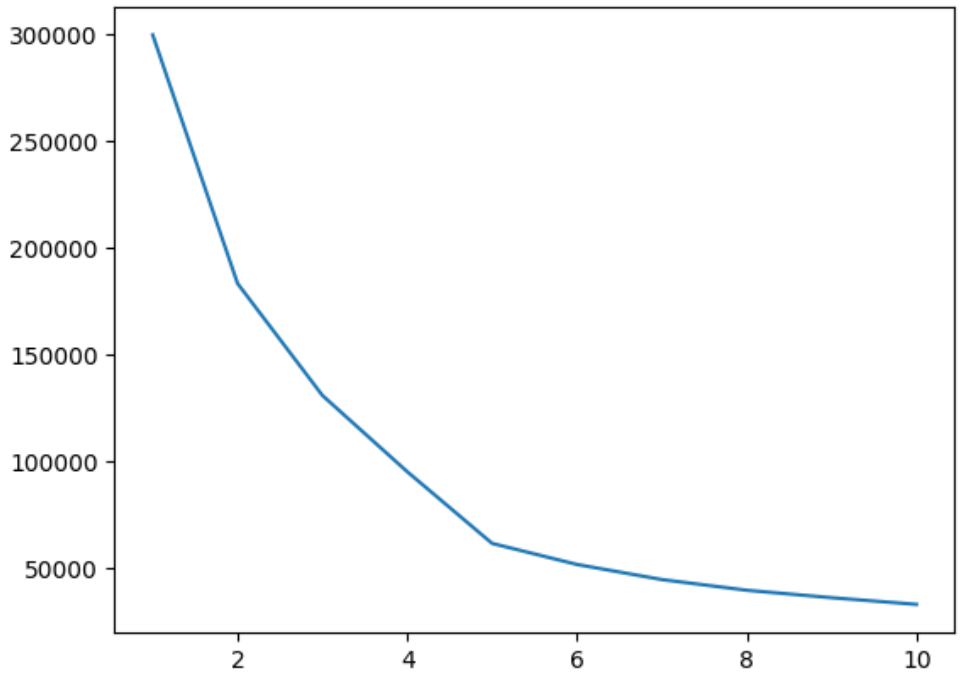
Part 1 - Clustering: Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.

We have to found out the WSS i.e. Within sum of square for finding the optimum number

We have find till 10th value and we got the values as shown below for WSS.

```
[299857.99999999866,
 183349.11866292177,
 130878.34240367389,
 95133.93066619692,
 61539.18919785395,
 51676.89681600456,
 44598.25849746805,
 39597.886661507146,
 36102.40483428025,
 32981.09334118491]
```

The below plot is elbow plot as shown below. We can see that the optimum number is 5 as we can see there is no much significant drop after n=5 . Please find the below elbow plot for the reference.



Part 1 - Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters.

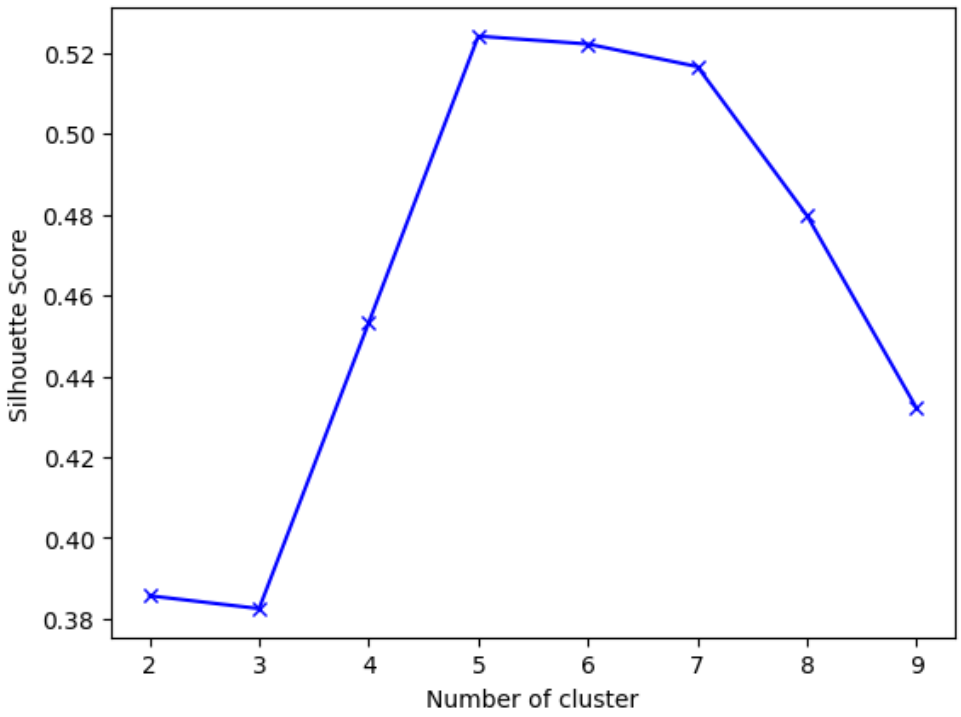
As we know that higher the silhouette score the better is the within clustering compared to the other cluster . For this we have calculated silhouette score till 10 . We got the below silhouette score for upto 10 clusters .

```
{2: 0.3857276961910107 7,
 3: 0.3825486036570082,
 4: 0.45324270552598256,
 5: 0.5240956940501831,
 6: 0.5221533662938636,
 7: 0.5165635029478517,
 8: 0.47972249893837277,
 9: 0.4320636564025043}
```

As we can see the optimum number is n=5 as it is the highest among all and as we know that higher the silhouette score the better is the within clustering compared to the other cluster.

We can see from the plot that silhouette score is highest for k=5. Well that makes it slightly easy for us and we can start with first understanding these 5 clusters

The silhouette score graph:



Part 1 - 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].

We can see from the plot that silhouette score is highest for k=5. Well that makes it slightly easy for us and we can start with first understanding these 5 clusters. So let's take the number of clusters as 5.

We have to identify the trend on different parameters mentioned above based on Device type.

We have added predicted labels to original data and scaled data

```
4    6524
0    6275
1    4676
2    4054
3    1537
Name: KMeans_Labels, dtype: int64
```

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

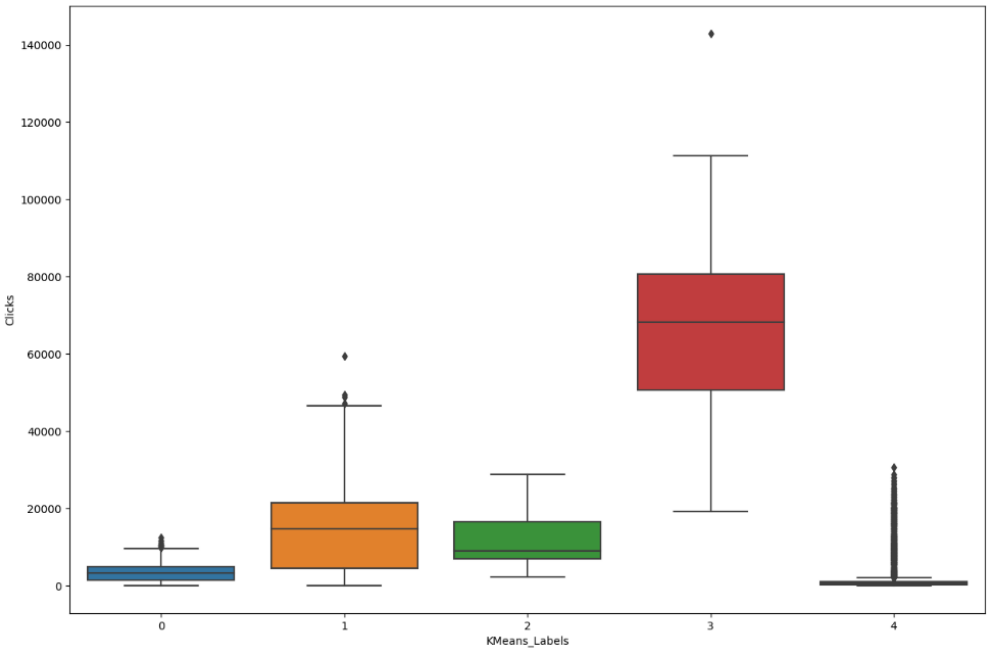
	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	KMeans_Labels
0	nter222	Video	Desktop	Display	1806	325	323	1	0.0	0.35	0.0	0.309598	0.0	0.0	0
1	nter227	App	Mobile	Video	1780	285	285	1	0.0	0.35	0.0	0.350877	0.0	0.0	0
2	nter222	Video	Desktop	Display	2727	356	355	1	0.0	0.35	0.0	0.281690	0.0	0.0	0
3	nter228	Video	Mobile	Video	2430	497	495	1	0.0	0.35	0.0	0.202020	0.0	0.0	0
4	nter217	Web	Desktop	Video	1218	242	242	1	0.0	0.35	0.0	0.413223	0.0	0.0	0

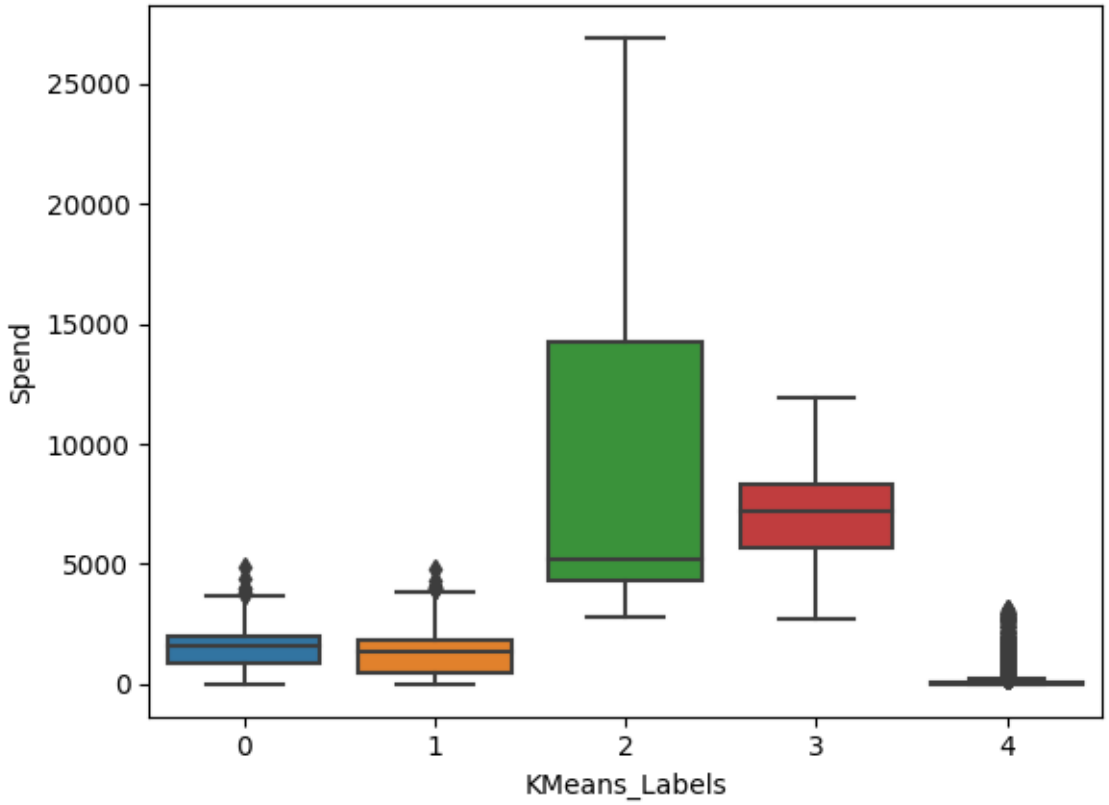
After that we have calculated mean and median of original data for each label

