BUSINESS DATA

ANALYSIS REPORT

Table of Contents

Problem1: Clustering 5

1.Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc. 5

2.Treat missing values in CPC, CTR and CPM using the formula given. You have to basically create a user defined function and then call the function for imputing. 9

3.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). 10

4.Perform z-score scaling and discuss how it affects the speed of the algorithm. 12

5.Perform clustering and do the following: 13

Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance. 13

6.Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm. 15

7.Print silhouette scores for up to 10 clusters and identify optimum number of clusters. 15

8.Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots.] 16

9. Clustering: Conclude the project by providing summary of your learnings. 19

Problem2 :PCA 19

1.Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc. 20

2.Perform detailed Exploratory analysis by creating certain questions like (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio? (Example Questions). Pick 5 variables out of the given 24 variables below for EDA: No_HH, TOT_M, TOT_F, M_06, F_06, M_SC, F_SC, M_ST, F_ST, M_LIT, F_LIT, M_ILL, F_ILL, TOT_WORK_M, TOT_WORK_F, MAINWORK_M, MAINWORK_F, MAIN_CL_M, MAIN_CL_F, MAIN_AL_M, MAIN_AL_F, MAIN_HH_M, MAIN_HH_F, MAIN_OT_M, MAIN_OT_F 23

3.We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary? 30

4.Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment. 30

5.Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector. 32

6.Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot. 35

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. 37

8.PCA: Write linear equation for first PC. 42

List of Figures

Figure 1: Clustering Clean Ads Dataset 5

Figure 2:Top 5 rows in the dataset 6
Figure 3:Bottom 5 rows in the dataset 6
Figure 4:Count plot of Platform, Device Type and Format 7
Figure 5: Count plot of Inventory and Ad type 7
Figure 6: Plotting Boxplot of Numerical Variables 8
Figure 7:Dataset after imputing the Null values for CTR,CPM and CPC 10
Figure 8:Boxplot of Numerical Variables 11
Figure 9: Boxplot after treating Outliers 12
Figure 10: Z score scaling of the Numerical Variables 13
Figure 11: Dendrogram using metric='Euclidean' and method='ward' with all the Clusters 14
Figure 12:Dendrogram using method='ward' with all the Clusters 14
Figure 13: Dendrogram using metric='Euclidean' and method='ward' last 10 Clusters 14
Figure 14: WSS Scores 15
Figure 15:Elbow Plot of WSS 15
Figure 16:Silhouette Scores. 16
Figure 17 Silhouette Plot. 16
Figure 18: Dataset with Clusters 17
Figure 19 :Graphical Representation of Cluster wise trends in clicks, spend, revenue, CPM, CTR, & CPC based on Device Type 18
Figure 20: Countplot of Device Types and Clusters 18
Figure 21:Top 5 observations of PCA India Data Census 20
Figure 22: Bottom 5 observations of PCA India Data Census 20
Figure 23: Distribution Plot and Boxplot of Number of Literate Males 24
Figure 24: Distribution Plot and Boxplot of Total Number of Males 24
Figure 25: Distribution Plot and Boxplot of Total Number of Females 24
Figure 26: Distribution Plot and Boxplot of Total Number of Male Workers 24
Figure 27: Distribution Plot and Boxplot of Total Number of Female Workers 24
Figure 28: Bar plot of Literate males in each state 25
Figure 29: bagplot of Total Males in each state 26
Figure 30: bar plot of Total Females in each state 26
Figure 31: bar plot of Total Working Males in each state 26
Figure 32: bar plot of Total Working Females in each state 27
Figure 33:Scatterplot of Numerical Variables 27
Figure 34: bar plot of Gender Ratio is Each State 28
Figure 35:District wise Ratio of Gender 28
Figure 36:State wise Gender Ratio of Male to Female Working Population 28
Figure 37: State wise ratio of literate Male to Total Male Population 29
Figure 38: State wise ratio of Working Male to Total Male Population 29
Figure 39: : State wise ratio of Working Female to Total Female Population 30
Figure 40 :Boxplot of Variables before scaling 31
Figure 41: Boxplot of Variables after scaling 31
Figure 42: Scaled Data 32
Figure 43:Correlation heatmap between 57 variables 34
Figure 44:Scree Plot with 57 Principal Components 35
Figure 45 PCA Variance with 10 components 36
Figure 46 PCA Variance Ratio with 10 components 36
Figure 47:Cumulative Variance of the Components 36
Figure 48: Scree Plot with 10 Principal Components 36
Figure 49: PCA Loading of PC1 Figure 50: PCA Loading of PC2 39
Figure 51: PCA Loading of PC3 Figure 52: PCA Loading of PC4 39
Figure 53: PCA Loading of PC5 and PC6 39

Figure 54:Correlation map of Principal Components and Actual Columns 40
Figure 55:Heatmap of PC's 42

List of Tables

Table 1: Clean Ads Dataset Columns Information and Data Types 6
Table 2:Data Summary of Categorical Variables 7
Table 3: Data Summary of Numerical Variables 8
Table 4: Count of Null Values 9
Table 5: Count of Null Values in CTR, CPC and CPM 9
Table 6: Data Summary Before and After Imputing Null Values 10
Table 7:Data Summary after Treating Outliers 12
Table 8: Data Summary after scaling using Z score method 13
Table 9: Grouping data by clusters with mean to identify trends in clicks, spend, revenue, CPM, CTR, & CPC based on Device Type 17
Table 10: Dataset Column Information1 21
Table 11: Dataset Column Information2 21
Table 12: Describing the Numerical Columns 22
Table 13: : Describing the Categorical Columns 22
Table 14: Null Values in the Dataset 23
Table 15: Data Description of the 5 variables 25
Table 16Snapshot of the Covariance Matrix of all the numerical Variables 33
Table 17:Eigen Vectors 34
Table 18:Eigen Values 35
Table 19 Variances and Standard Deviations of each PC 37
Table 20 PC scores 41
Table 21:Correlation between PC's 41
Table 22: linear equation for first PC. 43

Problem1: Clustering

Digital Ads Data:

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

The following three features are commonly used in digital marketing:

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

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

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

The Data Dictionary and the detailed description of the formulas for CPM, CPC and CTR are given in the sheet 2 of the Clustering Clean ads_data Excel File.

Perform the following in given order:

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

Ans: Load the sample data for number of observations and columns

	Timestamp	InventoryType	Ad - Length	Ad-Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.00	0.35	0.0000	0.309598	0.0	0.0
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.00	0.35	0.0000	0.350877	0.0	0.0
2	2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.00	0.35	0.0000	0.281690	0.0	0.0
3	2020-9-3-20	Format1	300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.00	0.35	0.0000	0.202020	0.0	0.0
4	2020-9-4-15	Format1	300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.00	0.35	0.0000	0.413223	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
23061	2020-9-13-7	Format5	720	300	216000	Inter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
23062	2020-11-2-7	Format5	720	300	216000	Inter224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.0260	NaN	NaN	NaN
23063	2020-9-14-22	Format5	720	300	216000	Inter218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	NaN	NaN	NaN
23064	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
23065	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

23066 rows × 19 columns

Figure 1: Clustering Clean Ads Dataset

As can be seen there are 23066 rows and 19 columns

Note: In the Original dataset , the Column CTR was expressed as % and not as per formula for CTR column. Hence Column in the excel has been modified as per the CTR formula mentioned before loading the dataset.

Let's see the top 5 and bottom 5 records

	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

Figure 2:Top 5 rows in the dataset

	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

Figure 3:Bottom 5 rows in the dataset

Let's see the columns description and their Data Types

```
RangeIndex: 23066 entries, 0 to 23065
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)
```

Table 1: Clean Ads Dataset Columns Information and Data Types

There are 13 Numeric type Columns of which are 6 are Float Type and 7 are Int type and 6 object type columns. Time Stamp column is marked as object type column though its Date Type Column, but timestamp column is not of any relevance to the clustering ,so we will not change the column Type.

Let's see the data summary of the Categorical variables

	InventoryType	Ad Type	Platform	Device Type	Format
count	23066	23066	23066	23066	23066
unique	7	14	3	2	2
top	Format4	Inter224	Video	Mobile	Video
freq	7165	1658	9873	14806	11552

Table 2:Data Summary of Categorical Variables

Let's see some graphical analysis on categorical variables

Univariate Analysis of Categorical Variables

Let's see the graphical representation of categorical variables

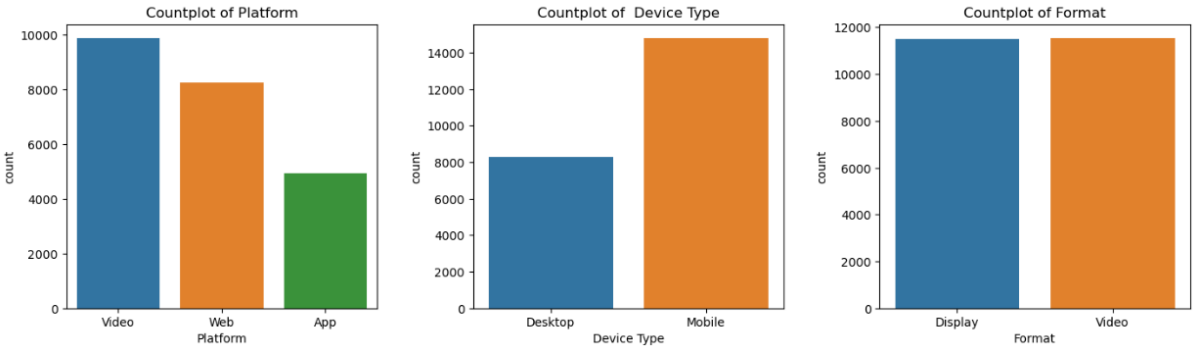


Figure 4:Count plot of Platform, Device Type and Format

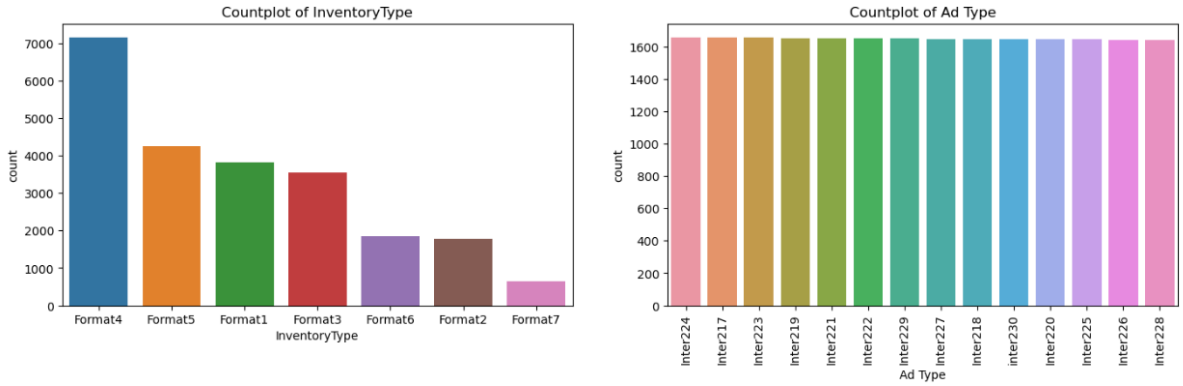


Figure 5: Count plot of Inventory and Ad type

Inferences from the above graphical representation:

1.There are 3 types of Platforms namely Video, Web and App. Video is the most used form of Platform for Ads and App is least used form for Ads.

2.There are 2 device types that support these Ads ,they are Mobile and Desktop. Mobile is most used Device for Ads .

3.Display and Video are 2 formats in which Ads are displayed . Video format of Ads are slightly more than Display.

4.There are 7 types of Inventories namely Format1,Format2,Format3,Format4,Format5,Format6,Format7.Format 4 is most used Inventory Type for Ads and Format 7 is least used .

5.There are 14 Ad types and Inter224 is the most used Ad Type.

Let's see the data summary of the numerical variables:

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
count	23066.000000	23066.000000	23066.000000	2.306600e+04	2.306600e+04	2.306600e+04	23066.000000	23066.000000	23066.000000	23066.000000	18330.000000	18330.000000	18330.000000
mean	385.163097	337.896037	96674.468048	2.432044e+06	1.295099e+06	1.241520e+06	10678.518816	2706.625689	0.335123	1924.252331	0.073661	7.672045	0.351061
std	233.651434	203.092885	61538.329557	4.742888e+06	2.512970e+06	2.429400e+06	17353.409363	4067.927273	0.031963	3105.238410	0.075160	6.481391	0.343334
min	120.000000	70.000000	33600.000000	1.000000e+00	1.000000e+00	1.000000e+00	1.000000	0.000000	0.210000	0.000000	0.000100	0.000000	0.000000
25%	120.000000	250.000000	72000.000000	3.367225e+04	1.828250e+04	7.990500e+03	710.000000	85.180000	0.330000	55.365375	0.002600	1.710000	0.090000
50%	300.000000	300.000000	72000.000000	4.837710e+05	2.580875e+05	2.252900e+05	4425.000000	1425.125000	0.350000	926.335000	0.082550	7.660000	0.160000
75%	720.000000	600.000000	84000.000000	2.527712e+06	1.180700e+06	1.112428e+06	12793.750000	3121.400000	0.350000	2091.338150	0.130000	12.510000	0.570000
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	81.560000	7.260000

Table 3: Data Summary of Numerical Variables

Univariate Analysis of Numerical Variables

Let's see the graphical representation of numerical variables

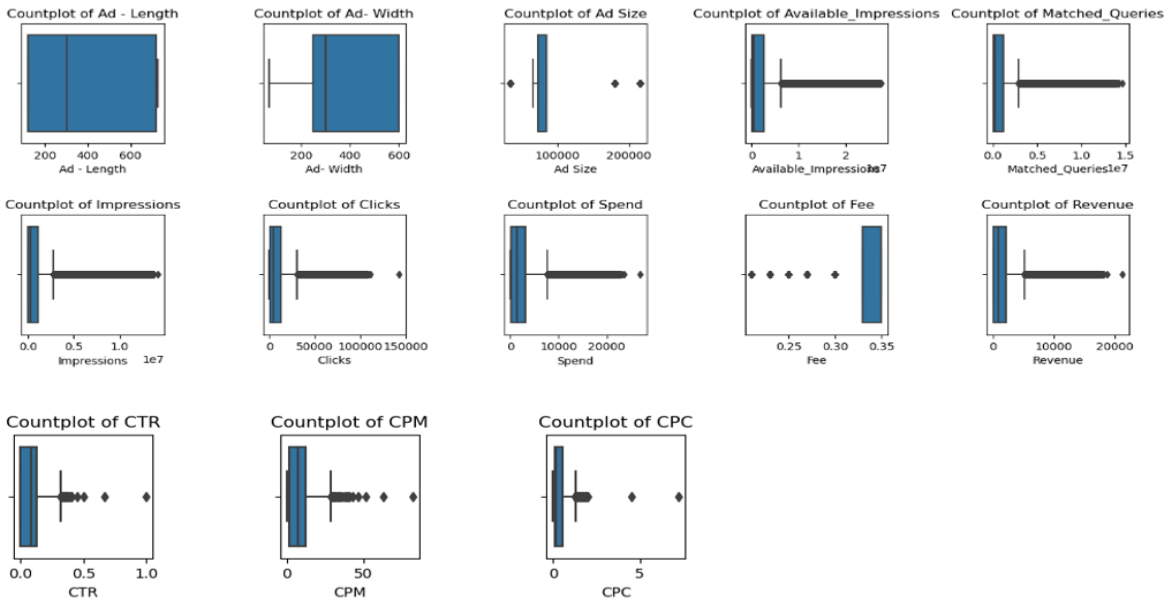


Figure 6: Plotting Boxplot of Numerical Variables

Inferences from the above graphical representation:

1.Except for Ad-width and Ad-Length, all other numerical variables have outliers .

