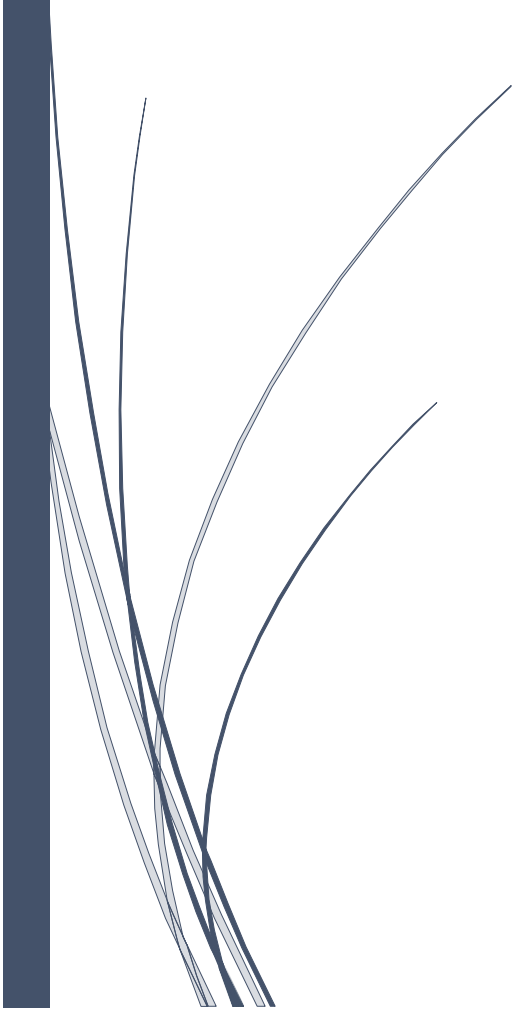
greatlearning
Learning for Life

11/13/2022

Data Mining Project



Contents:

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

CPC = Total Cost (spend) / Number of Clicks

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

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_Data1 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 Case Study 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.

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.
Data file - PCA India Data Census.xlsx

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

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

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

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

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

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

- Write linear equation for first PC.

Solutions:

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

CPC = Total Cost (spend) / Number of Clicks

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

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_Data1 Excel File.

Perform the following in given order:

Part 1 - Clustering: 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.:

- Top 5 rows of the Data frame:

	0	1	2	3	4
Timestamp	2020-9-2-17	2020-9-2-18	2020-9-3-16	2020-9-3-2	2020-9-3-13
InventoryType	Format1	Format1	Format6	Format1	Format1
Ad - Length	300	300	336	300	300
Ad- Width	250	250	250	250	250
Ad Size	75000	75000	84000	75000	75000
Ad Type	Inter222	Inter223	Inter217	Inter224	Inter225
Platform	Video	Web	Web	Web	Video
Device Type	Desktop	Mobile	Desktop	Desktop	Mobile
Format	Display	Display	Video	Display	Display
Available_Impressions	1806	1979	1566	643	1550
Matched_Queries	325	384	298	103	347
Impressions	323	380	297	102	345
Clicks	1	0	0	0	0
Spend	0.0	0.0	0.0	0.0	0.0
Fee	0.35	0.35	0.35	0.35	0.35
Revenue	0.0	0.0	0.0	0.0	0.0
CTR	0.0031	0.0	0.0	0.0	0.0
CPM	0.0	0.0	0.0	0.0	0.0
CPC	0.0	NaN	NaN	NaN	NaN

- Bottom 5 rows of the Data frame:

	25852	25853	25854	25855	25856
Timestamp	2020-10-1-5	2020-11-18-2	2020-9-14-0	2020-9-30-4	2020-10-17-3
InventoryType	Format5	Format4	Format5	Format7	Format5
Ad - Length	720	120	720	300	720
Ad- Width	300	600	300	600	300
Ad Size	216000	72000	216000	180000	216000
Ad Type	Inter222	inter230	Inter221	Inter228	Inter225
Platform	Video	Video	App	Video	Video
Device Type	Desktop	Mobile	Mobile	Mobile	Mobile
Format	Video	Video	Video	Display	Display
Available_Impressions	1	7	2	1	1
Matched_Queries	1	1	2	1	1
Impressions	1	1	2	1	1
Clicks	0	1	1	0	0
Spend	0.01	0.07	0.09	0.01	0.01
Fee	0.35	0.35	0.35	0.35	0.35
Revenue	0.0065	0.0455	0.0585	0.0065	0.0065
CTR	NaN	NaN	NaN	NaN	NaN
CPM	NaN	NaN	NaN	NaN	NaN
CPC	NaN	NaN	NaN	NaN	NaN

- Basic info about the Data frame:

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

- Data frame Summary:

	count	mean	std	min	25%	50%	75%	max
Ad - Length	25857.0	3.904312e+02	2.306961e+02	120.00	120.0000	300.0000	7.200000e+02	728.00
Ad- Width	25857.0	3.321828e+02	1.942609e+02	70.00	250.0000	300.0000	6.000000e+02	600.00
Ad Size	25857.0	9.968328e+04	6.264069e+04	33600.00	72000.0000	75000.0000	8.400000e+04	216000.00
Available_Impressions	25857.0	2.169621e+06	4.542680e+06	0.00	9133.0000	330968.0000	2.208484e+06	27592861.00
Matched_Queries	25857.0	1.155322e+06	2.407244e+06	0.00	5451.0000	189449.0000	1.008171e+06	14702025.00
Impressions	25857.0	1.107525e+06	2.326648e+06	0.00	2558.0000	162162.0000	9.496930e+05	14194774.00
Clicks	25857.0	9.525881e+03	1.672169e+04	0.00	305.0000	3457.0000	1.068100e+04	143049.00
Spend	25857.0	2.414473e+03	3.932835e+03	0.00	36.0300	1173.6600	2.692280e+03	26931.87
Fee	25857.0	3.367289e-01	3.053978e-02	0.21	0.3500	0.3500	3.500000e-01	0.35
Revenue	25857.0	1.716549e+03	2.993025e+03	0.00	23.4200	762.8800	1.749982e+03	21276.18
CTR	19392.0	6.962653e-02	7.497012e-02	0.00	0.0024	0.0077	1.283000e-01	1.00
CPM	19392.0	7.252900e+00	6.538314e+00	0.00	1.6300	3.0350	1.222000e+01	81.56
CPC	18330.0	3.510606e-01	3.433338e-01	0.00	0.0900	0.1600	5.700000e-01	7.26

- Data frame Null-value check:

```
Timestamp                  0
InventoryType              0
Ad - Length                0
Ad- Width                  0
Ad Size                    0
Ad Type                    0
Platform                   0
Device Type                0
Format                     0
Available_Impressions      0
Matched_Queries            0
Impressions                0
Clicks                     0
Spend                      0
Fee                        0
Revenue                    0
CTR                     6465
CPM                     6465
CPC                     7527
dtype: int64
```

- Data frame duplicate value check:

Total duplicate values: 0

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

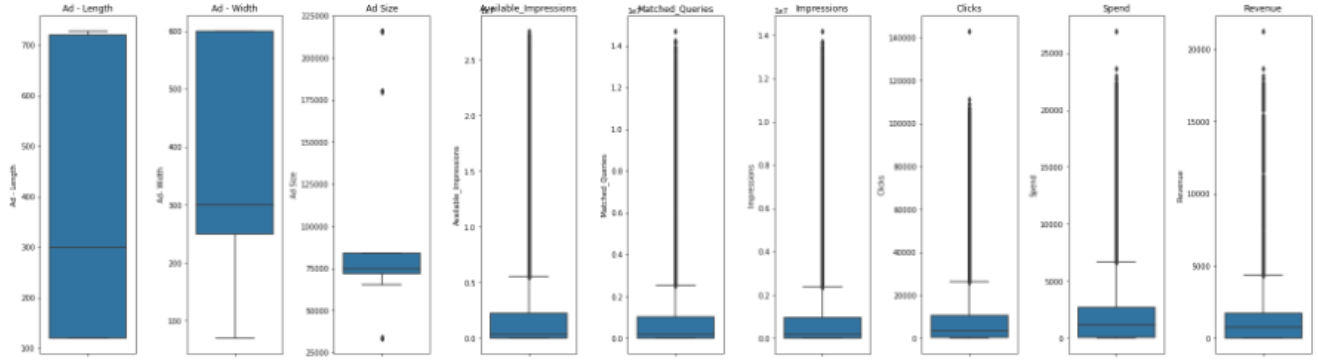
Ans.:

