Data Mining Project Group5

Submitted by:

Sruthi C

PGP DSBA – Apr 2023

Date: 27-Aug-2023

List of Figures

Figure 1 : Clustering dataset 6

Figure 2: Head (top 5) 6

Figure 3: Tail (Last 5) 7

Figure 4: Data information......8
Figure 5: Data description......9
Figure 6: Null values......9
Figure 7: Unique values......10
Figure 8: After imputing null values......10
Figure 9: Box plots......14
Figure 10: Box plots after treating outliers......18
Figure 11: Scaled data......19
Figure 12: Dendrogram......19
Figure 13: Merged Dendrogram......20
Figure 14: Elbow plot......20
Figure 15: Silhouette plot......21
Figure 16: Bar plot for clicks, spend and revenue......23
Figure 17: Bar plot for CTR, CPC, CPM......24
Figure 18: PCA dataset......27
Figure 19: Head (top 5)......28
Figure 20: Tail (Last 5)......29
Figure 21: Dataset Information......29
Figure 22: Null values......30
Figure 23: Duplicate values......30
Figure 24: Dataset description......31
Figure 25: Correlation table......33
Figure 26: Boxplots of any 5 variables......33
Figure 27: Barplot of any 5 variables......34
Figure 28: Pairplot of any 5 variables......35
Figure 29: Scaled data......36
Figure 30: Scaled data summary......38
Figure 31: Boxplots of scaled data......39
Figure 32: Heatmap of scaled data......39
Figure 33: Covariance matrix......40
Figure 34: Eigen vectors......40
Figure 35: Eigen values......41
Figure 36: Variance......43

Figure 37: Cumulative variance 43
Figure 38: Scree plot 44
Figure 39: Explained variance ratio vs PCs 44
Figure 40: Final Dimension of PCs 45
Figure 41: Heatmap of PCA vs Original columns 46
Figure 42: Heatmap of all PCs 48

Table of Contents

Problem Statement- Clustering 5

1. Read the data and perform basic analysis 5

2. Treat missing values in CPC, CTR and CPM using the formula given. 10

3. Check if there are any outliers. 9

4. Treat Outliers 11

5. Perform z-score scaling and discuss how it affects the speed of the algorithm. 188

6. Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance. 19

7. Elbow Plot 20

8. Silhouette score 21

9. Profile the ads 2121

10. Summary: 255

Problem Statement - PCA 26

1. Basic Data Analysis 26

2. Exploratory Data Analysis. 31

3. Data Scaling 36

4. Effect of scaling on outliers 38

5. Principal Component Analysis 40

6. Component wise inference. 46

7. Influence of actual variables on PC1 48

Problem Statement - Clustering

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

Perform the following in given order:

1. Read the data and perform basic analysis.

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

- Data frame-

	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1
1	2020-9-2-18	Format1	300	250	75000	Inter223	Web	Mobile	Display	1979	384	380	0
2	2020-9-3-16	Format6	336	250	84000	Inter217	Web	Desktop	Video	1566	298	297	0
3	2020-9-3-2	Format1	300	250	75000	Inter224	Web	Desktop	Display	643	103	102	0
4	2020-9-3-13	Format1	300	250	75000	Inter225	Video	Mobile	Display	1550	347	345	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
25852	2020-10-1-5	Format5	720	300	216000	Inter222	Video	Desktop	Video	1	1	1	0
25853	2020-11-18-2	Format4	120	600	72000	inter230	Video	Mobile	Video	7	1	1	1
25854	2020-9-14-0	Format5	720	300	216000	Inter221	App	Mobile	Video	2	2	2	1
25855	2020-9-30-4	Format7	300	600	180000	Inter228	Video	Mobile	Display	1	1	1	0
25856	2020-10-17-3	Format5	720	300	216000	Inter225	Video	Mobile	Display	1	1	1	0

25857 rows × 19 columns

Spend	Fee	Revenue	CTR	CPM	CPC
0.00	0.35	0.0000	0.0031	0.0	0.0
0.00	0.35	0.0000	0.0000	0.0	NaN
0.00	0.35	0.0000	0.0000	0.0	NaN
0.00	0.35	0.0000	0.0000	0.0	NaN
0.00	0.35	0.0000	0.0000	0.0	NaN
...	...	...	...	...	...
0.01	0.35	0.0065	NaN	NaN	NaN
0.07	0.35	0.0455	NaN	NaN	NaN
0.09	0.35	0.0585	NaN	NaN	NaN
0.01	0.35	0.0065	NaN	NaN	NaN
0.01	0.35	0.0065	NaN	NaN	NaN

Figure 1 : Clustering dataset

- Head – Top 5 rows

	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display
1	2020-9-2-18	Format1	300	250	75000	Inter223	Web	Mobile	Display
2	2020-9-3-16	Format6	336	250	84000	Inter217	Web	Desktop	Video
3	2020-9-3-2	Format1	300	250	75000	Inter224	Web	Desktop	Display
4	2020-9-3-13	Format1	300	250	75000	Inter225	Video	Mobile	Display

_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
1806	325	323	1	0.0	0.35	0.0	0.0031	0.0	0.0
1979	384	380	0	0.0	0.35	0.0	0.0000	0.0	NaN
1566	298	297	0	0.0	0.35	0.0	0.0000	0.0	NaN
643	103	102	0	0.0	0.35	0.0	0.0000	0.0	NaN
1550	347	345	0	0.0	0.35	0.0	0.0000	0.0	NaN

Figure 2: Head (top 5)

- Tail – Last 5 rows

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

Figure 3: Tail (Last 5)

- Shape – (25857,19)
  The dataset has 25857 rows and 19 columns.
- Info –

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

Figure 4: Data information

There are 6 floats, 7 integers and 6 objects.

- Describe –

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR
count	25857.000000	25857.000000	25857.000000	2.585700e+04	2.585700e+04	2.585700e+04	25857.000000	25857.000000	25857.000000	25857.000000	19392.000000
mean	390.431218	332.182774	99683.276482	2.169621e+06	1.155322e+06	1.107525e+06	9525.881386	2414.473115	0.336729	1716.548955	0.069627
std	230.696051	194.260924	62640.685612	4.542680e+06	2.407244e+06	2.326648e+06	16721.686071	3932.835240	0.030540	2993.025498	0.074970
min	120.000000	70.000000	33600.000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000	0.000000	0.210000	0.000000	0.000000
25%	120.000000	250.000000	72000.000000	9.133000e+03	5.451000e+03	2.558000e+03	305.000000	36.030000	0.350000	23.420000	0.002400
50%	300.000000	300.000000	75000.000000	3.309680e+05	1.894490e+05	1.621620e+05	3457.000000	1173.660000	0.350000	762.880000	0.007700
75%	720.000000	600.000000	84000.000000	2.208484e+06	1.008171e+06	9.496930e+05	10681.000000	2692.280000	0.350000	1749.982000	0.128300
max	728.000000	600.000000	216000.000000	2.759286e+07	1.470202e+07	1.419477e+07	143049.000000	26931.870000	0.350000	21276.180000	1.000000

	CPM	CPC
	19392.000000	18330.000000
	7.252900	0.351061
	6.538314	0.343334
	0.000000	0.000000
	1.630000	0.090000
	3.035000	0.160000
	12.220000	0.570000
	81.560000	7.260000

Figure 5: Data description

- Null values -

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

Figure 6: Null values

There are 6465 missing values in CTR, 6465 in CPM and 7527 in CPC.

- Duplicate values – There are no duplicate values in the dataset.
- Unique values –

```
Timestamp                  2018
InventoryType                 7
Ad - Length                   6
Ad- Width                     5
Ad Size                       7
Ad Type                      14
Platform                      3
Device Type                   2
Format                        2
Available_Impressions     22104
Matched_Queries           20978
Impressions               20454
Clicks                    12753
Spend                     20467
Fee                           7
Revenue                   20578
CTR                        2067
CPM                        2086
CPC                         194
dtype: int64
```

Figure 7: Unique values

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

The missing values were treated by writing a user defined function and calling it. The following formulas were used to treat then.

CPC = Total cost (Spend) / Number of clicks

CPM = (Total campaign spends / Number of impressions) *1000

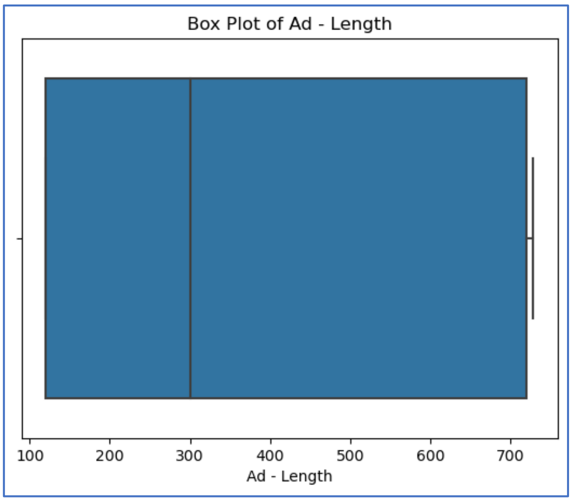
CTR = (Total measured clicks /Total measured ad impressions) *100

