2022

DATA MINING GRADED PROJECT REPORT

DSBA

Girish Chadha

13/11/2022

Contents

List of Figures

List of Tables

EXECUTIVE SUMMARY

Problem 1

Clustering:

Digital Ads Data:

The ads24x7 is a Digital Marketing company which has now got seed funding of $10 Million. They are expanding their wings in Marketing Analytics. They collected data from their Marketing Intelligence team and now wants you (their newly appointed data analyst) to segment type of ads based on the features provided. Use Clustering procedure to segment ads into homogeneous groups.

The following three features are commonly used in digital marketing:

CPM = (Total Campaign Spend / Number of Impressions) * 1,000. Note that the Total Campaign Spend refers to the 'Spend' Column in the dataset and the Number of Impressions refers to the 'Impressions' Column in the dataset.

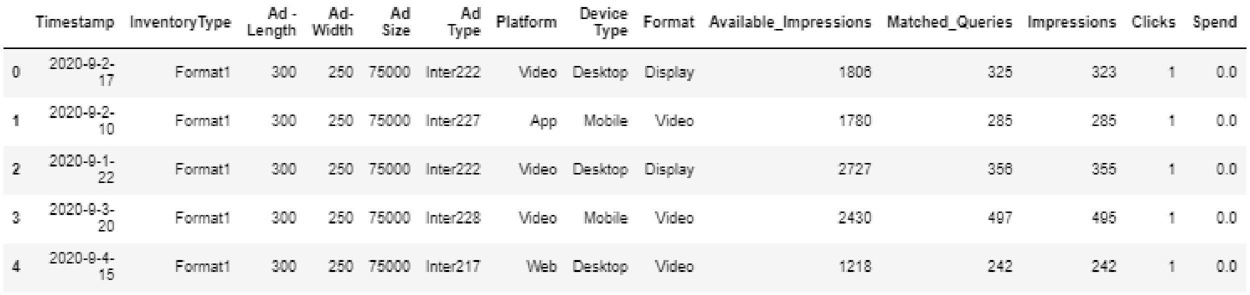
CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column in the dataset and the Number of Clicks refers to the 'Clicks' Column in the dataset.

CTR = Total Measured Clicks / Total Measured Ad Impressions x 100. Note that the Total Measured Clicks refers to the 'Clicks' Column in the dataset and the Total Measured Ad Impressions refers to the 'Impressions' Column in the dataset.

Perform the following in given order:

1.1 - 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.

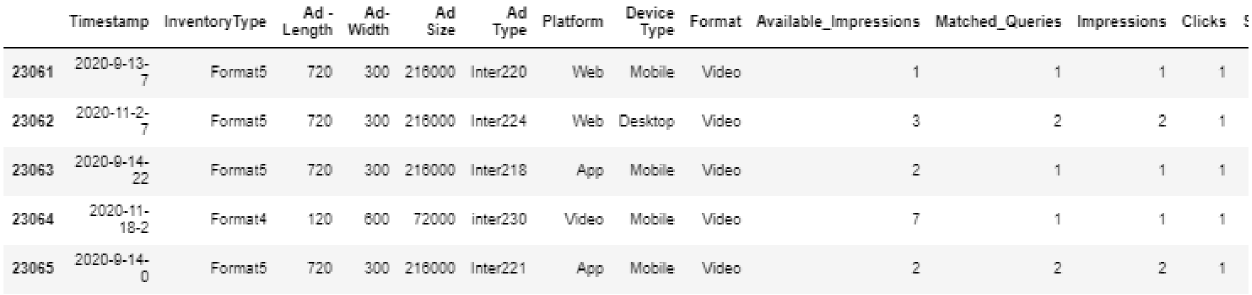
Solution : Reading the data



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

Table 1

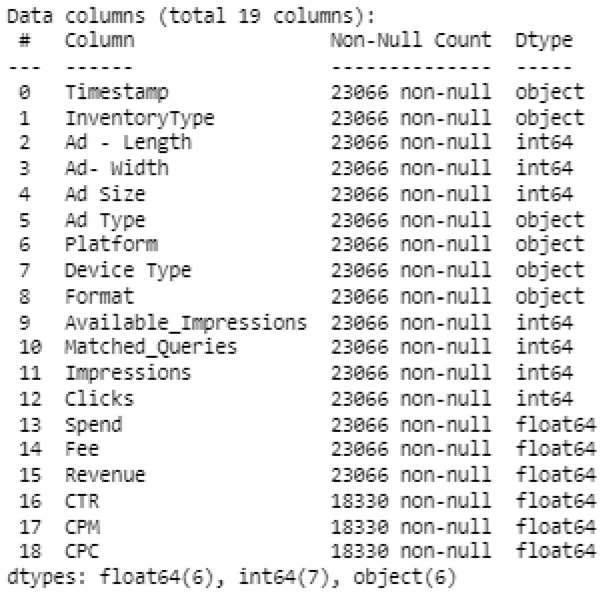
Printing few rows



	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks
23061	2020-9-13-7	Format5	720	300	216000	Inter220	Web	Mobile	Video	1	1	1	1
23062	2020-11-2-7	Format5	720	300	216000	Inter224	Web	Desktop	Video	3	2	2	1
23063	2020-9-14-22	Format5	720	300	216000	Inter218	App	Mobile	Video	2	1	1	1
23064	2020-11-18-2	Format4	120	600	72000	inter230	Video	Mobile	Video	7	1	1	1
23065	2020-9-14-0	Format5	720	300	216000	Inter221	App	Mobile	Video	2	2	2	1

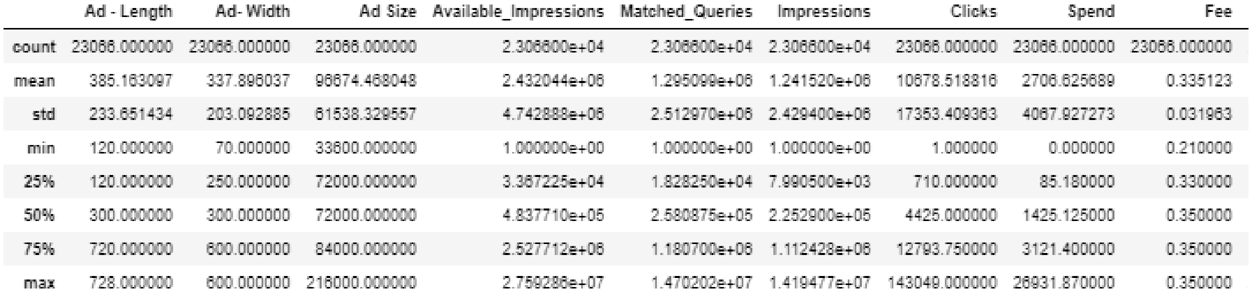
Table 2

Data information



```
Data columns (total 19 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   Timestamp              23066 non-null  object 
 1   InventoryType          23066 non-null  object 
 2   Ad - Length            23066 non-null  int64  
 3   Ad- Width              23066 non-null  int64  
 4   Ad Size                23066 non-null  int64  
 5   Ad Type                23066 non-null  object 
 6   Platform               23066 non-null  object 
 7   Device Type            23066 non-null  object 
 8   Format                 23066 non-null  object 
 9   Available_Impressions  23066 non-null  int64  
 10  Matched_Queries        23066 non-null  int64  
 11  Impressions            23066 non-null  int64  
 12  Clicks                 23066 non-null  int64  
 13  Spend                  23066 non-null  float64
 14  Fee                    23066 non-null  float64
 15  Revenue                23066 non-null  float64
 16  CTR                    18330 non-null  float64
 17  CPM                    18330 non-null  float64
 18  CPC                    18330 non-null  float64
dtypes: float64(6), int64(7), object(6)
```

Data Summary

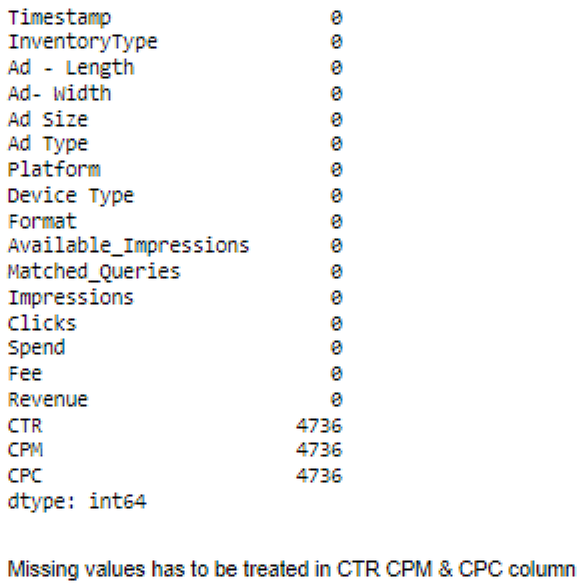


	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee
count	23066.000000	23066.000000	23066.000000	2.306600e+04	2.306600e+04	2.306600e+04	23066.000000	23066.000000	23066.000000
mean	385.163097	337.896037	96674.468048	2.432044e+06	1.295099e+06	1.241520e+06	10678.518816	2706.625689	0.335123
std	233.651434	203.092885	61538.329557	4.742888e+06	2.512970e+06	2.429400e+06	17353.409363	4067.927273	0.031963
min	120.000000	70.000000	33600.000000	1.000000e+00	1.000000e+00	1.000000e+00	1.000000	0.000000	0.210000
25%	120.000000	250.000000	72000.000000	3.367225e+04	1.828250e+04	7.990500e+03	710.000000	85.180000	0.330000
50%	300.000000	300.000000	72000.000000	4.837710e+05	2.580875e+05	2.252900e+05	4425.000000	1425.125000	0.350000
75%	720.000000	600.000000	84000.000000	2.527712e+06	1.180700e+06	1.112428e+06	12793.750000	3121.400000	0.350000
max	728.000000	600.000000	216000.000000	2.759286e+07	1.470202e+07	1.419477e+07	143049.000000	26931.870000	0.350000

Table 3

No duplicate entries

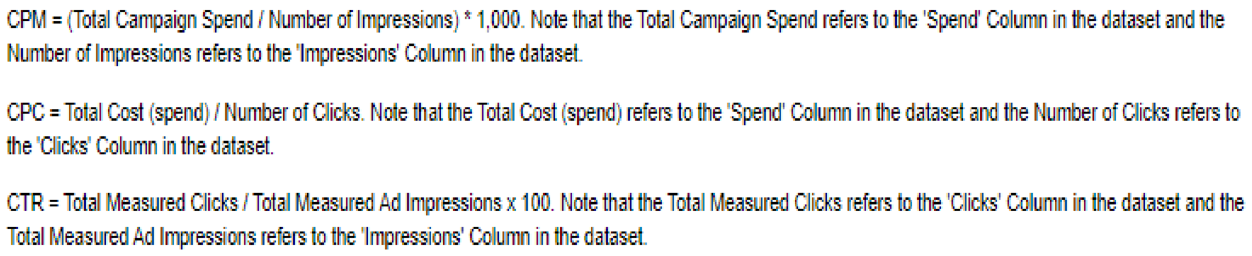
Missing values -



```
Timestamp                  0
InventoryType              0
Ad - Length                0
Ad- Width                  0
Ad Size                    0
Ad Type                    0
Platform                   0
Device Type                0
Format                     0
Available_Impressions      0
Matched_Queries            0
Impressions                0
Clicks                     0
Spend                      0
Fee                        0
Revenue                    0
CTR                     4736
CPM                     4736
CPC                     4736
dtype: int64
```

