Machine Learning - 1

Coded Project

SRINIVASAN T

Table of Content:

List of Tables:

List of Figures:

Rubric:

Criteria	Points
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	6.5
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.	2.5
Part 1: Clustering: Hierarchical Clustering - Construct a dendrogram using Ward linkage and Euclidean distance - Identify the optimum number of Clusters	4
Part 1: Clustering: K-means Clustering - Apply K-means Clustering - Plot the Elbow curve - Check Silhouette Scores - Figure out the appropriate number of clusters - Cluster Profiling	13
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.	6
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?	6.5
Part 2: PCA: Data Preprocessing	2.5

Criteria	Points
- 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	
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.	13
Quality of Business Report	6

Part1

1.1 Problem definition

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.

Loading Dataset:

Loading the dataset and check whether it is properly loaded using the head function.

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

Table 1: Loading the Dataset

The dataset is loaded properly

1.2 Check shape

The dataset has 23,066 rows and 19 columns

1.3 Data types

The datatypes can be identified using the info function.

```
<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 1: Data Types

The dataset has 1 - Date time variable, 5 - Categorical variables, 13 – Numerical variable. Except CTR, CPM & CPC variables all the other variables does not have null values.

1.4 Statistical summary

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	385.16	233.65	120.00	120.00	300.00	720.00	728.00
Ad- Width	23066.0	337.90	203.09	70.00	250.00	300.00	600.00	600.00
Ad Size	23066.0	96674.47	61538.33	33600.00	72000.00	72000.00	84000.00	216000.00
Available_Impressions	23066.0	2432043.67	4742887.76	1.00	33672.25	483771.00	2527711.75	27592861.00
Matched_Queries	23066.0	1295099.14	2512969.86	1.00	18282.50	258087.50	1180700.00	14702025.00
Impressions	23066.0	1241519.52	2429399.96	1.00	7990.50	225290.00	1112428.50	14194774.00
Clicks	23066.0	10678.52	17353.41	1.00	710.00	4425.00	12793.75	143049.00
Spend	23066.0	2706.63	4067.93	0.00	85.18	1425.12	3121.40	26931.87
Fee	23066.0	0.34	0.03	0.21	0.33	0.35	0.35	0.35
Revenue	23066.0	1924.25	3105.24	0.00	55.37	926.34	2091.34	21276.18
CTR	18330.0	0.07	0.08	0.00	0.00	0.08	0.13	1.00
CPM	18330.0	7.67	6.48	0.00	1.71	7.66	12.51	81.56
CPC	18330.0	0.35	0.34	0.00	0.09	0.16	0.57	7.26

Table 2: Statistical Summary – Numerical

- Ad Size has 25% & 50% are same
- Available Impressions , Matched Queries, Impressions & Clicks have min value as 1
- Available Impressions , Matched Queries, Impressions has higher standard deviation values
- All the numerical variables are at the different scale of measures

	count	unique	top	freq
Timestamp	23066	2018	2020-11-13-22	13
InventoryType	23066	7	Format4	7165
Ad Type	23066	14	Inter224	1658
Platform	23066	3	Video	9873
Device Type	23066	2	Mobile	14806