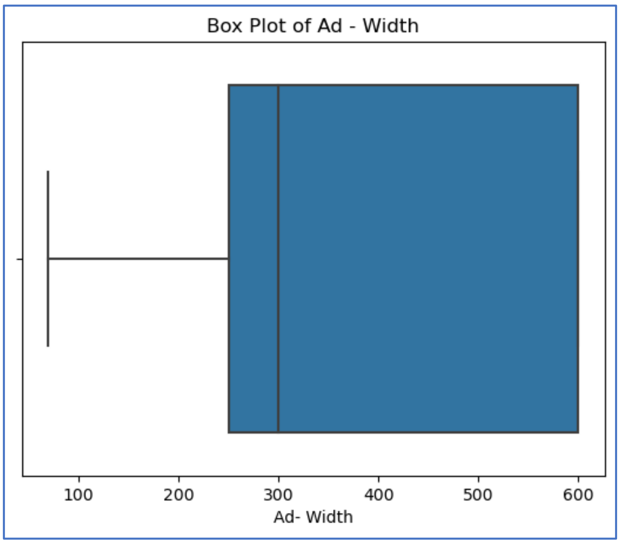
```
Timestamp                0
InventoryType            0
Ad - Length              0
Ad- Width                0
Ad Size                  0
Ad Type                  0
Platform                 0
Device Type              0
Format                   0
Available_Impressions    0
Matched_Queries          0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

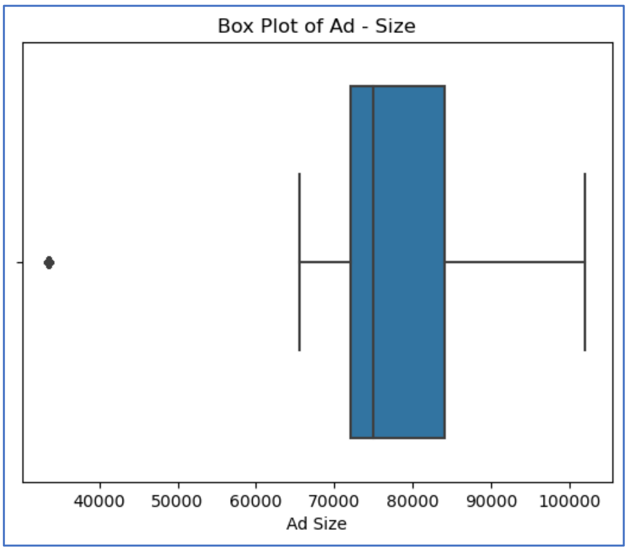
Figure 8: After imputing null values

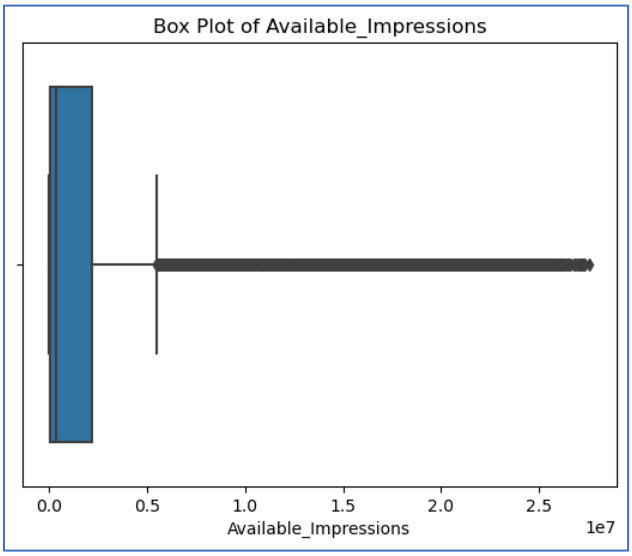
3. Check if there are any outliers.

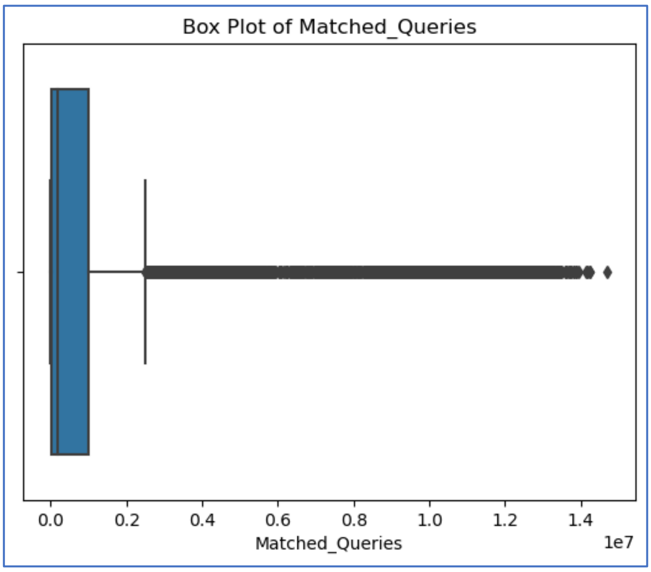
There are no significant outliers in Ad Length and AD width.
The other variables namely Ad size, Available impressions, Matched queries, Clicks, Speed, Free Revenue, CTR, CPM, CPC.

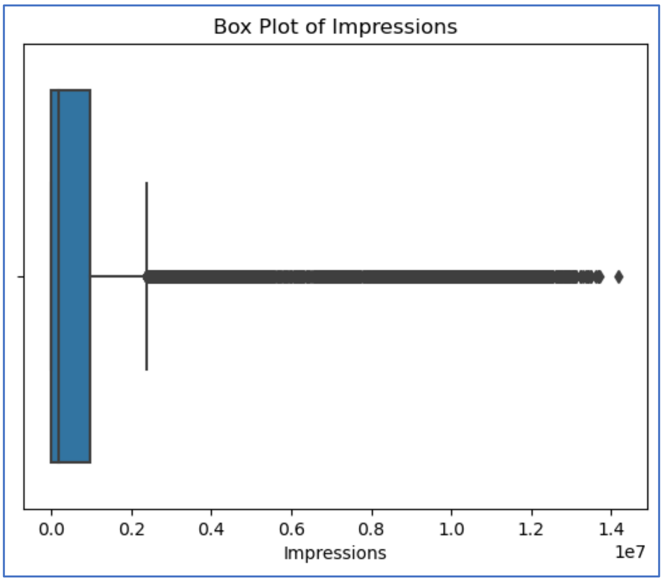


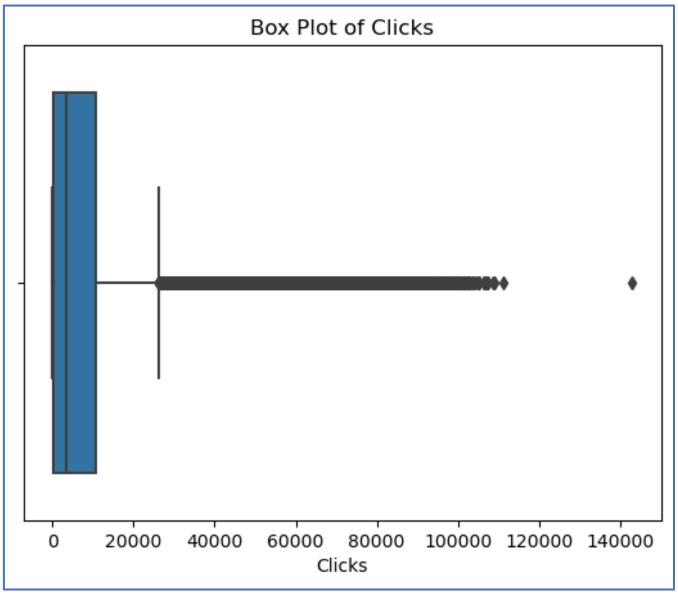


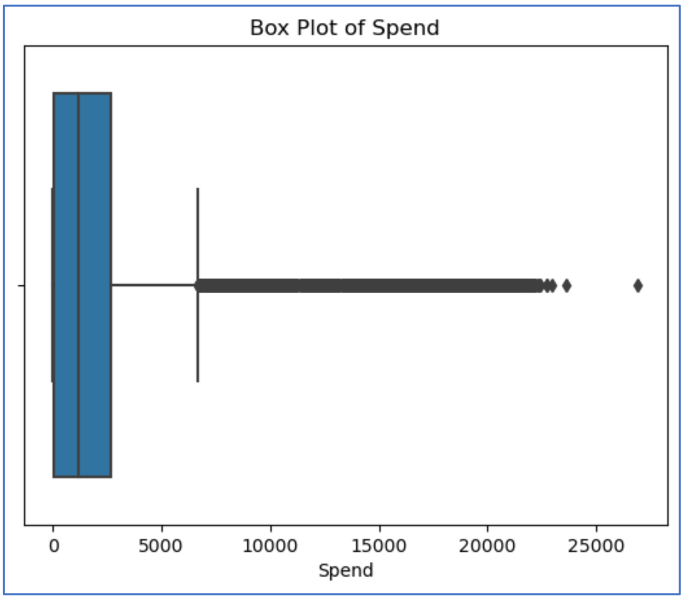


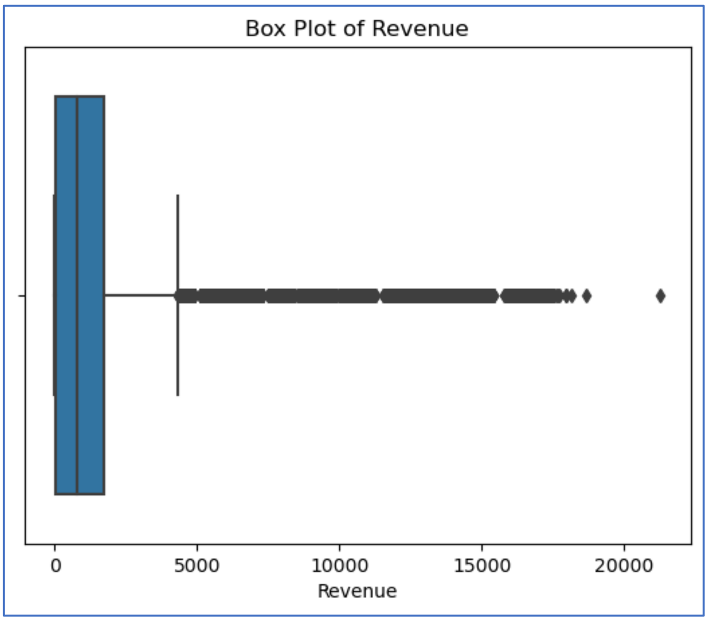

Box Plot of Available_Impressions
0.0
0.5
1.0
1.5
2.0
2.5
Available_Impressions
1e7


Box Plot of Matched_Queries
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Matched_Queries
1e7


Box Plot of Impressions
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Impressions
1e7


Box Plot of Clicks
0
20000
40000
60000
80000
100000
120000
140000
Clicks


Box Plot of Spend
0
5000
10000
15000
20000
25000
Spend


Box Plot of Revenue
0
5000
10000
15000
20000
Revenue

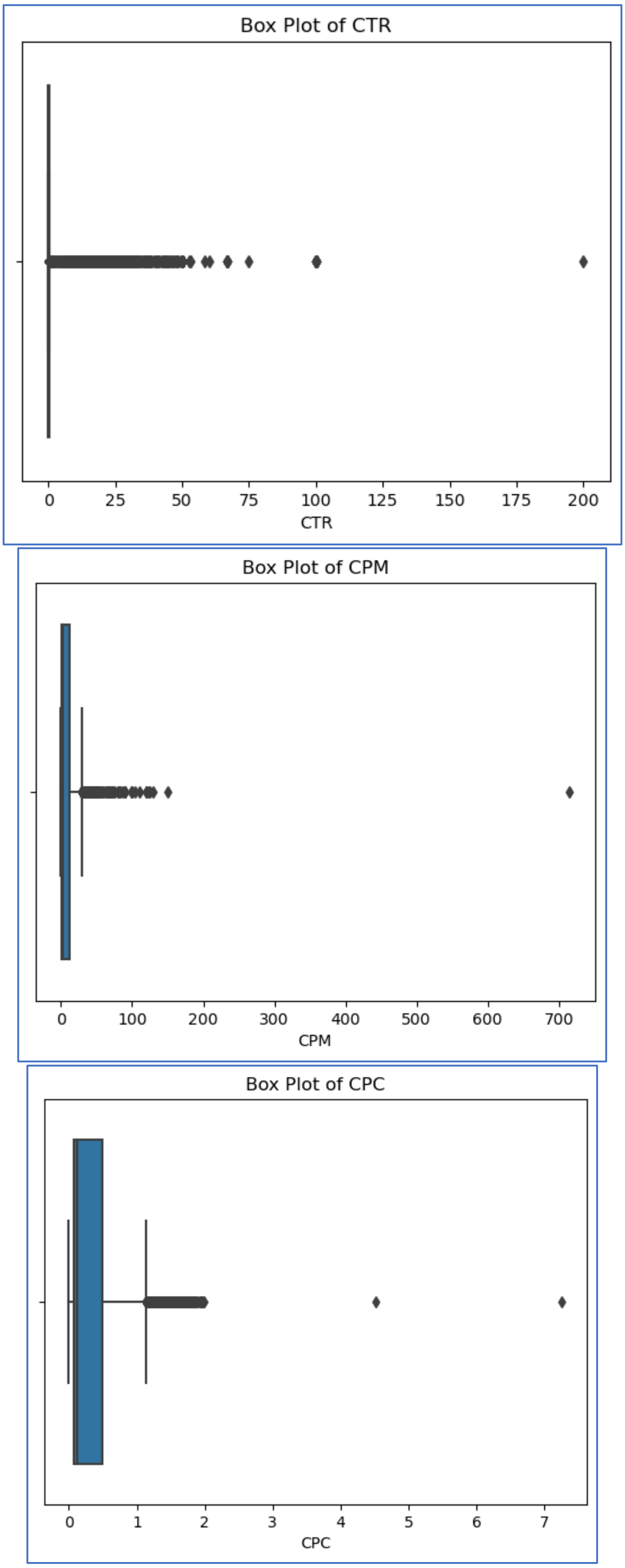


Figure 9: Box plots

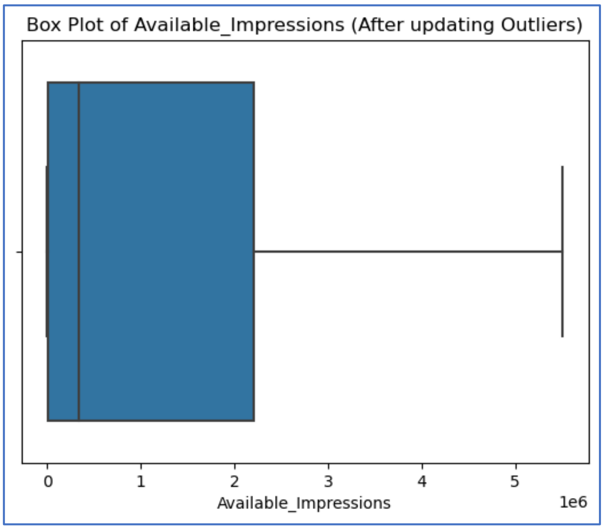
4. Treat Outliers

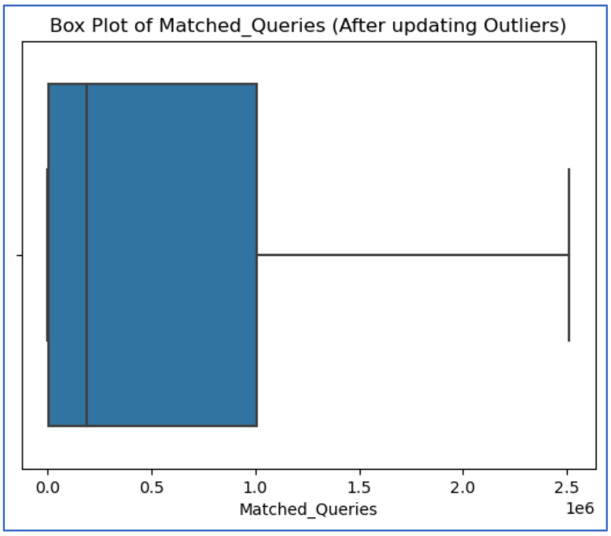
Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an analyst your judgement may be different from another analyst).

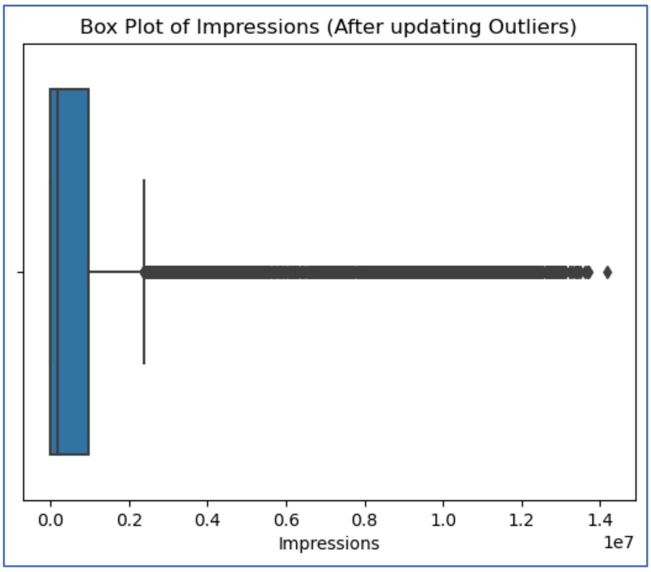
Treating the outliers is extremely important as K-means is highly sensitive to outliers. Therefore, we treat the outliers by the following method:

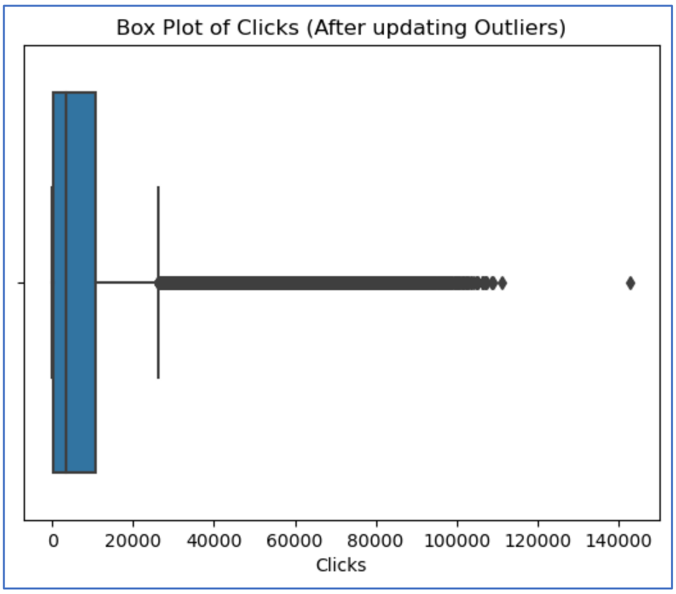
For lower-level outliers we treat it to get at 25th percentile value using the formula q25-(1.5XIQR) and the higher level outliers are treated to get at 95th percentile value using the formula q75+(1.5XIQR).

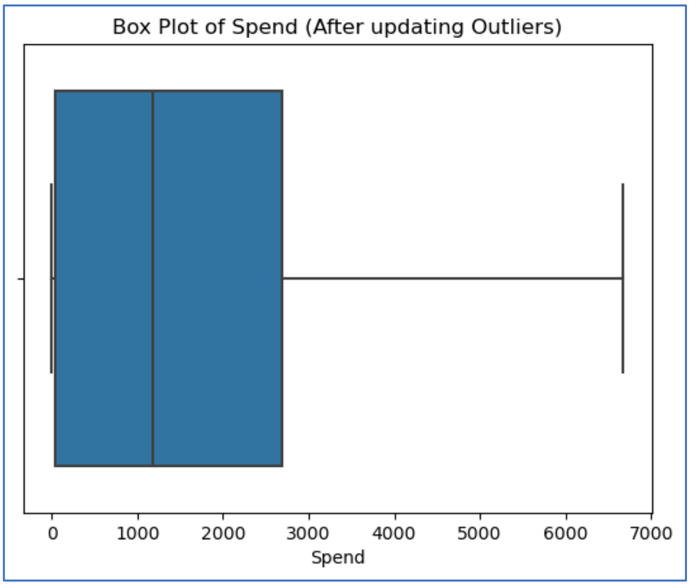
We need to understand that q25 is the median of first half of the data and q75 is the median of second half of the data. IQR (Interquartile range) is the difference of q75 and q25.

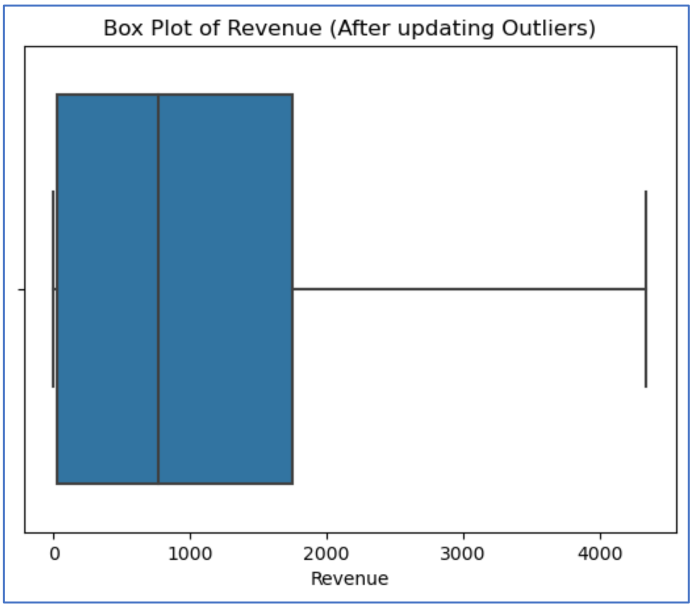


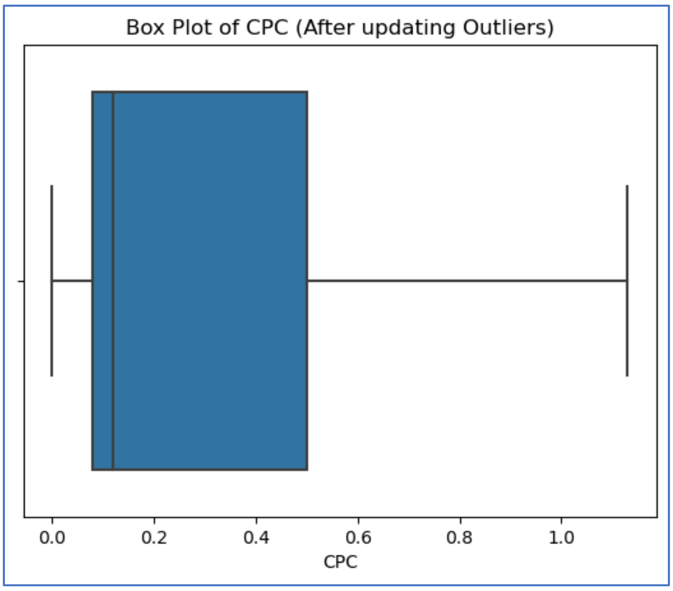


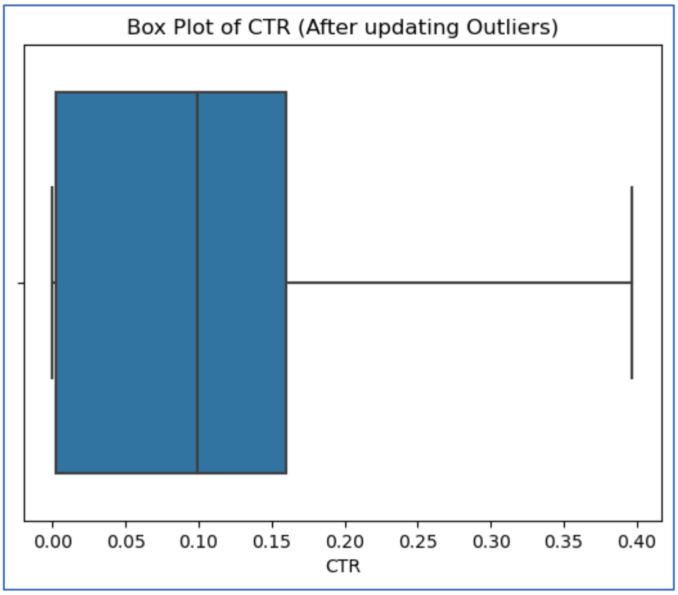

Box Plot of Impressions (After updating Outliers)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Impressions
1e7


Box Plot of Clicks (After updating Outliers)
0
20000
40000
60000
80000
100000
120000
140000
Clicks


Box Plot of Spend (After updating Outliers)
0
1000
2000
3000
4000
5000
6000
7000
Spend


Box Plot of Revenue (After updating Outliers)
0
1000
2000
3000
4000
Revenue


Box Plot of CPC (After updating Outliers)
0.0
0.2
0.4
0.6
0.8
1.0
CPC


Box Plot of CTR (After updating Outliers)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
CTR

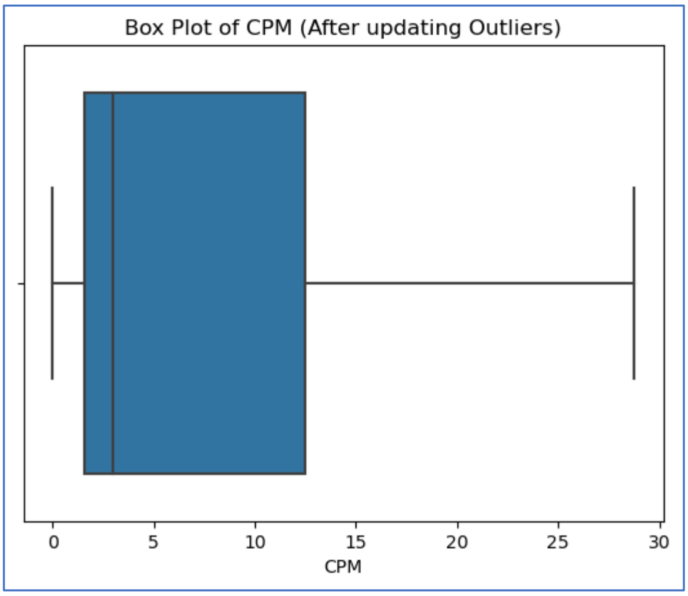


Figure 10: Box plots after treating outliers

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

We use standard scaler to scale the data.

	Ad - Length	Ad- Width	Available_Impressions	Matched_Queries	Impressions	Clicks
0	-0.392000	-0.423062	-0.714953	-0.744816	-0.475888	-0.569624