2.Most of the variables are right skewed .Fee Variable is left skewed.

3.Range of values of each of the variables is significantly different, we need to scale them so that all features are given equal weight.

4.There are ads with 0 Revenue and 0 Spend.

Now let's check for anomalies in the data

Let check for Null values in the columns

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

Table 4: Count of Null Values

As can be seen, columns CTR, CPM and CPC have 4736 Null values each.

There are no duplicate records in the Dataset.

```
: df_ad.duplicated().sum()
: 0
```

2.Treat missing values in CPC, CTR and CPM using the formula given. You have to basically create a user defined function and then call the function for imputing.

Using the formula for CPC, CTR and CPM given below , we can impute the missing values

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.

We create a user defined Lambda function that will compute CPC, CTR and CPM and then we will apply that function on the columns CPC, CTR and CPM and thus impute the Null values.

There are some anomalies in the original dataset . There are some CPM values that are non - zero for cases where spend is 0 and values for CTR in the original dataset are expressed as percentage, we will correct those values as per the formula .

After Imputing let's check the count of Null values in CPC, CTR and CPM .

As we can see there are no null values

CTR	0
CPM	0
CPC	0

Table 5: Count of Null Values in CTR, CPC and CPM

Let's again load dataset and see if the missing values are imputed

Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
Inter222	Video	Desktop	Display	1806	325	323	1	0.00	0.35	0.0000	0.309598	0.0	0.00
Inter227	App	Mobile	Video	1780	285	285	1	0.00	0.35	0.0000	0.350877	0.0	0.00
Inter222	Video	Desktop	Display	2727	356	355	1	0.00	0.35	0.0000	0.281690	0.0	0.00
Inter228	Video	Mobile	Video	2430	497	495	1	0.00	0.35	0.0000	0.202020	0.0	0.00
Inter217	Web	Desktop	Video	1218	242	242	1	0.00	0.35	0.0000	0.413223	0.0	0.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...
Inter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	100.000000	70.0	0.07
Inter224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.0260	50.000000	20.0	0.04
Inter218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	100.000000	50.0	0.05
inter230	Video	Mobile	Video	7	1	1	1	0.07	0.35	0.0455	100.000000	70.0	0.07
Inter221	App	Mobile	Video	2	2	2	1	0.09	0.35	0.0585	50.000000	45.0	0.09

Figure 7:Dataset after imputing the Null values for CTR,CPM and CPC

As we can see now the null values are imputed as per the formulas . The Data Summary of the three columns after we have imputed the Null Values looks very different from the original data

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	3.851631e+02	2.336514e+02	120.000000	120.000000	300.000000	7.200000e+02	728.00
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.000000	250.000000	300.000000	6.000000e+02	600.00
Ad Size	23066.0	9.667447e+04	6.153833e+04	33600.000000	72000.000000	72000.000000	8.400000e+04	216000.00
Available_Impressions	23066.0	2.432044e+06	4.742888e+06	1.000000	33672.250000	483771.000000	2.527712e+06	27592861.00
Matched_Queries	23066.0	1.295099e+06	2.512970e+06	1.000000	18282.500000	258087.500000	1.180700e+06	14702025.00
Impressions	23066.0	1.241520e+06	2.429400e+06	1.000000	7990.500000	225290.000000	1.112428e+06	14194774.00
Clicks	23066.0	1.067852e+04	1.735341e+04	1.000000	710.000000	4425.000000	1.279375e+04	143049.00
Spend	23066.0	2.706626e+03	4.067927e+03	0.000000	85.180000	1425.125000	3.121400e+03	26931.87
Fee	23066.0	3.351231e-01	3.196322e-02	0.210000	0.330000	0.350000	3.500000e-01	0.35
Revenue	23066.0	1.924252e+03	3.105238e+03	0.000000	55.365375	926.335000	2.091338e+03	21276.18
CTR	18330.0	7.366038e+00	7.515998e+00	0.010874	0.258301	8.257612	1.300110e+01	100.00
CPM	18330.0	7.672045e+00	6.481391e+00	0.000000	1.710000	7.660000	1.251000e+01	81.56
CPC	18330.0	3.510606e-01	3.433338e-01	0.000000	0.090000	0.160000	5.700000e-01	7.26

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	3.851631e+02	2.336514e+02	120.000000	120.000000	300.000000	7.200000e+02	7.280000e+02
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.000000	250.000000	300.000000	6.000000e+02	6.000000e+02
Ad Size	23066.0	9.667447e+04	6.153833e+04	33600.000000	72000.000000	72000.000000	8.400000e+04	2.160000e+05
Available_Impressions	23066.0	2.432044e+06	4.742888e+06	1.000000	33672.250000	483771.000000	2.527712e+06	2.759286e+07
Matched_Queries	23066.0	1.295099e+06	2.512970e+06	1.000000	18282.500000	258087.500000	1.180700e+06	1.470202e+07
Impressions	23066.0	1.241520e+06	2.429400e+06	1.000000	7990.500000	225290.000000	1.112428e+06	1.419477e+07
Clicks	23066.0	1.067852e+04	1.735341e+04	1.000000	710.000000	4425.000000	1.279375e+04	1.430490e+05
Spend	23066.0	2.706626e+03	4.067927e+03	0.000000	85.180000	1425.125000	3.121400e+03	2.693187e+04
Fee	23066.0	3.351231e-01	3.196322e-02	0.210000	0.330000	0.350000	3.500000e-01	3.500000e-01
Revenue	23066.0	1.924252e+03	3.105238e+03	0.000000	55.365375	926.335000	2.091338e+03	2.127618e+04
CTR	23066.0	8.409941e+00	9.262048e+00	0.010874	0.265107	9.391248	1.347057e+01	2.000000e+02
CPM	23066.0	8.396849e+00	9.057760e+00	0.000000	1.749084	8.371566	1.304202e+01	7.150000e+02
CPC	23066.0	3.366776e-01	3.412527e-01	0.000000	0.089736	0.139347	5.462421e-01	7.264000e+00

Table 6: Data Summary Before and After Imputing Null Values

3.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).

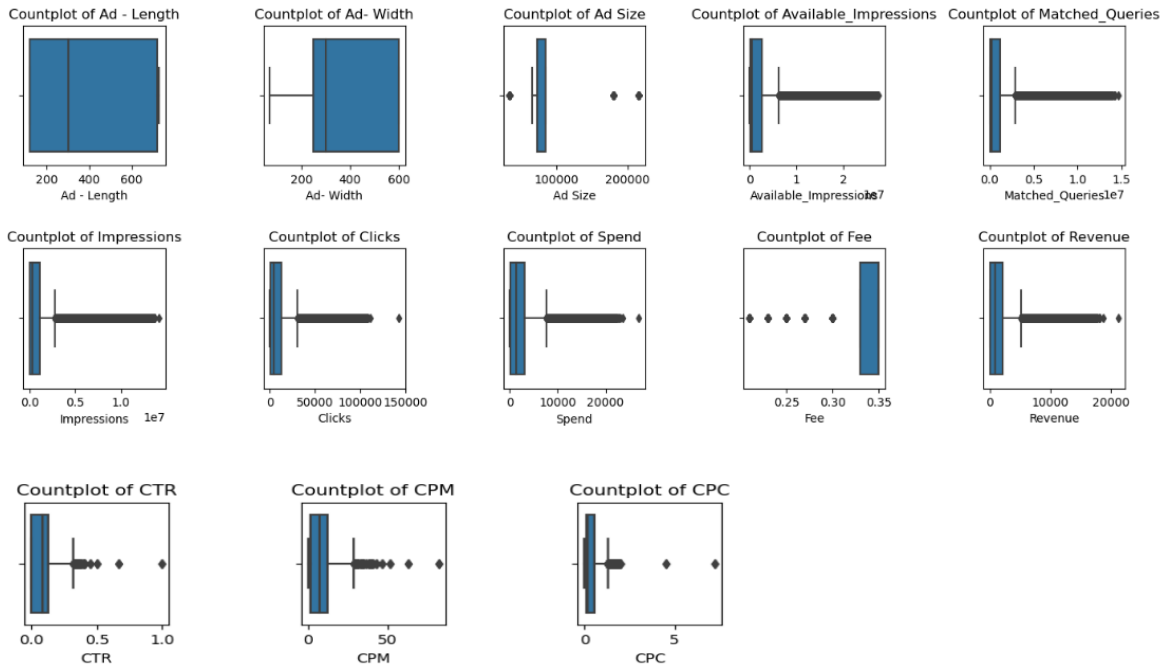


Figure 8:Boxplot of Numerical Variables

As can be seen from the boxplot above, there are outliers in many of the variables

After applying IQR method we see that there are outliers in some of the variables.

Outlier increases the mean of data .Since K-Means algorithm is about finding mean of clusters, the algorithm is influenced by outliers. The k-means algorithm updates the cluster centres by taking the average of all the data points that are closer to each cluster centre. When all the points are packed nicely together, the average makes sense. However, when you have outliers, this can affect the average calculation of the whole cluster. As a result, this will push cluster centre closer to the outlier. The benefit of removing outliers is to enhance the accuracy and stability of statistical models by reducing their impact on results. Outliers can distort statistical analyses and skew results as they are extreme values that differ from the rest of the data. Removing outliers makes the results more robust and accurate by eliminating their influence.

To treat outlier, we use the Capping and Flooring method where in data points that are exceeding the upper whisker or Q3+1.5*IQR (Q3 is the 75th percentile and IQR is Interquartile range)is equated to upper whisker limit of the distribution and the values less than lower whisker or Q1 -1.5*IQR(Q1 is the 25th percentile) are all equated to lower whisker limit value of the distribution. Let's apply this method and check if outliers are treated .

We take only the variables that need to be treated and create a user defined function to treat the outliers for each of those variables and apply that function of each of those variables

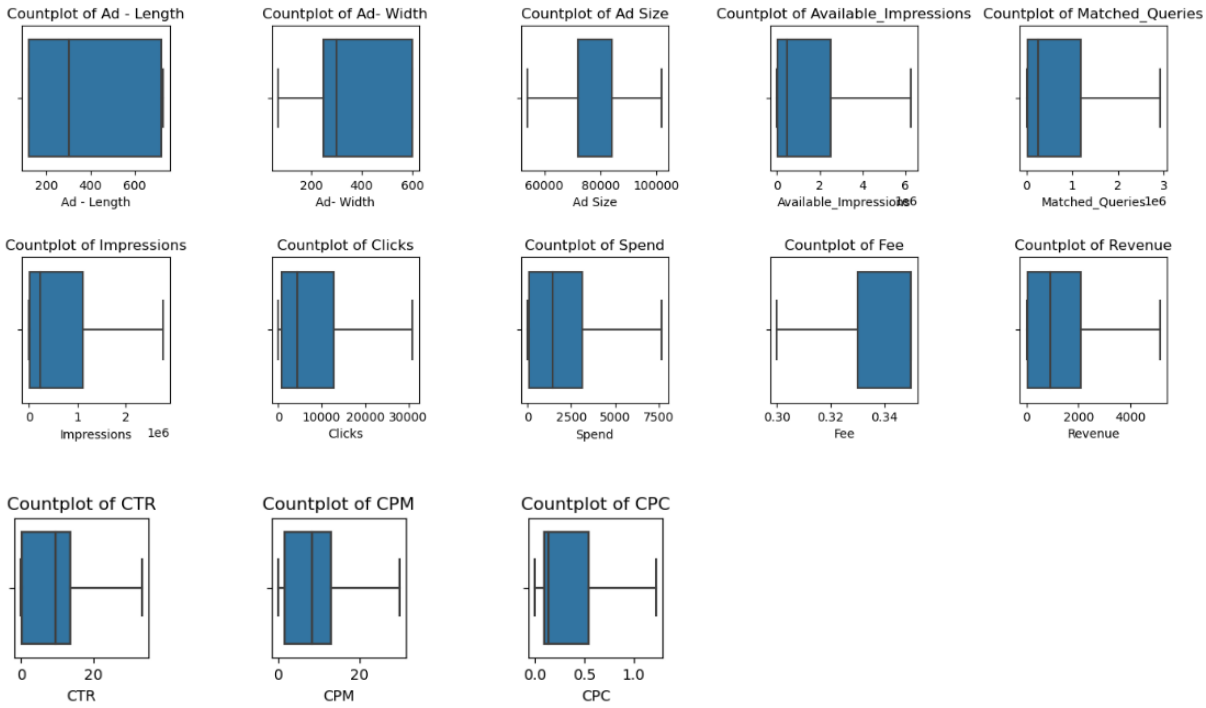


Figure 9: Boxplot after treating Outliers

After applying the function, we can see from the plots above , that the outliers have been treated and there are no outliers in the variables now .Let's see the data summary after treating outliers. The maximum and minimum values are now capped and floored

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	3.851631e+02	2.336514e+02	120.000000	120.000000	300.000000	7.200000e+02	7.280000e+02
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.000000	250.000000	300.000000	6.000000e+02	6.000000e+02
Ad Size	23066.0	7.657684e+04	1.538132e+04	54000.000000	72000.000000	72000.000000	8.400000e+04	1.020000e+05
Available_Impressions	23066.0	1.607253e+06	2.125528e+06	1.000000	33672.250000	483771.000000	2.527712e+06	6.268771e+06
Matched_Queries	23066.0	7.995380e+05	1.026037e+06	1.000000	18282.500000	258087.500000	1.180700e+06	2.924326e+06
Impressions	23066.0	7.536120e+05	9.802568e+05	1.000000	7990.500000	225290.000000	1.112428e+06	2.769086e+06
Clicks	23066.0	8.306828e+03	9.574779e+03	1.000000	710.000000	4425.000000	1.279375e+04	3.091938e+04
Spend	23066.0	2.166060e+03	2.425190e+03	0.000000	85.180000	1425.125000	3.121400e+03	7.675730e+03
Fee	23066.0	3.402683e-01	1.812855e-02	0.300000	0.330000	0.350000	3.500000e-01	3.500000e-01
Revenue	23066.0	1.449389e+03	1.646894e+03	0.000000	55.365375	926.335000	2.091338e+03	5.145297e+03
CTR	23066.0	8.223203e+00	8.253522e+00	0.010874	0.265107	9.391248	1.347057e+01	3.327877e+01
CPM	23066.0	8.219181e+00	6.881016e+00	0.000000	1.749084	8.371566	1.304202e+01	2.998142e+01
CPC	23066.0	3.300346e-01	3.165682e-01	0.000000	0.089736	0.139347	5.462421e-01	1.231002e+00

Table 7:Data Summary after Treating Outliers

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

In the machine learning algorithms if the values of the features are closer to each other there are chances for the algorithm to get trained well and faster instead of the data set where the data points or features values have high differences with each other .It will take more time to understand the data and the accuracy will be lower. So, if the data in any conditions has data points far from each other, scaling is a technique to make them closer to each other. Scaling is used for making data points generalized so that the distance between them will be lower. Scaling features restricts modules from being biased towards features having lower /higher magnitude.