Format	23066	2	Video	11552

Table 3: Statistical Summary – Categorical

- Ad Type has 14 types which the highest with Inter224 types as the highest no. of ad types
- There are only 3 platforms
- Mobile type is the highest usage among devices

1.5 Univariate Analysis

1.5.1 Univariate Analysis – Numerical

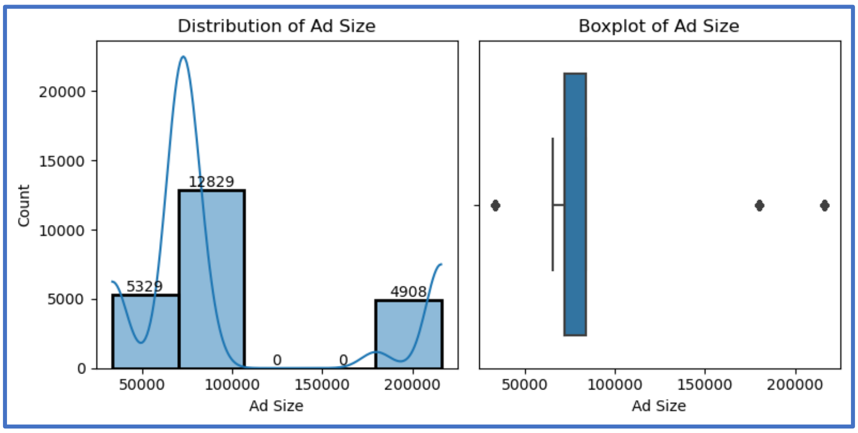


Figure 2: Distribution of Ad Size

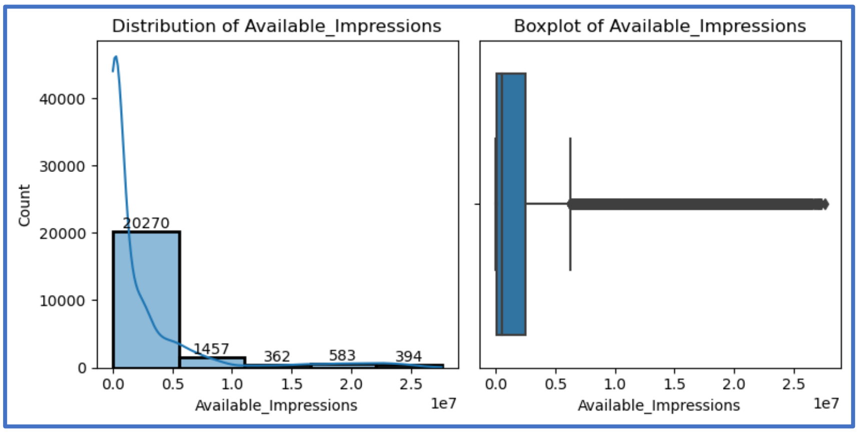


Figure 3: Distribution of Available Impressions

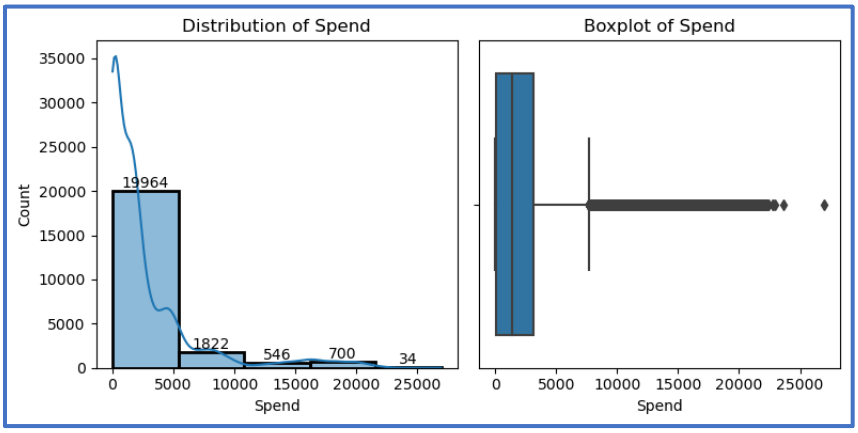


Figure 4: Distribution of Available Impressions

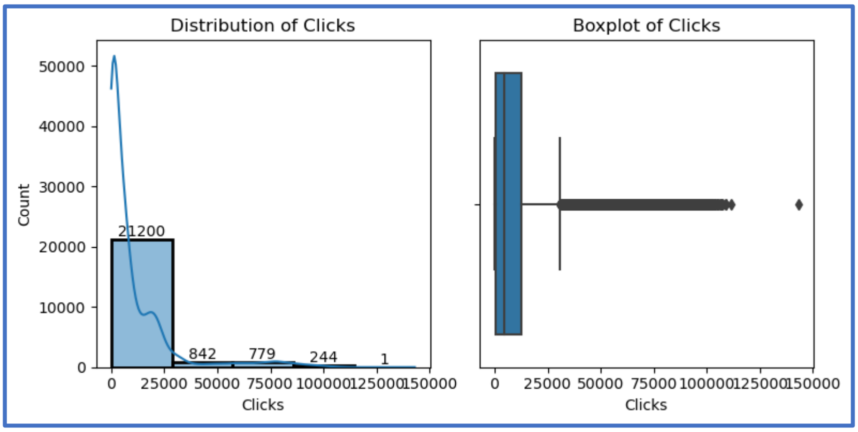


Figure 5: Distribution of Clicks

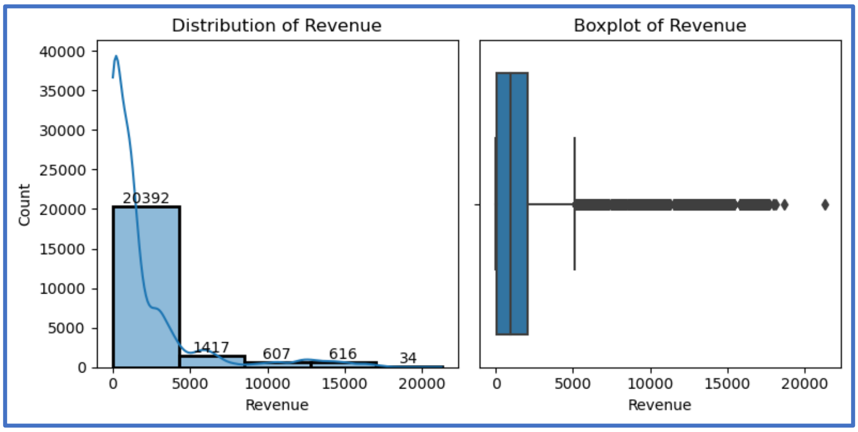


Figure 6: Distribution of Revenue

- Ad – Size: The distribution is right skewed. Does not have UL since the difference between the 75% and the max value is huge
- Available – Impressions: Right Skewed as mean>median. Having too many outliers
- Clicks: The distribution is right skewed. The Std is high suggests that the clicks of the ads vary notably around the mean.
- Spend: 75% of the spend was around 3K but the max is high as 27K
- Revenue: There are Ads with 0 revenue. The distribution is left skewed as the mean<median

1.5.2 Univariate Analysis – Categorical

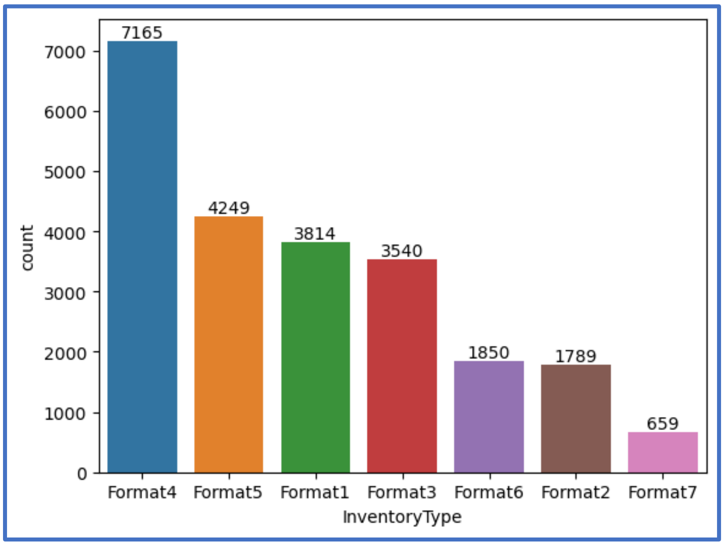


Figure 7: Count of Inventory Type

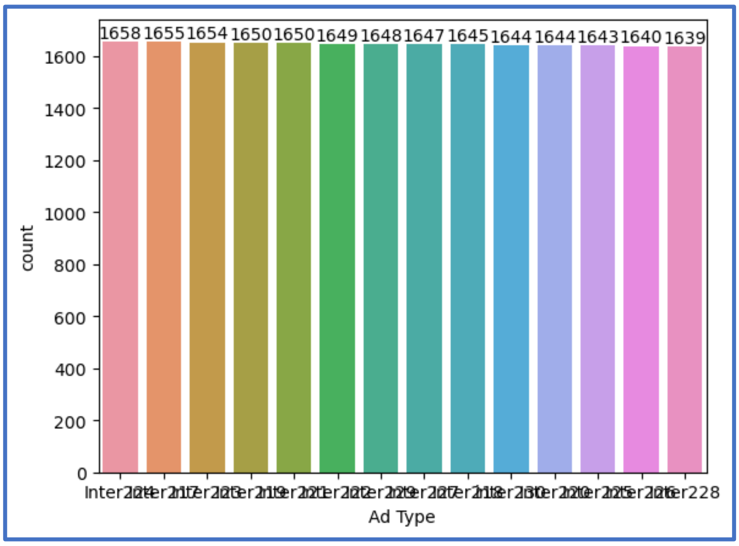


Figure 8: Count of Ad Type

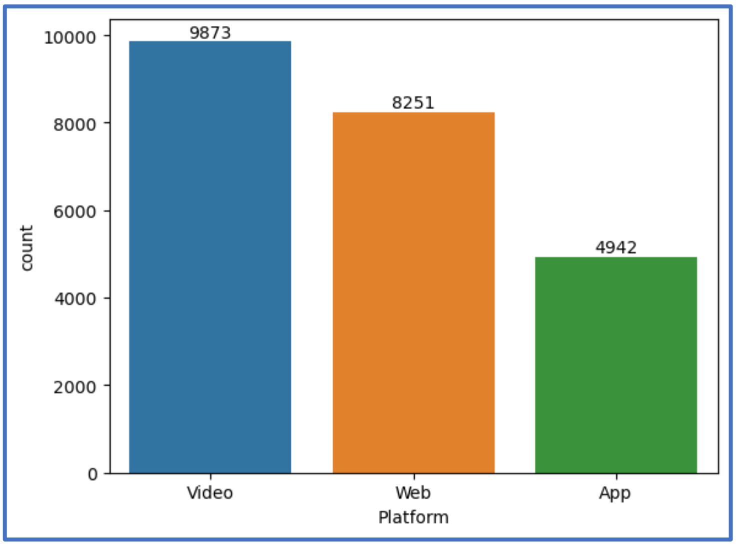


Figure 9: Count of Platform

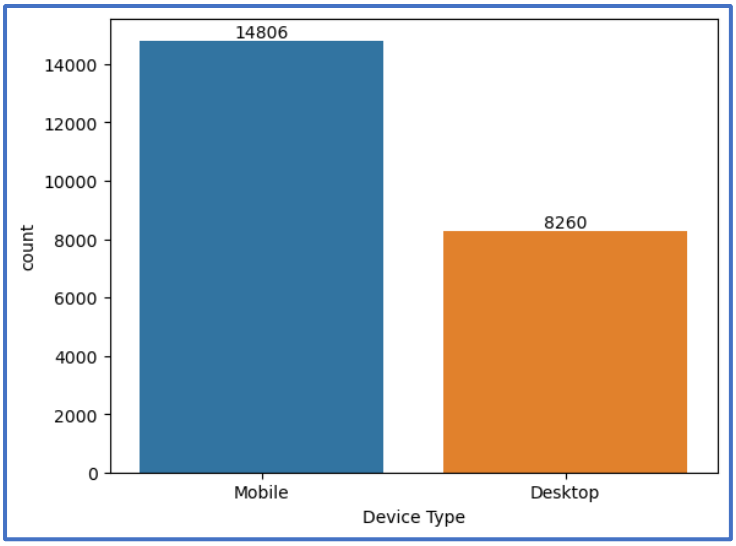


Figure 10: Count of Device Type

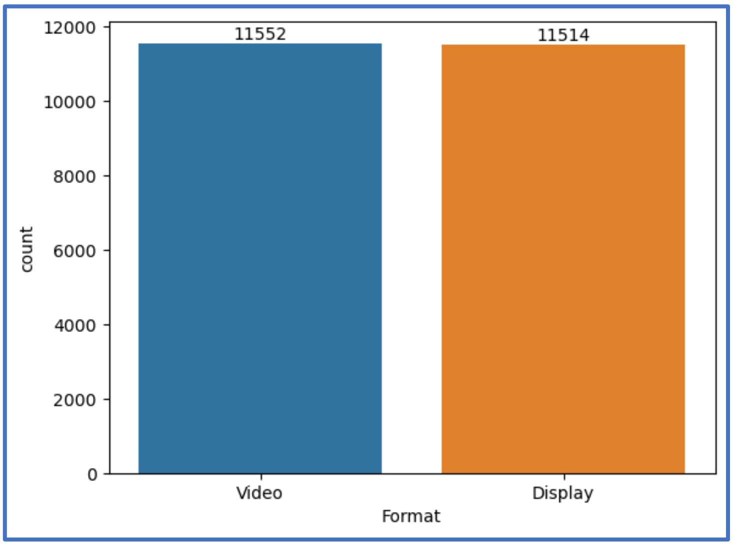


Figure 11: Count of Format

- Inventory Type: Format 4 is having the higher count while format 7 with the lowest count
- Ad Type: Almost all the types are same in count
- Platform: Video platform contributes to the highest count
- Device Type: Obviously mobile device has higher contribution

1.6 Bivariate analysis

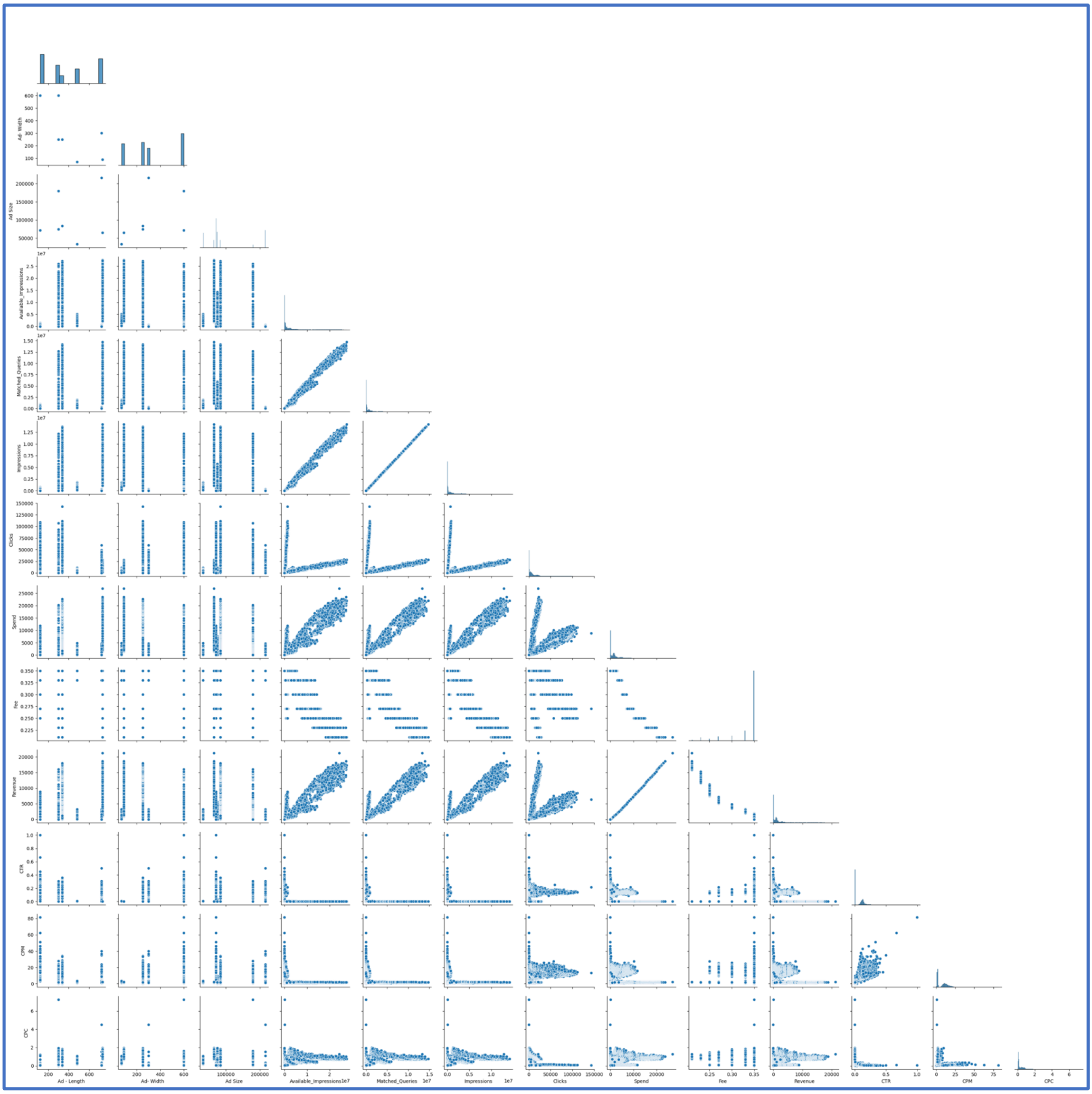


Figure 12: Bivariate Analysis – Numerical

Available_ Impression variable has positive correlation with Revenue, Spend, Matched Queries, Impressions variables
CTR is having positive correlation with CPM
Fee variable is having a negative correlation with Spend & Revenue variables

1.7 Missing Value 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                     4736
CPM                     4736
CPC                     4736
dtype: int64
```

Table 4: Missing Values

CTR, CPM & CPC have missing values

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

Missing values are replaced with NaN

d - th	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
20	300	216000	Inter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
20	300	216000	Inter224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.0260	NaN	NaN	NaN
20	300	216000	Inter218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	NaN	NaN	NaN
20	600	72000	inter230	Video	Mobile	Video	7	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
20	300	216000	Inter221	App	Mobile	Video	2	2	2	1	0.09	0.35	0.0585	NaN	NaN	NaN