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

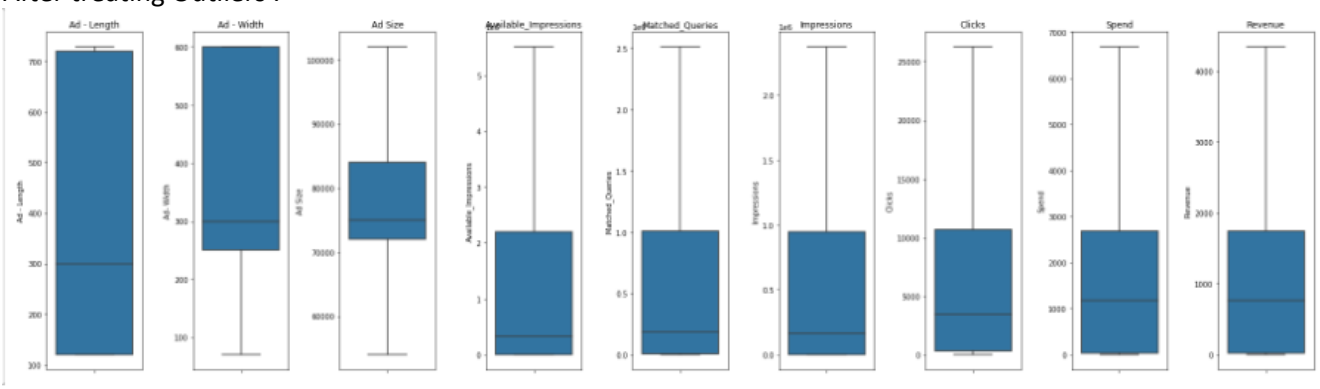
#We have no outliers now present in our 'df_cluster' Dataframe.

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).

Ans.:



After treating Outliers :



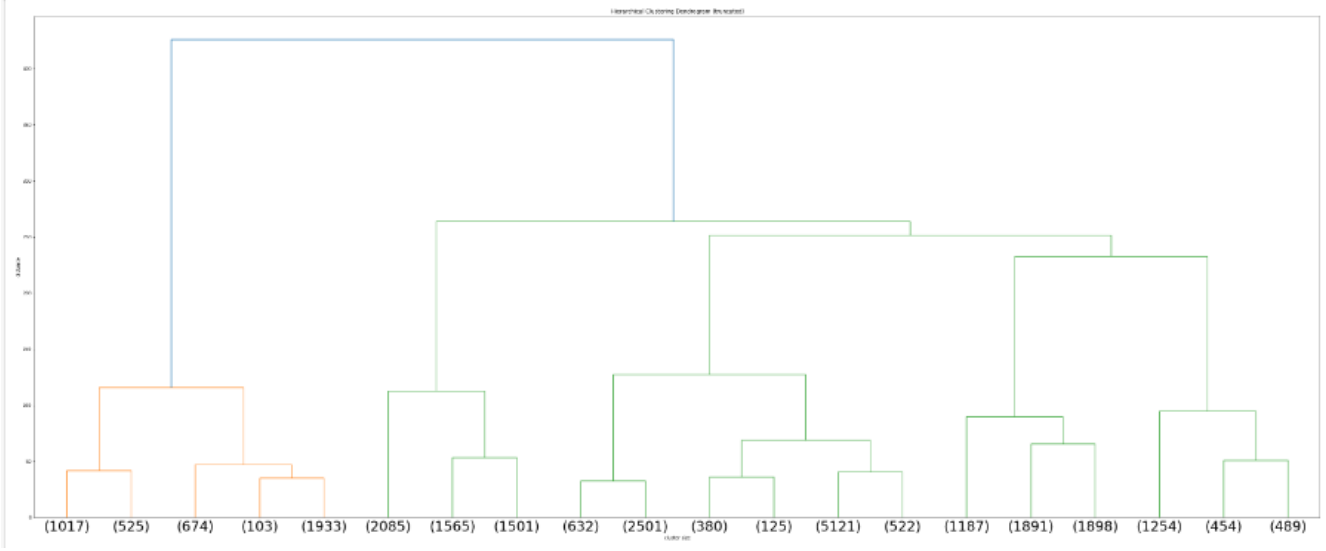
- **Yes, we need to treat Outliers as K-means Clustering is sensitive to outliers.**

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

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Revenue
0	-0.392000	-0.423062	-0.161806	-0.714953	-0.744816	-0.735050	-0.821889	-0.844382	-0.841307
1	-0.392000	-0.423062	-0.161806	-0.714862	-0.744749	-0.734983	-0.822006	-0.844382	-0.841307
2	-0.235948	-0.423062	0.424415	-0.715079	-0.744846	-0.735081	-0.822006	-0.844382	-0.841307
3	-0.392000	-0.423062	-0.161806	-0.715566	-0.745066	-0.735313	-0.822006	-0.844382	-0.841307
4	-0.392000	-0.423062	-0.161806	-0.715088	-0.744791	-0.735024	-0.822006	-0.844382	-0.841307

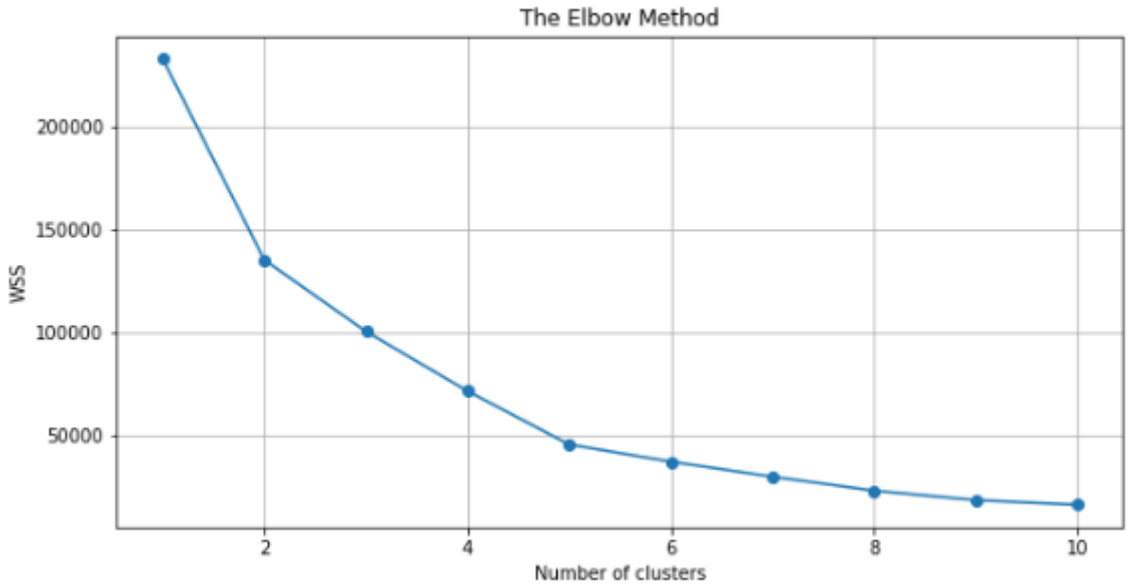
- **Without Scaling data, the algorithm may be biased towards higher value.**