1	-0.392000	-0.423062	-0.714862	-0.744749	-0.475864	-0.569683
2	-0.235948	-0.423062	-0.715079	-0.744846	-0.475899	-0.569683
3	-0.392000	-0.423062	-0.715566	-0.745066	-0.475983	-0.569683
4	-0.392000	-0.423062	-0.715088	-0.744791	-0.475879	-0.569683
...	...	...	...	...	...	...
25852	1.428612	-0.165671	-0.715905	-0.745182	-0.476026	-0.569683
25853	-1.172263	1.378674	-0.715901	-0.745182	-0.476026	-0.569624
25854	1.428612	-0.165671	-0.715904	-0.745180	-0.476026	-0.569624
25855	-0.392000	1.378674	-0.715905	-0.745182	-0.476026	-0.569683
25856	1.428612	-0.165671	-0.715905	-0.745182	-0.476026	-0.569683

25857 rows × 12 columns

Spend	Fee	Revenue	CTR	CPM	CPC
-0.613939	0.434559	-0.841307	-0.844481	-1.061370	-0.945141
-0.613939	0.434559	-0.841307	-0.867303	-1.061370	-0.945141
-0.613939	0.434559	-0.841307	-0.867303	-1.061370	-0.945141
-0.613939	0.434559	-0.841307	-0.867303	-1.061370	-0.945141
-0.613939	0.434559	-0.841307	-0.867303	-1.061370	-0.945141
...	...	...	...	...	...
-0.613936	0.434559	-0.841303	-0.867303	0.381478	-0.945141
-0.613921	0.434559	-0.841276	2.054648	3.089706	-0.718435
-0.613916	0.434559	-0.841267	2.054648	3.089706	-0.653662
-0.613936	0.434559	-0.841303	-0.867303	0.381478	-0.945141
-0.613936	0.434559	-0.841303	-0.867303	0.381478	-0.945141

Figure 11: Scaled data

Since the scale of the variables become same, the speed of the algorithm increases.

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

A dendrogram was obtained using Ward's method and Euclidean as metric.

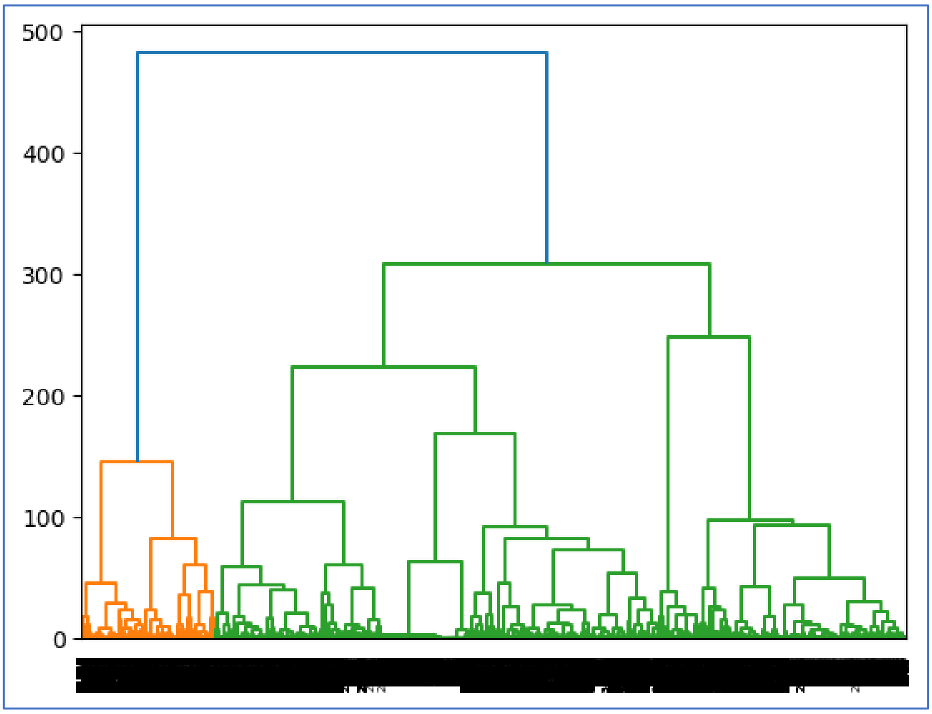


Figure 12: Dendrogram

To view only the last 10 merged clusters, we use truncate with p=10

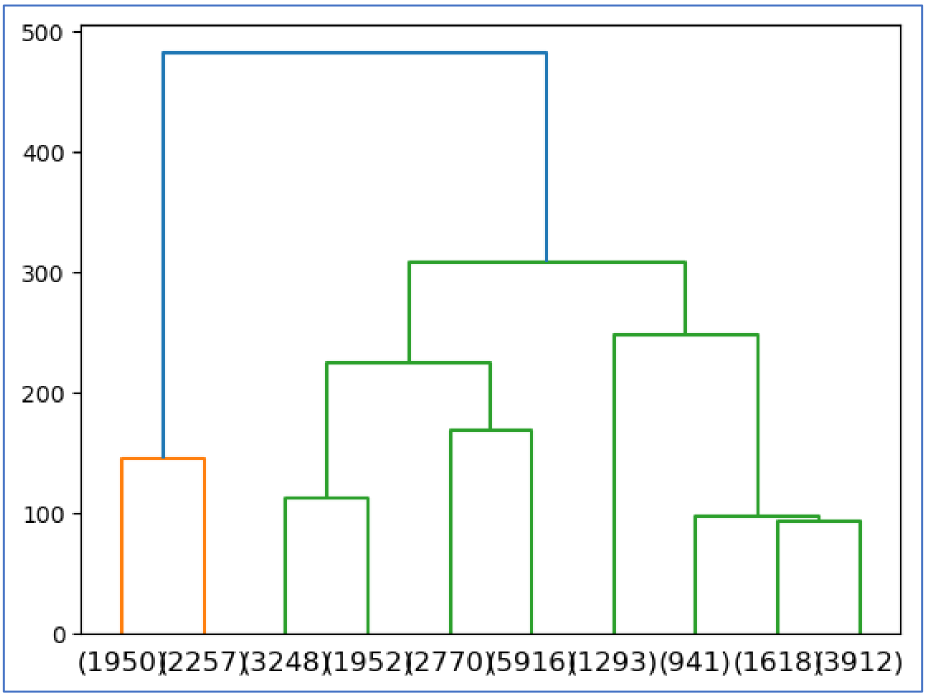


Figure 13: Merged Dendrogram

7. Elbow Plot

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

To build an elbow plot we initially import Kmeans to get WSS (within sum of square) values.

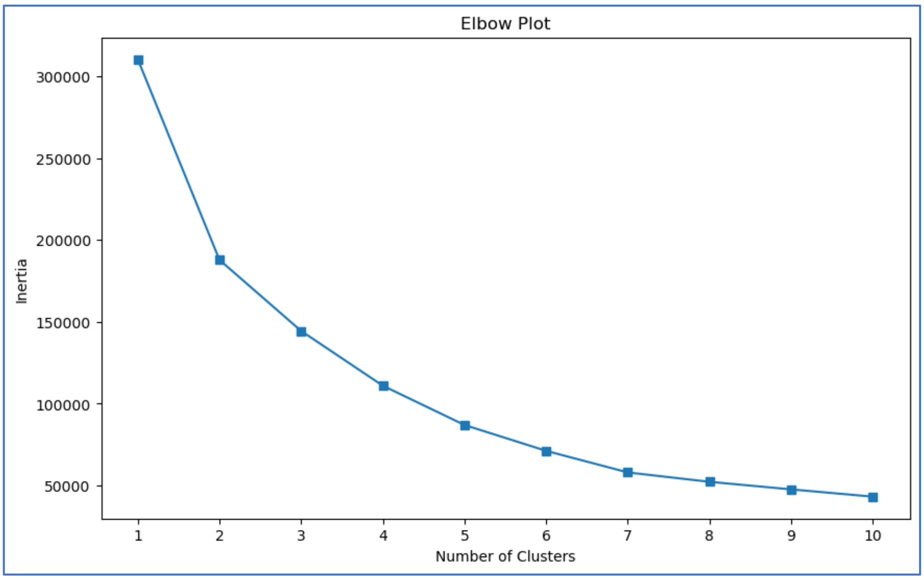


Figure 14: Elbow plot

When we look at WSS values for clusters, we notice that there is a large difference between clusters till cluster 5 but from cluster 5 to 6 there is no big difference.
Thus, optimum number of clusters would be 5.

8. Silhouette score

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

silhouette scores = 0.4249 which is positive. This indicates that the clustering is done properly.

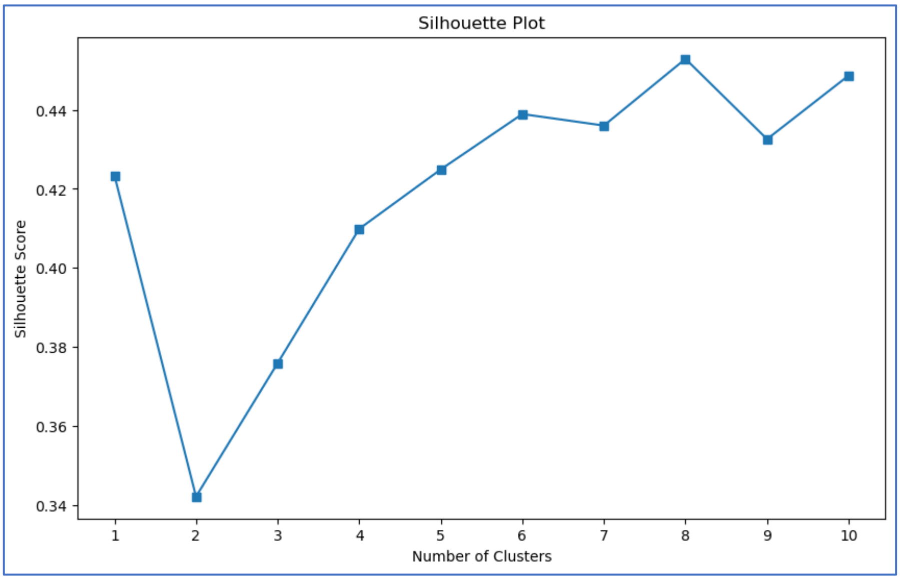
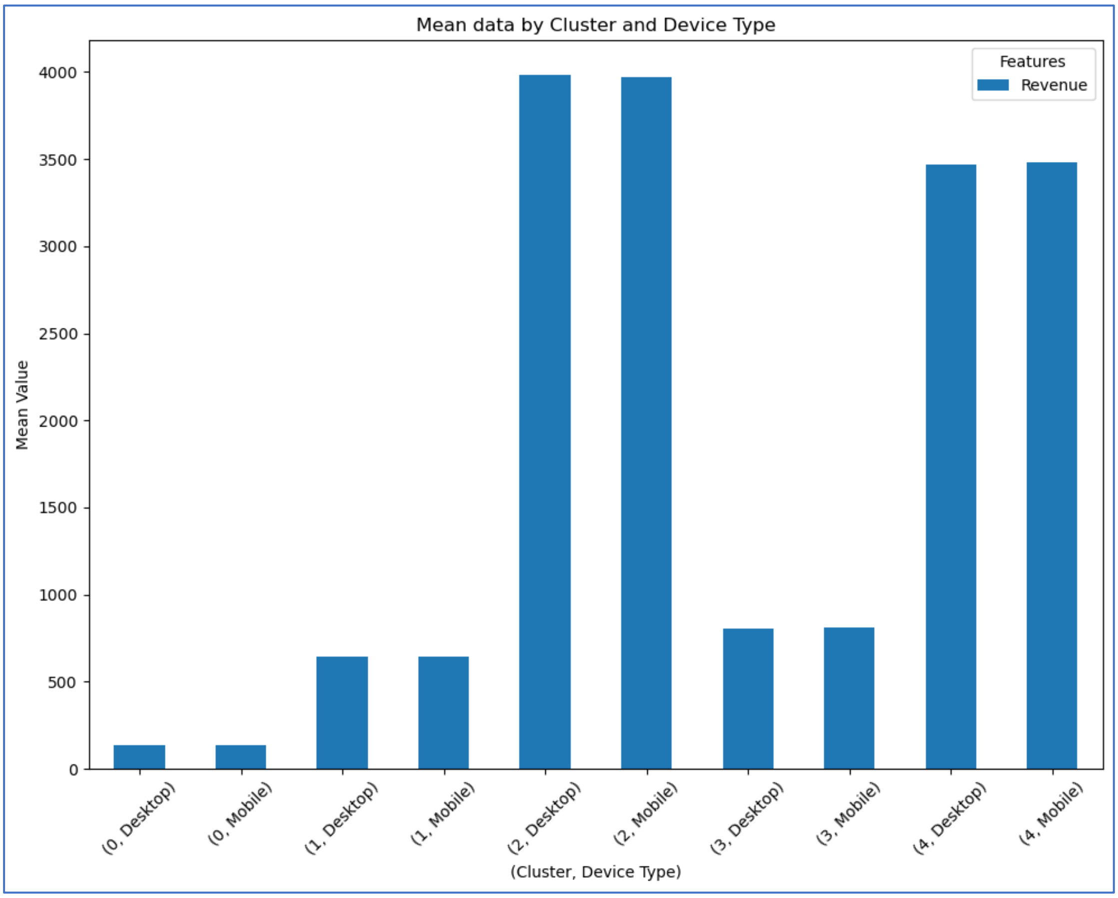


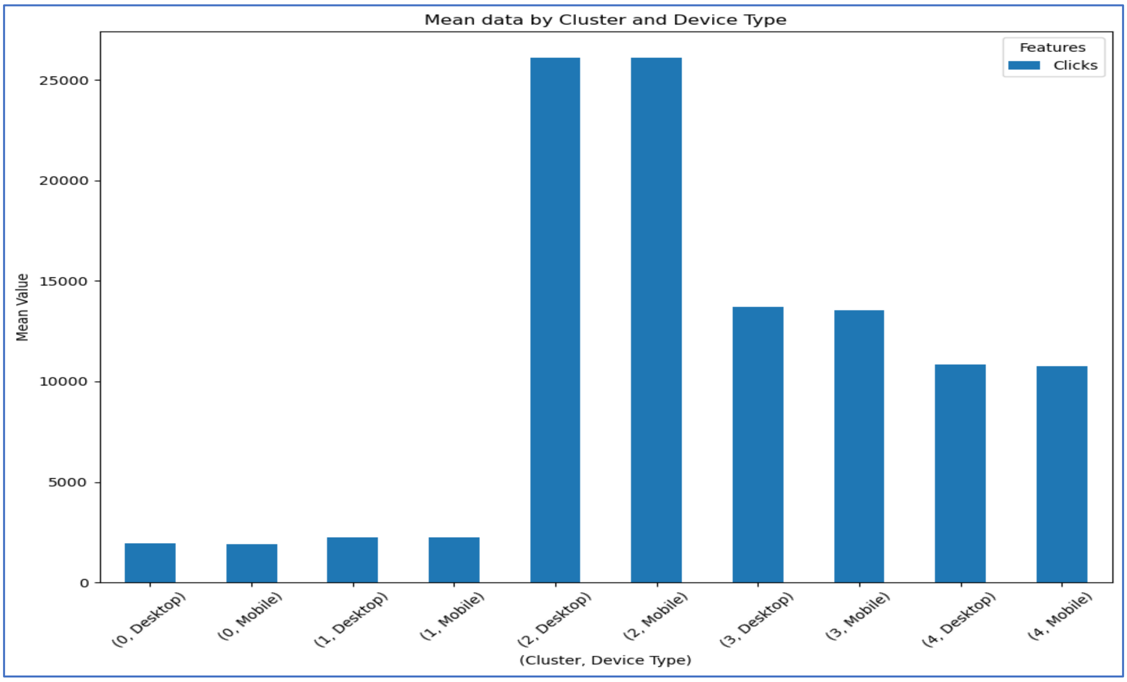
Figure 15: Silhouette plot

9. Profile the ads

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

Bar Plot for ‘Clicks’ ‘Spend’ and ‘Revenue’





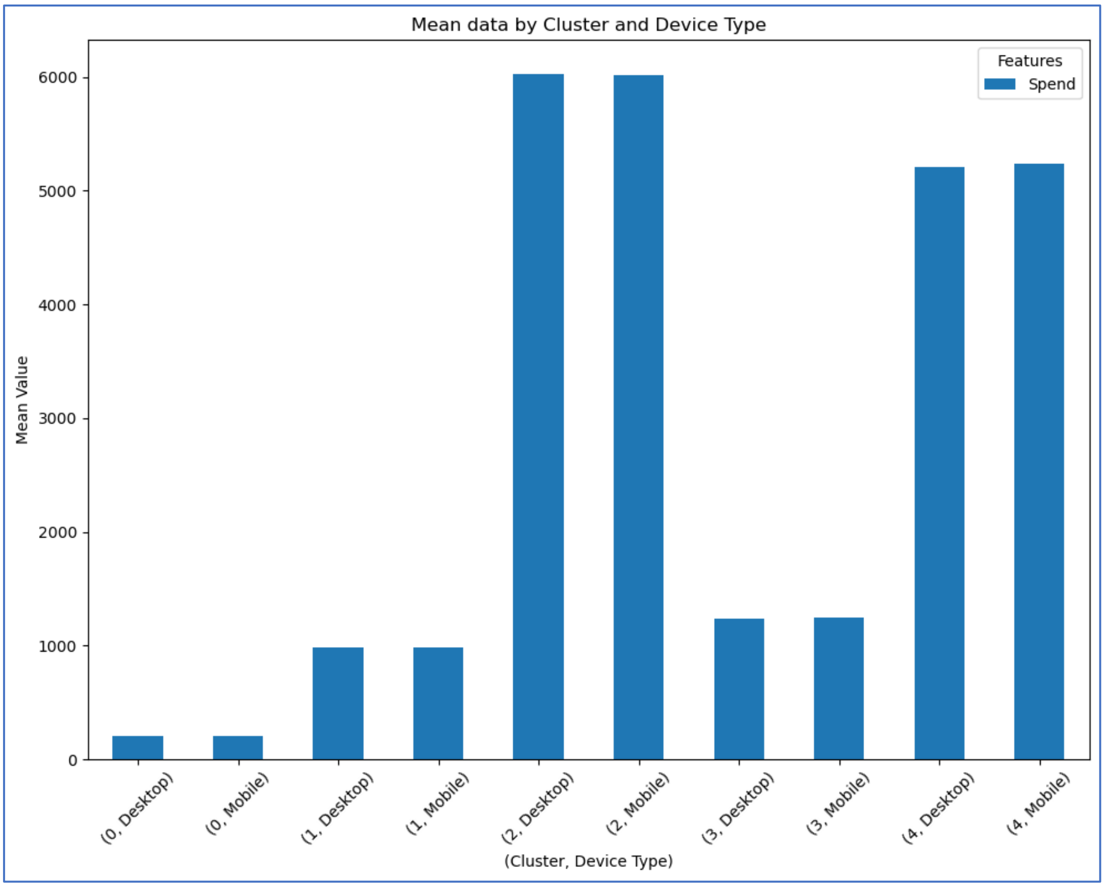
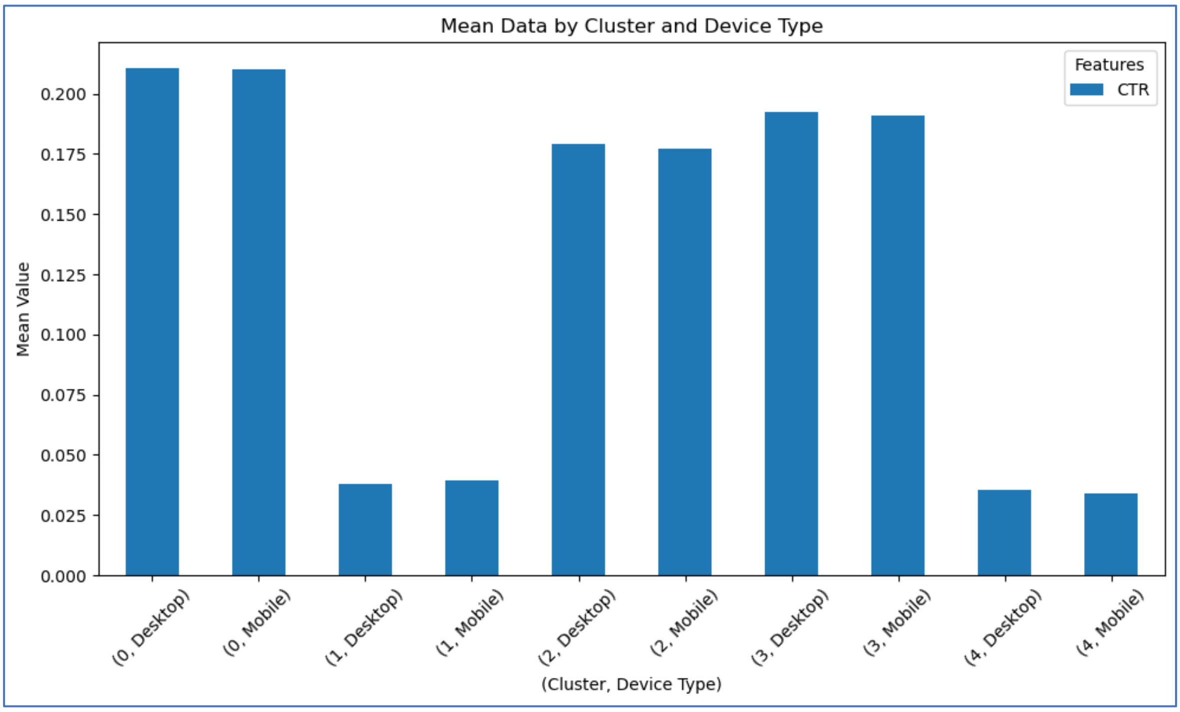
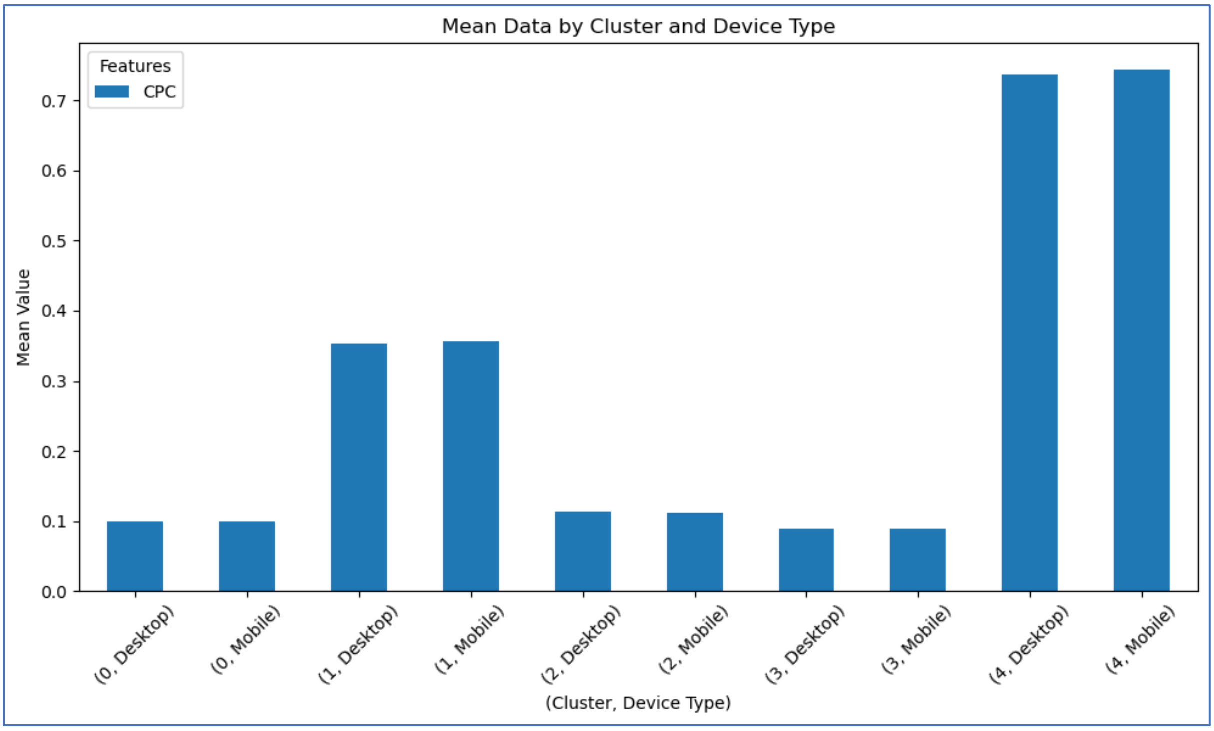


Figure 16: Bar plot for clicks, spend and revenue

Bar Plot for ‘CPC ‘CTR’ and ‘CPM’





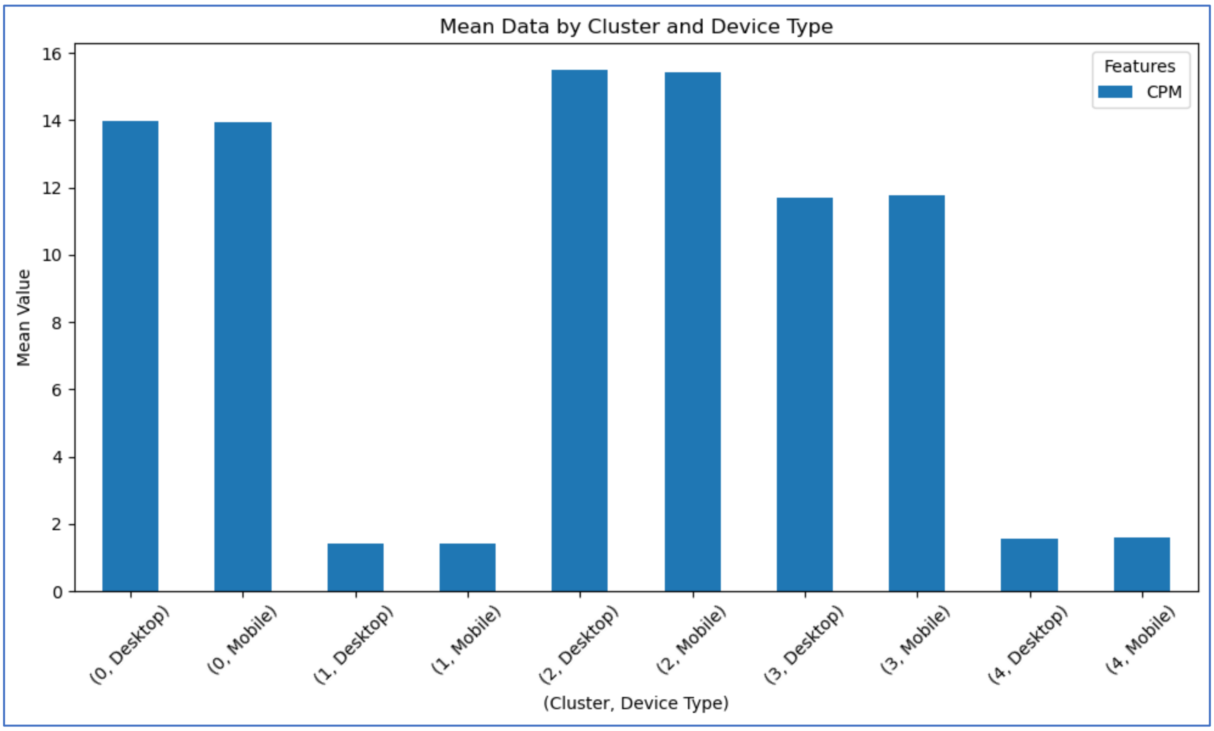


Figure 17: Bar plot for CTR, CPC, CPM

10. Summary:

- Cluster 2 has the highest number of clicks followed by cluster 3 & 4
- Cluster 2 has the highest spend and revenue, followed by cluster 4
- There is not much of a difference between the desktop and the mobile device types when it comes to clicks, spend and revenue and the device type does not make a difference
- Cluster4 has higher revenue relative to the number of clicks
- Cluster 2 has higher CTR & CPM with lower CPC indicating that this cluster could be targeting premium segments or niche audience