Missing values has to be treated in CTR CPM & CPC column

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



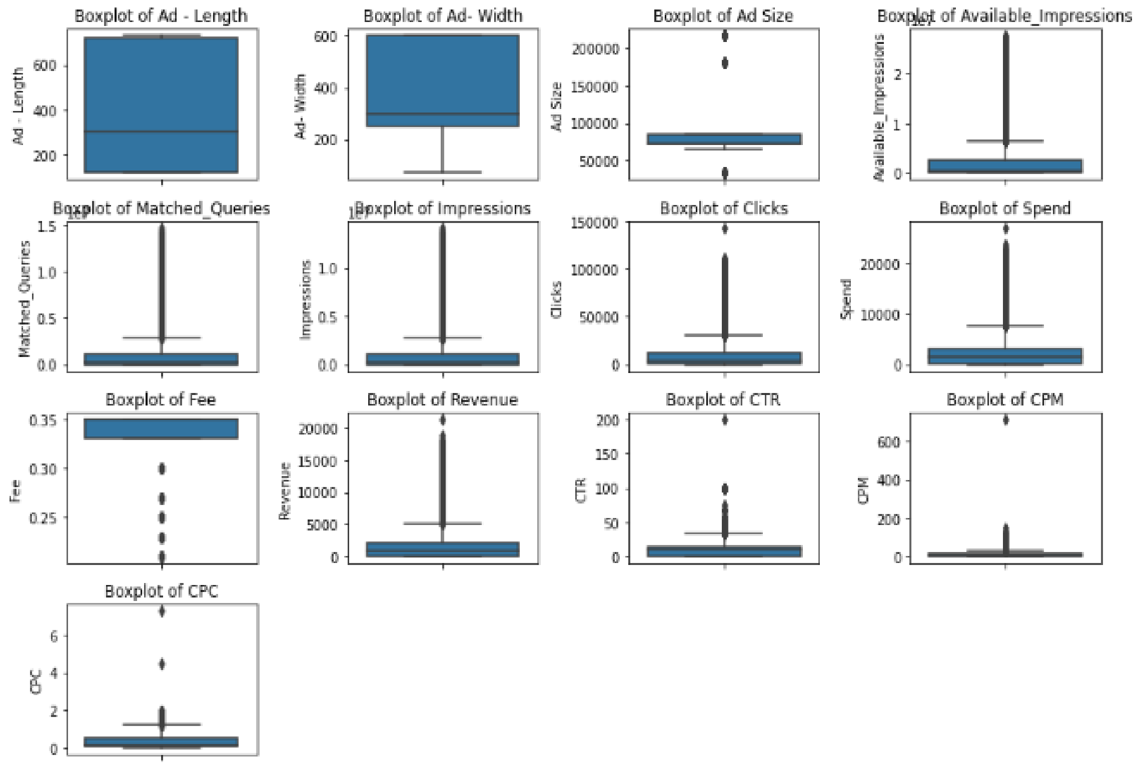
CPM = (Total Campaign Spend / Number of Impressions) * 1,000. Note that the Total Campaign Spend refers to the 'Spend' Column in the dataset and the Number of Impressions refers to the 'Impressions' Column in the dataset.

CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column in the dataset and the Number of Clicks refers to the 'Clicks' Column in the dataset.

CTR = Total Measured Clicks / Total Measured Ad Impressions x 100. Note that the Total Measured Clicks refers to the 'Clicks' Column in the dataset and the Total Measured Ad Impressions refers to the 'Impressions' Column in the dataset.

We have calculated the values using the calculate_CPM, calculate_CTR & calculate_CPM using the lambda function and all the null values are replaced by the true value using the formula given.

1.3 - 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).



As there are many outliers in the data hence it's important to treat the outliers for K-Means clustering. To treat outliers lets define a function 'remove_outlier' which returns the Upper and Lower limit to detect outliers for each feature. We have Capped & Floored the values beyond the outlier boundaries.

Figure 1

The following boxplot are obtained after treating outliers.

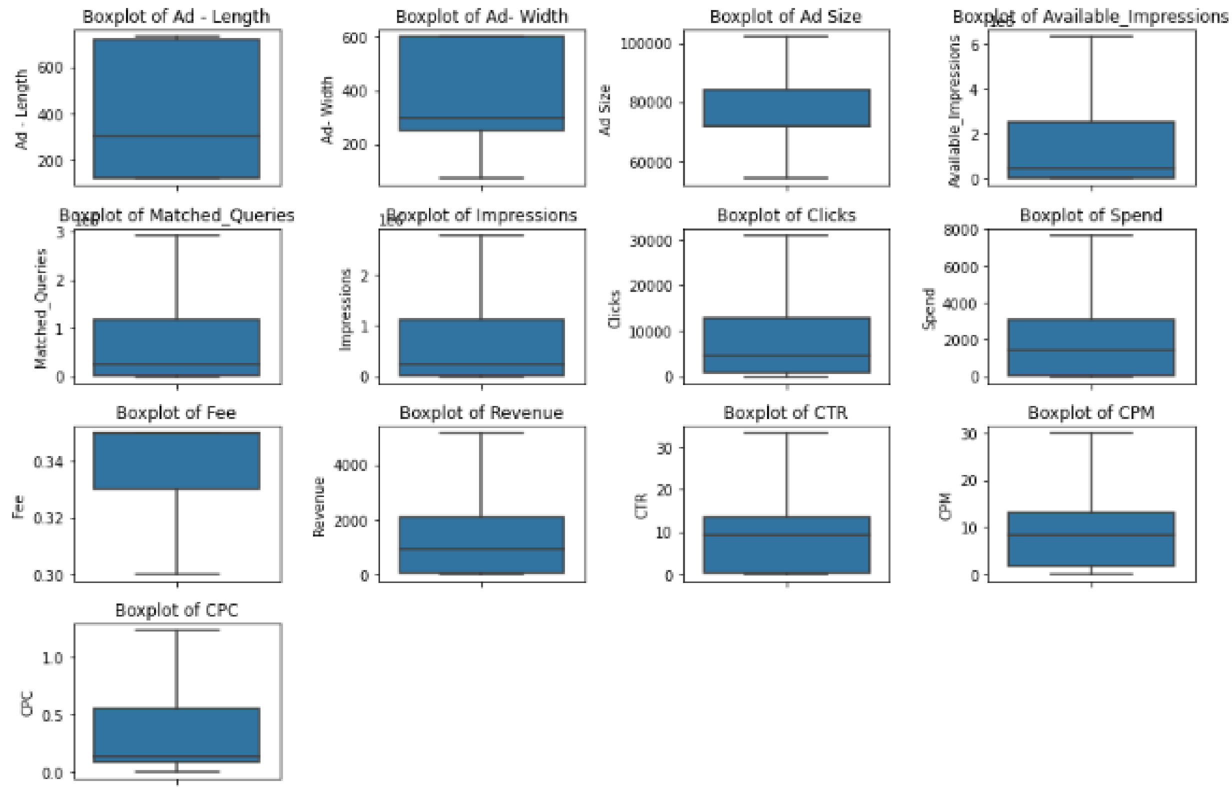
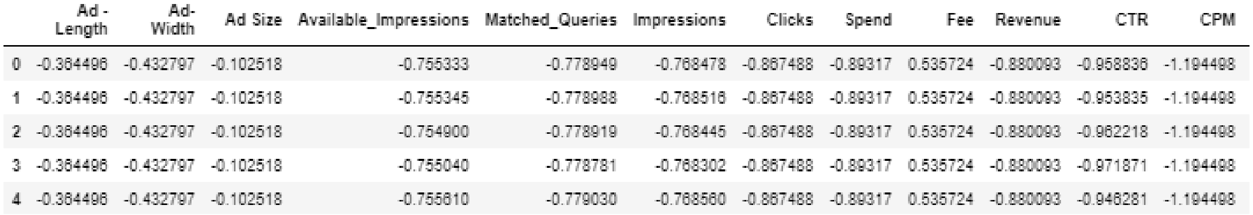


Figure 2

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

Yes, scaling is required in this data set as all features have different weights and to ensure that none of the feature is identified as important only because of the weight, scaling is mandatory for this data set. Standardization before clustering algorithm leads to obtain a better quality, efficient and accurate cluster result. Z-score is the most powerful method that will give more accurate and efficient result among the three methods in K-means clustering algorithm.

Scaled data -



	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM
0	-0.364496	-0.432797	-0.102518	-0.755333	-0.778949	-0.768478	-0.867488	-0.89317	0.535724	-0.880093	-0.958836	-1.194498
1	-0.364496	-0.432797	-0.102518	-0.755345	-0.778988	-0.768516	-0.867488	-0.89317	0.535724	-0.880093	-0.953835	-1.194498
2	-0.364496	-0.432797	-0.102518	-0.754900	-0.778919	-0.768445	-0.867488	-0.89317	0.535724	-0.880093	-0.962218	-1.194498
3	-0.364496	-0.432797	-0.102518	-0.755040	-0.778781	-0.768302	-0.867488	-0.89317	0.535724	-0.880093	-0.971871	-1.194498
4	-0.364496	-0.432797	-0.102518	-0.755610	-0.779030	-0.768560	-0.867488	-0.89317	0.535724	-0.880093	-0.946281	-1.194498

Table 4

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

Hierarchical Clustering by constructing a dendrogram using WARD

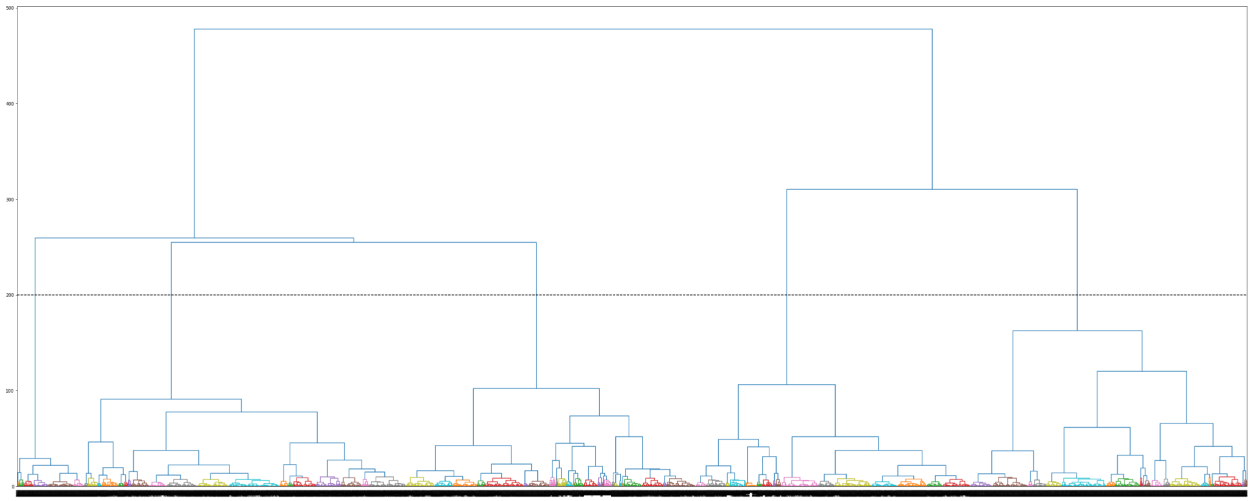


Figure 3

Hierarchical Clustering by constructing a dendrogram using Euclidean distance

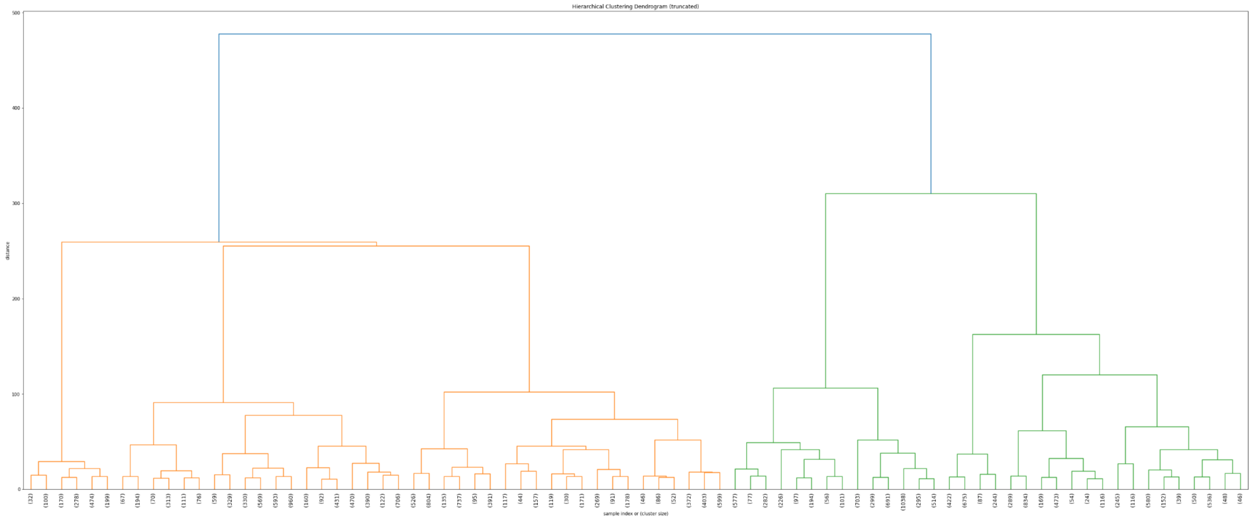
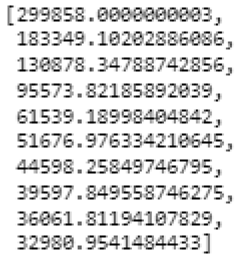