Part 1 - Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.
Ans.:



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

The WSS value for 1 cluster is 232713.00000000006
The WSS value for 2 clusters is 135274.9268314021
The WSS value for 3 clusters is 100590.2395311129
The WSS value for 4 clusters is 71656.59481682391
The WSS value for 5 clusters is 45771.31324276951
The WSS value for 6 clusters is 37438.815811017026
The WSS value for 7 clusters is 30149.7112338386
The WSS value for 8 clusters is 23382.874391416677
The WSS value for 9 clusters is 18790.993332464503
The WSS value for 10 clusters is 16544.499210561502



#So, from the Dendrogram we can say optimum number of clusters: ' 5 '

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

The Silhouette scores for 2 clusters is 0.43093038125940913
The Silhouette scores for 3 clusters is 0.4169029019588384
The Silhouette scores for 4 clusters is 0.4859045662423113
The Silhouette scores for 5 clusters is 0.5484421685630947
The Silhouette scores for 6 clusters is 0.5554079926857388
The Silhouette scores for 7 clusters is 0.5882973964429631
The Silhouette scores for 8 clusters is 0.6005106775133303
The Silhouette scores for 9 clusters is 0.6298955511943023
The Silhouette scores for 10 clusters is 0.6296839903311501

#So, from the Silhouette scores we can say optimum number of clusters: ' 5 '

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].
Ans.:

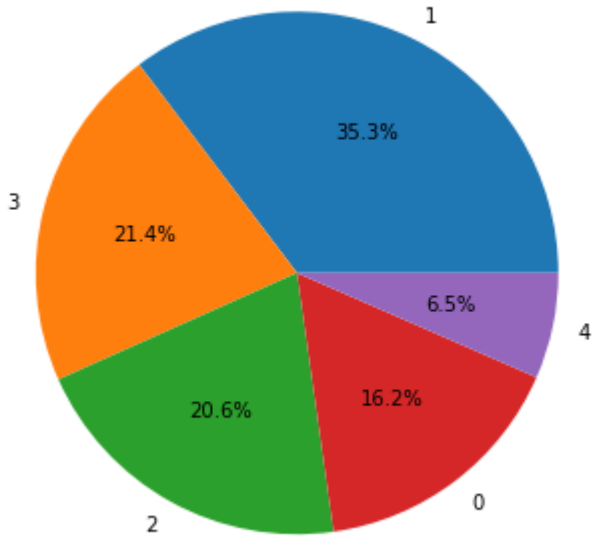
array ([1, 1, 1, ..., 3, 1, 3])

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Revenue	k_means_cluster_5	sil_width
0	300.0	250.0	75000.0	1806.0	325.0	323.0	1.0	0.0	0.0	1	0.484134
1	300.0	250.0	75000.0	1979.0	384.0	380.0	0.0	0.0	0.0	1	0.484118
2	336.0	250.0	84000.0	1566.0	298.0	297.0	0.0	0.0	0.0	1	0.455931
3	300.0	250.0	75000.0	643.0	103.0	102.0	0.0	0.0	0.0	1	0.484184
4	300.0	250.0	75000.0	1550.0	347.0	345.0	0.0	0.0	0.0	1	0.484141
5	300.0	250.0	75000.0	2641.0	493.0	491.0	0.0	0.0	0.0	1	0.484063
6	300.0	250.0	75000.0	469.0	104.0	103.0	0.0	0.0	0.0	1	0.484184
7	300.0	250.0	75000.0	1244.0	154.0	153.0	0.0	0.0	0.0	1	0.484173
8	300.0	250.0	75000.0	1961.0	287.0	287.0	0.0	0.0	0.0	1	0.484134
9	300.0	250.0	75000.0	1670.0	223.0	223.0	0.0	0.0	0.0	1	0.484155

- **Total Count per cluster:**

ster_5	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Revenue	sil_width	cluster count
0	460.364417	201.664679	73127.421894	5.161947e+06	2.450391e+06	2.324038e+06	11031.213212	5291.014889	3522.500695	0.585120	4193
1	190.589008	486.999124	75315.305452	7.423864e+04	3.971523e+04	3.467679e+04	1208.494964	166.682201	108.343391	0.625923	9134
2	442.612982	122.553151	61203.386642	1.939653e+06	9.174353e+05	8.754002e+05	3559.829163	1600.696988	1042.279585	0.501689	5315
3	693.438349	303.413000	101136.338946	2.174275e+05	1.177387e+05	1.002682e+05	11424.665037	1045.018470	680.529256	0.583809	5523
4	142.957447	572.281324	73925.531915	7.561640e+05	5.324346e+05	4.491000e+05	25720.598109	5734.283874	3785.384493	0.720601	1692

Cluster Profiling:



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

- 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.

•Dendrogram is the visualization and linkage are for computing the distances and merging the clusters from n to 1.

•The output of Linkage is visualized 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 fit – transform function and viewing the output - The data frame is now stored in an array.

•Using this array, we can now perform k-means

•The one requirement before we run the k-means algorithm, is to know how many clusters we require as output

•We map the elbow plot using wss values

•From the plot we have following observations:

•When we move from k=1 to k=2, we see that there is a significant drop in the value, also when we move from k=2 to k=3, k=3 to k=4 there is a significant drop as well.

•But from k=4 to k=5, k=5 to k=6, the drop in values reduces significantly.

•In other words, the wss is not significantly dropping beyond 5,

•So, 5 is optimal number of clusters.

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.
Data file - PCA India Data Census.xlsx

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

Ans.:

- Top 5 rows of the Data frame:

	0	1	2	3	4
State Code	1	1	1	1	1
Dist.Code	1	2	3	4	5
State	Jammu & Kashmir	Jammu & Kashmir	Jammu & Kashmir	Jammu & Kashmir	Jammu & Kashmir
Area Name	Kupwara	Badgam	Leh(Ladakh)	Kargil	Punch
No_HH	7707	6218	4452	1320	11654
TOT_M	23388	19585	6546	2784	20591
TOT_F	29796	23102	10964	4206	29981
M_06	5862	4482	1062	563	5157
F_06	6196	3733	1018	677	4587
M_SC	3	7	3	0	20
F_SC	0	6	6	0	33
M_ST	1999	427	5806	2686	7670
F_ST	2598	517	9723	3968	10843
M_LIT	13381	10513	4534	1842	13243
F_LIT	11364	7891	5840	1962	13477
M_ILL	10007	9072	2012	942	7348
F_ILL	18432	15211	5124	2244	16504
TOT_WORK_M	6723	6982	2775	1002	5717
TOT_WORK_F	3752	4200	4800	1118	7692
MAINWORK_M	2763	4628	1940	491	2523
MAINWORK_F	1275	1733	2923	408	2267
MAIN_CL_M	486	1098	519	35	743
MAIN_CL_F	235	357	1205	102	766
MAIN_AL_M	407	442	36	8	254
MAIN_AL_F	143	108	71	24	237
MAIN_HH_M	78	538	19	9	35
MAIN_HH_F	86	343	55	6	64
MAIN_OT_M	1792	2550	1366	439	1491
MAIN_OT_F	811	925	1592	276	1200
MARGWORK_M	3960	2354	835	511	3194
MARGWORK_F	2477	2467	1877	710	5425
MARG_CL_M	619	384	360	135	1327
MARG_CL_F	580	661	1250	286	2462
MARG_AL_M	2052	915	44	63	1037
MARG_AL_F	641	547	157	176	1069
MARG_HH_M	142	369	15	10	62
MARG_HH_F	244	627	32	43	319
MARG_OT_M	1147	686	416	303	768
MARG_OT_F	1012	632	438	205	1575
MARGWORK_3_6_M	16685	12603	3771	1782	14874
MARGWORK_3_6_F	26044	18902	6164	3088	22289
MARG_CL_3_6_M	2610	1829	721	317	2320
MARG_CL_3_6_F	1726	1752	1689	463	3497
MARG_AL_3_6_M	439	261	316	74	862
MARG_AL_3_6_F	343	432	1161	158	1419
MARG_HH_3_6_M	1372	729	41	50	832
MARG_HH_3_6_F	389	399	123	126	767
MARG_OT_3_6_M	110	293	15	6	38
MARG_OT_3_6_F	198	449	28	33	214
MARGWORK_0_3_M	889	546	349	187	588
MARGWORK_0_3_F	798	472	377	146	1097
MARG_CL_0_3_M	1150	525	114	194	874
MARG_CL_0_3_F	749	715	188	247	1928
MARG_AL_0_3_M	180	123	44	61	465
MARG_AL_0_3_F	237	229	89	128	1043
MARG_HH_0_3_M	680	186	3	13	205
MARG_HH_0_3_F	252	148	34	50	302
MARG_OT_0_3_M	32	76	0	4	24
MARG_OT_0_3_F	46	178	4	10	105
NON_WORK_M	258	140	67	116	180
NON_WORK_F	214	160	61	59	478