Therefore, it's important we scale the features so that they all are on same scale which in turn helps model to assign equal importance to all features and make predictions without bias

Now let's scale the numerical variables using Z score

Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
-0.364496	-0.432797	-0.102518	-0.755333	-0.778949	-0.768478	-0.867488	-0.893170	0.535724	-0.880093	-0.958836	-1.194498	-1.042561
-0.364496	-0.432797	-0.102518	-0.755345	-0.778988	-0.768516	-0.867488	-0.893170	0.535724	-0.880093	-0.953835	-1.194498	-1.042561
-0.364496	-0.432797	-0.102518	-0.754900	-0.778919	-0.768445	-0.867488	-0.893170	0.535724	-0.880093	-0.962218	-1.194498	-1.042561
-0.364496	-0.432797	-0.102518	-0.755040	-0.778781	-0.768302	-0.867488	-0.893170	0.535724	-0.880093	-0.971871	-1.194498	-1.042561
-0.364496	-0.432797	-0.102518	-0.755610	-0.779030	-0.768560	-0.867488	-0.893170	0.535724	-0.880093	-0.946281	-1.194498	-1.042561
...	...	...	...	...	...	...	...	...	...	...	...	...
1.433093	-0.186599	1.652896	-0.756182	-0.779265	-0.768806	-0.867488	-0.893141	0.535724	-0.880066	3.035808	3.162718	-0.821435
1.433093	-0.186599	1.652896	-0.756181	-0.779264	-0.768805	-0.867488	-0.893154	0.535724	-0.880078	3.035808	1.712113	-0.916204
1.433093	-0.186599	1.652896	-0.756182	-0.779265	-0.768806	-0.867488	-0.893150	0.535724	-0.880074	3.035808	3.162718	-0.884614
-1.134891	1.290590	-0.297564	-0.756179	-0.779265	-0.768806	-0.867488	-0.893141	0.535724	-0.880066	3.035808	3.162718	-0.821435
1.433093	-0.186599	1.652896	-0.756182	-0.779264	-0.768805	-0.867488	-0.893133	0.535724	-0.880058	3.035808	3.162718	-0.758256

Figure 10: Z score scaling of the Numerical Variables

As we can observe now all the variables are now scaled with mean close to 0 and standard deviation close to 1

	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	1.329072e-15	1.000022	-0.995031	-0.964227	0.141524	0.635787	3.035808
CPM	23066.0	5.791296e-17	1.000022	-1.194498	-0.940303	0.022146	0.700905	3.162718
CPC	23066.0	1.987283e-15	1.000022	-1.042561	-0.759091	-0.602371	0.682987	2.846105

Table 8: Data Summary after scaling using Z score method

5.Perform clustering and do the following:

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

A dendrogram is a visual representation of cluster-making. On the x-axis are the cluster names or numbers and the y-axis is the distance or height. The vertical straight lines denote the height where two clusters combine. The higher the level of combining, the distant the individual items or clusters are. In hierarchical clustering, all items must combine to make one cluster. Each cluster is a representative of a different population. After constructing the dendrogram, we decide the level where the resultant tree needs to be cut. If the number of clusters is large, the cluster size is small and the clusters homogeneous. If the number of clusters is small, each contains more item sand hence clusters are more heterogeneous. Depending on the distance measure and linkage used, the number of clusters and their composition maybe different. Now let us proceed to the clustering using Euclidean distance and ward linkage method. Using scipy.cluster.hierarchy library we create dendrograms forward and Euclidean distance

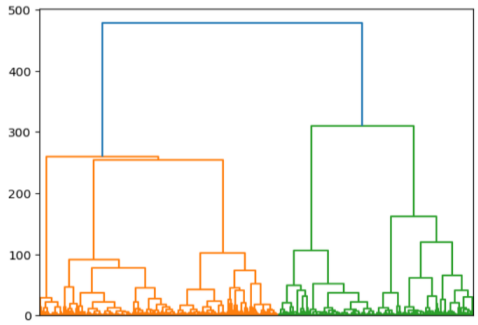


Figure 11: Dendrogram using metric='Euclidean' and method='ward' with all the Clusters

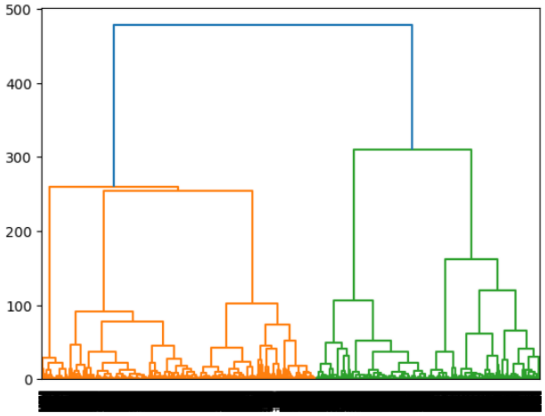


Figure 12:Dendrogram using method='ward' with all the Clusters

Dendrogram of Figure 11 includes all clusters . Going by color_threshold, there are 2 clusters. It is difficult for a business to understand segments from just 2 clusters. After multiple iterations, we can identify the suitable number of clusters.

Let's view the Last 10 clusters on the dendrogram

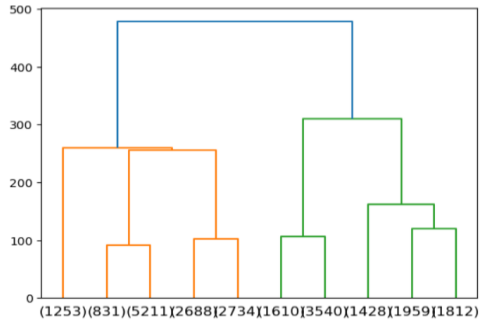


Figure 13: Dendrogram using metric='Euclidean' and method='ward' last 10 Clusters

Figure 13 does not show the full-grown tree but just the last 10 clusters. Cluster size is not equal for all 10 clusters, as indicated in the above diagram(X-axis give the cluster size). In the dendrogram we locate the largest vertical difference between nodes, and in the middle pass a horizontal line. The number of vertical lines intersecting it is the optimal number of clusters (when affinity is calculated using the method set in linkage). It is important to note that based on the method selected for linkage and affinity, cluster membership and cluster size could vary.

As we see the number of vertical lines intersecting is 5, if we draw a horizontal line at the largest vertical distance , so let's assume the cluster size as 5.We can further analyse using K-means clustering to conclude on the number of clusters .

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

k-means clustering is the most used non-hierarchical clustering technique. It aims to partition n observations into k clusters in which each observation belongs to the cluster whose mean (centroid) is nearest to it, serving as a prototype of the cluster. It minimizes within-cluster variances (squared Euclidean distances).For a given number of clusters, the total within-cluster sum of squares (WSS) is computed. That value of k is chosen to be optimum, where addition of one more cluster does not lower the value of total WSS appreciably. Following are WSS scores for each value of K starting from k=1 up to k=10.

```
[299858.0000000002,
 183349.1020288608,
 130878.34788742862,
 95133.93066619668,
 61539.18919785385,
 51676.892307099595,
 44598.27017775245,
 39597.84594043494,
 37179.24975829046,
 33619.0082162612]
```

Figure 14: WSS Scores

The Elbow method looks at the total WSS as a function of the number of clusters. Let's plot the WSS scores for each cluster

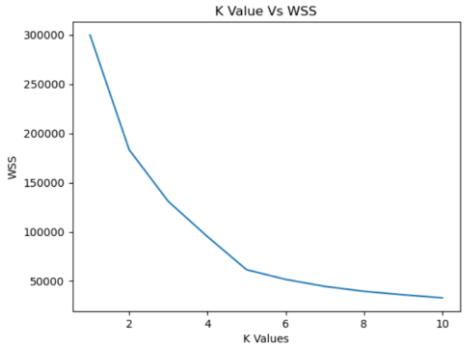


Figure 15:Elbow Plot of WSS

In this Elbow method, we are actually varying the number of clusters (K) from 1 – 10. For each value of K, we are calculating WCSS (Within-Cluster Sum of Square). WCSS is the sum of the squared distance between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot looks like an Elbow. As the number of clusters increases, the WCSS value will start to decrease. WCSS value is largest when K = 1. When we analyse the graph, we can see that the graph will rapidly change at a point and thus creating an elbow shape. From this point, the graph moves almost parallel to the X-axis. The K value corresponding to this point is the optimal value of K or an optimal number of clusters. As can be seen from the plot above , the point at which the elbow shape is created is 5; that is, our K value or an optimal number of clusters is 5.

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

Silhouette method measures how tightly the observations are clustered and the average distance between clusters. For each observation a silhouette score is constructed which is a function of the average distance between the point and all other points in the cluster to which it belongs, and the distance between the point and all other points in all other clusters, that it does not belong to. The maximum value of the statistic indicates the optimum value of k.

Following are the Silhouette scores up to 10 clusters

```
The Average Silhouette Score for 2clusters is 0.38573
The Average Silhouette Score for 3clusters is 0.38255
The Average Silhouette Score for 4clusters is 0.44534
The Average Silhouette Score for 5clusters is 0.5241
The Average Silhouette Score for 6clusters is 0.52215
The Average Silhouette Score for 7clusters is 0.51656
The Average Silhouette Score for 8clusters is 0.47975
The Average Silhouette Score for 9clusters is 0.43187
The Average Silhouette Score for 10clusters is 0.44463
```

Figure 16:Silhouette Scores.

As can be seen from the Silhouette scores, the maximum value of the statistic is at k value 5 after which scores are decreasing , therefore we can say that optimum value of k is 5.

Let's plot the Silhouette scores against each cluster and verify the same .

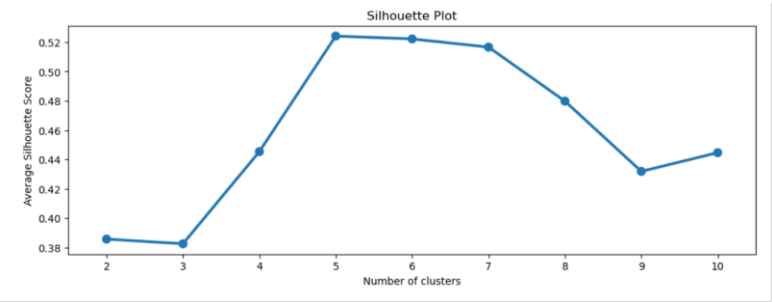


Figure 17 Silhouette Plot.

It is clear from Fig 17 that the maximum value of average silhouette score is achieved for k= 5, which, therefore, is considered to be the optimum number of clusters for this data.

8.Profile the ads based on optimum number of clusters using silhouette score and your domain understanding

[Hint: Group the data by clusters and take sum or mean to identify trends in clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots.]

We found that optimum number of clusters is 5 , so we can add a new column Clus_Kmeans to the Original Dataset which can now be profiled into 5 different clusters .The Figure below depicts the dataset assigned with 5 different clusters .There are 5 clusters 1,2,3,4,5 and each observation is grouped into one of these 5 clusters .

Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	Clus_kmeans
Inter222	Video	Desktop	Display	1806	325	323	1	0.00	0.35	0.0000	0.309598	0.0	0.00	4
Inter227	App	Mobile	Video	1780	285	285	1	0.00	0.35	0.0000	0.350877	0.0	0.00	4
Inter222	Video	Desktop	Display	2727	356	355	1	0.00	0.35	0.0000	0.281690	0.0	0.00	4
Inter228	Video	Mobile	Video	2430	497	495	1	0.00	0.35	0.0000	0.202020	0.0	0.00	4
Inter217	Web	Desktop	Video	1218	242	242	1	0.00	0.35	0.0000	0.413223	0.0	0.00	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Inter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	100.000000	70.0	0.07	3
Inter224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.0260	50.000000	20.0	0.04	3
Inter218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	100.000000	50.0	0.05	3
inter230	Video	Mobile	Video	7	1	1	1	0.07	0.35	0.0455	100.000000	70.0	0.07	2
Inter221	App	Mobile	Video	2	2	2	1	0.09	0.35	0.0585	50.000000	45.0	0.09	3

Figure 18: Dataset with Clusters

Let's group the data into clusters and observe the behaviour of the important variables in each of the clusters by taking the means of these variables to identify trends in clicks, spend, revenue, CPM, CTR, & CPC.