Table 5: Missing Replacement with NaN

Filling the CTR columns missing value using the formula CTR = (Clicks/Impressions)*100
Filling the CPM columns missing value using the formula CPM = (Spend/Impressions)*1000
Filling the CPC columns missing value using the formula CPC = Spend/Clicks

Ad-Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
300	216000	Inter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	100.0	70.0	0.07
300	216000	Inter224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.0260	50.0	20.0	0.04
300	216000	Inter218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	100.0	50.0	0.05
600	72000	inter230	Video	Mobile	Video	7	1	1	1	0.07	0.35	0.0455	100.0	70.0	0.07
300	216000	Inter221	App	Mobile	Video	2	2	2	1	0.09	0.35	0.0585	50.0	45.0	0.09

Table 6: Missing Replacement with formula given

1.9 Outlier Treatment

Before:

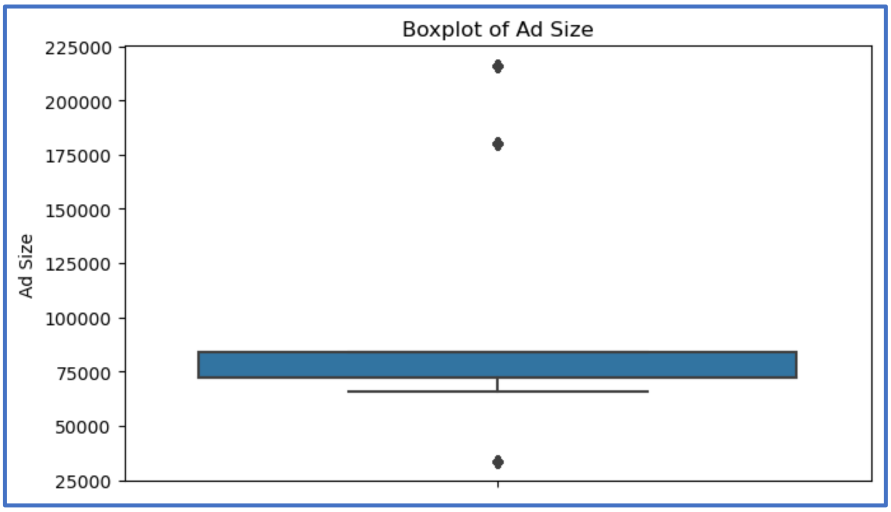


Figure 13: Outlier Treatment Before – Ad Size

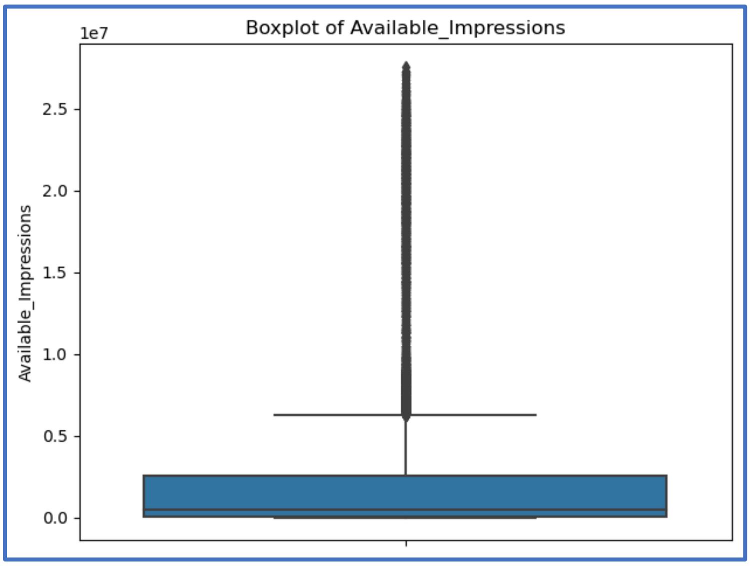


Figure 14: Outlier Treatment Before – Available Impressions

After:

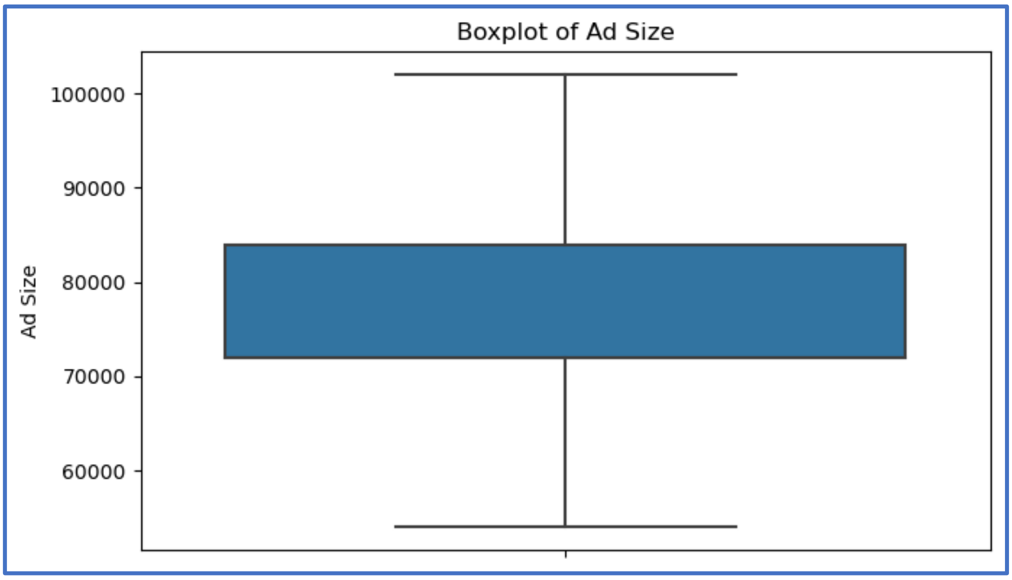


Figure 15: Outlier Treatment After – Ad Size

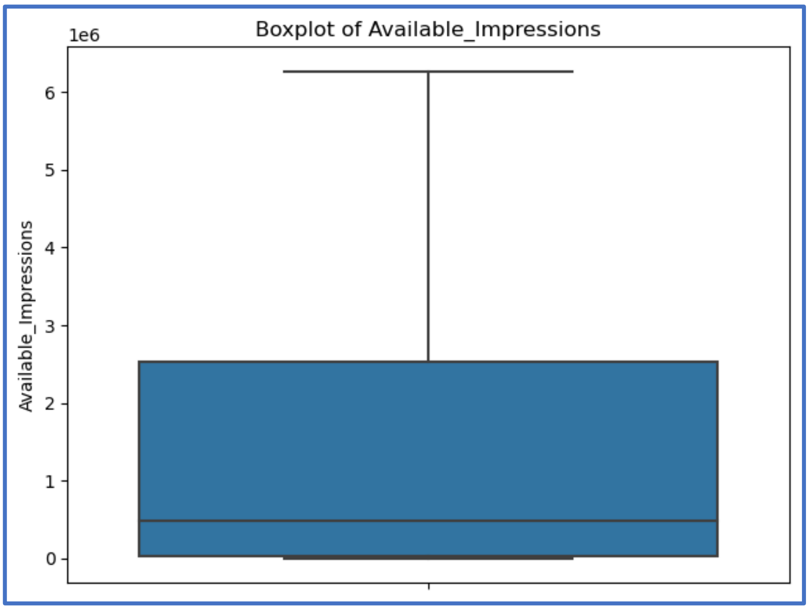


Figure 16: Outlier Treatment After – Available Impressions

Note: Only relevant numerical variables are used for the outlier treatment and scaling

1.10 Scaling using Zscore

	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	-0.102518	-0.755333	-0.778949	-0.768478	-0.867488	-0.89317	0.535724	-0.880093	-0.891201	-1.194562	-1.04114
1	-0.102518	-0.755345	-0.778988	-0.768516	-0.867488	-0.89317	0.535724	-0.880093	-0.888615	-1.194562	-1.04114
2	-0.102518	-0.754900	-0.778919	-0.768445	-0.867488	-0.89317	0.535724	-0.880093	-0.893142	-1.194562	-1.04114
3	-0.102518	-0.755040	-0.778781	-0.768302	-0.867488	-0.89317	0.535724	-0.880093	-0.898315	-1.194562	-1.04114
4	-0.102518	-0.755610	-0.779030	-0.768560	-0.867488	-0.89317	0.535724	-0.880093	-0.884734	-1.194562	-1.04114

