Data Mining Project Report

Preetam Sarmah

DSBA

Contents

List of Figures

Glossary

PCA	Principal Component Analysis
EDA	Exploratory Data Analysis
CPM	Cost Per 1000 Impressions
CPC	Cost Per Click
CTR	Click Through Rate

References

PreetamSarmah_30-Apr-2023.ipynb

Part 1

1.1 Basic Analysis

There are 23066 entries . non duplicated entries and 19 attributes for each entry

There are 4736 null entries for each of CTR,CPM,CPC

Refer PreetamSarmah_30-Apr-2023.ipynb Part 1

1.2 Null Value Treatment

To treat the null values the below formulas are used

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

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

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

Refer PreetamSarmah_30-Apr-2023.ipynb Part 1

1.3 Outlier Treatment

The K-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values.

We treat the outliers using the Inter Quartile Distance or IQR

IQR = 75th Percentile- 25th Percentile

Min Value = 25th Percentile - 1.5 * IQR

Max Value = 75th Percentile +1.5 * IQR

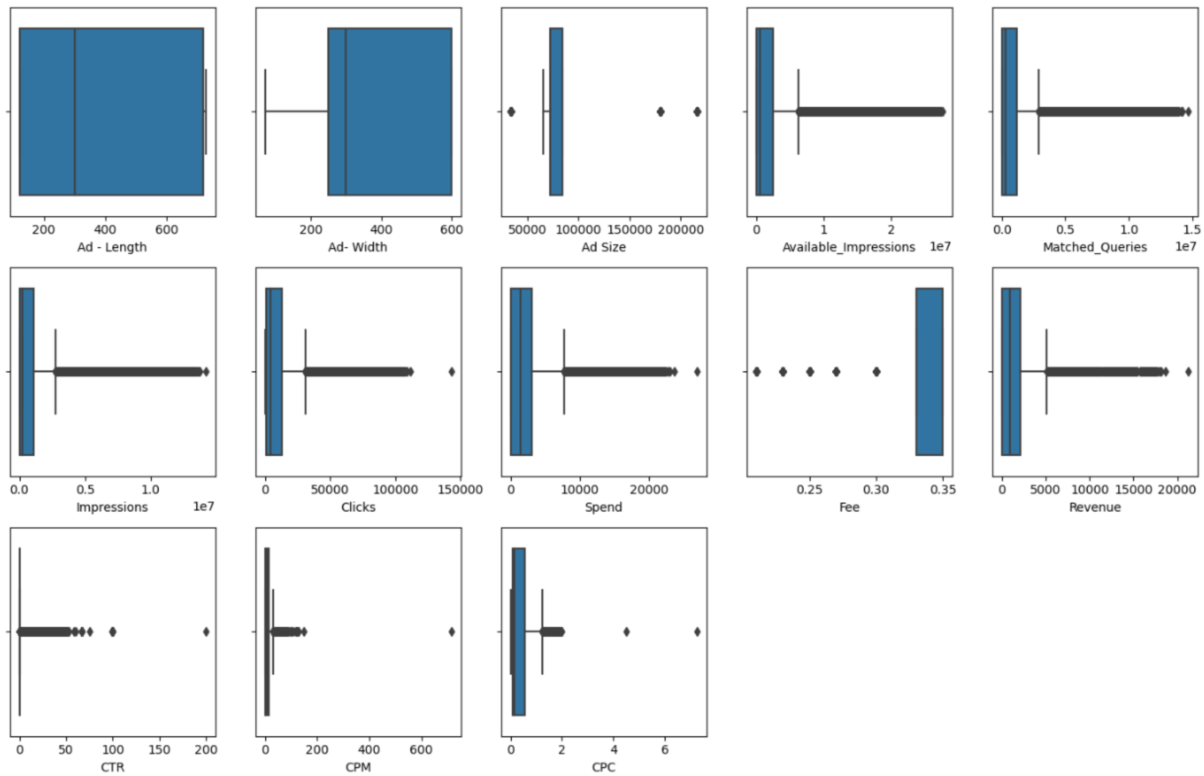

Fig 1.1 Before Outlier Treatment

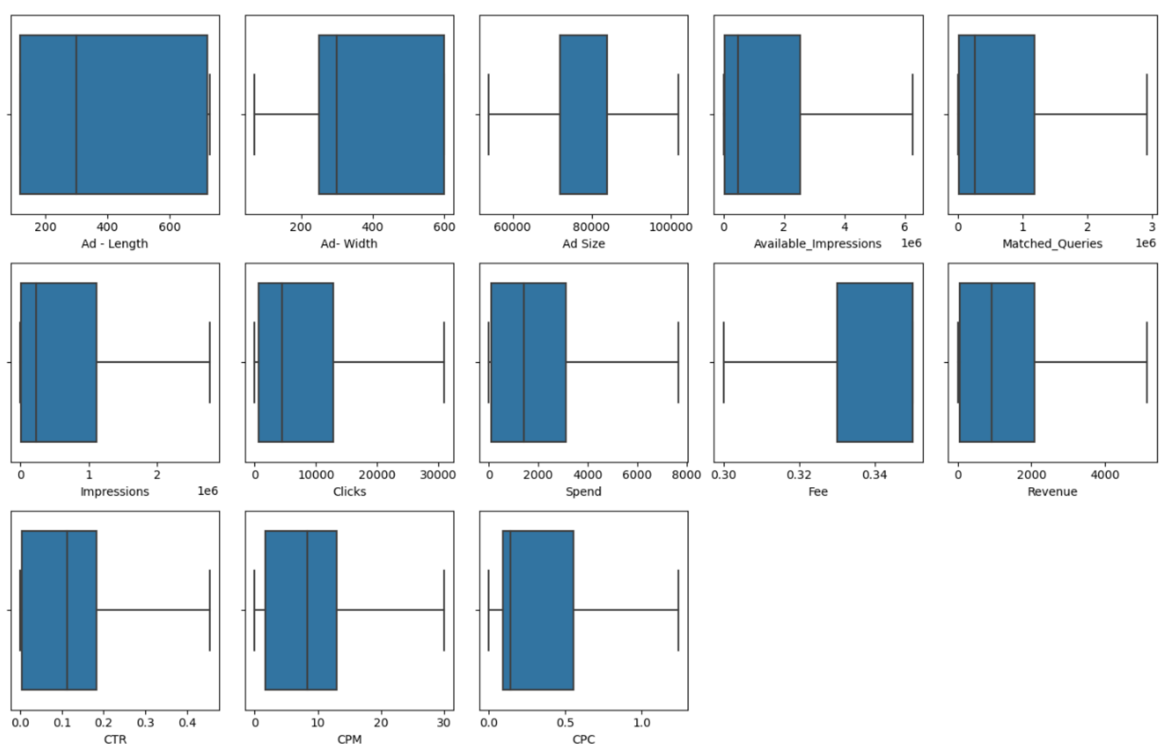

Fig 1.2 After Outlier Treatment

1.4 Z-Score Scaling

Z-Score Scaling is a variation of scaling that represents the number of standard deviations away from the mean. The distributions will have mean of 0 and standard deviation 1 (approximately)

$$\text{Z Score} = (x - \overline{x})/\sigma$$

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	-4.030447e-15	1.000022	-1.134891	-1.134891	-0.364496	1.433093	1.467332
Ad- Width	23066.0	5.390161e-15	1.000022	-1.319110	-0.432797	-0.186599	1.290590	1.290590
Ad Size	23066.0	-4.156304e-15	1.000022	-1.467840	-0.297564	-0.297564	0.482620	1.652896
Available_Impressions	23066.0	-3.617510e-15	1.000022	-0.756182	-0.740341	-0.528577	0.433059	2.193158
Matched_Queries	23066.0	1.341008e-15	1.000022	-0.779265	-0.761447	-0.527722	0.371498	2.070914
Impressions	23066.0	-1.224345e-15	1.000022	-0.768806	-0.760655	-0.538975	0.366051	2.056111
Clicks	23066.0	1.960656e-15	1.000022	-0.867488	-0.793438	-0.405431	0.468629	2.361729
Spend	23066.0	1.250852e-15	1.000022	-0.893170	-0.858046	-0.305523	0.393932	2.271900
Fee	23066.0	-2.322121e-14	1.000022	-2.222416	-0.567532	0.535724	0.535724	0.535724
Revenue	23066.0	3.136228e-15	1.000022	-0.880093	-0.846474	-0.317607	0.389803	2.244218
CTR	23066.0	-2.223858e-14	1.000022	-0.910603	-0.889261	-0.182714	0.277286	2.027108
CPM	23066.0	-6.707353e-16	1.000022	-1.194562	-0.940216	0.022045	0.700677	3.162016
CPC	23066.0	2.787153e-15	1.000022	-1.041140	-0.757396	-0.599760	0.692853	2.868227