		Clicks	Spend	Revenue	CPM	CTR	CPC
Clus_kmeans	Device Type						
1	Desktop	71686.146930	7602.611009	5498.218435	15.157941	13.801457	0.109803
	Mobile	71481.951066	7670.540928	5552.094863	15.248525	13.758742	0.110726
2	Desktop	3443.119147	380.899573	251.033256	14.679239	16.084188	0.102192
	Mobile	3534.177864	385.896597	254.378293	14.978210	16.122166	0.102293
3	Desktop	12891.169065	1115.317785	726.747997	12.398985	14.050985	0.091692
	Mobile	12747.637514	1116.921016	727.722050	12.368627	13.989152	0.091975
4	Desktop	3258.727372	1231.849318	800.772569	1.782531	0.449867	0.453324
	Mobile	3257.343185	1236.766577	804.006277	1.783221	0.444314	0.457464
5	Desktop	9580.749866	7354.112994	5373.367738	1.616326	0.217398	0.793896
	Mobile	9433.225990	7318.944568	5345.818105	1.633069	0.216845	0.804611

Table 9: Grouping data by clusters with mean to identify trends in clicks, spend, revenue, CPM, CTR, & CPC based on Device Type

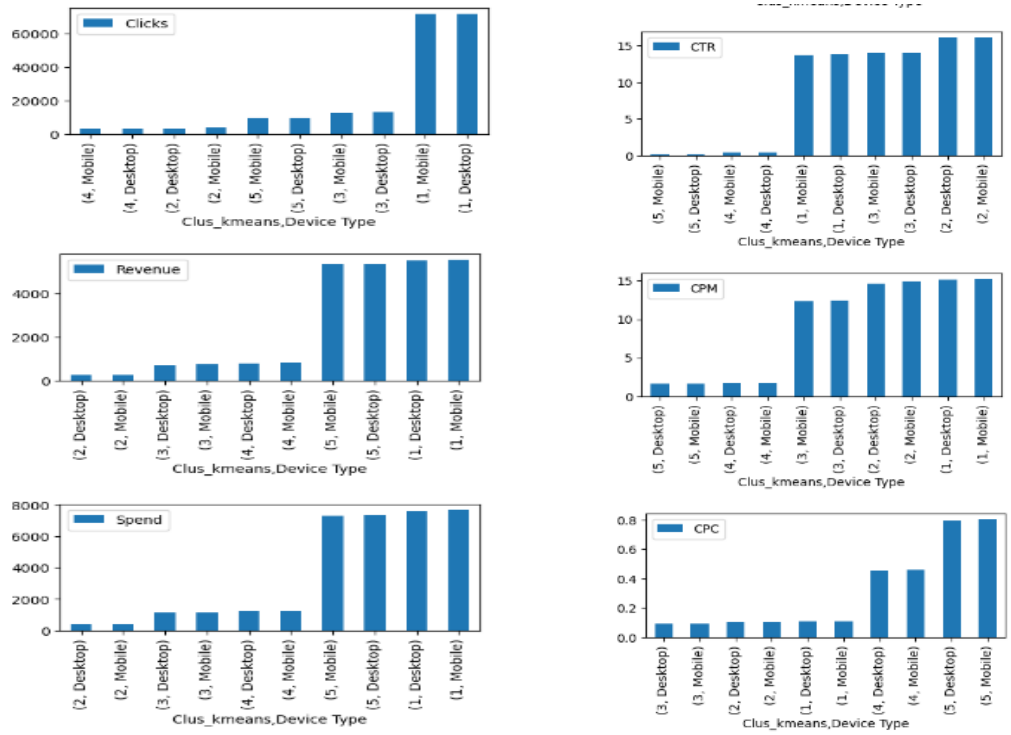


Figure 19 :Graphical Representation of Cluster wise trends in clicks, spend, revenue, CPM, CTR, & CPC based on Device Type

Some observations based on the data summary and Visualizations:

Cluster 1 has most Clicks and Cluster 4 has lowest Clicks

Cluster 1 has highest Revenue and Spend and Cluster 2 has lowest Revenue and Spend

Cluster 2 has highest CTR and Cluster 5 has lowest CTR

Cluster 1 has highest CPM and Cluster 5 has lowest CPM

Cluster 5 has highest CPC and Cluster 3 has lowest CPC

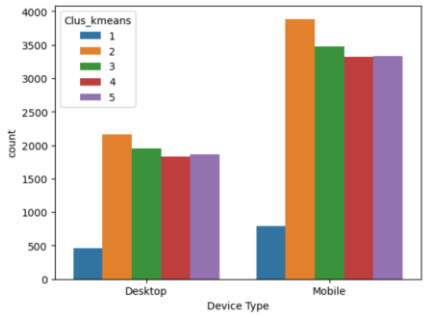


Figure 20: Countplot of Device Types and Clusters

There are more Ads played on Mobile device than on Desktop for all clusters

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

1.Cluster 1 has the highest Revenue among all clusters, also highest spends. But the Fee is lowest among all clusters . They also have the highest Clicks and highest CPM.Such ads should be preferred as they are having low fee but highest revenue generating Ads.

2.Ads in Cluster2 are the lowest on spending and revenue is also very low . Their Fee is also quite high compared to another Ad's. They have highest CTR (click to Impression ratio) and CPM (Spend per 1000 Impressions) is also high. CPC(spend per Click) is low but it's not increasing their revenue. They have second lowest clicks probably because Fee is quite high, so to increase revenue they can lower the Fees.

3.Ads in Cluster3 has the highest Ad size, with second least Revenue and High Fee and lowest CPC(spend per Click), CPM(Spend per 1000 Impressions) is second highest . Maybe they can reduce their Ad size as it's not guaranteeing more clicks or more revenue which could reduce some cost and spend more on creating better Ads.

4.Ads in Cluster4 has highest Fees among clusters and lowest Clicks and not a high revenue generating cluster . CPM(Spend per 1000 Impressions) is low . CPC (spend per Click) is second highest and CTR (click to Impression ratio) is very low. They have highest number of Impressions , but lowest clicks .May be reducing the Fees can increase number of clicks thereby increasing Revenue.

5.Ads in Cluster 5 has highest number of Impressions . Their Fees is also not very high with second highest Revenue and Spend. Highest CPC (spend per Click),Lowest CPM(Spend per 1000 Impressions). and lowest CTR(click to Impression ratio). They are not getting many clicks though the spend per click is high. Such Ads are generating good revenue amount but they have highest CPC so cost should be reduced to increase more revenue

Problem2 :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.

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

Ans: Lets load the PCA India Data Census

:

	State Code	Dist.Code	State	Area Name	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
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	0	1999	2598	13381	11364	10007	18432	6723
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	6	427	517	10513	7891	9072	15211	6982
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	6	5806	9723	4534	5840	2012	5124	2775
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	0	2666	3968	1842	1962	942	2244	1002
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	33	7670	10843	13243	13477	7348	16504	5717
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
635	34	636	Puducherry	Mahe	3333	8154	11781	1146	1203	21	30	0	0	6916	10184	1238	1597	3808
636	34	637	Puducherry	Karaikal	10612	12346	21691	1544	1533	2234	4155	0	0	10292	14225	2054	7466	6458
637	35	638	Andaman & Nicobar Island	Nicobars	1275	1549	2630	227	225	0	0	1012	1750	1187	1602	362	1028	715
638	35	639	Andaman & Nicobar Island	North & Middle Andaman	3762	5200	8012	723	664	0	0	28	50	4206	5273	994	2739	2707
639	35	640	Andaman & Nicobar Island	South Andaman	7975	11977	18049	1470	1358	0	0	161	264	10095	13362	1882	4687	6345

640 rows × 61 columns

There are 640 observations and 61 columns

Let's Analyse the Dataset provided PCA India Data Census.xlsx and print the top 5 observations.

	0	1	2	3	4
State Code	1	1	1	1	1
Dist.Code	1	2	3	4	5
State	Jammu & Kashmir	Jammu & Kashmir	Jammu & Kashmir	Jammu & Kashmir	Jammu & Kashmir
Area Name	Kupwara	Badgam	Leh(Ladakh)	Kargil	Punch
No_HH	7707	6218	4452	1320	11654
...	...	...	...	...	...
MARG_HH_0_3_F	252	148	34	50	302
MARG_OT_0_3_M	32	76	0	4	24
MARG_OT_0_3_F	46	178	4	10	105
NON_WORK_M	258	140	67	116	180
NON_WORK_F	214	160	61	59	478

Figure 21:Top 5 observations of PCA India Data Census

	635	636	637	638	639
State Code	34	34	35	35	35
Dist.Code	636	637	638	639	640
State	Puducherry	Puducherry	Andaman & Nicobar Island	Andaman & Nicobar Island	Andaman & Nicobar Island
Area Name	Mahe	Karaikal	Nicobars	North & Middle Andaman	South Andaman
No_HH	3333	10612	1275	3762	7975
...	...	...	...	...	...
MARG_HH_0_3_F	0	130	6	21	17
MARG_OT_0_3_M	0	4	17	1	2
MARG_OT_0_3_F	0	23	47	4	4
NON_WORK_M	32	110	76	100	148
NON_WORK_F	47	170	77	103	99

61 rows × 5 columns

Figure 22: Bottom 5 observations of PCA India Data Census

Printing the information about 61 columns

```
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
```

Table 10: Dataset Column Information1

```
 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
```

Table 11: Dataset Column Information2

As can be seen from the figures above there are 59 Integer type and 2 Object types Columns
State code and Dist Code are numerical values , but they are actually categorical values as they are not continuous variables
Printing the summary of the numerical columns(State code and Dist code are excluded)

	count	mean	std	min	25%	50%	75%	max
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
MAIN_AL_F	640.0	8925.995312	12864.287584	0.0	1408.75	3933.5	10617.50	87945.0
MAIN_HH_M	640.0	883.893750	1278.642345	0.0	187.50	498.5	1099.25	16429.0
MAIN_HH_F	640.0	1380.773438	3179.414449	0.0	248.75	540.5	1435.75	45979.0
MAIN_OT_M	640.0	18047.101562	26068.480886	36.0	3997.50	9598.0	21249.50	240855.0
MAIN_OT_F	640.0	12406.035938	18972.202369	153.0	3142.50	6380.5	14368.25	209355.0
MARGWORK_M	640.0	7787.960938	7410.791691	35.0	2937.50	5627.0	9800.25	47553.0
MARGWORK_F	640.0	13096.914062	10996.474528	117.0	5424.50	10175.0	18879.25	66915.0
MARG_CL_M	640.0	1040.737500	1311.546847	0.0	311.75	606.5	1281.00	13201.0
MARG_CL_F	640.0	2307.682813	3564.626095	0.0	630.25	1226.0	2659.25	44324.0
MARG_AL_M	640.0	3304.326562	3781.555707	0.0	873.50	2062.0	4300.75	23719.0
MARG_AL_F	640.0	6463.281250	6773.876298	0.0	1402.50	4020.5	9089.25	45301.0
MARG_HH_M	640.0	316.742188	462.661891	0.0	71.75	166.0	356.50	4298.0
MARG_HH_F	640.0	786.626562	1198.718213	0.0	171.75	429.0	962.50	15448.0
MARG_OT_M	640.0	3126.154687	3609.391821	7.0	935.50	2036.0	3985.25	24728.0
MARG_OT_F	640.0	3539.323438	4115.191314	19.0	1071.75	2349.5	4400.50	36377.0
MARGWORK_3_6_M	640.0	41948.168750	39045.316918	291.0	16208.25	30315.0	57218.75	300937.0
MARGWORK_3_6_F	640.0	81076.323438	82970.406216	341.0	26619.50	56793.0	107924.00	676450.0
MARG_CL_3_6_M	640.0	6394.987500	6019.806644	27.0	2372.00	4630.0	8167.00	39106.0
MARG_CL_3_6_F	640.0	10339.864063	8467.473429	85.0	4351.50	8295.0	15102.00	50065.0
MARG_AL_3_6_M	640.0	789.848438	905.639279	0.0	235.50	480.5	986.00	7426.0
MARG_AL_3_6_F	640.0	1749.584375	2496.541514	0.0	497.25	985.5	2059.00	27171.0
MARG_HH_3_6_M	640.0	2743.635938	3059.586387	0.0	718.75	1714.5	3702.25	19343.0
MARG_HH_3_6_F	640.0	5169.850000	5335.640960	0.0	1113.75	3294.0	7502.25	36253.0
MARG_OT_3_6_M	640.0	245.362500	358.728567	0.0	58.00	129.5	276.00	3535.0
MARG_OT_3_6_F	640.0	585.884375	900.025817	0.0	127.75	320.5	719.25	12094.0
MARGWORK_0_3_M	640.0	2616.140625	3036.964381	7.0	755.00	1681.5	3320.25	20648.0
MARGWORK_0_3_F	640.0	2834.545312	3327.836932	14.0	833.50	1834.5	3610.50	25844.0
MARG_CL_0_3_M	640.0	1392.973438	1489.707052	4.0	489.50	949.0	1714.00	9875.0
MARG_CL_0_3_F	640.0	2757.050000	2788.776676	30.0	957.25	1928.0	3599.75	21611.0
MARG_AL_0_3_M	640.0	250.889062	453.336594	0.0	47.00	114.5	270.75	5775.0
MARG_AL_0_3_F	640.0	558.098438	1117.642748	0.0	109.00	247.5	568.75	17153.0
MARG_HH_0_3_M	640.0	560.690625	762.578991	0.0	136.50	308.0	642.00	6116.0
MARG_HH_0_3_F	640.0	1293.431250	1585.377936	0.0	298.00	717.0	1710.75	13714.0
MARG_OT_0_3_M	640.0	71.379688	107.897627	0.0	14.00	35.0	79.00	895.0
MARG_OT_0_3_F	640.0	200.742188	309.740854	0.0	43.00	113.0	240.00	3354.0
NON_WORK_M	640.0	510.014063	610.603187	0.0	161.00	326.0	604.50	6456.0
NON_WORK_F	640.0	704.778125	910.209225	5.0	220.50	464.5	853.50	10533.0

Table 12: Describing the Numerical Columns

	count	unique	top	freq
State	640	35	Uttar Pradesh	71
Area Name	640	635	Raigarh	2

Table 13: : Describing the Categorical Columns

As we can see all the Numerical variables are have a right skewed distribution.

As we can among categorical variables ,Uttar Pradesh has the greatest number of observations and Raigarh is appearing twice because there both states Chhattisgarh and Maharashtra have area names called Raigarh

There are no Null values in the Dataset.