Bottom 5 rows of the Data frame:

	635	636	637	638	639
State Code	34	34	35	35	35
Dist.Code	636	637	638	639	640
State	Puducherry	Puducherry	Andaman & Nicobar Island	Andaman & Nicobar Island	Andaman & Nicobar Island
Area Name	Mahe	Karaikal	Nicobars	North & Middle Andaman	South Andaman
No_HH	3333	10612	1275	3762	7975
TOT_M	8154	12346	1549	5200	11977
TOT_F	11781	21691	2630	8012	18049
M_06	1146	1544	227	723	1470
F_06	1203	1533	225	664	1358
M_SC	21	2234	0	0	0
F_SC	30	4155	0	0	0
M_ST	0	0	1012	28	161
F_ST	0	0	1750	50	264
M_LIT	6916	10292	1187	4206	10095
F_LIT	10184	14225	1602	5273	13362
M_ILL	1238	2054	362	994	1882
F_ILL	1597	7466	1028	2739	4687
TOT_WORK_M	3808	6458	715	2707	6345
TOT_WORK_F	1328	5286	1031	2174	5278
MAINWORK_M	3459	5619	325	2098	5366
MAINWORK_F	997	4104	534	1666	4514
MAIN_CL_M	8	132	8	553	255
MAIN_CL_F	3	108	8	225	246
MAIN_AL_M	27	645	1	63	88
MAIN_AL_F	5	903	1	28	67
MAIN_HH_M	16	25	16	8	37
MAIN_HH_F	3	173	38	7	39
MAIN_OT_M	3408	4817	300	1474	4986
MAIN_OT_F	986	2920	487	1406	4162
MARGWORK_M	349	839	390	609	979
MARGWORK_F	331	1182	497	508	764
MARG_CL_M	1	26	19	108	69
MARG_CL_F	6	30	9	163	71
MARG_AL_M	3	272	11	69	62
MARG_AL_F	5	515	14	55	45
MARG_HH_M	2	11	78	4	13
MARG_HH_F	2	67	165	8	21
MARG_OT_M	343	530	282	428	835
MARG_OT_F	318	550	309	282	627
MARGWORK_3_6_M	4346	5868	834	2493	5632
MARGWORK_3_6_F	10453	16405	1599	5838	12771
MARG_CL_3_6_M	317	684	286	473	806
MARG_CL_3_6_F	284	845	363	336	642
MARG_AL_3_6_M	1	23	10	84	63
MARG_AL_3_6_F	6	16	5	119	69
MARG_HH_3_6_M	3	234	9	58	45
MARG_HH_3_6_F	5	385	8	34	28
MARG_OT_3_6_M	2	7	61	3	11
MARG_OT_3_6_F	2	64	118	4	17
MARGWORK_0_3_M	311	420	206	328	687
MARGWORK_0_3_F	271	380	232	179	528
MARG_CL_0_3_M	32	155	104	136	173
MARG_CL_0_3_F	47	337	134	172	122
MARG_AL_0_3_M	0	3	9	24	6
MARG_AL_0_3_F	0	14	4	44	2
MARG_HH_0_3_M	0	38	2	11	17
MARG_HH_0_3_F	0	130	6	21	17
MARG_OT_0_3_M	0	4	17	1	2
MARG_OT_0_3_F	0	23	47	4	4
NON_WORK_M	32	110	76	100	148
NON_WORK_F	47	170	77	103	99

- Basic info about the Data frame:

```
<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
```

- Data frame Summary:

	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	8863.0	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	43798.5	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
MAINWORK_M	640.0	30204.446875	31480.915680	65.0	9787.00	21250.5	40119.00	247911.0
MAINWORK_F	640.0	28198.846875	29998.262689	240.0	9502.25	18484.0	35063.25	226166.0
MAIN_CL_M	640.0	5424.342188	4739.161969	0.0	2023.50	4160.5	7695.00	29113.0
MAIN_CL_F	640.0	5486.042188	5326.362728	0.0	1920.25	3908.5	7286.25	36193.0
MAIN_AL_M	640.0	5849.109375	6399.507966	0.0	1070.25	3936.5	8067.25	40843.0
MAIN_AL_F	640.0	8925.995312	12864.287584	0.0	1408.75	3933.5	10617.50	87945.0
MAIN_HH_M	640.0	883.893750	1278.642345	0.0	187.50	498.5	1099.25	16429.0
MAIN_HH_F	640.0	1380.773438	3179.414449	0.0	248.75	540.5	1435.75	45979.0
MAIN_OT_M	640.0	18047.101562	26068.480886	36.0	3997.50	9598.0	21249.50	240855.0
MAIN_OT_F	640.0	12406.035938	18972.202369	153.0	3142.50	6380.5	14368.25	209355.0
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
MARG_CL_3_6_F	640.0	10339.864063	8467.473429	85.0	4351.50	8295.0	15102.00	50065.0
MARG_AL_3_6_M	640.0	789.848438	905.639279	0.0	235.50	480.5	986.00	7426.0
MARG_AL_3_6_F	640.0	1749.584375	2496.541514	0.0	497.25	985.5	2059.00	27171.0
MARG_HH_3_6_M	640.0	2743.635938	3059.586387	0.0	718.75	1714.5	3702.25	19343.0
MARG_HH_3_6_F	640.0	5169.850000	5335.640960	0.0	1113.75	3294.0	7502.25	36253.0
MARG_OT_3_6_M	640.0	245.362500	358.728567	0.0	58.00	129.5	276.00	3535.0
MARG_OT_3_6_F	640.0	585.884375	900.025817	0.0	127.75	320.5	719.25	12094.0
MARGWORK_0_3_M	640.0	2616.140625	3036.964381	7.0	755.00	1681.5	3320.25	20648.0
MARGWORK_0_3_F	640.0	2834.545312	3327.836932	14.0	833.50	1834.5	3610.50	25844.0
MARG_CL_0_3_M	640.0	1392.973438	1489.707052	4.0	489.50	949.0	1714.00	9875.0
MARG_CL_0_3_F	640.0	2757.050000	2788.776676	30.0	957.25	1928.0	3599.75	21611.0
MARG_AL_0_3_M	640.0	250.889062	453.336594	0.0	47.00	114.5	270.75	5775.0
MARG_AL_0_3_F	640.0	558.098438	1117.642748	0.0	109.00	247.5	568.75	17153.0
MARG_HH_0_3_M	640.0	560.690625	762.578991	0.0	136.50	308.0	642.00	6116.0
MARG_HH_0_3_F	640.0	1293.431250	1585.377936	0.0	298.00	717.0	1710.75	13714.0
MARG_OT_0_3_M	640.0	71.379688	107.897627	0.0	14.00	35.0	79.00	895.0
MARG_OT_0_3_F	640.0	200.742188	309.740854	0.0	43.00	113.0	240.00	3354.0
NON_WORK_M	640.0	510.014063	610.603187	0.0	161.00	326.0	604.50	6456.0
NON_WORK_F	640.0	704.778125	910.209225	5.0	220.50	464.5	853.50	10533.0