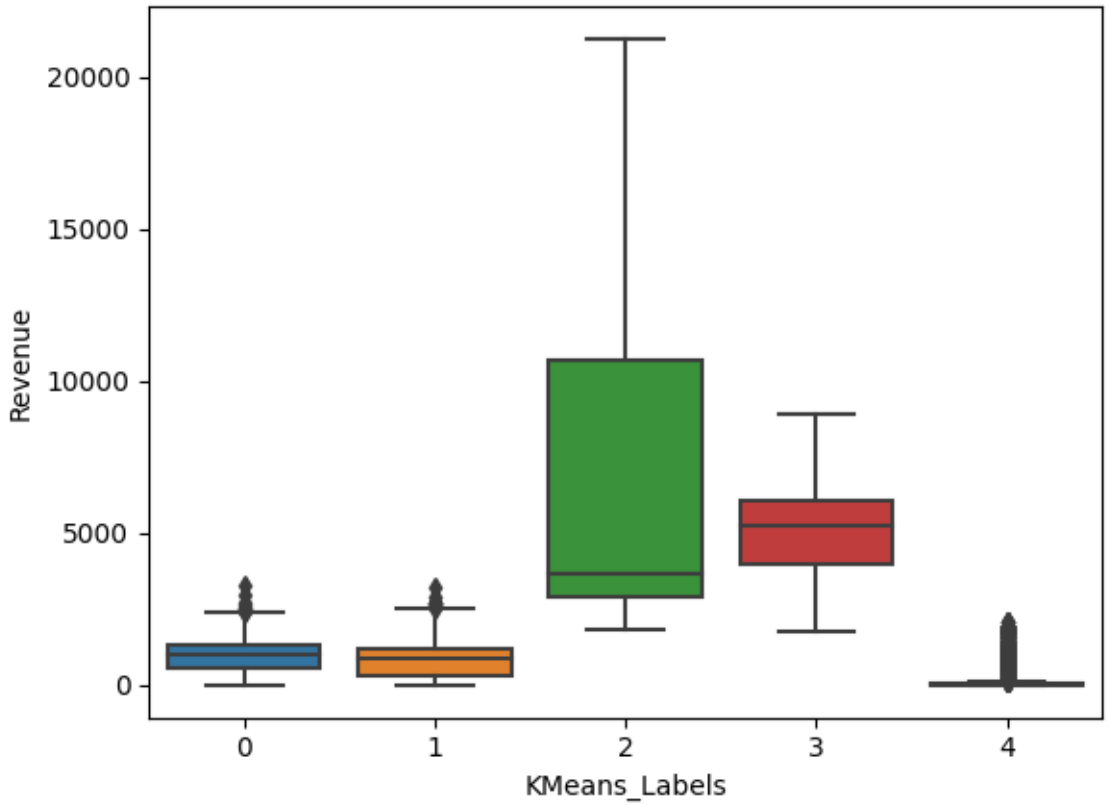
	group_0 Mean	group_1 Mean	group_2 Mean	group_3 Mean	group_4 Mean	group_0 Median	group_1 Median	group_2 Median	group_3 Median	
Ad - Length	4.216963e+02	683.825492	4.657819e+02	141.454782	143.280809	4.800000e+02	720.000000	3.000000e+02	120.000000	12
Ad- Width	1.520016e+02	303.785287	1.991490e+02	572.446324	572.103004	7.000000e+01	300.000000	2.500000e+02	600.000000	60
Ad Size	5.500884e+04	206160.821215	7.517657e+04	75614.834092	76597.026364	3.360000e+04	216000.000000	7.500000e+04	72000.000000	7200
Available_Impressions	1.810314e+06	251346.513687	1.038821e+07	806328.422902	32093.558860	1.833902e+06	213610.000000	7.052044e+06	831024.000000	1392
Matched_Queries	8.642623e+05	137550.912104	5.625808e+06	566864.050748	19624.057633	8.610010e+05	137284.000000	3.876646e+06	583232.000000	826
Impressions	8.262209e+05	116771.362703	5.447310e+06	478148.522446	13492.040313	8.156060e+05	115859.000000	3.794097e+06	490310.000000	363
Clicks	3.263132e+03	14406.540205	1.124575e+04	65315.176318	1914.448804	3.254000e+03	14723.500000	8.928500e+03	68257.000000	45
Spend	1.500091e+03	1252.285569	8.646648e+03	6990.360898	209.162609	1.548630e+03	1336.860000	5.229225e+03	7172.600000	4
Fee	3.492637e-01	0.349538	2.904391e-01	0.288302	0.349988	3.500000e-01	0.350000	3.000000e-01	0.270000	
Revenue	9.774242e+02	815.541831	6.373660e+03	5017.538285	135.993379	1.006610e+03	868.955000	3.660460e+03	5236.000000	3
CTR	4.043922e-01	13.857220	2.172420e-01	13.752664	16.037897	3.979831e-01	12.469471	2.242493e-01	13.575089	1
CPM	1.788731e+00	12.098200	1.573280e+00	15.385753	14.693481	1.817682e+00	11.087378	1.564732e+00	14.826422	1
CPC	5.446141e-01	0.090012	7.609292e-01	0.111918	0.102794	4.687385e-01	0.090971	7.148931e-01	0.111486	

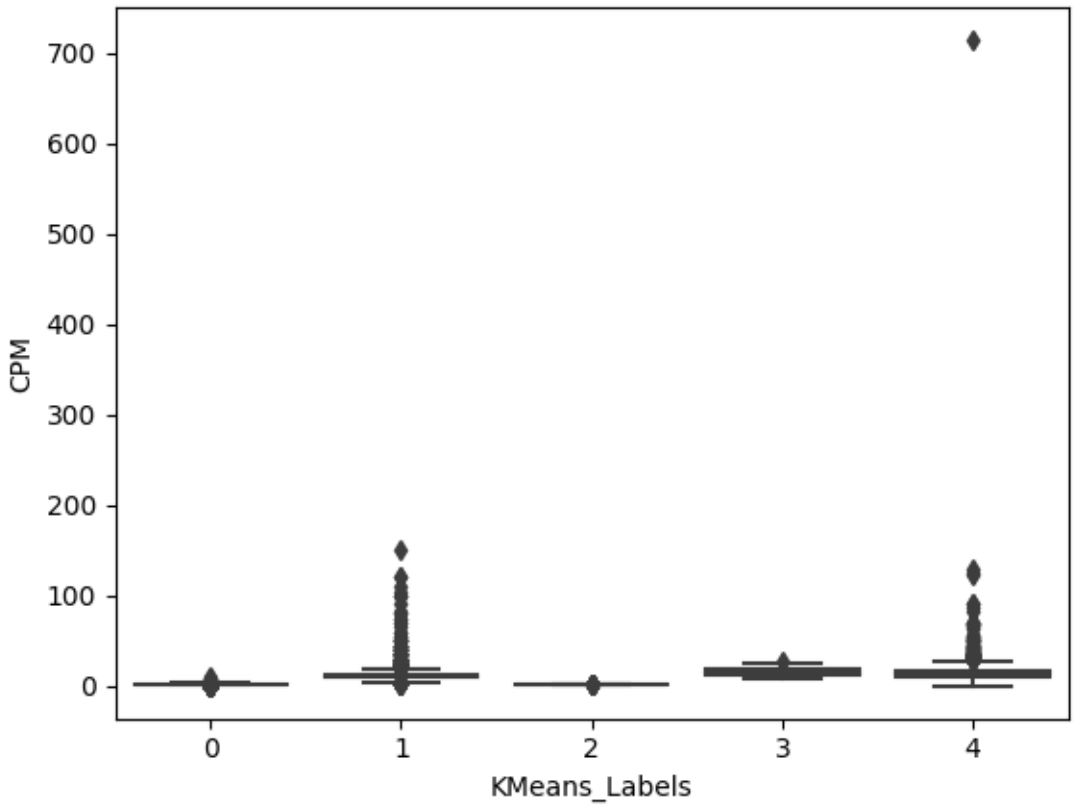
	group_0 Mean	group_1 Mean	group_2 Mean	group_3 Mean	group_4 Mean	group_0 Median	group_1 Median	group_2 Median	group_3 Median	group_4 Median
ength	4.216963e+02	683.825492	4.657819e+02	141.454782	143.280809	4.800000e+02	720.000000	3.000000e+02	120.000000	120.000000
Width	1.520016e+02	303.785287	1.991490e+02	572.446324	572.103004	7.000000e+01	300.000000	2.500000e+02	600.000000	600.000000
Size	5.500884e+04	206160.821215	7.517657e+04	75614.834092	76597.026364	3.360000e+04	216000.000000	7.500000e+04	72000.000000	72000.000000
mpressions	1.810314e+06	251346.513687	1.038821e+07	806328.422902	32093.558860	1.833902e+06	213610.000000	7.052044e+06	831024.000000	13920.500000
l_Queries	8.642623e+05	137550.912104	5.625808e+06	566864.050748	19624.057633	8.610010e+05	137284.000000	3.876646e+06	583232.000000	8263.500000
ssions	8.262209e+05	116771.362703	5.447310e+06	478148.522446	13492.040313	8.156060e+05	115859.000000	3.794097e+06	490310.000000	3631.500000
icks	3.263132e+03	14406.540205	1.124575e+04	65315.176318	1914.448804	3.254000e+03	14723.500000	8.928500e+03	68257.000000	454.000000
end	1.500091e+03	1252.285569	8.646648e+03	6990.360898	209.162609	1.548630e+03	1336.860000	5.229225e+03	7172.600000	47.460000
ee	3.492637e-01	0.349538	2.904391e-01	0.288302	0.349988	3.500000e-01	0.350000	3.000000e-01	0.270000	0.350000
enue	9.774242e+02	815.541831	6.373660e+03	5017.538285	135.993379	1.006610e+03	868.955000	3.660460e+03	5236.000000	30.848000
TR	4.043922e-01	13.857220	2.172420e-01	13.752664	16.037897	3.979831e-01	12.469471	2.242493e-01	13.575089	13.823976
PM	1.788731e+00	12.098200	1.573280e+00	15.385753	14.693481	1.817682e+00	11.087378	1.564732e+00	14.826422	13.381682
PC	5.446141e-01	0.090012	7.609292e-01	0.111918	0.102794	4.687385e-01	0.090971	7.148931e-01	0.111486	0.092500

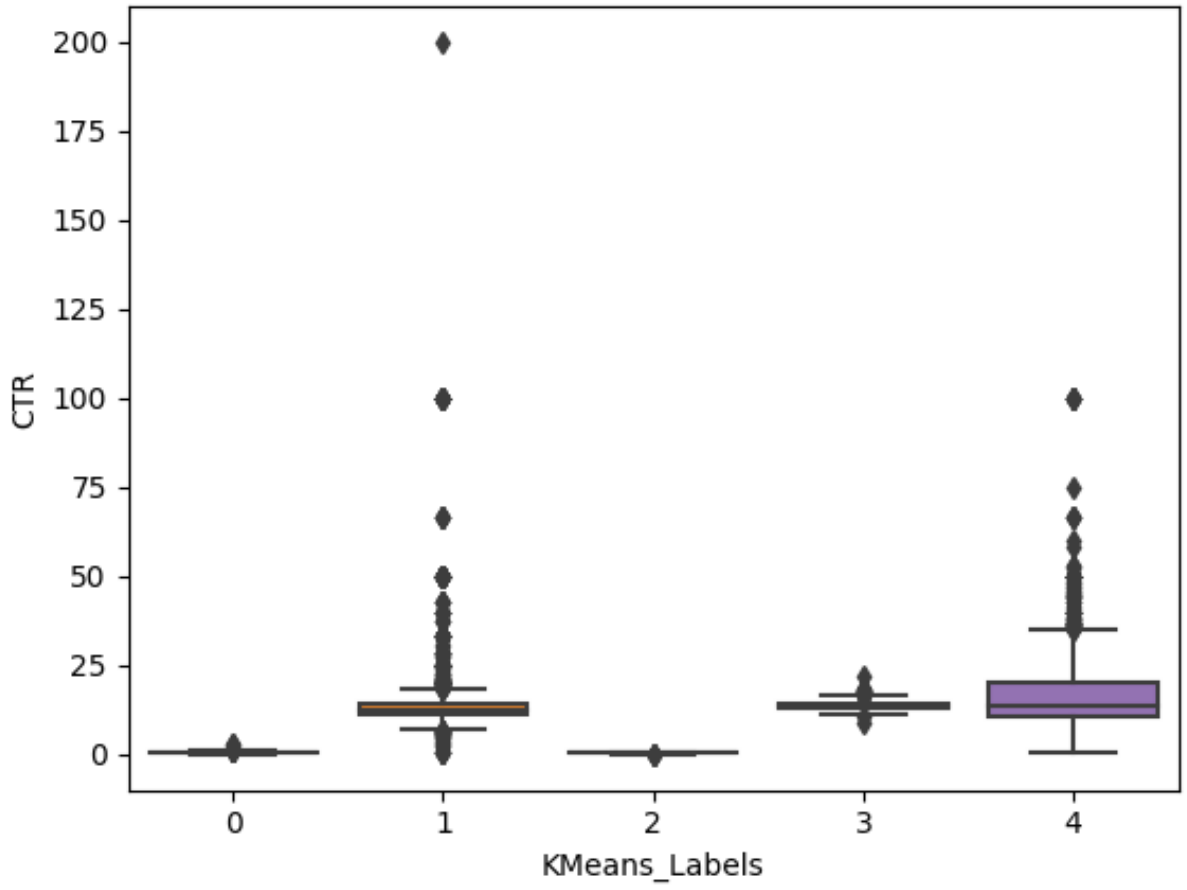
We have plotted boxplot of the parameters **in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type**