There are no duplicated observations 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
```

Table 14: Null Values in the Dataset

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

Let's pick 5 variables M_LIT, TOT_M, TOT_F, TOT_WORK_M, TOT_WORK_F for our analysis

M_LIT-Literates population Male

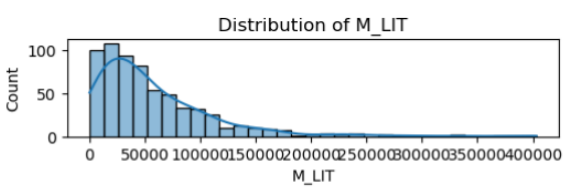
TOT_M- Total population Male

TOT_F- Total population Female

TOT_WORK_M- Total Worker Population Male

TOT_WORK_F- Total Worker Population Female

Univariate Analysis



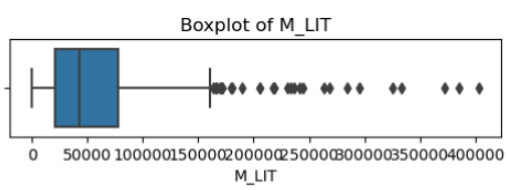


Figure 23: Distribution Plot and Boxplot of Number of Literate Males

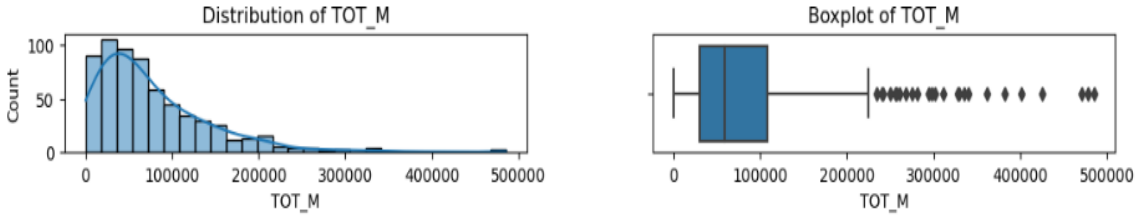


Figure 24: Distribution Plot and Boxplot of Total Number of Males

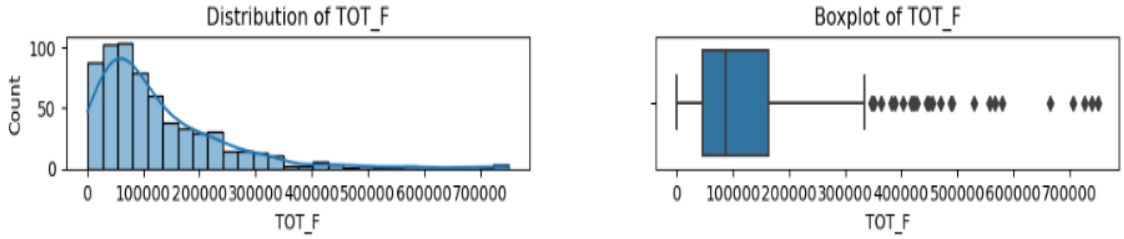


Figure 25: Distribution Plot and Boxplot of Total Number of Females

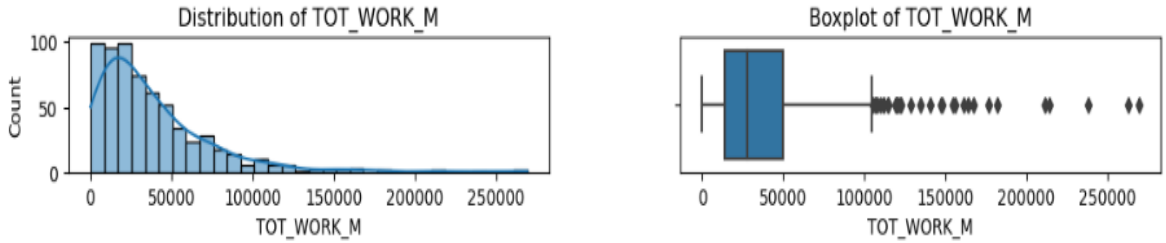


Figure 26: Distribution Plot and Boxplot of Total Number of Male Workers

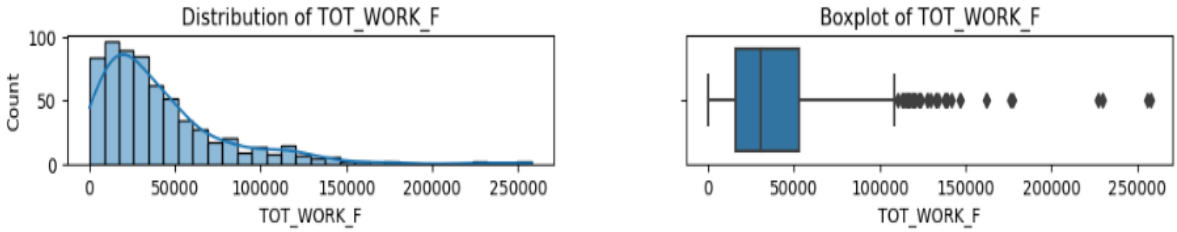


Figure 27: Distribution Plot and Boxplot of Total Number of Female Workers

	TOT_M	TOT_F	TOT_WORK_M	TOT_WORK_F	M_LIT
count	640.000000	640.000000	640.000000	640.000000	640.000000
mean	79940.576563	122372.084375	37992.407813	41295.760938	57967.979688
std	73384.511114	113600.717282	36419.537491	37192.360943	55910.282466
min	391.000000	698.000000	100.000000	357.000000	286.000000
25%	30228.000000	46517.750000	13753.500000	16097.750000	21298.000000
50%	58339.000000	87724.500000	27936.500000	30588.500000	42693.500000
75%	107918.500000	164251.750000	50226.750000	53234.250000	77989.500000
max	485417.000000	750392.000000	269422.000000	257848.000000	403261.000000

Table 15: Data Description of the 5 variables

Inferences from above representations of the 5 Numerical Variables

1.All the distributions are right skewed with outliers

2.The average number of females are more than the males.

3.There average number of females working population is also slightly more than male working population

4.Standard deviation of total females is very high compared to total males indicating the distribution is more spread out .

5.There are a minimum of 286 literate males in each district

Bivariate Analysis of Numerical and Categorical Variables

Let's Analyse the state wise numbers of these variables

Observation from the EDA:

When we plot the Average Number of Literate males against states, we see that Uttar Pradesh seems to be having highest whereas Dadra and Nagar Haveli the least number of Literate Males

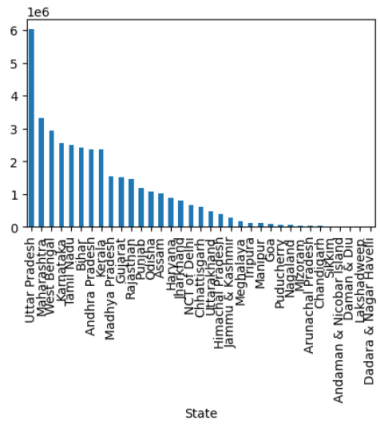


Figure 28: Bar plot of Literate males in each state

When we plot the Number of Males against states, we see that Uttar Pradesh seems to be having highest number of males whereas Dadar and Nagar Haveli has the least number of males

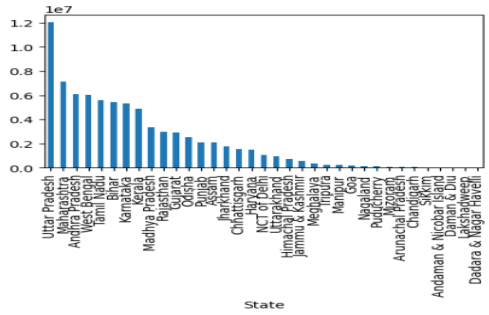


Figure 29: bagplot of Total Males in each state

When we plot the Number of Females against states, we see that Uttar Pradesh seems to be having highest number of females whereas Dadar and Nagar Haveli has the least number of females

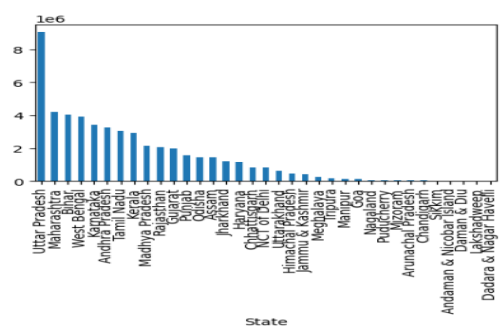


Figure 30: bar plot of Total Females in each state

When we plot the number of working males against states, we see that Uttar Pradesh seems to be having highest number of working males whereas Dadar and Nagar Haveli has the least number of working males

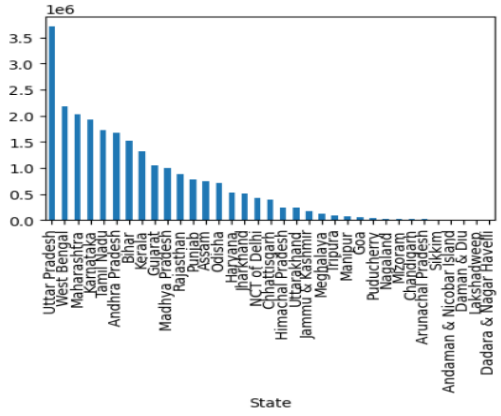


Figure 31: bar plot of Total Working Males in each state

When we plot the number of working females against states, we see that Uttar Pradesh seems to be having highest number of working females whereas Dadar and Nagar Haveli has the least number of working females

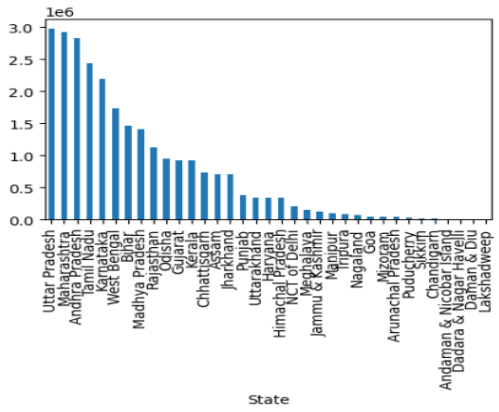


Figure 32: bar plot of Total Working Females in each state

Scatterplot of Numerical Variables to see the correlation among the numerical variables

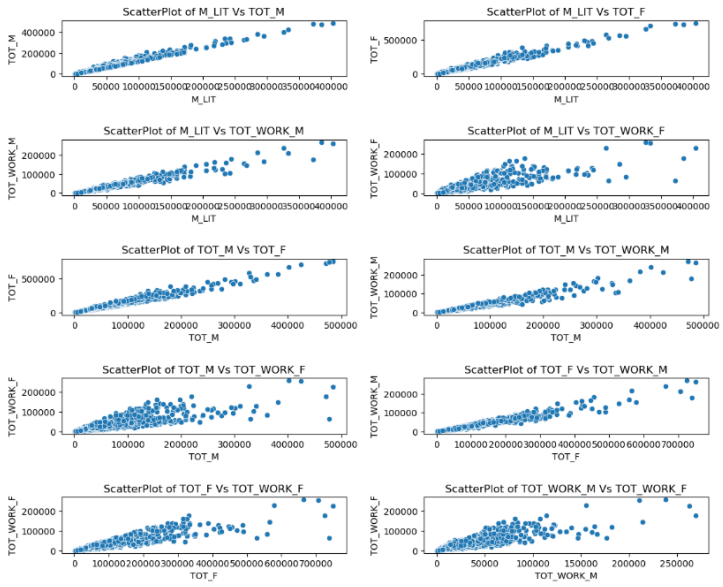


Figure 33:Scatterplot of Numerical Variables

We can see that there is a positive correlation among the numerical variables as seen in the scatterplot

When we plot the Gender Ratio against each state, we can observe that Lakshadweep seems to have a very high Male to Female Ratio whereas Andhra Pradesh seems to have least Male to Female Ratio, whereas when we compare the districts, Lakshadweep seems to have a high gender ratio of .87 and Krishna District in Andhra Pradesh seems to have the lowest ratio with .437 (Plotting will not possible because there are 640 districts).

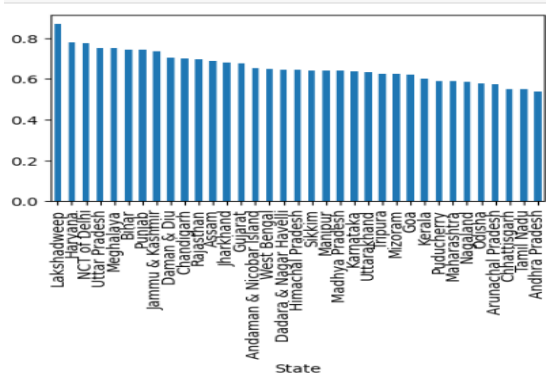


Figure 34: bar plot of Gender Ratio is Each State

Area Name	Ratio
Lakshadweep	0.868061
Badgam	0.847762
Mahamaya Nagar	0.847313
Dhaulpur	0.846911
Baghpat	0.844003
...	...
Baudh	0.451455
West Godavari	0.450076
Virudhunagar	0.449352
Koraput	0.440769
Krishna	0.437972

Figure 35:District wise Ratio of Gender

When we plot the Gender Ratio of working population against each state, we can observe that Lakshadweep seems to have a very high Male to Female working population Ratio whereas Arunachal Pradesh seems to have least Male to Female working population Ratio

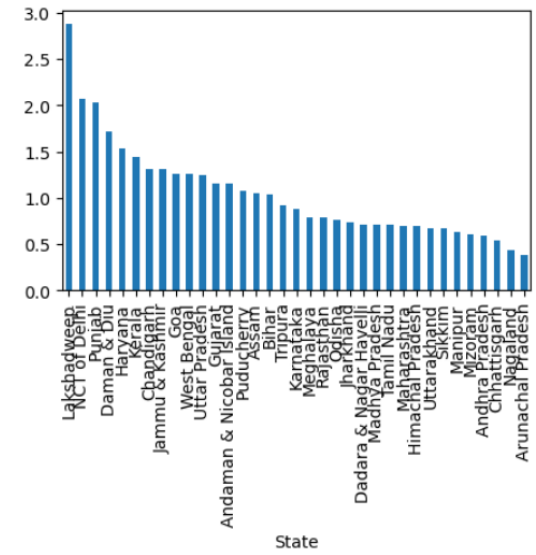


Figure 36:State wise Gender Ratio of Male to Female Working Population

When we plot the Ratio of Literate population of Males against total Male population in each state , we can observe that Goa seems to have a very high Ratio of Literate population of Males against total Male population whereas Bihar seems to have least Ratio of Literate population of Males against total Male population

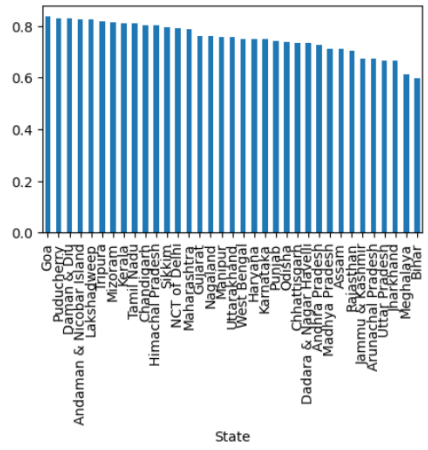


Figure 37: State wise ratio of literate Male to Total Male Population

When we plot the Ratio of working population of Males against total Male population in each state , we can observe that Karnataka seems to have a very high Ratio of working population of Males against total Male population whereas Arunachal Pradesh seems to have least Ratio of working population of Males against total Male population

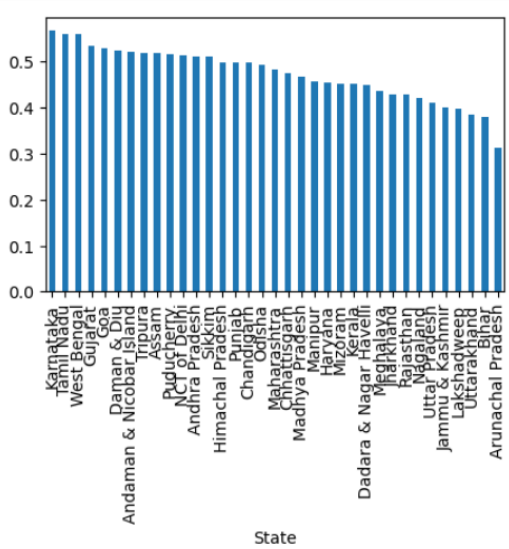


Figure 38: State wise ratio of Working Male to Total Male Population

When we plot the Ratio of working population of females against total females' population in each state , we can observe that Nagaland seems to have a very high Ratio of working population of females against total females' population whereas Lakshadweep seems to have least Ratio of working population of females against total females' population

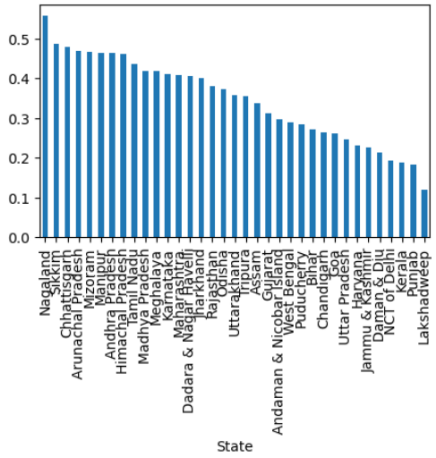


Figure 39: : State wise ratio of Working Female to Total Female Population

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

In a dataset, outliers are data points that strongly deviate from the norm and can impact model accuracy. Normally , treating outliers from a dataset before model training can improve the performance of the resulting model.

But some reasons why we should not remove outliers in a dataset:

- When there are a lot of observations in the dataset as it could mean something about the data that need to be analysed further
- When the model results are critical and can easily pose risks, for example, if dealing with sensitive use cases in the health or self-driving sectors
- When the outliers are natural to the data, in this case population of the entire country. They could have hidden patterns which otherwise wouldn't be unearthed if they were to be removed.