Fig 1.3 Zscore Scaled Data

Scaling features prevents models from being biased towards features having a higher or lower magnitude, as the entire feature is described as number of standard deviations away from the mean, and as seen from the above description , the zscore scaled data set, the overall min,max values are a lot smaller compared to the original data , this speeds up the algorithm

1.5 Hierarchical Clustering and Dendrogram.

Hierarchical Clustering works well for classifying data, and its based on a distance.i.e nearest points of data are assigned to one cluster and subsequently nearest clusters are combined to form a bigger cluster based on linkage

Linkage is the distance between points in two different clusters.

The following dendrogram shows the hierarchical relationship between objects.

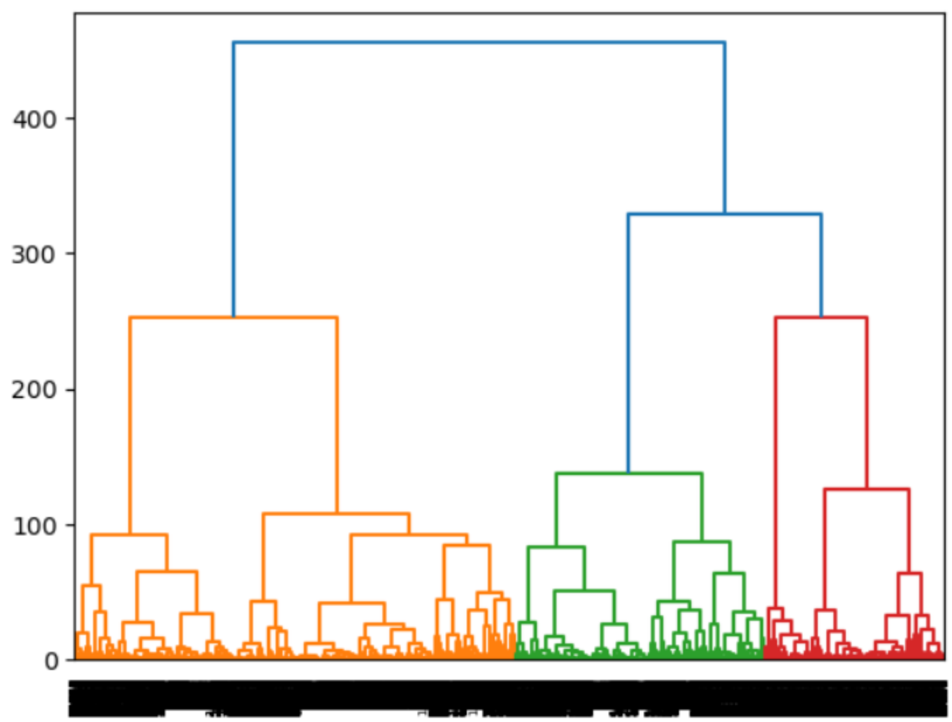

Fig 1.4 Hierarchical Clustering Dendrogram

As Hierarchical clustering is a top down approach , often the Dendrograms, tend to be boxy, so bellow is a truncated dendrodram, (which is a condensed dendrogram) of the last 10 merges.

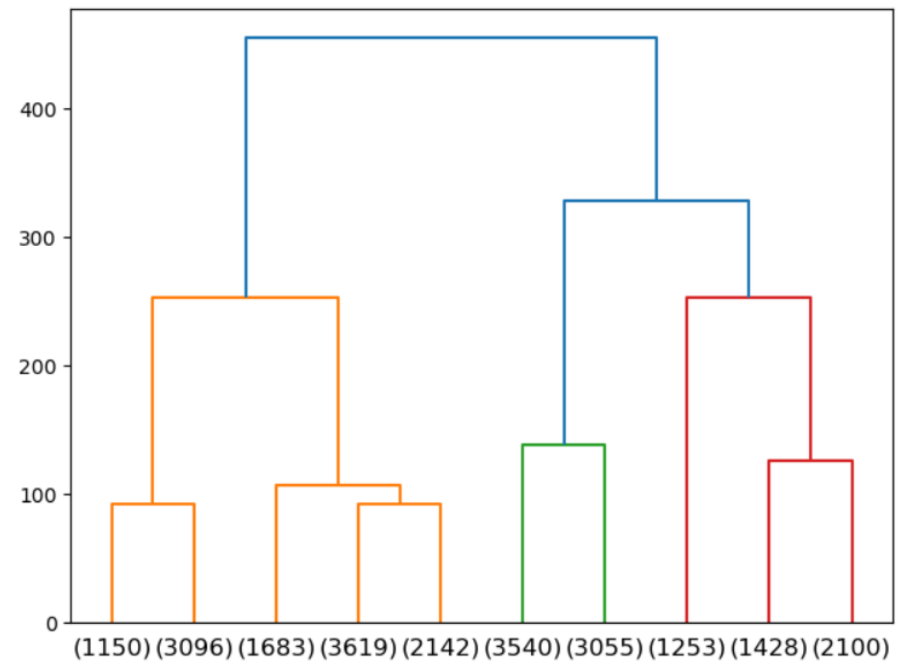

Fig 1.5 Truncated Hierarchical Clustering Dendrogram

Note the number in the () brackets represents the number of data points under each of those clusters

1.6 Elbow Plot and Optimum Number of Clusters

The elbow plot is a graphical representation of finding the optimal 'K' value in a K-means clustering. It works by finding WSS (Within Sum of Square) i.e. the sum of the square distance between points in a cluster and the cluster centroid.

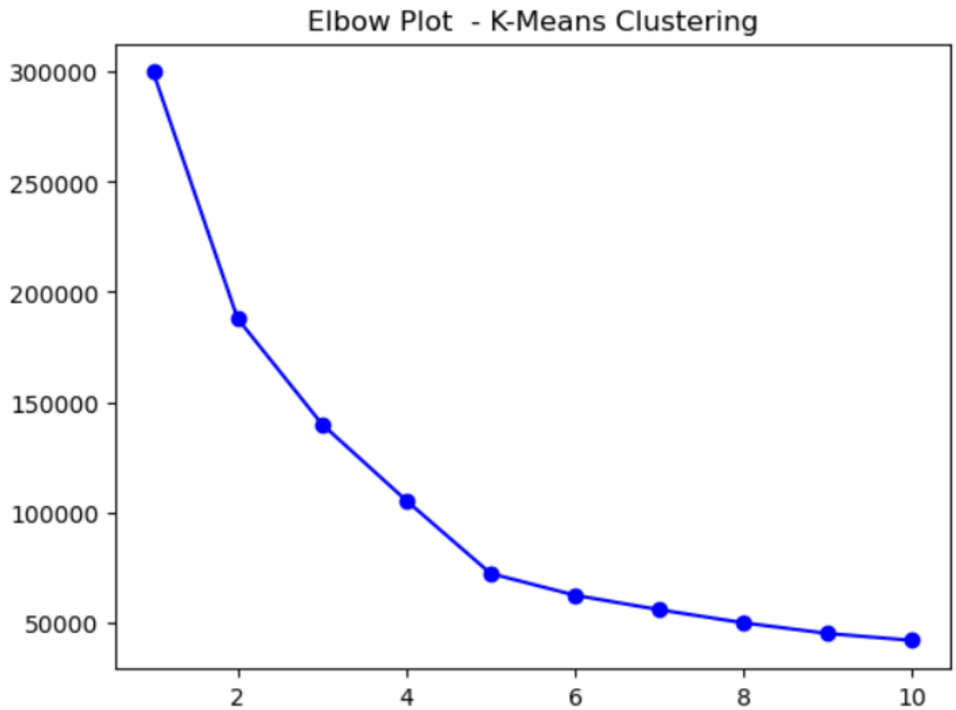

Fig 1.6 Elbow Plot - K-Means Clustering

From the above elbow, plot there is sharp drop from 1 cluster to 5 clusters, based on their respective within sum of squares value and after 5 clusters , even there is a drop but is slope is not that high , thus we conclude that 5 is the optimum number of clusters.