- Cluster 0 has higher CTR and lower CPC indicating the efficient spending of the Advertisement budget. The current spend is low, and if higher spending is done on cluster 0 it could possibly generate more revenue.
- Cluster 1 has lower CPM and CTR with higher CPC indicates that the spend on the advertisement is not justified and the advertisement strategy is not efficient
- Cluster 3 has higher CTR with lower CPC indicating that the advertisement is reaching the right audience and increasing the spend could help improve the revenue

Problem Statement – 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.

Basic Data Analysis:

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

- 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	M
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	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
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	

640 rows × 61 columns

MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
180	237	680	252	32	46	258	214
123	229	186	148	76	178	140	160
44	89	3	34	0	4	67	61
61	128	13	50	4	10	116	59
465	1043	205	302	24	105	180	478
...	...	...	...	...	...	...	...
0	0	0	0	0	0	32	47
3	14	38	130	4	23	110	170
9	4	2	6	17	47	76	77
24	44	11	21	1	4	100	103
6	2	17	17	2	4	148	99

Figure 18: PCA dataset

- Head – Top 5 rows

	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180	237	680	252	32	46	258	214
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123	229	186	148	76	178	140	160
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	...	114	188	44	89	3	34	0	4	67	61
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	...	194	247	61	128	13	50	4	10	116	59
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	...	874	1928	465	1043	205	302	24	105	180	478

Figure 19: Head (top 5)

- Tail – Last 5 rows

	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
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

MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
0	0	0	32	47
130	4	23	110	170
6	17	47	76	77
21	1	4	100	103
17	2	4	148	99

Figure 20: Tail (Last 5)

- Shape – (640,61)
  The dataset has 640 rows and 61 columns.
- Info –

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 640 entries, 0 to 639
Data columns (total 61 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   State Code      640 non-null    int64
 1   Dist.Code       640 non-null    int64
 2   State           640 non-null    object
 3   Area Name       640 non-null    object
 4   No_HH           640 non-null    int64
 5   TOT_M           640 non-null    int64
 6   TOT_F           640 non-null    int64
 7   M_06            640 non-null    int64
 8   F_06            640 non-null    int64
 9   M_SC            640 non-null    int64
 10  F_SC            640 non-null    int64
 11  M_ST            640 non-null    int64
 12  F_ST            640 non-null    int64
 13  M_LIT           640 non-null    int64
 14  F_LIT           640 non-null    int64
 15  M_ILL           640 non-null    int64
 16  F_ILL           640 non-null    int64
 17  TOT_WORK_M      640 non-null    int64
 18  TOT_WORK_F      640 non-null    int64
 19  MAINWORK_M      640 non-null    int64
 20  MAINWORK_F      640 non-null    int64
 21  MAIN_CL_M       640 non-null    int64
 22  MAIN_CL_F       640 non-null    int64
 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
```