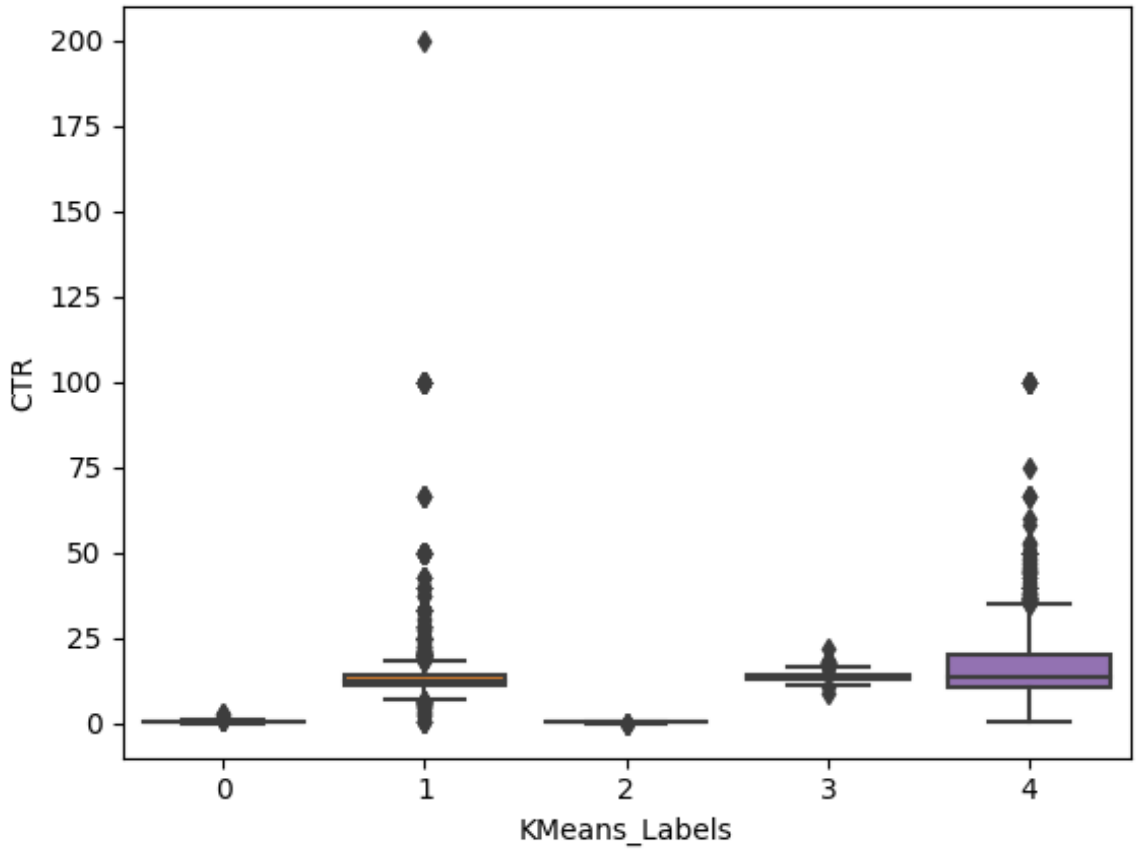



25000
20000
15000
10000
5000
0
Spend
0
1
2
3
4
KMeans_Labels


20000
15000
10000
5000
0
Revenue
0
1
2
3
4
KMeans_Labels


700
600
500
400
300
200
100
0
CPM
0
1
2
3
4
KMeans_Labels


200
175
150
125
100
75
50
25
0
CTR
0
1
2
3
4
KMeans_Labels



Part 1 - Clustering: Conclude the project by providing summary of your learnings.

- We have seen that in the provide dataset there are 23066 rows and 19 columns
- After that we found that there are null values that is present in 3 columns that is CPM,CTR, CPC
- The number of missing values that is present in each is as below
    1. CPM= 4736
    2. CTR=4736
    3. CPC=4736

- We have treated the null values of all these 3 parameters
- We have seen that there are outliers that are present in the columns AD size, Available impression, Matched queries, Impressions , Clicks , Spend, Revenue ,Fee, CTR,CPM,CPC
- We have treated the outliers after we found that there is outliers present in the dataset by using IQR method.
- We have applied zscore technique for scaling of our data and make a different dataset for the scaled data that is with the name scale_data
- We have plotted the dendrogram for p=10 because last 10 clusters are merged.
- We got optimum value for k means algorithm that is 5.
- After that we have plotted the elbow plot
- We have calculated the silhouette score and found that the optimum number for silhouette score is 5 as we know that the higher the silhouette score the better is the clustering within rather than the other cluster.

Conclusion after clustering

- We have seen that as Click on AD increases Revenue also increases
- When impression count on advertisement increases then revenue increases.
- Revenue will be increased then also when money spend on specific variation within a specific campaign increases.

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.
The data collected has so many variables thus making it difficult to find useful details without using Data Science Techniques. You are tasked to perform detailed EDA and identify Optimum Principal Components that explains the most variance in data. Use Sklearn only.
Note: The 24 variables given in the Rubric is just for performing EDA. You will have to consider the entire dataset, including all the variables for performing PCA.

Part 2 - PCA: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.

- Printing the first 5 rows of the data-frame .

	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_I
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180	237	
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123	229	
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	...	114	188	44	89	
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	...	194	247	61	128	
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	...	874	1928	465	1043	

- Printing the last 5 rows of the data-frame .

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

5 rows × 61 columns

- Rows and columns that is there in PCA set

```
(640, 61)
```

- Describe function to find out mean , std,count etc.

	count	mean	std	min	25%	50%	75%	max
State Code	640.0	17.114062	9.426486	1.0	9.00	18.0	24.00	35.0
Dist.Code	640.0	320.500000	184.896367	1.0	160.75	320.5	480.25	640.0
No_HH	640.0	51222.871875	48135.405475	350.0	19484.00	35837.0	68892.00	310450.0
TOT_M	640.0	79940.576563	73384.511114	391.0	30228.00	58339.0	107918.50	485417.0
TOT_F	640.0	122372.084375	113600.717282	698.0	46517.75	87724.5	164251.75	750392.0
M_06	640.0	12309.098438	11500.906881	56.0	4733.75	9159.0	16520.25	96223.0
F_06	640.0	11942.300000	11326.294567	56.0	4672.25	8663.0	15902.25	95129.0
M_SC	640.0	13820.946875	14426.373130	0.0	3466.25	9591.5	19429.75	103307.0
F_SC	640.0	20778.392188	21727.887713	0.0	5603.25	13709.0	29180.00	156429.0
M_ST	640.0	6191.807813	9912.668948	0.0	293.75	2333.5	7658.00	96785.0
F_ST	640.0	10155.640625	15875.701488	0.0	429.50	3834.5	12480.25	130119.0

M_06	640.0	12309.098438	11500.906881	56.0	4733.75	9159.0	16520.25	96223.0
F_06	640.0	11942.300000	11326.294567	56.0	4672.25	8663.0	15902.25	95129.0
M_SC	640.0	13820.946875	14426.373130	0.0	3466.25	9591.5	19429.75	103307.0
F_SC	640.0	20778.392188	21727.887713	0.0	5603.25	13709.0	29180.00	156429.0
M_ST	640.0	6191.807813	9912.668948	0.0	293.75	2333.5	7658.00	96785.0
F_ST	640.0	10155.640625	15875.701488	0.0	429.50	3834.5	12480.25	130119.0
M_LIT	640.0	57967.979688	55910.282466	286.0	21298.00	42693.5	77989.50	403261.0
F_LIT	640.0	66359.565625	75037.860207	371.0	20932.00	43796.5	84799.75	571140.0
M_ILL	640.0	21972.596875	19825.605268	105.0	8590.00	15767.5	29512.50	105961.0
F_ILL	640.0	56012.518750	47116.693769	327.0	22367.00	42386.0	78471.00	254160.0
TOT_WORK_M	640.0	37992.407813	36419.537491	100.0	13753.50	27936.5	50226.75	269422.0
TOT_WORK_F	640.0	41295.760938	37192.360943	357.0	16097.75	30588.5	53234.25	257848.0

MARGWORK_M	640.0	7787.960938	7410.791691	35.0	2937.50	5627.0	9800.25	47553.0
MARGWORK_F	640.0	13096.914062	10996.474528	117.0	5424.50	10175.0	18879.25	66915.0
MARG_CL_M	640.0	1040.737500	1311.546847	0.0	311.75	606.5	1281.00	13201.0
MARG_CL_F	640.0	2307.682813	3564.626095	0.0	630.25	1226.0	2659.25	44324.0
MARG_AL_M	640.0	3304.326562	3781.555707	0.0	873.50	2062.0	4300.75	23719.0
MARG_AL_F	640.0	6463.281250	6773.876298	0.0	1402.50	4020.5	9089.25	45301.0
MARG_HH_M	640.0	316.742188	462.661891	0.0	71.75	166.0	356.50	4298.0
MARG_HH_F	640.0	786.626562	1198.718213	0.0	171.75	429.0	962.50	15448.0
MARG_OT_M	640.0	3126.154687	3609.391821	7.0	935.50	2036.0	3985.25	24728.0
MARG_OT_F	640.0	3539.323438	4115.191314	19.0	1071.75	2349.5	4400.50	36377.0
MARGWORK_3_6_M	640.0	41948.168750	39045.316918	291.0	16208.25	30315.0	57218.75	300937.0
MARGWORK_3_6_F	640.0	81076.323438	82970.406216	341.0	26619.50	56793.0	107924.00	676450.0
MARG_CL_3_6_M	640.0	6394.987500	6019.806644	27.0	2372.00	4630.0	8167.00	39106.0

- Information of dataset that is giveb

```
<bound method DataFrame.info of      State Code  Dist.Code                    State                Area Name  \
0             1          1           Jammu & Kashmir                  Kupwara
1             1          2           Jammu & Kashmir                   Badgam
2             1          3           Jammu & Kashmir              Leh(Ladakh)
3             1          4           Jammu & Kashmir                   Kargil
4             1          5           Jammu & Kashmir                    Punch
..          ...        ...                       ...                      ...
635          34        636                Puducherry                    Mahe
636          34        637                Puducherry                Karaikal
637          35        638  Andaman & Nicobar Island                 Nicobars
638          35        639  Andaman & Nicobar Island  North  & Middle Andaman
639          35        640  Andaman & Nicobar Island            South Andaman

     No_HH  TOT_M  TOT_F  M_06  F_06  M_SC  ...  MARG_CL_0_3_M  MARG_CL_0_3_F  \
0     7707  23388  29796  5862  6196     3  ...           1150            749
1     6218  19585  23102  4482  3733     7  ...            525            715
2     4452   6546  10964  1082  1018     3  ...            114            188
3     1320   2784   4206   563   677     0  ...            194            247
4    11654  20591  29981  5157  4587    20  ...            874           1928
..     ...    ...    ...   ...   ...   ...  ...            ...            ...
635   3333   8154  11781  1146  1203    21  ...             32             47
636  10612  12346  21691  1544  1533  2234  ...            155            337
637   1275   1549   2630   227   225     0  ...            104            134
638   3762   5200   8012   723   664     0  ...            136            172
639   7975  11977  18049  1470  1358     0  ...            173            122
```

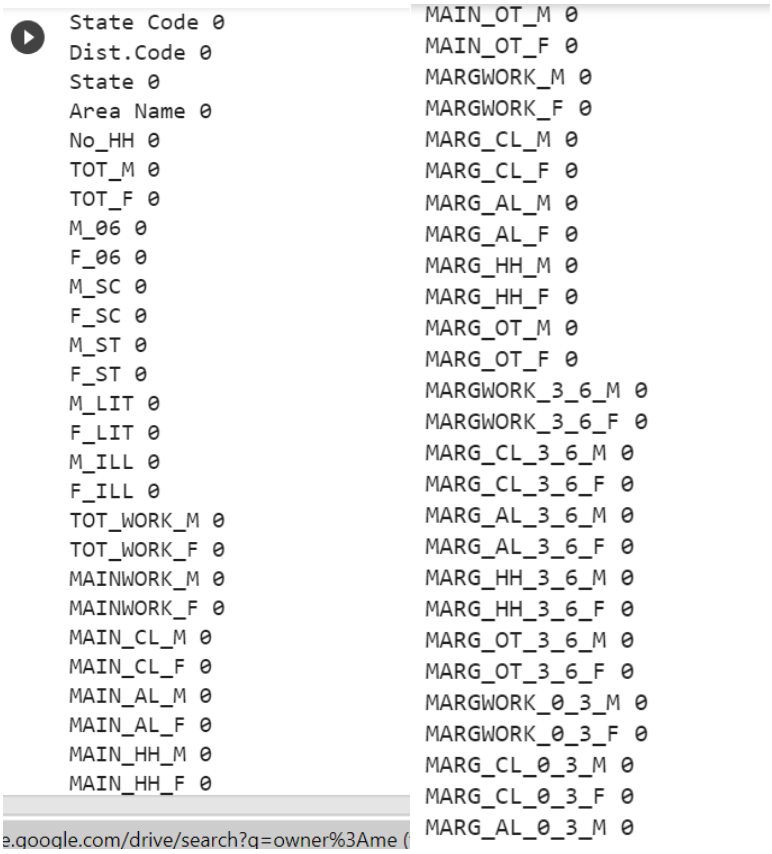
```
     MARG_AL_0_3_M  MARG_AL_0_3_F  MARG_HH_0_3_M  MARG_HH_0_3_F  \
0              180            237            680            252
1              123            229            186            148
2               44             89              3             34
3               61            128             13             50
4              465           1043            205            302
..             ...            ...            ...            ...
635              0              0              0              0
636              3             14             38            130
637              9              4              2              6
638             24             44             11             21
639              6              2             17             17

     MARG_OT_0_3_M  MARG_OT_0_3_F  NON_WORK_M  NON_WORK_F
0               32             46         258         214
1               76            178         140         160
2                0              4          67          61
3                4             10         116          59
4               24            105         180         478
..             ...            ...         ...         ...
635              0              0          32          47
636              4             23         110         170
637             17             47          76          77
638              1              4         100         103
639              2              4         148          99

[640 rows x 61 columns]>
```