Table 7: Z score Scaling

1.11 Hierarchical Clustering – Dendrogram using Ward link and Euclidean distance

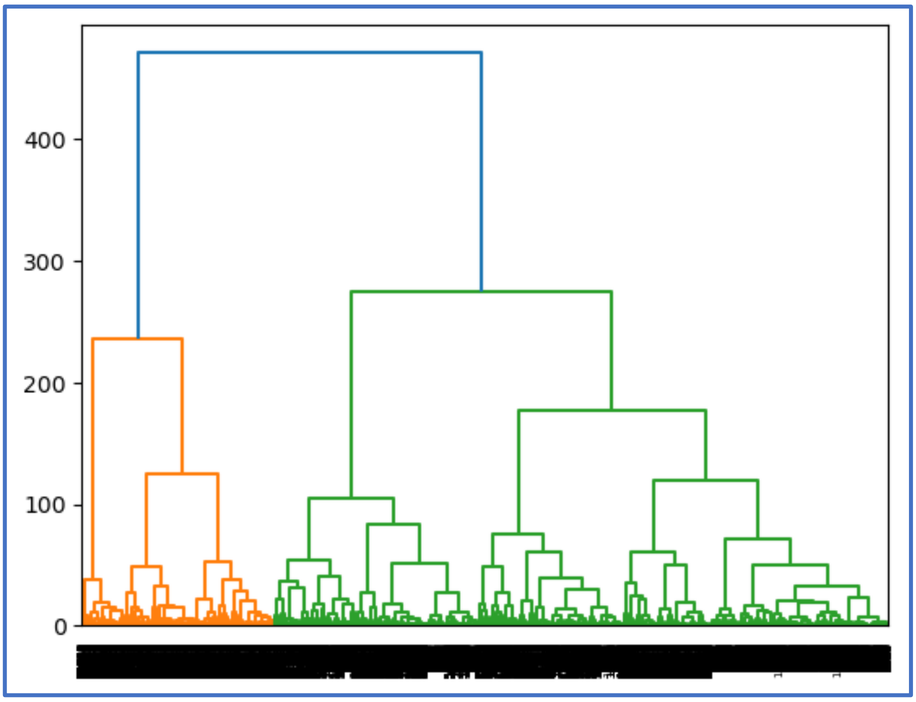


Figure 17: Dendrogram for Hierarchical Clustering

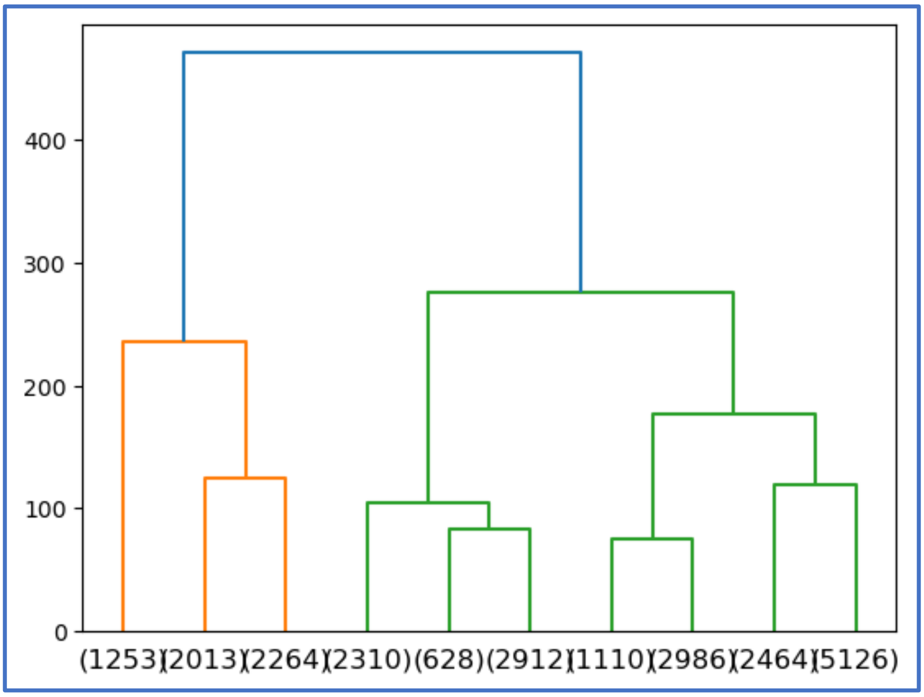


Figure 18: Truncated Dendrogram for Hierarchical Clustering

1.12 Hierarchical Clustering – Number of clusters

h_clusters	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	h_freq
1	386.08	285.04	75153.05	7947289.75	4336783.80	4182648.26	24635.15	8196.52	0.29	6009.55	0.54	4.67	0.61	5530
2	430.18	146.78	54068.02	1818648.70	860143.97	820524.22	3337.57	1492.46	0.35	970.70	0.06	1.85	0.56	5850
3	362.19	458.58	128187.49	129206.91	73463.99	60480.72	7748.89	716.53	0.35	468.37	4.87	13.43	0.10	11686

Table 8: Hierarchical Clusters and their means

1.13 K-means Clustering

Fixing the no. of clusters as 2 and finding the label & inertia

Label:

array([0, 0, 0, ..., 0, 0, 0], dtype=int32)

Inertia:

142414.4718775063

1.14 Plot the Elbow curve

Finding the inertia for a range for 10 clusters

```
[253726.00000000067,
 142414.39715260785,
 104382.57711174723,
 74718.29260909933,
 59611.25245495574,
 52041.85047517175,
 44828.35714042672,
 39347.87984288174,
 36409.57337157424,
 32970.992896818825]
```

Table 9: Inertia for 10 clusters

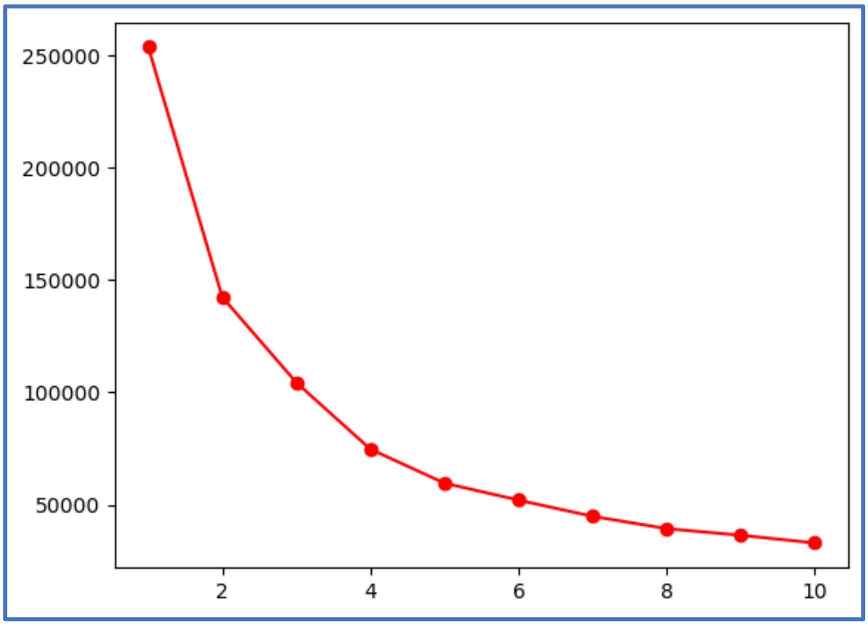


Figure 19: Elbow curve for 10 clusters in K means

1.15 Check Silhouette Scores

```
[0.4598081037587795,
 0.3692143940352544 7,
 0.42467380036745245,
 0.416371238916712 8,
 0.400836605751247 8,
 0.399774193434378 7,
 0.41005578380881896,
 0.426922527345338 5,
 0.411894803403356]
```

Table 10: Silhouette scores for 9 clusters

1.16 Figure out the appropriate number of clusters

k_clusters	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	k_freq
0	425.14	149.82	54588.81	1872875.16	894682.56	855411.68	3371.33	1553.04	0.35	1012.08	0.06	1.84	0.56	6103
1	364.74	457.79	128929.32	117382.32	65125.17	53392.02	6731.76	609.04	0.35	395.94	4.96	13.40	0.10	11290
2	466.27	199.01	75182.01	10416675.83	5641383.47	5462313.17	11274.42	8668.88	0.29	6391.03	0.03	1.57	0.76	4037
3	176.88	554.83	84117.36	788505.69	551944.35	465866.51	63703.59	6772.91	0.29	4851.70	2.33	15.20	0.11	1636

Table 11: K_Means Clusters and their means

1.18 Clustering: Actionable Insights & Recommendations

```
Format   Device Type  h_clusters
Display  Desktop      3             2076
                      2             1020
                      1             1018
         Mobile       3             3748
                      2             1871
                      1             1781
Video    Desktop      3             2118
                      2             1044
                      1              984
         Mobile       3             3744
                      2             1915
                      1             1747
Name: h_clusters, dtype: int64
```

```
Device Type  Platform  h_clusters
Desktop      Video     3             2501
                       2             1255
                       1             1191
             Web       3             1693
                       1              811
                       2              809
Mobile       App       3             2494
                       2             1265
                       1             1183
             Video     3             2487
                       2             1242
                       1             1197
             Web       3             2511
                       2             1279
                       1             1148
Name: h_clusters, dtype: int64
```

Hierarchical Clustering:

Cluster/Avg.	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
Cluster 1	High	High	High	Low	High	Medium	Medium	High
Cluster 2	Medium	Low	Medium	High	Medium	Low	Low	Medium
Cluster 3	Low	Medium	Low	High	Low	High	High	Low

Cluster 1: ‘High Impressions & High Revenue’
Cluster 2: ‘Moderate Impressions & Low CPM ‘
Cluster 3: ‘Low Impressions & High CTR’

Insights:

Cluster 1:

- Having high average impressions, clicks and revenue
- Mobile impressions of the static post is higher than desktop
- The cost per click is higher

Cluster 2:

- Having moderate Impressions & Low Clicks
- The CPM is low which will increase the ROI of the ads
- The mobile device contribution is higher

Cluster 3:

- Having Low spend & High CTR
- Higher mobile device usage across different formats od posts

- CPC is low with low spend

Recommendations

Cluster 1:
- ROI is 73% Higher among other clusters
- For higher sales(shop/order actions) it can be effective choice
- Focusing more on the mobile device type

Cluster 2:
- For reaching higher no. of audiences with lower customer reach cost
- For best ROI's these ads can be used
- These ads can be used to increase the reach with lower cost

Cluster 3:
- Increasing the spend can increase the CTR
- Objective of profile visits can be done using this types of ads
- CPC is low with low spend

K-Means Clustering:

k_clusters	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
Cluster 1	2	4	3	1	3	3	3	2
Cluster 2	4	3	4	1	4	1	2	4
Cluster 3	1	2	1	2	1	4	4	1
Cluster 4	3	1	2	2	2	2	1	3

1,2,3,4 – Rank wise values

Cluster 1: 'Moderate CPM & Average Spend'
Cluster 2: 'High CTR & High CPC'
Cluster 3: 'High Impressions & High Revenue'
Cluster 4: 'High Clicks & Low CPM'

Insights:

Cluster 1:

- Having average impressions with moderate revenue
- Since clicks are low with average impression will increase the CTR
- Moderate spend with average impressions will make the CPM high

Cluster 2:

- Having low Impressions & Low spend
- But the click through rate is higher
- With low spend and moderate clicks have high CPC

Cluster 3:

- High impressions and high revenue
- Average clicks with low CTR
- Higher spend with low CPM

Cluster 4:

- Higher clicks with moderate impressions
- Average revenue with moderate spend
- CPM is higher due to moderate impressions

Part2

2.1 Problem Definition

PCA FH (FT): Primary census abstract for female headed households excluding institutional households (India & States/Uts – District Level), Scheduled tribes – 2011 PCA for Female Headed Household Excluding Institutional Household. The Indian Census has the reputation of being one of the best in the world. The first Census in India was conducted in the year 1872. This was conducted at different points of time in different parts of the country. In 1881 a Census was taken for the entire country simultaneously. Since then, Census has been conducted every ten years, without a break. Thus, the Census of India 2011 was the fifteenth in this unbroken series since 1872, the seventh after independence and the second census of the third millennium and twenty first century. The census has been uninterruptedly continued despite of several adversities like wars, epidemics, natural calamities, political unrest, etc. The Census of India is conducted under the provisions of the Census Act 1948 and the Census Rules, 1990. The Primary Census Abstract which is important publication of 2011 Census gives basic information on Area, Total Number of Households, Total Population, Scheduled Castes, Scheduled Tribes Population, Population in the age group 0-6, Literates, Main Workers and Marginal Workers classified by the four broad industrial categories, namely, (i) Cultivators, (ii) Agricultural Laborers, (iii) Household Industry Workers, and (iv) Other Workers and also Non-Workers. The characteristics of the Total Population include Scheduled Castes, Scheduled Tribes, Institutional and Houseless Population and are presented by sex and rural-urban residence. Census 2011 covered 35 States/Union Territories, 640 districts, 5,924 sub-districts, 7,935 Towns and 6,40,867 Villages. The data collected has so many variables thus making it difficult to find useful details without using Data Science Techniques. You are tasked to perform detailed EDA and identify Optimum Principal Components that explains the most variance in data. Use Sklearn only.