Figure 21: Dataset Information

There are 59 integers and 2 objects.

- Null values – There are no null values in the given data

```
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
```

Figure 22: Null values

- Duplicate values – There are no duplicate values in the dataset.

```
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
```

```
df.duplicated().sum()
```

```
0
```

Figure 23: Duplicate values

- Describe –

	State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST
count	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000
mean	17.114062	320.500000	51222.871875	79940.576563	122372.084375	12309.098438	11942.300000	13820.946875	20778.392188	6191.807813
std	9.426486	184.896367	48135.405475	73384.511114	113600.717282	11500.906881	11326.294567	14426.373130	21727.887713	9912.668948
min	1.000000	1.000000	350.000000	391.000000	698.000000	56.000000	56.000000	0.000000	0.000000	0.000000
25%	9.000000	160.750000	19484.000000	30228.000000	46517.750000	4733.750000	4672.250000	3466.250000	5603.250000	293.750000
50%	18.000000	320.500000	35837.000000	58339.000000	87724.500000	9159.000000	8663.000000	9591.500000	13709.000000	2333.500000
75%	24.000000	480.250000	68892.000000	107918.500000	164251.750000	16520.250000	15902.250000	19429.750000	29180.000000	7658.000000
max	35.000000	640.000000	310450.000000	485417.000000	750392.000000	96223.000000	95129.000000	103307.000000	156429.000000	96785.000000

MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000
2757.050000	250.889062	558.098438	560.690625	1293.431250	71.379688	200.742188	510.014063	704.778125
2788.776676	453.336594	1117.642748	762.578991	1585.377936	107.897627	309.740854	610.603187	910.209225
30.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	5.000000
957.250000	47.000000	109.000000	136.500000	298.000000	14.000000	43.000000	161.000000	220.500000
1928.000000	114.500000	247.500000	308.000000	717.000000	35.000000	113.000000	326.000000	464.500000
3599.750000	270.750000	568.750000	642.000000	1710.750000	79.000000	240.000000	604.500000	853.500000
21611.000000	5775.000000	17153.000000	6116.000000	13714.000000	895.000000	3354.000000	6456.000000	10533.000000

Figure 24: Dataset description

The summary of the data shows that each column have varied range of data hence scaling would be needed to treat all the column with equal weightage.

Exploratory Data Analysis

2. Perform detailed Exploratory analysis by creating certain questions like the given example. Pick 5 variables out of the given 20 variables below.

We start by Removing Categorical Variables such as state code, dist code, state, and area name. Thus, now the dataframe has 57 columns all numerical and highly correlated among each other.

	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_
No_HH	1.000000	0.916170	0.970590	0.797559	0.796373	0.775309	0.823847	0.149627	0.165102	0.931938	...	0.556941	0.555
TOT_M	0.916170	1.000000	0.982640	0.950825	0.947792	0.839925	0.826299	0.091421	0.086180	0.989312	...	0.698310	0.595
TOT_F	0.970590	0.982640	1.000000	0.907975	0.906557	0.816959	0.832756	0.123626	0.128646	0.985441	...	0.655347	0.598
M_06	0.797559	0.950825	0.907975	1.000000	0.998151	0.781120	0.747530	0.055274	0.043948	0.912757	...	0.760610	0.646
F_06	0.796373	0.947792	0.906557	0.998151	1.000000	0.773135	0.741686	0.065138	0.054662	0.907641	...	0.763614	0.649
M_SC	0.775309	0.839925	0.816959	0.781120	0.773135	1.000000	0.985071	-0.045666	-0.047825	0.818484	...	0.673633	0.569
F_SC	0.823847	0.826299	0.832756	0.747530	0.741686	0.985071	1.000000	-0.014122	-0.009190	0.814150	...	0.650455	0.585
M_ST	0.149627	0.091421	0.123626	0.055274	0.065138	-0.045666	-0.014122	1.000000	0.988047	0.090541	...	0.122967	0.196
F_ST	0.165102	0.086180	0.128646	0.043948	0.054662	-0.047825	-0.009190	0.988047	1.000000	0.087375	...	0.121411	0.216
M_LIT	0.931938	0.989312	0.985441	0.912757	0.907641	0.818484	0.814150	0.090541	0.087375	1.000000	...	0.652507	0.560
F_LIT	0.928087	0.931708	0.957012	0.832509	0.829128	0.713939	0.728755	0.100488	0.100892	0.967956	...	0.547296	0.484
M_ILL	0.763041	0.911539	0.858199	0.945409	0.948609	0.800775	0.762560	0.083063	0.072589	0.841835	...	0.744658	0.625
F_ILL	0.862074	0.885361	0.886917	0.863324	0.865289	0.832714	0.847203	0.138031	0.149493	0.834384	...	0.708454	0.672
TOT_WORK_M	0.938199	0.970417	0.968955	0.855773	0.852793	0.824773	0.823689	0.122643	0.119171	0.976952	...	0.600872	0.514
TOT_WORK_F	0.925259	0.807895	0.876233	0.683494	0.685348	0.712971	0.776930	0.264749	0.284974	0.815368	...	0.492828	0.548
MAINWORK_M	0.926629	0.932832	0.941016	0.789694	0.784789	0.778492	0.782346	0.113607	0.109313	0.953300	...	0.472297	0.393
MAINWORK_F	0.891306	0.744368	0.822822	0.584979	0.585683	0.644142	0.712874	0.230810	0.246321	0.768040	...	0.302859	0.339
MAIN_CL_M	0.431402	0.531734	0.487657	0.561164	0.561599	0.608157	0.578519	0.099264	0.083126	0.468564	...	0.464847	0.391
MAIN_CL_F	0.382680	0.355887	0.385373	0.381994	0.383296	0.360798	0.388412	0.194493	0.199128	0.329627	...	0.309749	0.372
MAIN_AL_M	0.673638	0.593420	0.623724	0.549857	0.554182	0.625566	0.673433	0.142582	0.153601	0.543861	...	0.379632	0.394
MAIN_AL_F	0.585856	0.379748	0.472748	0.296250	0.298385	0.408169	0.507721	0.198814	0.226505	0.369241	...	0.115093	0.226
MAIN_HH_M	0.641375	0.740354	0.700957	0.659762	0.657411	0.705651	0.675973	-0.029207	-0.032221	0.725728	...	0.528910	0.405
MAIN_HH_F	0.490908	0.443512	0.466299	0.354727	0.357130	0.392656	0.418929	0.033600	0.036628	0.444565	...	0.243561	0.230

MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
0.555543	0.067425	0.046128	0.368591	0.417447	0.486747	0.536854	0.762384	0.735692
0.595696	0.167405	0.115580	0.495928	0.440360	0.651604	0.588180	0.844896	0.716061
0.598951	0.138763	0.099438	0.451011	0.443132	0.593805	0.571853	0.827653	0.746583
0.646998	0.266674	0.198338	0.601090	0.514055	0.690601	0.565030	0.784961	0.651143
0.649834	0.258670	0.189568	0.611568	0.523270	0.698636	0.574178	0.783727	0.651439
0.569579	0.184332	0.129750	0.523450	0.461898	0.664918	0.591437	0.735399	0.580020
0.585690	0.163427	0.116228	0.508006	0.488657	0.628470	0.589346	0.720099	0.600089
0.196571	0.027219	0.007077	0.126336	0.238813	-0.005482	0.090123	0.122986	0.146954
0.216741	0.017205	0.002556	0.136396	0.273307	-0.005880	0.099984	0.114131	0.150869
0.560065	0.144067	0.101355	0.421762	0.381727	0.600120	0.552371	0.852199	0.738672
0.484288	0.086819	0.063286	0.289115	0.272713	0.475349	0.464936	0.825726	0.772873
0.625526	0.213365	0.141987	0.646265	0.553483	0.719513	0.619406	0.724094	0.567371
0.672825	0.196296	0.138960	0.626967	0.634091	0.674655	0.638311	0.680465	0.569174
0.514116	0.074774	0.038574	0.386357	0.354322	0.573408	0.574262	0.826607	0.715260
0.548626	0.114920	0.102935	0.343976	0.457268	0.441843	0.490710	0.609380	0.591090
0.393265	-0.006430	-0.029046	0.245966	0.230231	0.463524	0.490312	0.767959	0.672722
0.339516	-0.029158	-0.030386	0.153024	0.260022	0.296494	0.365233	0.517041	0.500361
0.391683	0.239773	0.180130	0.430165	0.357735	0.458240	0.388795	0.337879	0.223493
0.372705	0.372446	0.354491	0.226309	0.293515	0.242593	0.183686	0.153681	0.132902
0.394190	-0.041769	-0.070824	0.429702	0.477320	0.369820	0.440337	0.355209	0.313488
0.226032	-0.107148	-0.101129	0.157280	0.333030	0.108155	0.171380	0.144808	0.178329
0.405362	0.075198	0.033128	0.376269	0.283512	0.733832	0.667100	0.634974	0.480480
0.230927	-0.026913	-0.043182	0.171140	0.153591	0.343462	0.570182	0.339777	0.299006

MARG_HH_F	0.538261	0.579697	0.565228	0.542395	0.550840	0.570181	0.567445	0.093886	0.103309	0.545231	...	0.628194	0.582
MARG_OT_M	0.800971	0.892429	0.871874	0.835365	0.834838	0.747179	0.728023	0.071022	0.062780	0.893902	...	0.712945	0.589
MARG_OT_F	0.833235	0.839530	0.857518	0.744644	0.745304	0.685561	0.694150	0.121542	0.123015	0.856023	...	0.603486	0.562
MARGWORK_3_6_M	0.846809	0.974313	0.943051	0.988824	0.985903	0.809307	0.784708	0.057429	0.050816	0.948130	...	0.751989	0.640
MARGWORK_3_6_F	0.914147	0.983256	0.976391	0.936791	0.934017	0.798961	0.791918	0.050588	0.048396	0.983740	...	0.676367	0.574
MARG_CL_3_6_M	0.692377	0.819869	0.778856	0.859413	0.866289	0.751967	0.730993	0.117438	0.119275	0.763697	...	0.917874	0.826
MARG_CL_3_6_F	0.723436	0.715267	0.736418	0.716630	0.721351	0.662006	0.694128	0.280432	0.307672	0.675965	...	0.789792	0.877
MARG_AL_3_6_M	0.210641	0.352883	0.306733	0.472981	0.475943	0.344283	0.316077	0.116489	0.108795	0.298784	...	0.767242	0.770
MARG_AL_3_6_F	0.094437	0.165073	0.149338	0.253610	0.247161	0.161941	0.148568	0.080390	0.077766	0.143411	...	0.493718	0.594
MARG_HH_3_6_M	0.448382	0.546598	0.507514	0.641075	0.653328	0.565129	0.556232	0.139290	0.153371	0.457879	...	0.800031	0.757
MARG_HH_3_6_F	0.496975	0.436428	0.470992	0.466309	0.474994	0.451241	0.505766	0.322888	0.365332	0.380993	...	0.563583	0.697
MARG_OT_3_6_M	0.501329	0.664219	0.603306	0.690955	0.697567	0.663500	0.625319	-0.016630	-0.018620	0.612484	...	0.772286	0.634
MARG_OT_3_6_F	0.532138	0.569662	0.556009	0.527947	0.536047	0.555866	0.552942	0.094028	0.103185	0.536081	...	0.591026	0.547
MARGWORK_0_3_M	0.798661	0.890768	0.869805	0.834998	0.834619	0.740155	0.720465	0.059681	0.051667	0.891050	...	0.693382	0.570
MARGWORK_0_3_F	0.829153	0.842307	0.856203	0.742728	0.743463	0.689118	0.694251	0.110104	0.110856	0.856519	...	0.575730	0.520
MARG_CL_0_3_M	0.556941	0.698310	0.655347	0.760610	0.763614	0.673633	0.650455	0.122967	0.121411	0.652507	...	1.000000	0.916
MARG_CL_0_3_F	0.555543	0.595696	0.598951	0.646998	0.649834	0.569579	0.585690	0.196571	0.216741	0.560065	...	0.916765	1.000
MARG_AL_0_3_M	0.067425	0.167405	0.138763	0.266674	0.258670	0.184332	0.163427	0.027219	0.017205	0.144067	...	0.585284	0.624
MARG_AL_0_3_F	0.046128	0.115580	0.099438	0.198338	0.189568	0.129750	0.116228	0.007077	0.002556	0.101355	...	0.494473	0.587
MARG_HH_0_3_M	0.368591	0.495928	0.451011	0.601090	0.611568	0.523450	0.508006	0.126336	0.136396	0.421762	...	0.875141	0.805
MARG_HH_0_3_F	0.417447	0.440360	0.443132	0.514055	0.523270	0.461898	0.488657	0.238813	0.273307	0.381727	...	0.766629	0.846
MARG_OT_0_3_M	0.486747	0.651604	0.593805	0.690601	0.698636	0.664918	0.628470	-0.005482	-0.005880	0.600120	...	0.829396	0.691
MARG_OT_0_3_F	0.536854	0.588180	0.571853	0.565030	0.574178	0.591437	0.589346	0.090123	0.099984	0.552371	...	0.713785	0.666
NON_WORK_M	0.762384	0.844896	0.827653	0.784961	0.783727	0.735399	0.720099	0.122986	0.114131	0.852199	...	0.765673	0.644
NON_WORK_F	0.735692	0.716061	0.746583	0.651143	0.651439	0.580020	0.600089	0.146954	0.150869	0.738672	...	0.623510	0.641

0.582993	0.145239	0.086603	0.578509	0.495631	0.760351	0.973383	0.567936	0.485369
0.589422	0.145294	0.098954	0.451809	0.374776	0.676455	0.593688	0.947725	0.829610
0.562354	0.084207	0.051721	0.344556	0.331405	0.571545	0.556523	0.878514	0.892865
0.640050	0.244887	0.181250	0.571706	0.497148	0.689823	0.569824	0.816939	0.678657
0.574138	0.138476	0.090005	0.463321	0.401748	0.614960	0.562999	0.860038	0.757237
0.826908	0.341167	0.262900	0.834578	0.749911	0.839817	0.733505	0.795369	0.654953
0.877599	0.402389	0.366206	0.703379	0.808523	0.662709	0.642034	0.632577	0.612455
0.770425	0.845747	0.793623	0.643325	0.568238	0.539538	0.419828	0.345163	0.253401
0.594711	0.915799	0.936248	0.272519	0.281378	0.226589	0.145112	0.144225	0.133034
0.757824	0.258258	0.173687	0.933827	0.881058	0.716579	0.645532	0.467244	0.354341
0.697808	0.148573	0.112136	0.693170	0.880890	0.464583	0.450024	0.316894	0.312860
0.634494	0.250372	0.174700	0.698978	0.559043	0.952411	0.807399	0.657035	0.481021
0.547196	0.125913	0.070179	0.538825	0.458845	0.729399	0.952273	0.546636	0.467115
0.570918	0.134291	0.088836	0.439092	0.363350	0.669358	0.583032	0.925301	0.808869
0.520033	0.064555	0.030645	0.328150	0.309687	0.574082	0.545667	0.845428	0.830600
0.916765	0.585284	0.494473	0.875141	0.766629	0.829396	0.713785	0.765673	0.623510
1.000000	0.624517	0.587746	0.805723	0.846438	0.691143	0.666214	0.644600	0.641180
0.624517	1.000000	0.958797	0.354791	0.301234	0.291239	0.196213	0.190933	0.144694
0.587746	0.958797	1.000000	0.251897	0.233344	0.208517	0.131240	0.143090	0.121797
0.805723	0.354791	0.251897	1.000000	0.902658	0.769568	0.673186	0.486815	0.358030
0.846438	0.301234	0.233344	0.902658	1.000000	0.629415	0.584842	0.408173	0.366075
0.691143	0.291239	0.208517	0.769568	0.629415	1.000000	0.823170	0.669458	0.485128
0.666214	0.196213	0.131240	0.673186	0.584842	0.823170	1.000000	0.609569	0.521097
0.644600	0.190933	0.143090	0.486815	0.408173	0.669458	0.609569	1.000000	0.880902
0.641180	0.144694	0.121797	0.358030	0.366075	0.485128	0.521097	0.880902	1.000000

Figure 25: Correlation table

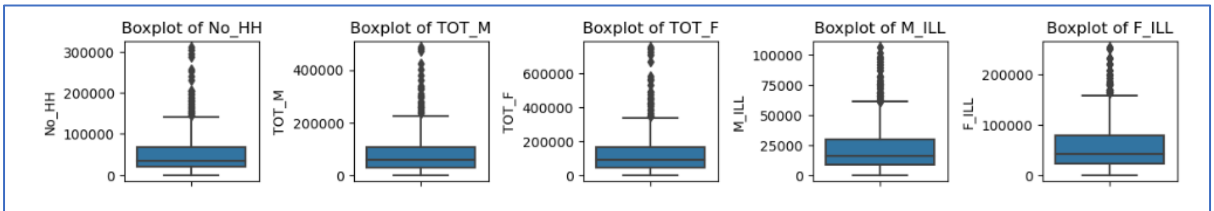


Figure 26: Boxplots of any 5 variables

- All of the chosen columns have several outliers
- As there are many outliers we decide not to treat them as it would be too much data modification and we might arrive at an inaccurate analysis.