- Check for the null values

We found that there is no null values present in dataset



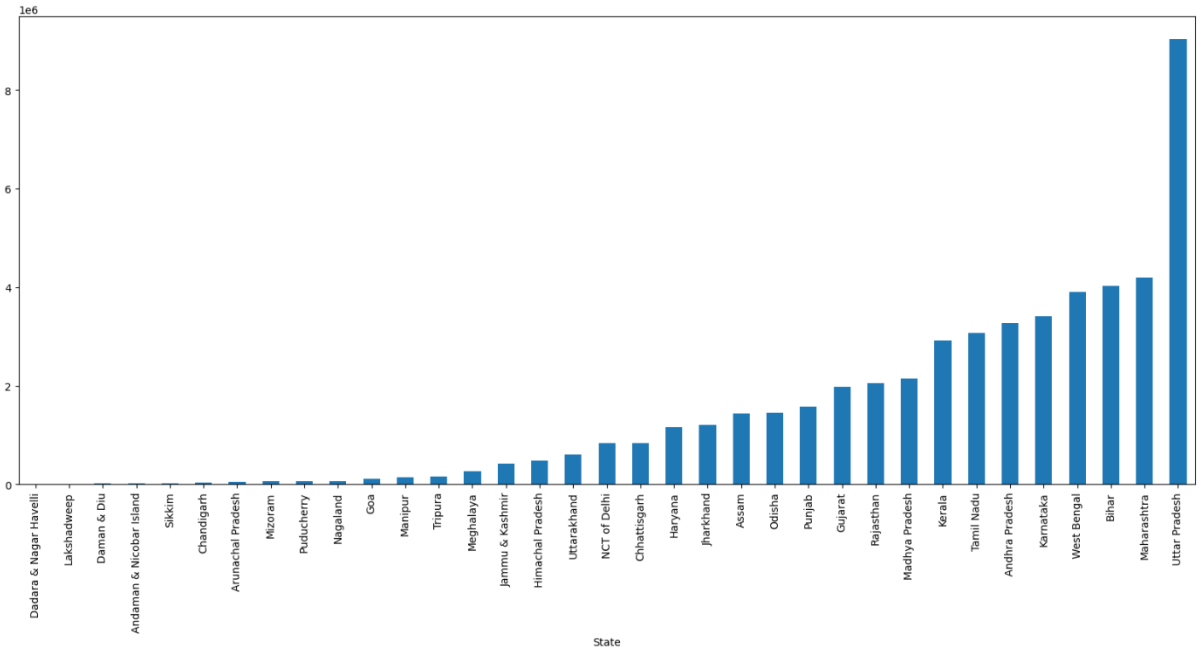
No duplicates rows found

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

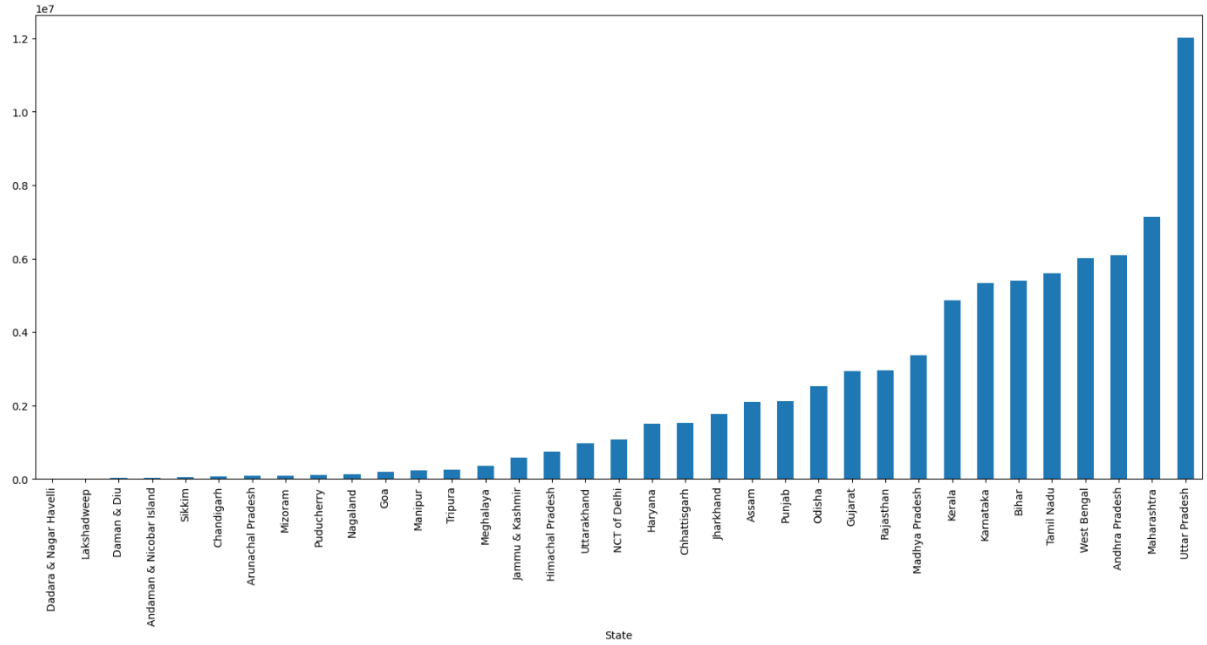
We have taken 5 parameters that is ‘TOT_M’ ,’ TOT_F, 'F_LIT','M_LIT', NON_WORK_M', NON_WORK_F

TOT_M = Total male
TOT_F =Total female
F_LIT= Female literate
M_LIT=Male literate
NON_WORK_M = Non-working male
NON_WORK_F = Non-working female

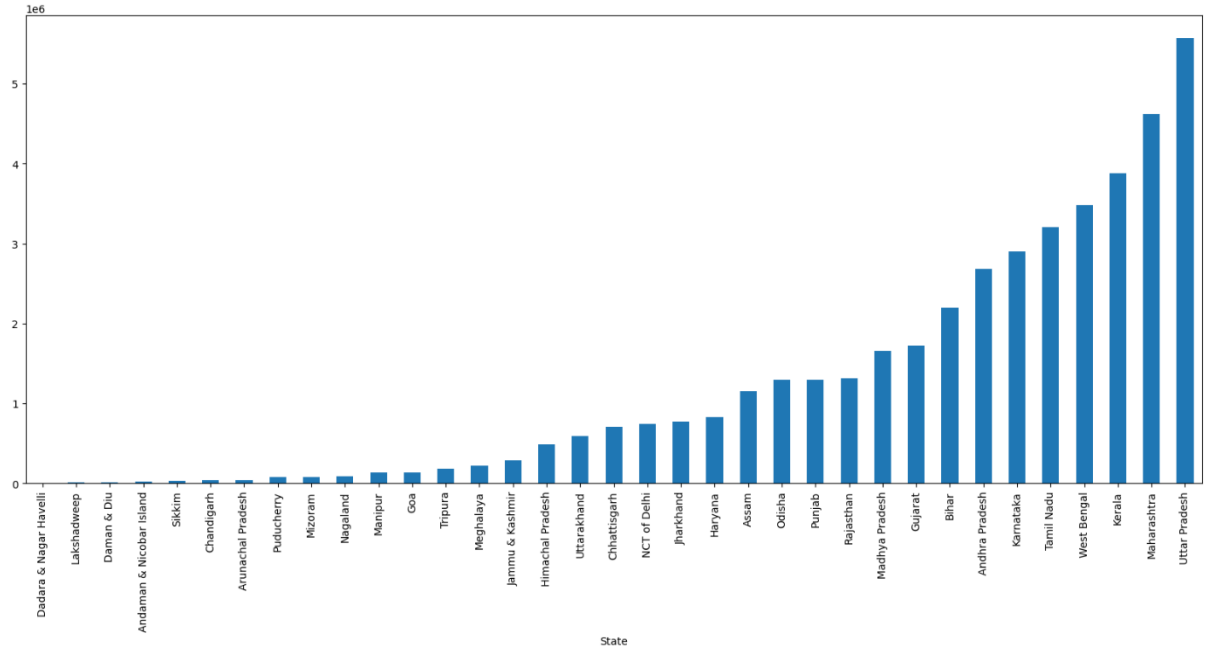
- which state has highest number of males ?
- which state has lowest number of males ?

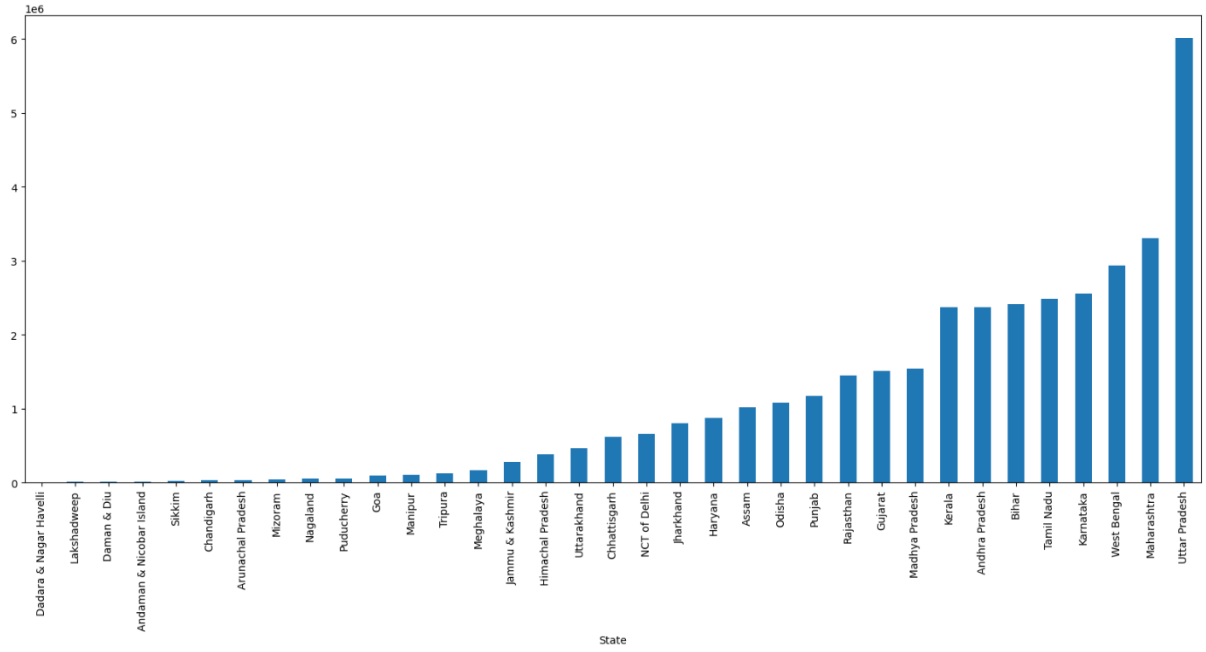


- which state has highest number of females ?
- which state has lowest number of females ?