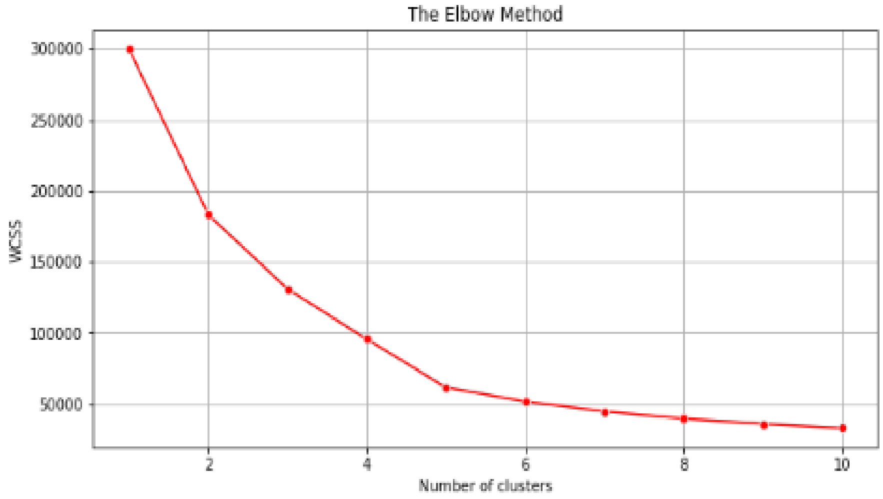
Figure 4

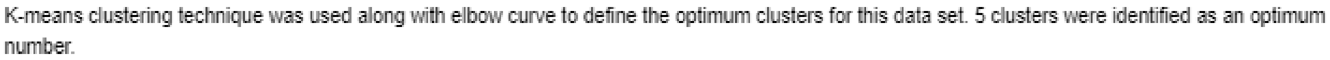
1.6 - Clustering: Make Elbow plot (up to n=10) and identify optimum number of clusters for the k-means algorithm.

Using KMeans Wcss is -



```
[299858.0000000003,
 183349.10202886086,
 130878.34788742856,
 95573.82185892039,
 61539.18998404842,
 51676.976334210645,
 44598.25849746795,
 39597.849558746275,
 36061.81194107829,
 32980.9541484433]
```





K-means clustering technique was used along with elbow curve to define the optimum clusters for this data set. 5 clusters were identified as an optimum number.

Figure 5

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

Silhouette scores for 10 clusters are -

For 2 clusters - 0.3857

For 3 clusters - 0.3825

For 4 clusters - 0.4453

For 5 clusters - 0.5240

For 6 clusters - 0.5221

For 7 clusters - 0.5165

For 8 clusters - 0.4797

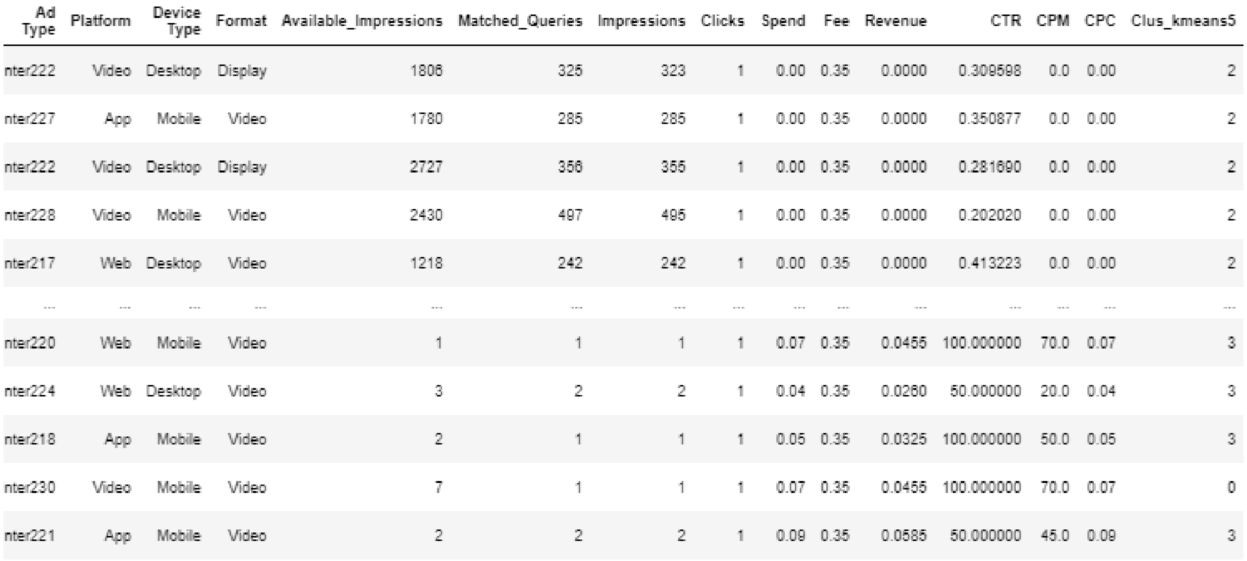
For 9 clusters - 0.4821

For 10 clusters - 0.4405

As the silhoutte score is better for 5 clusters thus the optimum number of clusters are 5.

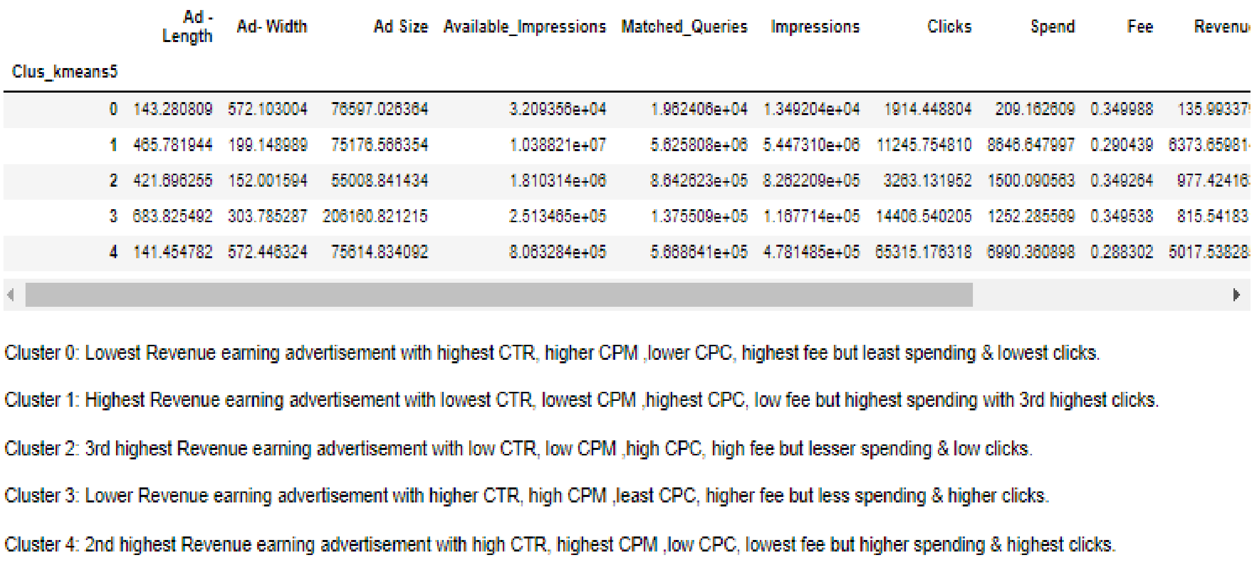
1.8 - Clustering: Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots].

Optimum number of clusters are 5 as sihoutte score is highest for 5 clusters.



Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	Clus_kmeans5
nter222	Video	Desktop	Display	1806	325	323	1	0.00	0.35	0.0000	0.309598	0.0	0.00	2
nter227	App	Mobile	Video	1780	285	285	1	0.00	0.35	0.0000	0.350877	0.0	0.00	2
nter222	Video	Desktop	Display	2727	356	355	1	0.00	0.35	0.0000	0.281690	0.0	0.00	2
nter228	Video	Mobile	Video	2430	497	495	1	0.00	0.35	0.0000	0.202020	0.0	0.00	2
nter217	Web	Desktop	Video	1218	242	242	1	0.00	0.35	0.0000	0.413223	0.0	0.00	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
nter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	100.000000	70.0	0.07	3
nter224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.0260	50.000000	20.0	0.04	3
nter218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	100.000000	50.0	0.05	3
nter230	Video	Mobile	Video	7	1	1	1	0.07	0.35	0.0455	100.000000	70.0	0.07	0
nter221	App	Mobile	Video	2	2	2	1	0.09	0.35	0.0585	50.000000	45.0	0.09	3

Table 5



	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenu
Clus_kmeans5										
0	143.280809	572.103004	76597.026364	3.209356e+04	1.962406e+04	1.349204e+04	1914.448804	209.162609	0.349988	135.99337
1	465.781944	199.148989	75176.566354	1.038821e+07	5.625808e+06	5.447310e+06	11245.754810	8646.647997	0.290439	6373.65981
2	421.696255	152.001594	55008.841434	1.810314e+06	8.642623e+05	8.262209e+05	3263.131952	1500.090563	0.349264	977.42416
3	683.825492	303.785287	206160.821215	2.513465e+05	1.375509e+05	1.167714e+05	14406.540205	1252.285569	0.349538	815.54183
4	141.454782	572.446324	75614.834092	8.083284e+05	5.688641e+05	4.781485e+05	65315.176318	6990.360898	0.288302	5017.53828

Cluster 0: Lowest Revenue earning advertisement with highest CTR, higher CPM ,lower CPC, highest fee but least spending & lowest clicks.

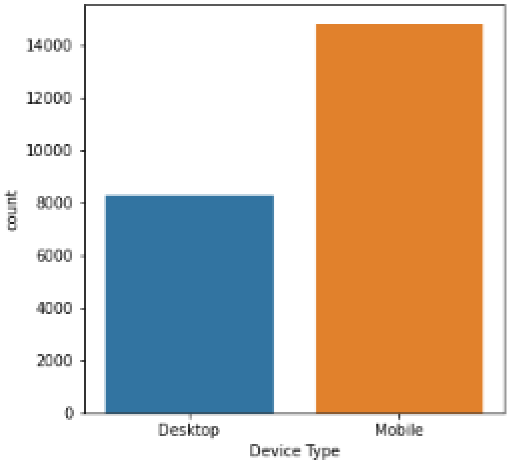
Cluster 1: Highest Revenue earning advertisement with lowest CTR, lowest CPM ,highest CPC, low fee but highest spending with 3rd highest clicks.

Cluster 2: 3rd highest Revenue earning advertisement with low CTR, low CPM ,high CPC, high fee but lesser spending & low clicks.

Cluster 3: Lower Revenue earning advertisement with higher CTR, high CPM ,least CPC, higher fee but less spending & higher clicks.

Cluster 4: 2nd highest Revenue earning advertisement with high CTR, highest CPM ,low CPC, lowest fee but higher spending & highest clicks.

Table 6



Mobile type advertisement has the most count hence leading to greater revenue and highest number of clicks.