2.2 Data loading

Loading the dataset and check whether it is properly loaded using the head function.

	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

Table 2.1: Loading the Dataset

2.3 Checking Shape

The dataset has 640 rows and 61 columns

2.4 Data Types

```
<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
```

All the variables are float datatype and does not have null values

2.5 Selecting only 5 variables

1. F_06 – Population in the age group 0-6 Female
2. F_SC – Scheduled Castes population Female
3. F_ST – Scheduled Tribes population Female
4. F_LIT – Literates population Female
5. F_ILL – Illiterate Female

2.6 Statistical Summary

	count	mean	std	min	25%	50%	75%	max
F_06	640.0	11942.300000	11326.294567	56.0	4672.25	8663.0	15902.25	95129.0
F_SC	640.0	20778.392188	21727.887713	0.0	5603.25	13709.0	29180.00	156429.0
F_ST	640.0	10155.640625	15875.701488	0.0	429.50	3834.5	12480.25	130119.0
F_LIT	640.0	66359.565625	75037.860207	371.0	20932.00	43796.5	84799.75	571140.0
F_ILL	640.0	56012.518750	47116.693769	327.0	22367.00	42386.0	78471.00	254160.0

Table 2.2: Statistical Summary

- Each district have at least 56 female child of age 0-6
- Few states/districts does not have scheduled caste and scheduled tribe female population
- On an average 59K female population are literate in each district/state
- Each district/state have at least 327 female population who are illiterate

2.7 Univariate Analysis

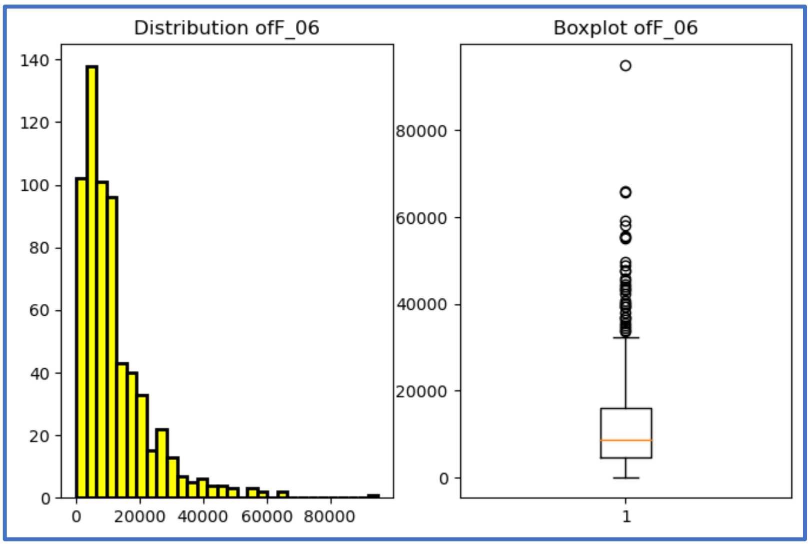


Figure 2.1: Distribution of F_06

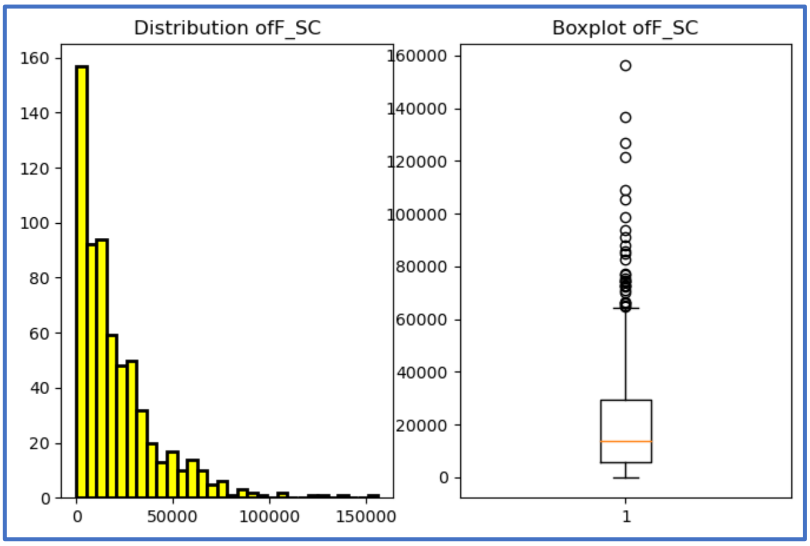


Figure 2.2: Distribution of F_SC

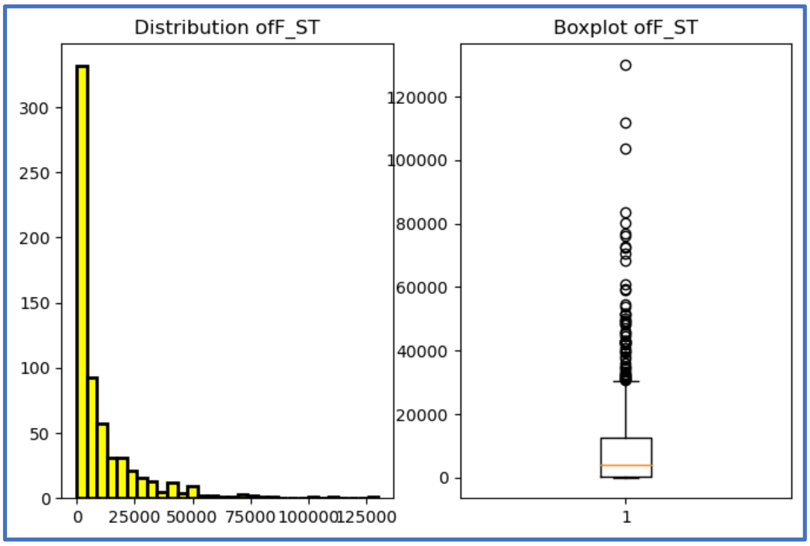


Figure 2.3: Distribution of F_ST

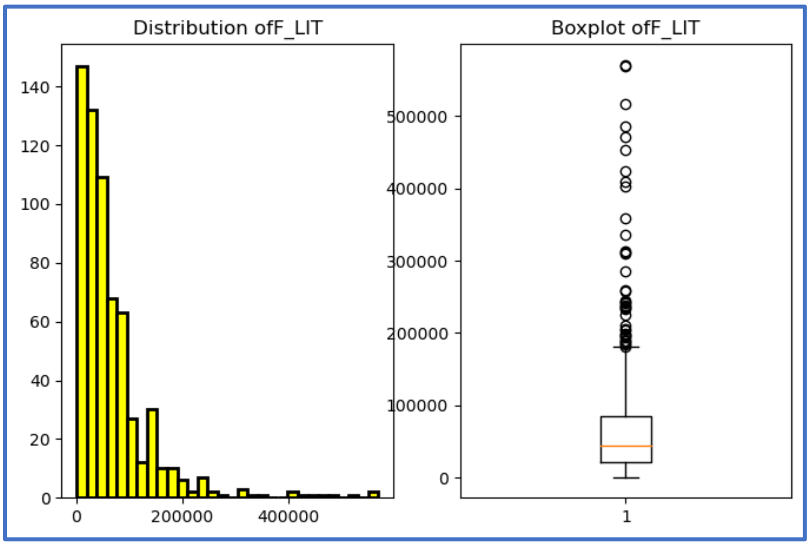


Figure 2.4: Distribution of F_LIT

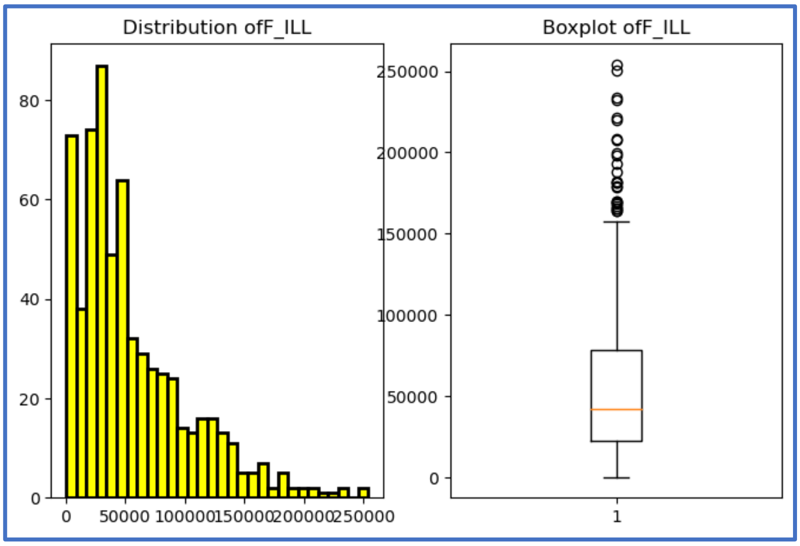


Figure 2.5: Distribution of F_ILL

- F_06 is right skewed and has outliers
- F_SC is right skewed and has outliers. The distance between 75% and max value is higher
- F_ST is right skewed and has outliers. The min value is 0 and the max value goes till 1.3 L
- F_LIT is right skewed and has outliers.
- F_ILL is right skewed and has outliers.