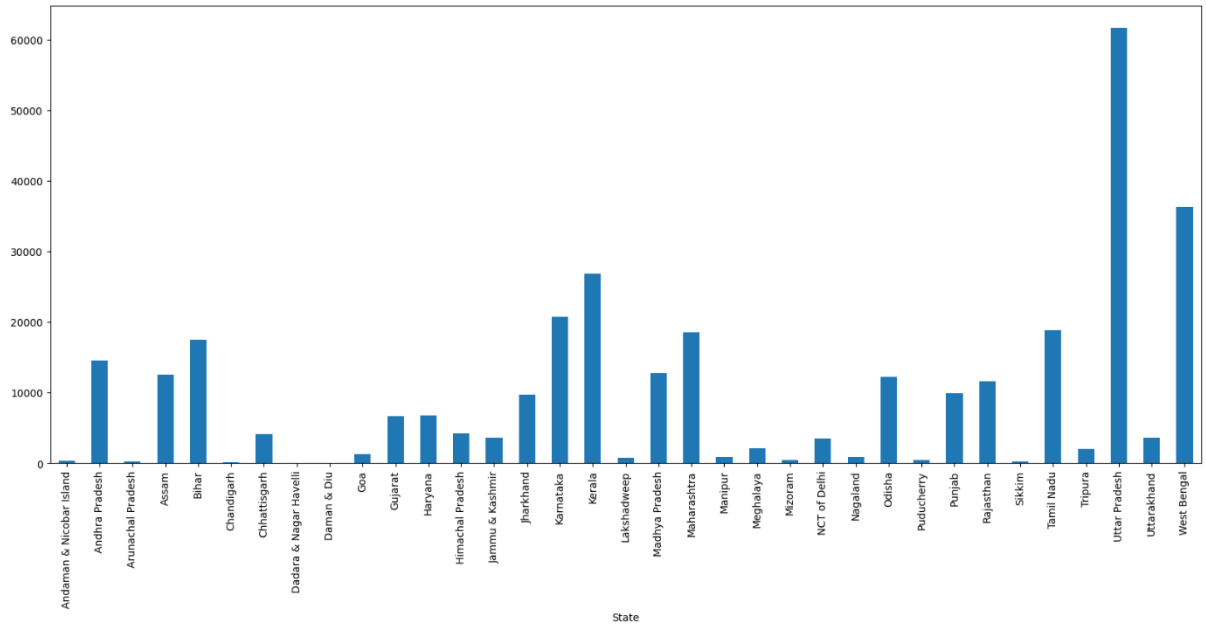


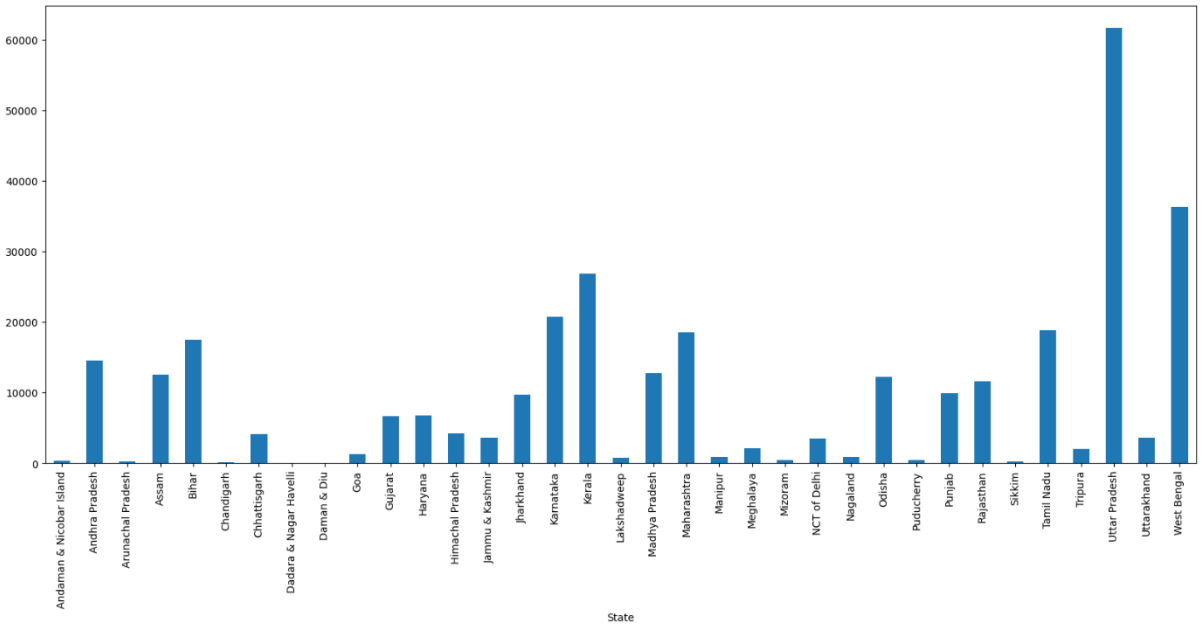
1) Which state has maximum number of literate females?
2)Which state has maximum number of literate males?





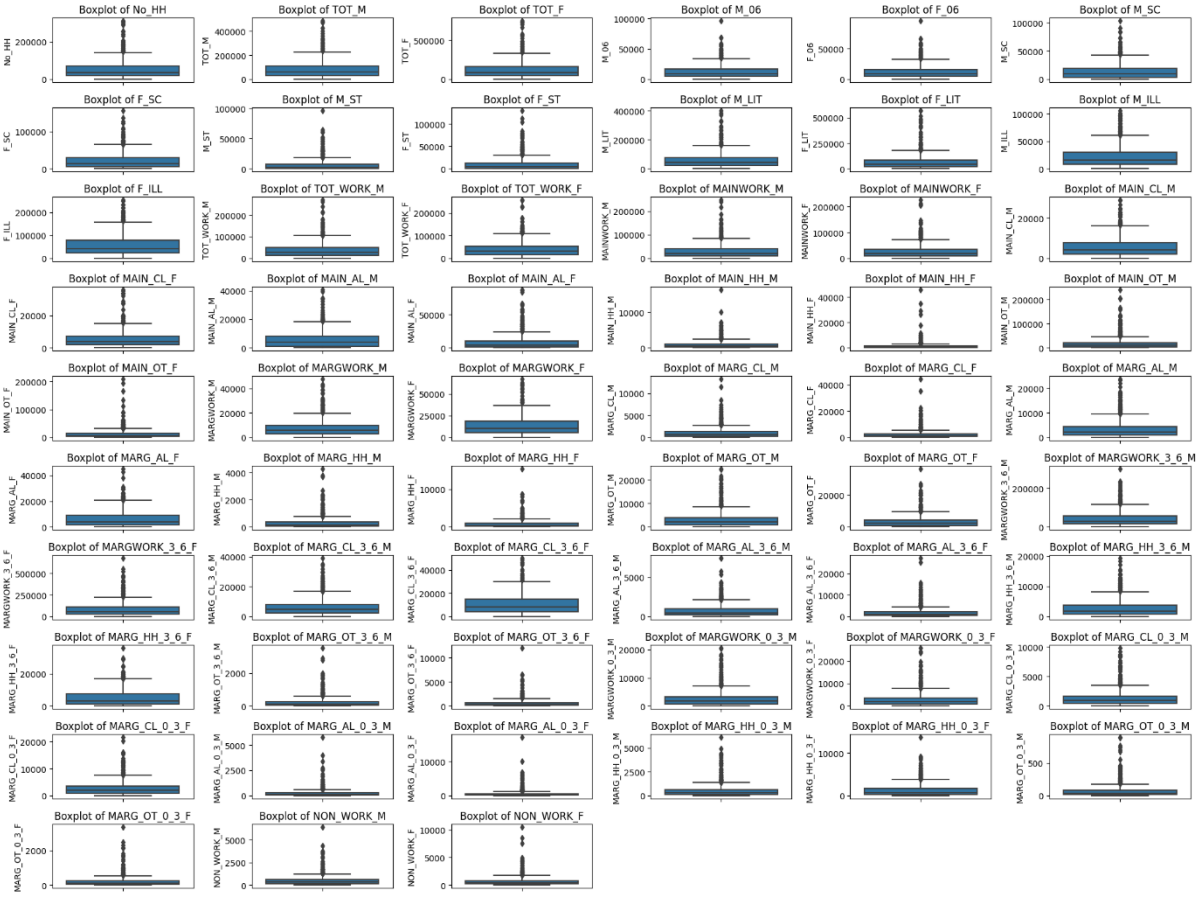
1)which state has most non working males?
2)Which state has most non working female?





Part 2 - PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?

There is outliers present in the dataset as shown below but we did not prefer to treat the outliers as in this we did not treat the outliers as mentioned



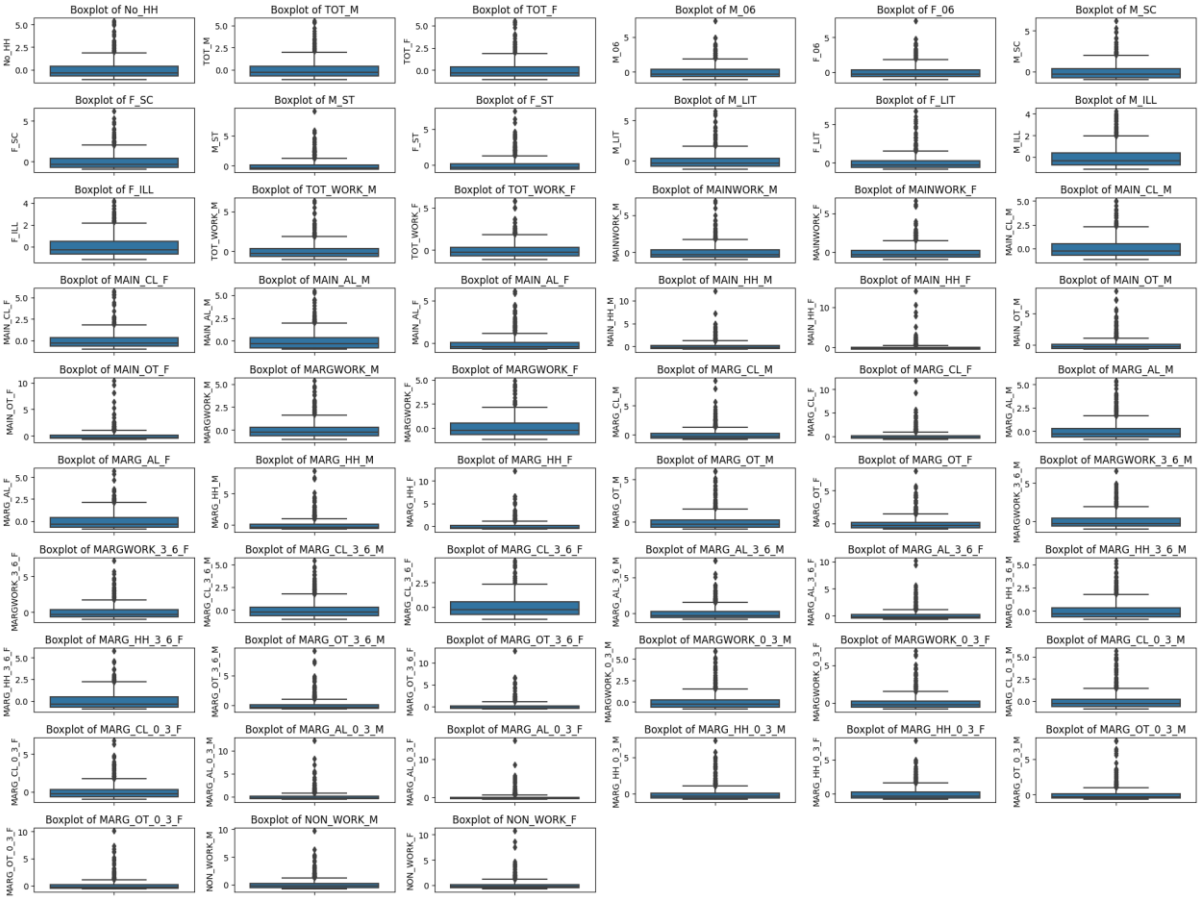
Part 2 - PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.

Z-score

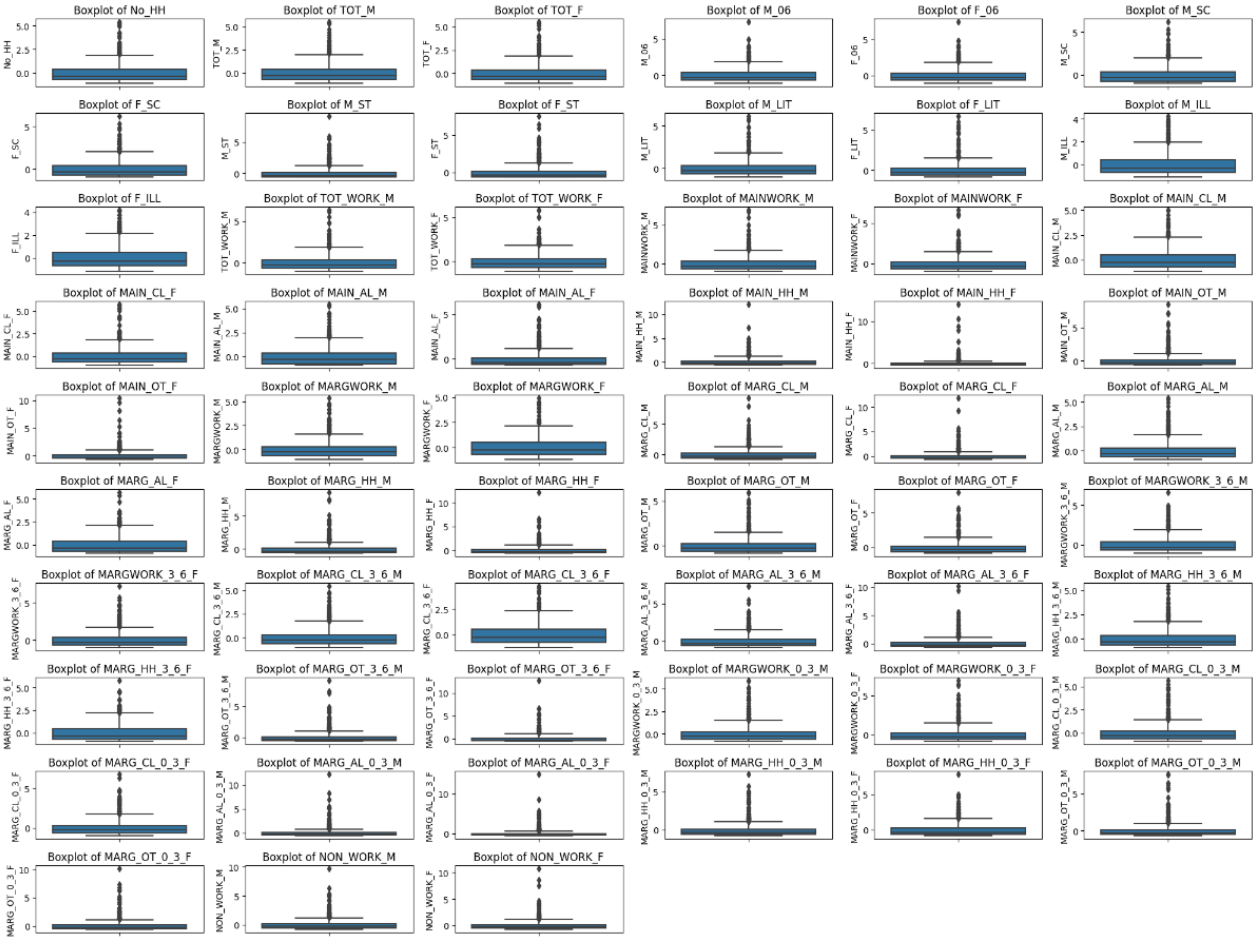
We have scaled the data using zscore method and after scaling we get the data. Scaling doesnot have any effect on outliers as shown in boxplots belows