Figure 6

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

1. Ads in Cluster 0 has lowest Revenue earning advertisement with highest CTR, higher CPM ,lower CPC, highest fee but least spending & lowest clicks. So it needs to improve in making the Ads more attractive. As the Fees is highest thus leading to lowest clicks . So, can promote offers on such Ads to increase revenue.
2. Ads in Cluster 1 has highest Revenue earning advertisement with lowest CTR, lowest CPM ,highest CPC, low fee but highest spending with 3rd highest clicks. Such ads should be prefered as they are having low fee but highest revenue generating Ads, these types of Ads are needed for mobile device type Ads to increase number of clicks hence to increase more revenue.
3. Ads in Cluster 2 has 3rd highest Revenue earning advertisement with low CTR, low CPM ,high CPC, high fee but lesser spending & low clicks. As the fee is high thus leading to low clicks , more attractive Ads are needed to increase clicks and increase revenue.
4. Ads in Cluster 3 has Lower Revenue earning advertisement with higher CTR, high CPM ,least CPC, higher fee but less spending & higher clicks. These type of Ads are not generating much revenue but they have higher clicks thus there is a need to provide more offers on such Ads to increase revenue.
5. Ads in Cluster 4 has 2nd highest Revenue earning advertisement with high CTR, highest CPM ,low CPC, lowest fee but higher spending & highest clicks. Such Ads are generating good revenue amount but they have highest CPM so cost should be reduced to increase more revenue.
6. Larger Ad Size doesn't seem to have greater impact on revenue. Cluster 3 Ads has extremely large Mean Ad size but still are lower revenue earning advertisement thus Ad size should be reduced as cost can be reduced thus leading to more revenue.

Problem 2

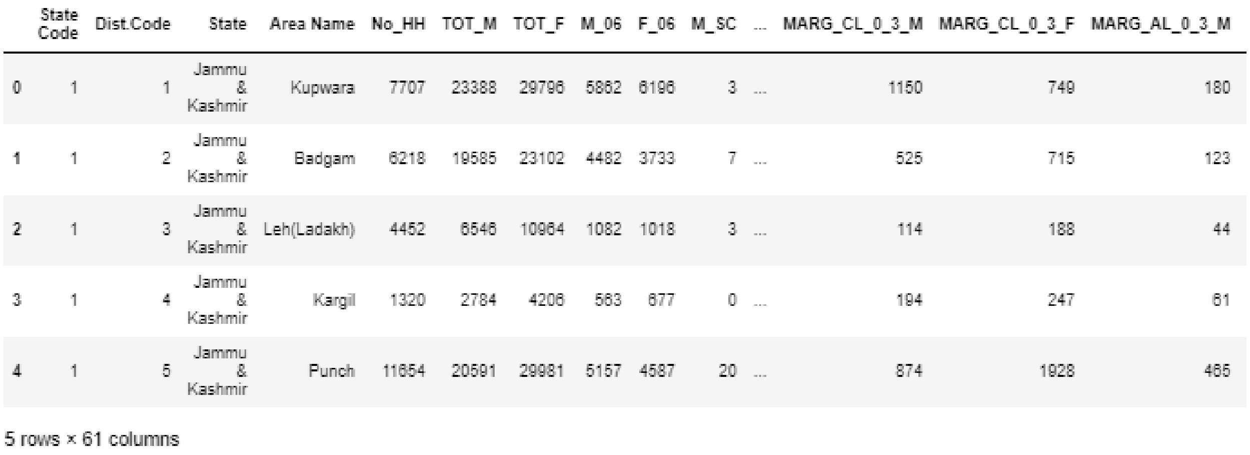
PCA:

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.

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

Reading the data -

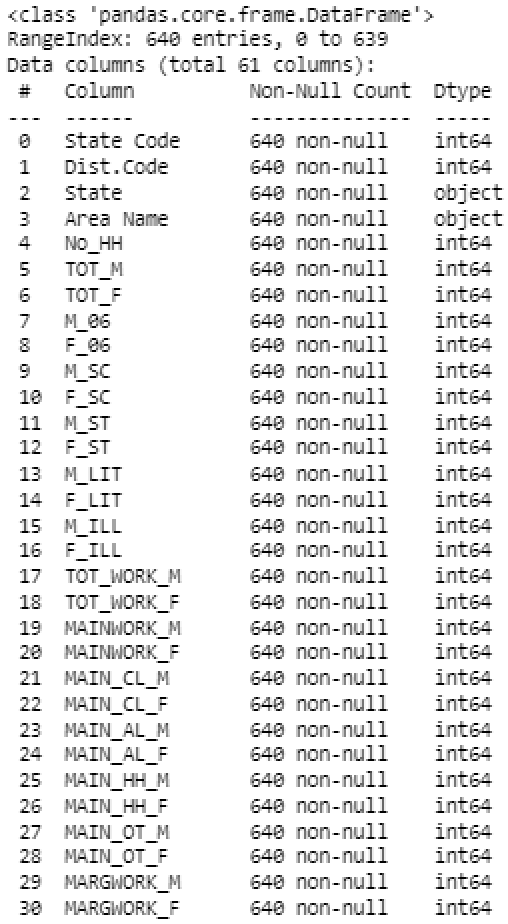


	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	...	114	188	44
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	...	194	247	61
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	...	874	1928	465

5 rows × 61 columns

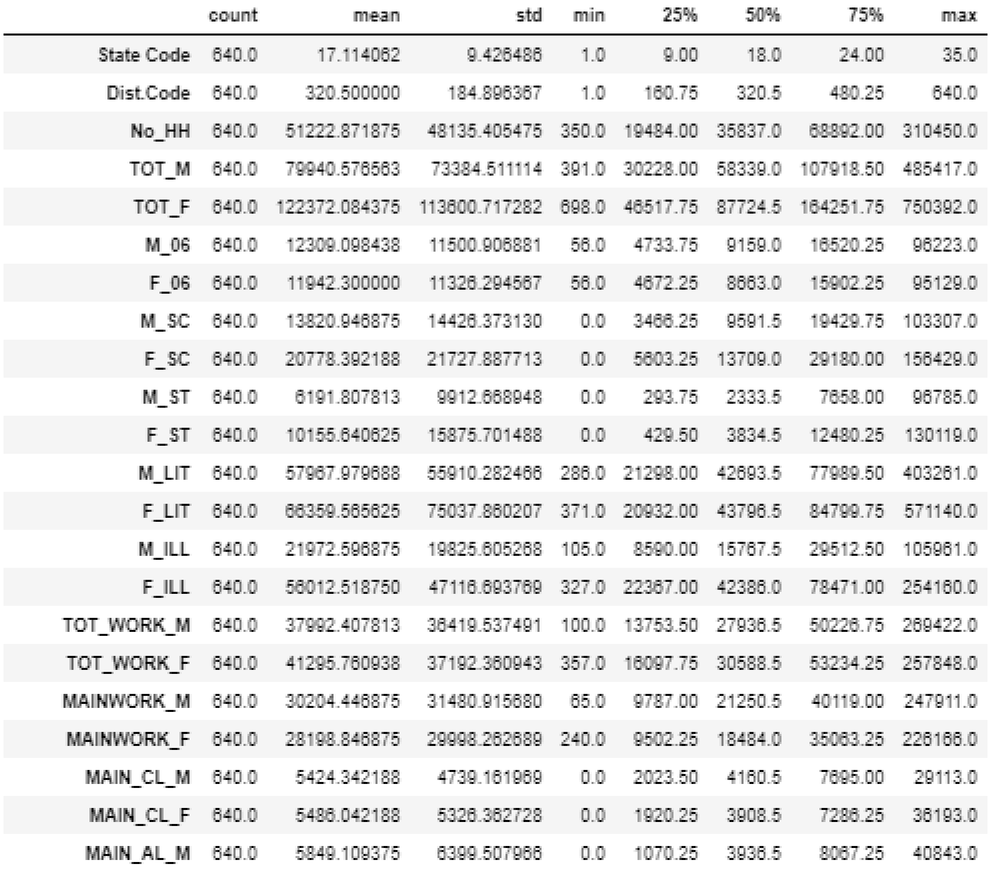
Table 7

Data Information -



```
<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 
```

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

Table 8

No duplicates

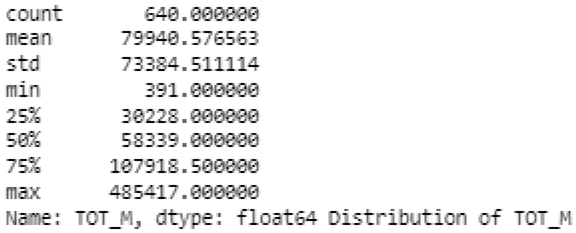
No missing values

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

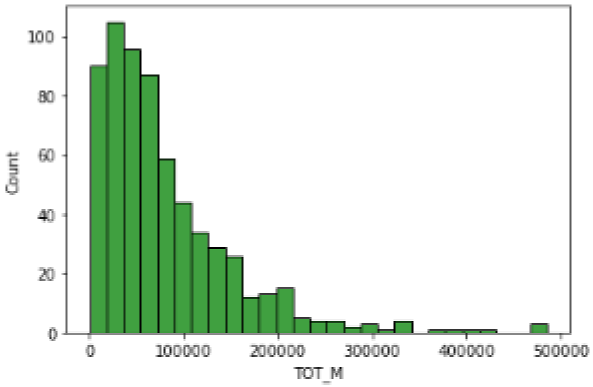
Performing detailed exploratory analysis on 5 variables such as

Total population Male,Total population Female,Literates population Male,Total Worker Population Male & Non Working Population Male.

Total population male -



```
count       640.000000
mean      79940.576563
std       73384.511114
min         391.000000
25%       30228.000000
50%       58339.000000
75%      107918.500000
max      485417.000000
Name: TOT_M, dtype: float64 Distribution of TOT_M
```



```
BoxPlot of TOT_M
-------------------------------------------------------
```

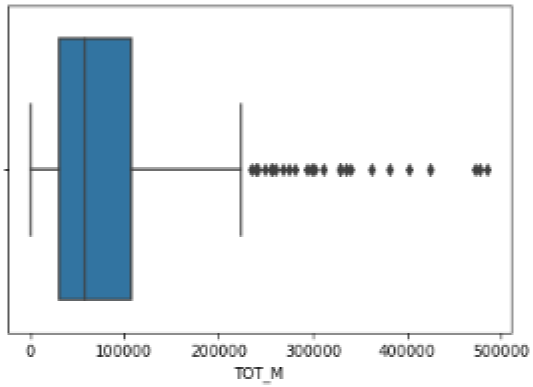
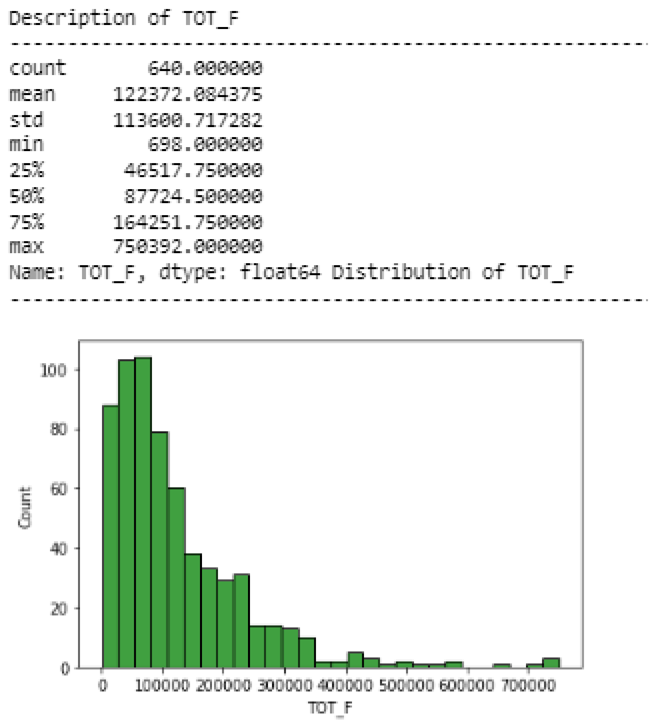
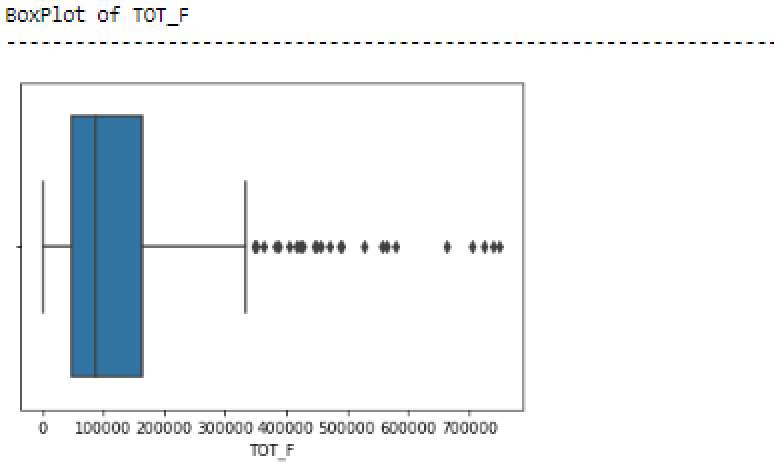