1.7 Silhouette Score and Optimum Number of Clusters

Silhouette Score is a metric to determine how well the clusters were formed, it ranges from -1 to 1. where 1 means , the clusters are well separated 0 means the clusters are not well separated and there is overlap whereas -1 indicates the clusters formed are incorrect

Silhouette Score = $b - a/max(b, a)$

Where,

a= average intra-cluster distance i.e the average distance between each point within a cluster.

b= average inter-cluster distance i.e the average distance between all clusters.

```
Number of Clusters =  2  and Silhouette Score =  0.40318725804432765
Number of Clusters =  3  and Silhouette Score =  0.3454647670915671 5
Number of Clusters =  4  and Silhouette Score =  0.403292158594085 5
Number of Clusters =  5  and Silhouette Score =  0.4802019193976827 5
Number of Clusters =  6  and Silhouette Score =  0.47613989974053916
Number of Clusters =  7  and Silhouette Score =  0.4688307485791759 5
Number of Clusters =  8  and Silhouette Score =  0.43228102175325334
Number of Clusters =  9  and Silhouette Score =  0.414126839182504 8
```

Fig 1.7 Clusters and their respective Silhouette Scores

From the values above its clear that when the number of clusters is 5, the silhouette score is the highest, hence the optimal number of clusters is 5