	count	mean	std	min	25%	50%	75%	max
No_HH	640.0	4.440892e-17	1.000782	-1.057697	-0.659882	-0.319887	0.367358	5.389586
TOT_M	640.0	-8.881784e-17	1.000782	-1.084858	-0.677956	-0.294592	0.381549	5.529690
TOT_F	640.0	-4.440892e-17	1.000782	-1.071906	-0.668250	-0.305233	0.368945	5.532633
M_06	640.0	-5.551115e-17	1.000782	-1.066236	-0.659189	-0.274114	0.366445	7.301993
F_06	640.0	6.661338e-17	1.000782	-1.050264	-0.642376	-0.289756	0.349898	7.350309
M_SC	640.0	5.551115e-18	1.000782	-0.958783	-0.718323	-0.293404	0.389092	6.207800
F_SC	640.0	-5.551115e-17	1.000782	-0.957049	-0.698964	-0.325615	0.386976	6.248040
M_ST	640.0	-4.440892e-17	1.000782	-0.625124	-0.595467	-0.389534	0.148027	9.146281
TOT_WORK_M	640.0	-4.440892e-17	1.000782	-1.041256	-0.666067	-0.276329	0.336191	6.359515
TOT_WORK_F	640.0	-8.881784e-17	1.000782	-1.101591	-0.678035	-0.288114	0.321244	5.827047
MAINWORK_M	640.0	-2.220446e-17	1.000782	-0.958137	-0.649073	-0.284647	0.315185	6.920918
MAINWORK_F	640.0	4.440892e-17	1.000782	-0.932745	-0.623743	-0.324100	0.229006	6.604449
MAIN_CL_M	640.0	-8.881784e-17	1.000782	-1.145474	-0.718165	-0.266889	0.479501	5.002401
MAIN_CL_F	640.0	-1.110223e-17	1.000782	-1.030785	-0.669985	-0.296408	0.338245	5.769599
MAIN_AL_M	640.0	0.000000e+00	1.000782	-0.914709	-0.747338	-0.299102	0.346882	5.472493
MAIN_AL_F	640.0	4.440892e-17	1.000782	-0.694401	-0.584807	-0.388393	0.131591	6.147314
MAIN_HH_M	640.0	1.665335e-17	1.000782	-0.691816	-0.545061	-0.301644	0.168557	12.167019
MAIN_HH_F	640.0	0.000000e+00	1.000782	-0.434625	-0.356326	-0.264492	0.017305	14.038154
MARGWORK_M	640.0	-1.665335e-17	1.000782	-1.046990	-0.655025	-0.291825	0.271747	5.370026
MARGWORK_F	640.0	2.220446e-17	1.000782	-1.181294	-0.698262	-0.265922	0.526247	4.897950
MARG_CL_M	640.0	0.000000e+00	1.000782	-0.794140	-0.556257	-0.331347	0.183333	9.278947
MARG_CL_F	640.0	-5.551115e-17	1.000782	-0.647891	-0.470946	-0.303687	0.098704	11.796239
MARG_AL_M	640.0	1.110223e-17	1.000782	-0.874484	-0.643314	-0.328780	0.263702	5.402708
MARG_AL_F	640.0	2.220446e-17	1.000782	-0.954894	-0.747687	-0.360900	0.387964	5.737940
MARG_HH_M	640.0	-5.551115e-18	1.000782	-0.685144	-0.529942	-0.326070	0.086000	8.611844
MARG_HH_F	640.0	1.110223e-17	1.000782	-0.656736	-0.513346	-0.298574	0.146833	12.240442
MARG_OT_M	640.0	1.110223e-17	1.000782	-0.864853	-0.607407	-0.302269	0.238203	5.989580
MARG_OT_F	640.0	-4.440892e-17	1.000782	-0.856115	-0.600094	-0.289356	0.209431	7.985865

Before scaling BOXPLOT::

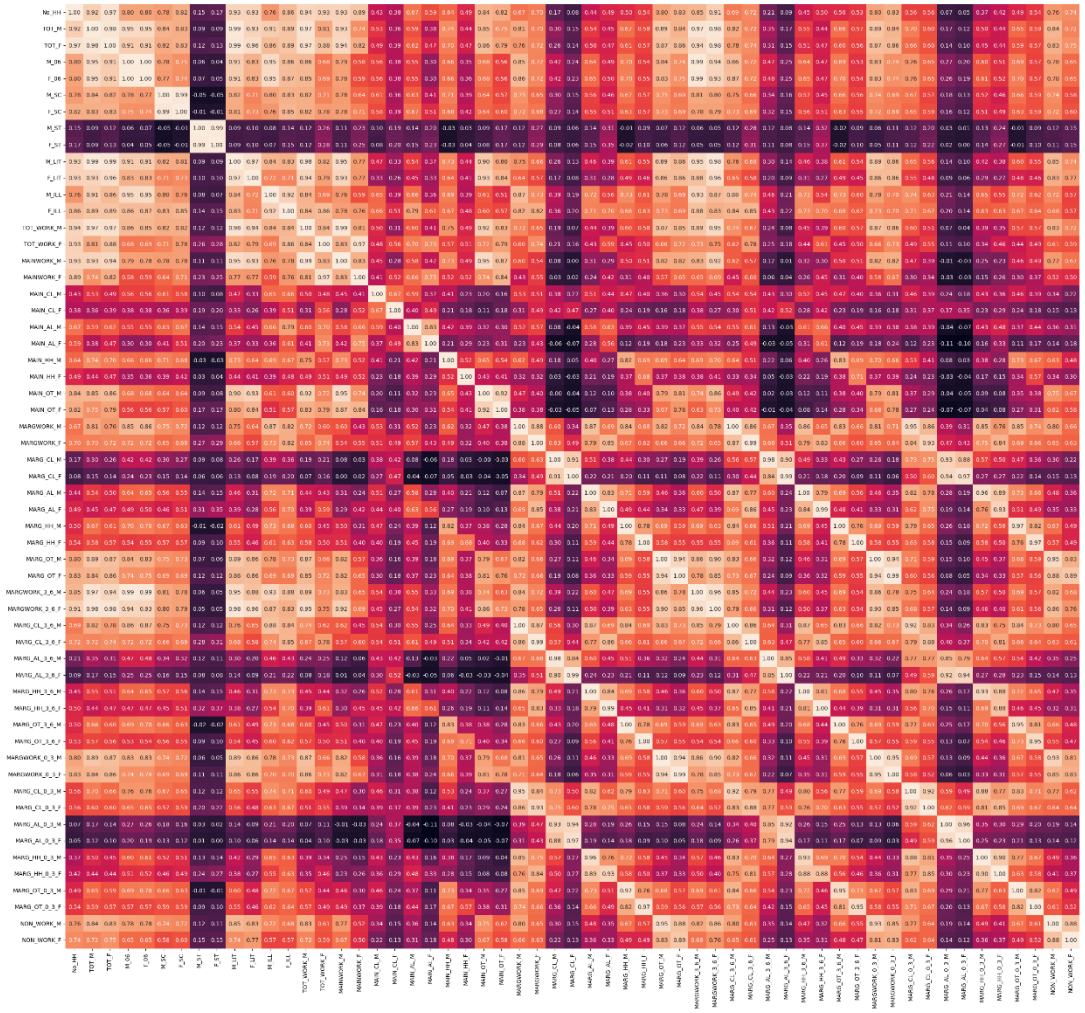


After scaling BOXPLOT::



Part 2 - PCA: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.

Presence of correlation present can be shown by the heatmap for the elements present as shown below



By Bartlett test we get the value as 0

```
0.0
```

By KMO Test we get the below output

Above 0.7 is good, below 0.5 is not acceptable So we got 0.8 that means we are good to go

```
0.8039889932781807
```

Covariance matrix

```
array([[1.00156495, 0.91760364, 0.97210871, ..., 0.53769433, 0.76357722,
        0.73684378],
       [0.91760364, 1.00156495, 0.98417823, ..., 0.5891007 , 0.84621844,
        0.71718181],
       [0.97210871, 0.98417823, 1.00156495, ..., 0.572748  , 0.82894851,
        0.74775097],
       ...,
       [0.53769433, 0.5891007 , 0.572748  , ..., 1.00156495, 0.61052325,
        0.52191235],
       [0.76357722, 0.84621844, 0.82894851, ..., 0.61052325, 1.00156495,
        0.88228018],
       [0.73684378, 0.71718181, 0.74775097, ..., 0.52191235, 0.88228018,
        1.00156495]])
```

We need to check the features

```
57
```

Applying PCA on all the features

Now extracting eigen vectors we got the below values

```
array([[ 0.15602058,  0.16711763,  0.16555318, ...,  0.13219224,
         0.15037558,  0.1310662 ],
       [-0.12634653, -0.08967655, -0.10491237, ...,  0.05081332,
        -0.06536455, -0.07384742],
       [-0.00269025,  0.05669762,  0.03874947, ..., -0.07871987,
         0.11182732,  0.1025525 ],
       ...,
       [ 0.        ,  0.37643683,  0.15058437, ...,  0.03363703,
        -0.07959556, -0.02552519],
       [-0.        ,  0.2448199 ,  0.09383958, ..., -0.02638552,
        -0.01672564,  0.03567243],
       [-0.        , -0.09325898, -0.0110033 , ...,  0.01165739,
        -0.01279215, -0.00377366]])
```

Eigen values

```
array([5.57260632e-01, 1.37844354e-01, 7.27529548e-02, 6.42641771e-02,
       3.86504944e-02, 3.39516923e-02, 2.06023855e-02, 1.31576386e-02,
       1.08085894e-02, 9.25395468e-03, 7.52911540e-03, 6.19101667e-03,
       5.18772384e-03, 4.92694855e-03, 3.36593119e-03, 2.38692984e-03,
       1.98617593e-03, 1.86206747e-03, 1.70414955e-03, 1.40317638e-03,
       1.00910494e-03, 7.77653131e-04, 6.63717190e-04, 5.19117774e-04,
       4.74341222e-04, 4.10687364e-04, 2.54183814e-04, 1.92422147e-04,
       1.63167083e-04, 1.42503342e-04, 1.38248605e-04, 8.80379297e-05,
       4.55026824e-05, 1.87057826e-05, 1.24990208e-05, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33])
```

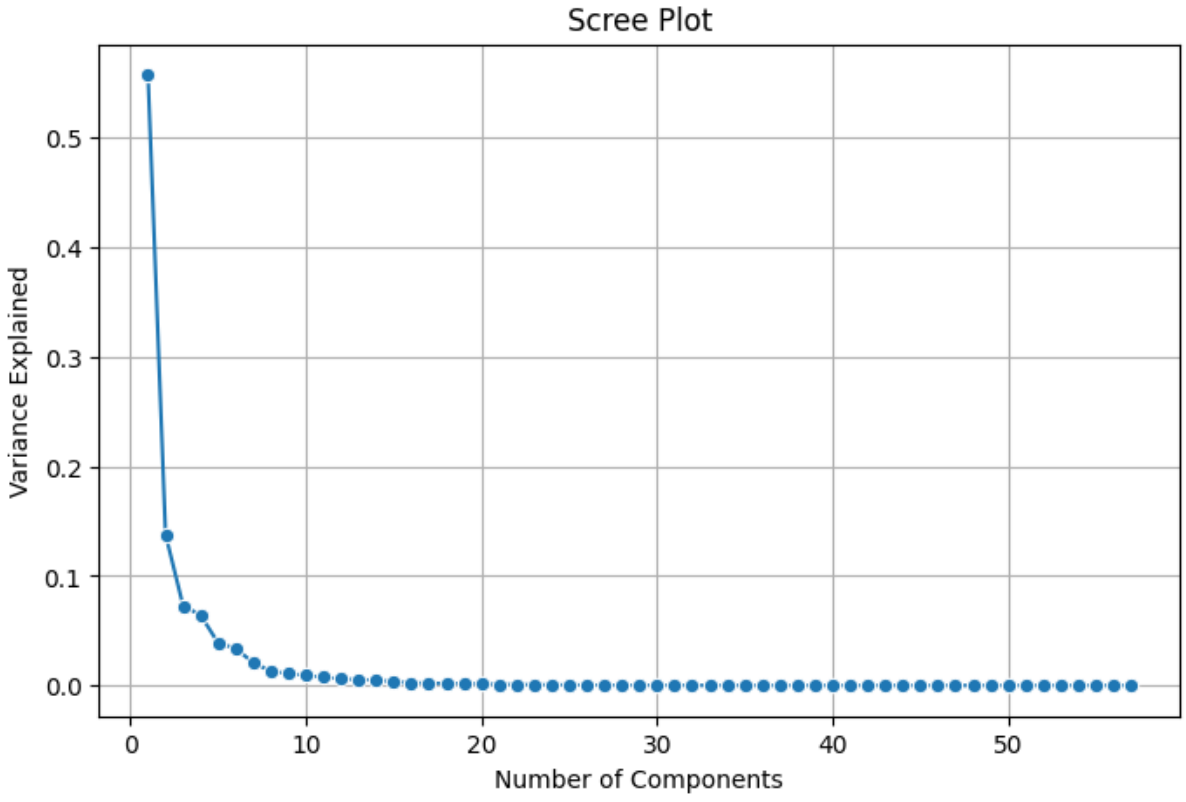
Part 2 - PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.

