Great Learning's

Machine Learning 1

PROJECT REPORT

By

Raghvendra Singh

Email: raghavsingh0027@gmail.com

Phone: +91-8130670022

Table of Contents

Data & Dictionary

Problem 1 Data Dictionary Clustering

Sl. No	Column Name	Column Description
1	Timestamp	The Timestamp of the particular Advertisement.
2	InventoryType	The Inventory Type of the particular Advertisement. Format 1 to 7. This is a Categorical Variable.
3	Ad - Length	The Length Dimension of the particular Adverstisement.
4	Ad- Width	The Width Dimension of the particular Advertisement.
5	Ad Size	The Overall Size of the particular Advertisement. Length*Width.
6	Ad Type	The type of the particular Advertisement. This is a Categorical Variable.
7	Platform	The platform in which the particular Advertisement is displayed. Web, Video or App. This is a Categorical Variable.
8	Device Type	The type of the device which supports the partciular Advertisement. This is a Categorical Variable.
9	Format	The Format in which the Advertisement is displayed. This is a Categorical Variable.
10	Available_Impressions	How often the particular Advertisement is shown. An impression is counted each time an Advertisement is shown on a search result page or other site on a Network.
11	Matched_Queries	Matched search queries data is pulled from Advertising Platform and consists of the exact searches typed into the search Engine that generated clicks for the particular Advertisement.
12	Impressions	The impression count of the particular Advertisement out of the total available impressions.
13	Clicks	It is a marketing metric that counts the number of times users have clicked on the particular advertisement to reach an online property.
14	Spend	It is the amount of money spent on specific ad variations within a specific campaign or ad set. This metric helps regulate ad performance.
15	Fee	The percentage of the Advertising Fees payable by Franchise Entities.
16	Revenue	It is the income that has been earned from the particular advertisement.
17	CTR	CTR stands for "Click through rate". CTR is the number of clicks that your ad receives divided by the number of times your ad is shown. Formula used here is CTR = Total Measured Clicks / Total Measured Ad Impressions x 100. Note that the Total Measured Clicks refers to the 'Clicks' Column and the Total Measured Ad Impressions refers to the 'Impressions' Column.

18	CPM	CPM stands for "cost per 1000 impressions." Formula used here is CPM = (Total Campaign Spend / Number of Impressions) * 1,000. Note that the Total Campaign Spend refers to the 'Spend' Column and the Number of Impressions refers to the 'Impressions' Column.
19	CPC	CPC stands for "Cost-per-click". Cost-per-click (CPC) bidding means that you pay for each click on your ads. The Formula used here is CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column and the Number of Clicks refers to the 'Clicks' Column.

Problem 4 Data Dictionary

Name	Description
State	State Code
District	District Code
Name	Name
TRU1	Area Name
No_HH	No of Household
TOT_M	Total population Male
TOT_F	Total population Female
M_06	Population in the age group 0-6 Male
F_06	Population in the age group 0-6 Female
M_SC	Scheduled Castes population Male
F_SC	Scheduled Castes population Female
M_ST	Scheduled Tribes population Male
F_ST	Scheduled Tribes population Female
M_LIT	Literates population Male
F_LIT	Literates population Female
M_ILL	Illiterate Male
F_ILL	Illiterate Female
TOT_WORK_M	Total Worker Population Male
TOT_WORK_F	Total Worker Population Female
MAINWORK_M	Main Working Population Male
MAINWORK_F	Main Working Population Female
MAIN_CL_M	Main Cultivator Population Male
MAIN_CL_F	Main Cultivator Population Female
MAIN_AL_M	Main Agricultural Labourers Population Male
MAIN_AL_F	Main Agricultural Labourers Population Female
MAIN_HH_M	Main Household Industries Population Male
MAIN_HH_F	Main Household Industries Population Female
MAIN_OT_M	Main Other Workers Population Male
MAIN_OT_F	Main Other Workers Population Female
MARGWORK_M	Marginal Worker Population Male
MARGWORK_F	Marginal Worker Population Female
MARG_CL_M	Marginal Cultivator Population Male

MARG_CL_F	Marginal Cultivator Population Female
MARG_AL_M	Marginal Agriculture Labourers Population Male
MARG_AL_F	Marginal Agriculture Labourers Population Female
MARG_HH_M	Marginal Household Industries Population Male
MARG_HH_F	Marginal Household Industries Population Female
MARG_OT_M	Marginal Other Workers Population Male
MARG_OT_F	Marginal Other Workers Population Female
MARGWORK_3_6_M	Marginal Worker Population 3-6 Male
MARGWORK_3_6_F	Marginal Worker Population 3-6 Female
MARG_CL_3_6_M	Marginal Cultivator Population 3-6 Male
MARG_CL_3_6_F	Marginal Cultivator Population 3-6 Female
MARG_AL_3_6_M	Marginal Agriculture Labourers Population 3-6 Male
MARG_AL_3_6_F	Marginal Agriculture Labourers Population 3-6 Female
MARG_HH_3_6_M	Marginal Household Industries Population 3-6 Male
MARG_HH_3_6_F	Marginal Household Industries Population 3-6 Female
MARG_OT_3_6_M	Marginal Other Workers Population Person 3-6 Male
MARG_OT_3_6_F	Marginal Other Workers Population Person 3-6 Female
MARGWORK_0_3_M	Marginal Worker Population 0-3 Male
MARGWORK_0_3_F	Marginal Worker Population 0-3 Female
MARG_CL_0_3_M	Marginal Cultivator Population 0-3 Male
MARG_CL_0_3_F	Marginal Cultivator Population 0-3 Female
MARG_AL_0_3_M	Marginal Agriculture Labourers Population 0-3 Male
MARG_AL_0_3_F	Marginal Agriculture Labourers Population 0-3 Female
MARG_HH_0_3_M	Marginal Household Industries Population 0-3 Male
MARG_HH_0_3_F	Marginal Household Industries Population 0-3 Female
MARG_OT_0_3_M	Marginal Other Workers Population 0-3 Male
MARG_OT_0_3_F	Marginal Other Workers Population 0-3 Female
NON_WORK_M	Non Working Population Male
NON_WORK_F	Non Working Population Female

List of Figures

PROBLEM 1

Part 1: Clustering: Define the problem and perform Exploratory Data Analysis

- Problem definition - Check shape, Data types, statistical summary - Univariate analysis - Bivariate analysis - Key meaningful observations on individual variables and the relationship between variables

Problem Definition:

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:

- Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.
- Treat missing values in CPC, CTR and CPM using the formula given. You may refer to the Bank_KMeans Solution File to understand the coding behind treating the missing values using a specific formula. You have to basically create an user defined function and then call the function for imputing.
- Check if there are any outliers.

- Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an analyst your judgement may be different from another analyst).
- Perform z-score scaling and discuss how it affects the speed of the algorithm.
- Perform clustering and do the following:
    1. Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.
    2. Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.
    3. Print silhouette scores for up to 10 clusters and identify optimum number of clusters.
    4. 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.]
    5. Conclude the project by providing summary of your learnings.

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

Head:

	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 1 Ads Data Head

Tail:

	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Sp
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 2 Ads Data Tail

Info:

```
<class 'pandas.core.frame.DataFrame'>
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)
memory usage: 3.3+ MB
```

Figure 3 Ads Data Info

The data set has 13 numerical and 6 categorical values. The Timestamp is not relevant data for analysis.

Shape:

```
(23066, 19)
```

There are 23066 columns and 19 rows in data set.

Columns:

```
Index(['Timestamp', 'InventoryType', 'Ad - Length', 'Ad- Width', 'Ad Si
ze',
       'Ad Type', 'Platform', 'Device Type', 'Format', 'Available_Impre
ssions',
       'Matched_Queries', 'Impressions', 'Clicks', 'Spend', 'Fee', 'Rev
enue',
       'CTR', 'CPM', 'CPC'],
      dtype='object')
```

The columns are described already in the Data dictionary.

Device Type:

There are two device types being used for ads. Mobile is most used compared to desktop to display ads. While planning any campaign this information must be kept in mind.

```
Device Type
Mobile     14806
Desktop     8260
Name: count, dtype: int64
```

Figure 4 Device types

Platform Type:

There are three major platforms being used. Vedio is most preffered platform. Web and App are other significant platforms.

```
Platform
Video     9873
Web       8251
App       4942
Name: count, dtype: int64
```

Figure 5 Platform Type

Describe:

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	C
count	23066.00	23066.00	23066.00	23066.00	23066.00	23066.00	23066.00	23066.00	23066.00	23066.00	18330.00	18330.00	1833(
mean	385.16	337.90	96674.47	2432043.67	1295099.14	1241519.52	10678.52	2706.63	0.34	1924.25	0.07	7.67	(
std	233.65	203.09	61538.33	4742887.76	2512969.86	2429399.96	17353.41	4067.93	0.03	3105.24	0.08	6.48	(
min	120.00	70.00	33600.00	1.00	1.00	1.00	1.00	0.00	0.21	0.00	0.00	0.00	(
25%	120.00	250.00	72000.00	33672.25	18282.50	7990.50	710.00	85.18	0.33	55.37	0.00	1.71	(
50%	300.00	300.00	72000.00	483771.00	258087.50	225290.00	4425.00	1425.12	0.35	926.34	0.08	7.66	(
75%	720.00	600.00	84000.00	2527711.75	1180700.00	1112428.50	12793.75	3121.40	0.35	2091.34	0.13	12.51	(
max	728.00	600.00	216000.00	27592861.00	14702025.00	14194774.00	143049.00	26931.87	0.35	21276.18	1.00	81.56	7

Figure 6 Description of Ads data

The statistical summary explains that:

1. Average ad length is 385.16 and width is 337.90
2. Mean CTR is 0.7 which is good as maximum is 1
3. Average Spend on campaigns is 2706.63 which is far below the maximum spend of 26931

Univariate Analysis:

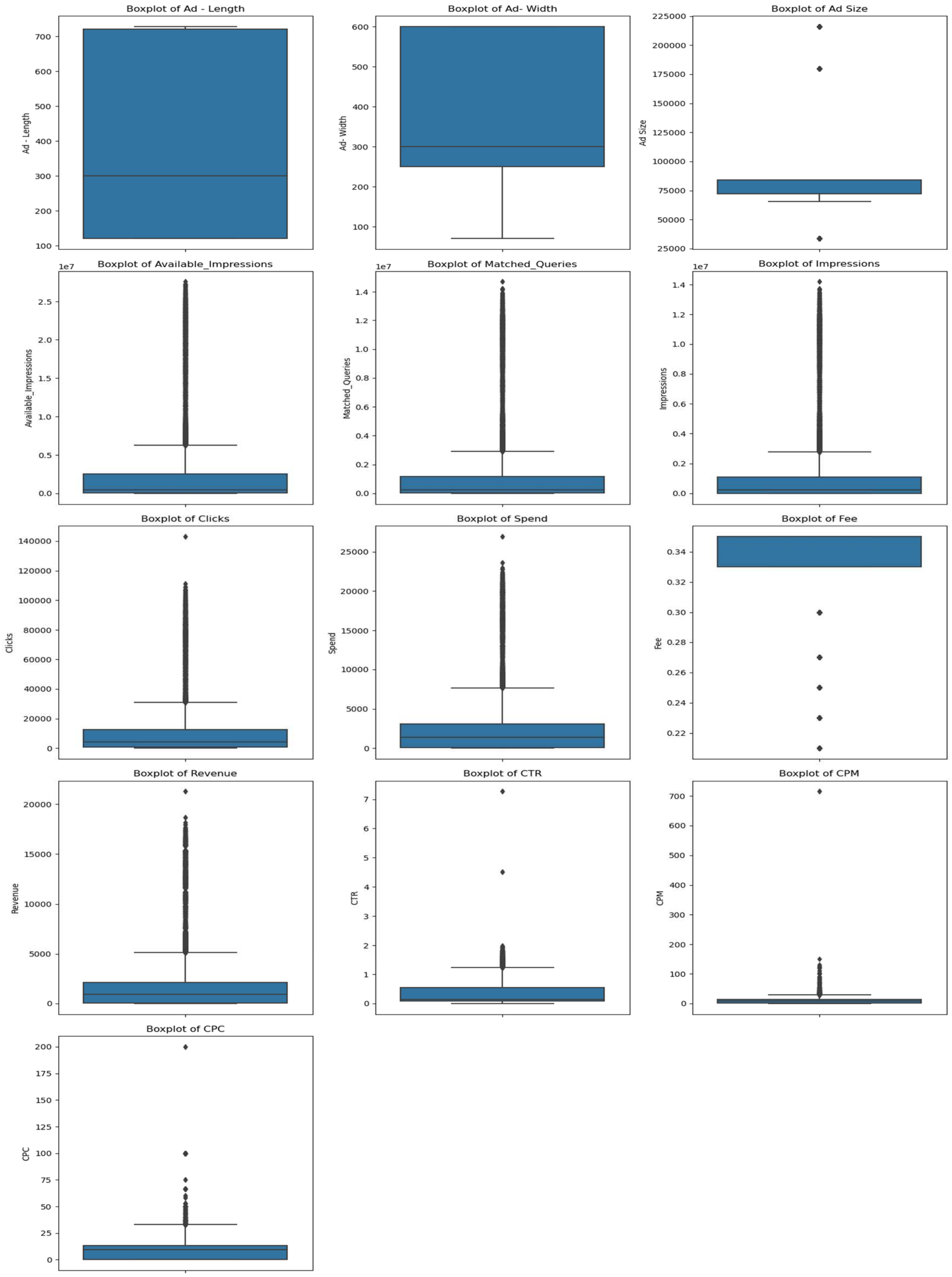


Figure 7 Ads Data Boxplot

- No outliers are present in ad length and ad width.
- Significant outliers are present in all other numerical field and needs to be treated.
- Fee Boxplot is mostly above .32 value
- Range of CPM value is very small and needs to be scaled in later stages to get comparable information

- Impressions, Clicks, Spends have large number of outliers present

Correlation Matrix:

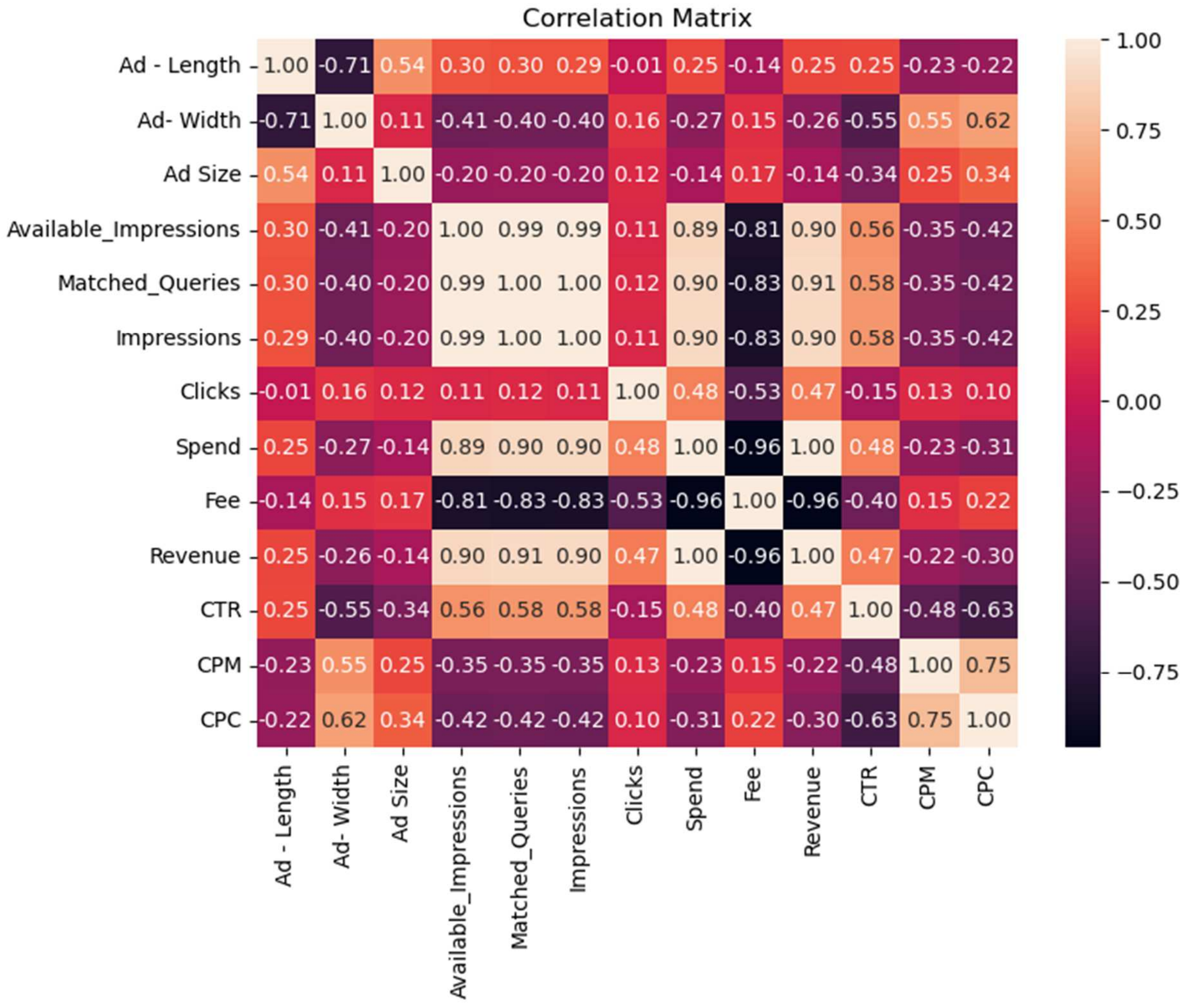


Figure 8 Correlation Matrix of Ads Data

If we create a correlation Matrix we can find below observations:

- Fee has very high negative correlation with spend and revenue -0.96
- CPM and CPC have very high positive correlation 0.75
- Ad width and CPM have decent correlation of 0.55
- Ad width and CPC has good correlation of 0.62
- Impressions and Fees have high negative correlation of -0.83

Bivariate Analysis:

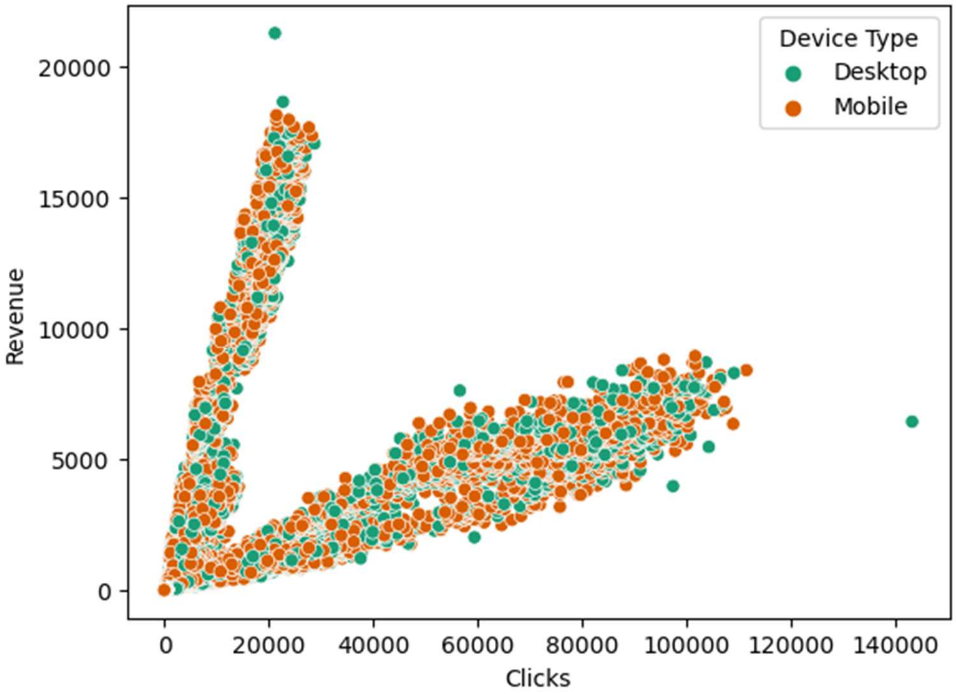


Figure 9 Revenue vs clicks

- There is bifurcation of scatter plot and shows two types of ads with positive correlation
- Type 1 ads have low clicks but high revenues
    - Type 2 ads have low revenue but very high clicks

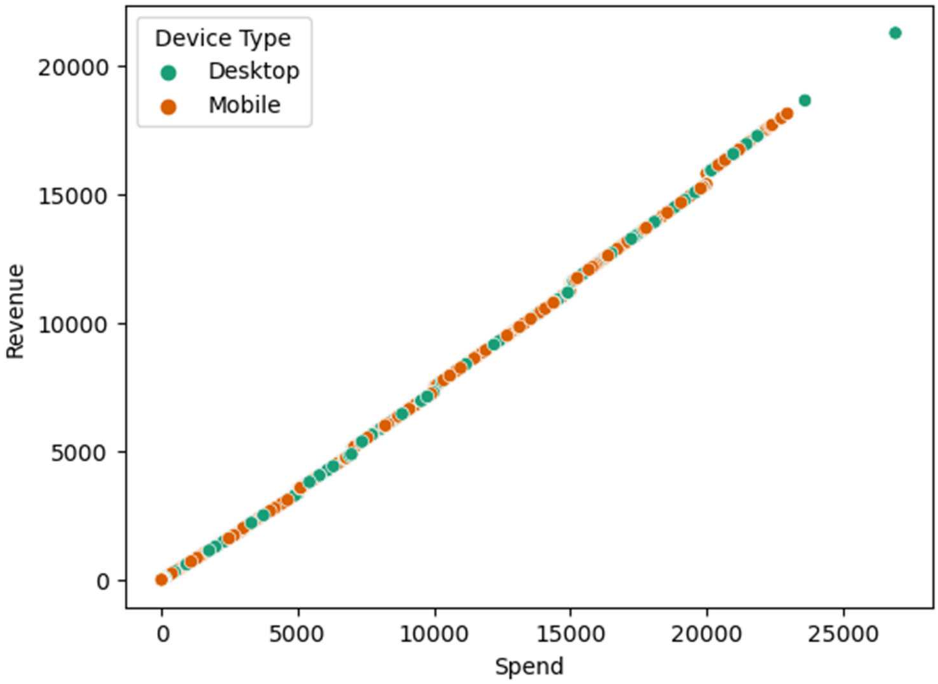


Figure 10 Revenue vs Spend

- For both device types the revenue and spend has proportional correlation suggesting revenues will always grow based on spending.

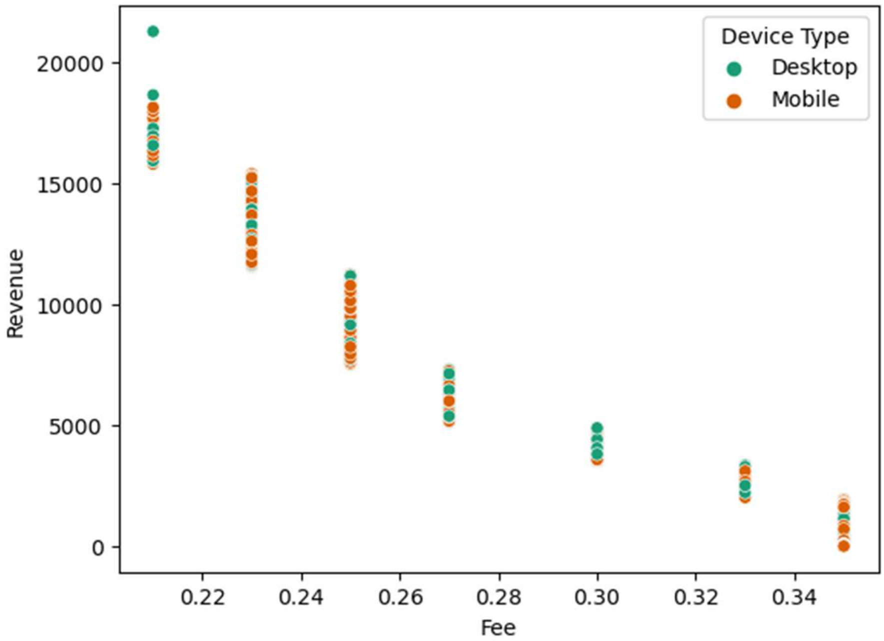


Figure 11 Revenue vs Fee

- As the fee of platform increases the revenue is going down.
- The highest fee is 0.35 which gives lowest revenue
- There are 7 different fee types

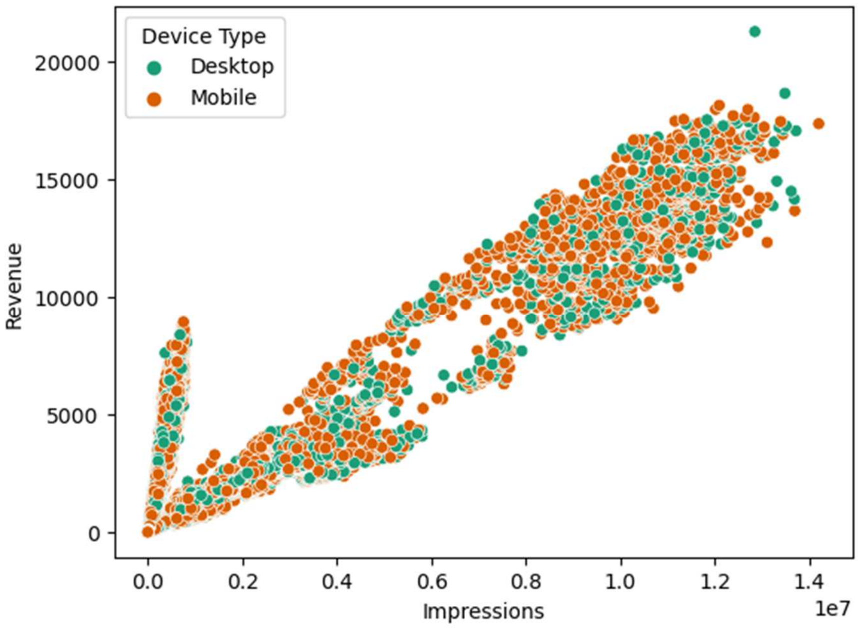


Figure 12 Revenue vs Impressions

- Revenue and impressions have high positive correlation
- We can see bifurcation of Impressions at low impressions
- Best revenues are genetated when Impressions are over 800000

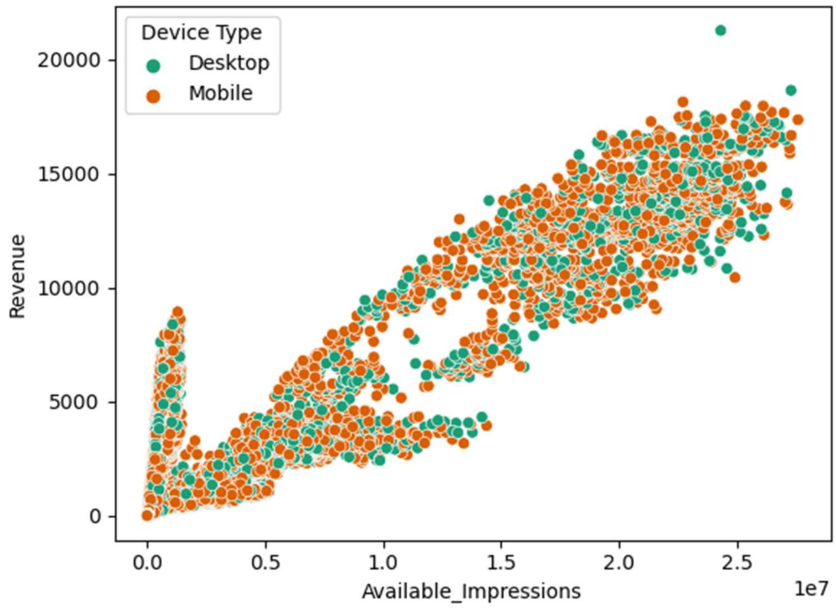


Figure 13Revenue vs Available Impressions

- Revenue and available impressions also have high positive correlation
- We can see bifurcation of Impressions at low impressions
- Best revenues are generated when available Impressions are over 1500000

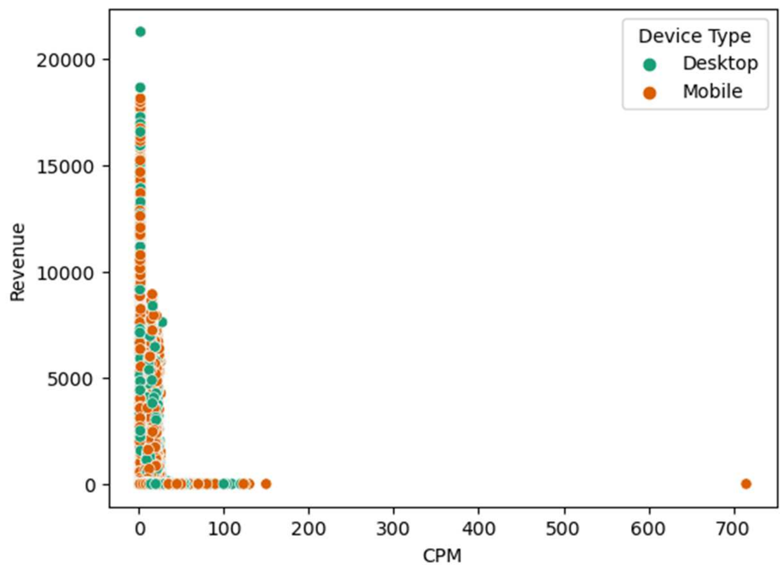


Figure 14 Revenue vs CPM

- Revenue vs CPM has no strong correlation
- Lower CPM results in higher CPM which is good for business

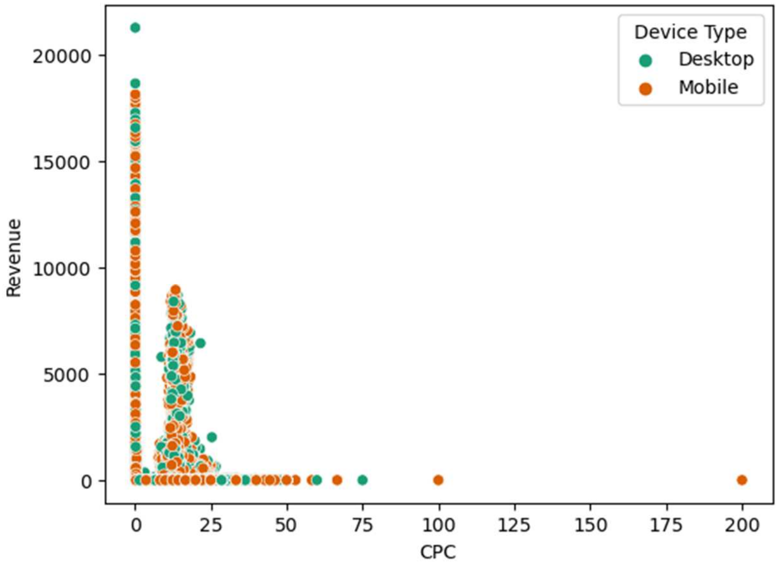


Figure 15 Revenue vs CPC

- No strong correlation exists between Revenue and CPC
- CPC when increasing is giving almost no revenue
- CPC range of 0 to 25 brings most of the revenue

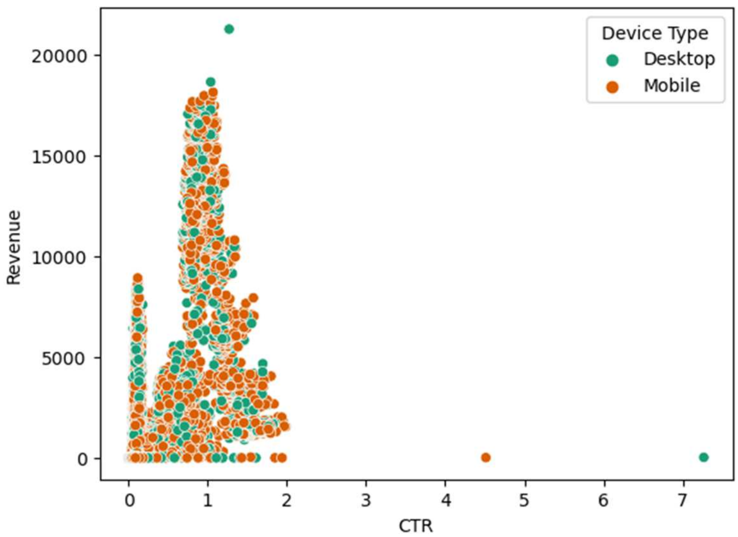


Figure 16 Revenue vs CTR

- No strong correlation is suggested from the scatter plot.
- CTR 1 gives maximum revenue based on the plot

Part 1: Clustering: Data Preprocessing

- Missing value check and treatment - Outlier Treatment - z-score scaling Note: Treat missing values in CPC, CTR and CPM using the formula given.

