DATA MINING PROJECT

A PROJECT REPORT

Submitted by

BALAJI S (PGP-DSBA-JUNE 2023 TO JUNE2024)

Problem Statement:

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.

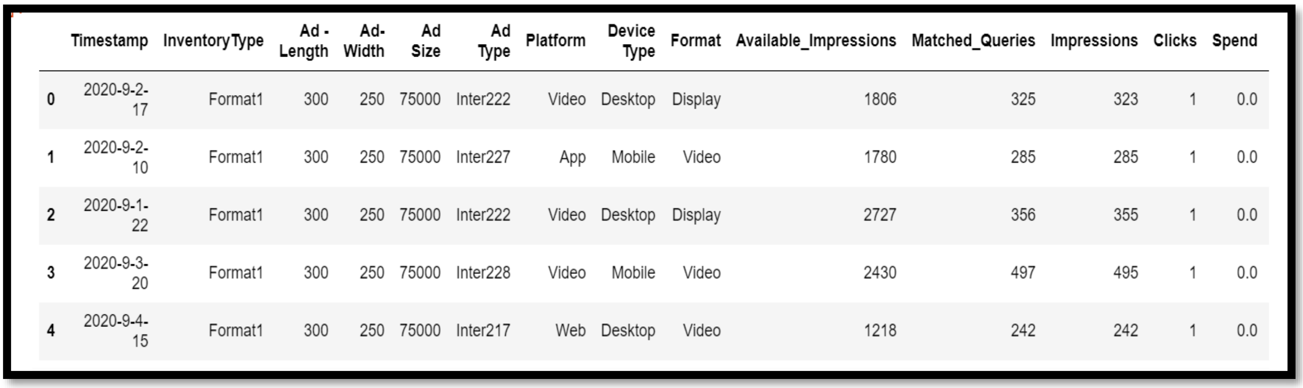
Perform the following in given order:

- Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.
- Treat missing values in CPC, CTR and CPM using the formula given. You may refer to the Bank_KMeans Solution File to understand the coding behind treating the missing values using a specific formula. You have to basically create an user defined function and then call the function for imputing.
- 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).
- Perform z-score scaling and discuss how it affects the speed of the algorithm.
- Perform clustering and do the following:
- Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.
- Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.
- Print silhouette scores for up to 10 clusters and identify optimum number of clusters.
- 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.]
- Conclude the project by providing summary of your learnings.

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.

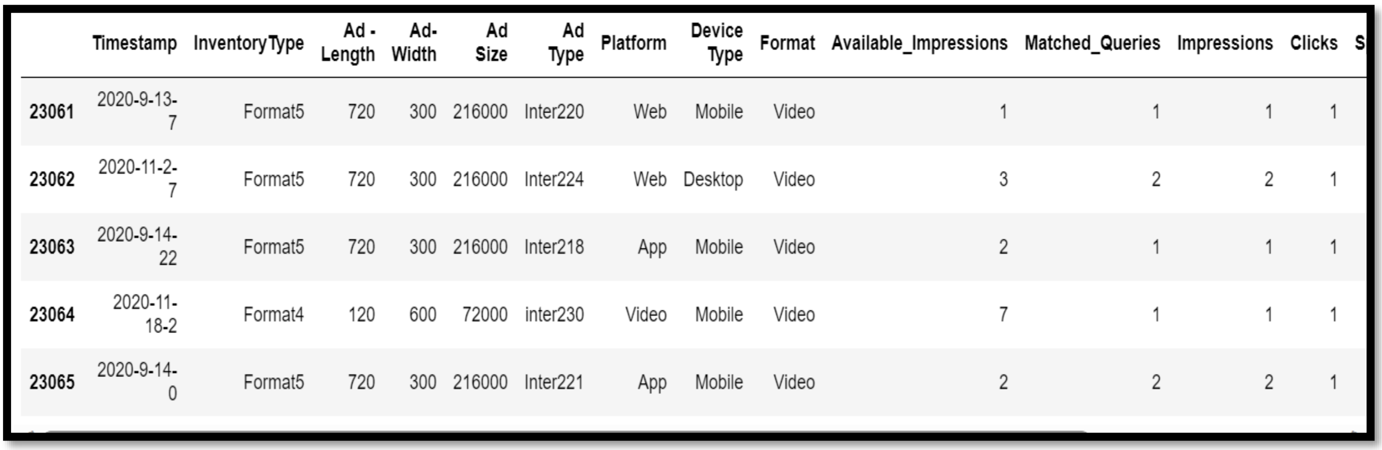
Ans:Dataset is loaded into the Dataframe

Print top 5



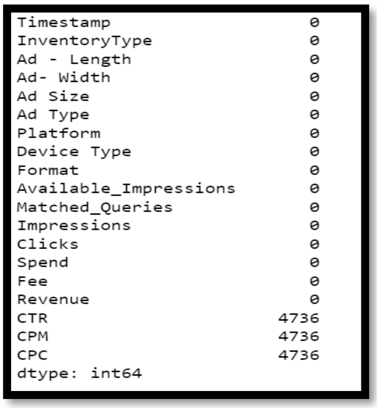
	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.0
2	2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0
3	2020-9-3-20	Format1	300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.0
4	2020-9-4-15	Format1	300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.0

Print last 5

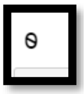


	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	S
23061	2020-9-13-7	Format5	720	300	216000	Inter220	Web	Mobile	Video	1	1	1	1	
23062	2020-11-2-7	Format5	720	300	216000	Inter224	Web	Desktop	Video	3	2	2	1	
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	

Checking NULL & Duplicate 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
```



```
0
```

NULL VALUES **DUPLICATED VALUES**

Q2) Treat missing values in CPC, CTR and CPM using the formula given.

Ans:

We created three functions such as ‘calculate_CPC’, ‘calculate_CTR’, and ‘calculate_CPM’ to treatmissing values in CPC, CTR, and CPM columns using the following formula.

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

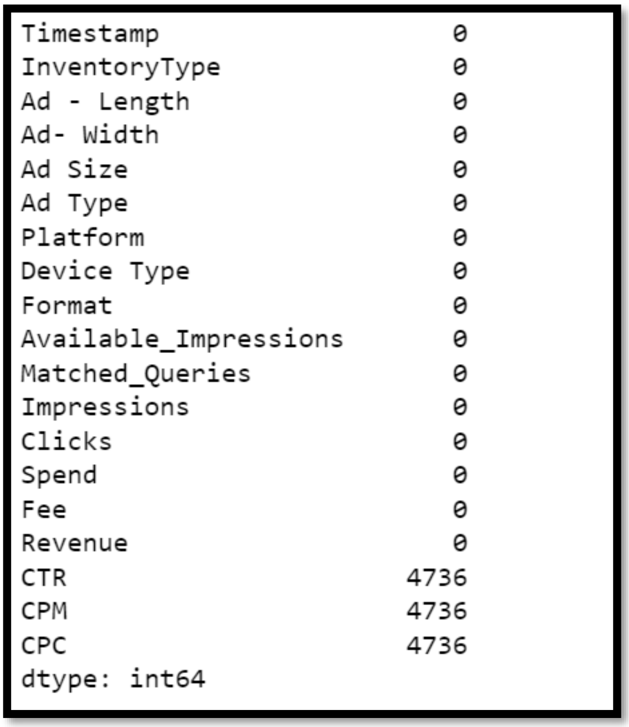
Note that the Total Campaign Spend refers to the 'Spend' Column in the dataset and the Number of Impressions refers to the 'Impressions' Column in the dataset.

CPC = Total Cost (spend) / Number of Clicks.

Note that the Total Cost (spend) refers to the 'Spend' Column in the dataset and the Number ofClicks refers to the 'Clicks' Column in the dataset.

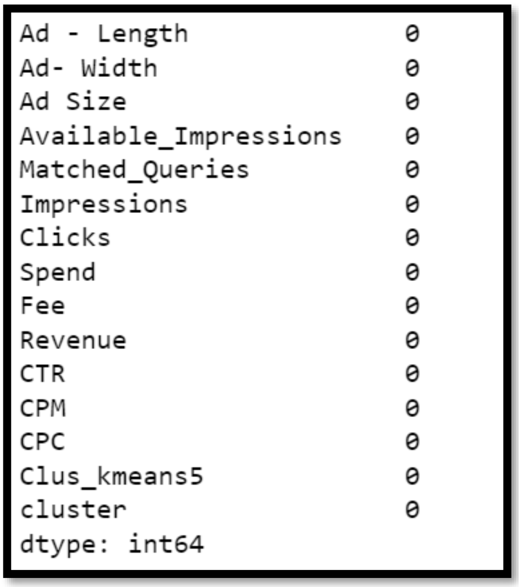
CTR = Total Measured Clicks / Total Measured Ad Impressions x 100.

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 database