State Wise Analysis for the chosen columns

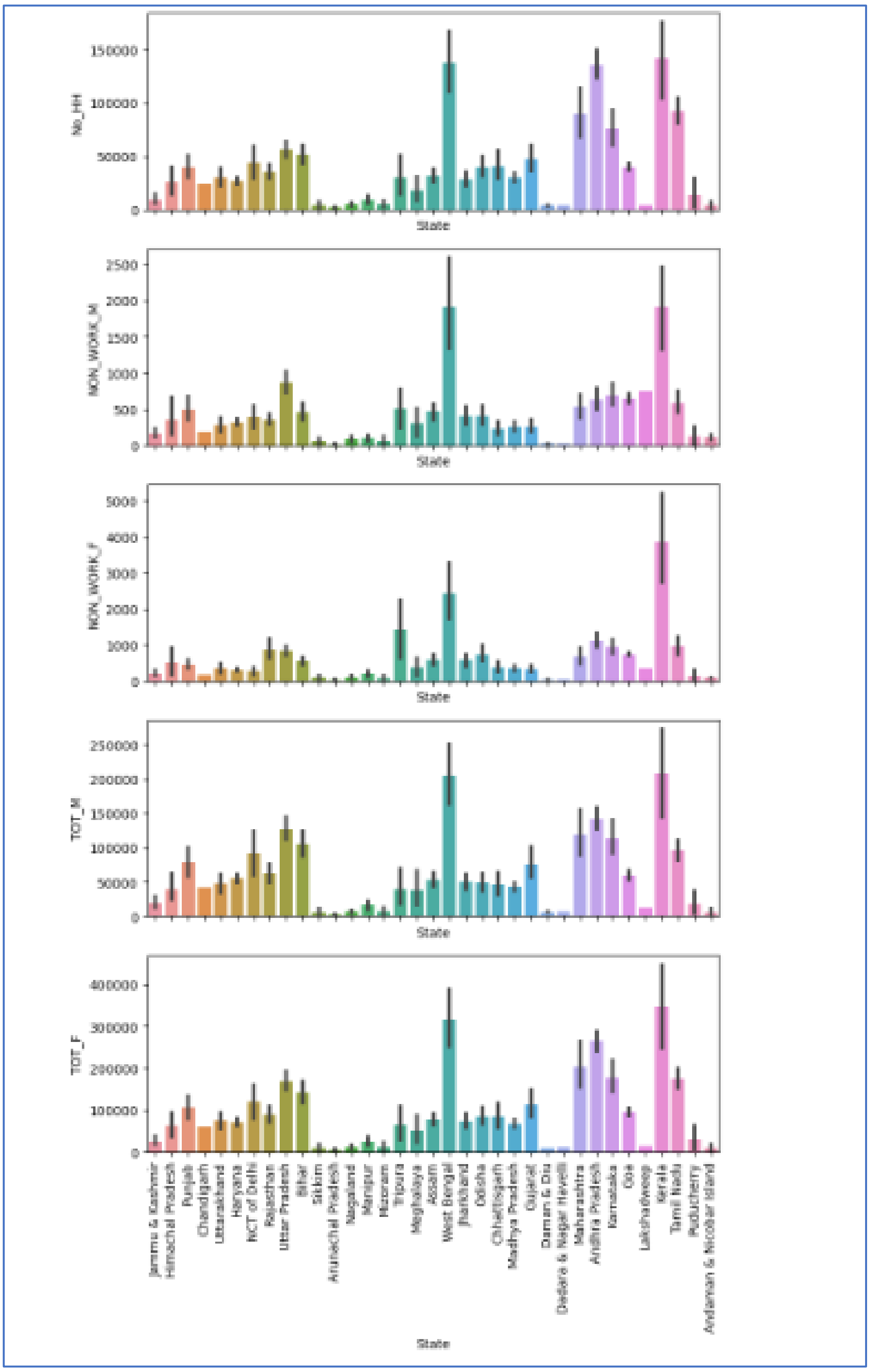


Figure 27: Barplot of any 5 variables

Pair Plot

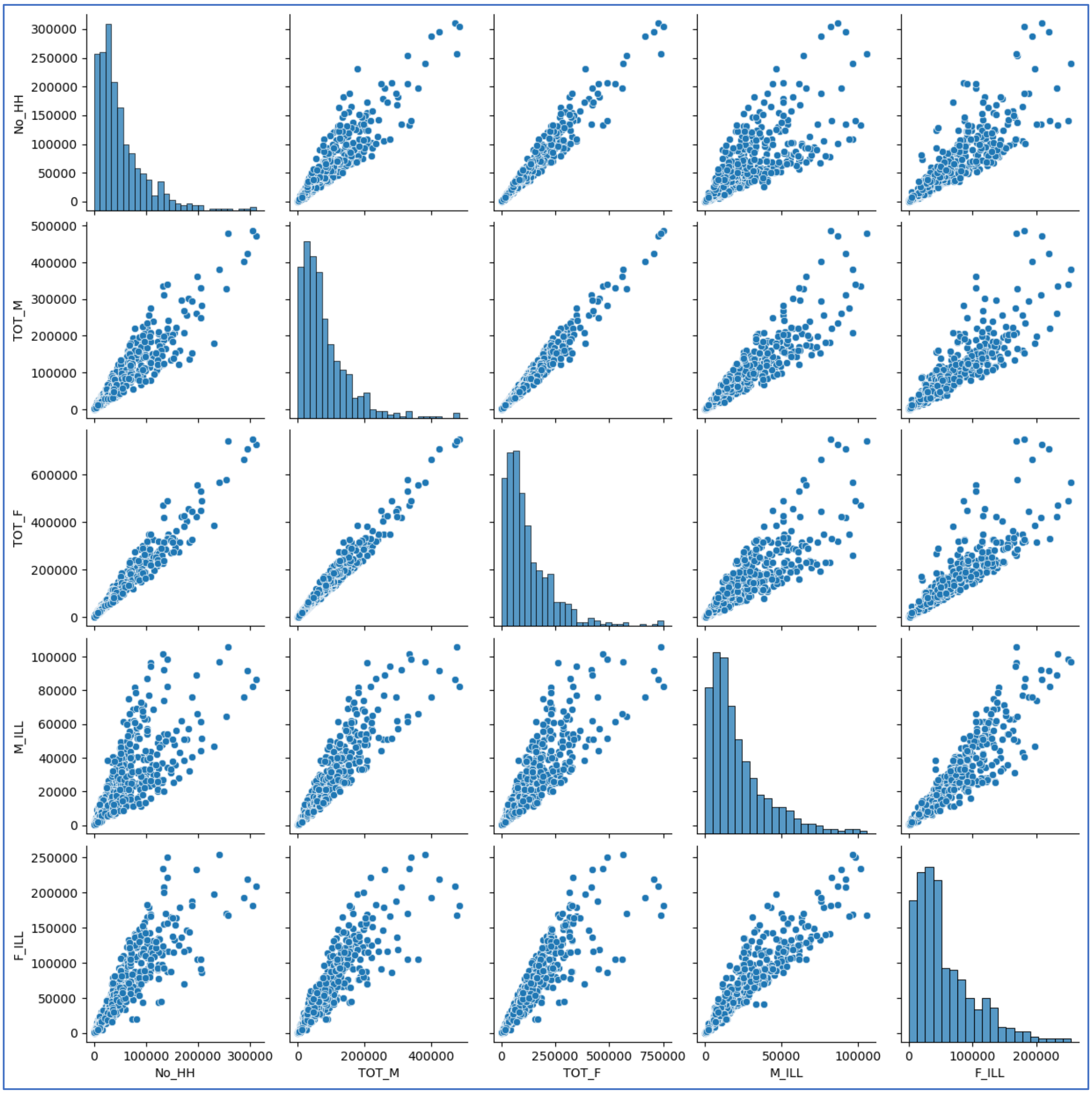


Figure 28: Pairplot of any 5 variables

- Pair plot also shows there is collinearity among the chosen columns.

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

   Since PCA is sensitive to outliers, we remove them instead of treating the outliers. Otherwise it may lead to manipulation of dataset.

Data Scaling

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

   In our basic data analysis itself it was found the data needs scaling as data values are far from each other.

Z-Score Scaling:

	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
0	-0.904738	-0.771236	-0.815563	-0.561012	-0.507738	-0.958575	-0.957049	-0.423306	-0.476423	-0.798097	...	-0.163229	-0.720610
1	-0.935695	-0.823100	-0.874534	-0.681096	-0.725367	-0.958297	-0.956772	-0.582014	-0.607607	-0.849434	...	-0.583103	-0.732811
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
...	...	...	...	...	...	...	...	...	...	...	...	...	...
635	-0.995677	-0.978990	-0.974268	-0.971387	-0.948916	-0.957326	-0.955667	-0.625124	-0.640197	-0.913820	...	-0.914299	-0.972530
636	-0.844340	-0.921822	-0.886965	-0.936754	-0.919757	-0.803806	-0.765670	-0.625124	-0.640197	-0.853390	...	-0.831668	-0.868461
637	-1.038465	-1.069066	-1.054885	-1.051356	-1.035331	-0.958783	-0.957049	-0.522953	-0.529880	-1.016367	...	-0.865930	-0.941309
638	-0.986758	-1.019276	-1.007472	-1.008195	-0.996541	-0.958783	-0.957049	-0.622297	-0.637046	-0.962328	...	-0.844432	-0.927673
639	-0.899166	-0.926854	-0.919050	-0.943193	-0.935220	-0.958783	-0.957049	-0.608870	-0.623555	-0.856916	...	-0.819576	-0.945616

MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
-0.720610	-0.156494	-0.287524	0.156577	-0.657412	-0.365258	-0.499977	-0.413053	-0.539614
-0.732811	-0.282327	-0.294688	-0.491731	-0.723062	0.042855	-0.073481	-0.606455	-0.598988
-0.921931	-0.456727	-0.420050	-0.731894	-0.795026	-0.662068	-0.635680	-0.726103	-0.707839
-0.900758	-0.419198	-0.385127	-0.718770	-0.784926	-0.624966	-0.616294	-0.645791	-0.710038
-0.297513	0.472670	0.434200	-0.466796	-0.625849	-0.439461	-0.309346	-0.540895	-0.249344
...	...	...	...	...	...	...	...	...
-0.972530	-0.553861	-0.499744	-0.735831	-0.816489	-0.662068	-0.648604	-0.783468	-0.723232
-0.868461	-0.547238	-0.487208	-0.685961	-0.734425	-0.624966	-0.574290	-0.655625	-0.587993
-0.941309	-0.533992	-0.496162	-0.733206	-0.812701	-0.504388	-0.496746	-0.711352	-0.690247
-0.927673	-0.500878	-0.460344	-0.721395	-0.803232	-0.652792	-0.635680	-0.672015	-0.661660
-0.945616	-0.540615	-0.497953	-0.713521	-0.805757	-0.643517	-0.635680	-0.593343	-0.666058

Figure 29: Scaled data

- Z-Score scaling makes mean close to zero and standard deviation close to 1

	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
F_ST	640.0	-2.220446e-17	1.000782	-0.640197	-0.613122	-0.398476	0.146640	7.582324
M_LIT	640.0	-4.440892e-17	1.000782	-1.032495	-0.656385	-0.273410	0.358381	6.180672
F_LIT	640.0	0.000000e+00	1.000782	-0.880091	-0.605869	-0.300924	0.245937	6.732272
M_ILL	640.0	3.885781e-17	1.000782	-1.103860	-0.675544	-0.313229	0.380609	4.239674
F_ILL	640.0	-4.440892e-17	1.000782	-1.182788	-0.714648	-0.289434	0.477029	4.208752
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
MAIN_OT_M	640.0	0.000000e+00	1.000782	-0.691455	-0.539371	-0.324365	0.122942	8.553708
MAIN_OT_F	640.0	-4.440892e-17	1.000782	-0.646347	-0.488651	-0.317847	0.103507	10.389042
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
MARGWORK_3_6_M	640.0	7.216450e-17	1.000782	-1.067727	-0.659748	-0.298173	0.391405	6.638220
MARGWORK_3_6_F	640.0	-2.220446e-17	1.000782	-0.973823	-0.656854	-0.292903	0.323834	7.181348
MARG_CL_3_6_M	640.0	-2.220446e-17	1.000782	-1.058667	-0.668815	-0.293426	0.294594	5.438148
MARG_CL_3_6_F	640.0	-8.881784e-17	1.000782	-1.212036	-0.707773	-0.241685	0.562843	4.695168
MARG_AL_3_6_M	640.0	-4.440892e-17	1.000782	-0.872827	-0.612586	-0.341847	0.216758	7.333319
MARG_AL_3_6_F	640.0	4.440892e-17	1.000782	-0.701351	-0.502020	-0.306297	0.124035	10.190617
MARG_HH_3_6_M	640.0	-7.216450e-17	1.000782	-0.897436	-0.662335	-0.336627	0.313560	5.429606
MARG_HH_3_6_F	640.0	-6.661338e-17	1.000782	-0.969686	-0.760784	-0.351845	0.437478	5.830127
MARG_OT_3_6_M	640.0	-5.551115e-18	1.000782	-0.684513	-0.522705	-0.323234	0.085473	9.177442
MARG_OT_3_6_F	640.0	3.330669e-17	1.000782	-0.651473	-0.509422	-0.295094	0.148296	12.796429
MARGWORK_0_3_M	640.0	0.000000e+00	1.000782	-0.869800	-0.613309	-0.307996	0.232028	5.942106
MARGWORK_0_3_F	640.0	0.000000e+00	1.000782	-0.848224	-0.601775	-0.300744	0.233353	6.919646
MARG_CL_0_3_M	640.0	-2.775558e-17	1.000782	-0.933110	-0.606952	-0.298260	0.215665	5.698208
MARG_CL_0_3_F	640.0	-5.551115e-17	1.000782	-0.978631	-0.645877	-0.297513	0.302412	6.765940
MARG_AL_0_3_M	640.0	2.220446e-17	1.000782	-0.553861	-0.450104	-0.301091	0.043845	12.194982
MARG_AL_0_3_F	640.0	-2.220446e-17	1.000782	-0.499744	-0.402141	-0.278122	0.009538	14.859741
MARG_HH_0_3_M	640.0	4.440892e-17	1.000782	-0.735831	-0.556693	-0.331622	0.106708	7.290595
MARG_HH_0_3_M	640.0	4.440892e-17	1.000782	-0.735831	-0.556693	-0.331622	0.106708	7.290595
MARG_HH_0_3_F	640.0	-1.110223e-17	1.000782	-0.816489	-0.628374	-0.363877	0.263436	7.840581
MARG_OT_0_3_M	640.0	-2.775558e-17	1.000782	-0.662068	-0.532213	-0.337432	0.070681	7.639320
MARG_OT_0_3_F	640.0	0.000000e+00	1.000782	-0.648604	-0.509670	-0.283498	0.126843	10.188272
NON_WORK_M	640.0	-2.220446e-17	1.000782	-0.835916	-0.572036	-0.301600	0.154863	9.745505
NON_WORK_F	640.0	-6.661338e-17	1.000782	-0.769412	-0.532468	-0.264188	0.163521	10.806207

Figure 30: Scaled data summary

Effect of Scaling on the outliers

- The below figure shows that scaling doesn't impact any of the outliers.