- Data frame Null-value check:

```
State Code          0
Dist.Code           0
State               0
Area Name           0
No_HH               0
TOT_M               0
TOT_F               0
M_06                0
F_06                0
M_SC                0
F_SC                0
M_ST                0
F_ST                0
M_LIT               0
F_LIT               0
M_ILL               0
F_ILL               0
TOT_WORK_M          0
TOT_WORK_F          0
MAINWORK_M          0
MAINWORK_F          0
MAIN_CL_M           0
MAIN_CL_F           0
MAIN_AL_M           0
MAIN_AL_F           0
MAIN_HH_M           0
MAIN_HH_F           0
MAIN_OT_M           0
MAIN_OT_F           0
MARGWORK_M          0
MARGWORK_F          0
MARG_CL_M           0
MARG_CL_F           0
MARG_AL_M           0
MARG_AL_F           0
MARG_HH_M           0
MARG_HH_F           0
MARG_OT_M           0
MARG_OT_F           0
MARGWORK_3_6_M      0
MARGWORK_3_6_F      0
MARG_CL_3_6_M       0
MARG_CL_3_6_F       0
MARG_AL_3_6_M       0
MARG_AL_3_6_F       0
MARG_HH_3_6_M       0
MARG_HH_3_6_F       0
MARG_OT_3_6_M       0
MARG_OT_3_6_F       0
MARGWORK_0_3_M      0
MARGWORK_0_3_F      0
MARG_CL_0_3_M       0
MARG_CL_0_3_F       0
MARG_AL_0_3_M       0
MARG_AL_0_3_F       0
MARG_HH_0_3_M       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
dtype: int64
```

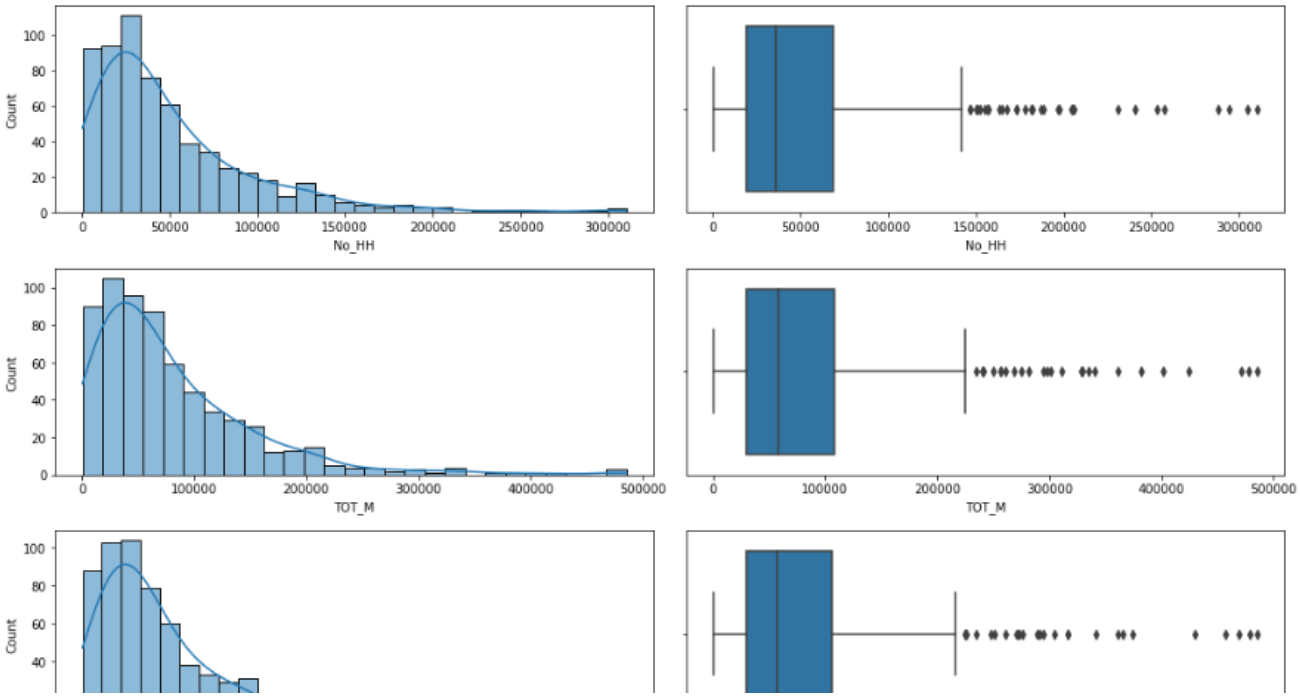
- Data frame duplicate value check:

Total duplicate values: 0

Part 2 - PCA: Perform detailed Exploratory analysis by creating certain questions like (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio? (Example Questions). Pick 5 variables out of the given 24 variables below for EDA: No_HH, TOT_M, TOT_F, M_06, F_06, M_SC, F_SC, M_ST, F_ST, M_LIT, F_LIT, M_ILL, F_ILL, TOT_WORK_M, TOT_WORK_F, MAINWORK_M, MAINWORK_F, MAIN_CL_M, MAIN_CL_F, MAIN_AL_M, MAIN_AL_F, MAIN_HH_M, MAIN_HH_F, MAIN_OT_M, MAIN_OT_F

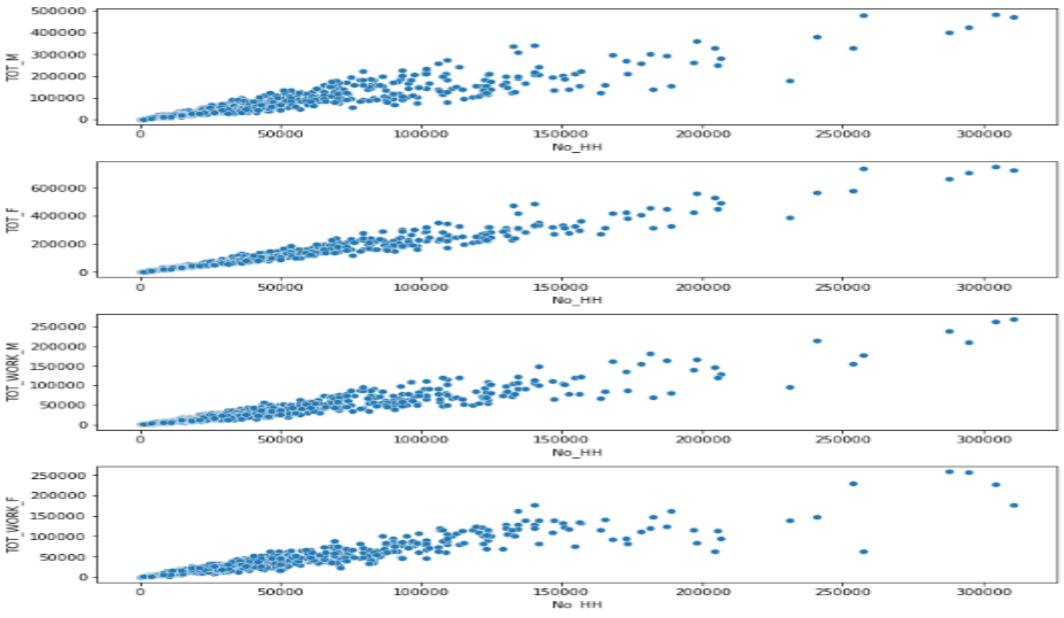
Ans.: Choose these 5 variables for EDA: 'No_HH', 'TOT_M', 'TOT_F', 'TOT_WORK_M', 'TOT_WORK_F'.