```
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
```

BEFORE TREATING MISSING VALUES



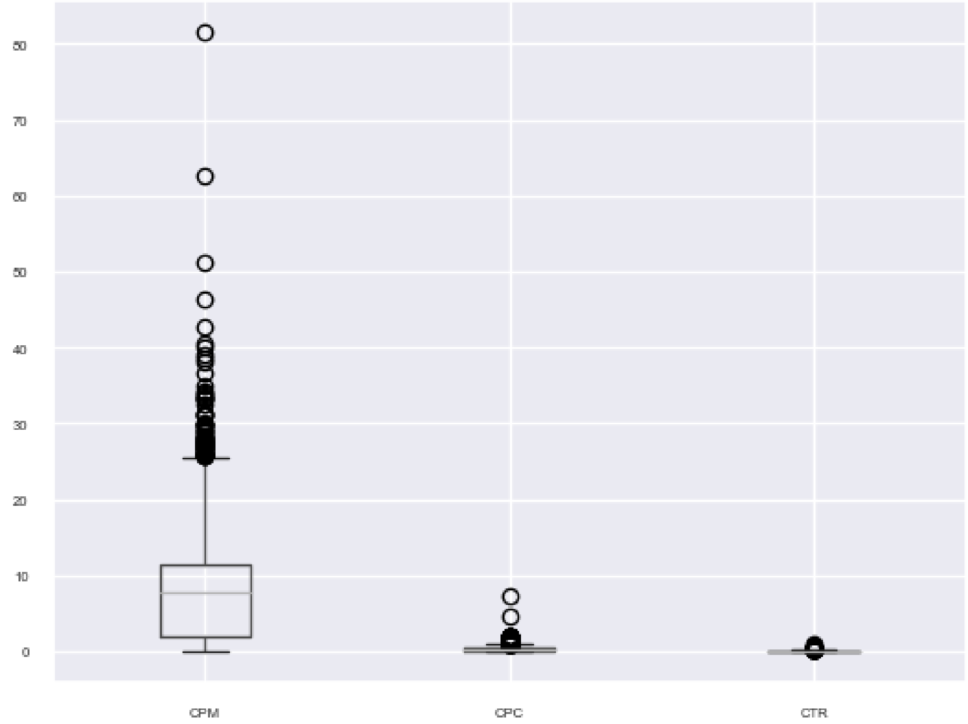
```
Ad - Length              0
Ad- Width                0
Ad Size                  0
Available_Impressions    0
Matched_Queries          0
Impressions              0
Clicks                   0
Spend                    0
Fee                      0
Revenue                  0
CTR                      0
CPM                      0
CPC                      0
Clus_kmeans5             0
cluster                  0
dtype: int64
```

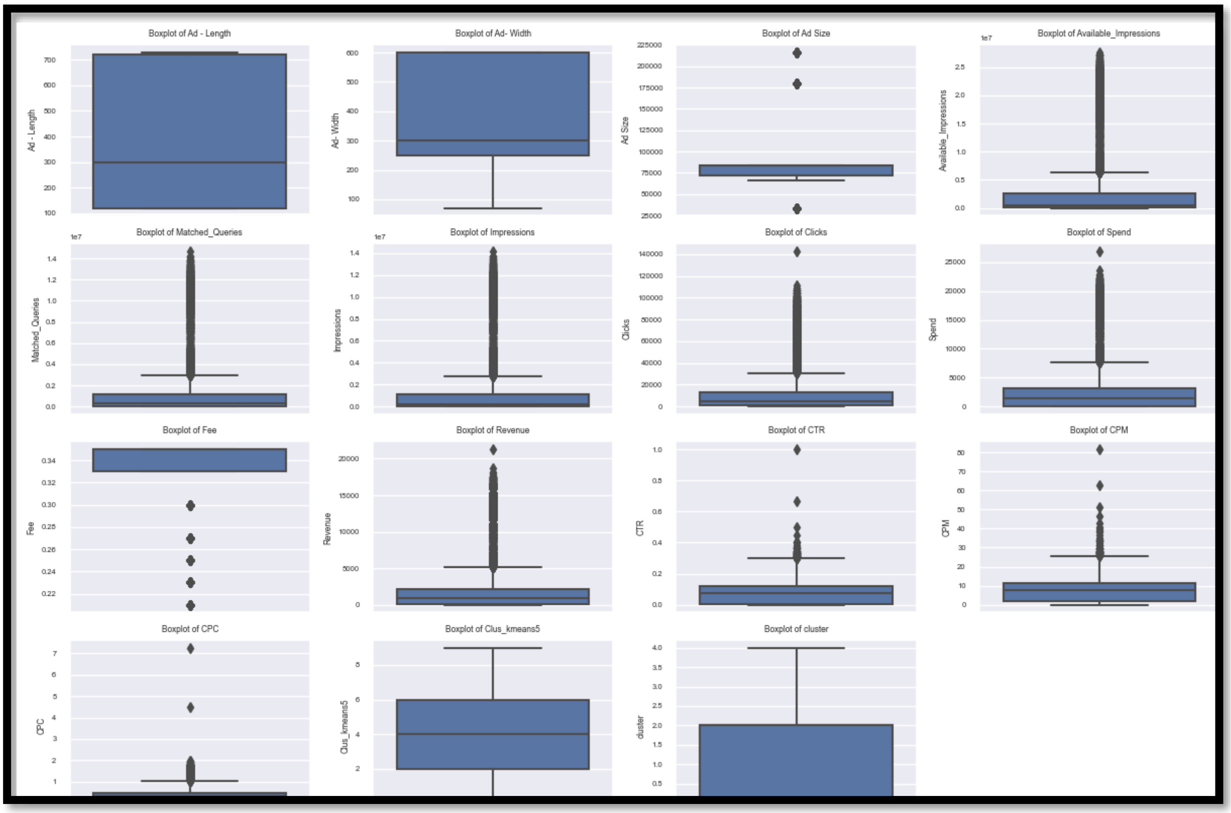
AFTER TREATING THE MISSING VALUES

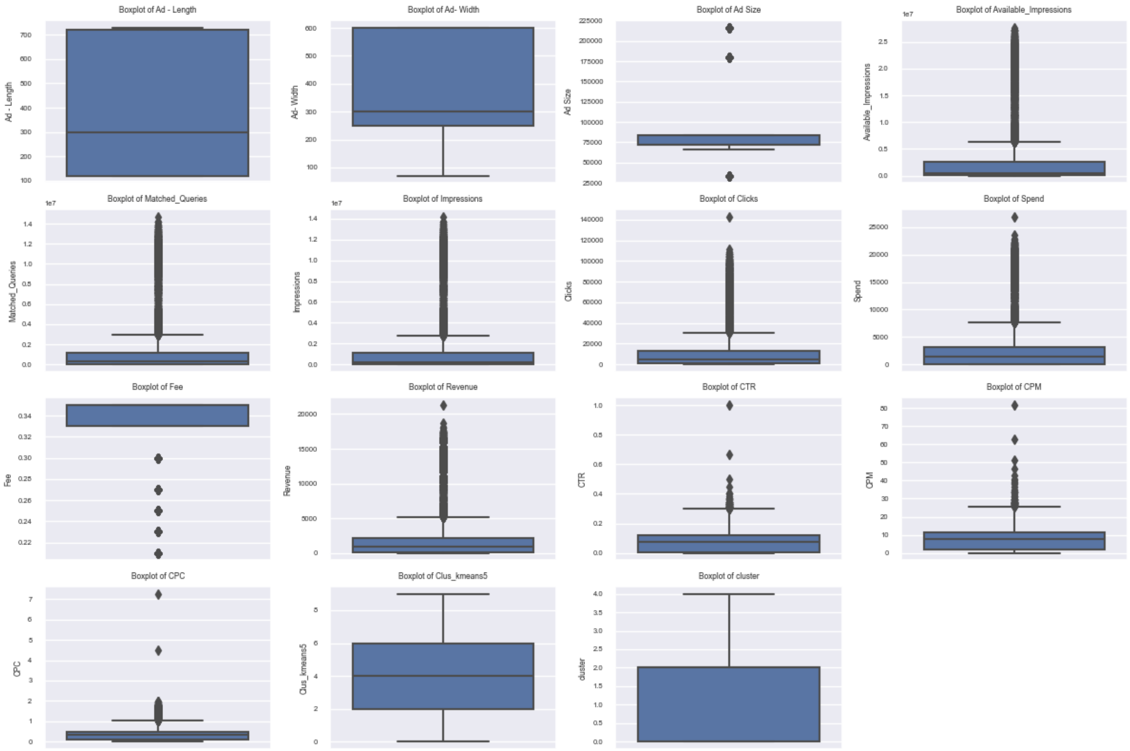
Q.3) Check if there are any outliers. Do you think treating outliers is necessary for K-Meansclustering? Based on your judgement decide whether to treat outliers and if yes, which method toemploy. (As an analyst your judgement may be different from another analyst).

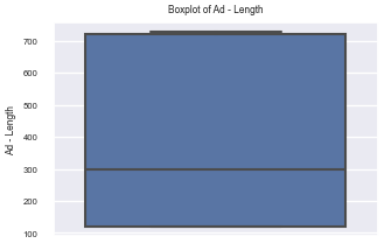
Ans:

I have checked with the data and it seems that there are Outliers. Below is the Boxplot figure of Features before Treating Outliers









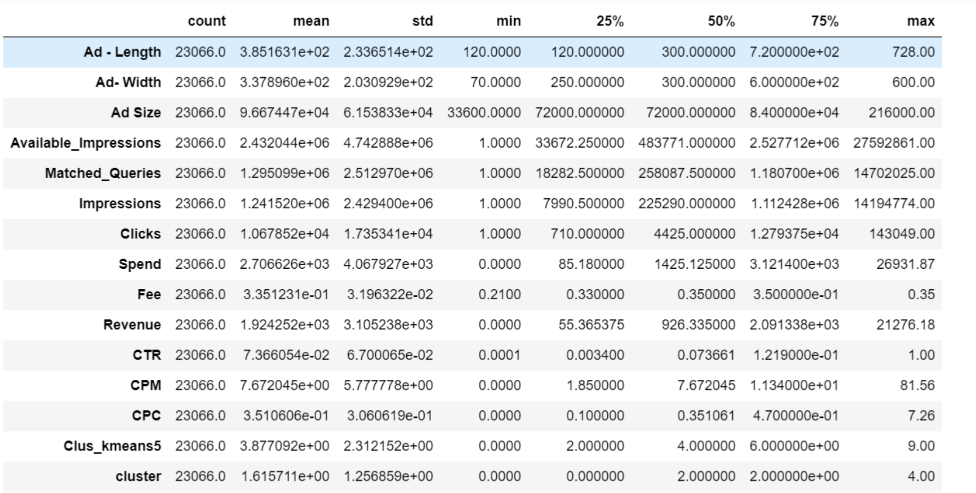
TREATING OUTLIERS:

It depends on the specific case and the domain knowledge. If the outliers are caused by errors in data collection or data entry, then it may be necessary to remove them. If the outliers are caused by actual extreme values in the data, then it may be necessary to keep them.

Here by I conclude that these outliers are caused by actual extreme values in data.

Q.4) Perform z-score scaling and discuss how it affects the speed of the algorithm.

Ans:

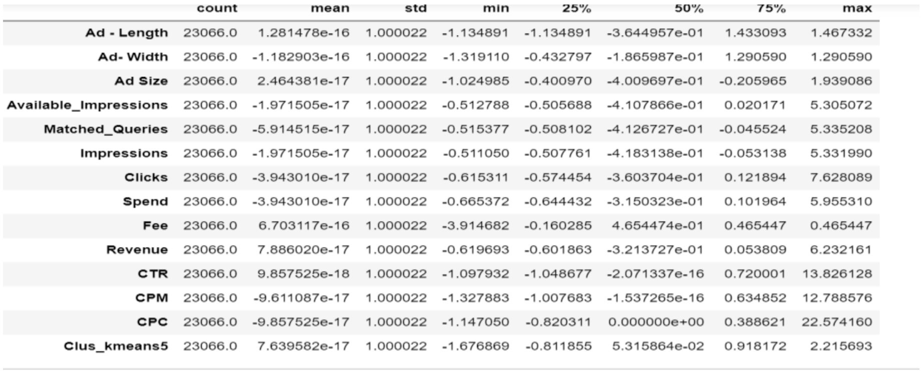


	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	3.851631e+02	2.336514e+02	120.0000	120.000000	300.000000	7.200000e+02	728.00
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.0000	250.000000	300.000000	6.000000e+02	600.00
Ad Size	23066.0	9.667447e+04	6.153833e+04	33600.0000	72000.000000	72000.000000	8.400000e+04	216000.00
Available_Impressions	23066.0	2.432044e+06	4.742888e+06	1.0000	33672.250000	483771.000000	2.527712e+06	27592861.00
Matched_Queries	23066.0	1.295099e+06	2.512970e+06	1.0000	18282.500000	258087.500000	1.180700e+06	14702025.00
Impressions	23066.0	1.241520e+06	2.429400e+06	1.0000	7990.500000	225290.000000	1.112428e+06	14194774.00
Clicks	23066.0	1.067852e+04	1.735341e+04	1.0000	710.000000	4425.000000	1.279375e+04	143049.00
Spend	23066.0	2.706626e+03	4.067927e+03	0.0000	85.180000	1425.125000	3.121400e+03	26931.87
Fee	23066.0	3.351231e-01	3.196322e-02	0.2100	0.330000	0.350000	3.500000e-01	0.35
Revenue	23066.0	1.924252e+03	3.105238e+03	0.0000	55.365375	926.335000	2.091338e+03	21276.18
CTR	23066.0	7.366054e-02	6.700065e-02	0.0001	0.003400	0.073661	1.219000e-01	1.00
CPM	23066.0	7.672045e+00	5.777778e+00	0.0000	1.850000	7.672045	1.134000e+01	81.56
CPC	23066.0	3.510606e-01	3.060619e-01	0.0000	0.100000	0.351061	4.700000e-01	7.26