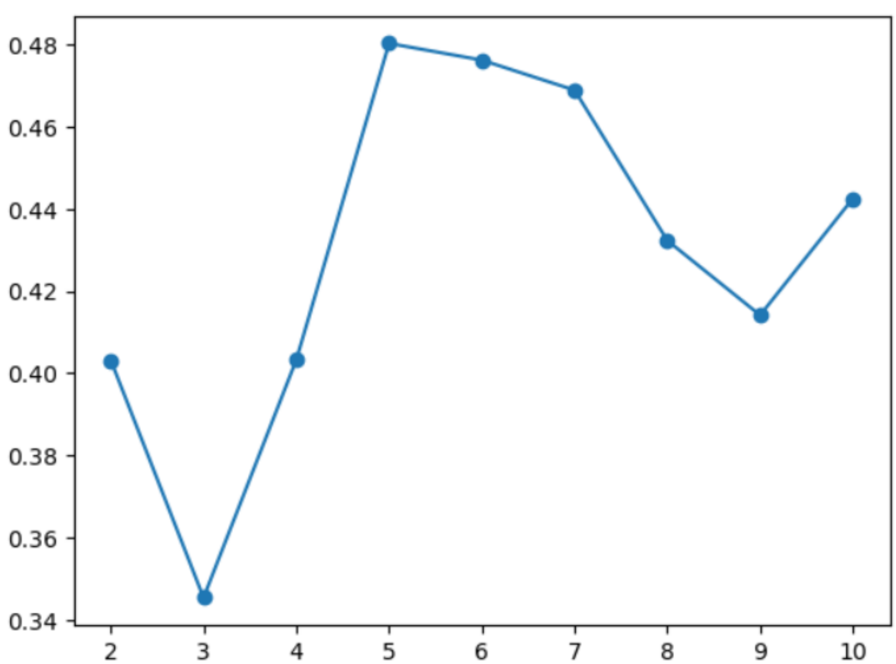

Fig 1.8 Number of Clusters vs Silhouette Score

1.8 Ad Profiling

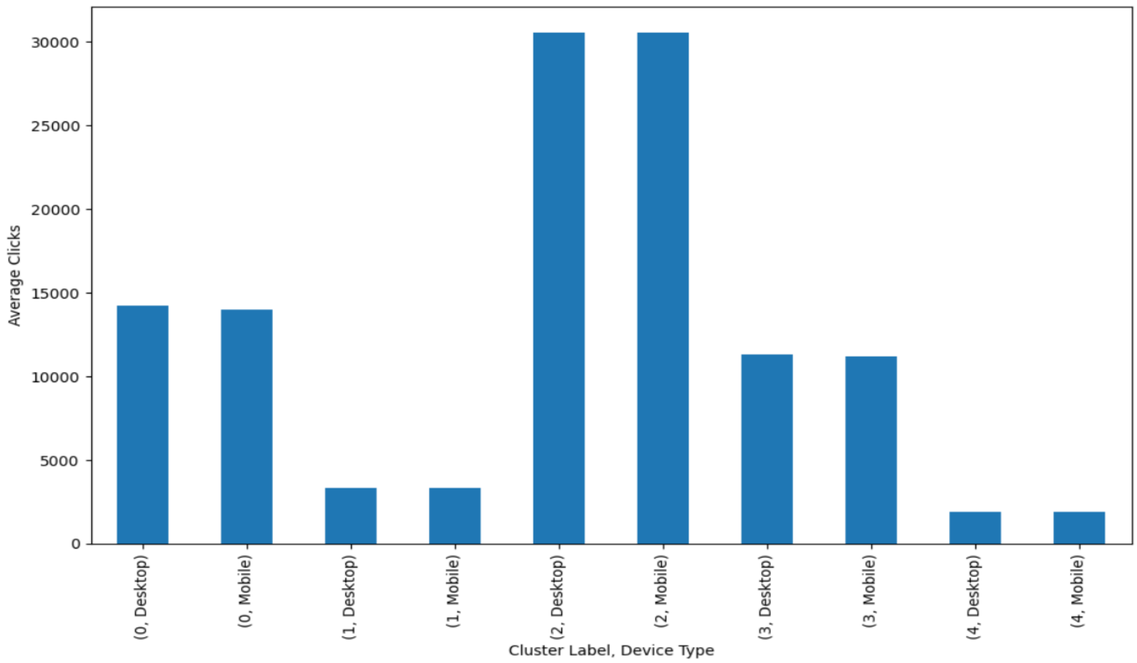

Fig 1.9 Average Clicks - grouped by Cluster Label, Device Type

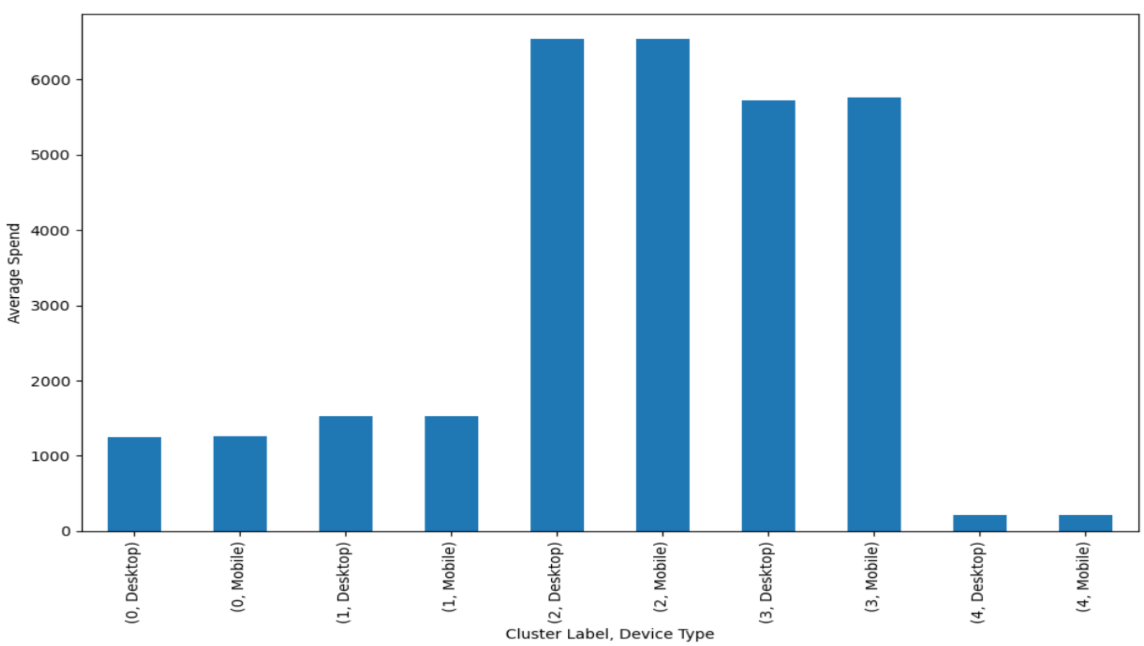

Fig 1.10 Average Spend - grouped by Cluster Label, Device Type

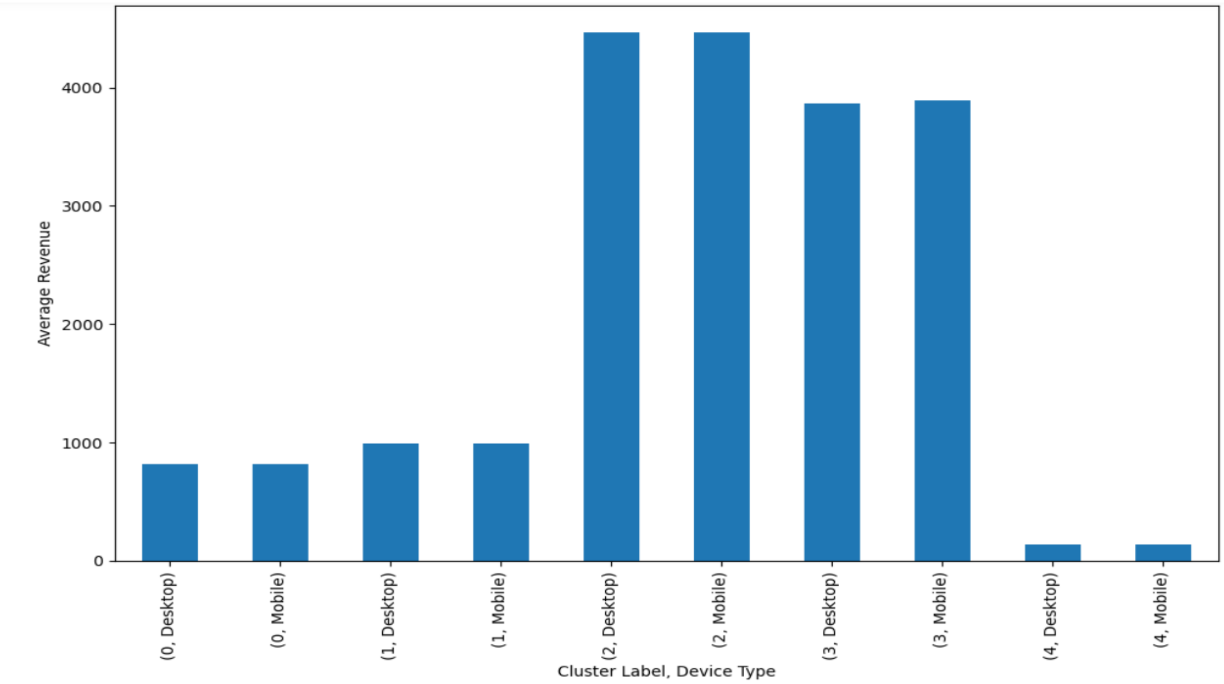

Fig 1.11 Average Revenue - grouped by Cluster Label, Device Type

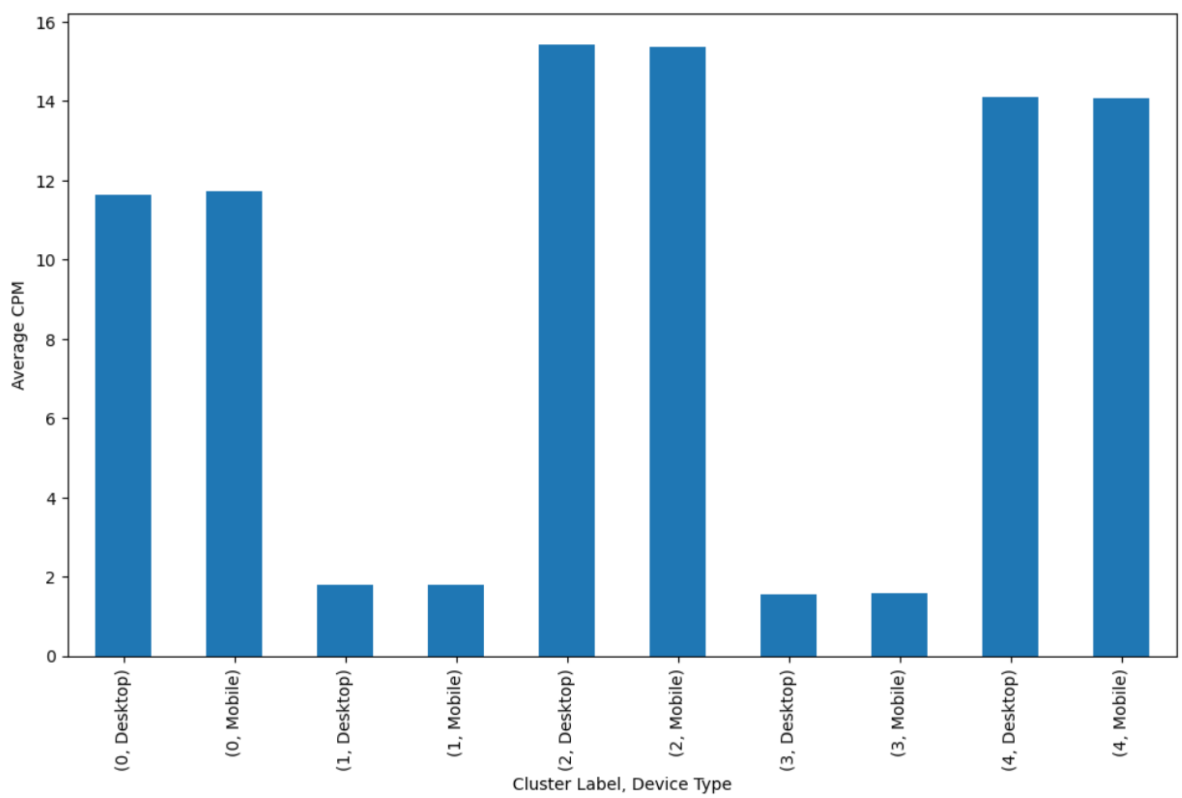

Fig 1.12 Average CPM - grouped by Cluster Label, Device Type

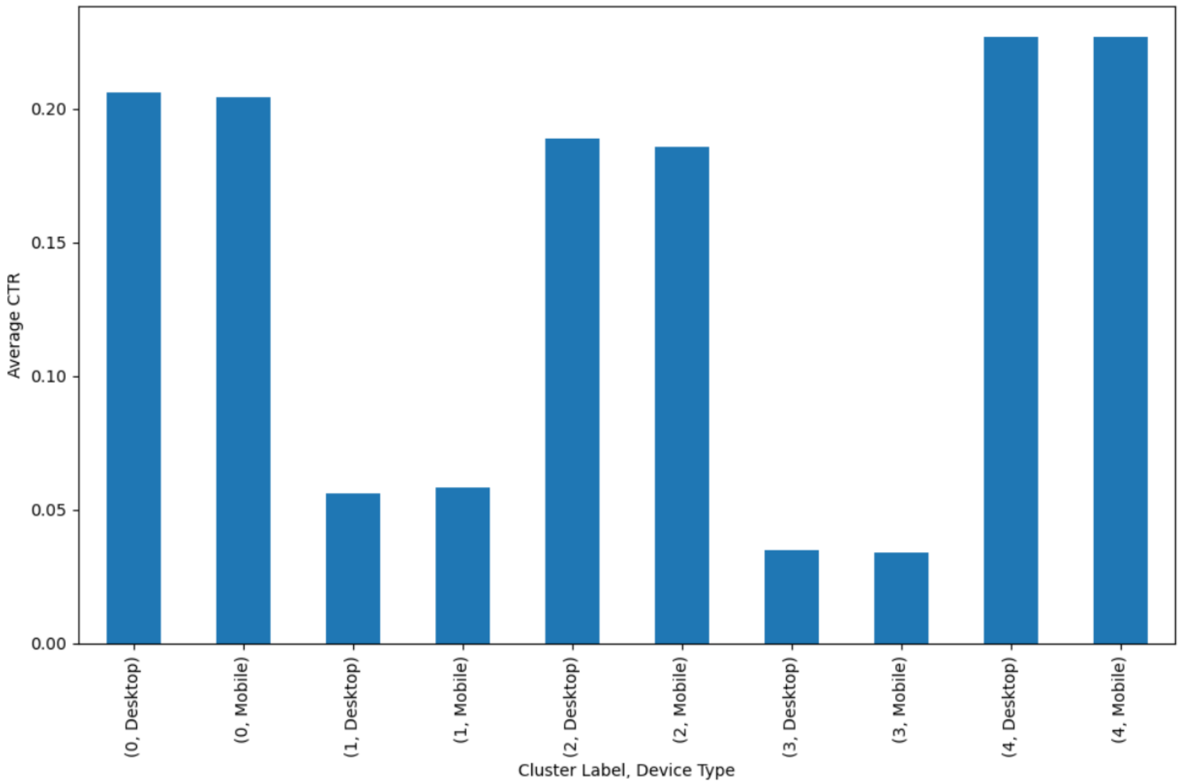

Fig 1.13 Average CTR - grouped by Cluster Label, Device Type

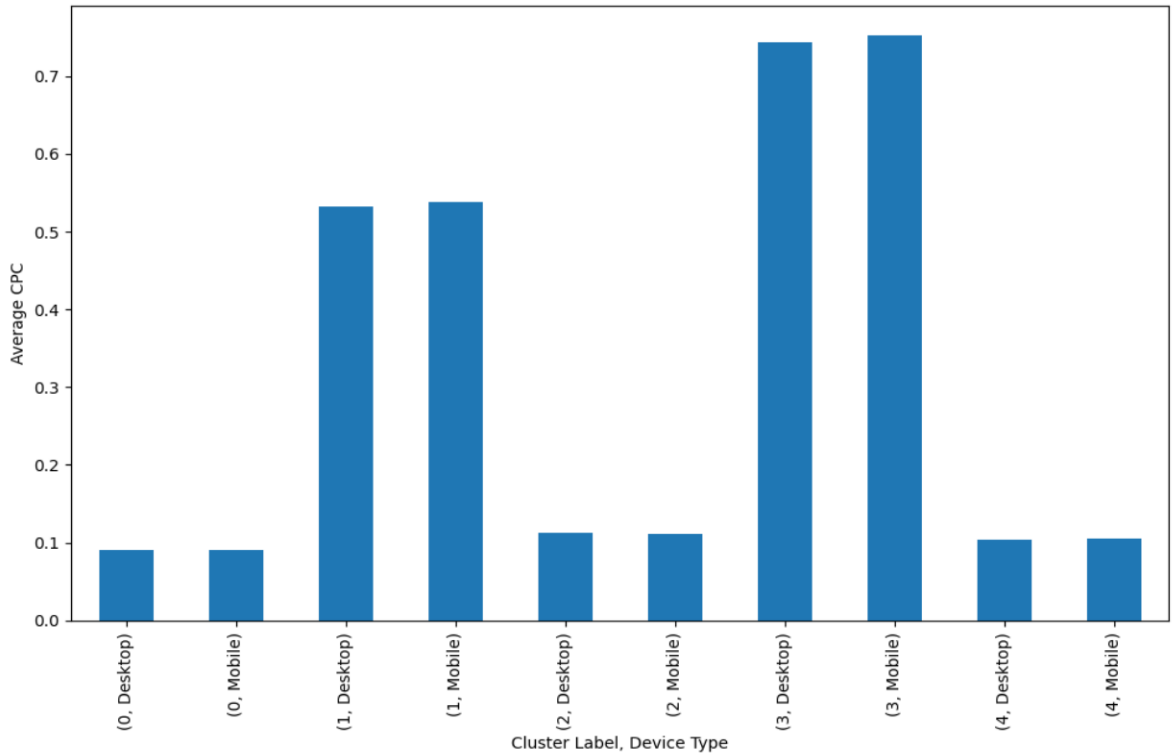

Fig 1.14 Average CPC - grouped by Cluster Label, Device Type

1.9 Conclusion.

- Cluster 3 has the highest Cost per Click , but also the lowest CTR and CPM ie the spend is high , but the impression of the ads and click through rate is worst
- Cluster 4 has the lowest Cost per Cick and the highest click through Rate , thats is the these ads are most likely to be clicked and the cost for these ads is low , ie the best performing ads
- Cluster 2 has the highest Cost per 1000 impressions , with a low cost per click and high click through rate , these the performing good
- Cluster 1 is moderate cost per click , low cost per 1000 impressions and second lowest cost per click
- Cluster 0 has the highest cost per click and moderate CTR and CPM

Part 2

2.1 Basic Analysis

There are 640 non null and non duplicated entries spread across 61 attributes.

Refer PreetamSarmah_30-Apr-2023.ipynb Part 2

2.2 Exploratory analysis

2.2.1 Which state has the highest gender ratio and which state has the lowest?

Lakshadweep has the highest Male to Female gender ratio , where as Andhra Pradesh has the lowest Male to Female gender ratio

2.2.2 Which district has the highest gender ratio and which district has the lowest?

Lakshadweep district of Lakshadweep has the highest Male to Female gender ratio , where as Krishna district of Andhra Pradesh has the lowest Male to Female gender ratio