So, in this case treating outliers is not a good option

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

From the data description we see that variances of the variables are so widely different; it is not a good idea to perform PCA on the unscaled variables. PCA works on the total variance which is the sum of the variances in the data. If one variance or more variance(s) is very high compared to the rest, it will dominate the construction of the PCs and all variables will not have proper representation. When sample variances of the original variables show differences by large magnitude, variables need to be normalized.

Let's Visualize data before scaling

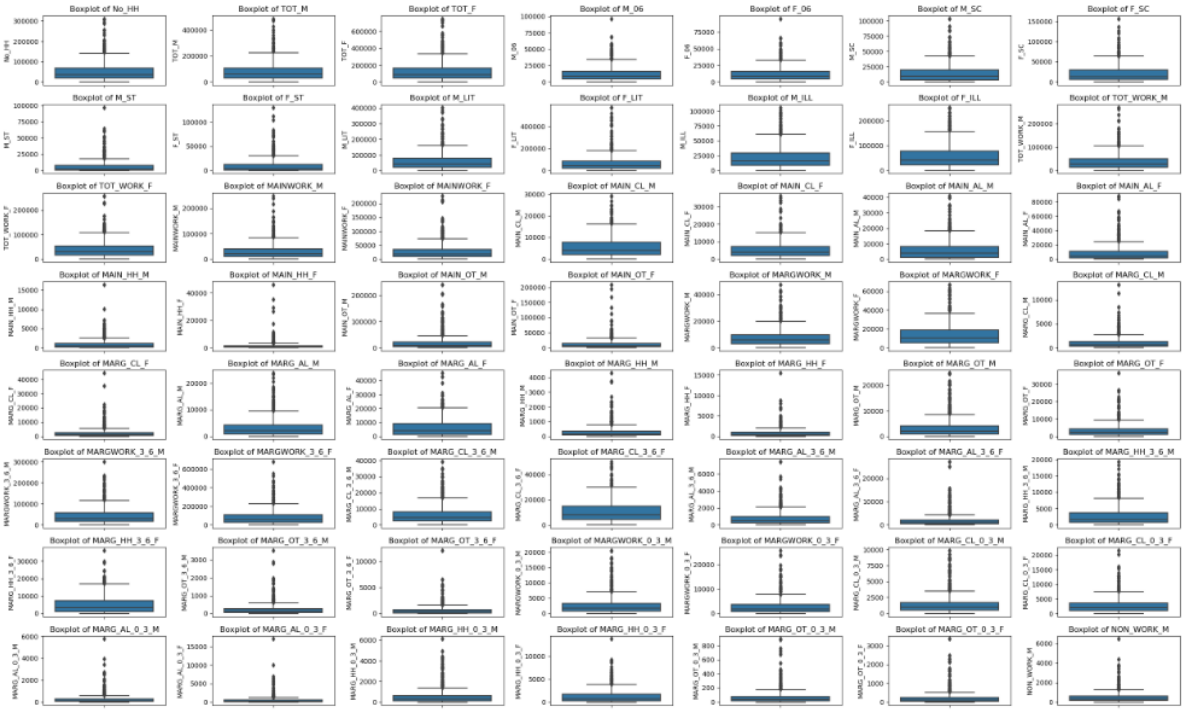


Figure 40 :Boxplot of Variables before scaling

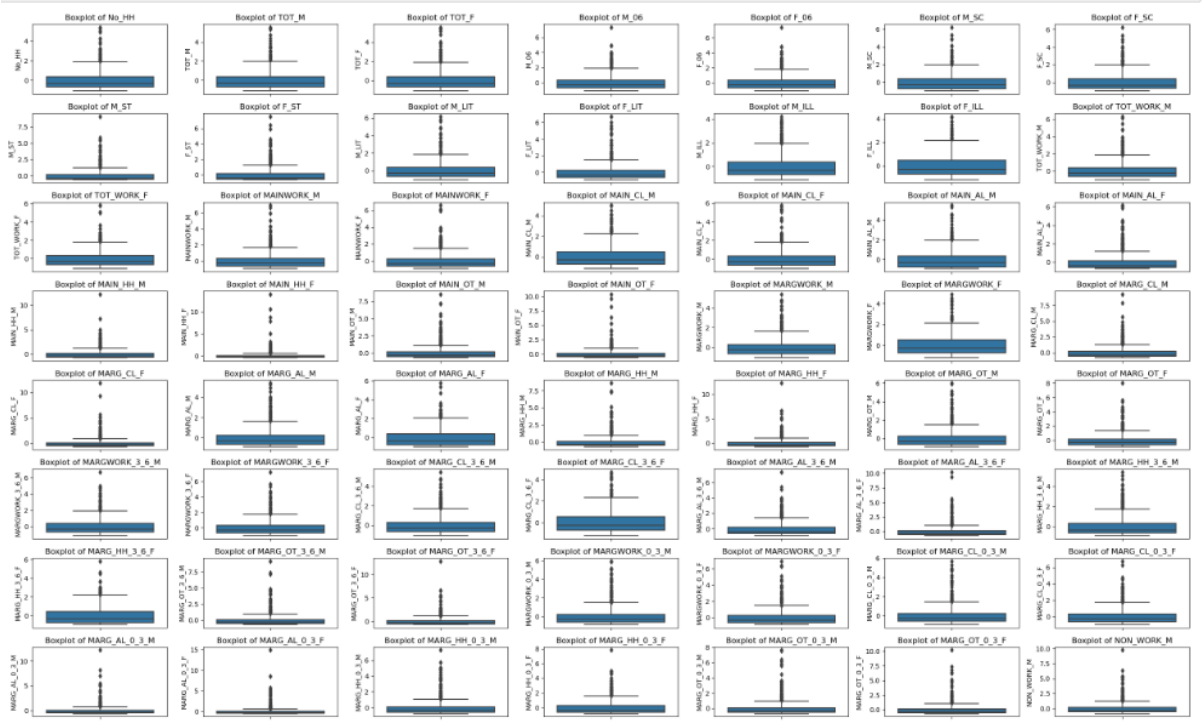


Figure 41: Boxplot of Variables after scaling

We used the Z score method to scale the data .

As we can see from the boxplot before and after scaling , there is no impact of scaling on the outliers.

Printing the Dataset after scaling

	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
0	-0.904738	-0.771236	-0.815563	-0.561012	-0.507738	-0.958575	-0.957049	-0.423306	-0.476423	-0.798097	-0.733477	-0.604015	-0.798229	-0.859260
1	-0.935695	-0.823100	-0.874534	-0.681096	-0.725367	-0.958297	-0.956772	-0.582014	-0.607607	-0.849434	-0.779797	-0.651213	-0.866645	-0.852143
2	-0.972412	-1.000919	-0.981466	-0.976956	-0.965262	-0.958575	-0.956772	-0.038951	-0.027273	-0.956457	-0.807151	-1.007596	-1.080898	-0.967749
3	-1.037530	-1.052224	-1.041001	-1.022118	-0.995393	-0.958783	-0.957049	-0.355965	-0.390060	-1.004643	-0.858872	-1.061609	-1.142070	-1.016469
4	-0.822676	-0.809381	-0.813933	-0.622359	-0.649908	-0.957395	-0.955529	0.149238	0.043330	-0.800568	-0.705296	-0.738239	-0.839181	-0.886905
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
635	-0.995677	-0.978990	-0.974268	-0.971387	-0.948916	-0.957326	-0.955667	-0.625124	-0.640197	-0.913820	-0.749215	-1.046667	-1.155813	-0.939362
636	-0.844340	-0.921822	-0.886965	-0.936754	-0.919757	-0.803806	-0.765670	-0.625124	-0.640197	-0.853390	-0.695320	-1.005476	-1.031152	-0.866542
637	-1.038465	-1.069066	-1.054885	-1.051356	-1.035331	-0.958783	-0.957049	-0.522953	-0.529880	-1.016367	-0.863674	-1.090887	-1.167899	-1.024356
638	-0.986758	-1.019276	-1.007472	-1.008195	-0.996541	-0.958783	-0.957049	-0.622297	-0.637046	-0.962328	-0.814713	-1.058984	-1.131556	-0.969617
639	-0.899166	-0.926854	-0.919050	-0.943193	-0.935220	-0.958783	-0.957049	-0.608870	-0.623555	-0.856916	-0.706830	-1.014159	-1.090180	-0.869648

Figure 42: Scaled Data

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

PCA is a method that: Measures how each variable is associated with one another using a Covariance matrix. Understands the directions of the spread of our data using Eigenvectors. Brings out the relative importance of these directions using Eigenvalues.

PCA uses this concept of eigen decomposition. We center the n predictors to their respective means and then get an *n* x *n* covariance matrix. This covariance matrix is then decomposed into eigenvalues and eigenvectors. So, a covariance matrix has variances (covariance of a predictor with itself) and covariances (between predictors).*Eigenvectors* are unit vectors with length or magnitude equal to 1. They are often referred to as right vectors, which simply means a column vector. *Eigenvalues* are coefficients applied to eigenvectors that give the vectors their length or magnitude

Here a snapshot of Covariance Matrix for the all 57 variables

	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_
No_HH	1.00	0.92	0.97	0.80	0.80	0.78	0.83	0.15	0.17	0.93	0.93	0.76	0.86	0.94	0.93	0.
TOT_M	0.92	1.00	0.98	0.95	0.95	0.84	0.83	0.09	0.09	0.99	0.93	0.91	0.89	0.97	0.81	0.
TOT_F	0.97	0.98	1.00	0.91	0.91	0.82	0.83	0.12	0.13	0.99	0.96	0.86	0.89	0.97	0.88	0.
M_06	0.80	0.95	0.91	1.00	1.00	0.78	0.75	0.06	0.04	0.91	0.83	0.95	0.86	0.86	0.68	0.
F_06	0.80	0.95	0.91	1.00	1.00	0.77	0.74	0.07	0.05	0.91	0.83	0.95	0.87	0.85	0.69	0.
M_SC	0.78	0.84	0.82	0.78	0.77	1.00	0.99	-0.05	-0.05	0.82	0.72	0.80	0.83	0.83	0.71	0.
F_SC	0.83	0.83	0.83	0.75	0.74	0.99	1.00	-0.01	-0.01	0.82	0.73	0.76	0.85	0.82	0.78	0.
M_ST	0.15	0.09	0.12	0.06	0.07	-0.05	-0.01	1.00	0.99	0.09	0.10	0.08	0.14	0.12	0.27	0.
F_ST	0.17	0.09	0.13	0.04	0.05	-0.05	-0.01	0.99	1.00	0.09	0.10	0.07	0.15	0.12	0.29	0.
M_LIT	0.93	0.99	0.99	0.91	0.91	0.82	0.82	0.09	0.09	1.00	0.97	0.84	0.84	0.98	0.82	0.
F_LIT	0.93	0.93	0.96	0.83	0.83	0.72	0.73	0.10	0.10	0.97	1.00	0.72	0.72	0.94	0.79	0.
M_ILL	0.76	0.91	0.86	0.95	0.95	0.80	0.76	0.08	0.07	0.84	0.72	1.00	0.93	0.84	0.69	0.
F_ILL	0.86	0.89	0.89	0.86	0.87	0.83	0.85	0.14	0.15	0.84	0.72	0.93	1.00	0.85	0.86	0.
TOT_WORK_M	0.94	0.97	0.97	0.86	0.85	0.83	0.82	0.12	0.12	0.98	0.94	0.84	0.85	1.00	0.84	0.
TOT_WORK_F	0.93	0.81	0.88	0.68	0.69	0.71	0.78	0.27	0.29	0.82	0.79	0.69	0.86	0.84	1.00	0.
MAINWORK_M	0.93	0.93	0.94	0.79	0.79	0.78	0.78	0.11	0.11	0.95	0.93	0.77	0.78	0.99	0.83	1.
MAINWORK_F	0.89	0.75	0.82	0.59	0.59	0.65	0.71	0.23	0.25	0.77	0.77	0.59	0.77	0.81	0.97	0.
MAIN_CL_M	0.43	0.53	0.49	0.56	0.56	0.61	0.58	0.10	0.08	0.47	0.33	0.65	0.66	0.50	0.48	0.
MAIN_CL_F	0.38	0.36	0.39	0.38	0.38	0.36	0.39	0.19	0.20	0.33	0.26	0.39	0.51	0.31	0.57	0.
MAIN_AL_M	0.67	0.59	0.62	0.55	0.56	0.63	0.67	0.14	0.15	0.54	0.45	0.66	0.79	0.60	0.70	0.
MAIN_AL_F	0.59	0.38	0.47	0.30	0.30	0.41	0.51	0.20	0.23	0.37	0.33	0.36	0.61	0.41	0.73	0.
MAIN_HH_M	0.64	0.74	0.70	0.66	0.66	0.71	0.68	-0.03	-0.03	0.73	0.64	0.69	0.68	0.76	0.57	0.
MAIN_HH_F	0.49	0.44	0.47	0.36	0.36	0.39	0.42	0.03	0.04	0.45	0.41	0.39	0.48	0.49	0.51	0.
MAIN_OT_M	0.85	0.85	0.86	0.69	0.68	0.64	0.64	0.09	0.08	0.90	0.93	0.61	0.60	0.92	0.72	0.
MAIN_OT_F	0.82	0.75	0.80	0.56	0.56	0.58	0.60	0.17	0.17	0.80	0.85	0.51	0.57	0.83	0.79	0.
MARGWORK_M	0.68	0.81	0.77	0.85	0.86	0.75	0.73	0.12	0.12	0.75	0.64	0.87	0.83	0.72	0.60	0.
MARGWORK_F	0.70	0.70	0.72	0.72	0.72	0.66	0.68	0.27	0.29	0.66	0.57	0.73	0.82	0.65	0.74	0.
MARG_CL_M	0.17	0.30	0.26	0.42	0.42	0.30	0.28	0.09	0.08	0.26	0.17	0.39	0.36	0.19	0.21	0.
MARG_CL_F	0.08	0.15	0.14	0.24	0.23	0.15	0.14	0.06	0.06	0.13	0.08	0.19	0.20	0.07	0.16	0.
MARG_AL_M	0.44	0.54	0.50	0.64	0.65	0.56	0.55	0.14	0.15	0.46	0.31	0.72	0.71	0.45	0.43	0.
MARG_AL_F	0.49	0.45	0.48	0.49	0.50	0.46	0.51	0.31	0.35	0.39	0.28	0.56	0.70	0.39	0.59	0.
MARG_HH_M	0.50	0.67	0.61	0.70	0.70	0.67	0.63	-0.01	-0.02	0.62	0.49	0.74	0.69	0.60	0.45	0.

Table 16Snapshot of the Covariance Matrix of all the numerical Variables

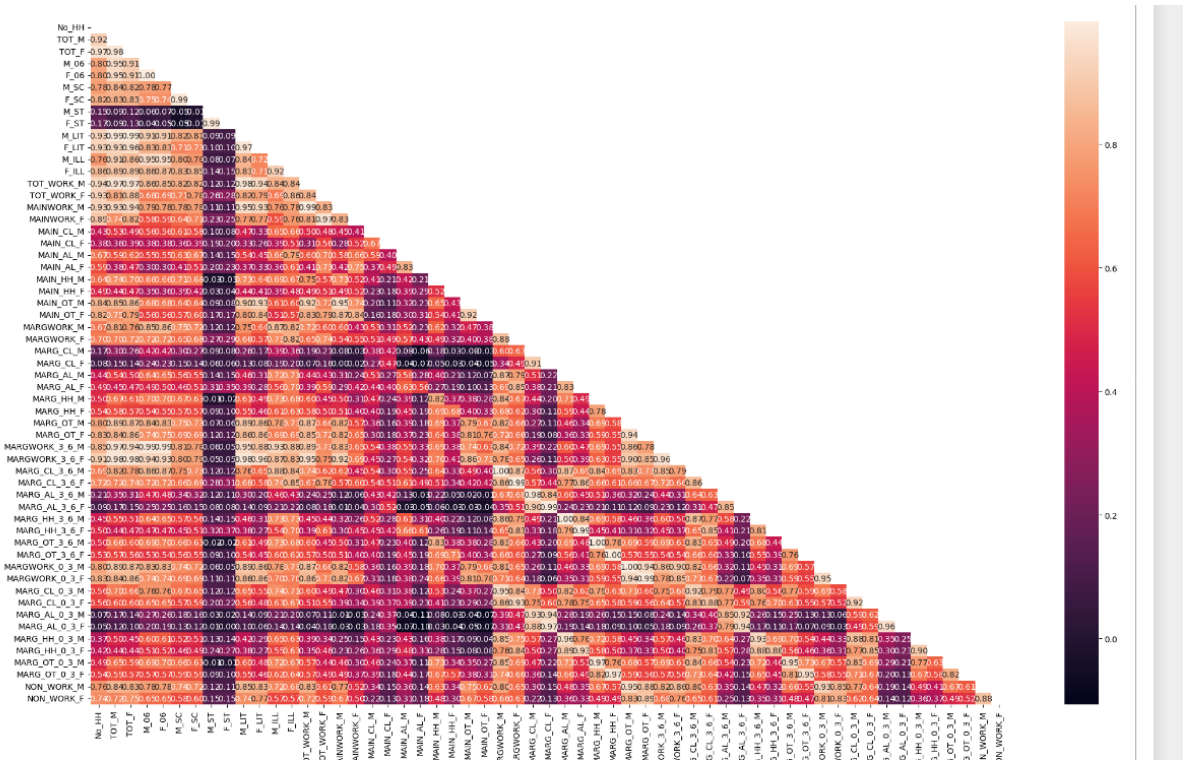


Figure 43:Correlation heatmap between 57 variables

The correlation matrix and heatmap indicate high correlation among some variables such as no of Households and No of Males./Females, No of working Males/Females. Moderate correlation may be detected between several pairs of variables, such as; between number of Males/Females and Main Work population etc. Existence of such pairs of high and moderate correlations indicate that dimension reduction must be considered for the data.

There are 57 variables , hence we get 57 Principal components

Following is a snapshot of the Eigen Vectors for all 57 Principal components .Since it's not possible to list all the 57 eigen vectors for 57 PCA, below is just a snapshot