Clus_kmeans5	23066.0	3.877092e+00	2.312152e+00	0.0000	2.000000	4.000000	6.000000e+00	9.00
cluster	23066.0	1.615711e+00	1.256859e+00	0.0000	0.000000	2.000000	2.000000e+00	4.00

BEFORE SCALING

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

AFTER SCALING:



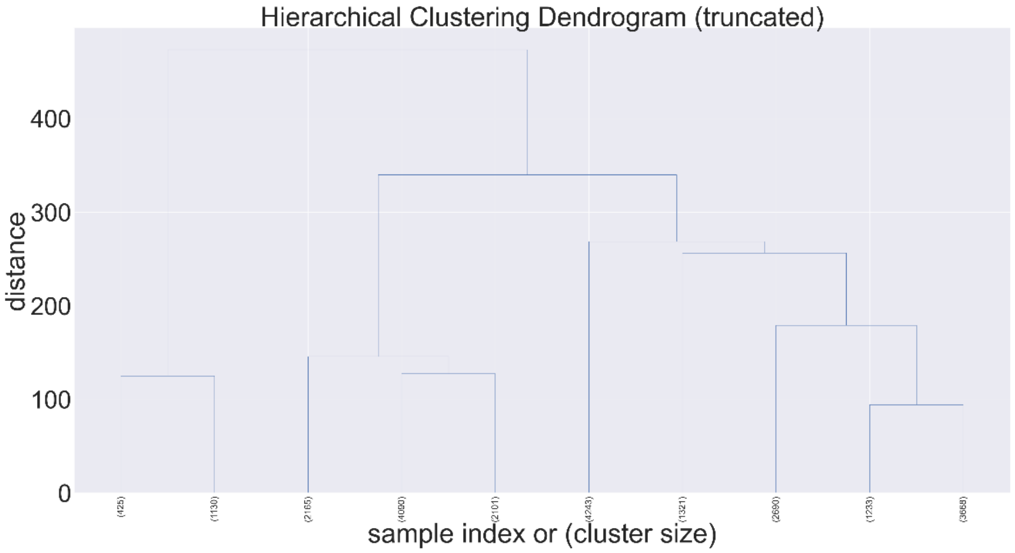
	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	1.281478e-16	1.000022	-1.134891	-1.134891	-3.644957e-01	1.433093	1.467332
Ad- Width	23066.0	-1.182903e-16	1.000022	-1.319110	-0.432797	-1.865987e-01	1.290590	1.290590
Ad Size	23066.0	2.464381e-17	1.000022	-1.024985	-0.400970	-4.009697e-01	-0.205965	1.939086
Available_Impressions	23066.0	-1.971505e-17	1.000022	-0.512788	-0.505688	-4.107866e-01	0.020171	5.305072
Matched_Queries	23066.0	-5.914515e-17	1.000022	-0.515377	-0.508102	-4.126727e-01	-0.045524	5.335208
Impressions	23066.0	-1.971505e-17	1.000022	-0.511050	-0.507761	-4.183138e-01	-0.053138	5.331990
Clicks	23066.0	-3.943010e-17	1.000022	-0.615311	-0.574454	-3.603704e-01	0.121894	7.628089
Spend	23066.0	-3.943010e-17	1.000022	-0.665372	-0.644432	-3.150323e-01	0.101964	5.955310
Fee	23066.0	6.703117e-16	1.000022	-3.914682	-0.160285	4.654474e-01	0.465447	0.465447
Revenue	23066.0	7.886020e-17	1.000022	-0.619693	-0.601863	-3.213727e-01	0.053809	6.232161
CTR	23066.0	9.857525e-18	1.000022	-1.097932	-1.048677	-2.071337e-16	0.720001	13.826128
CPM	23066.0	-9.611087e-17	1.000022	-1.327883	-1.007683	-1.537265e-16	0.634852	12.788576
CPC	23066.0	-9.857525e-17	1.000022	-1.147050	-0.820311	0.000000e+00	0.388621	22.574160
Clus_kmeans5	23066.0	7.639582e-17	1.000022	-1.676869	-0.811855	5.315864e-02	0.918172	2.215693

Q.5) Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.

Ans:

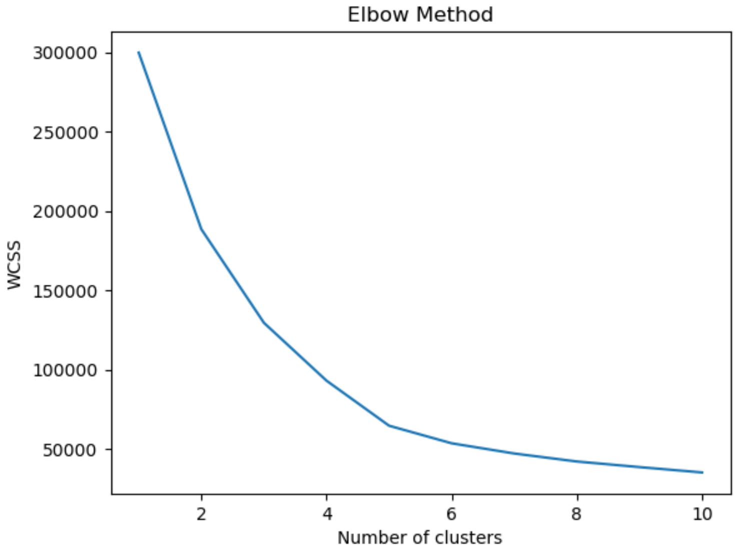
Please find below Dendrogram performed for Hierarchical using WARD and Euclidean Distance on the Scaled Data such as “data_scaled”.

In this Dendrogram, value of P = 10, which means that only the last 10 merged clusters are show

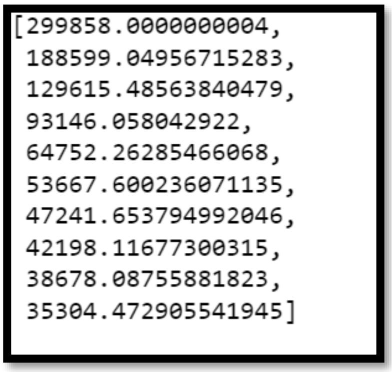


Q.6) Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.

Ans: Elbow Plot (up to n=10)



For checking the Optimal number of clusters



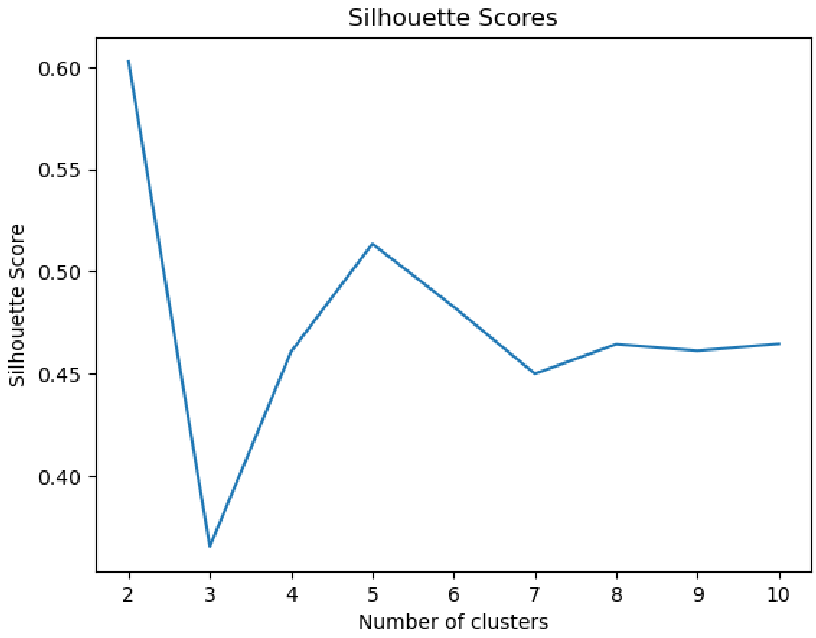
```
[299858.0000000004,
 188599.04956715283,
 129615.48563840479,
 93146.058042922,
 64752.26285466068,
 53667.600236071135,
 47241.653794992046,
 42198.11677300315,
 38678.08755881823,
 35304.472905541945]
```

WSS

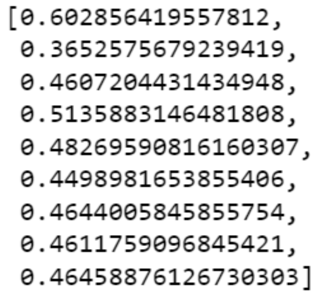
we use WSS (Within Sum Of Square)As per the check When we move from K=1 to K=2, We see that there is a significant drop in the value.

Also when we move from k=2 to k-3, k=3 to k=4, k=4 to k=5 there is a significant drop As well's k =5 to k=6, the drop in values reduces significantly. Hence, In this case, the WSS is not significantly dropping beyond 5, so 5 is an optimal number of clusters.

Q7) Print silhouette scores for up to 10 clusters and identify an optimum number of clusters.



Silhouette scores for up to 10 clusters:



```
[0.602856419557812,
 0.3652575679239419,
 0.4607204431434948,
 0.5135883146481808,
 0.48269590816160307,
 0.449898165385540б,
 0.4644005845855754,
 0.4611759096845421,
 0.46458876126730303]
```

Optimal number of clusters: 5

I have calculated Silhouette Score for scaled data using the silhouette score() function.

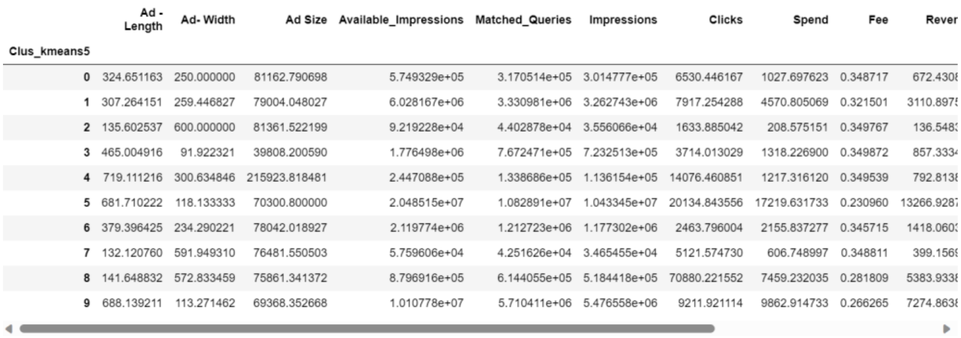
The Silhouette Score is a measure of how similar an object is to its own cluster compared to other clusters, and it ranges from -1 to 1, with higher values indicating better clustering.

As per the Elbow plot/scree plot, we concluded that the optimal number of clusters should be 5. Because 2 would be very less number of clusters.

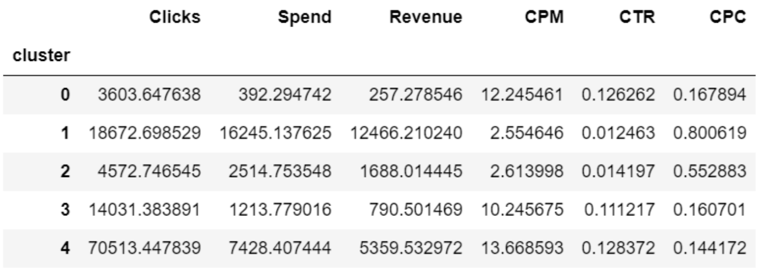
Q.8) Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots].

Ans:

- I have performed K-Means Clustering on scaled data.
- The K-Means function from sci-kit-learn is used to create a K-Means object with n-clusters=5(i.e., 5 clusters).
- Created clusters for the Ads based on the optimum number of clusters using silhouette score
- The group by method from Pandas is used to group the data by the K-Means cluster labels, and the mean, are used to compute the mean of each feature within each cluster. The resulting data frames are stored in the variables mean.

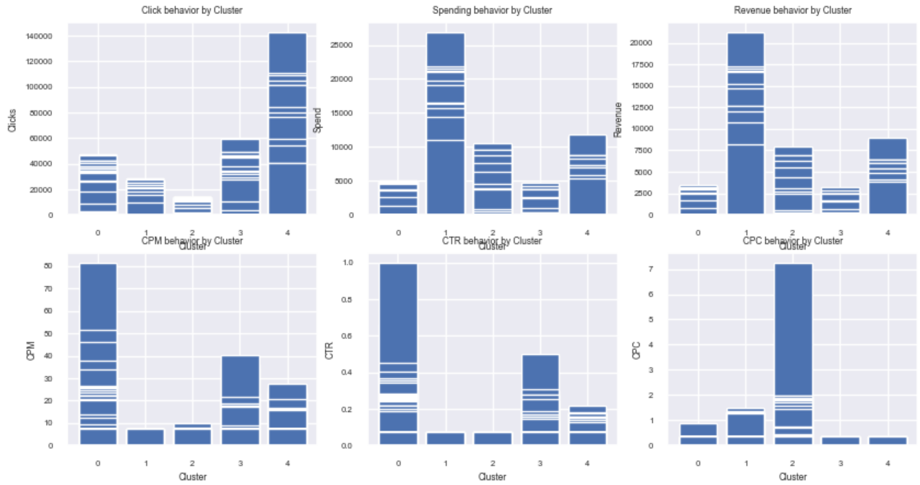


	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Rever
Clus_kmeans5										
0	324.651163	250.000000	81162.790698	5.749329e+05	3.170514e+05	3.014777e+05	6530.446167	1027.697623	0.348717	672.4308
1	307.264151	259.446827	79004.048027	6.028167e+06	3.330981e+06	3.262743e+06	7917.254288	4570.805069	0.321501	3110.8975
2	135.602537	600.000000	81361.522199	9.219228e+04	4.402878e+04	3.556066e+04	1633.885042	208.575151	0.349767	136.5483
3	465.004916	91.922321	39808.200590	1.776498e+06	7.672471e+05	7.232513e+05	3714.013029	1318.226900	0.349872	857.3334
4	719.111216	300.634846	215923.818481	2.447088e+05	1.338686e+05	1.136154e+05	14076.460851	1217.316120	0.349539	792.8138
5	681.710222	118.133333	70300.800000	2.048515e+07	1.082891e+07	1.043345e+07	20134.843556	17219.631733	0.230960	13266.9287
6	379.396425	234.290221	78042.018927	2.119774e+06	1.212723e+06	1.177302e+06	2463.796004	2155.837277	0.345715	1418.0603
7	132.120760	591.949310	76481.550503	5.759606e+04	4.251626e+04	3.465455e+04	5121.574730	606.748997	0.348811	399.1569
8	141.648832	572.833459	75861.341372	8.796916e+05	6.144055e+05	5.184418e+05	70880.221552	7459.232035	0.281809	5383.9338
9	688.139211	113.271462	69368.352668	1.010778e+07	5.710411e+06	5.476558e+06	9211.921114	9862.914733	0.266265	7274.8638



	Clicks	Spend	Revenue	CPM	CTR	CPC
cluster						
0	3603.647638	392.294742	257.278546	12.245461	0.126262	0.167894
1	18672.698529	16245.137625	12466.210240	2.554646	0.012463	0.800619
2	4572.746545	2514.753548	1688.014445	2.613998	0.014197	0.552883
3	14031.383891	1213.779016	790.501469	10.245675	0.111217	0.160701
4	70513.447839	7428.407444	5359.532972	13.668593	0.128372	0.144172

Mean values of clustering data



GRAPH OF CLUSTERING DATA

Q.9) Conclude the project by providing summary of your learnings

- The dataset has 25857 rows and 19 columns.
- The missing values in CPC, CTR and CPM are treated by using the formulae given and writing a user-defined function, and calling it.
- We check for outliers; we can see there are outliers in the variables.
- The dendrogram is the visualization and linkage is for computing the distances and merging the clusters from n to 1.
- The output of Linkage is visualised by Dendrogram.
- We will create linkage using Ward's method and run linkage function on the usable columns of the data.

- The linkage now stores the various distance at which the n clusters are sequentially merged into a single cluster.
- Using this array, we can now perform k-means
- The one requirement before we run the k-means algorithm is to know how many clusters we require as output
- We map the elbow plot using wss values
- From the plot we have the following observations:
- When we move from k=1 to k=2, we see that there is a significant drop in the value, also when we move k=2 to k=3, k=3 to k=4 there is a significant drop as well.
- But from k=4 to k=5, k=5 to k=6, the drop in values reduces significantly other words, the wss is not significantly dropping beyond 5,
- So, 5 is the optimal number of clusters

CONCLUSION AFTER CLUSTERING:

- In this project, we used clustering techniques to segment digital ads data into homogeneous groups based on the features of CPM, CPC and CTR.

- We first performed basic data analysis, treated missing values, checked for outliers and scaled the data using z-score scaling.

- We then performed hierarchical clustering and used the elbow plot
- and silhouette scores to identify the optimum number of clusters.
- When Clicking on Ads gets increases then Revenue is also increases.
- When the amount of money spent on specific ad variations within a specific campaign or ad set increases then Revenue also increases.
- When the impression count of a particular Advertisement increases then Revenue also increases

- Finally, we profiled the ads based on the optimum number of clusters and identified trends in clicks, spend, revenue, CPM, CTR and CPC based on Device Type.

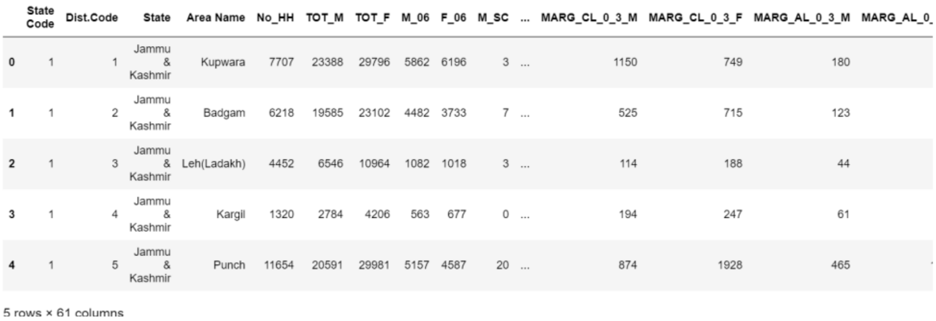
PCA:

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.

Q.1) Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.

Ans:

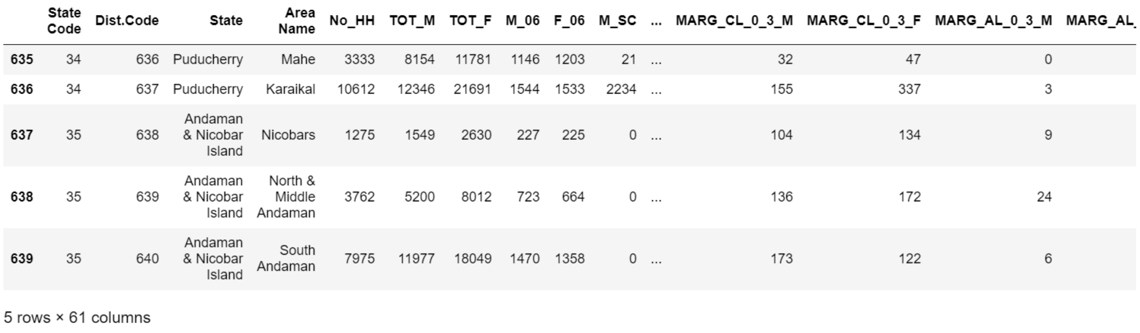
PCA India Data Census.xlsxdataset is loaded into the dataframe.Data Frame printing rows with Head (Prints top 5 rows) function as below :



	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

HEAD

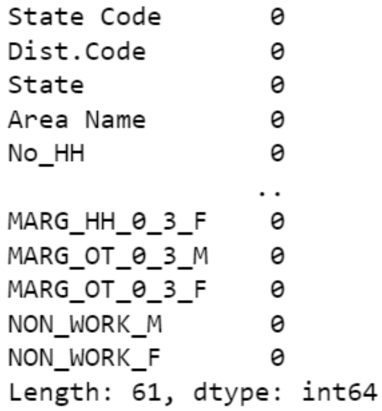


	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_
635	34	636	Puducherry	Mahe	3333	8154	11781	1146	1203	21	...	32	47	0	
636	34	637	Puducherry	Karaikal	10612	12346	21691	1544	1533	2234	...	155	337	3	
637	35	638	Andaman & Nicobar Island	Nicobars	1275	1549	2630	227	225	0	...	104	134	9	
638	35	639	Andaman & Nicobar Island	North & Middle Andaman	3762	5200	8012	723	664	0	...	136	172	24	
639	35	640	Andaman & Nicobar Island	South Andaman	7975	11977	18049	1470	1358	0	...	173	122	6	

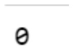
5 rows × 61 columns

TAIL

CHECKING NULL AND DUPLICATE 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
```



```
0
```