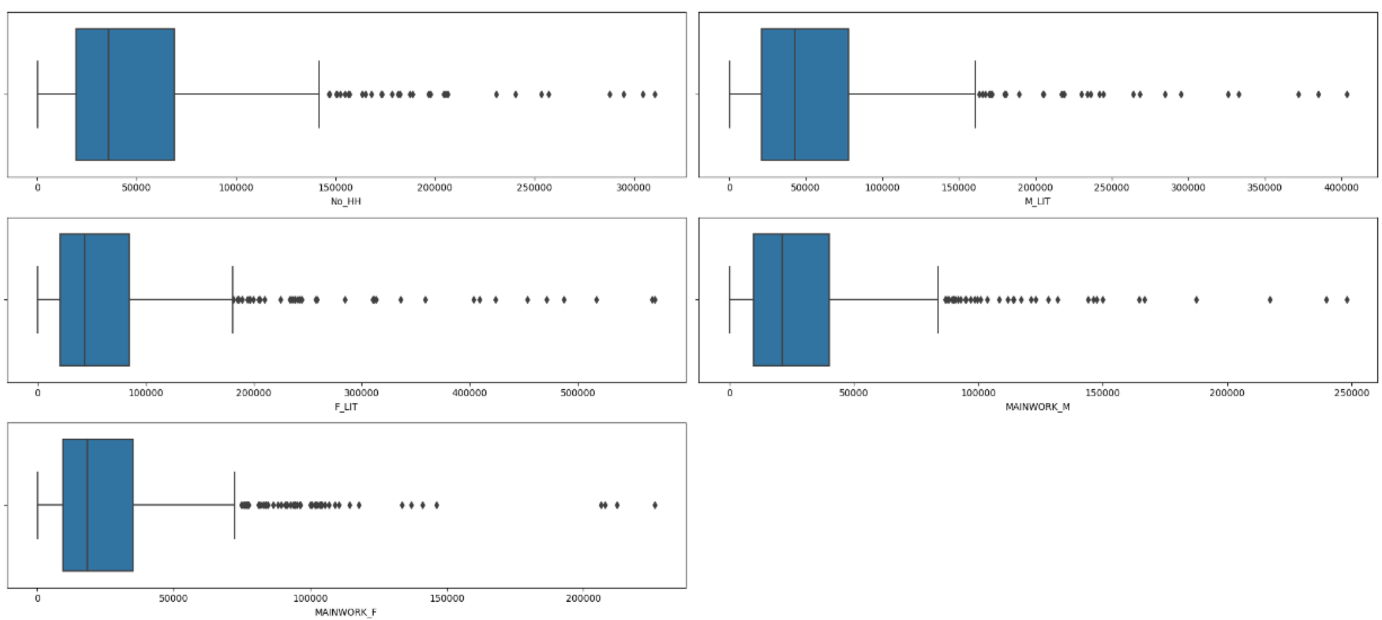

Fig 2.1 : EDA Data

- From the above we can conclude that on average the Number of Females literate are more than the number of Literate Males, but as the average male population is significantly less than the average female population across states the average Male Literacy rate is significantly higher than the average Female Literacy rate.

- On average there are around 51000 households
- The main worker population , on average is significantly male

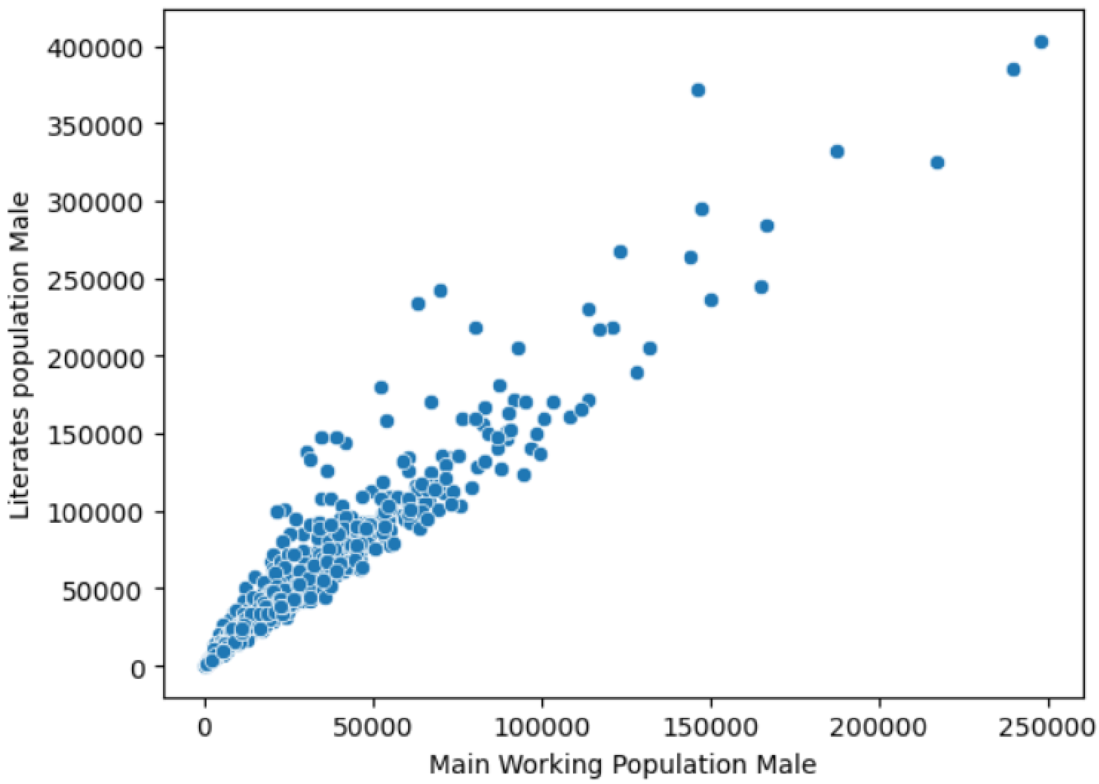

Fig 2.2 Literate Male Population vs Main Working Population Male

We can conclude from the above plot that the Literacy population of Males has a high correlation with the Main Working population Male

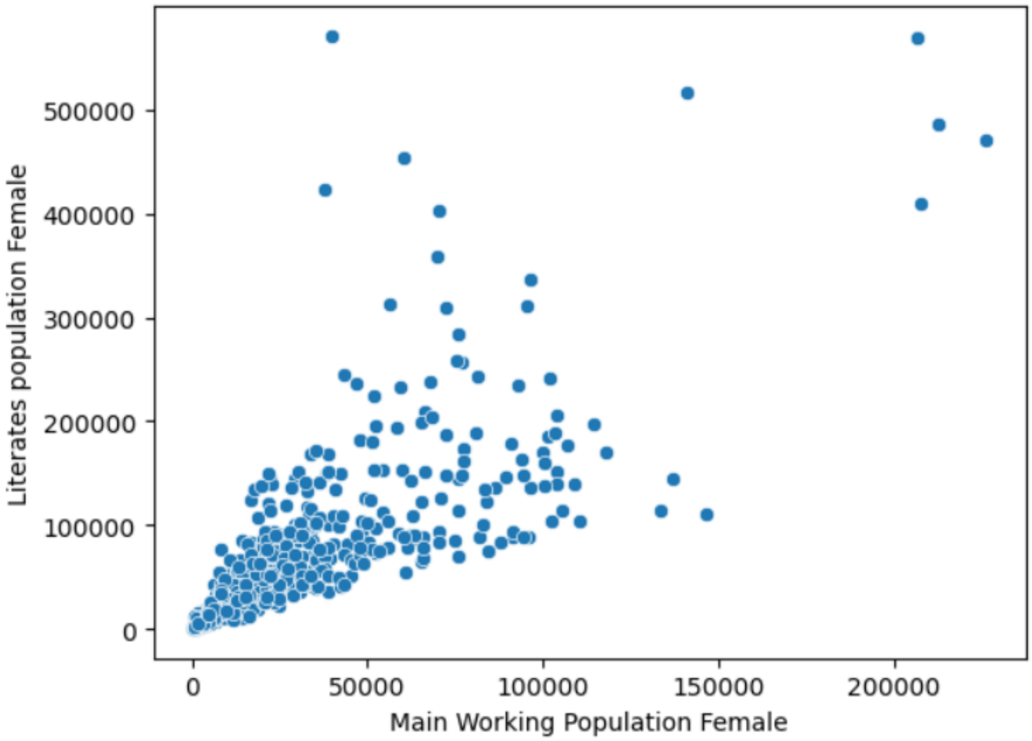

Fig 2.3 Literate Female Population vs Main Working Population Female

We can conclude from the above plot that the Literacy population of Females has some correlation with the Main Working population Female , but the correlation is not as strong compared to Literate Male Population vs Main Working Population Male

2.3 Outlier Treatment

As this is census data of a population , this data is likely to follow a normal distribution and often with such populations , we are bound to get outliers but these outliers are genuine outliers.

Treating outliers is not always necessary , as in census data the outliers are critical to the results and also there seems to be large percentage of outliers

2.4 Z-Score Scaling

Z-Score scaling does'nt affect outliers as by using Z-Score we move the data points to certain standard deviations away from the mean , so the original distance just gets scaled. Hence there are no affect on the outliers.