Checking Missing Values:

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

Figure 17 Ads Data missing value

We see that CTR, CPM and CPC have large number of missing values. Treating Missing values by imputing values based on formula:

Formula used:-

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

After Treatment:

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

Figure 18 No missing values after imputing values

Checking Null Values:

No null values are present in the data set.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23066 entries, 0 to 23065
Data columns (total 13 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   Ad - Length            23066 non-null  float64
 1   Ad- Width              23066 non-null  float64
 2   Ad Size                23066 non-null  float64
 3   Available_Impressions  23066 non-null  float64
 4   Matched_Queries        23066 non-null  float64
 5   Impressions            23066 non-null  float64
 6   Clicks                 23066 non-null  float64
 7   Spend                  23066 non-null  float64
 8   Fee                    23066 non-null  float64
 9   Revenue                23066 non-null  float64
 10  CTR                    23066 non-null  float64
 11  CPM                    23066 non-null  float64
 12  CPC                    23066 non-null  float64
dtypes: float64(13)
memory usage: 2.3 MB
```

Figure 19 Null Values

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

Ans: Yes, treating outliers is a crucial step before performing clustering. Clustering algorithms, such as K-means and Hierarchical Clustering, are sensitive to outliers. Outliers can significantly change the shape and scale of the data distribution, which can lead to misleading clusters.

Therefore, we need to remove outliers and scale the data before performing clustering. We can remove outliers using Quartile method.

IQR =Q3-Q1

lower_range =Q1-(1.5*IQR)

upper_range = Q3 + (1.5*IQR)

Boxplot After Treatment:

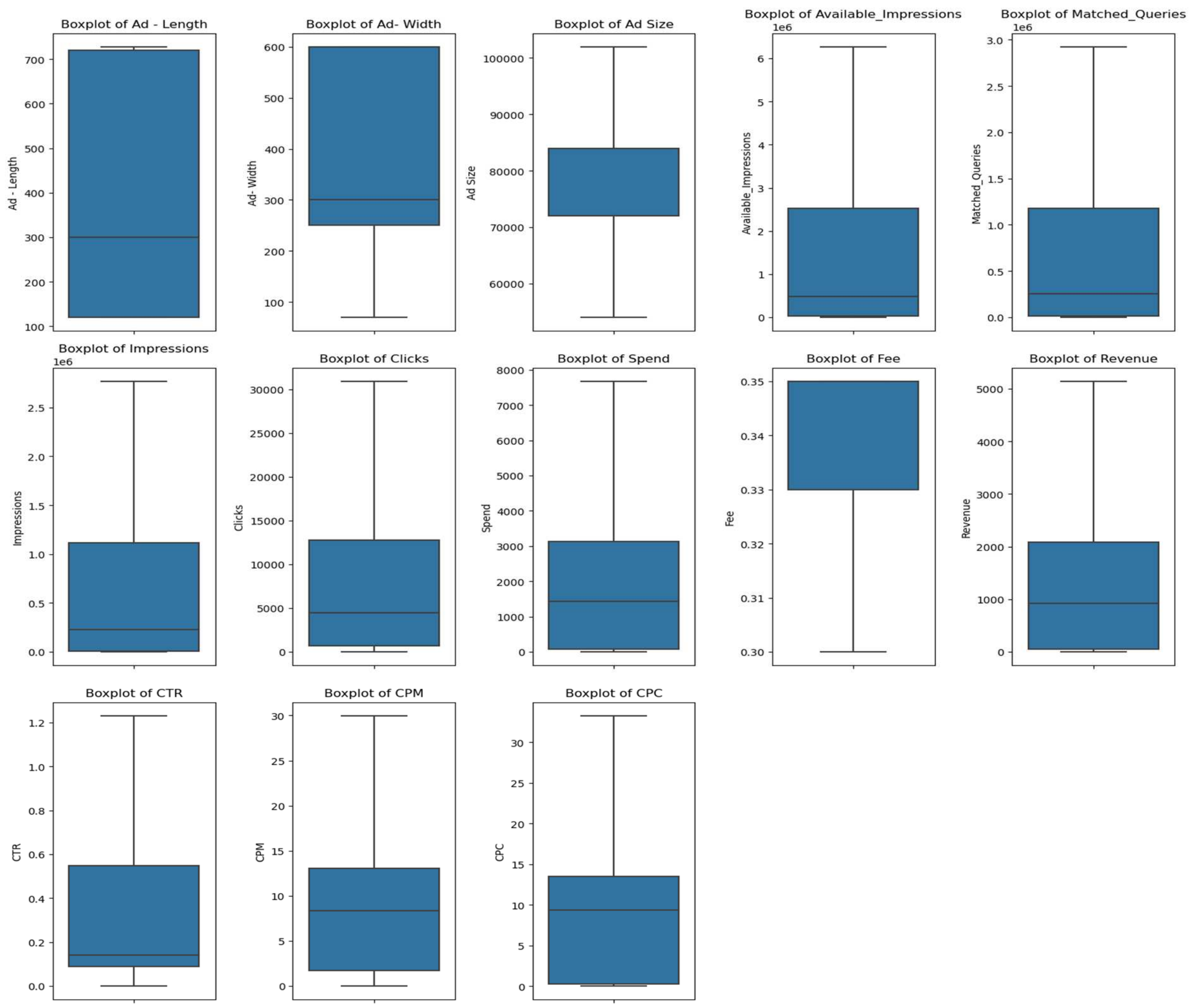


Figure 20 Boxplot ads data after treatment

Q. Perform z-score scaling and discuss how it affects the speed of the algorithm.
After Z scaling:

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CP
0	-0.364496	-0.432797	-0.102518	-0.755333	-0.778949	-0.768478	-0.867488	-0.89317	0.535724	-0.880093	-1.042561	-1.194498	-0.95883
1	-0.364496	-0.432797	-0.102518	-0.755345	-0.778988	-0.768516	-0.867488	-0.89317	0.535724	-0.880093	-1.042561	-1.194498	-0.95383
2	-0.364496	-0.432797	-0.102518	-0.754900	-0.778919	-0.768445	-0.867488	-0.89317	0.535724	-0.880093	-1.042561	-1.194498	-0.96221
3	-0.364496	-0.432797	-0.102518	-0.755040	-0.778781	-0.768302	-0.867488	-0.89317	0.535724	-0.880093	-1.042561	-1.194498	-0.97187
4	-0.364496	-0.432797	-0.102518	-0.755610	-0.779030	-0.768560	-0.867488	-0.89317	0.535724	-0.880093	-1.042561	-1.194498	-0.94628

Figure 21 Scaling ADs data

Data description after scaling:

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
count	23066.00	23066.00	23066.00	23066.00	23066.00	23066.00	23066.00	23066.00	23066.00	23066.00	23066.00	23066.00	23066.00
mean	0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	0.00
std	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
min	-1.13	-1.32	-1.47	-0.76	-0.78	-0.77	-0.87	-0.89	-2.22	-0.88	-1.04	-1.19	-1.00
25%	-1.13	-0.43	-0.30	-0.74	-0.76	-0.76	-0.79	-0.86	-0.57	-0.85	-0.76	-0.94	-0.96
50%	-0.36	-0.19	-0.30	-0.53	-0.53	-0.54	-0.41	-0.31	0.54	-0.32	-0.60	0.02	0.14
75%	1.43	1.29	0.48	0.43	0.37	0.37	0.47	0.39	0.54	0.39	0.68	0.70	0.64
max	1.47	1.29	1.65	2.19	2.07	2.06	2.36	2.27	0.54	2.24	2.85	3.16	3.04

Figure 22 Data Description after scaling

The difference between runtime of algorithm before and after scaling to describe the date is **0.00026.** We can say that after scaling the algorithm has become faster.

Part 1: Clustering: Hierarchical Clustering

<u>- Construct a dendrogram using Ward linkage and Euclidean distance - Identify the optimum number of Clusters</u>

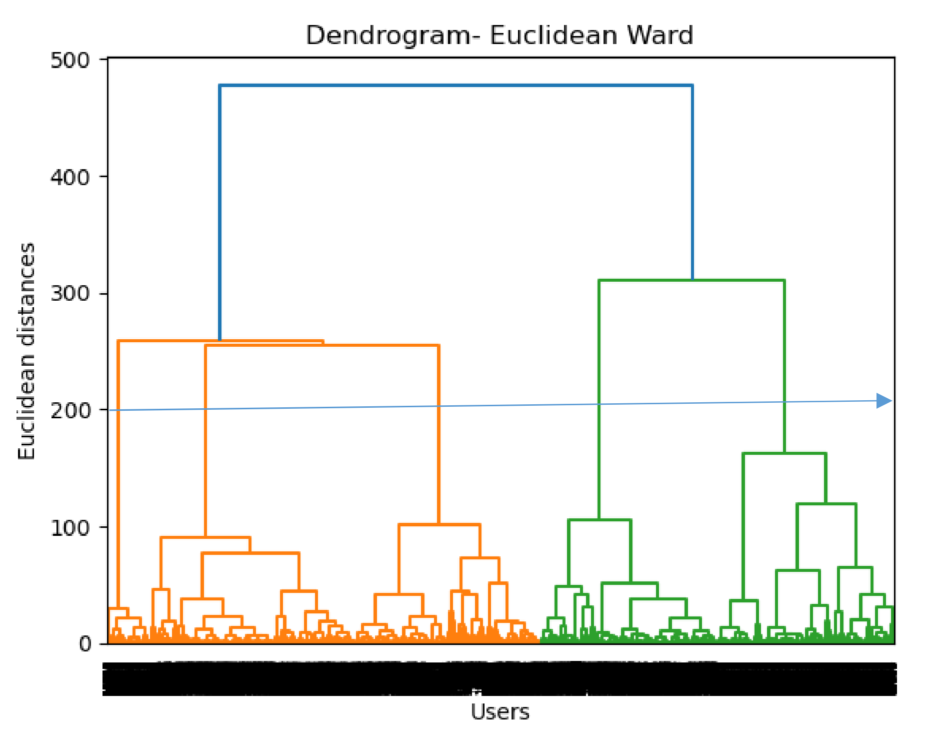


Figure 23 Dendogram (Eucledian-Ward)

As per Wards Eucledian we have 2 clusters. These two are depicted in orange and green colour. However, only two cluster are not enough for segmentation. We will try to find more number of segments. If we cut the distance at 200 we will have 5 clusters. Therefore, K=5 is suitable number of clusters based on Dendogram.

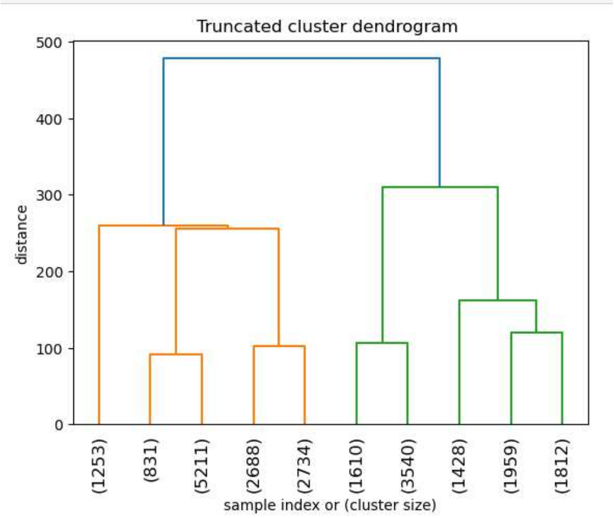


Figure 24 Truncated Cluster Dendogram

Part 1: Clustering: K-means Clustering

<u>- Apply K-means Clustering - Plot the Elbow curve - Check Silhouette Scores - Figure out the appropriate number of clusters - Cluster Profiling</u>

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

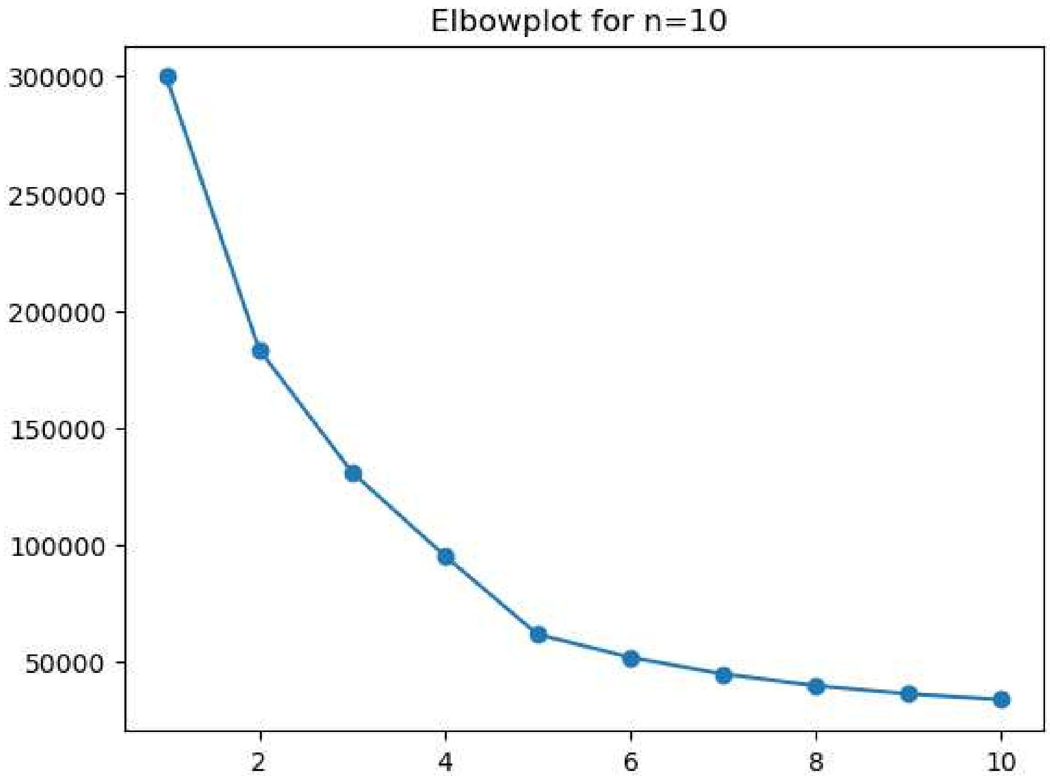


Figure 25 Elbow Plot for N=10

As we can see the differential becomes smaller and flattens after we start moving from n=5. Therefore we can take k=10. We will also calculate Silhouette score for k=10 as well and identify ideal number of clusters.

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