NULL VALUES **DUPLICATED VALUES**

```
<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
```

INFORMATION ABOUT THE DATA SET

Q.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_

ANS:

I have picked 5 Variables such as 'TOT_M', 'TOT_F','M_LIT','F_LIT', and , 'TOT_WORK_M'. And comparing those 5 variables against 'State' and 'Dist.Code'

.TOT_M - Total population Male

TOT_F -Total population Female

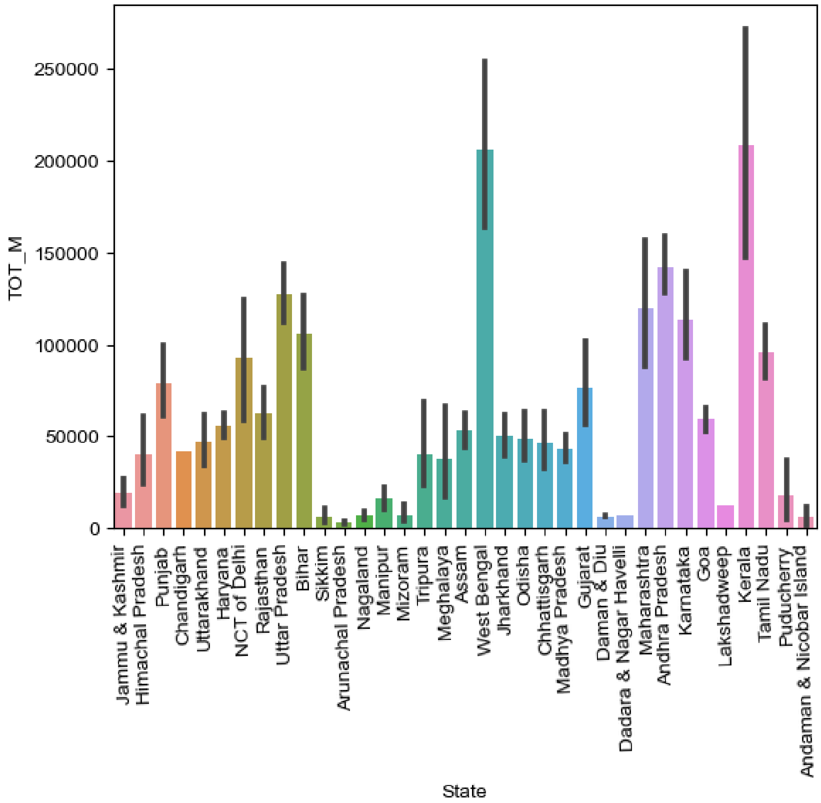
M_LIT - Literates population Male

F_LIT -Literates population Female

TOT_WORK_M -Total Worker Population Male

State -State

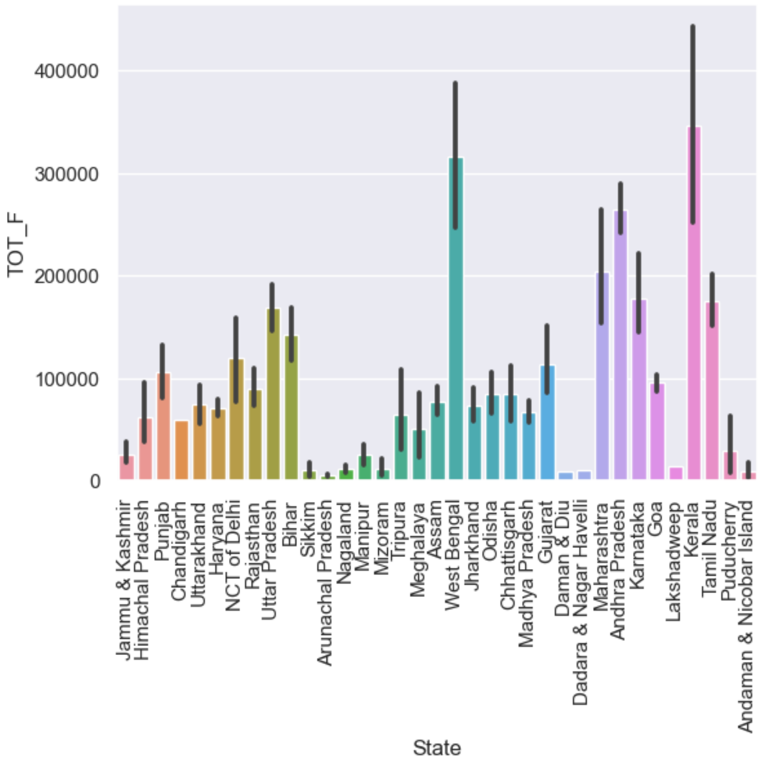
District- District code



By Above Bar plot (State, TOT_M), We can get the following Questions such as:

1) Which state has the highest Total population of Males?

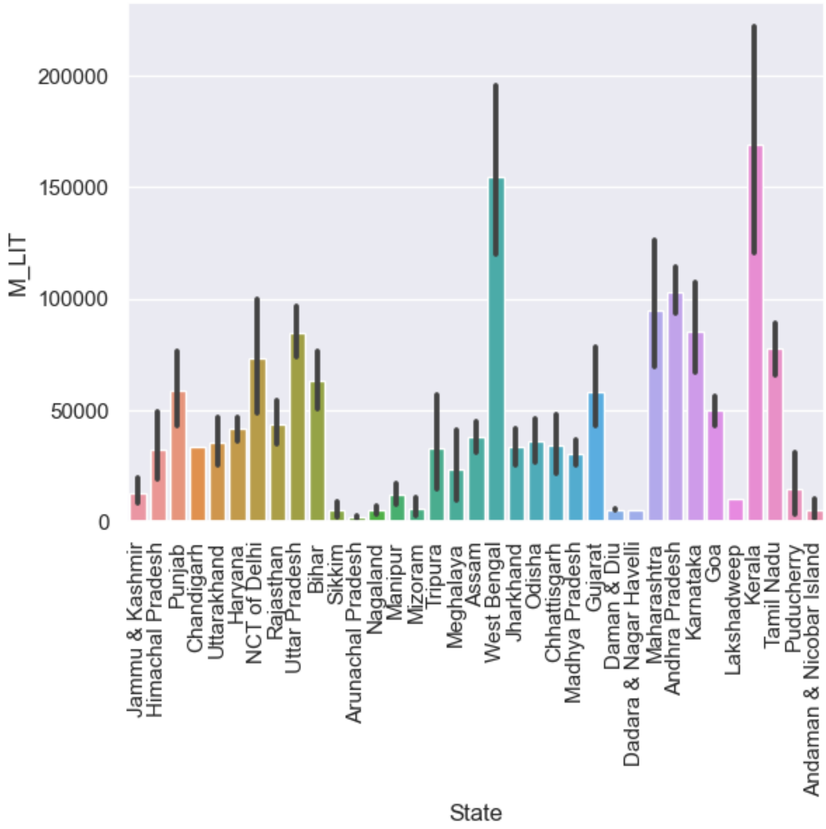
2) Which state has the lowest Total population of Males?



By Above Bar plot (State, TOT_F), We can get the following Questions as

1) Which state has highest Total population of Female?

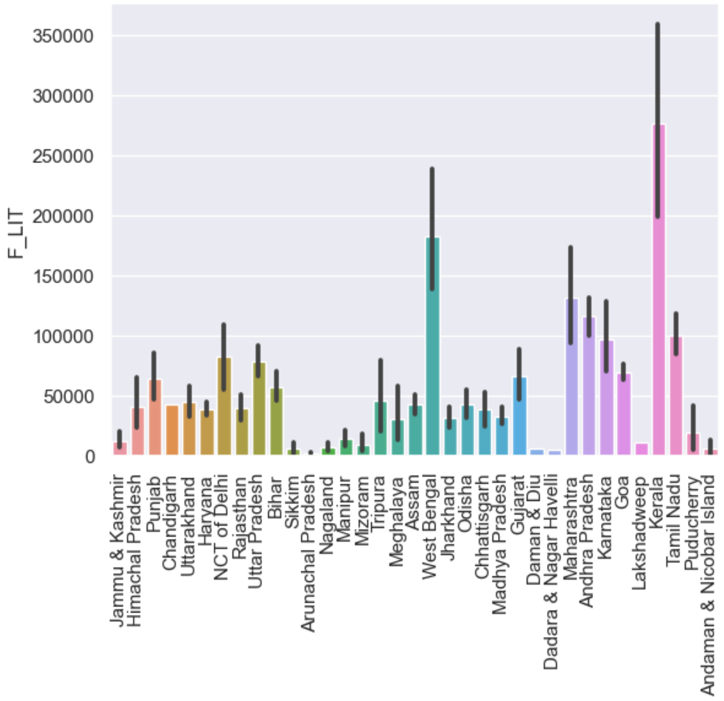
2) Which state has the lowest Total population of Female?



By Above Bar plot (State, M_LIT), We can get the following Questions such as:

1) Which state has the highest literate population, Male?

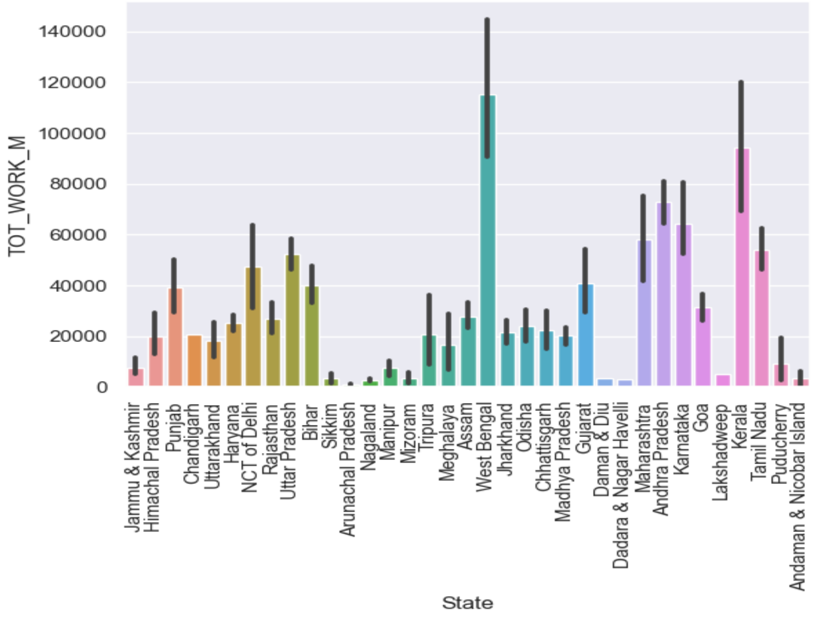
2) Which state has the lowest literate population, Male?



By Above Bar plot (State, F_LIT), We can get the following Questions such as:

1) Which state has the highest literate population, Female?

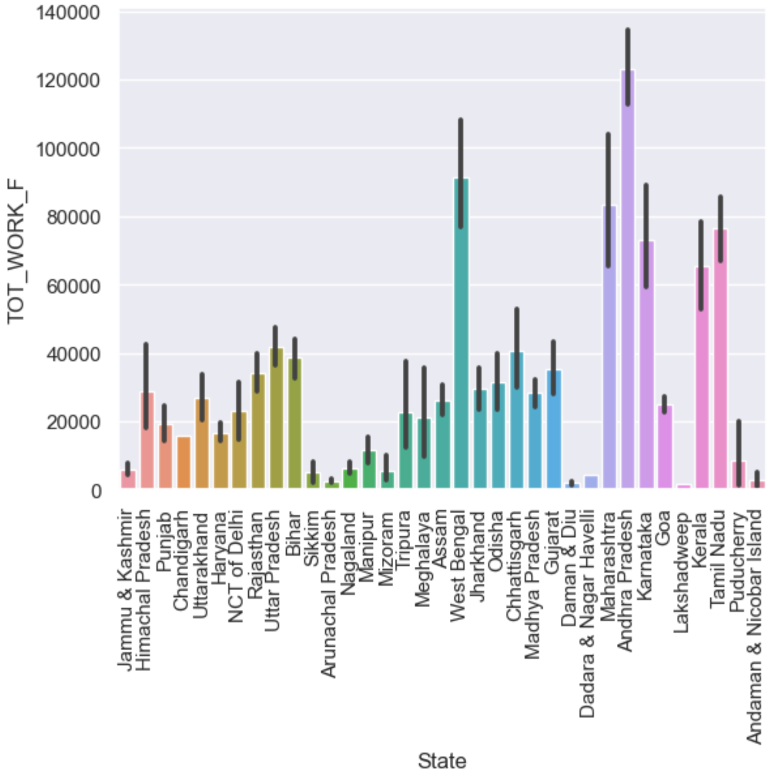
2) Which state has the lowest literate population, Female?



By Above Barplot (State, TOT_WORK_M), We can get the following Questions such as:

1) Which state has the highest Total Worker Population of Males?

2) Which state has the lowest Total Worker Population of Males?



By Above Barplot (State, TOT_WORK_F), We can get the following Questions such as:

1) Which state has the highest Total Worker Population of Females?

2) Which state has the lowest Total Worker Population of Females?