2.8 Bivariate Analysis

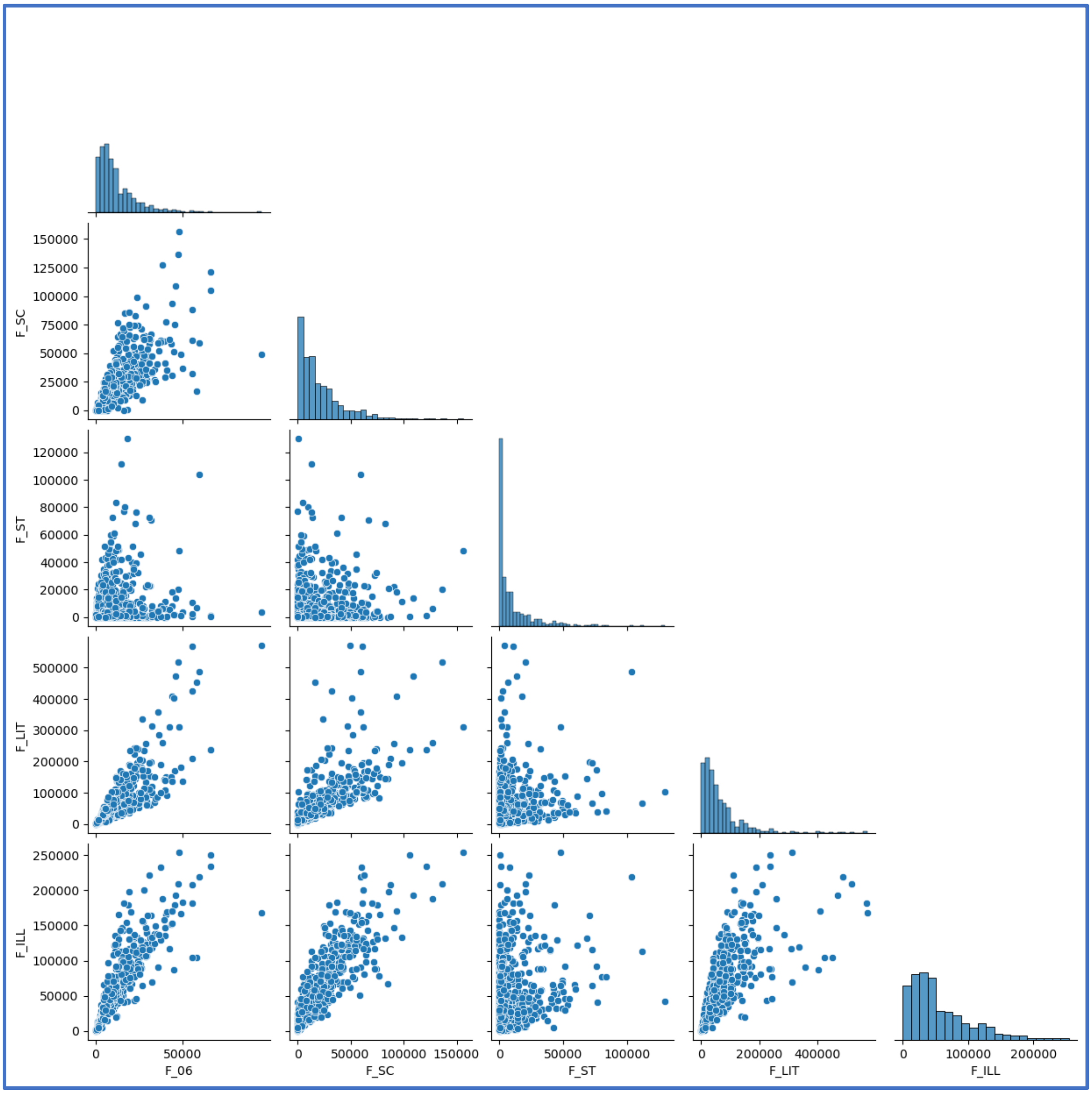


Figure 2.6: Bivariate Analysis

- F_06 with other variables except F_ST has positive correlation
- F_SC has positive correlation with F_ILL, F_LIT & F_06.
- F_ST has less F_LIT rate among the female population
- F_LIT has positive correlation with F_ILL

2.9 Multivariate Analysis

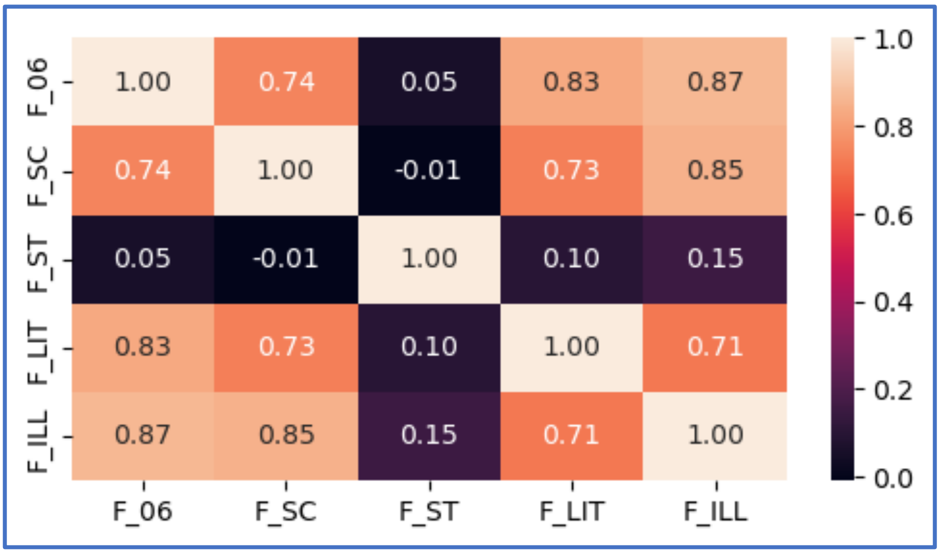


Figure 2.7: Multivariate Analysis

- F_06 has a strong correlation with F_ILL, F_LIT & F_SC
- F_SC has a strong correlation with F_ILL, F_LIT & F_06
- F_ST does not have strong correlation with other 4 variables

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

```
State
Lakshadweep                 0.868061
Haryana                     0.779129
NCT of Delhi                0.775077
Uttar Pradesh               0.752167
Meghalaya                   0.752160
Bihar                       0.744596
Punjab                      0.744502
Jammu & Kashmir             0.735154
Daman & Diu                 0.703143
Chandigarh                  0.700037
Rajasthan                   0.695286
Assam                       0.686561
Jharkhand                   0.681804
Gujarat                     0.674844
Andaman & Nicobar Island    0.652679
West Bengal                 0.650345
Dadara & Nagar Havelli      0.644631
Himachal Pradesh            0.642741
Sikkim                      0.642227
Manipur                     0.641179
Madhya Pradesh              0.639695
Karnataka                   0.637802
Uttarakhand                 0.630865
Tripura                     0.625881
Mizoram                     0.623634
Goa                         0.621648
Kerala                      0.601238
Puducherry                  0.591111
Maharashtra                 0.587812
Nagaland                    0.583682
Odisha                      0.575500
Arunachal Pradesh           0.574365
Chhattisgarh                0.549200
Tamil Nadu                  0.547921
Andhra Pradesh              0.537024
dtype: float64
```

Table 2.3: State Wise Gender Ratio

Highest Gender Ratio – State: Lakshadweep
Lowest Gender Ratio – State: Andhra Pradesh

2.11 Which district has the highest & lowest gender ratio?

```
Area Name
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
Length: 635, dtype: float64
```

Table 2.4: District Wise Gender Ratio

Highest Gender Ratio – District: Lakshadweep
Lowest Gender Ratio – District: Krishna

2.12 Data Pre-processing

Checking the Statistical Summary and Datatypes for the entire dataset

2.13 Checking for and treat duplicates

```
The no. of duplicated rows are 0
```

2.14 Checking for and treat bad data

No bad data found

2.15 Checking for and treat anomalies

No anomalies found

2.16 Checking for and treat missing values

```
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 2.5: Missing Values

2.17 Dropping of irrelevant variables/columns

Dropping the variables

State Code, Dist.Code, State, Area Name

2.18 Check for and treat outliers

Before Scaling

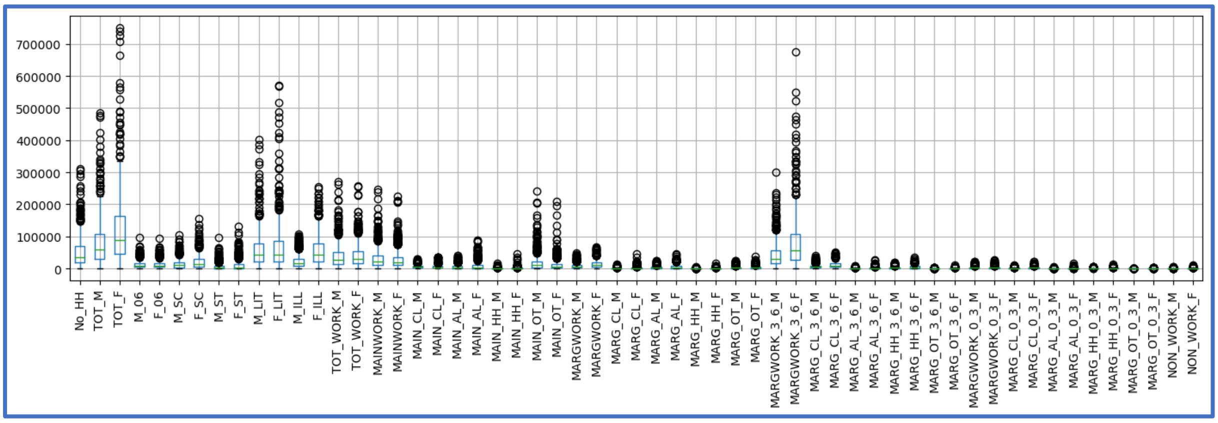


Figure 2.8: Outlier Before Scaling

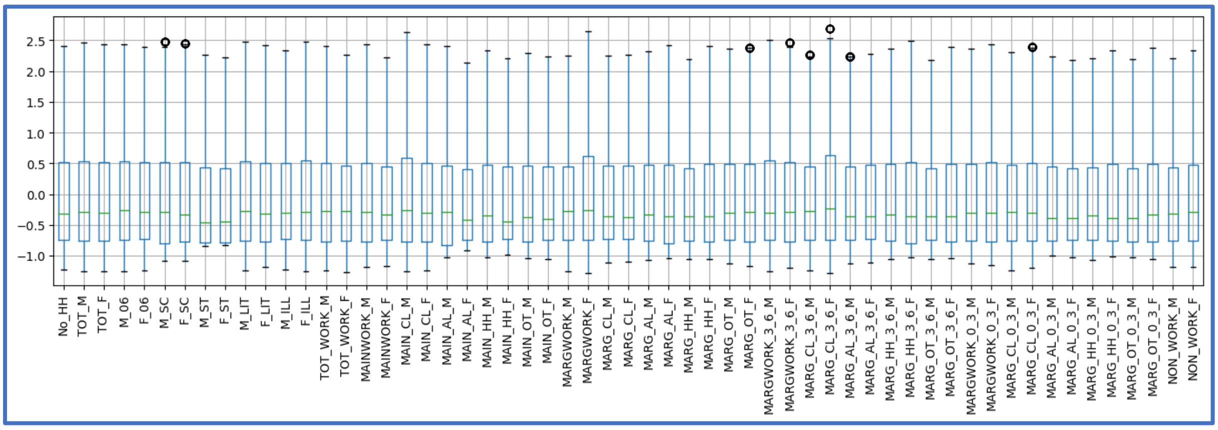


Figure 2.9: Outlier After Scaling

There are very few outliers after scaling the data comparatively before scaling

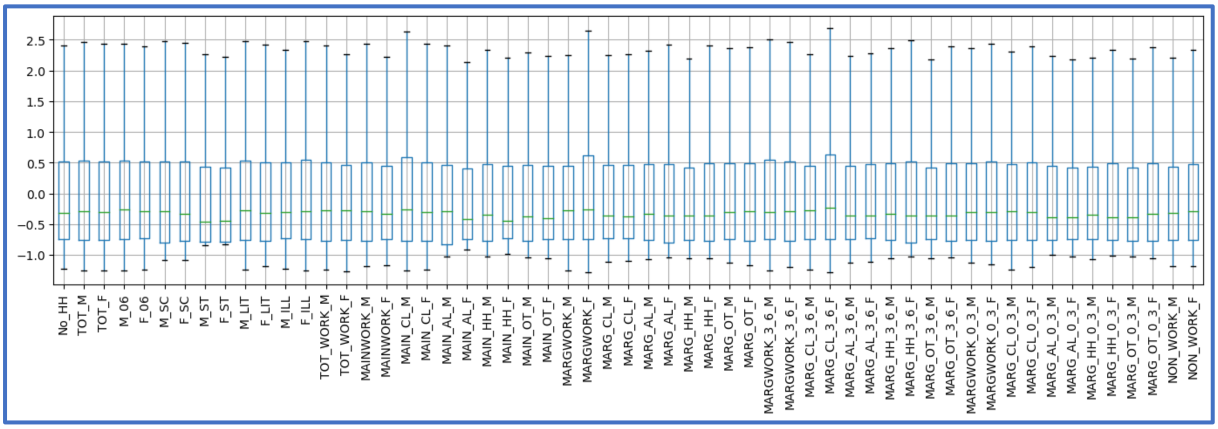


Figure 2.10: After Outlier Treatment

Outliers are capped to their Lower & Upper Limit

2.19 Zscore Scaling

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_
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	

5 rows × 57 columns

Table 2.6: Zscore Scaling

PCA

2.20 Checking the statistical significance of correlations

H0: Correlations are not significant
H1: There are significant correlations

Reject H0 if p-value < 0.05