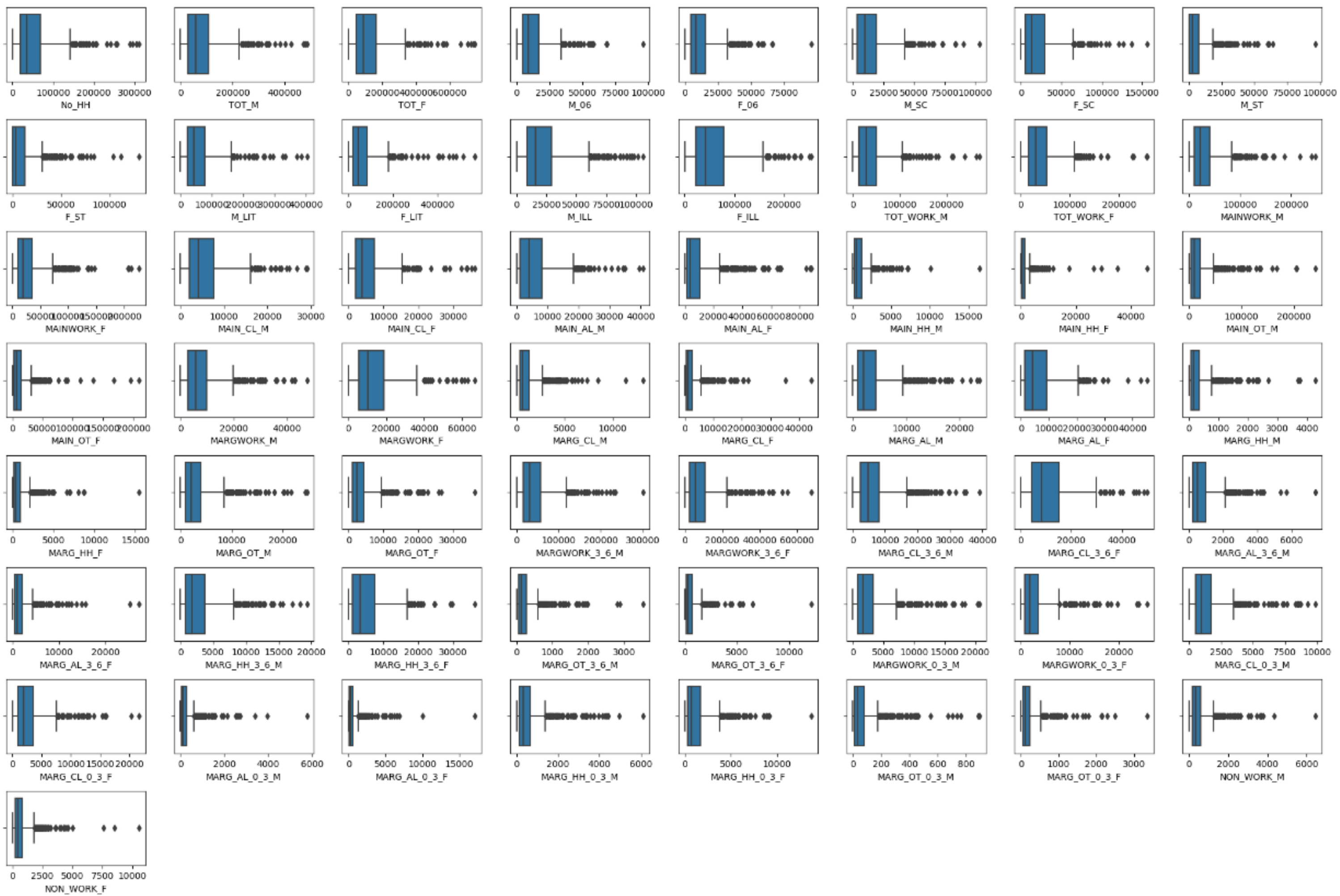

Fig 2.4 Before Z-Score Scaling

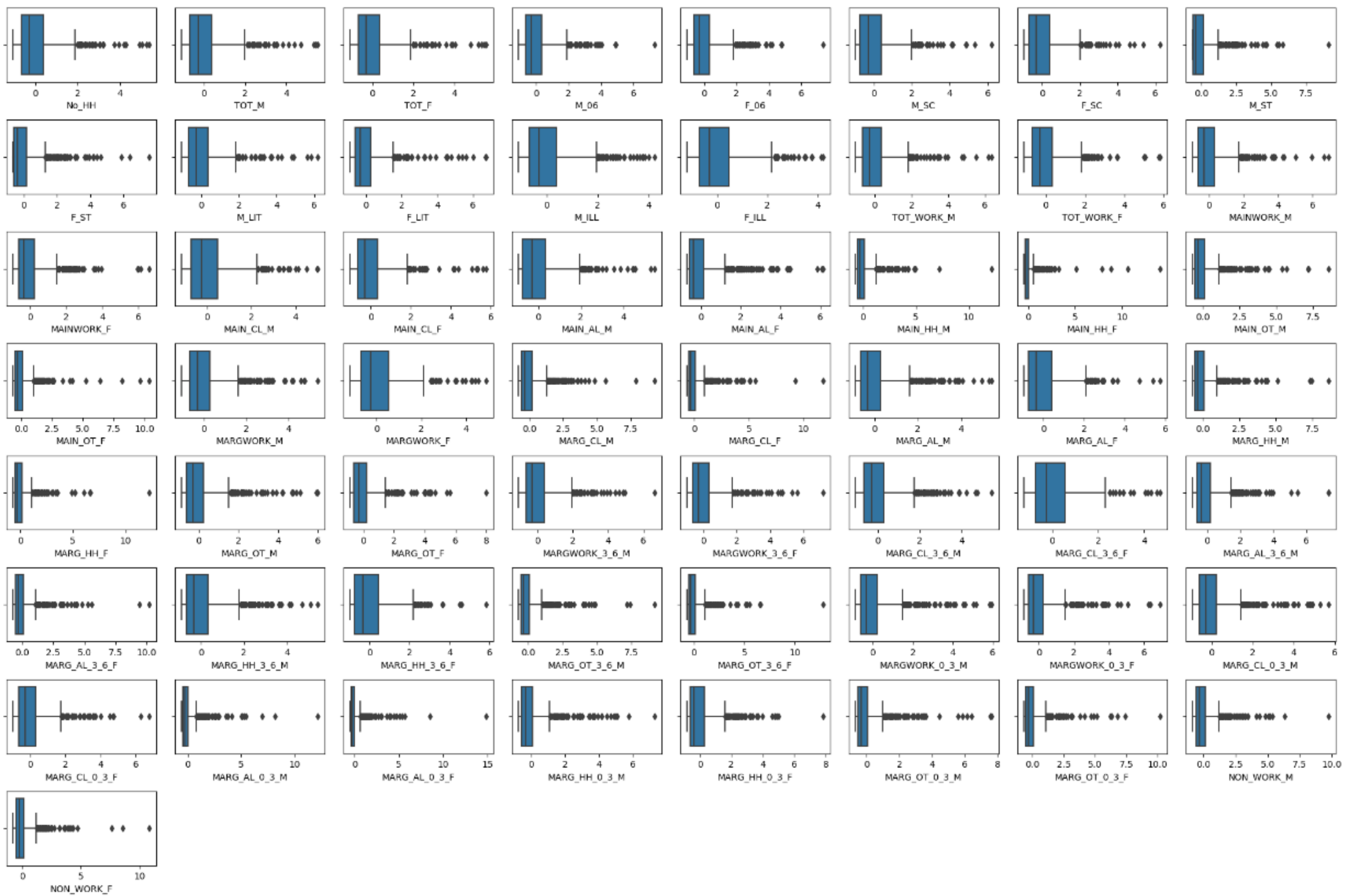

Fig 2.5 After Z-Score Scaling

As its pretty evident from the above plots , the outliers are'nt affected by Z-score scaling