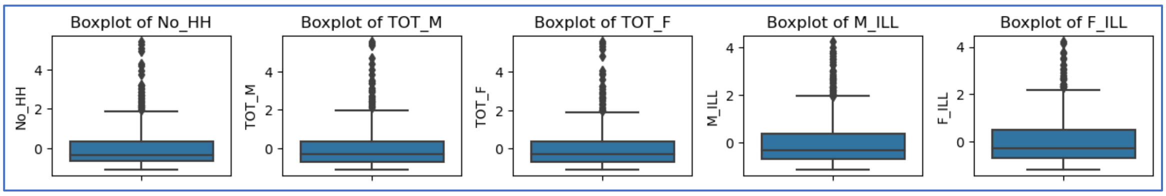


Figure 31: Boxplots of scaled data

- Heat Map for All Given Columns (Scaled Data)

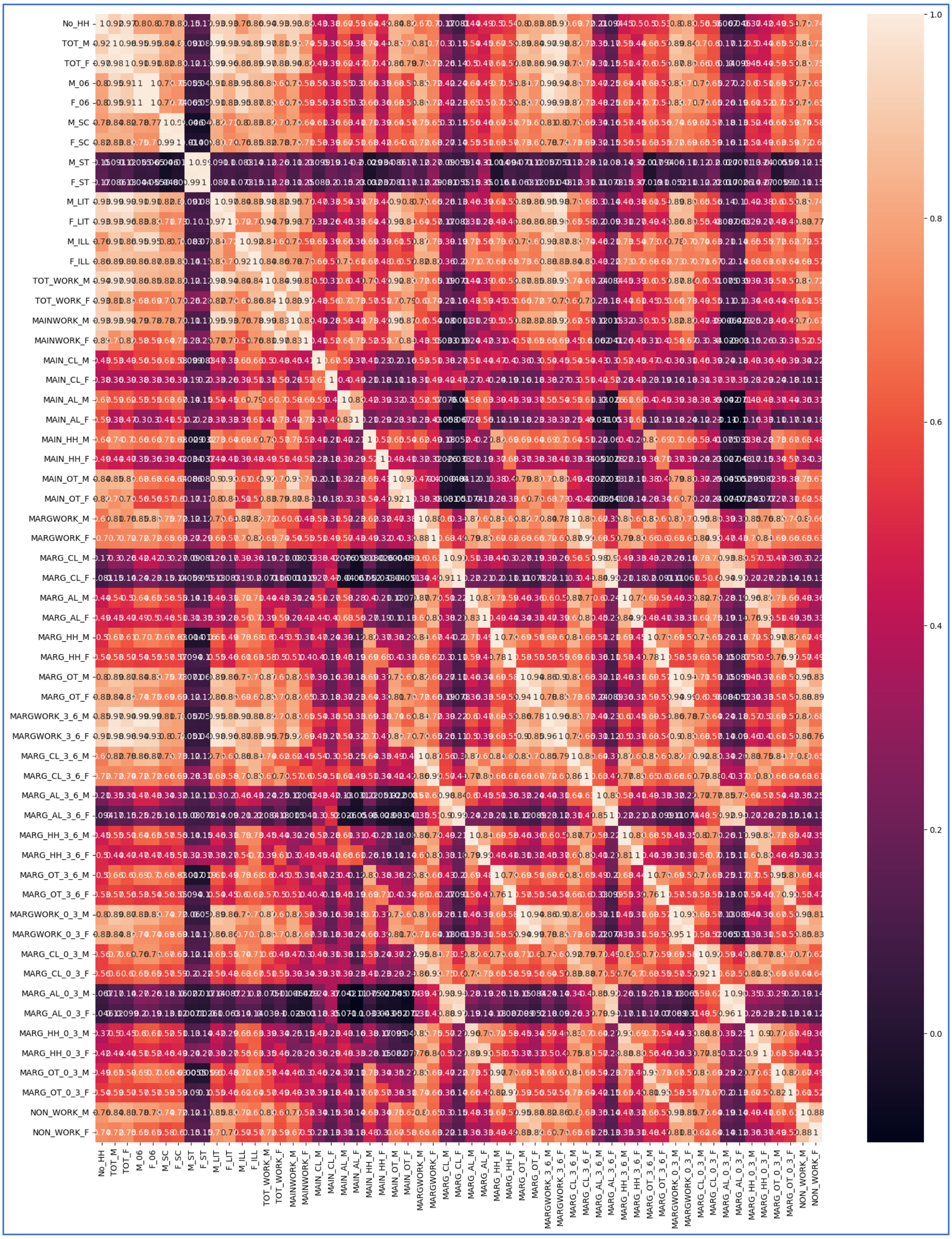


Figure 32: Heatmap of scaled data

- Heat map shows high correlation among many columns with the other.

Principal Component Analysis

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

Covariance Matrix:

```
Covariance Matrix
 [[1.         0.91616988 0.97058979 ... 0.53685418 0.76238413 0.73569246]
 [0.91616988 1.         0.98264045 ... 0.58818023 0.84489622 0.71606121]
 [0.97058979 0.98264045 1.         ... 0.57185308 0.82765328 0.74658261]
 ...
 [0.53685418 0.58818023 0.57185308 ... 1.         0.6095693  0.52109686]
 [0.76238413 0.84489622 0.82765328 ... 0.6095693  1.         0.88090162]
 [0.73569246 0.71606121 0.74658261 ... 0.52109686 0.88090162 1.        ]]
```

Figure 33: Covariance matrix

Eigen Vector and Eigen Value:

```
Eigen Vectors
 [[-1.56020579e-01+0.00000000e+00j  1.26346525e-01+0.00000000e+00j
  -2.69025037e-03+0.00000000e+00j ... -2.18416172e-14+0.00000000e+00j
   9.82414958e-15-1.22816598e-14j  9.82414958e-15+1.22816598e-14j]
 [-1.67117635e-01+0.00000000e+00j  8.96765481e-02+0.00000000e+00j
   5.66976191e-02+0.00000000e+00j ...  7.26493909e-02+0.00000000e+00j
   6.19403635e-02-1.49494655e-02j  6.19403635e-02+1.49494655e-02j]
 [-1.65553179e-01+0.00000000e+00j  1.04912371e-01+0.00000000e+00j
   3.87494746e-02+0.00000000e+00j ...  3.91091522e-01+0.00000000e+00j
   1.92123097e-01-1.04060242e-01j  1.92123097e-01+1.04060242e-01j]
 ...
 [-1.32192245e-01+0.00000000e+00j -5.08133220e-02+0.00000000e+00j
  -7.87198691e-02+0.00000000e+00j ... -5.74108207e-03+0.00000000e+00j
  -2.57200586e-02-1.20192830e-02j -2.57200586e-02+1.20192830e-02j]
 [-1.50375578e-01+0.00000000e+00j  6.53645529e-02+0.00000000e+00j
   1.11827318e-01+0.00000000e+00j ... -5.35261432e-03+0.00000000e+00j
  -6.65894129e-02-1.22250480e-02j -6.65894129e-02+1.22250480e-02j]
 [-1.31066203e-01+0.00000000e+00j  7.38474208e-02+0.00000000e+00j
   1.02552501e-01+0.00000000e+00j ... -1.95662634e-02+0.00000000e+00j
  -1.19467868e-01-1.05430357e-02j -1.19467868e-01+1.05430357e-02j]]
```

Figure 34: Eigen vectors

```
Eigen Values
[ 3.17638560e+01+0.00000000e+00j  7.85712818e+00+0.00000000e+00j
  4.14691842e+00+0.00000000e+00j  3.66305809e+00+0.00000000e+00j
  2.20307818e+00+0.00000000e+00j  1.93524646e+00+0.00000000e+00j
  1.17433597e+00+0.00000000e+00j  7.49985400e-01+0.00000000e+00j
  6.16089597e-01+0.00000000e+00j  5.27475417e-01+0.00000000e+00j
  4.29159578e-01+0.00000000e+00j  3.52887950e-01+0.00000000e+00j
  2.95700259e-01+0.00000000e+00j  2.80836067e-01+0.00000000e+00j
  1.91858078e-01+0.00000000e+00j  1.36055001e-01+0.00000000e+00j
  1.13212028e-01+0.00000000e+00j  1.06137846e-01+0.00000000e+00j
  9.71365243e-02+0.00000000e+00j  7.99810535e-02+0.00000000e+00j
  5.75189814e-02+0.00000000e+00j  4.43262285e-02+0.00000000e+00j
  3.78318798e-02+0.00000000e+00j  2.95897131e-02+0.00000000e+00j
  2.70374497e-02+0.00000000e+00j  2.34091798e-02+0.00000000e+00j
  1.44884774e-02+0.00000000e+00j  7.12444187e-04+0.00000000e+00j
  1.06622961e-03+0.00000000e+00j  2.59365290e-03+0.00000000e+00j
  5.01816199e-03+0.00000000e+00j  1.09680624e-02+0.00000000e+00j
  9.30052372e-03+0.00000000e+00j  8.12269047e-03+0.00000000e+00j
  7.88017049e-03+0.00000000e+00j  1.11825550e-15+0.00000000e+00j
 -1.07120849e-15+3.32448656e-16j -1.07120849e-15-3.32448656e-16j
 -1.16230609e-15+0.00000000e+00j -1.06546917e-15+0.00000000e+00j
  7.93165234e-16+0.00000000e+00j  7.21408988e-16+7.57361847e-17j
  7.21408988e-16-7.57361847e-17j -8.02459721e-16+0.00000000e+00j
 -7.34145147e-16+0.00000000e+00j  4.24731429e-16+4.58780256e-17j
  4.24731429e-16-4.58780256e-17j -5.74053941e-16+0.00000000e+00j
 -4.99577381e-16+0.00000000e+00j  3.08320515e-16+0.00000000e+00j
 -3.46407818e-16+0.00000000e+00j -2.78792906e-16+0.00000000e+00j
 -1.17833161e-16+6.71989630e-17j -1.17833161e-16-6.71989630e-17j
  1.20032068e-18+0.00000000e+00j  9.53536475e-17+1.96465885e-17j
  9.53536475e-17-1.96465885e-17j]
```

Figure 35: Eigen values

First Eigen Vector equation:

The first eigen vector is:
(-0.1560205785856788+0j) * No_HH +
(0.12634652545112177+0j) * TOT_M +
(-0.002690250367895609+0j) * TOT_F +
(-0.12529337156421827+0j) * M_06 +
(0.007022081300930482+0j) * F_06 +
(-0.004082812708624174+0j) * M_SC +
(0.11811039900370703+0j) * F_SC +
(-0.05723830782871984+0j) * M_ST +
(0.004264737531171913+0j) * F_ST +
(0.01998510330402382+0j) * M_LIT +
(-0.01059187665377821+0j) * F_LIT +
(-0.08619327075293665+0j) * M_ILL +
(-0.1041749256470337+0j) * F_ILL +
(0.028891545615620327+0j) * TOT_WORK_M +
(0.05731964114311972+0j) * TOT_WORK_F +
(0.022262954176667 4+0j) * MAINWORK_M +
(0.07927843396646114+0j) * MAINWORK_F +
(0.13279831420623323+0j) * MAIN_CL_M +
(0.09950616875207853+0j) * MAIN_CL_F +
(-0.06153349699574671+0j) * MAIN_AL_M +

(0.09141067674655443+0j) * MAIN_AL_F +
(-0.3912638142970992+0j) * MAIN_HH_M +
(0.32033486100615255+0j) * MAIN_HH_F +
(-0.0020116374778708715+0j) * MAIN_OT_M +
(0.09665622840546396+0j) * MAIN_OT_F +
(-0.135743225605903+0j) * MARGWORK_M +
(0.20645117771828925+0j) * MARGWORK_F +
(0.16309535079634924+0j) * MARG_CL_M +
(-0.2126981254754507+0j) * MARG_CL_F +
(-0.6418772071297786+0j) * MARG_AL_M +
(0.08346450006334961+0j) * MARG_AL_F +
(0.03877758203462681+0j) * MARG_HH_M +
(0.11460174850970233+0j) * MARG_HH_F +
(0.16049099951976822+0j) * MARG_OT_M +
(-0.00037100948914314276+0j) * MARG_OT_F +
(-1.1264965458035132e-13+0j) * MARGWORK_3_6_M +
(-3.622715671187225e-14-2.5865991357172073e-14j) * MARGWORK_3_6_F +
(-3.622715671187225e-14+2.5865991357172073e-14j) * MARG_CL_3_6_M +
(-6.144849989290794e-14+0j) * MARG_CL_3_6_F +
(-4.9032897077474184e-14+0j) * MARG_AL_3_6_M +
(3.374370787315205e-14+0j) * MARG_AL_3_6_F +
(-4.241566837391235e-14+3.2008674758378366e-14j) * MARG_HH_3_6_M +
(-4.241566837391235e-14-3.2008674758378366e-14j) * MARG_HH_3_6_F +
(9.593069045378147e-14+0j) * MARG_OT_3_6_M +
(4.8549123100476903e-14+0j) * MARG_OT_3_6_F +
(5.863311252902347e-14+2.7132010787965834e-15j) * MARGWORK_0_3_M +
(5.863311252902347e-14-2.7132010787965834e-15j) * MARGWORK_0_3_F +
(-4.517917855236678e-14+0j) * MARG_CL_0_3_M +
(3.7311611273099406e-14+0j) * MARG_CL_0_3_F +
(5.887485159490267e-14+0j) * MARG_AL_0_3_M +
(4.4603628032996474e-15+0j) * MARG_AL_0_3_F +
(-1.1026364526829432e-14+0j) * MARG_HH_0_3_M +
(1.444642432624533e-15-1.404106268381001e-14j) * MARG_HH_0_3_F +
(1.444642432624533e-15+1.404106268381001e-14j) * MARG_OT_0_3_M +
(-2.1841617181179887e-14+0j) * MARG_OT_0_3_F +
(9.824149575382851e-15-1.2281659828338364e-14j) * NON_WORK_M +
(9.824149575382851e-15+1.2281659828338364e-14j) * NON_WORK_F

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

Variance Experienced by Each of the Eigen Value:

```
The variance explained by each of eigen values in order is
 [(55.726063245483395+0j), (13.784435398867204+0j), (7.275295475037663+0j), (6.426417707902724+0j), (3.8650494385437404+0j),
(3.395169233122141+0j), (2.060238546173363+0j), (1.3157638603636073+0j), (1.0808589417423942+0j), (0.9253954683322126+0j), (0.7
529115396464453+0j), (0.6191016673019892+0j), (0.5187723837113795+0j), (0.49269485485278275+0j), (0.3365931194569853+0j), (0.2
386929842998293+0j), (0.19861759344696095+0j), (0.18620674680204716+0j), (0.17041495488866652+0j), (0.1403176376500245+0j), (0.
10091049360553642+0j), (0.07776531307693414+0j), (0.0663717189868593+0j), (0.05191177735689486+0j), (0.04743412222616297+0j),
(0.04106873644998113+0j), (0.0254183814352128+0j), (0.01924221472767971+0j), (0.016316708282405244+0j), (0.014250334150863289+0
j), (0.013824860517365593+0j), (0.008803792969803456+0j), (0.004550268241441011+0j), (0.0018705782641093816+0j), (0.00124990208
32031325+0j), (1.961851747531199e-15+0j), (1.3915179550653592e-15+0j), (1.2656298040215477e-15+1.328704994755039e-16j), (1.2656
298040215477e-15-1.328704994755039e-16j), (7.451428578863546e-16+8.048776416468543e-17j), (7.451428578863546e-16-8.048776416468
543e-17j), (5.409131850335533e-16+0j), (1.6728710079455306e-16+3.4467699085397086e-17j), (1.6728710079455306e-16-3.446769908539
7086e-17j), (2.105825746495905e-18+0j), (-2.0672484355892472e-16+1.178929176272042e-16j), (-2.0672484355892472e-16-1.178929176
272042e-16j), (-4.891103608731506e-16+0j), (-6.077330146050475e-16+0j), (-8.764515458418412e-16+0j), (-1.007112177184315e-15+0
j), (-1.287973941524004e-15+0j), (-1.4078240711128474e-15+0j), (-1.8692441630379535e-15+0j), (-1.879313132478615 5e-15+5.8324325
62225718e-16j), (-1.8793131324786155e-15-5.832432562225718e-16j), (-2.0391334948456777e-15+0j)]
```

Figure 36: Variance

Cumulative Variance:

```
Cumulative Variance Explained

  [ 55.72606325+0.00000000e+00j  69.51049864+0.00000000e+00j
  76.78579412+0.00000000e+00j  83.21221183+0.00000000e+00j
  87.07726127+0.00000000e+00j  90.4724305 +0.00000000e+00j
  92.53266905+0.00000000e+00j  93.84843291+0.00000000e+00j
  94.92929185+0.00000000e+00j  95.85468732+0.00000000e+00j
  96.60759886+0.00000000e+00j  97.22670052+0.00000000e+00j
  97.74547291+0.00000000e+00j  98.23816776+0.00000000e+00j
  98.57476088+0.00000000e+00j  98.81345386+0.00000000e+00j
  99.01207146+0.00000000e+00j  99.19827821+0.00000000e+00j
  99.36869316+0.00000000e+00j  99.5090108 +0.00000000e+00j
  99.60992129+0.00000000e+00j  99.6876866 +0.00000000e+00j
  99.75405832+0.00000000e+00j  99.8059701 +0.00000000e+00j
  99.85340422+0.00000000e+00j  99.89447296+0.00000000e+00j
  99.91989134+0.00000000e+00j  99.93913356+0.00000000e+00j
  99.95545026+0.00000000e+00j  99.9697006 +0.00000000e+00j
  99.98352546+0.00000000e+00j  99.99232925+0.00000000e+00j
  99.99687952+0.00000000e+00j  99.9987501 +0.00000000e+00j
 100.        +0.00000000e+00j 100.        +0.00000000e+00j
 100.        +0.00000000e+00j 100.        +1.32870499e-16j
 100.        +0.00000000e+00j 100.        +8.04877642e-17j
 100.        +0.00000000e+00j 100.        +0.00000000e+00j
 100.        +3.44676991e-17j 100.        +0.00000000e+00j
 100.        +0.00000000e+00j 100.        +1.17892918e-16j
 100.        +0.00000000e+00j 100.        +0.00000000e+00j
 100.        +0.00000000e+00j 100.        +0.00000000e+00j
 100.        +0.00000000e+00j 100.        +0.00000000e+00j
 100.        +0.00000000e+00j 100.        +0.00000000e+00j
 100.        +5.83243256e-16j 100.        +0.00000000e+00j
 100.        +0.00000000e+00j]
```

Figure 37: Cumulative variance

The above figures shows that the optimum number of components needed would be 6 to cover 90% of the explained variance.