```
The p value is 0.0
```

Since the p value is less than we reject the H0 and conclude there are significant corelations between the independent variables. So we can proceed with PCA

2.21 Confirm the adequacy of sample size

Condition: Above 0.7 is acceptable, below 0.5 is not acceptable

```
The KMO value is  0.9361896166652944
```

2.22 Fit and Transform PCA Model

Fit and transform the scaled data using the PCA from SKlearn library

2.23 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
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
M_ILL	0.782405	0.933452	0.880243	0.967971	0.972547	0.822290	0.784357	0.023378	0.010249	0.869070	...	0.781156	0.
F_ILL	0.896107	0.917169	0.928913	0.896778	0.900544	0.842658	0.858401	0.112222	0.112487	0.877996	...	0.728973	0.
TOT_WORK_M	0.938328	0.977458	0.974326	0.898655	0.893232	0.868242	0.866029	0.057298	0.049061	0.982191	...	0.655936	0.
TOT_WORK_F	0.948620	0.825119	0.904224	0.732839	0.734787	0.733823	0.803562	0.250209	0.257052	0.842559	...	0.566210	0.
MAINWORK_M	0.926588	0.936031	0.943223	0.833607	0.825308	0.838925	0.842746	0.047749	0.040740	0.954067	...	0.531633	0.
MAINWORK_F	0.921397	0.772433	0.858357	0.650808	0.651110	0.690579	0.764551	0.217172	0.224043	0.805023	...	0.397956	0.
MAIN_CL_M	0.522335	0.629559	0.586212	0.649146	0.650964	0.645914	0.616419	0.073674	0.055829	0.585503	...	0.511180	0.
MAIN_CL_F	0.457357	0.413760	0.452244	0.430757	0.437133	0.398906	0.435648	0.245013	0.242003	0.396327	...	0.373515	0.
MAIN_AL_M	0.742109	0.684407	0.718934	0.646443	0.655998	0.666125	0.707189	0.138552	0.145525	0.651223	...	0.437050	0.
MAIN_AL_F	0.680048	0.489614	0.588526	0.415484	0.424655	0.484640	0.578741	0.261592	0.277316	0.495143	...	0.215343	0.
MAIN_HH_M	0.772796	0.881542	0.844553	0.833838	0.826709	0.842285	0.812071	-0.067456	-0.075597	0.873604	...	0.619146	0.
MAIN_HH_F	0.811980	0.776562	0.807333	0.689734	0.691654	0.727225	0.755018	0.041680	0.043034	0.782405	...	0.506555	0.
MAIN_OT_M	0.850983	0.844854	0.857130	0.720698	0.704311	0.737664	0.740948	0.016286	0.008891	0.890509	...	0.433279	0.

Table 2.7:Covariance Matrix

2.24 Extracting Eigen Vectors & Eigen Values

Eigen Vectors

```
array([[ 0.15,  0.16,  0.16, ...,  0.14,  0.15,  0.14],
       [-0.12, -0.08, -0.09, ...,  0.04, -0.05, -0.04],
       [ 0.1 , -0.04,  0.03, ..., -0.1 , -0.13, -0.03],
       ...,
       [ 0.  , -0.01,  0.02, ..., -0.01,  0.06, -0.01],
       [ 0.  ,  0.05,  0.  , ...,  0.01, -0.08, -0.  ],
       [-0.  , -0.  ,  0.01, ...,  0.  ,  0.01,  0.  ]])
```

Table 2.8: Eigen 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])
```

Table 2.9: Eigen Values

2.25 Extracting the Variability of the PC's

Check the explained variance for each PC

Explained variance = (eigen value of each PC)/(sum of eigen values of all PCs)

```
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])
```

Table 2.10: Variability of the PC's

2.26 Creating a Scree Plot

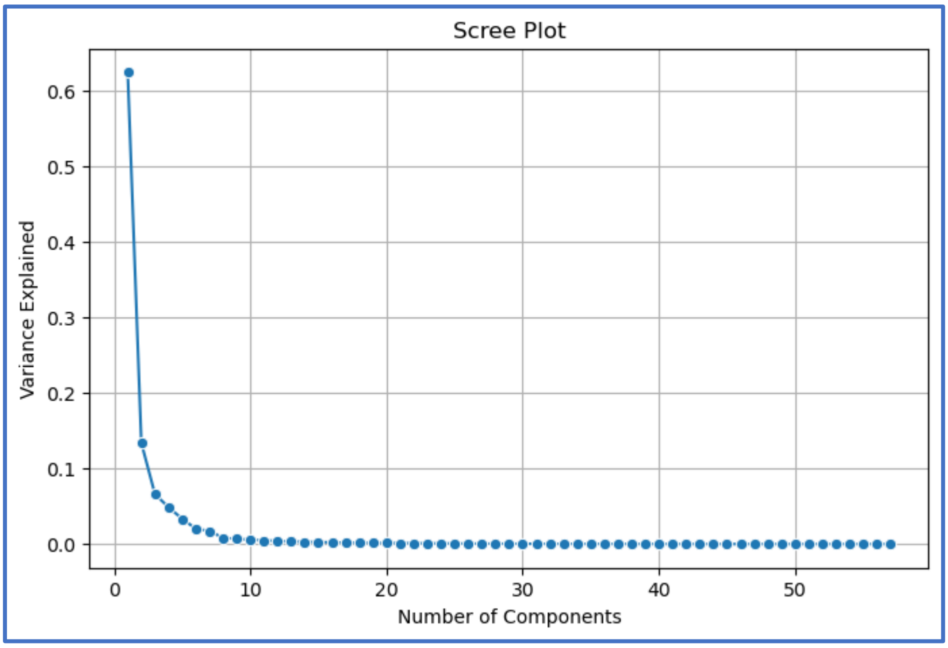


Figure 2.11: Scree Plot

2.27 Building PCA model with the 5 PC's

```
array([[ 0.14922158,  0.15916917,  0.15820921,  0.15634043,  0.1568144 ,
         0.14335015,  0.14353705,  0.01884873,  0.01787797,  0.15515239,
         0.14544984,  0.1545511 ,  0.15828347,  0.15407627,  0.14252995,
         0.14193201,  0.12573163,  0.11169244,  0.08303496,  0.11929067,
         0.09008881,  0.14184969,  0.13388011,  0.1227618 ,  0.1168656 ,
         0.15665637,  0.14869489,  0.08816344,  0.06516026,  0.1272781 ,
         0.11588826,  0.14536607,  0.14230182,  0.15087675,  0.14801846,
         0.15790761,  0.15583101,  0.15764021,  0.1495015 ,  0.0947852 ,
         0.06715842,  0.12818439,  0.11395923,  0.14510769,  0.14102942,
         0.15092232,  0.14753416,  0.14298675,  0.13378373,  0.06296394,
         0.05674058,  0.11910165,  0.11304417,  0.14213963,  0.14136961,
         0.14762899,  0.14210263],
       [-0.11548673, -0.08023879, -0.09371751, -0.02034061, -0.01431023,
        -0.07966701, -0.08709832,  0.06910144,  0.06731586, -0.10598636,
        -0.13323356, -0.00945956, -0.02179345, -0.12091195, -0.07600253,
        -0.16669997, -0.14224991,  0.04255228,  0.09589258, -0.05334228,
        -0.07246688, -0.10183528, -0.11325661, -0.2036023 , -0.20589888,
         0.07903864,  0.10881279,  0.2715224 ,  0.27539755,  0.15657864,
         0.13504767,  0.04097368,  0.00668481, -0.07344039, -0.08836101,
        -0.04404402, -0.09238317,  0.06620762,  0.08965133,  0.26126801,
         0.26669101,  0.14983097,  0.12064763,  0.03676265, -0.00368515,
        -0.0777393 , -0.10114106,  0.13683939,  0.16641612,  0.28188148,
         0.28754091,  0.18234077,  0.17711216,  0.05292484,  0.03510934,
        -0.04912234, -0.03984815],
       [ 0.1015276 , -0.03866173,  0.0289595 , -0.07441918, -0.06822314,
        -0.03761902,  0.02134973,  0.32382724,  0.33870545, -0.03210704,
        -0.00513336, -0.04705352,  0.07934454, -0.0011159 ,  0.19412998,
         0.01982148,  0.20997642,  0.03313125,  0.1888222 ,  0.22583087,
         0.35656643, -0.10220234,  0.02161302, -0.02814398,  0.06903375,
        -0.06868497,  0.10495656, -0.10474484, -0.03632536,  0.0704345 ,
         0.25998651, -0.14434657, -0.09383805, -0.13141498, -0.05388345,
        -0.06687743, -0.05871826, -0.06017243,  0.1257919 , -0.09655088,
        -0.01825633,  0.07819427,  0.28323496, -0.14251113, -0.08935617,
        -0.13068659, -0.05848926, -0.10356452,  0.03342285, -0.1202934 ,
        -0.08809749,  0.02617609,  0.16477413, -0.14441938, -0.10217491,
        -0.12667281, -0.02854464],
       [ 0.07681409,  0.05297633,  0.07002217,  0.02851986,  0.01639807,
         0.01021041,  0.01624416,  0.09114279,  0.07955449,  0.08918669,
         0.12541201, -0.03466478, -0.01057813,  0.06904579,  0.11105656,
         0.10018791,  0.13301329,  0.07885146,  0.2650219 , -0.12137878,
        -0.02098921, -0.02196919, -0.04543644,  0.14702469,  0.15591746,
        -0.07857186,  0.01578813,  0.15710396,  0.28502411, -0.25059413,
        -0.15379789, -0.16753968, -0.15146925,  0.02119534,  0.05996115,
         0.03931895,  0.04613025, -0.09131505,  0.01886534,  0.13159069,
         0.29284517, -0.2503371 , -0.14304544, -0.16600189, -0.14259884,
         0.01988712,  0.0600874 , -0.01822291,  0.0059541 ,  0.20894141,
         0.2404994 , -0.24041564, -0.18940781, -0.16755357, -0.16901995,
         0.02403566,  0.05740164],
       [-0.01209003, -0.04234376, -0.02292653, -0.08033939, -0.07832648,
        -0.16789316, -0.15809156,  0.41841183,  0.4159652 , -0.01403251,
         0.02908422, -0.10407302, -0.11033167, -0.02310352, -0.01893052,
        -0.04322541, -0.054674  , -0.30337639, -0.25792534, -0.25313081,
        -0.19921997, -0.06081182, -0.0230627 ,  0.06990677,  0.10677437,
         0.06581161,  0.07762414, -0.01800453, -0.05515214, -0.04720013,
        -0.01264328,  0.00557458,  0.04361632,  0.1451087 ,  0.19075649,
        -0.0598864 , -0.02247554,  0.05907845,  0.06434924, -0.01388688,
        -0.06101878, -0.05866475, -0.02538622,  0.00331493,  0.04167758,
         0.13279387,  0.17059608,  0.0942929 ,  0.11235112, -0.01807012,
        -0.03629271,  0.01698094,  0.04753801,  0.01418678,  0.04750424,
         0.19178951,  0.24976544]])
```