```
For n_clusters=2, the silhouette score is 0.38572769619101116
For n_clusters=3, the silhouette score is 0.3825486036570086
For n_clusters=4, the silhouette score is 0.4532427055259838
For n_clusters=5, the silhouette score is 0.5240956940501847
For n_clusters=6, the silhouette score is 0.5221495642670954
For n_clusters=7, the silhouette score is 0.5165635029478534
For n_clusters=8, the silhouette score is 0.47973343359439863
For n_clusters=9, the silhouette score is 0.43190290852533963
For n_clusters=10, the silhouette score is 0.44470762445447837
```

Figure 26 Sihouette score

The silhouette score is highest **for n=5 which is 0.5240** among all values on n. Therefore, we can take cluster size as **K=5**.

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.]

Cluster Based Analysis:

Row Labels	% Share	Average of Ad - Length	Average of Ad- Width	Average of Ad Size	Average of CTR	Average of CPM	Average of CPC	Average of Revenue	Ad Lenth to With Ratio
5	7%	141.45	572.45	73686.40	0.11	15.39	13.75	4471.78	0.25
4	18%	465.78	199.15	72963.94	0.75	1.57	0.22	3878.75	2.34
1	27%	421.70	152.00	64300.00	0.53	1.79	0.40	977.42	2.77
2	20%	683.83	303.79	100775.88	0.09	11.73	13.29	815.54	2.25
3	28%	143.28	572.10	73966.74	0.10	14.33	15.78	135.99	0.25

Figure 27 Cluster based Analysis

1. We **have 5 clusters** among which 5th cluster seems to be like a video ad as it has horizontal dimensions and average size of 73686.
2. Cluster 5 **has lowest legth to width ratio of .25** and brings highest average revenue.
3. Cluster 3 **has highest share of ads** and is horizontal kind of ads but it is most expensive with CPC of 15.8 and brings least amount of average revenue of 135.99
4. Cluster 4 has 18% share of ads and bring second highest average revenue 0f 3878.78
5. Although Cluster 5 and Cluster 4 have similar size, Cluster 4 has highest Click through Ratio.
6. Cluster 4 can be considered best for of ad based on its dimension it seems to be a mobile ad as it is vertical with L/W ratio of 2.34. It also has lowest cost per click
7. Cluster 1 is third best grossing ad with avg revenue of 977. It has highest share of ads with 28%
8. Cluster 2 is a vertical mobile based ads with highest size. This means Bigger size ad are not that fruitful.
9. Cluster 3 is least performing ad with highest share of 28%. It also has the lowest Click through Ratio.
10. We can say **that Cluster 5, 4 and 1 are the best choice for ads**.

Device Type Summary:

Row Labels	% Share	Average of Ad - Length	Average of Ad- Width	Average of Ad Size	Average of CTR	Average of CPM	Average of CPC	Average of Revenue
Desktop	35.81%	385.02	338.00	76608.32	0.33	8.22	8.26	1452.85
Mobile	64.19%	385.24	337.84	76559.27	0.33	8.22	8.20	1447.46
Grand Total	**100.00%**	**385.16**	**337.90**	**76576.84**	**0.33**	**8.22**	**8.22**	**1449.39**

Figure 28 Device based summary

Observations:-

- Desktop has 35% share of total ads which means it should be a secondary channel to reach customers
- Both device Type have similar ad sizes, CTR, CPM and Revenue figures on average.
- Mobile brings more customers and is less costly as CPM which is cost per mile is 8.20 compared to desktop.

Platform Type Summary:

Row Labels	% Share	Average of Ad - Length	Average of Ad-Width	Average of Ad Size	Average of CTR	Average of CPM	Average of CPC	Average of Revenue
App	21.43%	385.55	337.35	76549.32	0.33	8.24	8.20	1453.27
Video	42.80%	384.97	338.05	76556.66	0.33	8.21	8.22	1455.27
Web	35.77%	385.16	338.04	76617.46	0.33	8.22	8.24	1440.02
Grand Total	**100.00%**	**385.16**	**337.90**	**76576.84**	**0.33**	**8.22**	**8.22**	**1449.39**

Figure 29 Platform based summary

- Video is the best platform to reach customer due to high share of 42%
- We bring more average revenue than video or App
- CPM is highest for APP based ads

Part 1: Clustering: Actionable Insights & Recommendations

- Extract meaningful insights (atleast 3) from the clusters to identify the most effective types of ads, target audiences, or marketing strategies that can be inferred from each segment. - Based on the clustering analysis and key insights, provide actionable recommendations (atleast 3) to Ads24x7 on how to optimize their digital marketing efforts, allocate budgets efficiently, and tailor ad content to specific audience segments.

Top 3 Business Insights:

1. Vedio based ads needs to be run more as they contribute highest average revenue among all clusters.
2. Prioritizing mobile based customers is important because it has highest share of ads and low CPM
3. Poster size ads have highest conversion rate with CTR of .75 and also brings maximum revenue

Marketing Strategy & Business Recommendation:-

1. Platform share should be prioritized as follows: 70% Web 20% and App 10% share can bring better revenues.
2. Poster size ads with avg dimensions of 465 x 200 should be used to increase the revenues. They can be used on mobiles.
3. Vedio based ads with dimensions of 141 x 572 has only share of 7% but brings maximum average revenue. The share must be increase to 50% to get maximum results.
4. Mobile based ads cost less compared to Web based ads. Prioritizing Mobile based Vedio ads is a must have strategy for any campaign.

PROBLEM STATEMENT 2

PCA:

PCA FH (FT): Primary census abstract for female headed households excluding institutional households (India & States/UTs - District Level), Scheduled tribes - 2011 PCA for Female Headed Household Excluding Institutional Household. The Indian Census has the reputation of being one of the best in the world. The first Census in India was conducted in the year 1872. This was conducted at different points of time in different parts of the country.

In 1881 a Census was taken for the entire country simultaneously. Since then, Census has been conducted every ten years, without a break. Thus, the Census of India 2011 was the fifteenth in this unbroken series since 1872, the seventh after independence and the second census of the third millennium and twenty first century. The census has been uninterruptedly continued despite of several adversities like wars, epidemics, natural calamities, political unrest, etc. The Census of India is conducted under the provisions of the Census Act 1948 and the Census Rules, 1990.

The Primary Census Abstract which is important publication of 2011 Census gives basic information on Area, Total Number of Households, Total Population, Scheduled Castes, Scheduled Tribes Population, Population in the age group 0-6, Literates, Main Workers and Marginal Workers classified by the four broad industrial categories, namely,

- Cultivators,
- Agricultural Laborers,
- Household Industry Workers, and
- 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.

- **Note: The 24 variables given in the Rubric is just for performing EDA. You will have to consider the entire dataset, including all the variables for performing PCA.**

Part 2: PCA: Define the problem and perform Exploratory Data Analysis

- Problem Definition - Check shape, Data types, statistical summary –
-Perform an EDA on the data to extract useful insights
Note:
1. 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

2. Example questions to answer from EDA –
(i) Which state has highest gender ratio and which has the lowest?
(ii) Which district has the highest & lowest gender ratio?

Solution 2

Head:

	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180	
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123	
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	...	114	188	44	
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	...	194	247	61	
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	...	874	1928	465	

5 rows × 61 columns

Figure 30 Census Head

Shape:

(640, 61)

Info:

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