```
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]])
```

Table 17:Eigen Vectors

Following are the 57 Eigen values for all the 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])
```

Table 18:Eigen Values

In the above output, eigenvectors give the PCA components and eigenvalues give the explained variances of the components. The first few PC components explain the maximum variance .If there are eigenvalues close to zero, they represent components that may be discarded because they don't explain much of the data.

Let's draw the Scree Plot with PCA =57

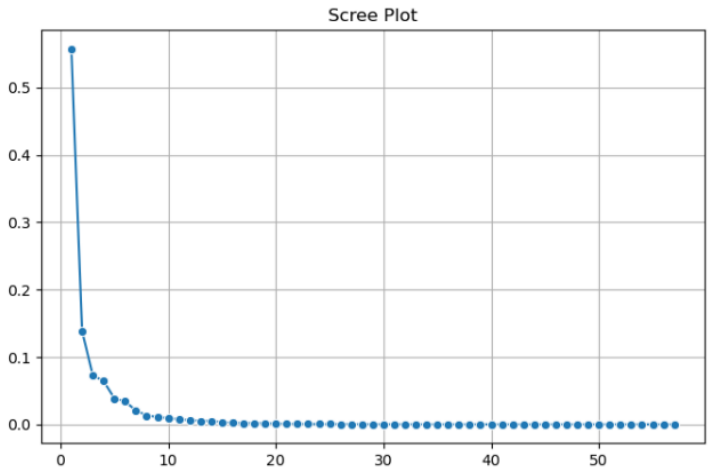


Figure 44:Scree Plot with 57 Principal Components

Above Scree plot is a graphic that shows the explained variance per principal component. The measure of the plot can be the percentage or the absolute value of the explained variance (eigenvalues). The first few principal components explain the major amount of variance.

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

The scree plot is a useful visual tool to select optimum number of PC. On the X-axis are shown the indices of the PCs and on the Y-axis are shown the variances. If there is a distinct break point in the line joining the variances (elbow point) beyond which the line becomes approximately horizontal, then that point may be taken as the value of k, provided other conditions are also satisfied.

Since its not feasible to have 57 PC's and first few components explain the maximum variance let's take number of PC's as 10 and get the variances of each of them . Let's get the variances ratio of each component and the cumulative variance . As we can see PC1(first value in the array) explains maximum variance with 31.8 as variance

```
array([31.81356474,  7.86942415,  4.15340812,  3.66879058,  2.20652588,
        1.93827502,  1.17617374,  0.75115909,  0.61705374,  0.52830088])
```

Figure 45 PCA Variance with 10 components

Let's get the explained variance ratios. Explained Variance Ratio is explained variance of component / (total of all explained variances).Here PC1 has ratio of 55.7

```
array([0.55726063, 0.13784435, 0.07275295, 0.06426418, 0.03865049,
       0.03395169, 0.02060239, 0.01315764, 0.01080859, 0.00925395])
```

Figure 46 PCA Variance Ratio with 10 components

Let's get the cumulative variance ratios.PC1 has maximum ratio of 55.7. PC 1 and 2 together explain 69.5% variance .PC1 ,2 and 3 explain 76.7% variance and so on.

```
array([0.55726063, 0.69510499, 0.76785794, 0.83212212, 0.87077261,
       0.9047243 , 0.92532669, 0.93848433, 0.94929292, 0.95854687])
```

Figure 47:Cumulative Variance of the Components

Let plot a scree plot of PC components equal to 10

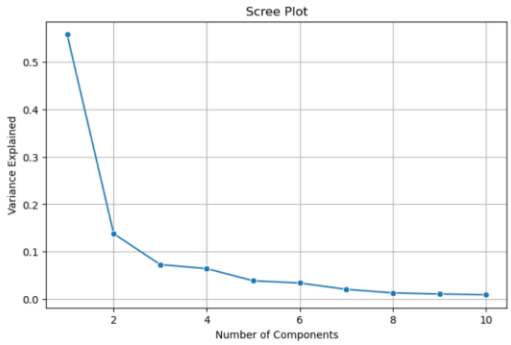


Figure 48: Scree Plot with 10 Principal Components

In Fig 48, there is a distinct break at 2. However, k cannot be taken to be 2 since the first two PCs explain only 69.5% of total variance. The PCs must be taken so as to explain at least 90% of the total variance. If k=6, then the first 6 PCs explain 90.4% of the total variance. One choice of k could be 6. Also, from the scree plot we can see that after 6 Principal components , the Variance explained isn't changing much , so we can select 6 PCs to represent the entire dataset

	PCs	Proportion Of Variance	Standard Deviation	Cumulative Proportion
0	PC1	0.56	5.64	0.56
1	PC2	0.14	2.81	0.70
2	PC3	0.07	2.04	0.77
3	PC4	0.06	1.92	0.83
4	PC5	0.04	1.49	0.87
5	PC6	0.03	1.39	0.90
6	PC7	0.02	1.08	0.93
7	PC8	0.01	0.87	0.94
8	PC9	0.01	0.79	0.95
9	PC10	0.01	0.73	0.96

Table 19 Variances and Standard Deviations of each PC

The principal components are constructed in decreasing order of magnitude of their standard deviations, which is equivalent to decreasing order of magnitude of their variances. In Table 19 the variances of the constructed principal components and their sum total is given.

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.

Since we have selected 6 as optimum number of principal components , the new dataset will now be represented by these 6 variables instead of 57 variables . Principal components are linear combinations of the original variables or scaled variables, as the case may be. It is possible that some of the coefficients are very small numbers or close to 0. We present the linear combinations that make up the first 6 PC's

Let's have a look at the PCA loadings of each of these PC components with each of the columns

	PC1	PC2	PC3	PC4	PC5	PC6
No_HH	0.156021	-0.126347	-0.002690	-0.125293	-0.007022	0.004083
TOT_M	0.167118	-0.089677	0.056698	-0.019942	-0.033026	-0.073389
TOT_F	0.165553	-0.104912	0.038749	-0.070873	-0.012847	-0.043647
M_06	0.162193	-0.022095	0.057788	0.011917	-0.050248	-0.157957
F_06	0.162566	-0.020271	0.050126	0.014844	-0.043848	-0.154436
M_SC	0.151358	-0.045111	0.002569	0.012485	-0.173007	-0.064295
F_SC	0.151567	-0.051924	-0.025101	-0.029893	-0.159803	-0.040518
M_ST	0.027234	0.027679	-0.123504	-0.222247	0.433163	0.222591
F_ST	0.028183	0.030223	-0.139769	-0.229754	0.438792	0.225531
M_LIT	0.161993	-0.115355	0.082168	-0.035163	-0.009101	-0.055465
F_LIT	0.146873	-0.153109	0.117098	-0.059559	0.055844	-0.048021
M_ILL	0.161749	-0.006625	-0.021855	0.025348	-0.096580	-0.115234
F_ILL	0.165248	-0.009107	-0.093062	-0.076023	-0.119911	-0.028757
TOT_WORK_M	0.159872	-0.133529	0.045176	-0.040154	-0.019553	-0.001801
TOT_WORK_F	0.145936	-0.085087	-0.059450	-0.225160	-0.040437	0.105162
MAINWORK_M	0.146201	-0.176368	0.054295	-0.068351	-0.036802	0.019283
MAINWORK_F	0.123970	-0.151413	-0.055609	-0.246640	-0.082834	0.123832
MAIN_CL_M	0.103127	0.062415	-0.067399	-0.089769	-0.286039	-0.006170
MAIN_CL_F	0.074540	0.086477	-0.009238	-0.288965	-0.241936	0.102951
MAIN_AL_M	0.113356	-0.031040	-0.247917	-0.136082	-0.205723	-0.031068
MAIN_AL_F	0.073882	-0.058688	-0.251932	-0.290042	-0.177605	0.019240
MAIN_HH_M	0.131573	-0.076021	0.026569	0.152366	-0.134089	0.174465
MAIN_HH_F	0.083383	-0.082477	-0.060523	0.048950	-0.139441	0.422309
MAIN_OT_M	0.123526	-0.212984	0.137378	-0.040289	0.064638	0.023477
MAIN_OT_F	0.111021	-0.210071	0.095634	-0.120391	0.080743	0.083079
MARGWORK_M	0.164615	0.092994	-0.008628	0.093018	0.060244	-0.090762
MARGWORK_F	0.155396	0.125270	-0.049370	-0.088707	0.089202	0.017868
MARG_AL_M	0.128599	0.165831	-0.189868	0.091787	0.019422	-0.141605
MARG_AL_F	0.114305	0.140958	-0.267768	-0.106365	0.080527	-0.085120
MARG_HH_M	0.140853	0.068068	-0.021257	0.237985	-0.059971	0.089533
MARG_HH_F	0.127670	0.024216	-0.082504	0.196321	-0.033602	0.365112
MARG_OT_M	0.155263	-0.089442	0.111713	0.087119	0.119121	-0.061066
MARG_OT_F	0.147287	-0.117899	0.100046	0.026729	0.166882	0.001739
MARGWORK_3_6_M	0.164972	-0.043995	0.064423	-0.000026	-0.043834	-0.136253
MARGWORK_3_6_F	0.161253	-0.105502	0.079704	0.003894	0.000537	-0.106900
MARG_CL_3_6_M	0.165502	0.077193	-0.024205	0.092875	0.054073	-0.096708
MARG_CL_3_6_F	0.155647	0.103174	-0.072013	-0.107860	0.073050	0.023773
MARG_AL_3_6_M	0.093014	0.264409	0.153518	-0.038488	-0.007789	0.013477
MARG_AL_3_6_F	0.051536	0.244261	0.256213	-0.179691	-0.061303	0.093993
MARG_HH_3_6_M	0.128576	0.158783	-0.200119	0.080411	0.008457	-0.144061
MARG_HH_3_6_F	0.110646	0.125287	-0.279866	-0.136240	0.064109	-0.076709
MARG_OT_3_6_M	0.139593	0.062262	-0.020618	0.237745	-0.066400	0.097058
MARG_OT_3_6_F	0.124546	0.014766	-0.082794	0.190511	-0.044810	0.384552
MARGWORK_0_3_M	0.154294	-0.093159	0.110285	0.086479	0.108829	-0.062043
MARGWORK_0_3_F	0.146286	-0.125596	0.095667	0.027275	0.141190	0.008962
MARG_CL_0_3_M	0.150126	0.150681	0.054892	0.087433	0.081185	-0.060715
MARG_CL_0_3_F	0.140157	0.180690	0.023982	-0.022290	0.129936	-0.001727
MARG_AL_0_3_M	0.052542	0.251328	0.268330	-0.104686	-0.048849	0.065409
MARG_AL_0_3_F	0.041786	0.240720	0.284956	-0.135716	-0.051895	0.083743
MARG_HH_0_3_M	0.121840	0.185277	-0.138628	0.132544	0.062380	-0.124209
MARG_HH_0_3_F	0.116011	0.180616	-0.202198	0.004051	0.128308	-0.105530
MARG_OT_0_3_M	0.139869	0.084869	-0.022599	0.230038	-0.036390	0.061228
MARG_OT_0_3_F	0.132192	0.050813	-0.078720	0.206201	0.000165	0.295600
NON_WORK_M	0.150376	-0.065365	0.111827	0.084854	0.162862	-0.052387
NON_WORK_F	0.131066	-0.073847	0.102553	0.021124	0.238292	-0.024901

Here is a Graphical representation of Principal components with the corresponding loadings for each column of the dataset