Figure 7

<u>Total population Female</u>



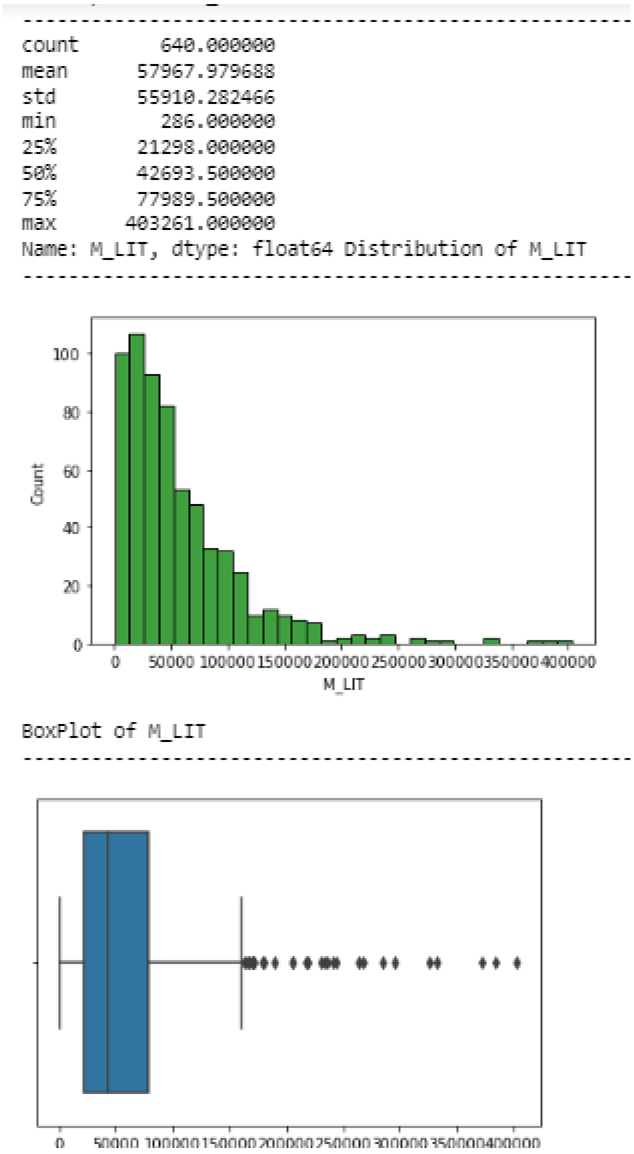
```
Description of TOT_F
-------------------------------------------------------
count       640.000000
mean     122372.084375
std      113600.717282
min         698.000000
25%       46517.750000
50%       87724.500000
75%      164251.750000
max      750392.000000
Name: TOT_F, dtype: float64 Distribution of TOT_F
-------------------------------------------------------
```



```
BoxPlot of TOT_F
------------------------------------------------------------------
```

Figure 8

Literates population Male

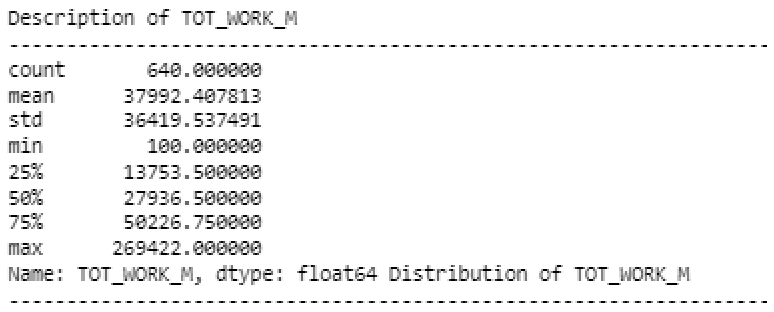


```
-----------------------------------------------------
count       640.000000
mean      57967.979688
std       55910.282466
min         286.000000
25%       21298.000000
50%       42693.500000
75%       77989.500000
max      403261.000000
Name: M_LIT, dtype: float64 Distribution of M_LIT
-----------------------------------------------------
```

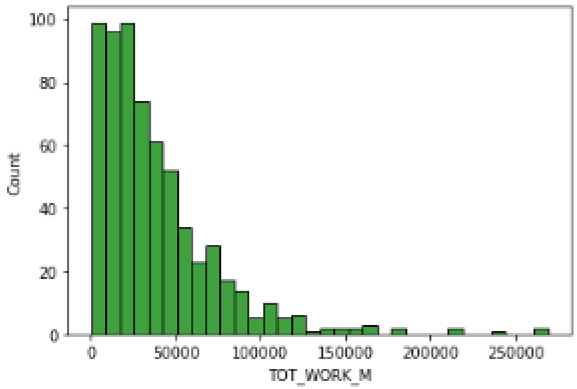
```
BoxPlot of M_LIT
-----------------------------------------------------
```

Figure 9

Total Worker Population Male



```
Description of TOT_WORK_M
-----------------------------------------------------------------
count       640.000000
mean      37992.407813
std       36419.537491
min         100.000000
25%       13753.500000
50%       27936.500000
75%       50226.750000
max      269422.000000
Name: TOT_WORK_M, dtype: float64 Distribution of TOT_WORK_M
-----------------------------------------------------------------
```



```
BoxPlot of TOT_WORK_M
--------------------------------------------------
```

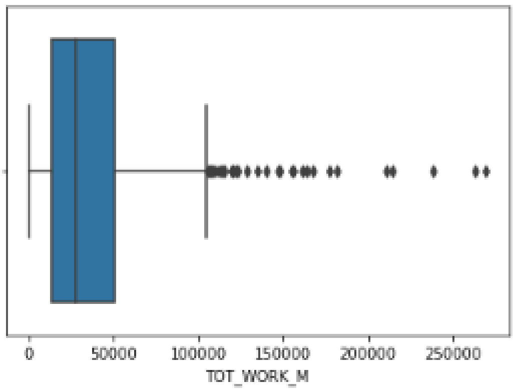
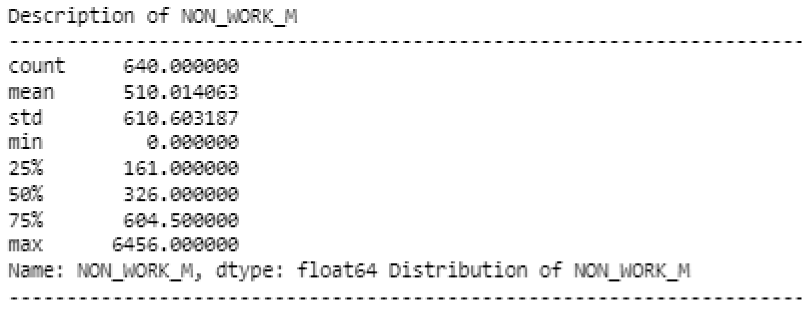
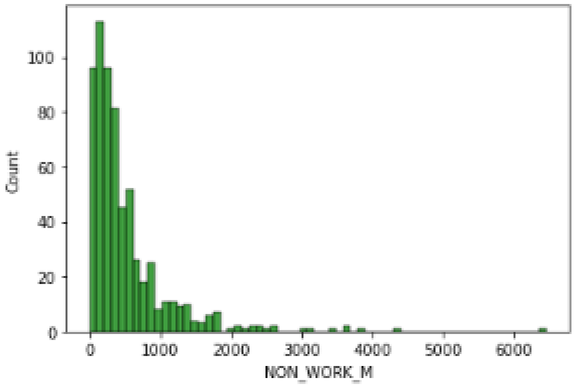


Figure 10

Non Working Population Male



```
Description of NON_WORK_M
--------------------------------------------------------------------
count     640.000000
mean      510.014063
std       610.603187
min         0.000000
25%       161.000000
50%       326.000000
75%       604.500000
max      6456.000000
Name: NON_WORK_M, dtype: float64 Distribution of NON_WORK_M
--------------------------------------------------------------------
```



```
BoxPlot of NON_WORK_M
----------------------------------------------------------------
```

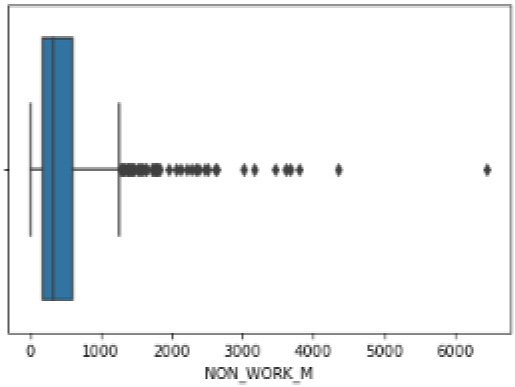


Figure 11

<u>State Count -</u>

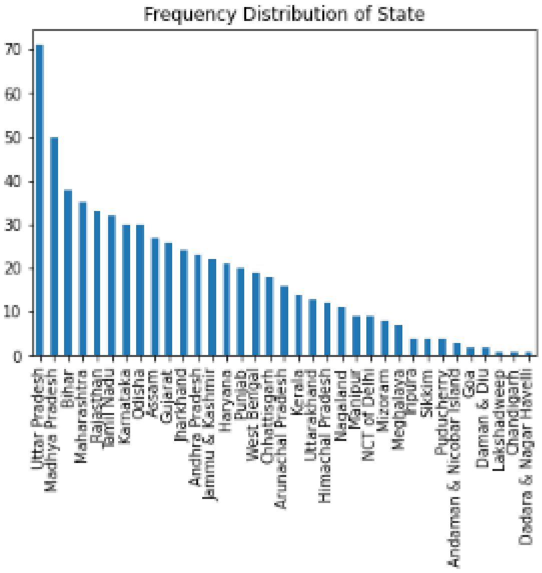


Figure 12

Uttar Pradesh has the highest number of count and Chandigarh and Dadara has the least count.