2.5 PCA - Covariance Matrix, Eigen Vectors and Eigen Values

The below is the covariance matrix for first 10 Principal Components

	0	1	2	3	4	5	6	7	8	9
0	-4.617263	0.138116	0.328545	1.543697	0.353737	-0.420947	-0.010393	0.479105	0.049653	-0.035607
1	-4.771662	-0.105865	0.244449	1.963215	-0.153884	0.417310	-0.023119	-0.006797	0.424390	-0.190761
2	-5.964836	-0.294347	0.367393	0.619543	0.478199	0.276580	0.069554	0.040713	0.162092	0.013163
3	-6.280796	-0.500384	0.212701	1.074516	0.300799	0.051158	-0.250541	0.084362	0.150616	0.123880
4	-4.478566	0.894154	1.078277	0.535556	0.804065	0.341677	-0.092331	0.376964	-0.067445	0.196151

Fig 2.6 Covariance Matrix - 57 PCs

The below is the Eigen Vectors for a few Principal Components

```
array([[ 0.15602058,  0.16711763,  0.16555318, ...,  0.13219224,
         0.15037558,  0.1310662 ],
       [-0.12634653, -0.08967655, -0.10491237, ...,  0.05081332,
        -0.06536455, -0.07384742],
       [-0.00269025,  0.05669762,  0.03874947, ..., -0.07871987,
         0.11182732,  0.1025525 ],
       ...,
       [ 0.        ,  0.2077636 ,  0.24647657, ..., -0.07217993,
         0.00399206, -0.06929081],
       [ 0.        ,  0.2887035 , -0.20596721, ...,  0.04019745,
        -0.03192722,  0.00778048],
       [-0.        ,  0.18790022,  0.02642675, ..., -0.02597314,
        -0.13972835, -0.02147533]])
```

Fig 2.7 Eigen Vectors - 57 PCs

The Following is the list of all the Eigen Values of all 57 Principal Components.

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

Fig 2.8 Eigen Values - 57 PCs

2.6 Scree Plot - Optimum Number of Principal Components

Constructing the cumulative variance ratio, we note the it crosses 90% explained variance on the 6 th Principal Components , so the optimum number of Principal Components is 6

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

Fig 2.9 Cummulative Variance Ratio - 57 PCs

This is further demonstrated by the Scree Plot below

Scree plot tells us how much variation each PC captures

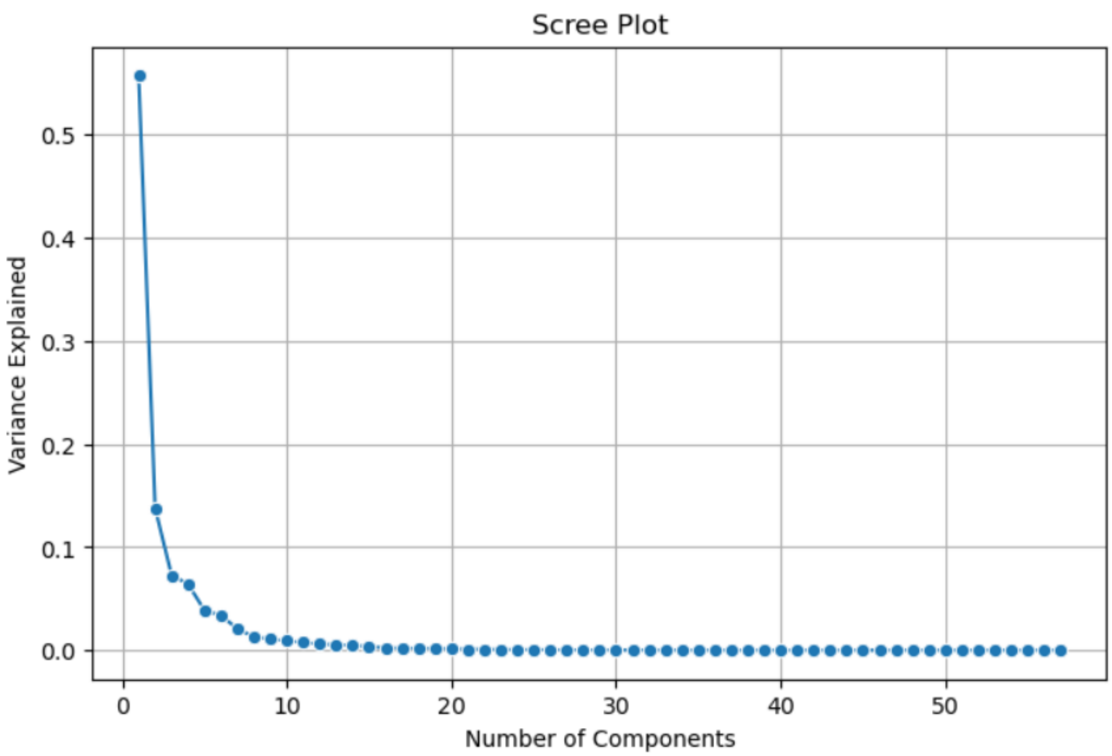

Fig 2.10 Scree Plot

2.7 Original features influence on various PCs

2.8 Linear equation for first PC

The following is the linear equation of the first PC.

```
( 0.16 ) * No_HH + ( 0.17 ) * TOT_M + ( 0.17 ) * TOT_F + ( 0.16 ) * M_06 + ( 0.16 ) * F_06 + ( 0.15 ) * M_SC + ( 0.15 ) * F_SC
+ ( 0.03 ) * M_ST + ( 0.03 ) * F_ST + ( 0.16 ) * M_LIT + ( 0.15 ) * F_LIT + ( 0.16 ) * M_ILL + ( 0.17 ) * F_ILL + ( 0.16 ) * TO
T_WORK_M + ( 0.15 ) * TOT_WORK_F + ( 0.15 ) * MAINWORK_M + ( 0.12 ) * MAINWORK_F + ( 0.1 ) * MAIN_CL_M + ( 0.07 ) * MAIN_CL_F +
( 0.11 ) * MAIN_AL_M + ( 0.07 ) * MAIN_AL_F + ( 0.13 ) * MAIN_HH_M + ( 0.08 ) * MAIN_HH_F + ( 0.12 ) * MAIN_OT_M + ( 0.11 ) * M
AIN_OT_F + ( 0.16 ) * MARGWORK_M + ( 0.16 ) * MARGWORK_F + ( 0.08 ) * MARG_CL_M + ( 0.05 ) * MARG_CL_F + ( 0.13 ) * MARG_AL_M +
( 0.11 ) * MARG_AL_F + ( 0.14 ) * MARG_HH_M + ( 0.13 ) * MARG_HH_F + ( 0.16 ) * MARG_OT_M + ( 0.15 ) * MARG_OT_F + ( 0.16 ) * M
ARGWORK_3_6_M + ( 0.16 ) * MARGWORK_3_6_F + ( 0.17 ) * MARG_CL_3_6_M + ( 0.16 ) * MARG_CL_3_6_F + ( 0.09 ) * MARG_AL_3_6_M + (
0.05 ) * MARG_AL_3_6_F + ( 0.13 ) * MARG_HH_3_6_M + ( 0.11 ) * MARG_HH_3_6_F + ( 0.14 ) * MARG_OT_3_6_M + ( 0.12 ) * MARG_OT_3_
6_F + ( 0.15 ) * MARGWORK_0_3_M + ( 0.15 ) * MARGWORK_0_3_F + ( 0.15 ) * MARG_CL_0_3_M + ( 0.14 ) * MARG_CL_0_3_F + ( 0.05 ) *
MARG_AL_0_3_M + ( 0.04 ) * MARG_AL_0_3_F + ( 0.12 ) * MARG_HH_0_3_M + ( 0.12 ) * MARG_HH_0_3_F + ( 0.14 ) * MARG_OT_0_3_M + (
0.13 ) * MARG_OT_0_3_F + ( 0.15 ) * NON_WORK_M + ( 0.13 ) * NON_WORK_F +
```