```
21  MAIN_CL_M       640 non-null    int64
22  MAIN_CL_F       640 non-null    int64
23  MAIN_AL_M       640 non-null    int64
24  MAIN_AL_F       640 non-null    int64
25  MAIN_HH_M       640 non-null    int64
26  MAIN_HH_F       640 non-null    int64
27  MAIN_OT_M       640 non-null    int64
28  MAIN_OT_F       640 non-null    int64
29  MARGWORK_M      640 non-null    int64
30  MARGWORK_F      640 non-null    int64
31  MARG_CL_M       640 non-null    int64
32  MARG_CL_F       640 non-null    int64
33  MARG_AL_M       640 non-null    int64
34  MARG_AL_F       640 non-null    int64
35  MARG_HH_M       640 non-null    int64
36  MARG_HH_F       640 non-null    int64
37  MARG_OT_M       640 non-null    int64
38  MARG_OT_F       640 non-null    int64
39  MARGWORK_3_6_M  640 non-null    int64
40  MARGWORK_3_6_F  640 non-null    int64
41  MARG_CL_3_6_M   640 non-null    int64
42  MARG_CL_3_6_F   640 non-null    int64
43  MARG_AL_3_6_M   640 non-null    int64
44  MARG_AL_3_6_F   640 non-null    int64
45  MARG_HH_3_6_M   640 non-null    int64
46  MARG_HH_3_6_F   640 non-null    int64
47  MARG_OT_3_6_M   640 non-null    int64
48  MARG_OT_3_6_F   640 non-null    int64
49  MARGWORK_0_3_M  640 non-null    int64
50  MARGWORK_0_3_F  640 non-null    int64
51  MARG_CL_0_3_M   640 non-null    int64
52  MARG_CL_0_3_F   640 non-null    int64
53  MARG_AL_0_3_M   640 non-null    int64
54  MARG_AL_0_3_F   640 non-null    int64
55  MARG_HH_0_3_M   640 non-null    int64
56  MARG_HH_0_3_F   640 non-null    int64
57  MARG_OT_0_3_M   640 non-null    int64
58  MARG_OT_0_3_F   640 non-null    int64
59  NON_WORK_M      640 non-null    int64
60  NON_WORK_F      640 non-null    int64
dtypes: int64(59), object(2)
memory usage: 305.1+ KB
```

Figure 31 Census Info

Describe:

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

8 rows × 59 columns

Figure 32 Census description

- Mean and standard variation are different. The scale of measurement of each row is also varying

Null Values:
No Null values found

Duplicated values:
No duplicated values found

Selecting below 5 variables for EDA:
1. State : Name of State
2. No_HH: No. of Households
3. TOT_M: Total Male Population
4. TOT_F: Total Female Population
5. TOT_WORK_M: Total Working population Male
6. TOT_WORK_F: Total Working population Female

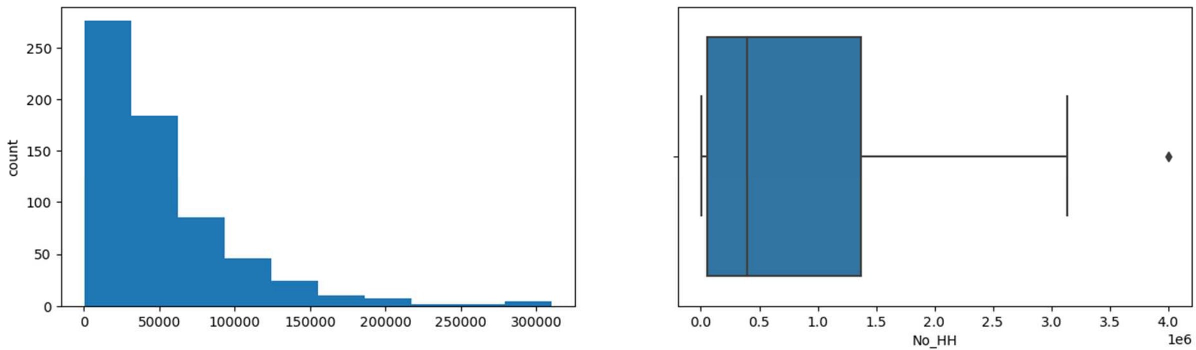


Figure 33 Boxplot: Total households

- Total no. of household in states ranges upto 30000
- Data has skeness of 1.24
- Dadra and Nagar Haveli has lowest no. of households 4288
- Top 3 states with highest no. of households are as follows:

Uttar Pradesh	4006871
Maharashtra	3136214
Andhra Pradesh	3127287

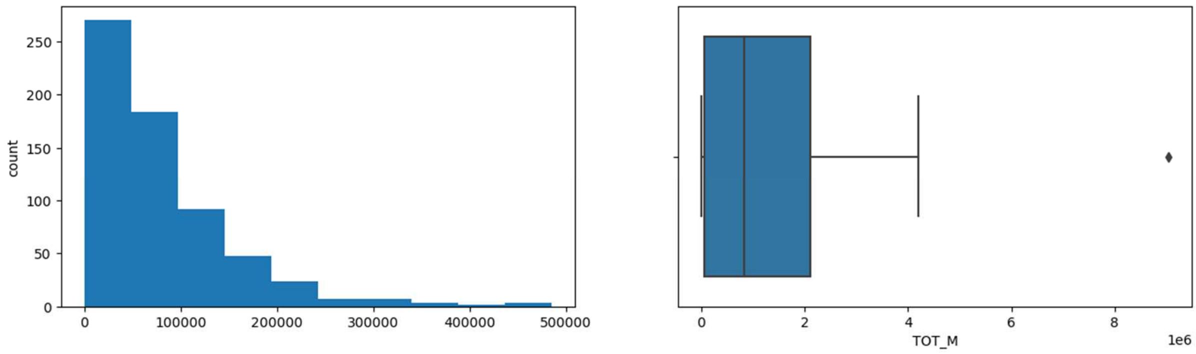


Figure 34 Boxplot: Total Male population

- Total Male Population in states goes upto 400000
- The data has skewness of `2.17`
- Uttar Pradesh has highest male population of 9043969
- Dadra & Nagar Haveli has lowest male population of 6982

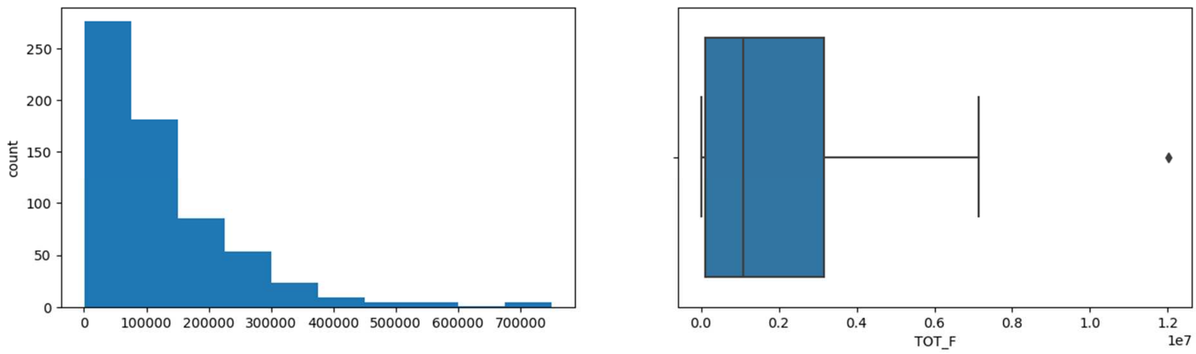


Figure 35 Boxplot: Total Female population

- Total Female population in state varies upto 700000
- Data has skewness of 1.67
- Top 3 states with highest female population:

Uttar Pradesh	12023885
Maharashtra	7138557
Andhra Pradesh	6097235

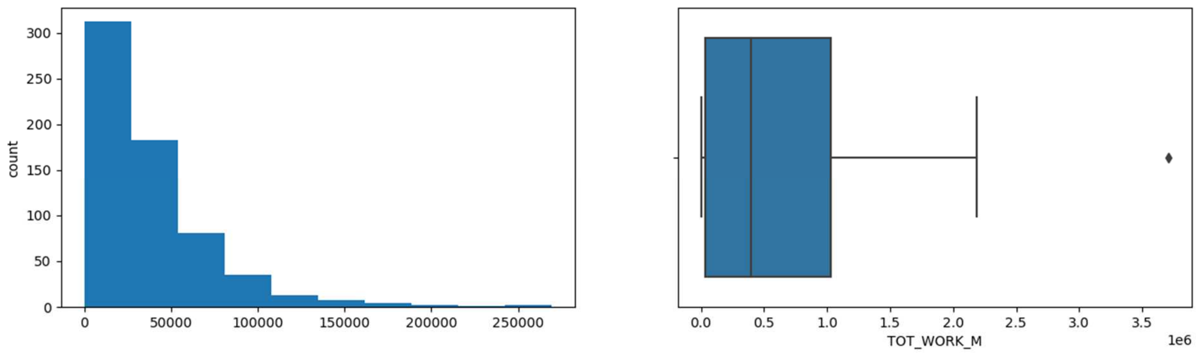


Figure 36 Boxplot Total Working Male

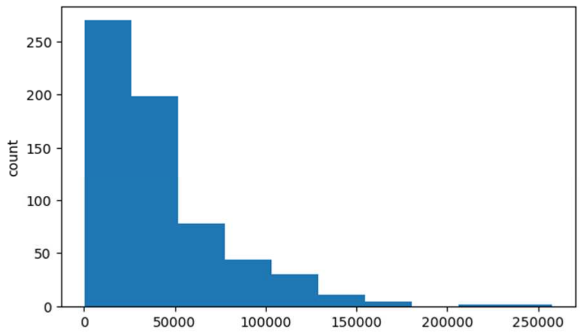
- Total working male population goes upto 200000
- Data has skewness of 1.66
- It also has outliers and represents Uttar Pradesh

- Top states with highest working population are

Uttar Pradesh	3710433
West Bengal	2187371
Maharashtra	2028630

- States with lowest working population of males are:

Dadara & Nagar Havelli	3138
Lakshadweep	5115
Daman & Diu	6884



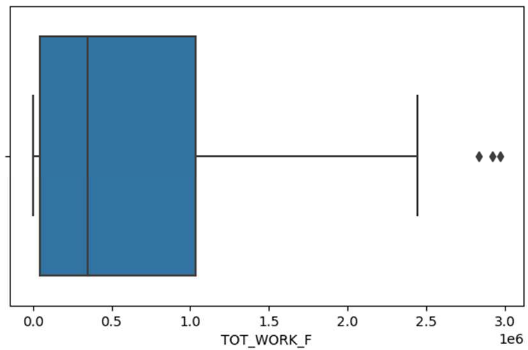


Figure 37 Boxplot Total working female

- Total Working Females ranges upto 250000
- Data has skewness of 1.31
- It has outlies representing below sates which have highest working female population

Uttar Pradesh	2972243
Maharashtra	2918694
Andhra Pradesh	2833719

Sex ratio:

Top 5 states with highest sex ratio are:

Andhra Pradesh	1862
Tamil Nadu	1825
Chhattisgarh	1821
Arunachal Pradesh	1741
Odisha	1738

Figure 38 Top 5 States: Highest Sex ratio

- Andhra Pradesh has highest sex ratio with 1862 women per thousand males

Top 5 states with lowest sex ratio

State	Sex Ratio
Lakshadweep	1152
Haryana	1283
NCT of Delhi	1290
Uttar Pradesh	1329
Meghalaya	1330

Figure 39Top 5 States: Lowest Sex ratio

Top 5 Areas with highest Male working population:

Area Name	No_HH	TOT_M	TOT_F	TOT_WORK_M	TOT_WORK_F	Sex Ratio
North Twenty Four Parganas	310450	471482	725514	269422	176430	1538.794694
Mumbai Suburban	304502	485417	750392	262638	227123	1545.870870
Bangalore	287841	401545	664595	238323	257848	1655.094697
Barddhaman	240421	381529	565766	214205	146860	1482.891209
Thane	294698	424759	706327	211026	255770	1662.888838

Figure 40 Districts with highest working males

Top Areas with highest Sex ratio:

Area Name	No_HH	TOT_M	TOT_F	TOT_WORK_M	TOT_WORK_F	Sex Ratio
Krishna	182404	137603	314182	69810	146724	2283.249638
Koraput	46307	38026	86272	17870	44720	2268.763478
Virudhunagar	90241	66704	148445	38587	76432	2225.428760
West Godavari	163437	123111	273534	66492	116180	2221.848576
Baudh	10665	8672	19209	4295	10319	2215.059963

Figure 41 Districts with highest working females

Top Areas with lowest Sex Ratio are:

Area Name	No_HH	TOT_M	TOT_F	TOT_WORK_M	TOT_WORK_F	Sex Ratio
Lakshadweep	4445	12823	14772	5115	1780	1151.992513
Badgam	6218	19585	23102	6982	4200	1179.576206
Mahamaya Nagar	27728	67258	79378	30629	17018	1180.201612
Dhaulpur	15153	31904	37671	15679	15274	1180.761033
Baghpat	21966	54807	64937	24649	12429	1184.830405

Figure 42 District with lowest sex ratio

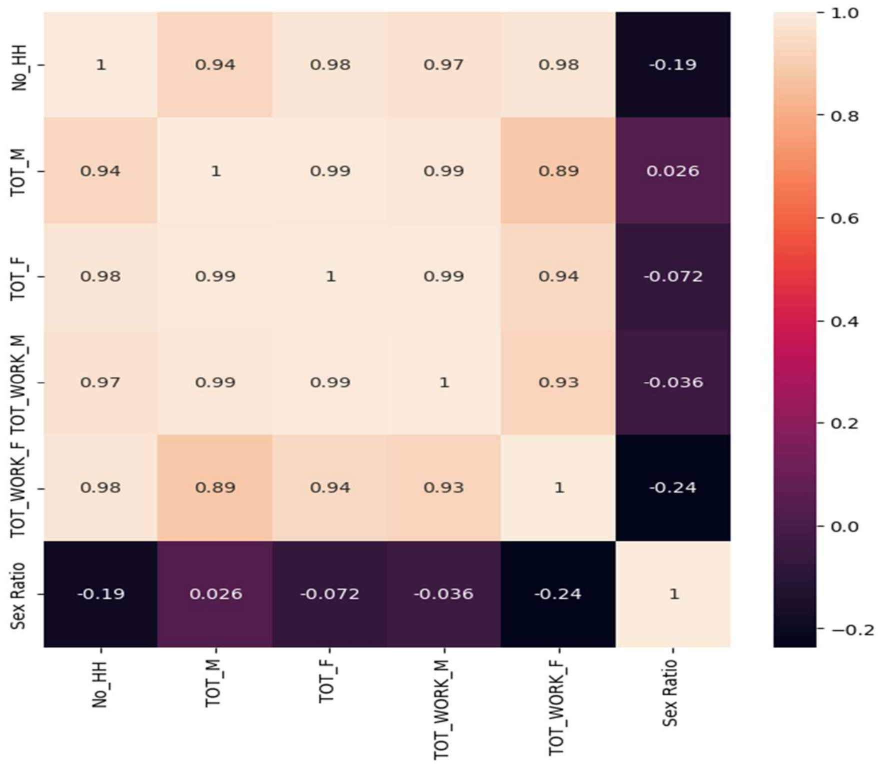


Figure 43 Correlation matrix of 5 variables selected

Observations:

- We can see very high correlations exist between Total male and female populations with 0.99
- However, negative correlating can be seen between Sex ratio and No. of households
- When sex ratio increase total male population has positive total male population which is not very strong.

After conducting EDA for selected variable and understanding data structure and getting key insights we are ready for further steps to Data Pre processing and PCA analysis.

Part 2: PCA: Data Pre-processing

- Check for and treat (if needed) missing values - Check for and treat (if needed) data irregularities - Scale the Data using the z-score method - Visualize the data before and after scaling and comment on the impact on outliers

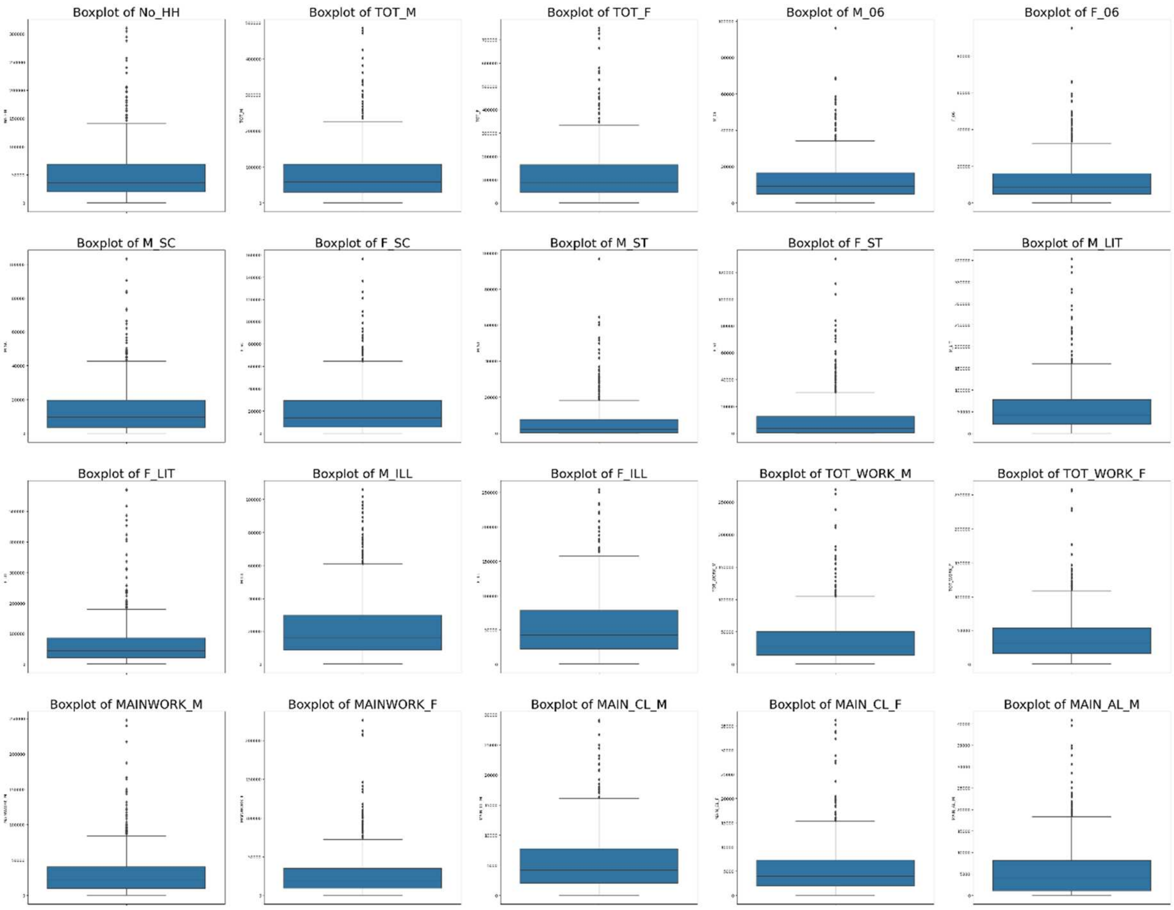


Figure 44 Boxplot Census

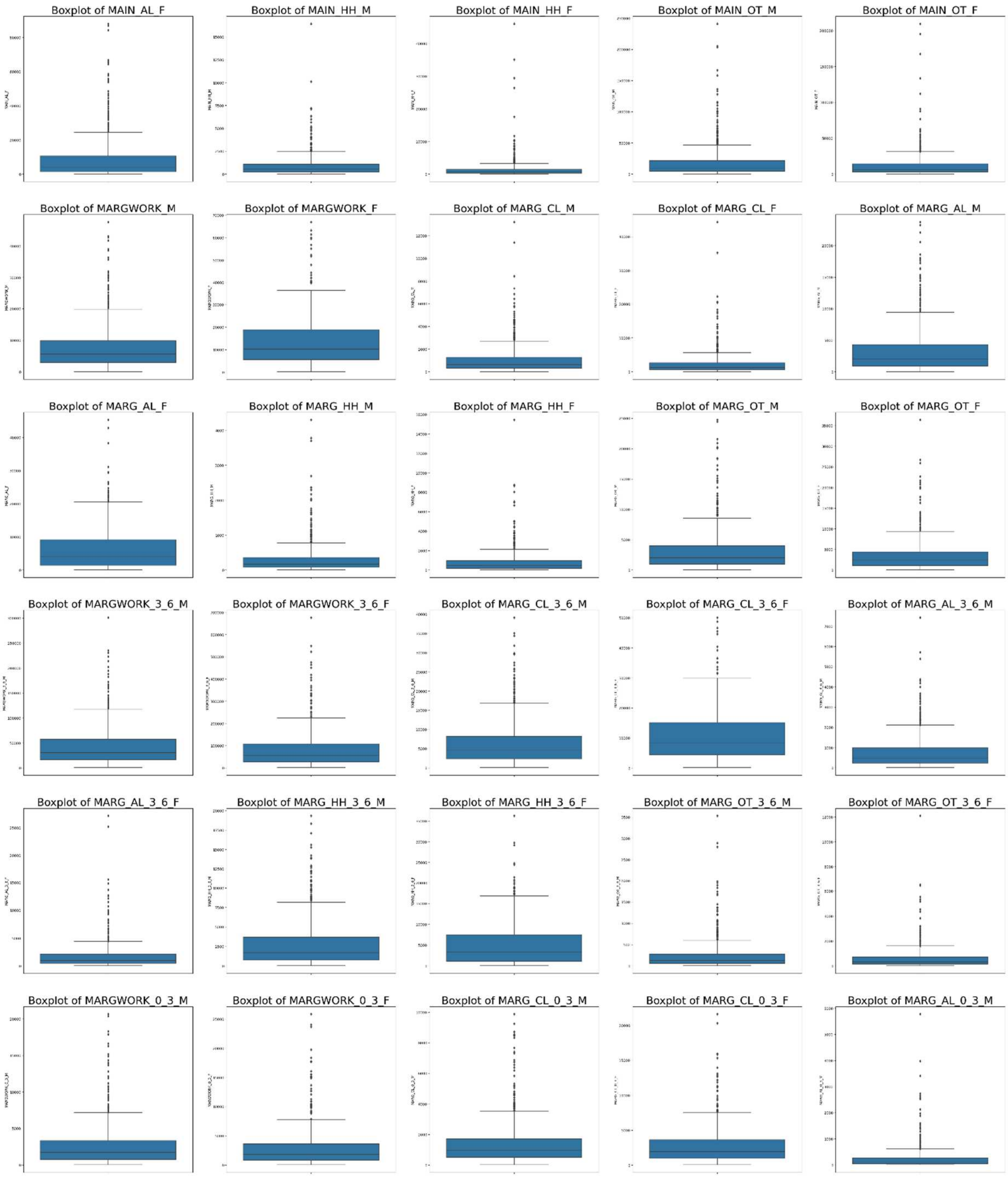


Figure 45 Boxplot Census

Boxplot after removing outliers

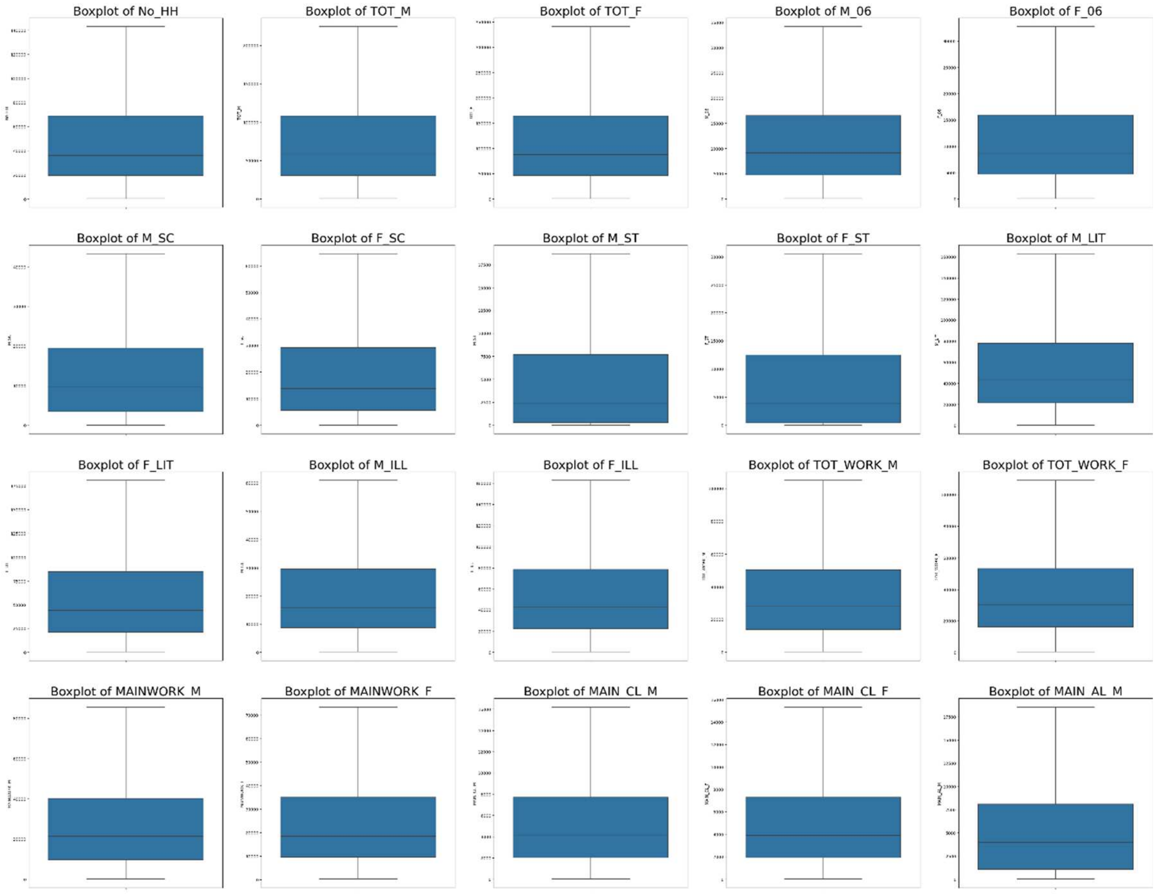


Figure 46 Boxplot after removing outliers

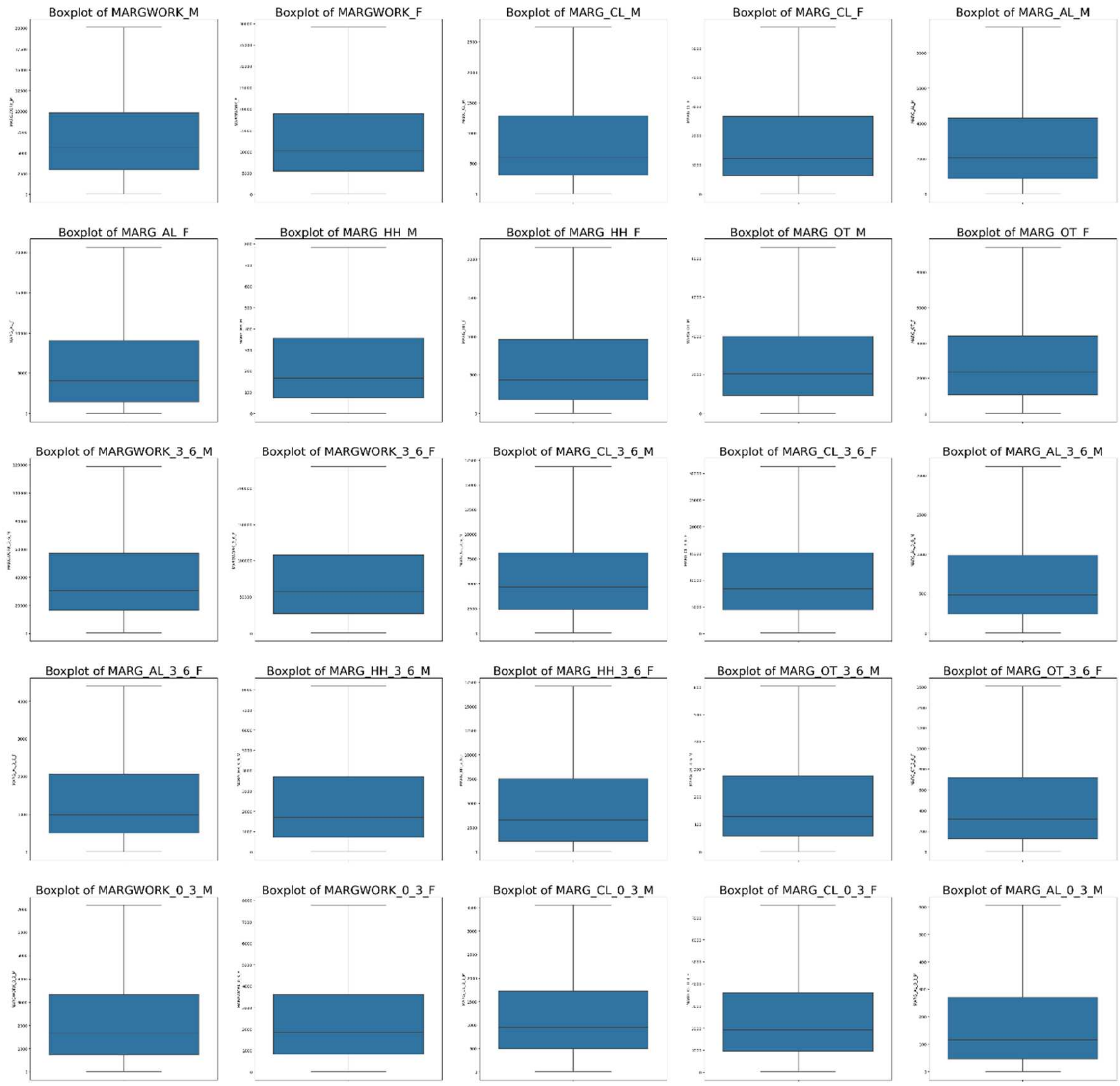


Figure 47 Boxplot after removing outliers

Applying Z score

Head:

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL
0	-1.038986	-0.874837	-0.937027	-0.624685	-0.561282	-1.080201	-1.079963	-0.510440	-0.574198	-0.939617	...	-0.093587	-0.860882	-(
1	-1.076896	-0.938023	-1.009723	-0.773932	-0.835657	-1.079873	-1.079635	-0.771833	-0.782092	-1.005083	...	-0.719169	-0.877096	-(
2	-1.121858	-1.154665	-1.141539	-1.141642	-1.138104	-1.080201	-1.079635	0.122588	0.137599	-1.141561	...	-1.130551	-1.128423	-(
3	-1.201599	-1.217171	-1.214930	-1.197772	-1.176091	-1.080447	-1.079963	-0.399531	-0.437333	-1.203009	...	-1.050477	-1.100286	-(
4	-0.938495	-0.921309	-0.935018	-0.700931	-0.740523	-1.078807	-1.078160	0.432534	0.249489	-0.942767	...	-0.369844	-0.298617	1

5 rows × 57 columns

Figure 48 Zscore applied

Describe:

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_
count	640.00	640.00	640.00	640.00	640.00	640.00	640.00	640.00	640.00	640.00	...	640.00	640.00	640.00	64(
mean	-0.00	-0.00	-0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00	...	-0.00	-0.00	-0.00	-(
std	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	...	1.00	1.00	1.00	1
min	-1.23	-1.26	-1.25	-1.25	-1.25	-1.08	-1.08	-0.84	-0.83	-1.24	...	-1.24	-1.20	-1.01	-1
25%	-0.74	-0.76	-0.76	-0.75	-0.73	-0.80	-0.77	-0.79	-0.79	-0.76	...	-0.75	-0.76	-0.75	-(
50%	-0.32	-0.29	-0.31	-0.27	-0.29	-0.29	-0.33	-0.45	-0.45	-0.27	...	-0.29	-0.30	-0.39	-(
75%	0.52	0.53	0.52	0.53	0.52	0.51	0.51	0.43	0.41	0.54	...	0.47	0.50	0.44	(
max	2.41	2.47	2.44	2.44	2.40	2.48	2.45	2.27	2.22	2.48	...	2.31	2.39	2.24	2

8 rows × 57 columns

Figure 49 Description after Zscore

After scaling we can see that mean has become zero and standard deviation has become 1. Lets create boxplot and get a snapshot view of the data.

Box plot after scaling the data and changes observed (snapshot)

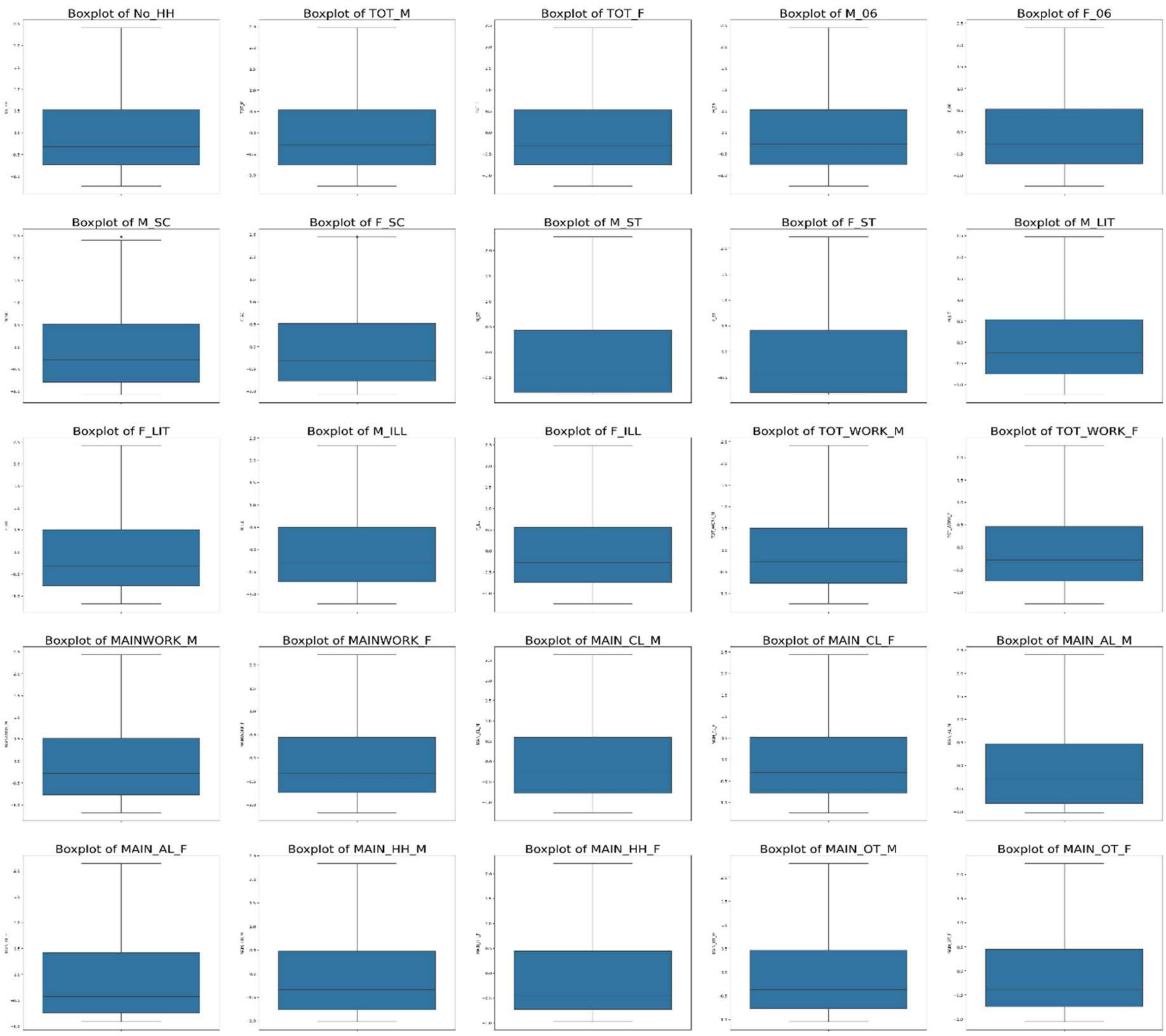


Figure 50 Boxplot after applying Zscore

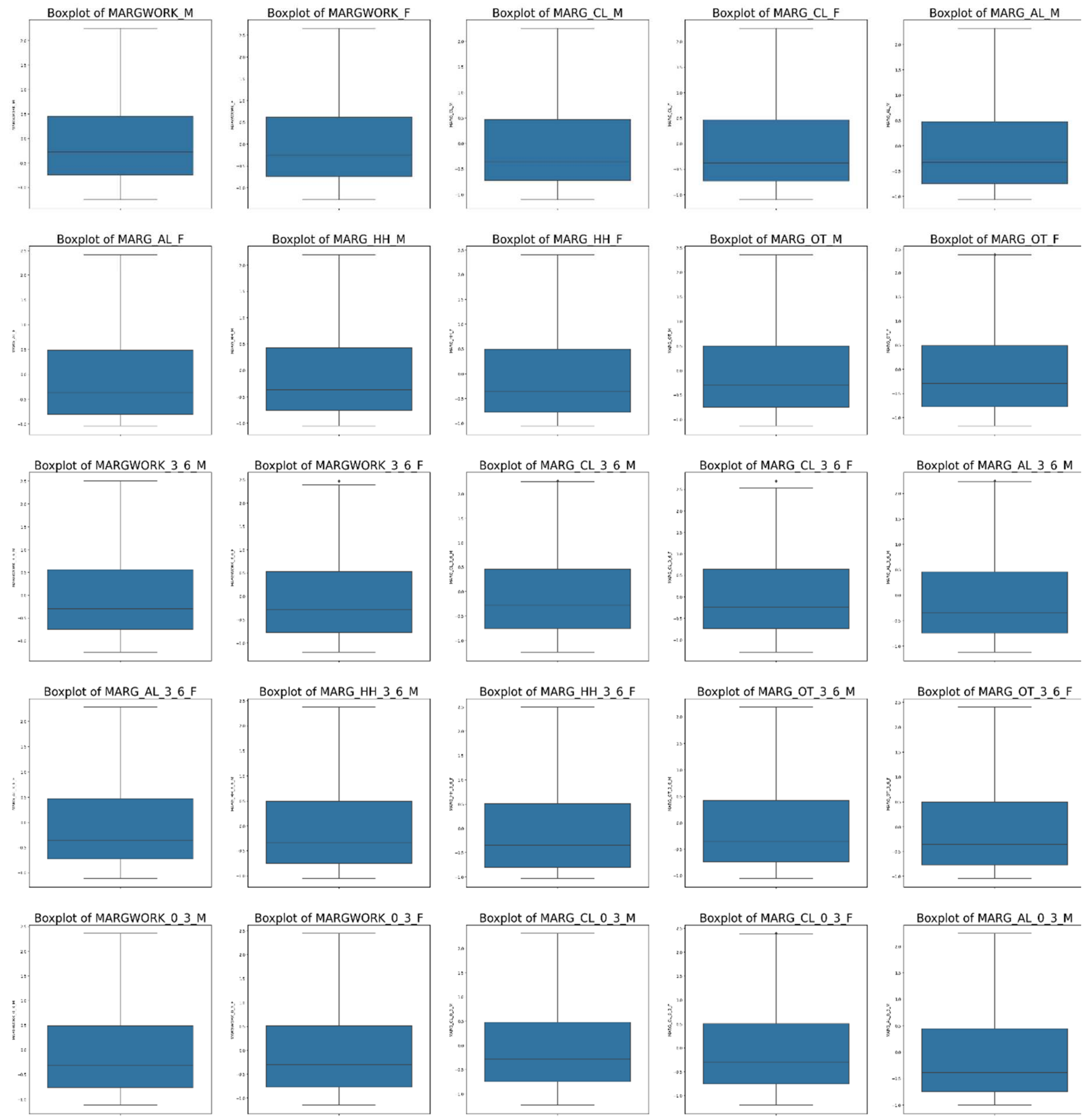


Figure 51 Boxplot after applying Zscore

Part 2: PCA: PCA

- Create the covariance matrix

- Get eigen values and eigen vectors

- Identify the optimum number of PCs

- Show Scree plot

- Compare PCs with Actual Columns and identify which is explaining most variance

- Write inferences about all the PCs in terms of actual variables

- Write linear equation for first PC

Note: For the scope of this project, take at least 90% explained variance.

Bartlett Sphericity Test

We need to confirm the statistical significance of correlations using Bartlett Sphericity Test

H0: Correlations are not significant,

H1: There are significant correlations

#Reject H0 if p-value < 0.05

P Value = 00, Thus we reject null hypothesis and accept alternate hypothesis suggesting correlations are significant.

KMO Test:

We need to confirm the adequacy of sample size using KMO test.

Note: Above 0.7 is good, below 0.5 is not acceptable