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

Although there are many outliers as shown in the below figure

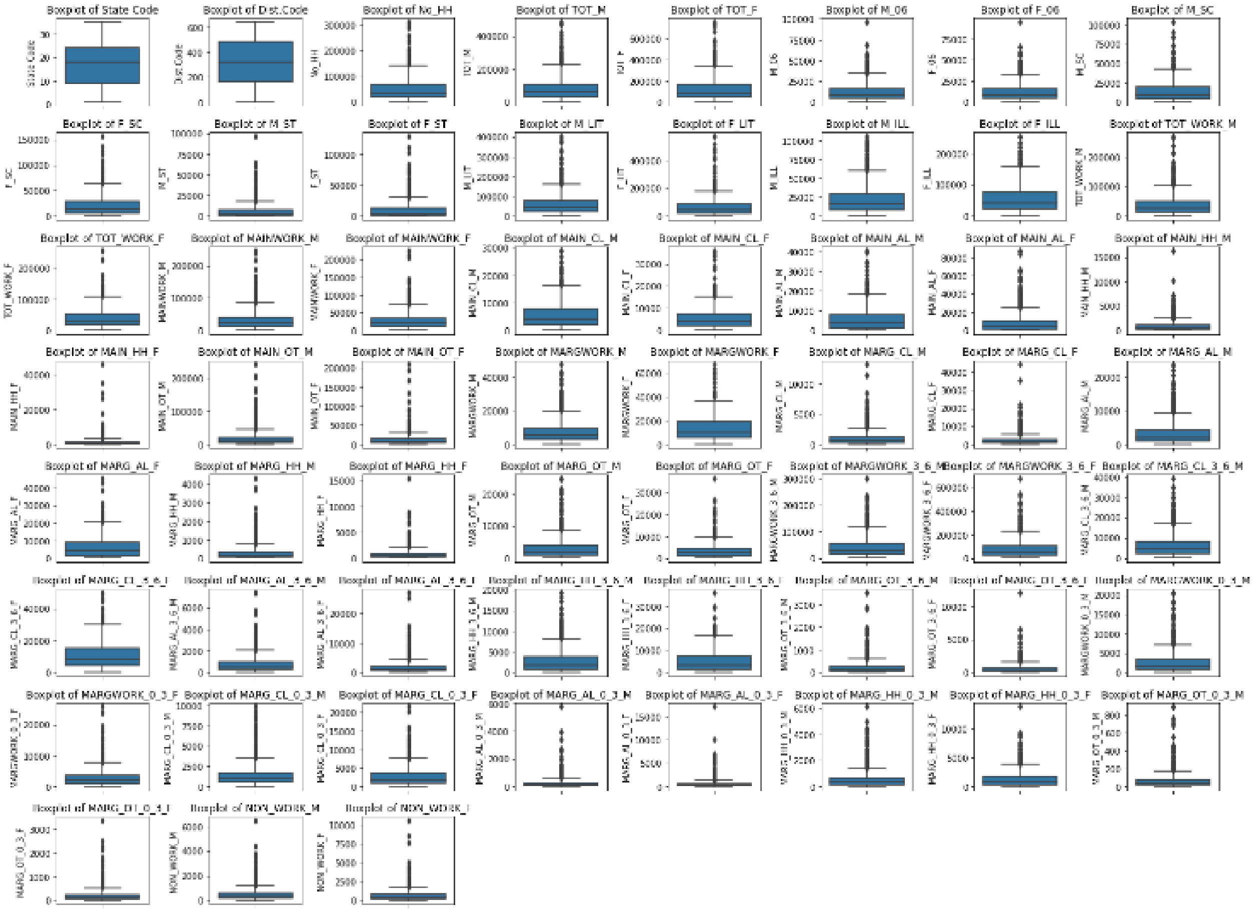
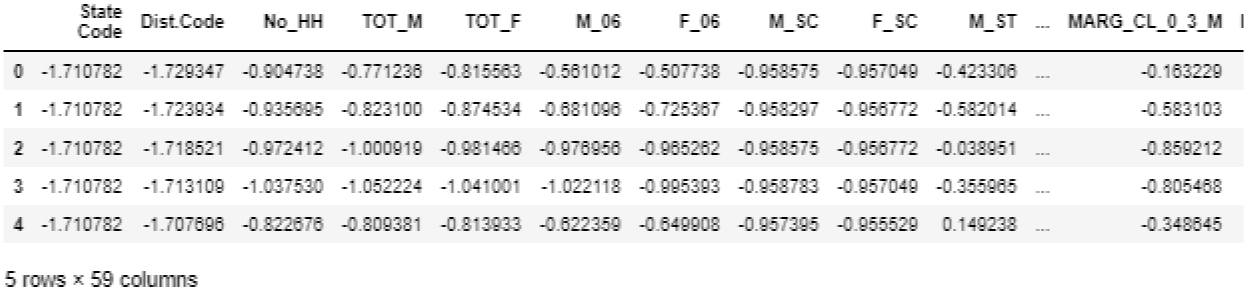


Figure 13

No treating outliers is not necessary in this case as you never eliminate outliers unless they are the result from a processing mistake or wrong measurement. True outliers must be kept in the data while doing PCA.

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

Scaled data using z score method -



	State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MARG_CL_0_3_M
0	-1.710782	-1.729347	-0.904738	-0.771236	-0.815563	-0.561012	-0.507738	-0.958575	-0.957049	-0.423306	...	-0.163229
1	-1.710782	-1.723934	-0.935695	-0.823100	-0.874534	-0.681096	-0.725367	-0.958297	-0.956772	-0.582014	...	-0.583103
2	-1.710782	-1.718521	-0.972412	-1.000919	-0.981466	-0.976956	-0.965262	-0.958575	-0.956772	-0.038951	...	-0.859212
3	-1.710782	-1.713109	-1.037530	-1.052224	-1.041001	-1.022118	-0.995393	-0.958783	-0.957049	-0.355965	...	-0.805468
4	-1.710782	-1.707696	-0.822676	-0.809381	-0.813933	-0.622359	-0.649908	-0.957395	-0.955529	0.149238	...	-0.348645

5 rows × 59 columns

Table 9

Boxplot before scaling -

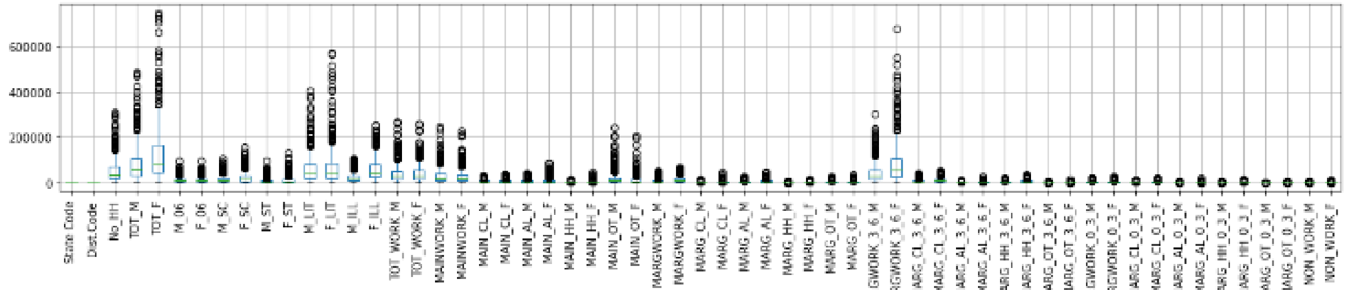


Figure 14

Boxplot after scaling -

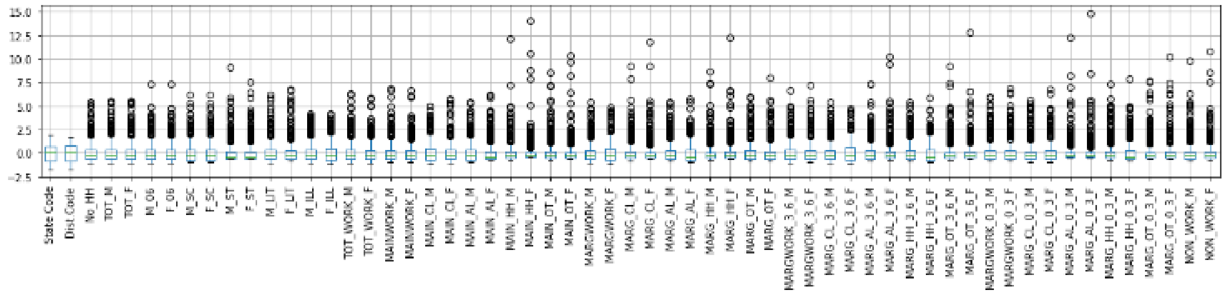


Figure 15

As we can see scaling does have an impact on outliers.

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

Showing the heatmap -

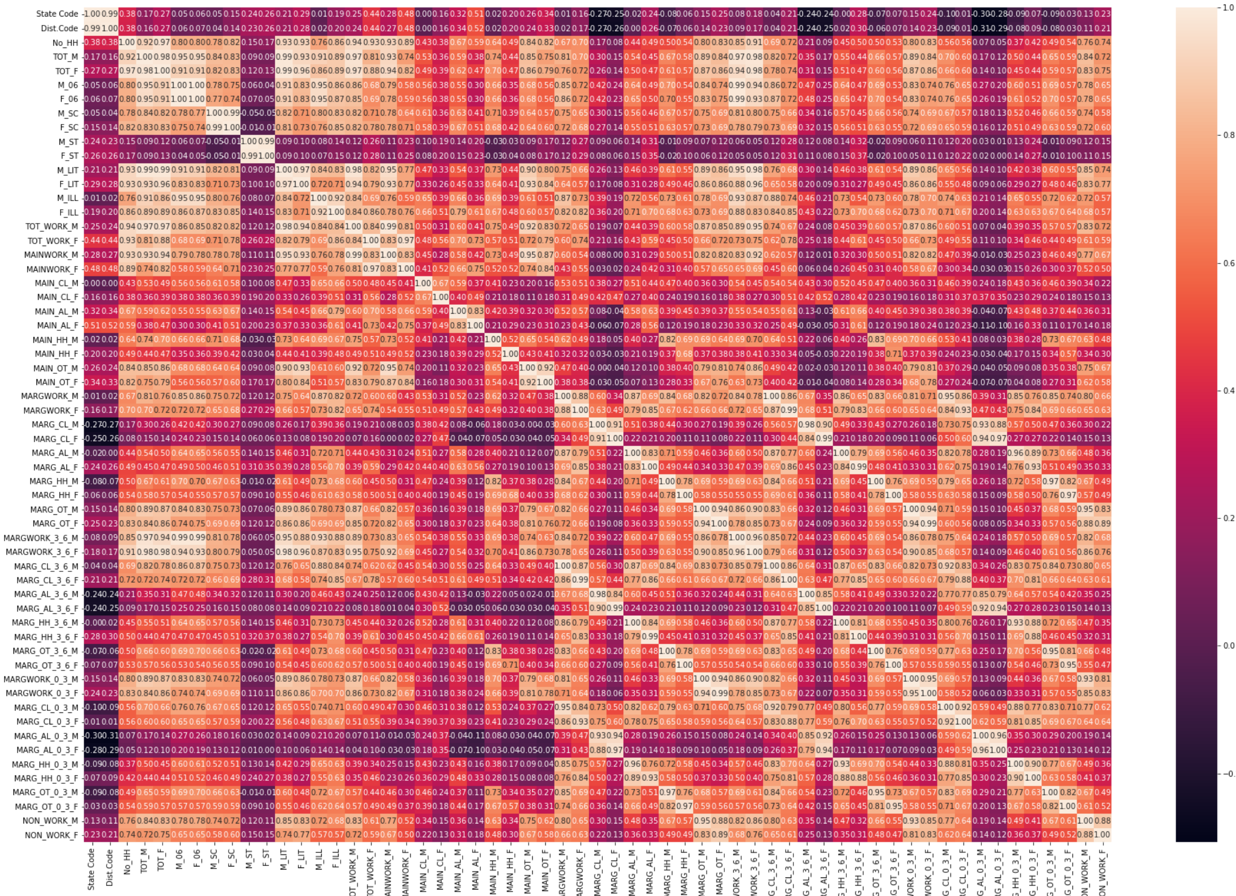
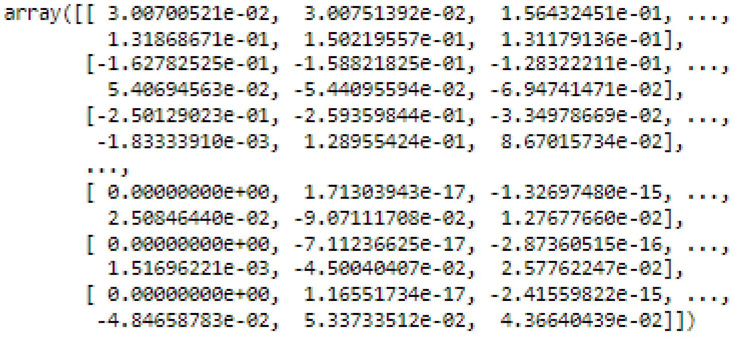


Figure 16

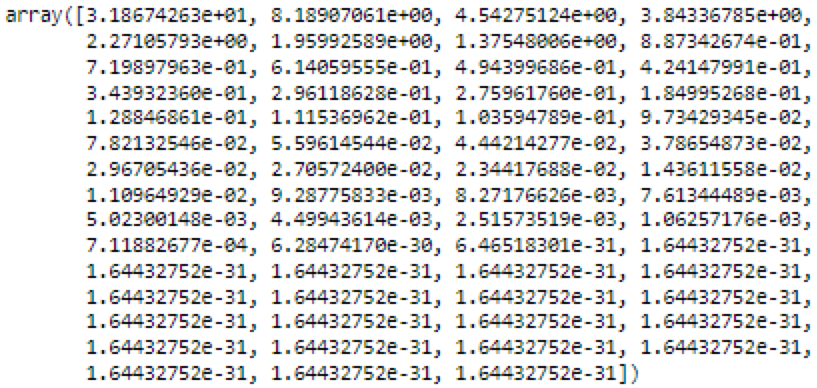
Applying PCA

Eigen vectors are -



```
array([[ 3.00700521e-02,  3.00751392e-02,  1.56432451e-01, ...,
         1.31868671e-01,  1.50219557e-01,  1.31179136e-01],
       [-1.62782525e-01, -1.58821825e-01, -1.28322211e-01, ...,
         5.40694563e-02, -5.44095594e-02, -6.94741471e-02],
       [-2.50129023e-01, -2.59359844e-01, -3.34978669e-02, ...,
        -1.83333910e-03,  1.28955424e-01,  8.67015734e-02],
       ...,
       [ 0.00000000e+00,  1.71303943e-17, -1.32697480e-15, ...,
         2.50846440e-02, -9.07111708e-02,  1.27677660e-02],
       [ 0.00000000e+00, -7.11236625e-17, -2.87360515e-16, ...,
         1.51696221e-03, -4.50040407e-02,  2.57762247e-02],
       [ 0.00000000e+00,  1.16551734e-17, -2.41559822e-15, ...,
        -4.84658783e-02,  5.33733512e-02,  4.36640439e-02]])
```