Scree Plot:

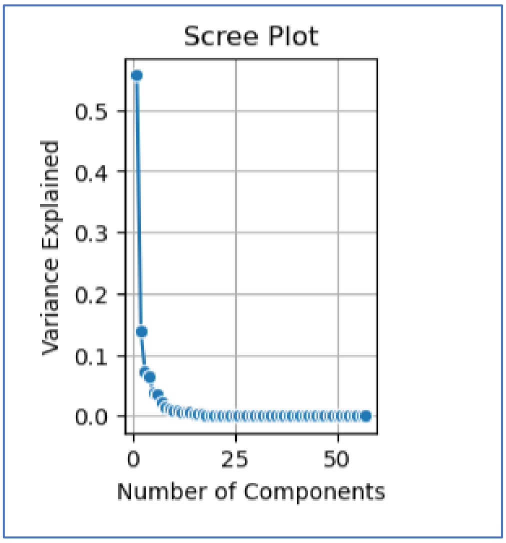


Figure 38: Scree plot

Cumulative Explained Variance:

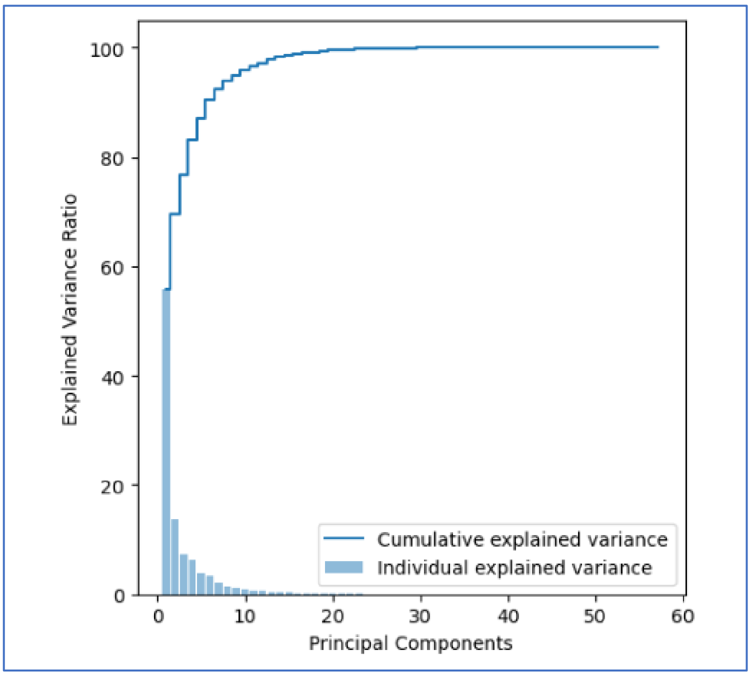


Figure 39: Explained variance ratio vs PCs

Final Dimension:

	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
...	...	...	...	...	...	...
635	-6.262088	-0.854414	0.242575	1.174113	0.063816	-0.159470
636	-5.767714	-0.900436	0.168051	1.102774	0.055179	-0.156458
637	-6.294625	-0.638127	0.107483	1.368187	0.153745	0.141145
638	-6.223192	-0.672320	0.271325	1.143493	0.060440	-0.115682
639	-5.896236	-0.937170	0.349218	1.114861	0.149104	-0.154544

640 rows × 6 columns

Figure 40: Final Dimension of PCs

- The given data is now reduced to 6 columns, PCA doesn't do any modification to the number of rows, it remains the same 640.

Heat Map of the PCA VS Original Columns:

- The below heat map shows how each component of PCA is influenced by the original data columns.

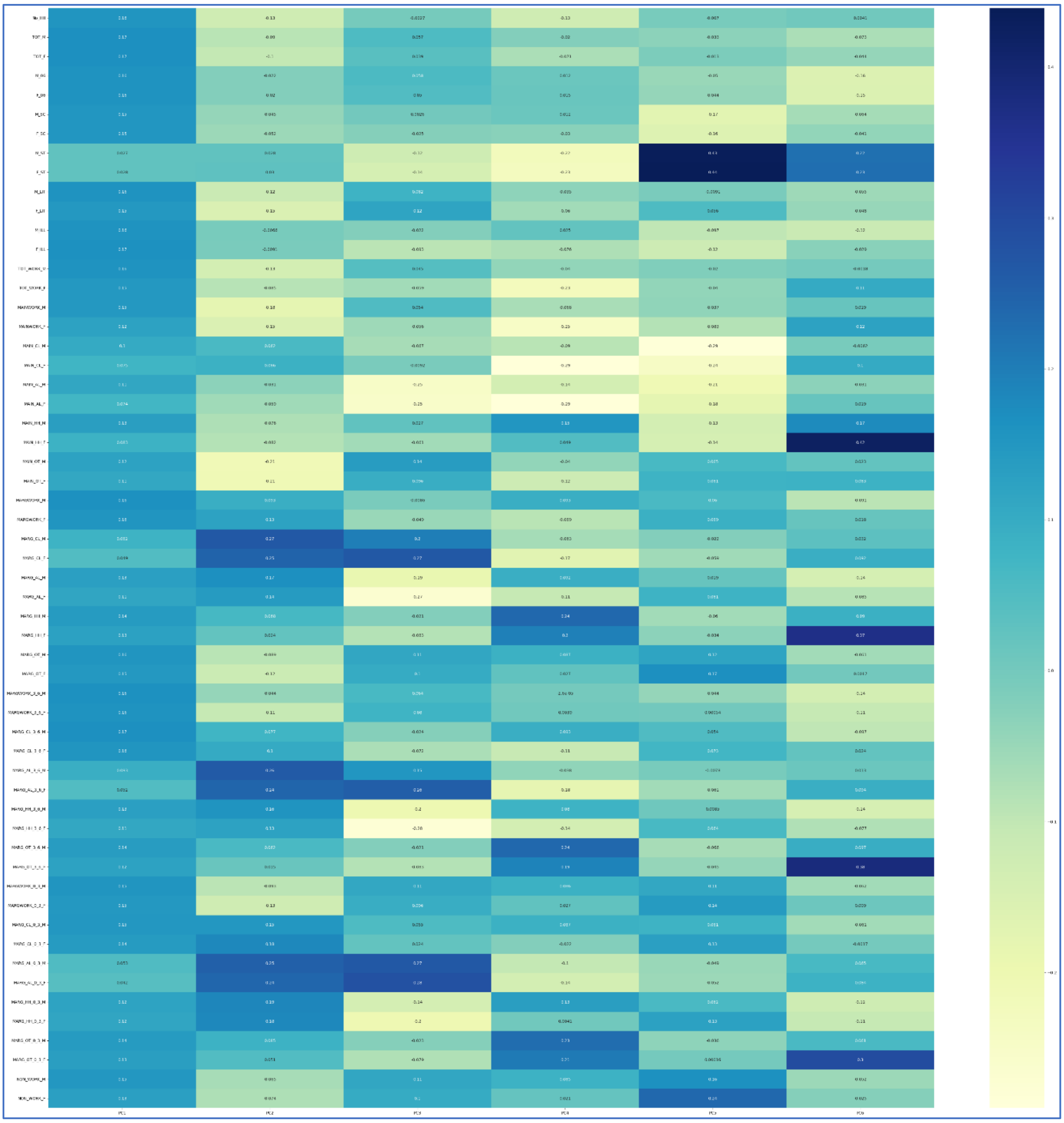


Figure 41: Heatmap of PCA vs Original columns

The above figure shows that the first PCA components is most influenced by almost all columns given, and the influenced keeps reducing in the subsequent components, such that in the 5th and 6th components some of the columns are having close to zero correlation. So PCA1 explains most variance which is 55%.

Component wise inference

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

- PC1 explains the most variance which is 55% of the given data. This Principal component covers mostly about Marginal Literate Agricultural Laborers. It has strong correlation with the actual columns like Marginal Worker Population 3-6 Female, Marginal Worker Population 3-6 Male, Marginal Worker Population Male, Literates population Female, Literates population Male. Hence, we have labelled this component as **Marginal Literate Agricultural Labourers**.

- PC2 is about 13% of the given data. This principal component covers Marginal Cultivators and agriculture labourer, with most of its influence from the actual variable Marginal Agriculture Labourers Population 0-3 Female, Marginal Cultivator Population Female, Marginal Agriculture Labourers Population 0-3 Male, Marginal Agriculture Labourers Population 3-6 Female, Marginal Cultivator Population Male, Marginal Agriculture Labourers Population 3-6 Male who own the land and involve themselves in farming. Hence labelling them as **Marginal Cultivators.**

- PC3 is about 7% of the given data. We have labelled them as **Marginal Agricultural Labourers Female** as most influencing variables are Marginal Agriculture Labourers Population 0-3 Female, Marginal Cultivator Population Female, Marginal Agriculture Labourers Population 3-6 Female.

- PC4 is about 6.4% of the given data, this component covers **Marginal Household industry and Other workers** hence the same label has been given to it. The variable influencing the component mostly is Marginal Other Workers Population Person 3-6 Male, Marginal Household Industries Population Male, Marginal Other Workers Population 0-3 Male, Marginal Other Workers population 0-3 Female, Marginal Household Industries Population Female

- PC5 is about 3.8% of the given data. The most influencing actual variable for this component is Non-Working Population Male, Non-Working Population Female, Marginal Other Workers Population Female, Marginal Worker Population 0-3 Female, Scheduled Tribes population Male and Scheduled Tribes population Female. Hence labelling this component as **Scheduled Tribe Population.**

- PC6 is about 3.3% of the given data which is mostly influenced by actual column like Main Working Population Female, Main Household Industries Population Female, Marginal Other Workers Population 0-3 Female, Marginal Household Industries Population Female, Marginal Other Workers Population Person 3-6 Female, Scheduled Tribes population Female. This component is mainly about the female population hence labelling it **as Main & Marginal Female population.**

Heat Map PCA components:

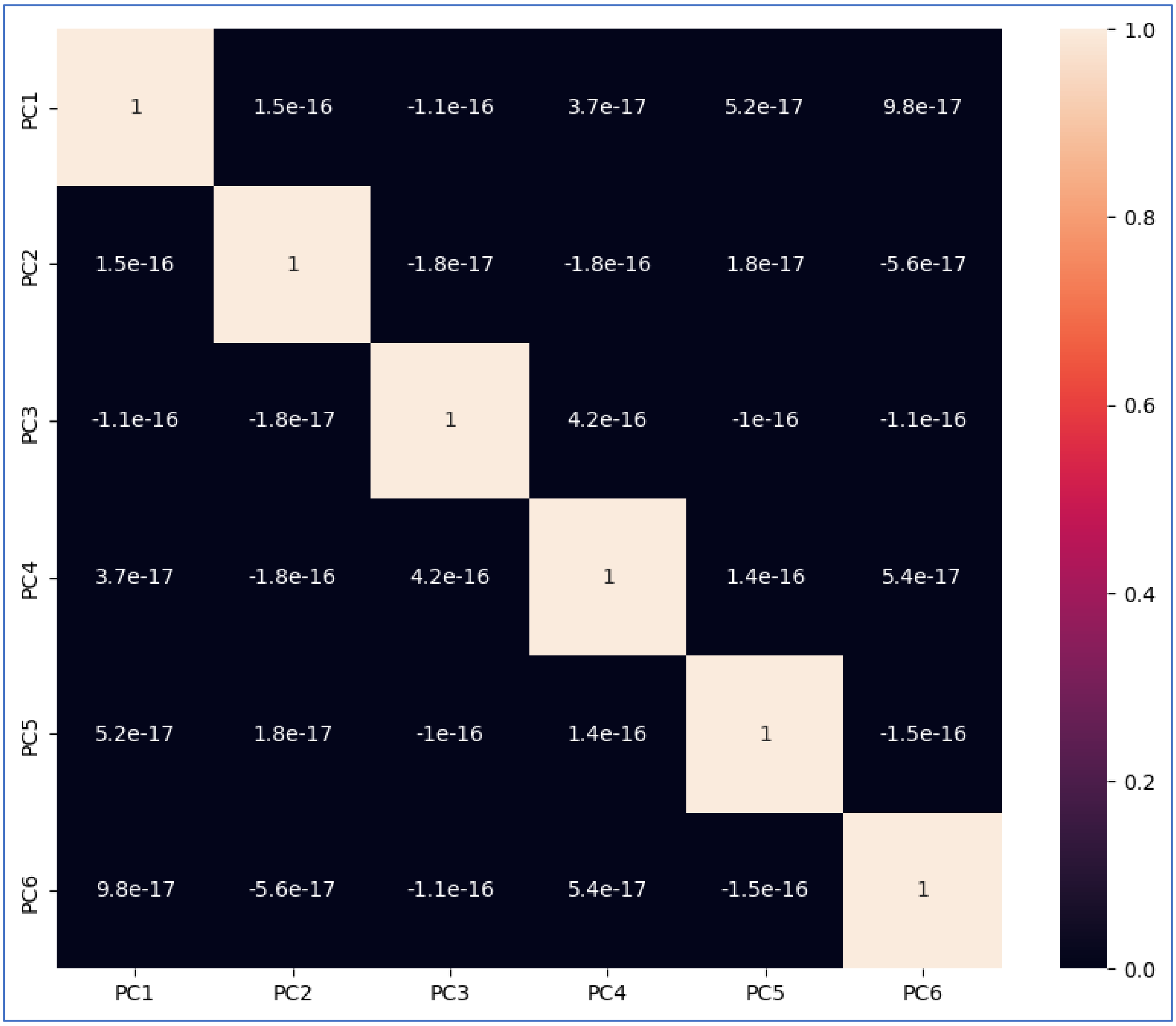


Figure 42: Heatmap of all PCs

The above figure shows that none of the PCA component is corelated to the other components as the covariance value is very close to zero which shows our PCA is successful.

Influence of actual variables on PC1

8. Write linear equation for first PC

The influence of actual variable and the PC can be explained by a linear equation.

```
The Equation of the First PC (PC1) is:
```

```
0.15602057858558915 * No_HH +
0.1671176348853478 * TOT_M +
0.16555317909057707 * TOT_F +
0.16219294820457575 * M_06 +
0.16256639565726402 * F_06 +
0.15135784909062253 * M_SC +
0.1515665001920242 * F_SC +
0.027234194570986126 * M_ST +
0.02818331501586022 * F_ST +
0.16199283733627975 * M_LIT +
0.1468726803012997 * F_LIT +
0.16174944463479718 * M_ILL +
0.16524818736832483 * F_ILL +
0.15987198816211723 * TOT_WORK_M +
0.14593580377247212 * TOT_WORK_F +
0.14620072976315196 * MAINWORK_M +
0.12397028357273014 * MAINWORK_F +
0.10312715882997171 * MAIN_CL_M +
0.07453978555514146 * MAIN_CL_F +
0.11335571218162897 * MAIN_AL_M +
0.07388215903143862 * MAIN_AL_F +
0.13157258402261926 * MAIN_HH_M +
0.0833826396742786 * MAIN_HH_F +
0.1235262419226749 * MAIN_OT_M +
0.11102126391320055 * MAIN_OT_F +
0.16461547856023126 * MARGWORK_M +
0.15539561810834482 * MARGWORK_F +
0.08238854140677228 * MARG_CL_M +
0.049195395678877776 * MARG_CL_F +
0.1285985629468215 * MARG_AL_M +
0.11430507278921848 * MARG_AL_F +
0.14085322696180522 * MARG_HH_M +
0.12766959801481645 * MARG_HH_F +
0.15526287162332747 * MARG_OT_M +
0.14728658356507177 * MARG_OT_F +
0.16497194993707148 * MARGWORK_3_6_M +
0.161253432575217 * MARGWORK_3_6_F +
0.1655016110259308 * MARG_CL_3_6_M +
0.15564704914486402 * MARG_CL_3_6_F +
0.09301420640152648 * MARG_AL_3_6_M +
0.051535863970250805 * MARG_AL_3_6_F +
0.12857611642886113 * MARG_HH_3_6_M +
0.11064584323703604 * MARG_HH_3_6_F +
0.13959276252154193 * MARG_OT_3_6_M +
0.12454590917265383 * MARG_OT_3_6_F +
0.15429378578934816 * MARGWORK_0_3_M +
0.14628565406202168 * MARGWORK_0_3_F +
0.15012570610272008 * MARG_CL_0_3_M +
0.1401570468900264 * MARG_CL_0_3_F +
0.052541782853976135 * MARG_AL_0_3_M +
0.04178595301223521 * MARG_AL_0_3_F +
0.12184035387919034 * MARG_HH_0_3_M +
0.11601141016809986 * MARG_HH_0_3_F +
0.1398687741103844 * MARG_OT_0_3_M +
0.13219224458201573 * MARG_OT_0_3_F +
0.15037557804442866 * NON_WORK_M +
0.13106620313178857 * NON_WORK_F
```