Q.3) We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?

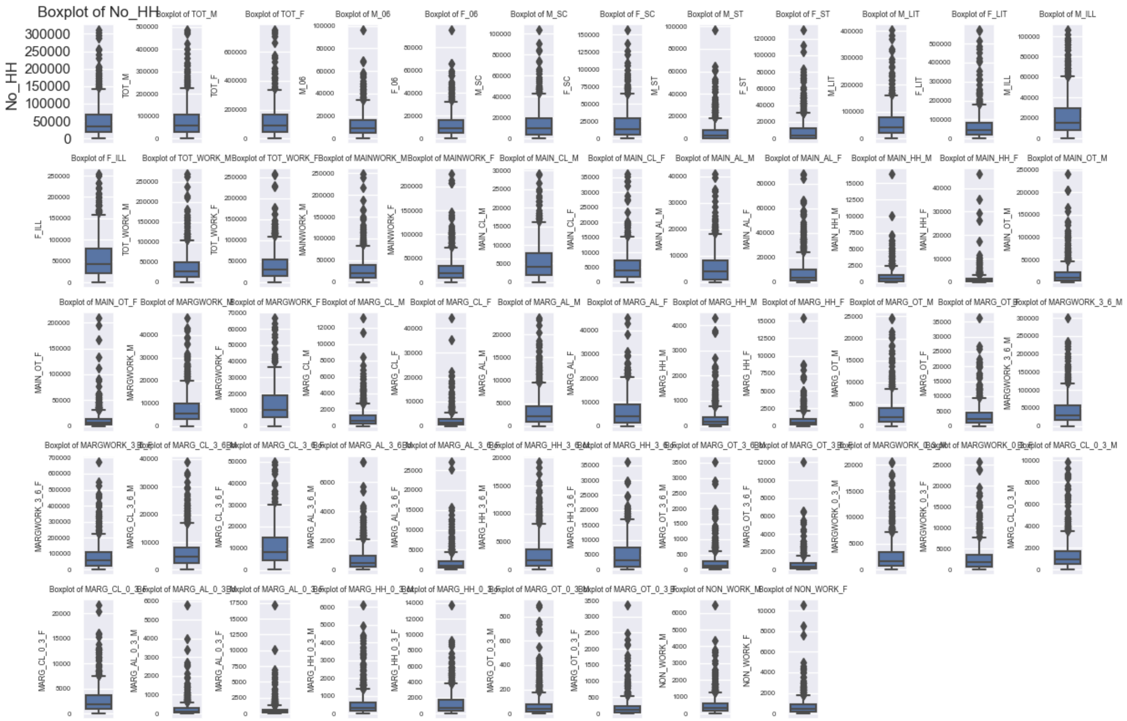
Ans:

Yes. Treating Outliers in this case is necessary.

I have dropped categorical features from Dataset to verify and treat Outliers.

I have verified Outliers with the help of Boxplot.

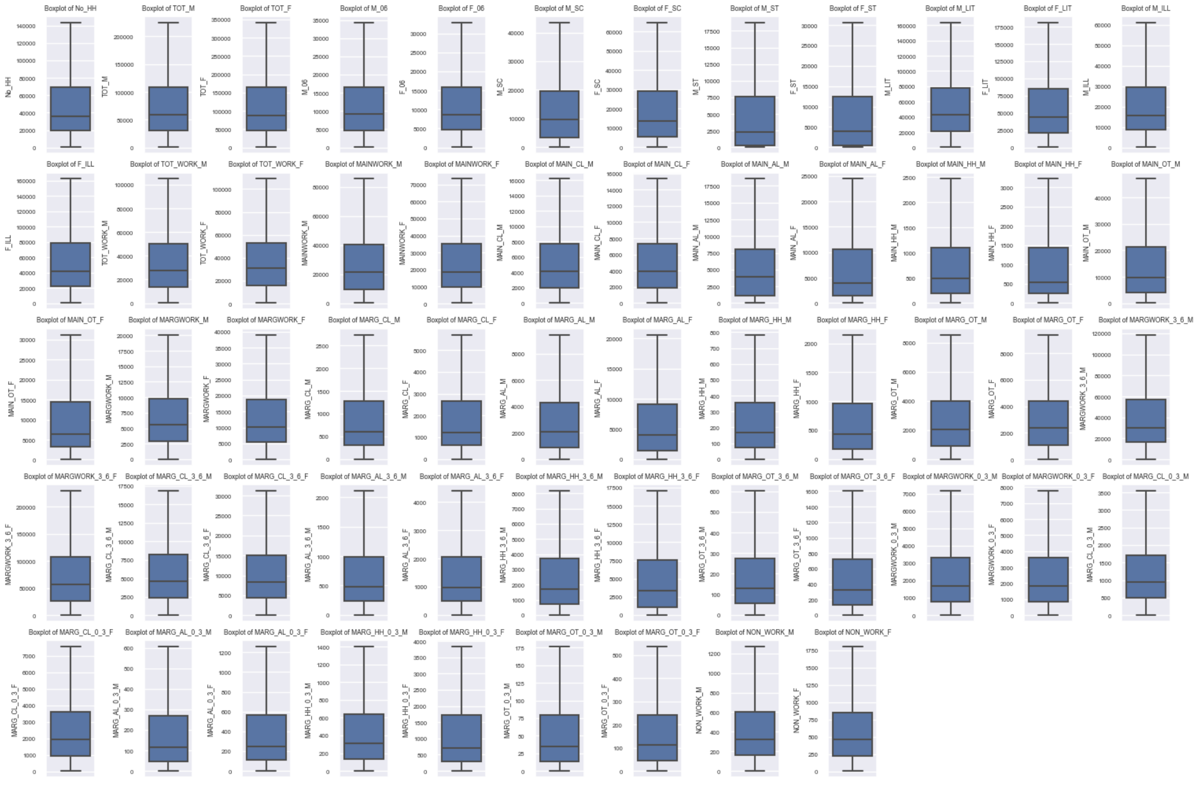
Before Treating Outliers.



We are going to treat outliers by IQR Method. (IQR : Interquartile Range).

After applying the remove outlier function on the dataset

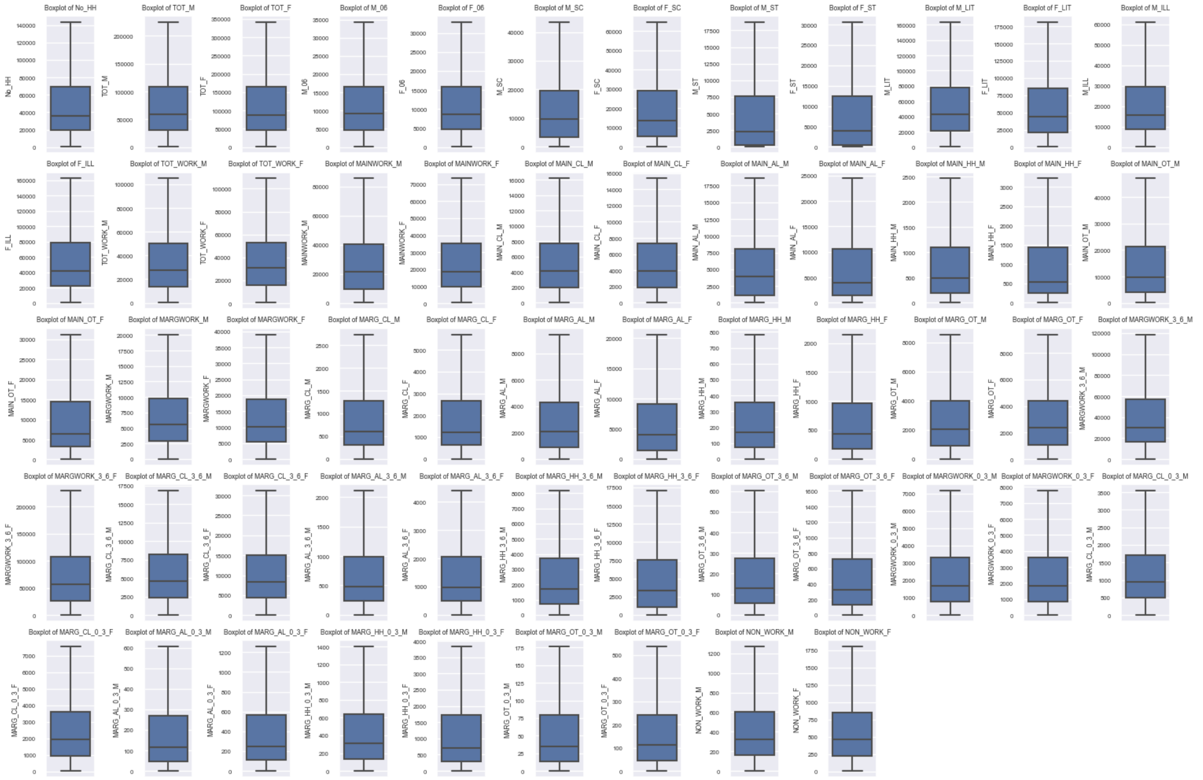
we get the following output.

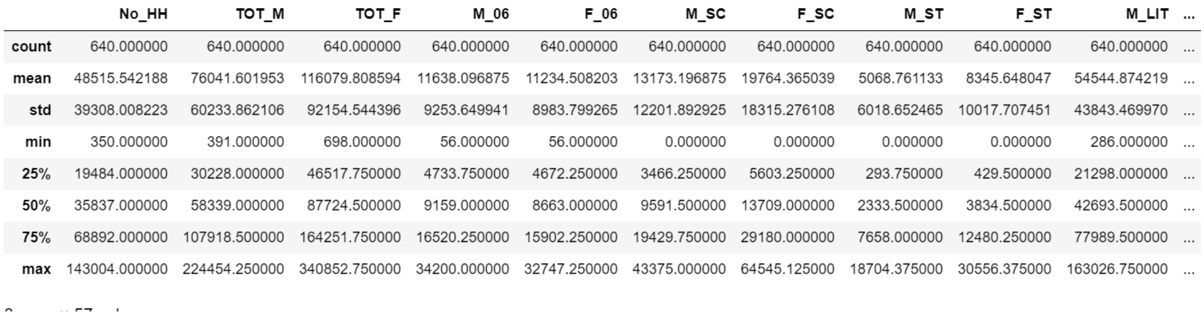


Q.4) Scale the Data using the z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment

I have already treated Outliers in Q.3) only. But still, I applied the Z-score for the scaling of the dataset. Please find below outputs by Boxplot and Describe function for 'Before' and after

BEFORE SCALING:

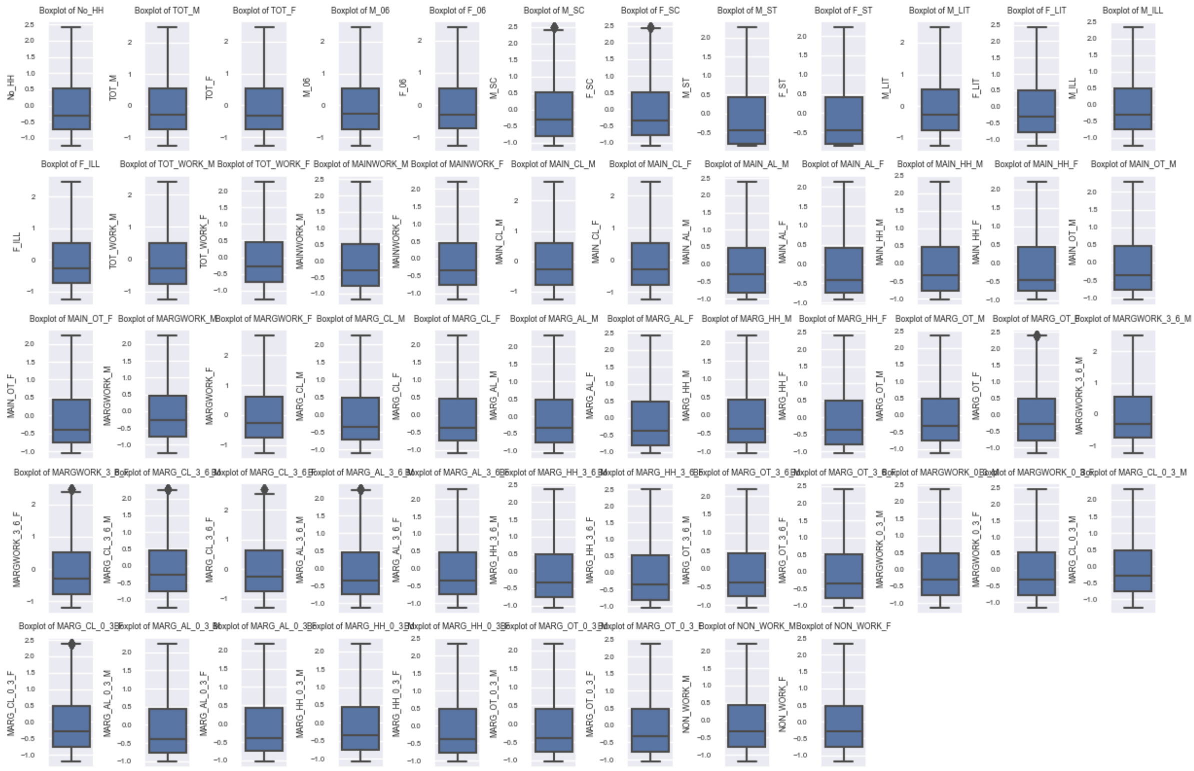


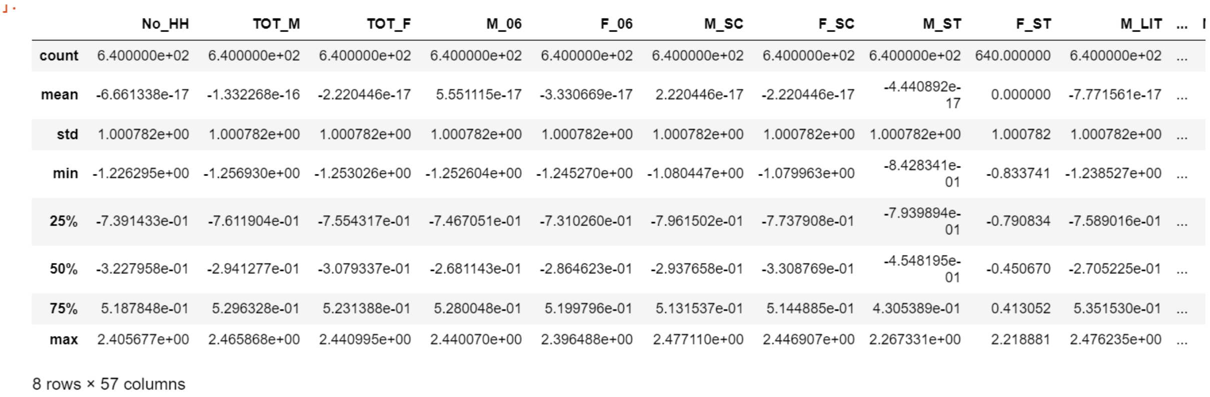


	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...
count	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	...
mean	48515.542188	76041.601953	116079.808594	11638.096875	11234.508203	13173.196875	19764.365039	5068.761133	8345.648047	54544.874219	...
std	39308.008223	60233.862106	92154.544396	9253.649941	8983.799265	12201.892925	18315.276108	6018.652465	10017.707451	43843.469970	...
min	350.000000	391.000000	698.000000	56.000000	56.000000	0.000000	0.000000	0.000000	0.000000	286.000000	...
25%	19484.000000	30228.000000	46517.750000	4733.750000	4672.250000	3466.250000	5603.250000	293.750000	429.500000	21298.000000	...
50%	35837.000000	58339.000000	87724.500000	9159.000000	8663.000000	9591.500000	13709.000000	2333.500000	3834.500000	42693.500000	...
75%	68892.000000	107918.500000	164251.750000	16520.250000	15902.250000	19429.750000	29180.000000	7658.000000	12480.250000	77989.500000	...
max	143004.000000	224454.250000	340852.750000	34200.000000	32747.250000	43375.000000	64545.125000	18704.375000	30556.375000	163026.750000	...

8 rows × 57 columns

AFTER SCALING:





	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...
count	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	640.000000	6.400000e+02	...
mean	-6.661338e-17	-1.332268e-16	-2.220446e-17	5.551115e-17	-3.330669e-17	2.220446e-17	-2.220446e-17	-4.440892e-17	0.000000	-7.771561e-17	...
std	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782	1.000782e+00	...
min	-1.226295e+00	-1.256930e+00	-1.253026e+00	-1.252604e+00	-1.245270e+00	-1.080447e+00	-1.079963e+00	-8.428341e-01	-0.833741	-1.238527e+00	...
25%	-7.391433e-01	-7.611904e-01	-7.554317e-01	-7.467051e-01	-7.310260e-01	-7.961502e-01	-7.737908e-01	-7.939894e-01	-0.790834	-7.589016e-01	...
50%	-3.227958e-01	-2.941277e-01	-3.079337e-01	-2.681143e-01	-2.864623e-01	-2.937658e-01	-3.308769e-01	-4.548195e-01	-0.450670	-2.705225e-01	...
75%	5.187848e-01	5.296328e-01	5.231388e-01	5.280048e-01	5.199796e-01	5.131537e-01	5.144885e-01	4.305389e-01	0.413052	5.351530e-01	...
max	2.405677e+00	2.465868e+00	2.440995e+00	2.440070e+00	2.396488e+00	2.477110e+00	2.446907e+00	2.267331e+00	2.218881	2.476235e+00	...

8 rows × 57 columns

Q.5) Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Geteigen values and eigen vector.

Ans:

Find below steps for PCA (Principal Component Analysis)

1) Performing Outlier Treatment

2) Scaling of the data

3) Create Covariance Matrix

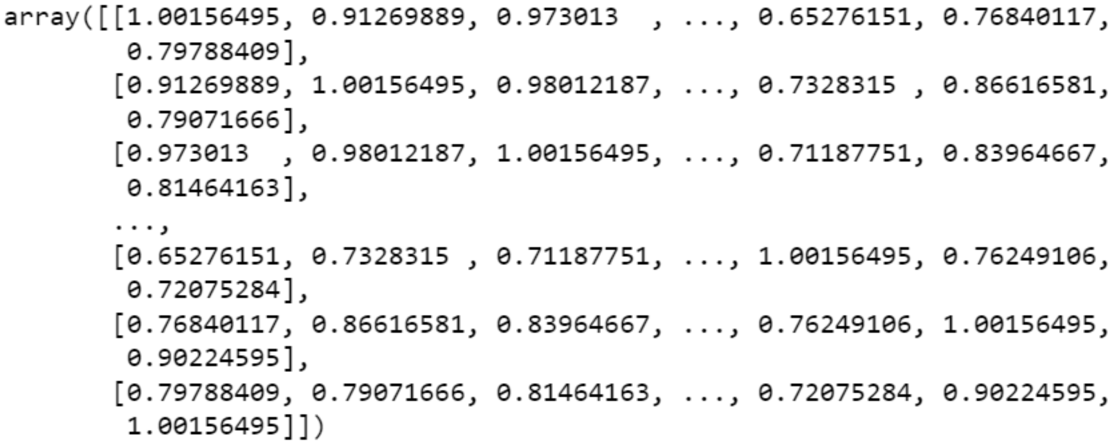
4) Extract Eigen Vector

5) Find Eigen Value

6) Create WSS Scree plot for variance

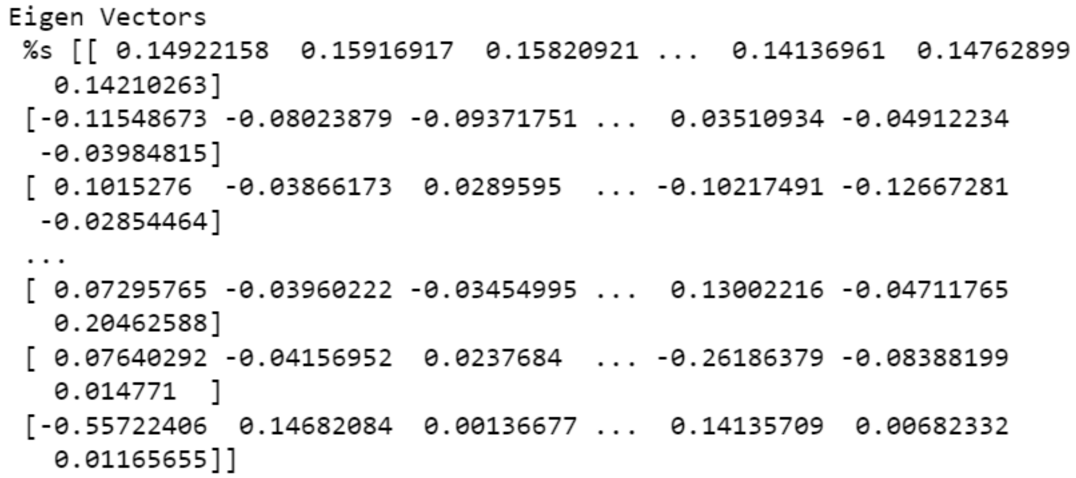
7) Find a cut off for selecting the number of PCs

Covariance Matrix:



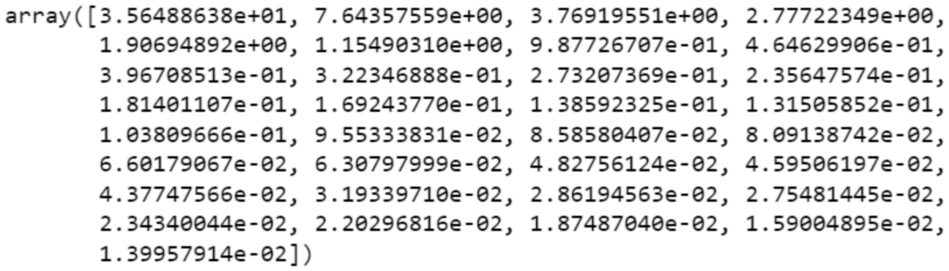
```
array([[1.00156495, 0.91269889, 0.973013  , ..., 0.65276151, 0.76840117,
        0.79788409],
       [0.91269889, 1.00156495, 0.98012187, ..., 0.7328315 , 0.86616581,
        0.79071666],
       [0.973013  , 0.98012187, 1.00156495, ..., 0.71187751, 0.83964667,
        0.81464163],
       ...,
       [0.65276151, 0.7328315 , 0.71187751, ..., 1.00156495, 0.76249106,
        0.72075284],
       [0.76840117, 0.86616581, 0.83964667, ..., 0.76249106, 1.00156495,
        0.90224595],
       [0.79788409, 0.79071666, 0.81464163, ..., 0.72075284, 0.90224595,
        1.00156495]])
```