Explained variance for each PC

Explained variance=(Eigen value of each PC)/(sum of eigen values of all PC's)

We got the explained variance

```
array([5.57260632e-01, 1.37844354e-01, 7.27529548e-02, 6.42641771e-02,
       3.86504944e-02, 3.39516923e-02, 2.06023855e-02, 1.31576386e-02,
       1.08085894e-02, 9.25395468e-03, 7.52911540e-03, 6.19101667e-03,
       5.18772384e-03, 4.92694855e-03, 3.36593119e-03, 2.38692984e-03,
       1.98617593e-03, 1.86206747e-03, 1.70414955e-03, 1.40317638e-03,
       1.00910494e-03, 7.77653131e-04, 6.63717190e-04, 5.19117774e-04,
       4.74341222e-04, 4.10687364e-04, 2.54183814e-04, 1.92422147e-04,
       1.63167083e-04, 1.42503342e-04, 1.38248605e-04, 8.80379297e-05,
       4.55026824e-05, 1.87057826e-05, 1.24990208e-05, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33])
```



We got the optimum number of PC's is n=5

So we are proceeding with the same.

```
array([[ 1.56020579e-01,  1.67117635e-01,  1.65553179e-01,
         1.62192948e-01,  1.62566396e-01,  1.51357849e-01,
         1.51566500e-01,  2.72341946e-02,  2.81833150e-02,
         1.61992837e-01,  1.46872680e-01,  1.61749445e-01,
         1.65248187e-01,  1.59871988e-01,  1.45935804e-01,
         1.46200730e-01,  1.23970284e-01,  1.03127159e-01,
         7.45397856e-02,  1.13355712e-01,  7.38821590e-02,
         1.31572584e-01,  8.33826397e-02,  1.23526242e-01,
         1.11021264e-01,  1.64615479e-01,  1.55395618e-01,
         8.23885414e-02,  4.91953957e-02,  1.28598563e-01,
         1.14305073e-01,  1.40853227e-01,  1.27669598e-01,
         1.55262872e-01,  1.47286584e-01,  1.64971950e-01,
         1.61253433e-01,  1.65501611e-01,  1.55647049e-01,
         9.30142064e-02,  5.15358640e-02,  1.28576116e-01,
         1.10645843e-01,  1.39592763e-01,  1.24545909e-01,
         1.54293786e-01,  1.46285654e-01,  1.50125706e-01,
         1.40157047e-01,  5.25417829e-02,  4.17859530e-02,
         1.21840354e-01,  1.16011410e-01,  1.39868774e-01,
         1.32192245e-01,  1.50375578e-01,  1.31066203e-01],
       [-1.26346525e-01, -8.96765481e-02, -1.04912371e-01,
        -2.20945086e-02, -2.02705495e-02, -4.51109032e-02,
        -5.19237543e-02,  2.76790387e-02,  3.02225550e-02,
        -1.15354767e-01, -1.53109487e-01, -6.62537318e-03,
        -9.10743681e-03, -1.33529221e-01, -8.50869689e-02,
        -1.76368057e-01, -1.51412544e-01,  6.24149874e-02,
         8.64767269e-02, -3.10403498e-02, -5.86880214e-02,
```

```
  9.56339840e-02, -8.62782532e-03, -4.93697036e-02,
  1.98754143e-01,  2.68786906e-01, -1.89867566e-01,
 -2.67767729e-01, -2.12567389e-02, -8.25040484e-02,
  1.11712747e-01,  1.00045670e-01,  6.44232082e-02,
  7.97035639e-02, -2.42054167e-02, -7.20134424e-02,
  1.53517557e-01,  2.56212919e-01, -2.00118572e-01,
 -2.79866019e-01, -2.06182532e-02, -8.27935650e-02,
  1.10285441e-01,  9.56665987e-02,  5.48919412e-02,
  2.39815624e-02,  2.68330072e-01,  2.84955665e-01,
 -1.38627893e-01, -2.02198401e-01, -2.25985194e-02,
 -7.87198691e-02,  1.11827318e-01,  1.02552501e-01],
[-1.25293372e-01, -1.99415702e-02, -7.08726203e-02,
  1.19171727e-02,  1.48442006e-02,  1.24850958e-02,
 -2.98925082e-02, -2.22247412e-01, -2.29754420e-01,
 -3.51625571e-02, -5.95594177e-02,  2.53483369e-02,
 -7.60233572e-02, -4.01544117e-02, -2.25160033e-01,
 -6.83507466e-02, -2.46639865e-01, -8.97686820e-02,
 -2.88964883e-01, -1.36082339e-01, -2.90042169e-01,
  1.52366335e-01,  4.89504701e-02, -4.02891830e-02,
 -1.20391064e-01,  9.30182651e-02, -8.87071351e-02,
 -6.27609057e-02, -1.68401590e-01,  9.17874511e-02,
 -1.06365430e-01,  2.37984720e-01,  1.96320743e-01,
  8.71191185e-02,  2.67292472e-02, -2.55415009e-05,
  3.89358958e-03,  9.28748902e-02, -1.07860188e-01,
```

```
 -1.38627893e-01, -2.02198401e-01, -2.25985194e-02,
 -7.87198691e-02,  1.11827318e-01,  1.02552501e-01],
[-1.25293372e-01, -1.99415702e-02, -7.08726203e-02,
  1.19171727e-02,  1.48442006e-02,  1.24850958e-02,
 -2.98925082e-02, -2.22247412e-01, -2.29754420e-01,
 -3.51625571e-02, -5.95594177e-02,  2.53483369e-02,
 -7.60233572e-02, -4.01544117e-02, -2.25160033e-01,
 -6.83507466e-02, -2.46639865e-01, -8.97686820e-02,
 -2.88964883e-01, -1.36082339e-01, -2.90042169e-01,
  1.52366335e-01,  4.89504701e-02, -4.02891830e-02,
 -1.20391064e-01,  9.30182651e-02, -8.87071351e-02,
 -6.27609057e-02, -1.68401590e-01,  9.17874511e-02,
 -1.06365430e-01,  2.37984720e-01,  1.96320743e-01,
  8.71191185e-02,  2.67292472e-02, -2.55415009e-05,
  3.89358958e-03,  9.28748902e-02, -1.07860188e-01,
 -3.84875764e-02, -1.79691341e-01,  8.04108516e-02,
 -1.36240262e-01,  2.37744958e-01,  1.90510604e-01,
  8.64794097e-02,  2.72754576e-02,  8.74333683e-02,
 -2.22902305e-02, -1.04686027e-01, -1.35715829e-01,
  1.32544187e-01,  4.05131008e-03,  2.30037988e-01,
  2.06200724e-01,  8.48540391e-02,  2.11244737e-02],
[-7.02208129e-03, -3.30261798e-02, -1.28467026e-02,
 -5.02475120e-02, -4.38479686e-02, -1.73006735e-01,
 -1.59803417e-01,  4.33163419e-01,  4.38791922e-01,
 -9.10133100e-03,  5.58436993e-02, -9.65797534e-02,
```

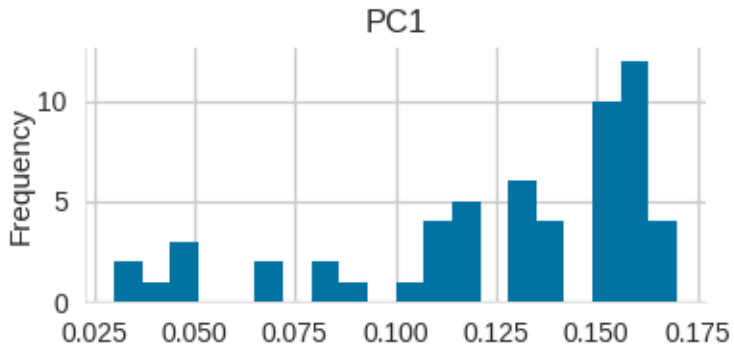
Part 2 - PCA: Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.

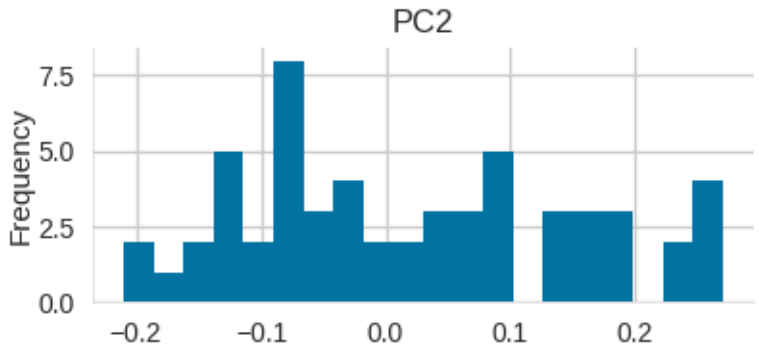
We have taken PC as PC1,PC2,PC3,PC4,PC5

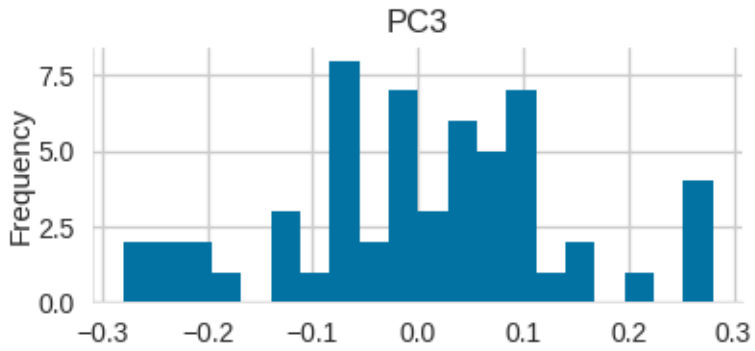
index	PC1	PC2	PC3	PC4	PC5
No_HH	0.15602057858568008	-0.12634652545112005	-0.00269025036769594	-0.12529337156470965	-0.007022081288543556
TOT_M	0.16711763488533543	-0.08967654811143609	0.056697619050839435	-0.019941570190169226	-0.03302617975398989
TOT_F	0.16555317909064954	-0.10491237103791488	0.03874947459668204	-0.07087262026984385	-0.01284670261749123 3
M_06	0.16219294820465557	-0.02209450856947588	0.05778815178389401	0.0119171727334428	-0.05024751195824964
F_06	0.16256639565734846	-0.020270549505351183	0.050125567728648346	0.014844200618405657	-0.04384796856512247
M_SC	0.1513578490906061	-0.04511090318250624	0.0025689039783159708	0.01248509578111028	-0.17300673529926056
F_SC	0.15156650019208903	-0.051923754284057606	-0.025100879567880173	-0.029892508223490172	-0.15980341709190526
M_ST	0.02723419457100404	0.02767903871202621	-0.12350445325241302	-0.2222474123682455	0.4331634192461761
F_ST	0.028183315015872474	0.030222555010887876	-0.13976883253358743	-0.22975441968735436	0.4387919216336268
M_LIT	0.16199283733629224	-0.11535476702476159	0.08216766769670565	-0.035162557126882 67	-0.009101331000967118
F_LIT	0.14687268030140369	-0.15310948711103942	0.11709768332858393	-0.05955941770079014	0.05584369933487008
M_ILL	0.16174944463471633	-0.006625373178056333	-0.021855093414654646	0.025348336918529352	-0.0965797534176452
F_ILL	0.16524818736833372	-0.009107436812887493	-0.09306237626146524	-0.0760233571637385	-0.11991050059368492
TOT_WORK_M	0.15987198816201376	-0.13352922127206826	0.04517636922338625	-0.04015441172477221	-0.019552883798007518
TOT_WORK_F	0.14593580377247664	-0.08508696893837382	-0.05944954569860727	-0.22516003272248072	-0.04043736771003749
MAINWORK_M	0.14620072976306095	-0.1763680574888248	0.0542945289852904	-0.06835074656606742	-0.03680196214818792
MAINWORK_F	0.12397028357273733	-0.15141254378037938	-0.055609096135149645	-0.24663986486161005	-0.08283385942429368
MAIN_CL_M	0.1031271588301983	0.06241498736569892	-0.06739929440338172	-0.08976868196627247	-0.2860390779339572