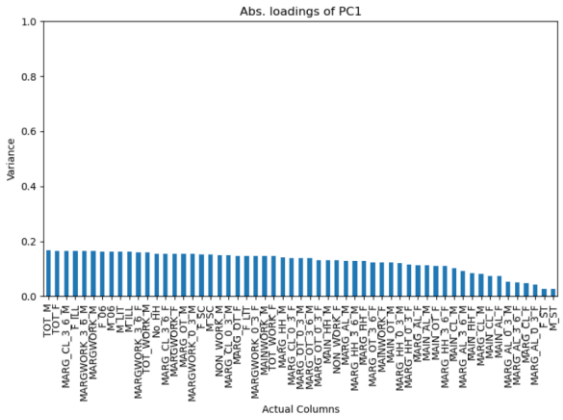


Figure 49: PCA Loading of PC1

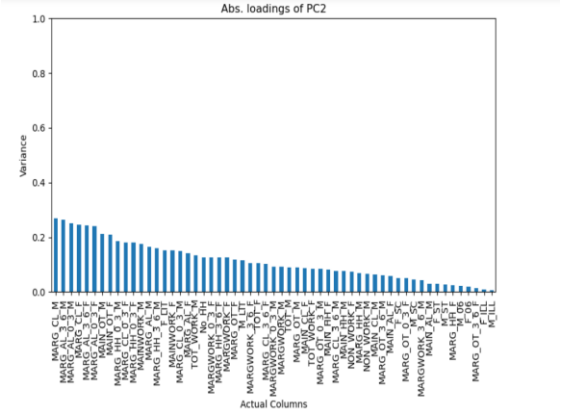


Figure 50: PCA Loading of PC2

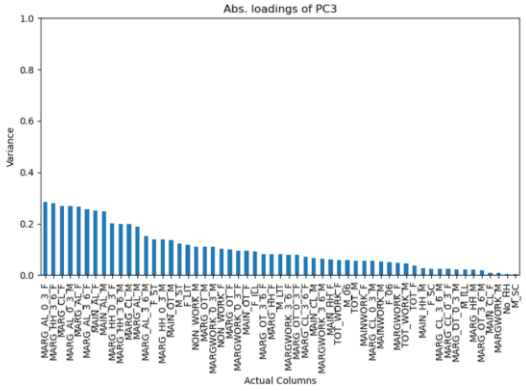


Figure 51: PCA Loading of PC3

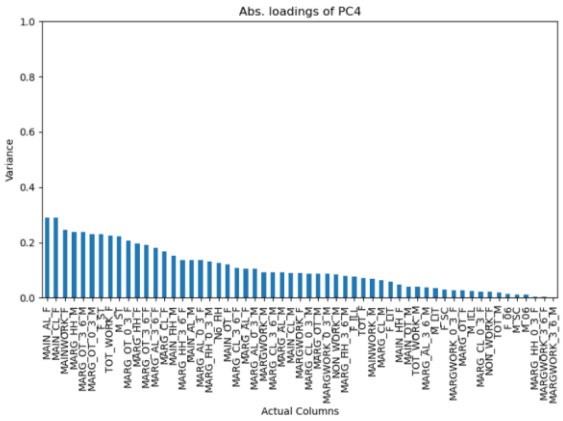
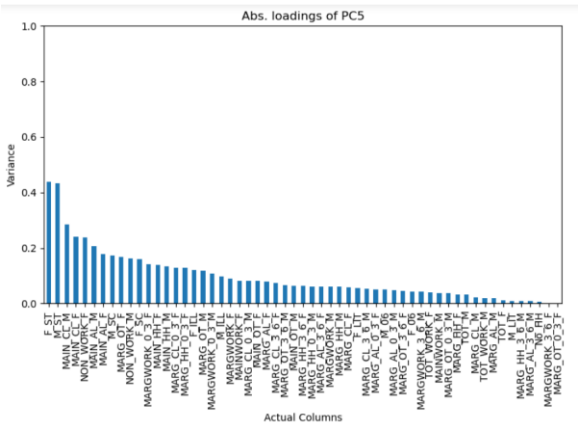


Figure 52: PCA Loading of PC4



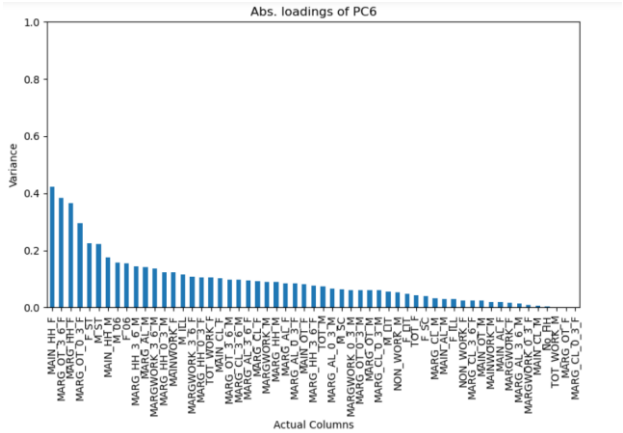


Figure 53: PCA Loading of PC5 and PC6

Heatmap of PC loading shows the correlation of PC with actual columns

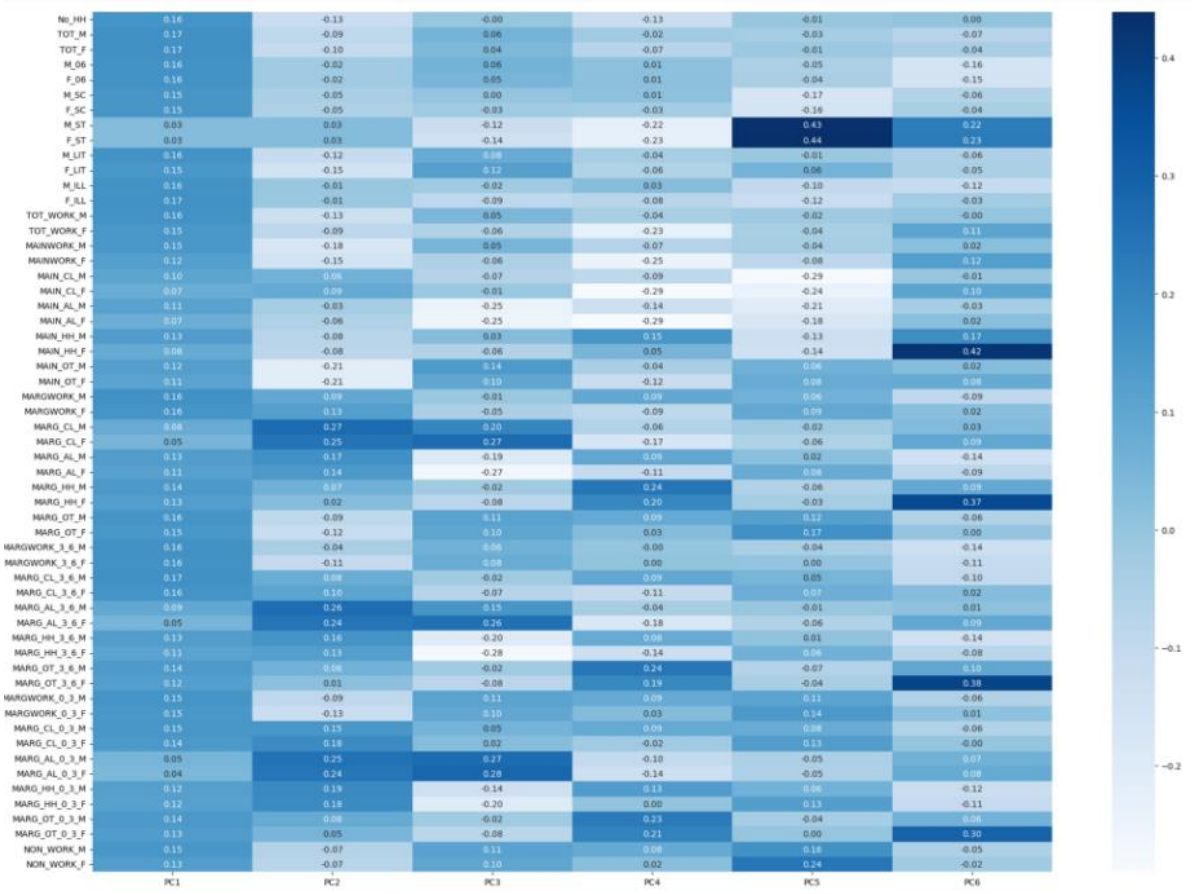


Figure 54:Correlation map of Principal Components and Actual Columns

PC1 is showing the positive correlation with many variables

PC2 is showing the very high correlation with variables MAIN_OT_M,MAIN_OT_F,MARG_CL_M,MARG_CL_F,MARG_AL_3_6_M,MARG_AL_3_6_F,MARG_AL_0_3_M,MARG_AL_0_3_F representing Main Other Workers Population Male and Female, Marginal Cultivator Population Male and Female, Marginal Agriculture Labourers Population 3-6 Male and Female, Marginal Agriculture Labourers Population 0-3 Male and Female

PC3 is showing the maximum correlation with variables MAIN_AL_M,MAIN_AL_F,MARG_CL_F,MARG_AL_3_6_F,MARG_HH_3_6_F,MARG_AL_0_3_M,MARG_AL_0_3_F indicating columns representing Main Agricultural Labourers Population Male and Female and
Marginal Agriculture Labourers Population 0-3 Male and Female, Marginal Household Industries Population 3-6 Female, Marginal Agriculture Labourers Population 3-6 Female, Marginal Cultivator Population Female

PC4 is showing the maximum correlation with variables MAIN_AL_F,MAIN_CL_F, MARG_OT_3_6_M , MARG_OT_0_3_M , MARG_HH_M,MAIN_WORK_F,TOT_WORK_F indicating columns representing Main Agricultural Labourers Population Female and Main Cultivator Population Female, Main Working Population Female, Total Worker Population Female

PC5 is showing the maximum correlation with variables M_ST, F_ST , MAIN_CL_M, MAIN_CL_F , M_SC, F_SC indicating a column representing Scheduled Tribes Population , Scheduled Caste Population and Cultivator Population.

PC6 is showing the maximum correlation with variables MAIN_HH_F,MARG_OT_3_6_F,MARG_HH_F,MARG_OT_0_3_F indicating a column representing Marginal Working Population

Principal components are linear combinations of the original variables. Each PC is a linear combination of all variables, or scaled variables, as the case may be .Once the original variables are replaced by the PCs, the latter are used for any further analysis. Just as each observed unit has a particular value of each variable, similarly each observation has a particular value for each PC. These values are called PC scores. These scores are obtained by putting scaled values of the variables in the expression of PCs as shown below

	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

Table 20 PC scores

To check that the PCs are orthogonal, correlation matrix is computed. As can be seen the correlation between the PC's is 0, indicating there is no correlation between them.

	PC1	PC2	PC3	PC4	PC5	PC6
PC1	1.0	0.0	-0.0	-0.0	0.0	-0.0
PC2	0.0	1.0	0.0	-0.0	-0.0	-0.0
PC3	-0.0	0.0	1.0	0.0	-0.0	0.0
PC4	-0.0	-0.0	0.0	1.0	0.0	0.0
PC5	0.0	-0.0	-0.0	0.0	1.0	-0.0
PC6	-0.0	-0.0	0.0	0.0	-0.0	1.0

Table 21:Correlation between PC's

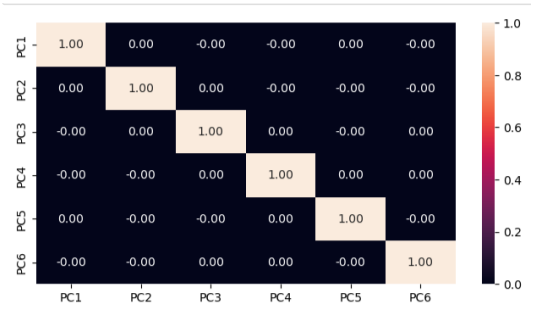


Figure 55:Heatmap of PC's

8.PCA: Write linear equation for first PC.

Formula for 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.

For each PC, these are the coefficients with which the corresponding variables need to be multiplied to get the PC. Note that the weights can be positive or negative

	PC1	PC2	PC3	PC4	PC5	PC6
No_HH	0.156021	-0.126347	-0.002690	-0.125293	-0.007022	0.004083
TOT_M	0.167118	-0.089677	0.056698	-0.019942	-0.033026	-0.073389
TOT_F	0.165553	-0.104912	0.038749	-0.070873	-0.012847	-0.043647
M_06	0.162193	-0.022095	0.057788	0.011917	-0.050248	-0.157957
F_06	0.162566	-0.020271	0.050126	0.014844	-0.043848	-0.154436
M_SC	0.151358	-0.045111	0.002569	0.012485	-0.173007	-0.064295
F_SC	0.151567	-0.051924	-0.025101	-0.029893	-0.159803	-0.040518
M_ST	0.027234	0.027679	-0.123504	-0.222247	0.433163	0.222591
F_ST	0.028183	0.030223	-0.139769	-0.229754	0.438792	0.225531
M_LIT	0.161993	-0.115355	0.082168	-0.035163	-0.009101	-0.055465
F_LIT	0.146873	-0.153109	0.117098	-0.059559	0.055844	-0.048021
M_ILL	0.161749	-0.006625	-0.021855	0.025348	-0.096580	-0.115234
F_ILL	0.165248	-0.009107	-0.093062	-0.076023	-0.119911	-0.028757
TOT_WORK_M	0.159872	-0.133529	0.045176	-0.040154	-0.019553	-0.001801
TOT_WORK_F	0.145936	-0.085087	-0.059450	-0.225160	-0.040437	0.105162
MAINWORK_M	0.146201	-0.176368	0.054295	-0.068351	-0.036802	0.019283
MAINWORK_F	0.123970	-0.151413	-0.055609	-0.246640	-0.082834	0.123832
MAIN_CL_M	0.103127	0.062415	-0.067399	-0.089769	-0.286039	-0.006170
MAIN_CL_F	0.074540	0.086477	-0.009238	-0.288965	-0.241936	0.102951
MAIN_AL_M	0.113356	-0.031040	-0.247917	-0.136082	-0.205723	-0.031068
MAIN_AL_F	0.073882	-0.058688	-0.251932	-0.290042	-0.177605	0.019240
MAIN_HH_M	0.131573	-0.076021	0.026569	0.152366	-0.134089	0.174465
MAIN_HH_F	0.083383	-0.082477	-0.060523	0.048950	-0.139441	0.422309
MAIN_OT_M	0.123526	-0.212984	0.137378	-0.040289	0.064638	0.023477
MAIN_OT_F	0.111021	-0.210071	0.095634	-0.120391	0.080743	0.083079
MARGWORK_M	0.164615	0.092994	-0.008628	0.093018	0.060244	-0.090762
MARGWORK_F	0.155396	0.125270	-0.049370	-0.088707	0.089202	0.017868

Using the above weights, we arrive at the linear equation for PC1 as below . Note the coefficient have been rounded off to 2 decimal places

```
( 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 +
```

Table 22: linear equation for first PC.