Eigen Values are -



```
array([3.18674263e+01, 8.18907061e+00, 4.54275124e+00, 3.84336785e+00,
       2.27105793e+00, 1.95992589e+00, 1.37548006e+00, 8.87342674e-01,
       7.19897963e-01, 6.14059555e-01, 4.94399686e-01, 4.24147991e-01,
       3.43932360e-01, 2.96118628e-01, 2.75961760e-01, 1.84995268e-01,
       1.28846861e-01, 1.11536962e-01, 1.03594789e-01, 9.73429345e-02,
       7.82132546e-02, 5.59614544e-02, 4.44214277e-02, 3.78654873e-02,
       2.96705436e-02, 2.70572400e-02, 2.34417688e-02, 1.43611558e-02,
       1.10964929e-02, 9.28775833e-03, 8.27176626e-03, 7.61344489e-03,
       5.02300148e-03, 4.49943614e-03, 2.51573519e-03, 1.06257176e-03,
       7.11882677e-04, 6.28474170e-30, 6.46518301e-31, 1.64432752e-31,
       1.64432752e-31, 1.64432752e-31, 1.64432752e-31, 1.64432752e-31,
       1.64432752e-31, 1.64432752e-31, 1.64432752e-31, 1.64432752e-31,
       1.64432752e-31, 1.64432752e-31, 1.64432752e-31, 1.64432752e-31,
       1.64432752e-31, 1.64432752e-31, 1.64432752e-31, 1.64432752e-31,
       1.64432752e-31, 1.64432752e-31, 1.64432752e-31])
```

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

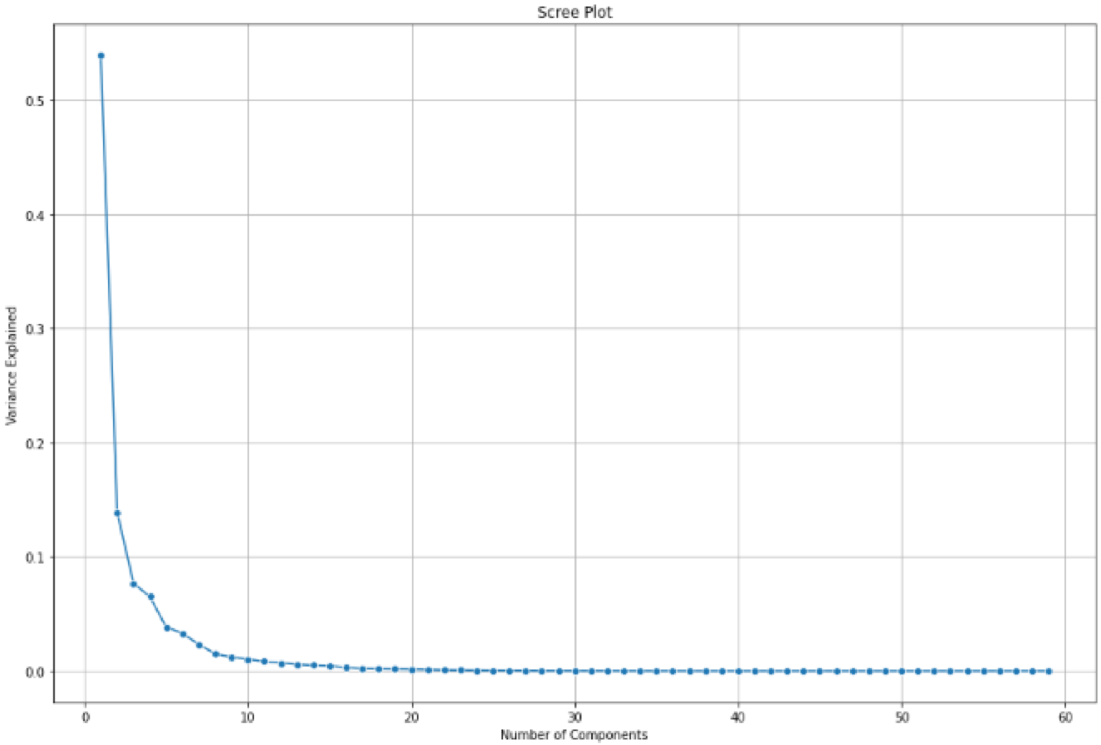
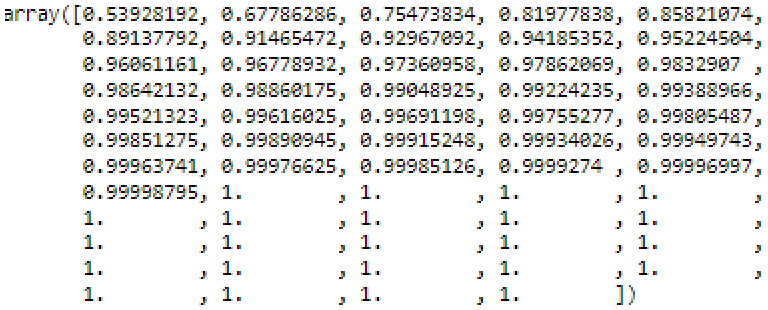


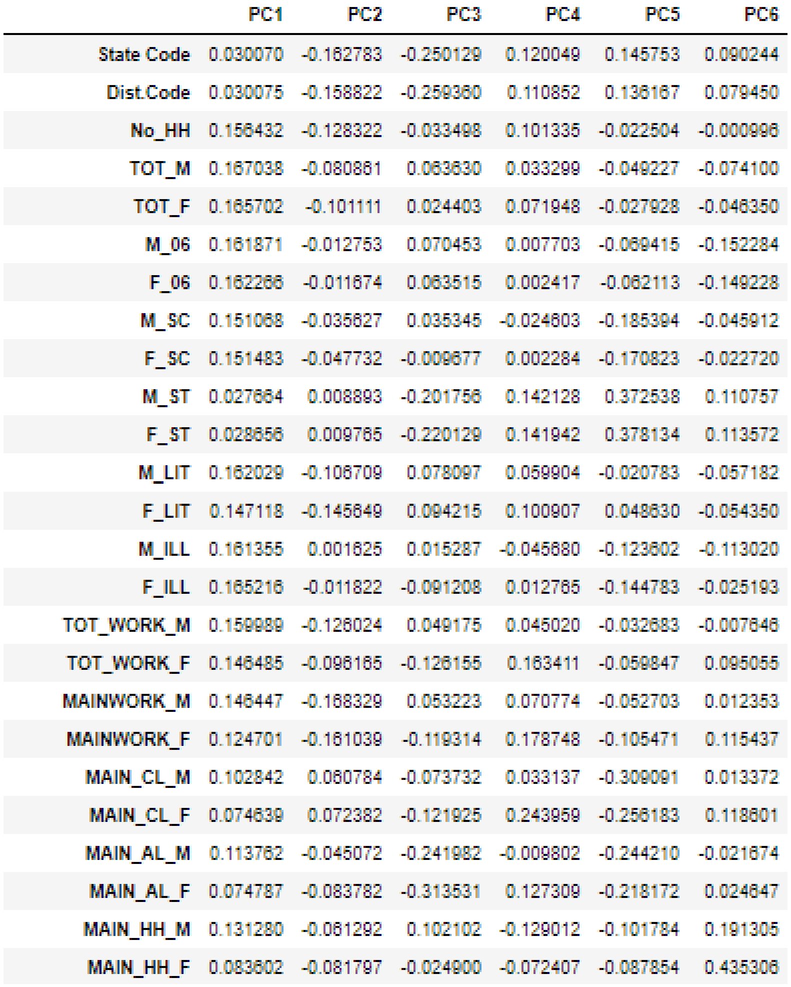
Figure 17

Cumulative explained variance



```
array([0.53928192, 0.67786286, 0.75473834, 0.81977838, 0.85821074,
       0.89137792, 0.91465472, 0.92967092, 0.94185352, 0.95224504,
       0.96061161, 0.96778932, 0.97360958, 0.97862069, 0.9832907 ,
       0.98642132, 0.98860175, 0.99048925, 0.99224235, 0.99388966,
       0.99521323, 0.99616025, 0.99691198, 0.99755277, 0.99805487,
       0.99851275, 0.99890945, 0.99915248, 0.99934026, 0.99949743,
       0.99963741, 0.99976625, 0.99985126, 0.9999274 , 0.99996997,
       0.99998795, 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        ])
```

As 6 PCA components are explaining atleast 90% explained variance thus 6 PCs are the optimum number of PCs.



	PC1	PC2	PC3	PC4	PC5	PC6
State Code	0.030070	-0.162783	-0.250129	0.120049	0.145753	0.090244
Dist.Code	0.030075	-0.158822	-0.259360	0.110852	0.136167	0.079450
No_HH	0.156432	-0.128322	-0.033498	0.101335	-0.022504	-0.000996
TOT_M	0.167038	-0.080861	0.063630	0.033299	-0.049227	-0.074100
TOT_F	0.165702	-0.101111	0.024403	0.071948	-0.027928	-0.046350
M_06	0.161871	-0.012753	0.070453	0.007703	-0.069415	-0.152284
F_06	0.162266	-0.011674	0.063515	0.002417	-0.062113	-0.149228
M_SC	0.151068	-0.035627	0.035345	-0.024603	-0.185394	-0.045912
F_SC	0.151483	-0.047732	-0.009677	0.002284	-0.170823	-0.022720
M_ST	0.027664	0.008893	-0.201756	0.142128	0.372538	0.110757
F_ST	0.028656	0.009765	-0.220129	0.141942	0.378134	0.113572
M_LIT	0.162029	-0.106709	0.078097	0.059904	-0.020783	-0.057182
F_LIT	0.147118	-0.145649	0.094215	0.100907	0.048630	-0.054350
M_ILL	0.161355	0.001625	0.015287	-0.045680	-0.123602	-0.113020
F_ILL	0.165216	-0.011822	-0.091208	0.012765	-0.144783	-0.025193
TOT_WORK_M	0.159989	-0.126024	0.049175	0.045020	-0.032683	-0.007646
TOT_WORK_F	0.146485	-0.096165	-0.126155	0.163411	-0.059847	0.095055
MAINWORK_M	0.146447	-0.168329	0.053223	0.070774	-0.052703	0.012353
MAINWORK_F	0.124701	-0.161039	-0.119314	0.178748	-0.105471	0.115437
MAIN_CL_M	0.102842	0.060784	-0.073732	0.033137	-0.309091	0.013372
MAIN_CL_F	0.074639	0.072382	-0.121925	0.243959	-0.256183	0.118601
MAIN_AL_M	0.113762	-0.045072	-0.241982	-0.009802	-0.244210	-0.021674
MAIN_AL_F	0.074787	-0.083782	-0.313531	0.127309	-0.218172	0.024647
MAIN_HH_M	0.131280	-0.061292	0.102102	-0.129012	-0.101784	0.191305
MAIN_HH_F	0.083602	-0.081797	-0.024900	-0.072407	-0.087854	0.435306