- Univariate Analysis:



- **From the Univariate Analysis we can say all variables are Left Skewed here and all are having Outliers.**

- Bi-variate Analysis:



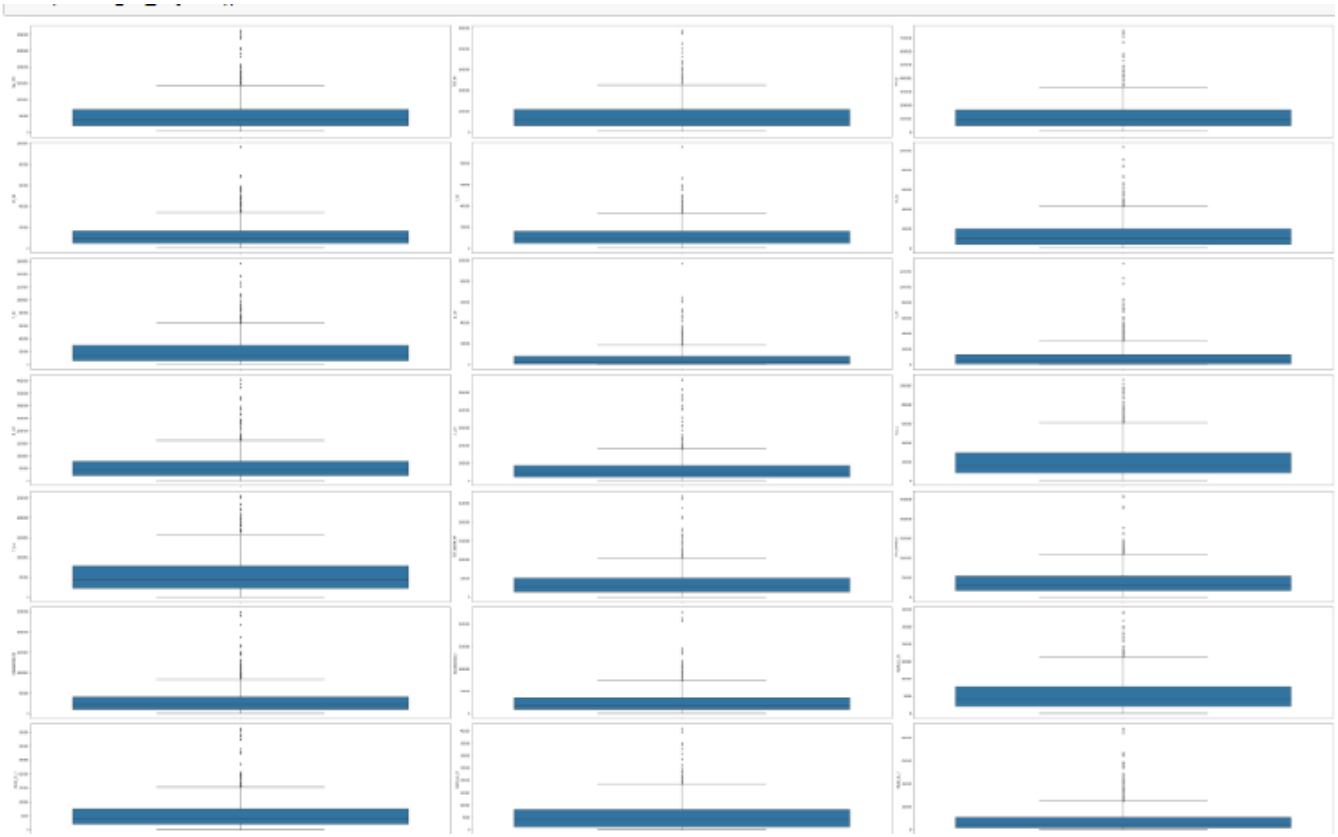
- **From the Bivariate Analysis we can say all variables are Positively Co-related to each other.**

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

Ans.: Outliers treatment is not necessary unless the they are the result from a processing mistake or wrong measurement. True outliers must be kept in the data.

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.

Ans.: Presence of outliers before z-score method :

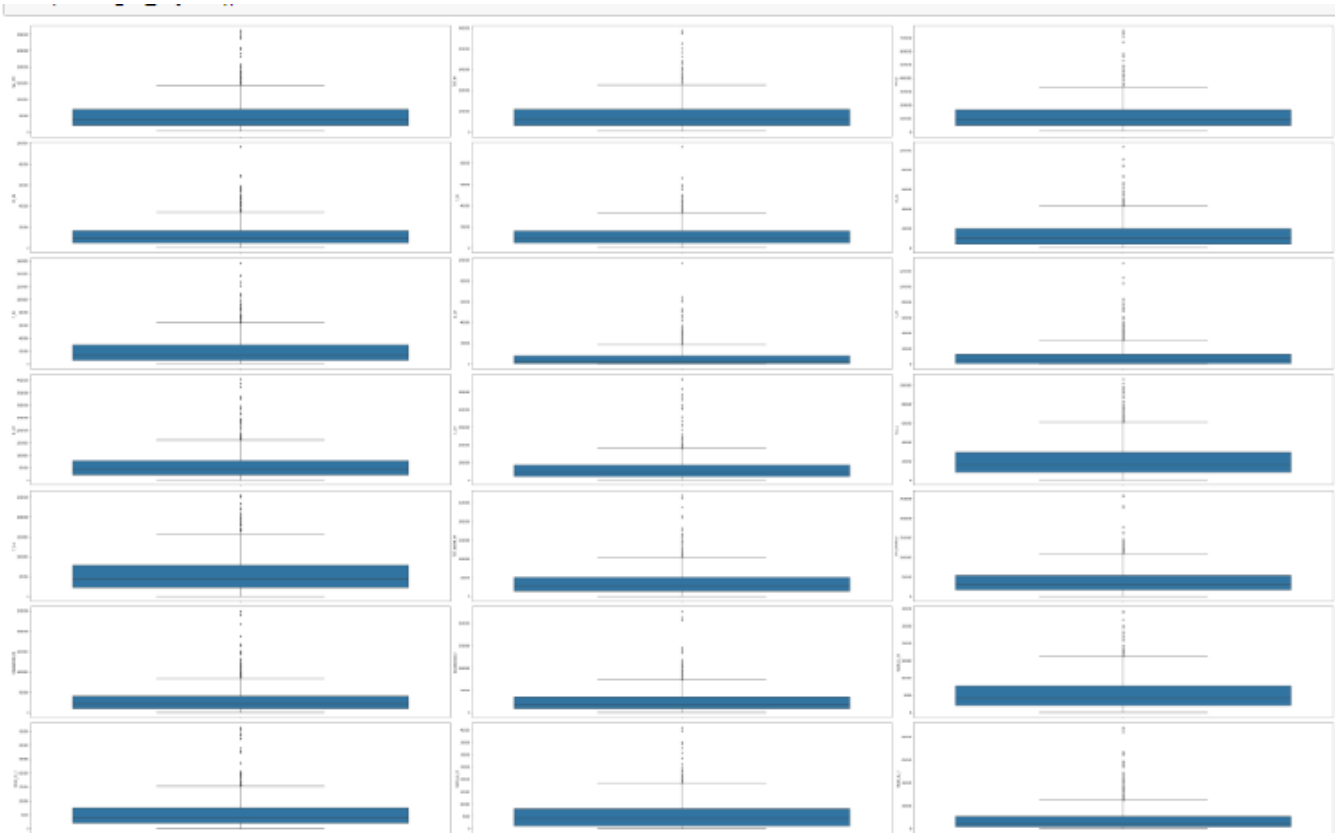


Scaled Data frame after applying Z-score :