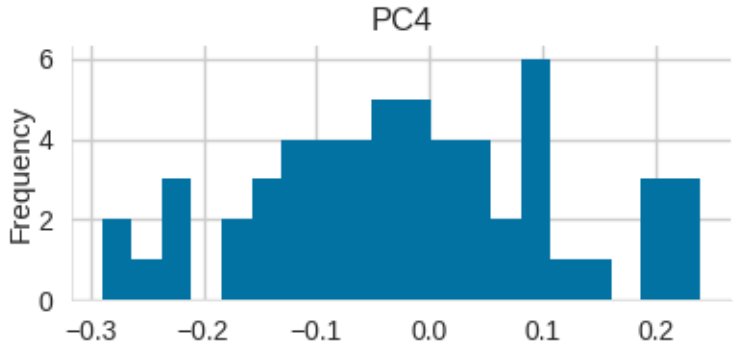
M_ILL	0.16174944463471633	-0.006625373178056333	-0.021855093414654646	0.02534833691852935	
F_ILL	0.16524818736833372	-0.009107436812887493	-0.09306237626146524	-0.076023357163738	
TOT_WORK_M	0.15987198816201376	-0.13352922127206826	0.04517636922338625	-0.04015441172477221	-0.019552883798007518
TOT_WORK_F	0.14593580377247664	-0.08508696893837382	-0.05944954569860727	-0.22516003272248072	-0.04043736771003749
MAINWORK_M	0.14620072976306095	-0.1763680574888248	0.0542945289852904	-0.06835074656606742	-0.03680196214818792
MAINWORK_F	0.12397028357273733	-0.15141254378037938	-0.055609096135149645	-0.24663986486161005	-0.08283385942429368
MAIN_CL_M	0.1031271588301983	0.06241498736569892	-0.06739929440338172	-0.08976868196627247	-0.2860390779339572
MAIN_CL_F	0.0745397855548363	0.08647672694112218	-0.009238089503975205	-0.2889648831930554	-0.24193636807820876
MAIN_AL_M	0.11335571218156744	-0.031040349752566983	-0.2479170553146319	-0.13608233948757903	-0.20572350102112308
MAIN_AL_F	0.07388215903155917	-0.05868802144896378	-0.2519322957756121	-0.29004216853655573	-0.17760476624801932
MAIN_HH_M	0.1315725840227565	-0.07602106774592018	0.026568938525516204	0.1523663351269932	-0.13408883244745878
MAIN_HH_F	0.08338263967435816	-0.08247663752152952	-0.060523299227770486	0.0489504700629877	-0.13944088269530588
MAIN_OT_M	0.12352624192253213	-0.21298425418136235	0.13737788089722278	-0.040289183042922914	0.06463770953094834
MAIN_OT_F	0.11102126391320258	-0.2100711663147429	0.09563398404275447	-0.12039106435535807	0.08074276753313354

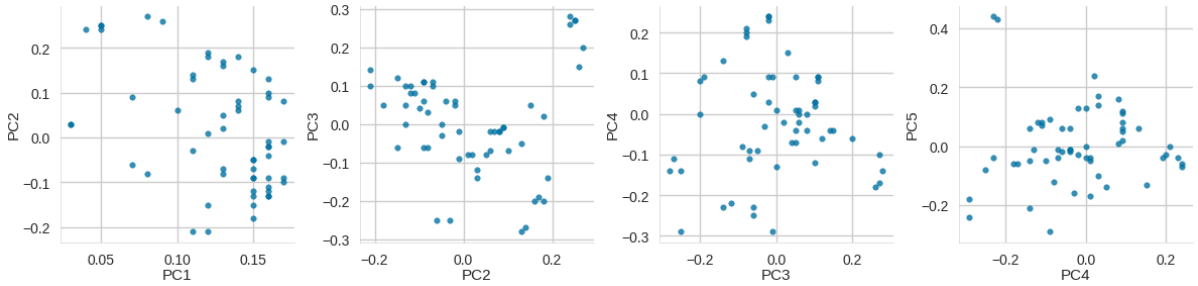
index	PC1	PC2	PC3	PC4	PC5
No_HH	0.16	-0.13	-0.0	-0.13	-0.01
TOT_M	0.17	-0.09	0.06	-0.02	-0.03
TOT_F	0.17	-0.1	0.04	-0.07	-0.01
M_06	0.16	-0.02	0.06	0.01	-0.05
F_06	0.16	-0.02	0.05	0.01	-0.04
M_SC	0.15	-0.05	0.0	0.01	-0.17
F_SC	0.15	-0.05	-0.03	-0.03	-0.16
M_ST	0.03	0.03	-0.12	-0.22	0.43
F_ST	0.03	0.03	-0.14	-0.23	0.44
M_LIT	0.16	-0.12	0.08	-0.04	-0.01
F_LIT	0.15	-0.15	0.12	-0.06	0.06
M_ILL	0.16	-0.01	-0.02	0.03	-0.1
F_ILL	0.17	-0.01	-0.09	-0.08	-0.12
TOT_WORK_M	0.16	-0.13	0.05	-0.04	-0.02
TOT_WORK_F	0.15	-0.09	-0.06	-0.23	-0.04
MAINWORK_M	0.15	-0.18	0.05	-0.07	-0.04
MAINWORK_F	0.12	-0.15	-0.06	-0.25	-0.08





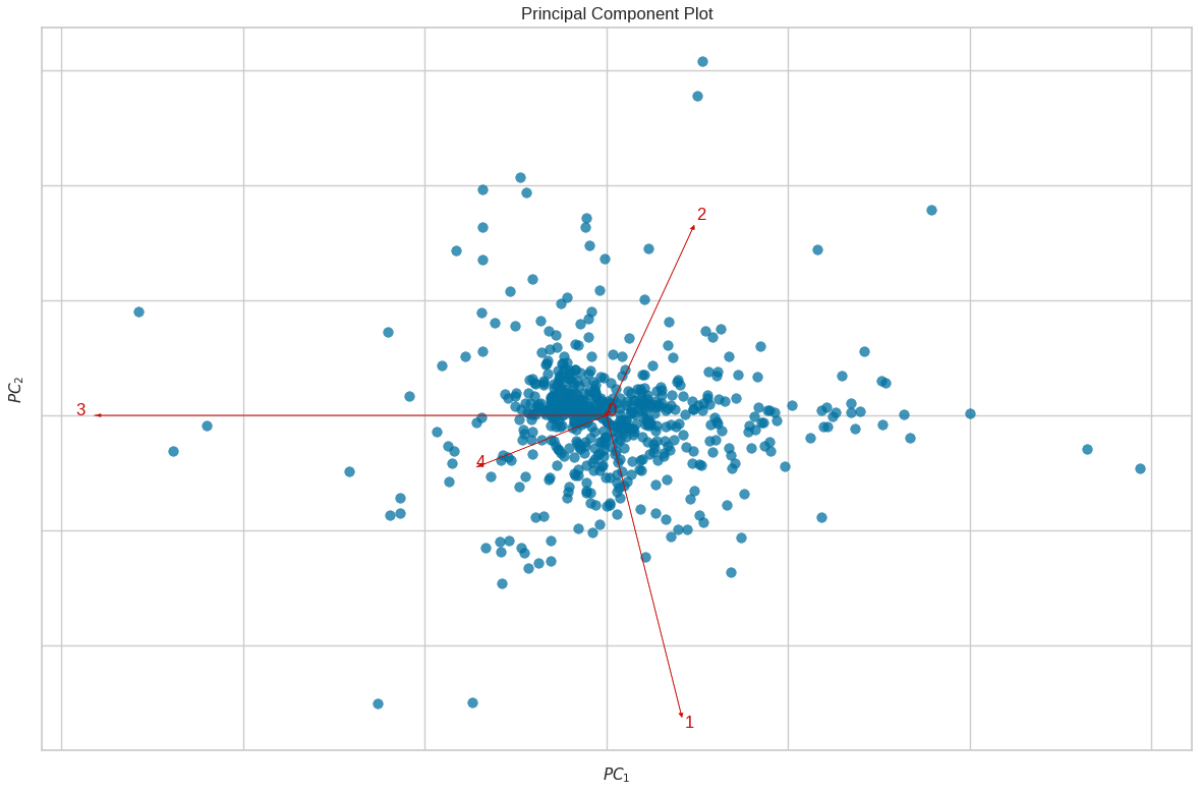

PC3
Frequency
7.5
5.0
2.5
0.0
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3


PC4
Frequency
6
4
2
0
−0.3
−0.2
−0.1
0.0
0.1
0.2


PC2
PC1
PC3
PC2
PC4
PC3
PC5
PC4

PC1 has most variance

	PC1	PC2	PC3	PC4	PC5
No_HH	0.160000	-0.130000	-0.000000	-0.130000	-0.010000
TOT_M	0.170000	-0.090000	0.060000	-0.020000	-0.030000
TOT_F	0.170000	-0.100000	0.040000	-0.070000	-0.010000
M_06	0.160000	-0.020000	0.060000	0.010000	-0.050000
F_06	0.160000	-0.020000	0.050000	0.010000	-0.040000
M_SC	0.150000	-0.050000	0.000000	0.010000	-0.170000
F_SC	0.150000	-0.050000	-0.030000	-0.030000	-0.160000
M_ST	0.030000	0.030000	-0.120000	-0.220000	0.430000
F_ST	0.030000	0.030000	-0.140000	-0.230000	0.440000
M_LIT	0.160000	-0.120000	0.080000	-0.040000	-0.010000
F_LIT	0.150000	-0.150000	0.120000	-0.060000	0.060000
M_ILL	0.160000	-0.010000	-0.020000	0.030000	-0.100000
F_ILL	0.170000	-0.010000	-0.090000	-0.080000	-0.120000
TOT_WORK_M	0.160000	-0.130000	0.050000	-0.040000	-0.020000
TOT_WORK_F	0.150000	-0.090000	-0.060000	-0.230000	-0.040000
MAINWORK_M	0.150000	-0.180000	0.050000	-0.070000	-0.040000
MAINWORK_F	0.120000	-0.150000	-0.060000	-0.250000	-0.080000
MAIN_CL_M	0.100000	0.060000	-0.070000	-0.090000	-0.290000
MAIN_CL_F	0.070000	0.090000	-0.010000	-0.290000	-0.240000
MAIN_AL_M	0.110000	-0.030000	-0.250000	-0.140000	-0.210000
MAIN_AL_F	0.070000	-0.060000	-0.250000	-0.290000	-0.180000
MAIN_HH_M	0.130000	-0.080000	0.030000	0.150000	-0.130000
MAIN_HH_F	0.080000	-0.080000	-0.060000	0.050000	-0.140000
MAIN_OT_M	0.120000	-0.210000	0.140000	-0.040000	0.060000
MAIN_OT_F	0.110000	-0.210000	0.100000	-0.120000	0.080000
MARGWORK_M	0.160000	0.090000	-0.010000	0.090000	0.060000
MARGWORK_F	0.160000	0.130000	-0.050000	-0.090000	0.090000
MARG_CL_M	0.080000	0.270000	0.200000	-0.060000	-0.020000
MARG_CL_F	0.050000	0.250000	0.270000	-0.170000	-0.060000
MARG_AL_M	0.130000	0.170000	-0.190000	0.090000	0.020000
MARG_AL_F	0.110000	0.140000	-0.270000	-0.110000	0.080000
MARG_HH_M	0.140000	0.070000	-0.020000	0.240000	-0.060000
MARG_HH_F	0.130000	0.020000	-0.080000	0.200000	-0.030000
MARG_OT_M	0.160000	-0.090000	0.110000	0.090000	0.120000
MARG_OT_F	0.150000	-0.120000	0.100000	0.030000	0.170000
MARGWORK_3_6_M	0.160000	-0.040000	0.060000	-0.000000	-0.040000
MARGWORK_3_6_F	0.160000	-0.110000	0.080000	0.000000	0.000000
MARG_CL_3_6_M	0.170000	0.080000	-0.020000	0.090000	0.050000
MARG_CL_3_6_F	0.160000	0.100000	-0.070000	-0.110000	0.070000



Part 2 - PCA: Write linear equation for first PC.

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_ST (+) 0.1