EIGEN VECTOR:



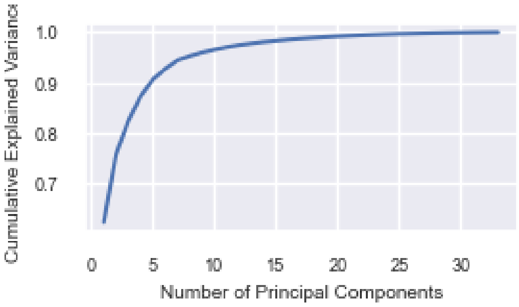
```
Eigen Vectors
 %s [[ 0.14922158  0.15916917  0.15820921 ...  0.14136961  0.14762899
   0.14210263]
 [-0.11548673 -0.08023879 -0.09371751 ...  0.03510934 -0.04912234
  -0.03984815]
 [ 0.1015276  -0.03866173  0.0289595  ... -0.10217491 -0.12667281
  -0.02854464]
 ...
 [ 0.07295765 -0.03960222 -0.03454995 ...  0.13002216 -0.04711765
   0.20462588]
 [ 0.07640292 -0.04156952  0.0237684  ... -0.26186379 -0.08388199
   0.014771  ]
 [-0.55722406  0.14682084  0.00136677 ...  0.14135709  0.00682332
   0.01165655]]
```

EIGEN VALUES:



```
array([3.56488638e+01, 7.64357559e+00, 3.76919551e+00, 2.77722349e+00,
       1.90694892e+00, 1.15490310e+00, 9.87726707e-01, 4.64629906e-01,
       3.96708513e-01, 3.22346888e-01, 2.73207369e-01, 2.35647574e-01,
       1.81401107e-01, 1.69243770e-01, 1.38592325e-01, 1.31505852e-01,
       1.03809666e-01, 9.55333831e-02, 8.58580407e-02, 8.09138742e-02,
       6.60179067e-02, 6.30797999e-02, 4.82756124e-02, 4.59506197e-02,
       4.37747566e-02, 3.19339710e-02, 2.86194563e-02, 2.75481445e-02,
       2.34340044e-02, 2.20296816e-02, 1.87487040e-02, 1.59004895e-02,
       1.39957914e-02])
```

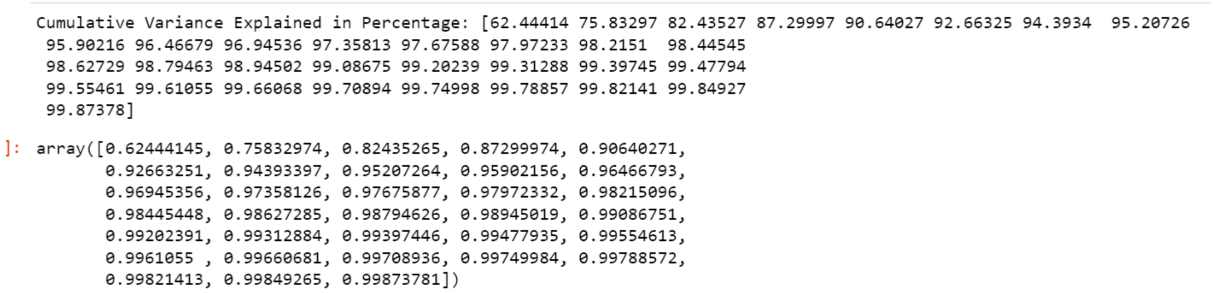
Q.6) Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.



Number of components needed to explain 90% of the variance: 6

The optimum number of PCs is 6

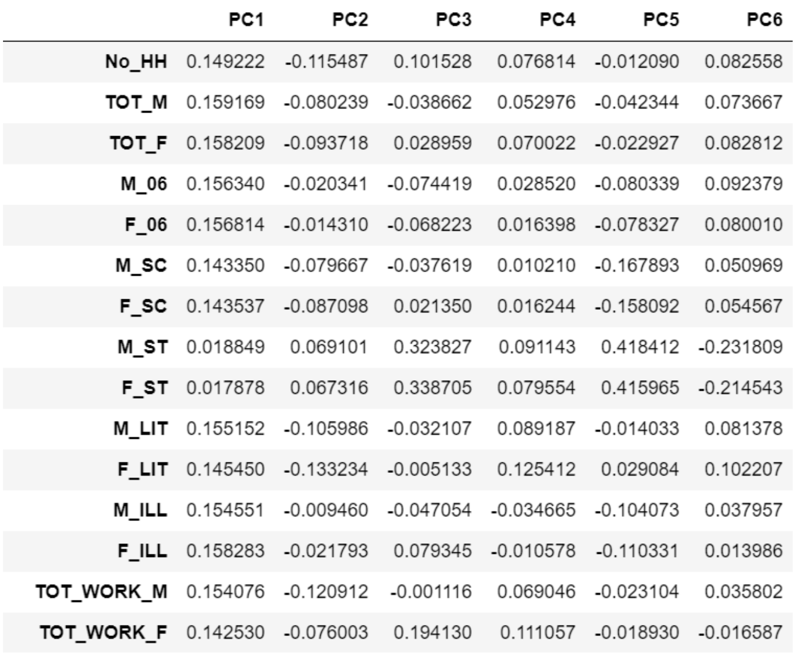
Cumulative explained variance ratio to find a cut off for selecting the number of PCs:



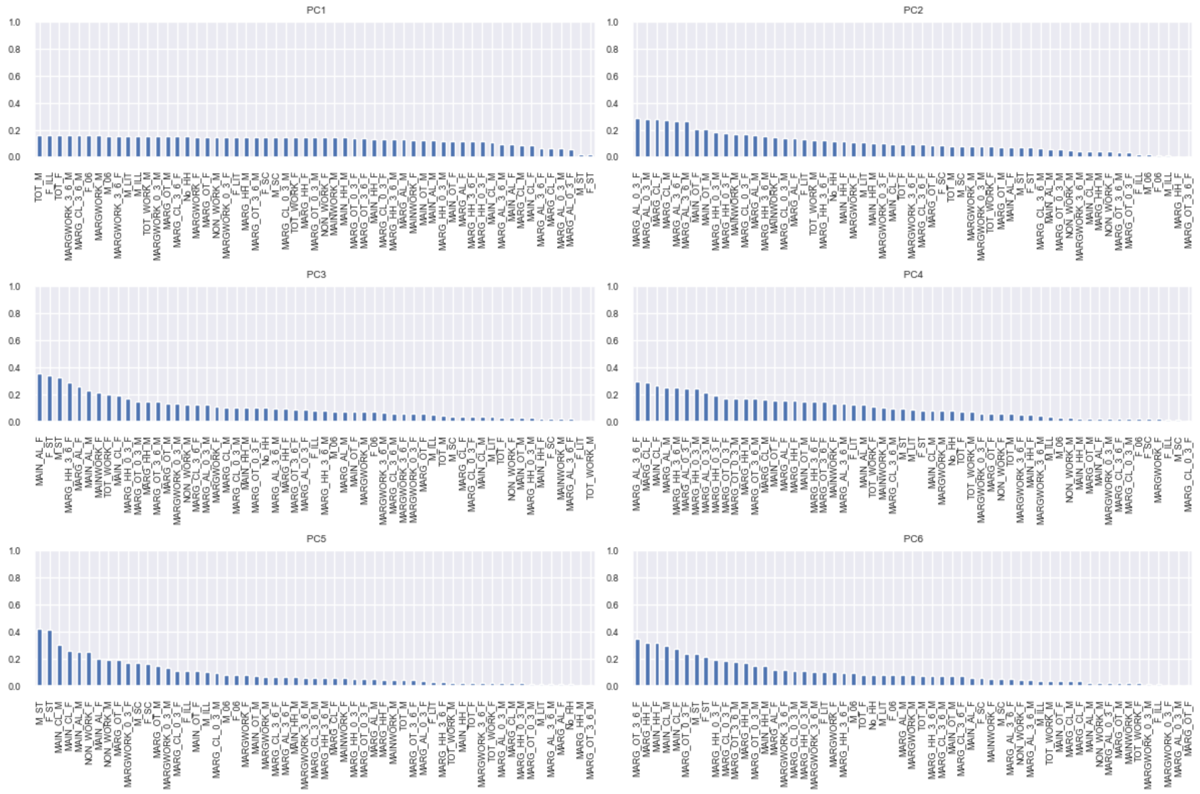
```
Cumulative Variance Explained in Percentage: [62.44414 75.83297 82.43527 87.29997 90.64027 92.66325 94.3934  95.20726
 95.90216 96.46679 96.94536 97.35813 97.67588 97.97233 98.2151  98.44545
 98.62729 98.79463 98.94502 99.08675 99.20239 99.31288 99.39745 99.47794
 99.55461 99.61055 99.66068 99.70894 99.74998 99.78857 99.82141 99.84927
 99.87378]

]: array([0.62444145, 0.75832974, 0.82435265, 0.87299974, 0.90640271,
        0.92663251, 0.94393397, 0.95207264, 0.95902156, 0.96466793,
        0.96945356, 0.97358126, 0.97675877, 0.97972332, 0.98215096,
        0.98445448, 0.98627285, 0.98794626, 0.98945019, 0.99086751,
        0.99202391, 0.99312884, 0.99397446, 0.99477935, 0.99554613,
        0.9961055 , 0.99660681, 0.99708936, 0.99749984, 0.99788572,
        0.99821413, 0.99849265, 0.99873781])
```

Q.7) Compare PCs with Actual Columns and identify which explains most variance. Write inferences about all the Principal components in terms of actual variables.



	PC1	PC2	PC3	PC4	PC5	PC6
No_HH	0.149222	-0.115487	0.101528	0.076814	-0.012090	0.082558
TOT_M	0.159169	-0.080239	-0.038662	0.052976	-0.042344	0.073667
TOT_F	0.158209	-0.093718	0.028959	0.070022	-0.022927	0.082812
M_06	0.156340	-0.020341	-0.074419	0.028520	-0.080339	0.092379
F_06	0.156814	-0.014310	-0.068223	0.016398	-0.078327	0.080010
M_SC	0.143350	-0.079667	-0.037619	0.010210	-0.167893	0.050969
F_SC	0.143537	-0.087098	0.021350	0.016244	-0.158092	0.054567
M_ST	0.018849	0.069101	0.323827	0.091143	0.418412	-0.231809
F_ST	0.017878	0.067316	0.338705	0.079554	0.415965	-0.214543
M_LIT	0.155152	-0.105986	-0.032107	0.089187	-0.014033	0.081378
F_LIT	0.145450	-0.133234	-0.005133	0.125412	0.029084	0.102207
M_ILL	0.154551	-0.009460	-0.047054	-0.034665	-0.104073	0.037957
F_ILL	0.158283	-0.021793	0.079345	-0.010578	-0.110331	0.013986
TOT_WORK_M	0.154076	-0.120912	-0.001116	0.069046	-0.023104	0.035802
TOT_WORK_F	0.142530	-0.076003	0.194130	0.111057	-0.018930	-0.016587



Q.8) Write linear equation for first PC

. Ans: $PC1 = a1x1 + a2x2 + a3x3 + a4x4 + a5x5 + \cdots + anxn$