Table 10

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

Comparing PCs with actual columns

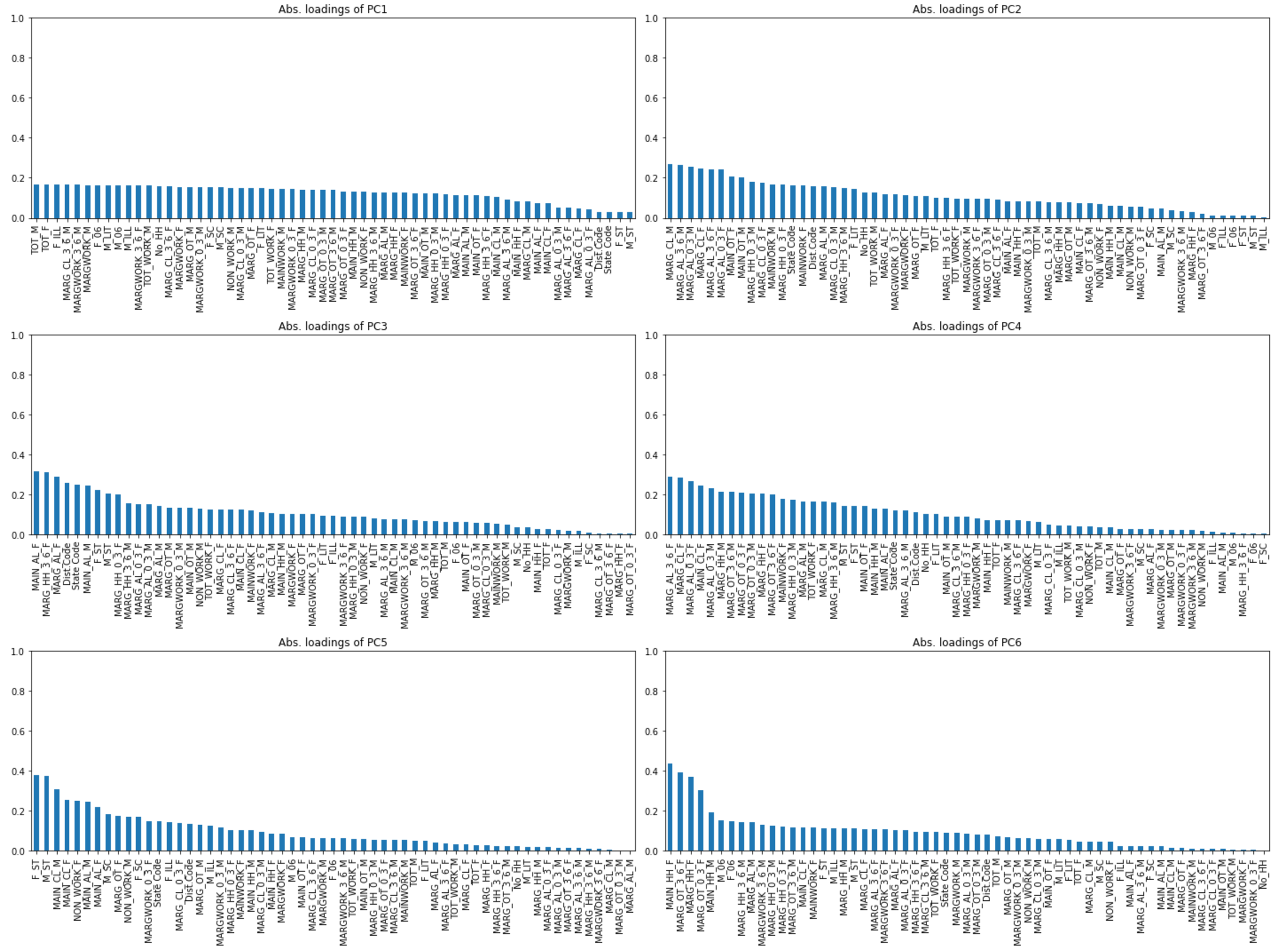
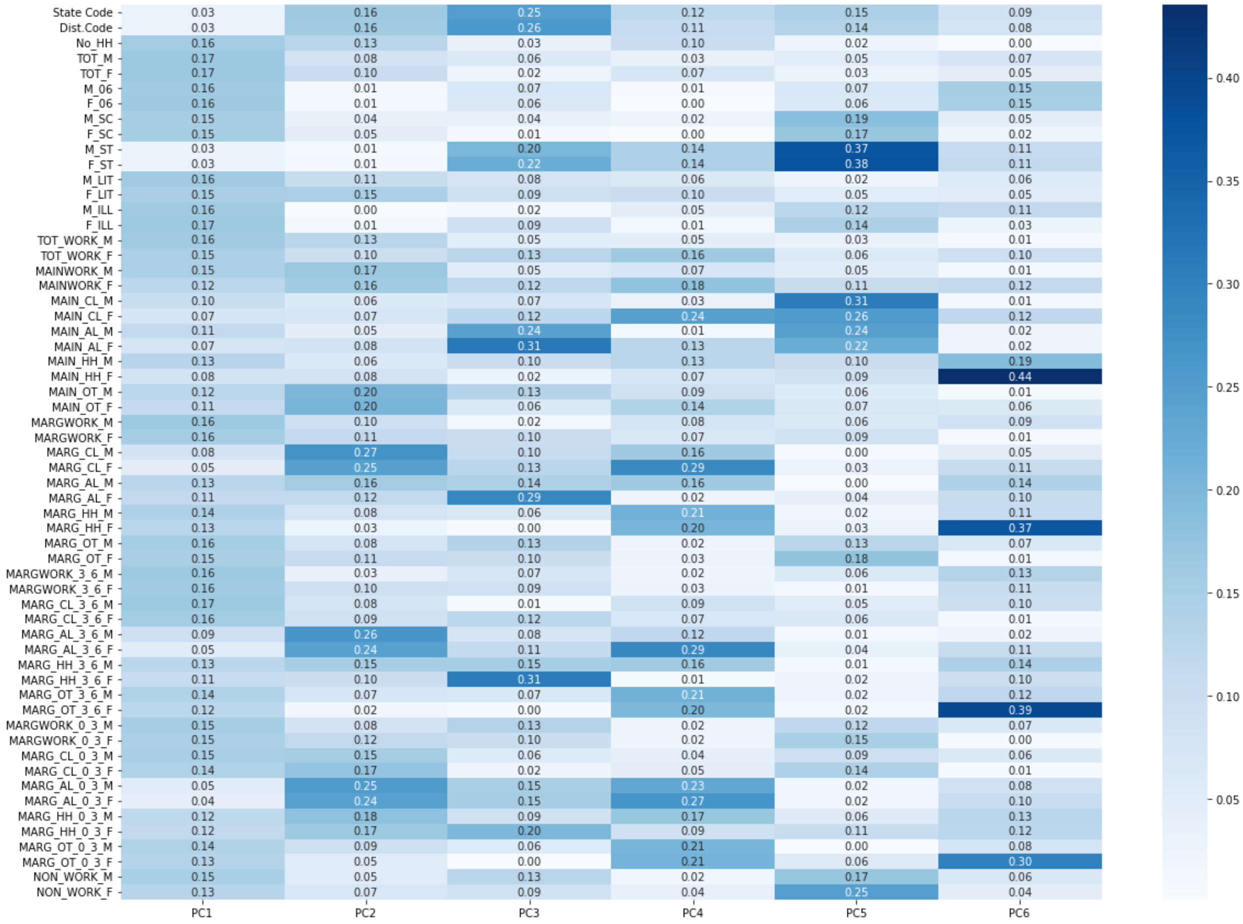


Figure 18

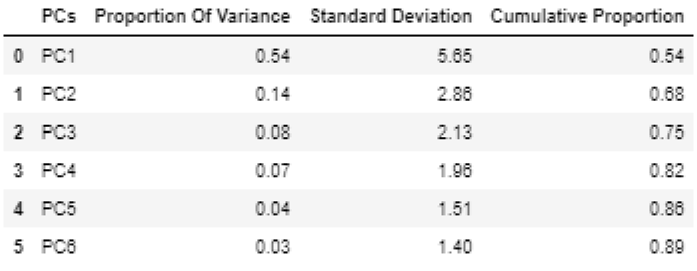
Heatmap



This heatmap shows the correlation between PCs and actual variables. Higher correlation shows that PC6 is highly correlated with Main Household Industries Population Female variable.

Figure 19

6 PCs with their standard deviation shows that the most explained variance is of PC1 i.e. square of standard deviation



	PCs	Proportion Of Variance	Standard Deviation	Cumulative Proportion
0	PC1	0.54	5.65	0.54
1	PC2	0.14	2.86	0.68
2	PC3	0.08	2.13	0.75
3	PC4	0.07	1.96	0.82
4	PC5	0.04	1.51	0.86
5	PC6	0.03	1.40	0.89

Table 11

To check that the PCs are orthogonal, correlation matrix is computed

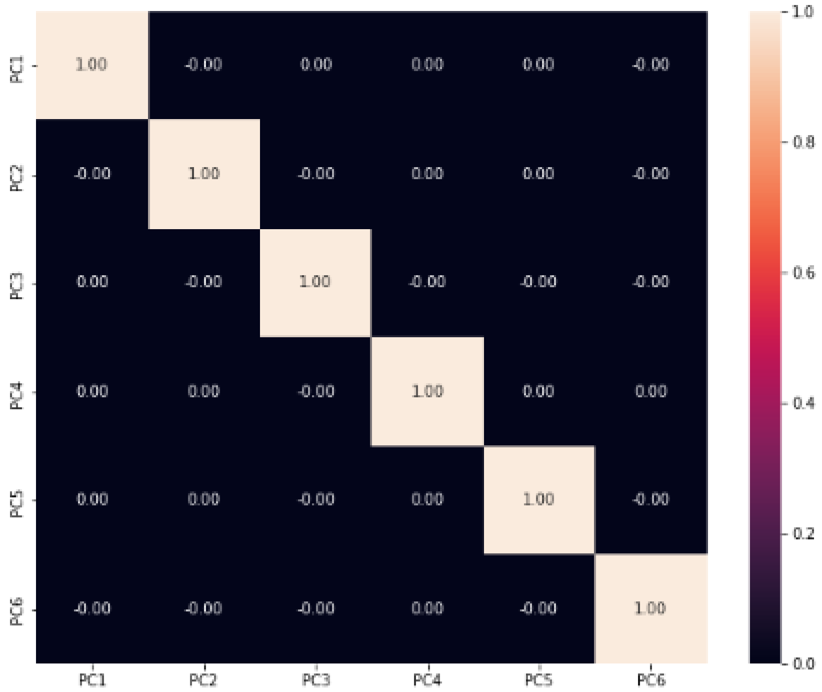
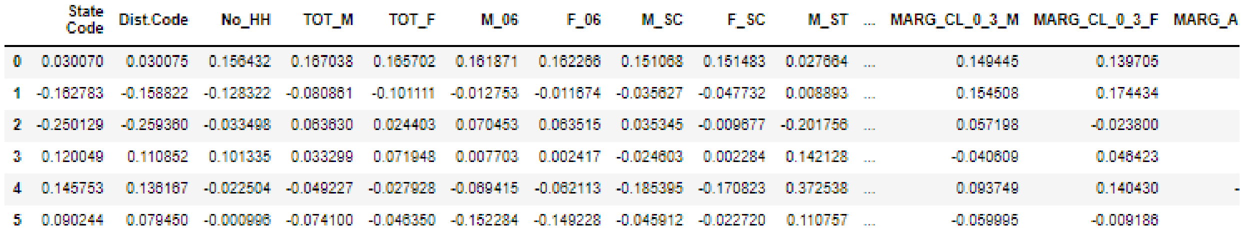


Figure 20

2.8 - PCA: Write linear equation for first PC.

Linear equation for PC1 = a1x1 + a2x2 + a3x3 + ...+ anXn , where a1,a2....aN are the coefficients or loadings and x1,x2,x3.....xn are the observed data .

Data according to 6 PCs with actual variables -



	State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_A
0	0.030070	0.030075	0.156432	0.167038	0.165702	0.161871	0.162266	0.151068	0.151483	0.027664	...	0.149445	0.139705	
1	-0.162783	-0.158822	-0.128322	-0.080861	-0.101111	-0.012753	-0.011674	-0.035627	-0.047732	0.008893	...	0.154508	0.174434	
2	-0.250129	-0.259360	-0.033498	0.063630	0.024403	0.070453	0.063515	0.035345	-0.009677	-0.201756	...	0.057198	-0.023800	
3	0.120049	0.110852	0.101335	0.033299	0.071948	0.007703	0.002417	-0.024603	0.002284	0.142128	...	-0.040609	0.046423	
4	0.145753	0.136167	-0.022504	-0.049227	-0.027928	-0.069415	-0.062113	-0.185395	-0.170823	0.372538	...	0.093749	0.140430	-
5	0.090244	0.079450	-0.000996	-0.074100	-0.046350	-0.152284	-0.149228	-0.045912	-0.022720	0.110757	...	-0.059995	-0.009186	

Table 12

Linear equation for **PC1 = (0.03070)X1 + (0.030075)X2 + (0.156432)X3 + ...** where X1 is state code, X2 is dist.Code etc.

THANK YOU