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_A
0	-0.904738	-0.771236	-0.815563	-0.561012	-0.507738	-0.958575	-0.957049	-0.423308	-0.476423	-0.798097	...	-0.163229	-0.720610	
1	-0.935695	-0.823100	-0.874634	-0.681096	-0.725367	-0.958297	-0.956772	-0.582014	-0.607607	-0.849434	...	-0.583103	-0.732811	
2	-0.972412	-1.000919	-0.981466	-0.976956	-0.965262	-0.958575	-0.956772	-0.038951	-0.027273	-0.956457	...	-0.859212	-0.921931	
3	-1.037530	-1.052224	-1.041001	-1.022118	-0.995393	-0.958783	-0.957049	-0.355965	-0.390060	-1.004643	...	-0.805468	-0.900758	
4	-0.822676	-0.809381	-0.813933	-0.622359	-0.649908	-0.957395	-0.955529	0.149238	0.043330	-0.800568	...	-0.348645	-0.297513	

5 rows × 57 columns

Presence of outliers after z-score method :



#So, we can clearly see that scaling have no impact on outliers.

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

Ans.: 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]])
```

Eigen vectors:

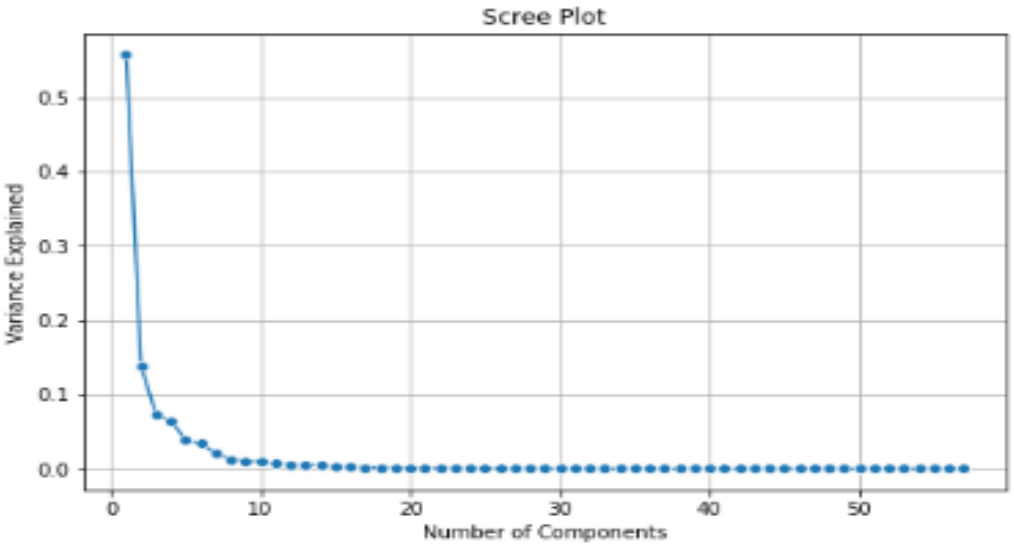
```
array([[ 0.15602058,  0.16711763,  0.16555318, ...,  0.13219224,
         0.15037558,  0.1310662 ],
       [-0.12634653, -0.08967655, -0.10491237, ...,  0.05081332,
        -0.06536455, -0.07384742],
       [-0.00269025,  0.05669762,  0.03874947, ..., -0.07871987,
         0.11182732,  0.1025525 ],
       ...,
       [ 0.        ,  0.14884588,  0.21643081, ...,  0.01740567,
        -0.0135858 ,  0.00152872],
       [ 0.        ,  0.11336536, -0.01111799, ..., -0.02196029,
        -0.08140651, -0.01767078],
       [ 0.        , -0.24963875,  0.38221285, ...,  0.02957246,
         0.04681613,  0.10209731]])
```

Eigen values:

```
array([3.18135647e+01, 7.86942415e+00, 4.15340812e+00, 3.66879058e+00,
       2.20652588e+00, 1.93827502e+00, 1.17617374e+00, 7.51159086e-01,
       6.17053743e-01, 5.28300887e-01, 4.29831189e-01, 3.53440201e-01,
       2.96163013e-01, 2.81275560e-01, 1.92158325e-01, 1.36267920e-01,
       1.13389199e-01, 1.06303946e-01, 9.72885376e-02, 8.01062194e-02,
       5.76089954e-02, 4.43955966e-02, 3.78910846e-02, 2.96360194e-02,
       2.70797618e-02, 2.34458139e-02, 1.45111511e-02, 1.09852268e-02,
       9.31507853e-03, 8.13540203e-03, 7.89250253e-03, 5.02601514e-03,
       2.59771182e-03, 1.06789820e-03, 7.13559124e-04, 1.39427796e-30,
       4.37180870e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       5.60923467e-32])
```

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

Ans.:



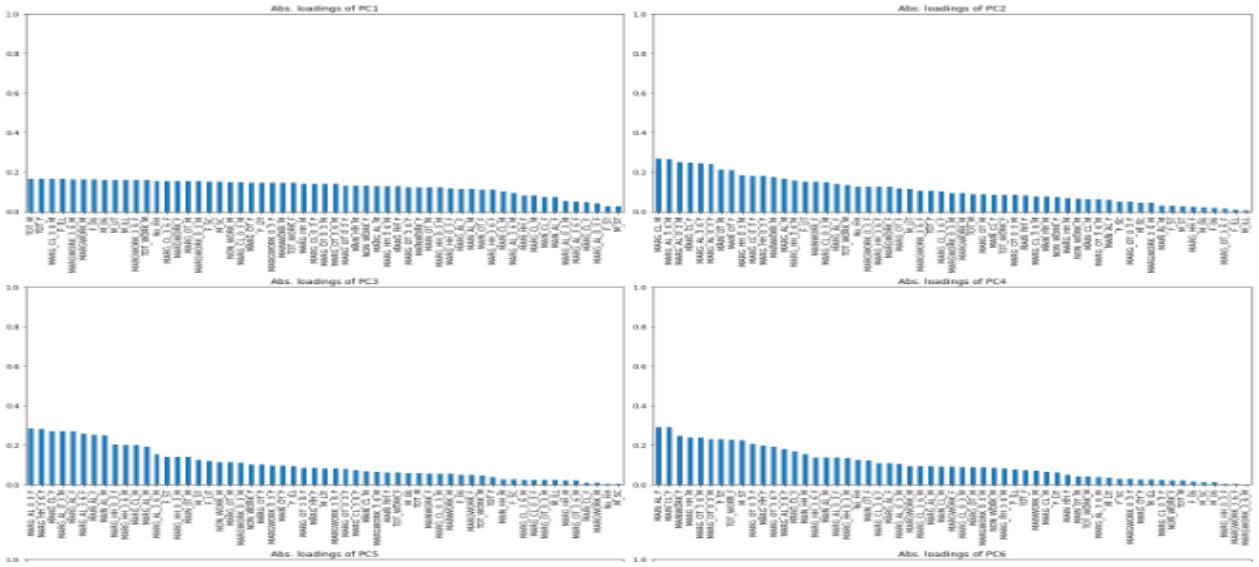
Cumulative explained variance ratio to find a cut off for selecting the number of PCs:

```
array([0.55726063, 0.69510499, 0.76785794, 0.83212212, 0.87077261,
       0.9047243 , 0.92532669, 0.93848433, 0.94929292, 0.95854687,
       0.96607599, 0.97226701, 0.97745473, 0.98238168, 0.98574761,
       0.98813454, 0.99012071, 0.99198278, 0.99368693, 0.99509011,
       0.99609921, 0.99687687, 0.99754058, 0.9980597 , 0.99853404,
       0.99894473, 0.99919891, 0.99939134, 0.9995545 , 0.99969701,
       0.99983525, 0.99992329, 0.9999688 , 0.9999875 , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        ])
```

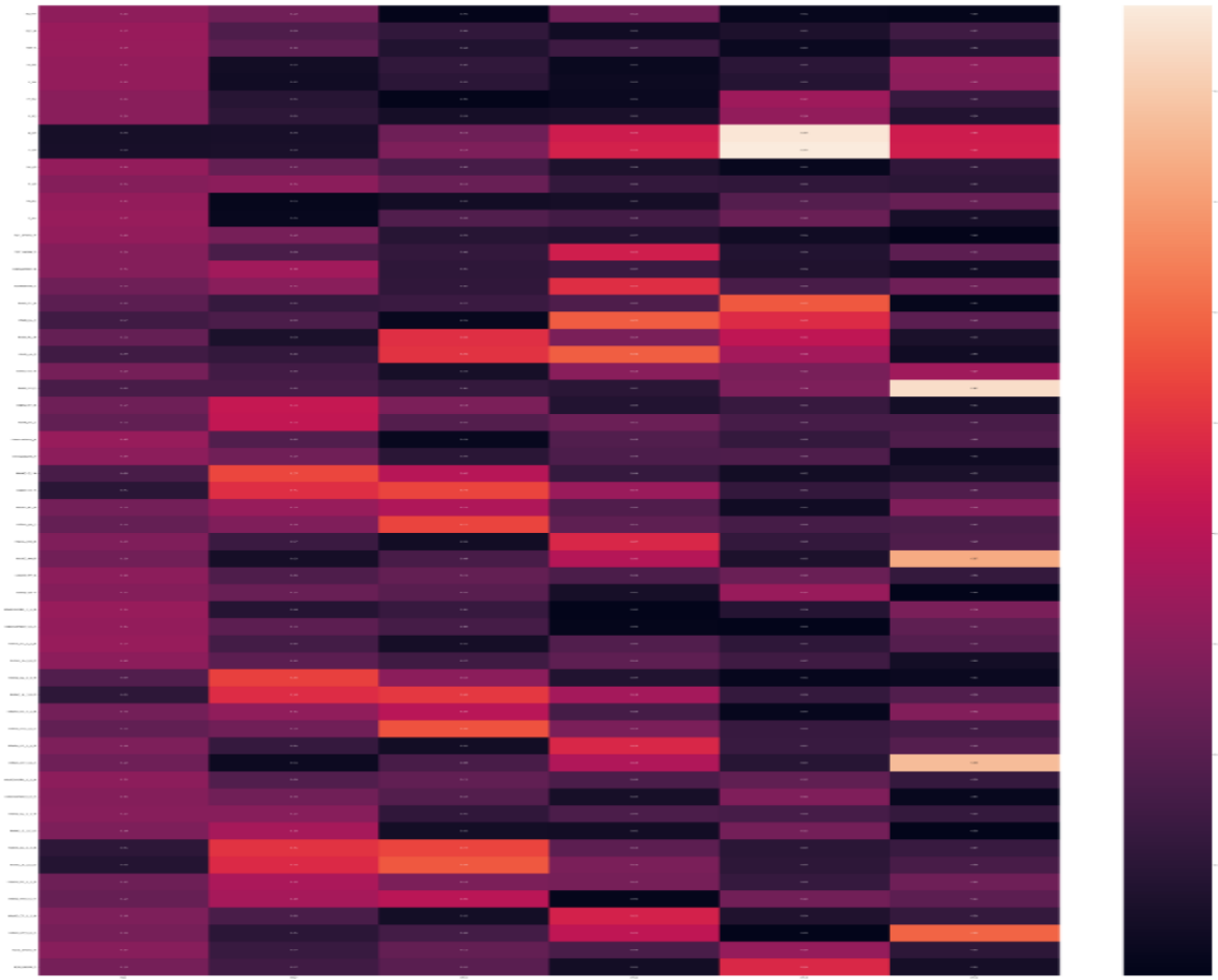
- ***For this project, we need to consider at least 90% explained variance, so cut off for selecting the number of PCs is: '6'***

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.

Ans.: Compare how the original features influence various PCs:



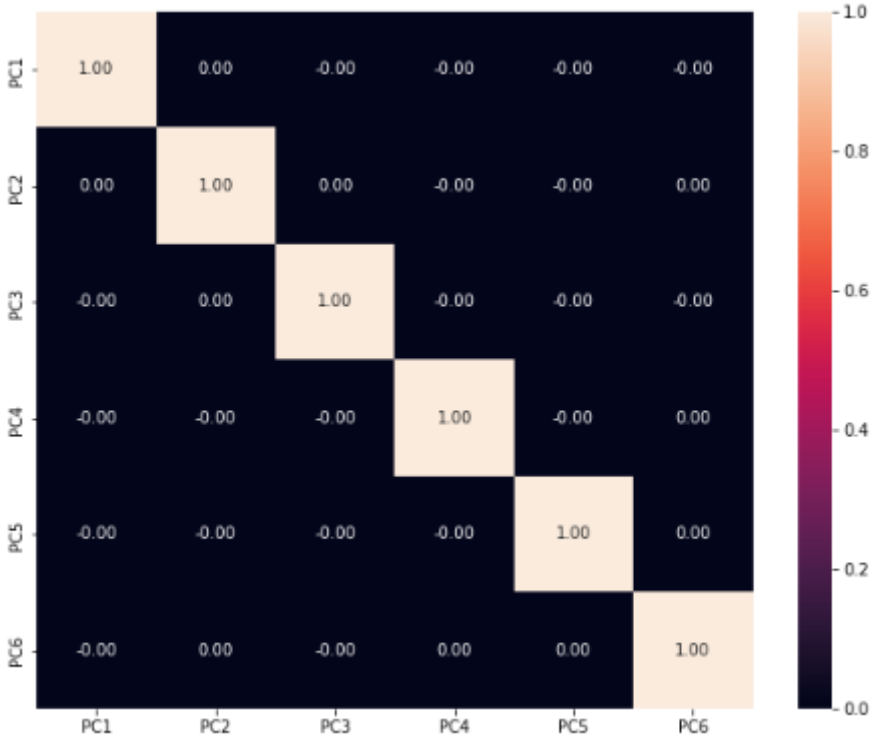
#Heatmap,Compare how the original features influence various PCs



#Extract the required number of PCs (6 in our case):

	PC1	PC2	PC3	PC4	PC5	PC6
0	-4.617263	0.138116	0.328545	1.543697	0.353736	-0.420948
1	-4.771662	-0.105865	0.244449	1.963215	-0.153884	0.417308
2	-5.964836	-0.294347	0.367394	0.619543	0.478199	0.276581
3	-6.280796	-0.500384	0.212701	1.074515	0.300799	0.051157
4	-4.478566	0.894154	1.078277	0.535557	0.804065	0.341678
5	-3.319963	2.823865	3.058460	-0.447904	0.742445	0.634676
6	-5.021393	-0.346359	0.650378	0.981072	-0.059778	-0.246957
7	-4.608709	0.022370	0.398755	1.576995	0.171316	-0.139444
8	-5.186703	-0.059097	0.184397	1.735440	0.169174	0.455039
9	-4.226190	-1.335080	0.697838	1.470509	0.269146	-0.002576

#Check for presence of correlations among the PCs:



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

Ans.: PC1 = a1x1 + a2x2 + a3X3 +a4X4 + ……. + a57x57