```
KMO Test Value = 0.936189616665265
Thus, sample size is adequate enough.
```

Creating covariance Matrix:

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG
No_HH	1.001565	0.912699	0.973013	0.812856	0.809883	0.806713	0.858562	0.116300	0.122722	0.931350	...	0.604943	0.617144	
TOT_M	0.912699	1.001565	0.980122	0.965044	0.960153	0.877158	0.861703	0.023439	0.013301	0.989312	...	0.739665	0.637775	
TOT_F	0.973013	0.980122	1.001565	0.914418	0.911167	0.857664	0.876435	0.076189	0.074248	0.983281	...	0.697119	0.652550	
M_06	0.812856	0.965044	0.914418	1.001565	0.999032	0.833344	0.796794	-0.006081	-0.021166	0.924761	...	0.799076	0.683667	
F_06	0.809883	0.960153	0.911167	0.999032	1.001565	0.823888	0.790043	0.006803	-0.007896	0.915929	...	0.805050	0.689114	
M_SC	0.806713	0.877158	0.857664	0.833344	0.823888	1.001565	0.984688	-0.096913	-0.099226	0.868007	...	0.647698	0.554284	
F_SC	0.858562	0.861703	0.876435	0.796794	0.790043	0.984688	1.001565	-0.052859	-0.048597	0.862923	...	0.620049	0.572684	
M_ST	0.116300	0.023439	0.076189	-0.006081	0.006803	-0.096913	-0.052859	1.001565	0.994481	0.026290	...	0.094899	0.202219	
F_ST	0.122722	0.013301	0.074248	-0.021166	-0.007896	-0.099226	-0.048597	0.994481	1.001565	0.017617	...	0.083930	0.207070	
M_LIT	0.931350	0.989312	0.983281	0.924761	0.915929	0.868007	0.862923	0.026290	0.017617	1.001565	...	0.694535	0.603799	
F_LIT	0.940747	0.937579	0.963424	0.844453	0.835104	0.805082	0.823245	0.047388	0.043933	0.974173	...	0.615830	0.555857	

Figure 52 Covariance matrix

Eigen Vectors:

```
array([[ 0.14922158,  0.15916917,  0.15820921, ...,  0.14136961,
         0.14762899,  0.14210263],
       [-0.11548673, -0.08023879, -0.09371751, ...,  0.03510934,
        -0.04912234, -0.03984815],
       [ 0.1015276 , -0.03866173,  0.0289595 , ..., -0.10217491,
        -0.12667281, -0.02854464],
       ...,
       [ 0.00112879, -0.00673066,  0.02298648, ..., -0.01159627,
         0.05608352, -0.00610478],
       [ 0.00070908,  0.04637872,  0.00402434, ...,  0.01406358,
        -0.07729171, -0.00056173],
       [-0.00461221, -0.00370327,  0.00963954, ...,  0.00227908,
         0.00539901,  0.00130606]])
```

Figure 53 Eigent Vectors

Eigen Values:

```
array([3.56488638e+01, 7.64357559e+00, 3.76919551e+00, 2.77722349e+00,
       1.90694892e+00, 1.15490310e+00, 9.87726707e-01, 4.64629906e-01,
       3.96708513e-01, 3.22346888e-01, 2.73207369e-01, 2.35647574e-01,
       1.81401107e-01, 1.69243770e-01, 1.38592325e-01, 1.31505852e-01,
       1.03809666e-01, 9.55333831e-02, 8.58580407e-02, 8.09138742e-02,
       6.60179067e-02, 6.30797999e-02, 4.82756124e-02, 4.59506197e-02,
       4.37747566e-02, 3.19339710e-02, 2.86194563e-02, 2.75481445e-02,
       2.34340044e-02, 2.20296816e-02, 1.87487040e-02, 1.59004895e-02,
       1.39957919e-02, 1.18916465e-02, 1.11133495e-02, 9.07842645e-03,
       7.25127869e-03, 6.27213692e-03, 4.95541908e-03, 4.60667097e-03,
       3.45902033e-03, 2.18408510e-03, 2.13514664e-03, 1.92111328e-03,
       1.43840980e-03, 1.09968912e-03, 9.65752052e-04, 8.62630267e-04,
       6.51634478e-04, 5.76658846e-04, 4.35790607e-04, 3.70037468e-04,
       3.06660171e-04, 2.07854170e-04, 1.38286484e-04, 8.97034441e-05,
       4.61745385e-05])
```

Figure 54 Eigen Values

Explained Variance Ratio:

```
array([6.24441446e-01, 1.33888289e-01, 6.60229147e-02, 4.86470891e-02,
       3.34029704e-02, 2.02297994e-02, 1.73014629e-02, 8.13866529e-03,
       6.94892379e-03, 5.64637229e-03, 4.78562250e-03, 4.12770833e-03,
       3.17750294e-03, 2.96454958e-03, 2.42764517e-03, 2.30351534e-03,
       1.81837655e-03, 1.67340548e-03, 1.50392785e-03, 1.41732362e-03,
       1.15639919e-03, 1.10493400e-03, 8.45617224e-04, 8.04891611e-04,
       7.66778221e-04, 5.59369722e-04, 5.01311201e-04, 4.82545623e-04,
       4.10480504e-04, 3.85881758e-04, 3.28410688e-04, 2.78520087e-04,
       2.45156553e-04, 2.08299401e-04, 1.94666401e-04, 1.59021779e-04,
       1.27016642e-04, 1.09865556e-04, 8.68013375e-05, 8.06925096e-05,
       6.05897475e-05, 3.82574118e-05, 3.74001838e-05, 3.36510796e-05,
       2.51958296e-05, 1.92626466e-05, 1.69165450e-05, 1.51102177e-05,
       1.14143210e-05, 1.01010143e-05, 7.63350323e-06, 6.48174183e-06,
       5.37159674e-06, 3.64086663e-06, 2.42228792e-06, 1.57128566e-06,
       8.08813873e-07])
```

Figure 55 Explained Variance ratio

Scree Plot:

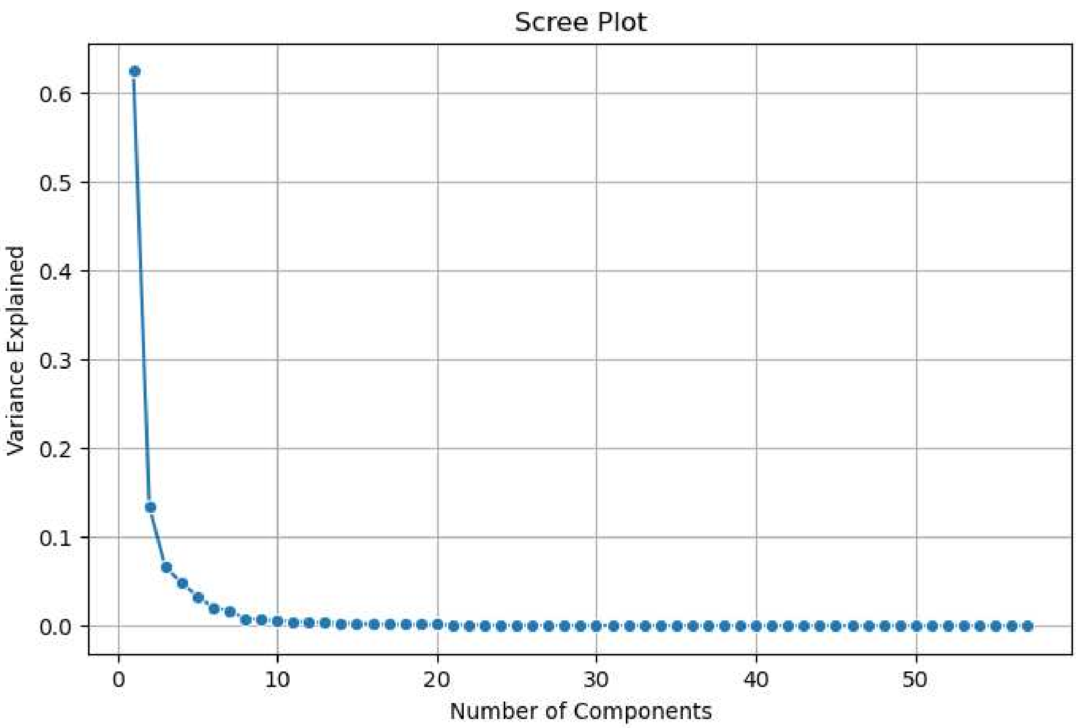


Figure 56 Scree plot

As per the Scree Plot, the graph flattens completely after 10. If we consider initial 10 components then we need to do further reduction until the explained variance is atleast 90%.

We will calculate explained variance ratio to find how many Principal components are able to explain atleast 90% of variance.

Cumulative Explained Variance Ratio:

```
array([0.62444145, 0.75832974, 0.82435265, 0.87299974, 0.90640271,
       0.92663251, 0.94393397, 0.95207264, 0.95902156, 0.96466793,
       0.96945356, 0.97358126, 0.97675877, 0.97972332, 0.98215096,
       0.98445448, 0.98627285, 0.98794626, 0.98945019, 0.99086751,
       0.99202391, 0.99312884, 0.99397446, 0.99477935, 0.99554613,
       0.9961055 , 0.99660681, 0.99708936, 0.99749984, 0.99788572,
       0.99821413, 0.99849265, 0.99873781, 0.99894611, 0.99914077,
       0.99929979, 0.99942681, 0.99953668, 0.99962348, 0.99970417,
       0.99976476, 0.99980302, 0.99984042, 0.99987407, 0.99989927,
       0.99991853, 0.99993544, 0.99995055, 0.99996197, 0.99997207,
       0.9999797 , 0.99998619, 0.99999156, 0.9999952 , 0.99999762,
       0.99999919, 1.        ])
```

Figure 57 Cumulative explained variance ratio

PC 5 explains 90.6 % of variance. Therefore, we will take 5 principal components.

Taking top 5 Principal Components that explain 90% of covariance:

0	1	2	3	4	
No_HH	0.149222	-0.115487	0.101528	0.076814	-0.012090
TOT_M	0.159169	-0.080239	-0.038662	0.052976	-0.042344
TOT_F	0.158209	-0.093718	0.028959	0.070022	-0.022927
M_06	0.156340	-0.020341	-0.074419	0.028520	-0.080339
F_06	0.156814	-0.014310	-0.068223	0.016398	-0.078326
M_SC	0.143350	-0.079667	-0.037619	0.010210	-0.167893
F_SC	0.143537	-0.087098	0.021350	0.016244	-0.158092
M_ST	0.018849	0.069101	0.323827	0.091143	0.418412
F_ST	0.017878	0.067316	0.338705	0.079554	0.415965
M_LIT	0.155152	-0.105986	-0.032107	0.089187	-0.014033
F_LIT	0.145450	-0.133234	-0.005133	0.125412	0.029084
M_ILL	0.154551	-0.009460	-0.047054	-0.034665	-0.104073

0	1	2	3	4	
F_ILL	0.158283	-0.021793	0.079345	-0.010578	-0.110332
TOT_WORK_M	0.154076	-0.120912	-0.001116	0.069046	-0.023104
TOT_WORK_F	0.142530	-0.076003	0.194130	0.111057	-0.018931
MAINWORK_M	0.141932	-0.166700	0.019821	0.100188	-0.043225
MAINWORK_F	0.125732	-0.142250	0.209976	0.133013	-0.054674
MAIN_CL_M	0.111692	0.042552	0.033131	0.078851	-0.303376
MAIN_CL_F	0.083035	0.095893	0.188822	0.265022	-0.257925
MAIN_AL_M	0.119291	-0.053342	0.225831	-0.121379	-0.253131
MAIN_AL_F	0.090089	-0.072467	0.356566	-0.020989	-0.199220
MAIN_HH_M	0.141850	-0.101835	-0.102202	-0.021969	-0.060812
MAIN_HH_F	0.133880	-0.113257	0.021613	-0.045436	-0.023063
MAIN_OT_M	0.122762	-0.203602	-0.028144	0.147025	0.069907
MAIN_OT_F	0.116866	-0.205899	0.069034	0.155917	0.106774
MARGWORK_M	0.156656	0.079039	-0.068685	-0.078572	0.065812
MARGWORK_F	0.148695	0.108813	0.104957	0.015788	0.077624
MARG_CL_M	0.088163	0.271522	-0.104745	0.157104	-0.018005
MARG_CL_F	0.065160	0.275398	-0.036325	0.285024	-0.055152
MARG_AL_M	0.127278	0.156579	0.070434	-0.250594	-0.047200
MARG_AL_F	0.115888	0.135048	0.259987	-0.153798	-0.012643
MARG_HH_M	0.145366	0.040974	-0.144347	-0.167540	0.005575

0	1	2	3	4	
MARG_HH_F	0.142302	0.006685	-0.093838	-0.151469	0.043616
MARG_OT_M	0.150877	-0.073440	-0.131415	0.021195	0.145109
MARG_OT_F	0.148018	-0.088361	-0.053883	0.059961	0.190756
MARGWORK_3_6_M	0.157908	-0.044044	-0.066877	0.039319	-0.059886
MARGWORK_3_6_F	0.155831	-0.092383	-0.058718	0.046130	-0.022476
MARG_CL_3_6_M	0.157640	0.066208	-0.060172	-0.091315	0.059078
MARG_CL_3_6_F	0.149501	0.089651	0.125792	0.018865	0.064349
MARG_AL_3_6_M	0.094785	0.261268	-0.096551	0.131591	-0.013887
MARG_AL_3_6_F	0.067158	0.266691	-0.018256	0.292845	-0.061019
MARG_HH_3_6_M	0.128184	0.149831	0.078194	-0.250337	-0.058665
MARG_HH_3_6_F	0.113959	0.120648	0.283235	-0.143045	-0.025386
MARG_OT_3_6_M	0.145108	0.036763	-0.142511	-0.166002	0.003315
MARG_OT_3_6_F	0.141029	-0.003685	-0.089356	-0.142599	0.041678
MARGWORK_0_3_M	0.150922	-0.077739	-0.130687	0.019887	0.132794
MARGWORK_0_3_F	0.147534	-0.101141	-0.058489	0.060087	0.170596
MARG_CL_0_3_M	0.142987	0.136839	-0.103565	-0.018223	0.094293
MARG_CL_0_3_F	0.133784	0.166416	0.033423	0.005954	0.112351
MARG_AL_0_3_M	0.062964	0.281881	-0.120293	0.208941	-0.018070
MARG_AL_0_3_F	0.056741	0.287541	-0.088097	0.240499	-0.036293
MARG_HH_0_3_M	0.119102	0.182341	0.026176	-0.240416	0.016981

	0	1	2	3	4
MARG_HH_0_3_F	0.113044	0.177112	0.164774	-0.189408	0.047538
MARG_OT_0_3_M	0.142140	0.052925	-0.144419	-0.167554	0.014187
MARG_OT_0_3_F	0.141370	0.035109	-0.102175	-0.169020	0.047504
NON_WORK_M	0.147629	-0.049122	-0.126673	0.024036	0.191790
NON_WORK_F	0.142103	-0.039848	-0.028545	0.057402	0.249765

Linear Equation for calculating PC:

- We know that principal component is a linear combination of original features. We can write the equation for PC1 in the following manner:

PC1 = (No_HH x 0.149222) + (TOT_M x 0.159169) + (TOT_F x 0.158209) + (M_06 x 0.15634) +

(F_06 x 0.156814) + (M_SC x 0.14335) + (F_SC x 0.143537) + (M_ST x 0.018849) + (F_ST x 0.017878) + (M_LIT x 0.155152) + (F_LIT x 0.14545) + (M_ILL x 0.154551) + (F_ILL x 0.158283) + (TOT_WORK_M x 0.154076) + (TOT_WORK_F x 0.14253) + (MAINWORK_M x 0.141932) + (MAINWORK_F x 0.125732) + (MAIN_CL_M x 0.111692) + (MAIN_CL_F x 0.083035) + (MAIN_AL_M x 0.119291) + (MAIN_AL_F x 0.090089) + (MAIN_HH_M x 0.14185) + (MAIN_HH_F x 0.13388) + (MAIN_OT_M x 0.122762) + (MAIN_OT_F x 0.116866) + (MARGWORK_M x 0.156656) + (MARGWORK_F x 0.148695) + (MARG_CL_M x 0.088163) + MARG_CL_F x 0.06516 + MARG_AL_M x 0.127278 + MARG_AL_F x 0.115888 + MARG_HH_M x 0.145366 + MARG_HH_F x 0.142302 + MARG_OT_M x 0.150877 + MARG_OT_F x 0.148018 + MARGWORK_3_6_M x 0.157908 + MARGWORK_3_6_F x 0.155831 + MARG_CL_3_6_M x 0.15764 + MARG_CL_3_6_F x 0.149501 + MARG_AL_3_6_M x 0.094785 + MARG_AL_3_6_F x 0.067158 + MARG_HH_3_6_M x 0.128184 + MARG_HH_3_6_F x 0.113959 + MARG_OT_3_6_M x 0.145108 + MARG_OT_3_6_F x 0.141029 + MARGWORK_0_3_M x 0.150922 + MARGWORK_0_3_F x 0.147534 + MARG_CL_0_3_M x 0.142987 + MARG_CL_0_3_F x 0.133784 + MARG_AL_0_3_M x 0.062964 + MARG_AL_0_3_F x 0.056741 + MARG_HH_0_3_M x 0.119102 + MARG_HH_0_3_F x 0.113044 + MARG_OT_0_3_M x 0.14214 + MARG_OT_0_3_F x 0.14137 + + NON_WORK_M x 0.147629 + NON_WORK_F x 0.142103

To understand this formula and simplify for understanding,

	PC1	PC2	PC3
A	X1	Y1	Z1
B	X2	Y2	Z2
C	X3	Y3	Z3
D	X4	Y4	Z4
E	X5	Y5	Z5

Then, **PC 1 = (A x X1) + (B x X2) + (C x X3) + (D x X4) + (E x X5)**

Identifying correlations in columns in Principal components using heat map:

Identifying PC's based on highest correlation found in Heat Map:

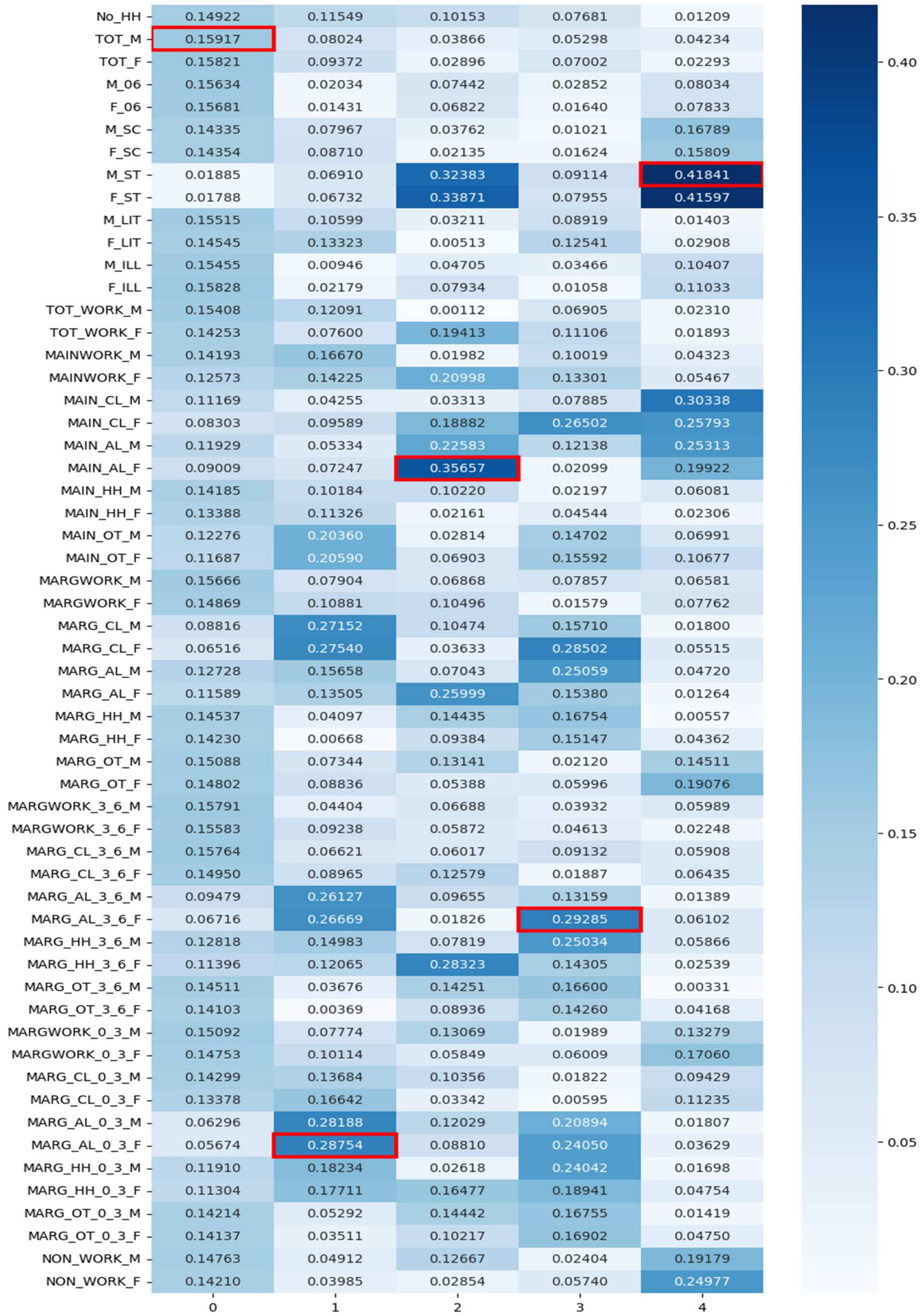
PC1: TOT_M has highest weightage in the columns which represents Total Male population

PC2: MARG_AL_0_3_F has highest weightage which represents Marginal Agriculture Labourers Population 0-3 Female

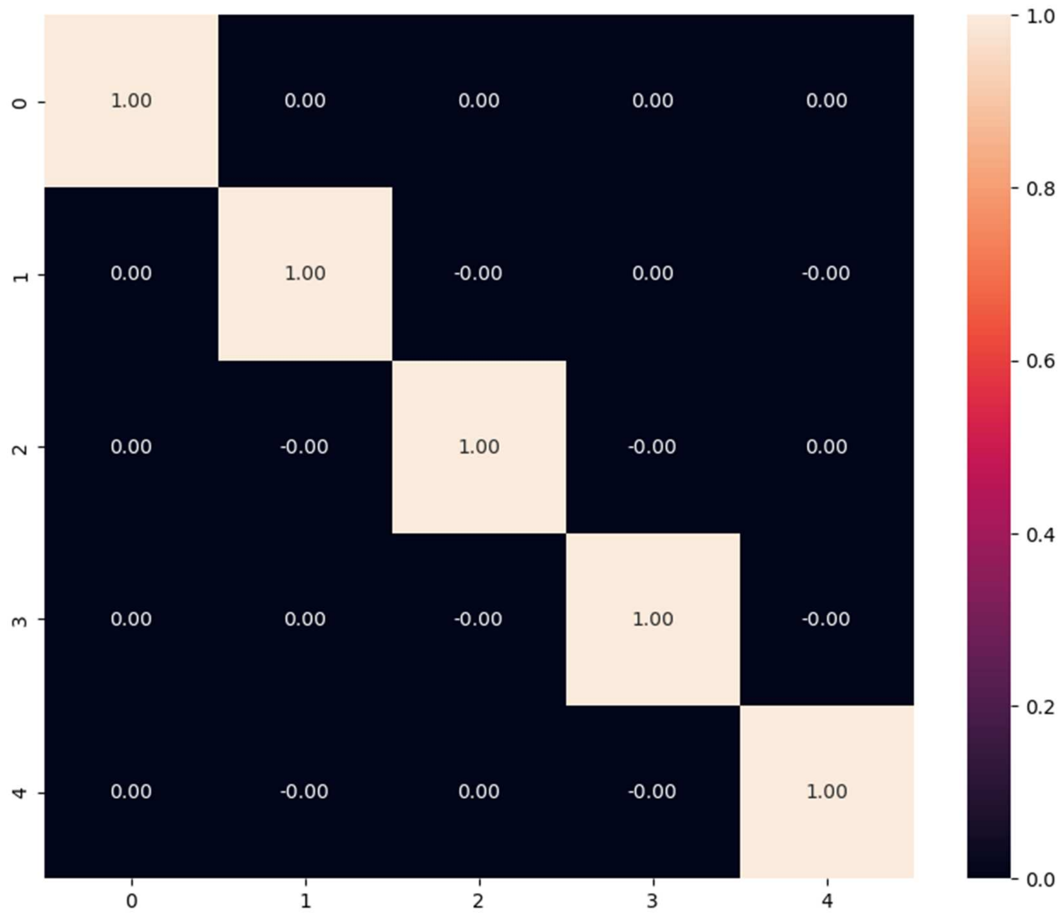
PC3: MAIN_AL_F has highest weightage which represents Main Agricultural Labourers Population Female

PC4: MARG_AL_3_6_F has highest weightage which represents Marginal Agriculture Labourers Population 3-6 Female

PC5: M_ST has highest weightage which represents Scheduled Tribes population Male


No_HH 0.14922 0.11549 0.10153 0.07681 0.01209
TOT_M 0.15917 0.08024 0.03866 0.05298 0.04234
TOT_F 0.15821 0.09372 0.02896 0.07002 0.02293
M_06 0.15634 0.02034 0.07442 0.02852 0.08034
F_06 0.15681 0.01431 0.06822 0.01640 0.07833
M_SC 0.14335 0.07967 0.03762 0.01021 0.16789
F_SC 0.14354 0.08710 0.02135 0.01624 0.15809
M_ST 0.01885 0.06910 0.32383 0.09114 0.41841
F_ST 0.01788 0.06732 0.33871 0.07955 0.41597
M_LIT 0.15515 0.10599 0.03211 0.08919 0.01403
F_LIT 0.14545 0.13323 0.00513 0.12541 0.02908
M_ILL 0.15455 0.00946 0.04705 0.03466 0.10407
F_ILL 0.15828 0.02179 0.07934 0.01058 0.11033
TOT_WORK_M 0.15408 0.12091 0.00112 0.06905 0.02310
TOT_WORK_F 0.14253 0.07600 0.19413 0.11106 0.01893
MAINWORK_M 0.14193 0.16670 0.01982 0.10019 0.04323
MAINWORK_F 0.12573 0.14225 0.20998 0.13301 0.05467
MAIN_CL_M 0.11169 0.04255 0.03313 0.07885 0.30338
MAIN_CL_F 0.08303 0.09589 0.18882 0.26502 0.25793
MAIN_AL_M 0.11929 0.05334 0.22583 0.12138 0.25313
MAIN_AL_F 0.09009 0.07247 0.35657 0.02099 0.19922
MAIN_HH_M 0.14185 0.10184 0.10220 0.02197 0.06081
MAIN_HH_F 0.13388 0.11326 0.02161 0.04544 0.02306
MAIN_OT_M 0.12276 0.20360 0.02814 0.14702 0.06991
MAIN_OT_F 0.11687 0.20590 0.06903 0.15592 0.10677
MARGWORK_M 0.15666 0.07904 0.06868 0.07857 0.06581
MARGWORK_F 0.14869 0.10881 0.10496 0.01579 0.07762
MARG_CL_M 0.08816 0.27152 0.10474 0.15710 0.01800
MARG_CL_F 0.06516 0.27540 0.03633 0.28502 0.05515
MARG_AL_M 0.12728 0.15658 0.07043 0.25059 0.04720
MARG_AL_F 0.11589 0.13505 0.25999 0.15380 0.01264
MARG_HH_M 0.14537 0.04097 0.14435 0.16754 0.00557
MARG_HH_F 0.14230 0.00668 0.09384 0.15147 0.04362
MARG_OT_M 0.15088 0.07344 0.13141 0.02120 0.14511
MARG_OT_F 0.14802 0.08836 0.05388 0.05996 0.19076
MARGWORK_3_6_M 0.15791 0.04404 0.06688 0.03932 0.05989
MARGWORK_3_6_F 0.15583 0.09238 0.05872 0.04613 0.02248
MARG_CL_3_6_M 0.15764 0.06621 0.06017 0.09132 0.05908
MARG_CL_3_6_F 0.14950 0.08965 0.12579 0.01887 0.06435
MARG_AL_3_6_M 0.09479 0.26127 0.09655 0.13159 0.01389
MARG_AL_3_6_F 0.06716 0.26669 0.01826 0.29285 0.06102
MARG_HH_3_6_M 0.12818 0.14983 0.07819 0.25034 0.05866
MARG_HH_3_6_F 0.11396 0.12065 0.28323 0.14305 0.02539
MARG_OT_3_6_M 0.14511 0.03676 0.14251 0.16600 0.00331
MARG_OT_3_6_F 0.14103 0.00369 0.08936 0.14260 0.04168
MARGWORK_0_3_M 0.15092 0.07774 0.13069 0.01989 0.13279
MARGWORK_0_3_F 0.14753 0.10114 0.05849 0.06009 0.17060
MARG_CL_0_3_M 0.14299 0.13684 0.10356 0.01822 0.09429
MARG_CL_0_3_F 0.13378 0.16642 0.03342 0.00595 0.11235
MARG_AL_0_3_M 0.06296 0.28188 0.12029 0.20894 0.01807
MARG_AL_0_3_F 0.05674 0.28754 0.08810 0.24050 0.03629
MARG_HH_0_3_M 0.11910 0.18234 0.02618 0.24042 0.01698
MARG_HH_0_3_F 0.11304 0.17711 0.16477 0.18941 0.04754
MARG_OT_0_3_M 0.14214 0.05292 0.14442 0.16755 0.01419
MARG_OT_0_3_F 0.14137 0.03511 0.10217 0.16902 0.04750
NON_WORK_M 0.14763 0.04912 0.12667 0.02404 0.19179
NON_WORK_F 0.14210 0.03985 0.02854 0.05740 0.24977
0 1 2 3 4
0.40 0.35 0.30 0.25 0.20 0.15 0.10 0.05

Checking correlation between the 5 PC:



We can see that there is no correlation among the principal components and forms an orthogonal identity matrix.

Thus, we were able to reduce the variables successfully from 57 to only 5. These 5 Principal Compenents are able to explain atleast 90% of variance in the data.

Thank you!