2.28 Extracting factor loadings

	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_
0	0.149222	0.159169	0.158209	0.156340	0.156814	0.143350	0.143537	0.018849	0.017878	0.155152	...	0.142987	0.133784	
1	-0.115487	-0.080239	-0.093718	-0.020341	-0.014310	-0.079667	-0.087098	0.069101	0.067316	-0.105986	...	0.136839	0.166416	
2	0.101528	-0.038662	0.028959	-0.074419	-0.068223	-0.037619	0.021350	0.323827	0.338705	-0.032107	...	-0.103565	0.033423	
3	0.076814	0.052976	0.070022	0.028520	0.016398	0.010210	0.016244	0.091143	0.079554	0.089187	...	-0.018223	0.005954	
4	-0.012090	-0.042344	-0.022927	-0.080339	-0.078326	-0.167893	-0.158092	0.418412	0.415965	-0.014033	...	0.094293	0.112351	

5 rows × 57 columns

Table 2.11: Extracted Factor Loadings

2.29 Identifying the actual columns wirh most variance

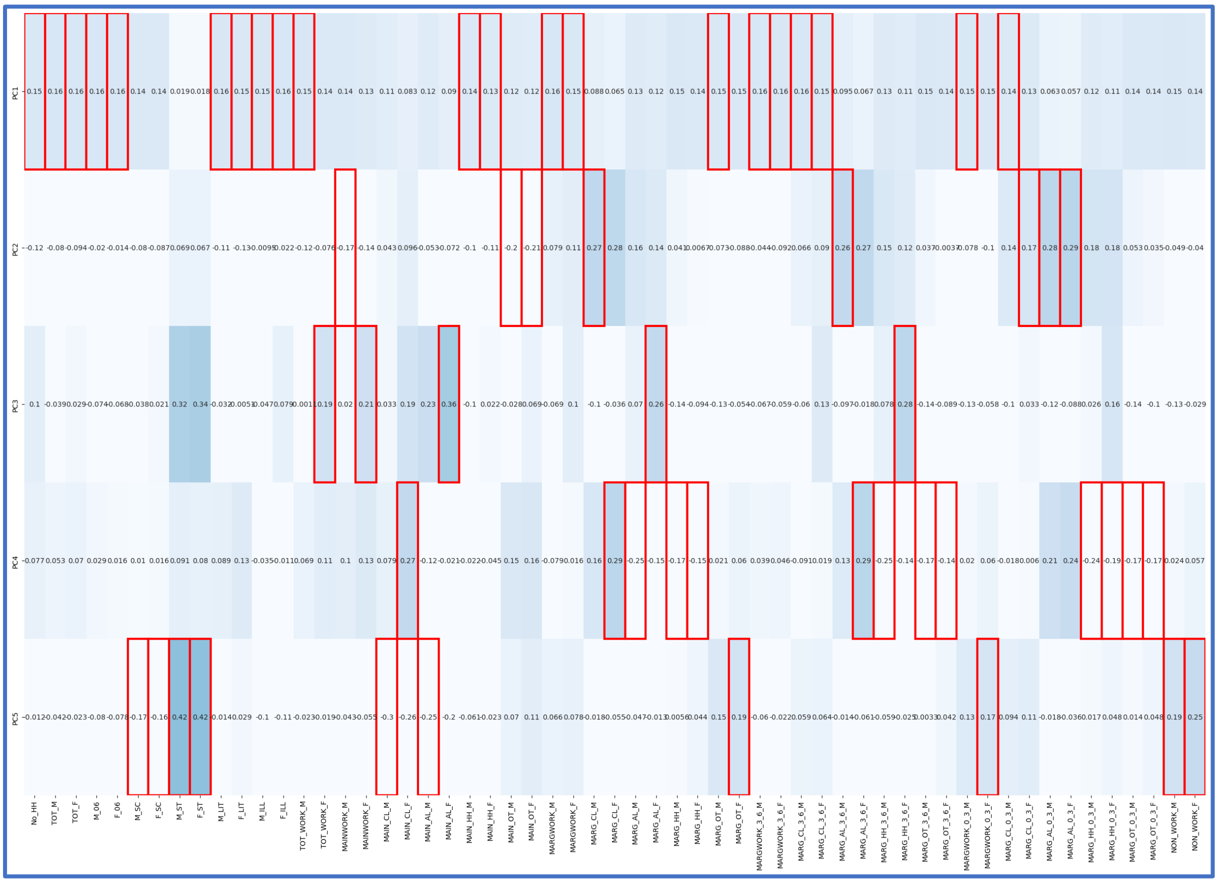


Figure 2.12: Most Variance in PC's

2.30 Checking the presence of correlations among the PCs

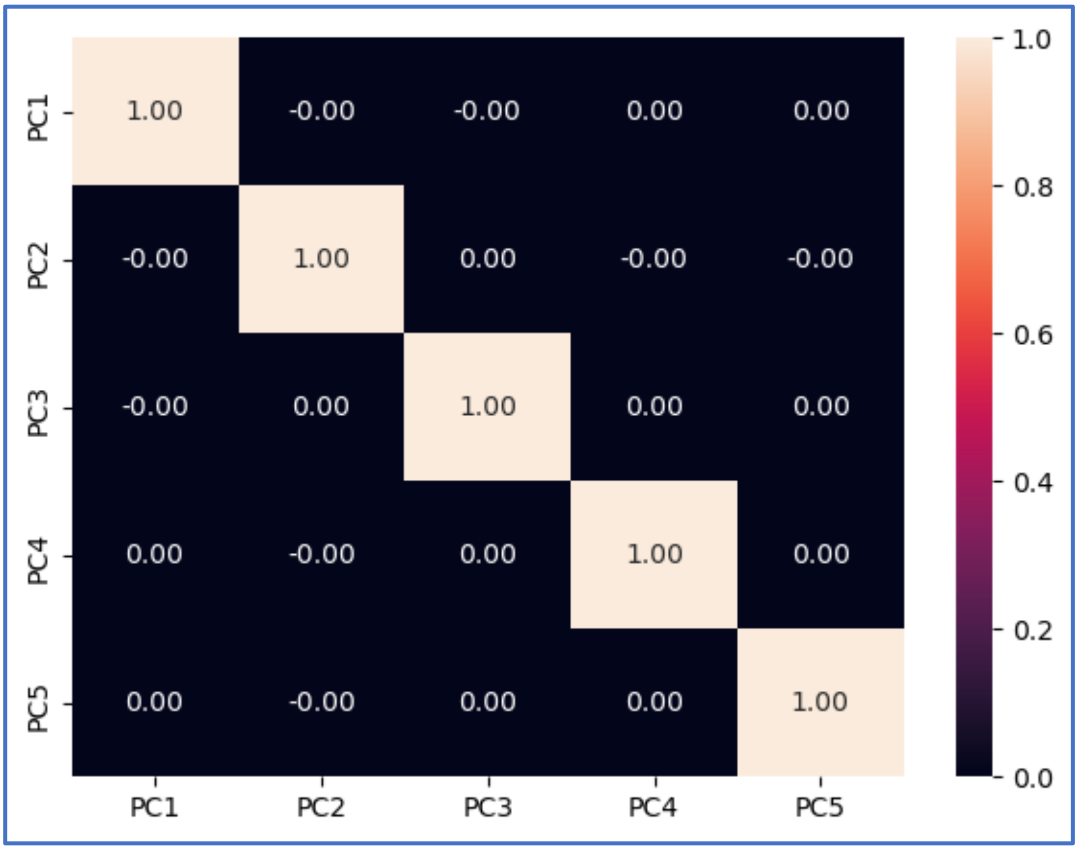


Figure 2.13: Correlation among PC's

2.31 Identifying the actual variables

Name	Description	PC	Name
TOT_M	Total population Male	PC1	Total Population Male & Female
TOT_F	Total population Female	PC1	
M_ST	Scheduled Tribes population Male	PC5	Total Population Male & Female ST
F_ST	Scheduled Tribes population Female	PC5	
MAIN_AL_F	Main Agricultural Labourers Population Female	PC3	Main Agricultural Labourers Population Female
MARG_CL_F	Marginal Cultivator Population Female	PC4	Female Population - Cultivator & Agricultural (3-6)
MARG_AL_3_6_F	Marginal Agriculture Labourers Population 3-6 Female	PC4	
MARG_AL_0_3_M	Marginal Agriculture Labourers Population 0-3 Male	PC2	Total Population Male & Female of Marginla Agricultural Labourers
MARG_AL_0_3_F	Marginal Agriculture Labourers Population 0-3 Female	PC2	

Table 2.12: Identifying the actual variables

	pc_total_male_female	pc_total_male-female_marginal_agricultural_labourers	pc_main_agricultural_labourers_female_population	pc_female_population_cultivator_agricultural_3-6
0	-5.528161	0.430378	-1.473827	-1.278049
1	-5.492016	-0.106110	-2.015641	-1.750168
2	-7.474643	-0.217194	-0.247428	0.006079
3	-7.919737	-0.652311	-0.659220	-0.735550
4	-5.175695	2.304059	-1.157327	1.060796

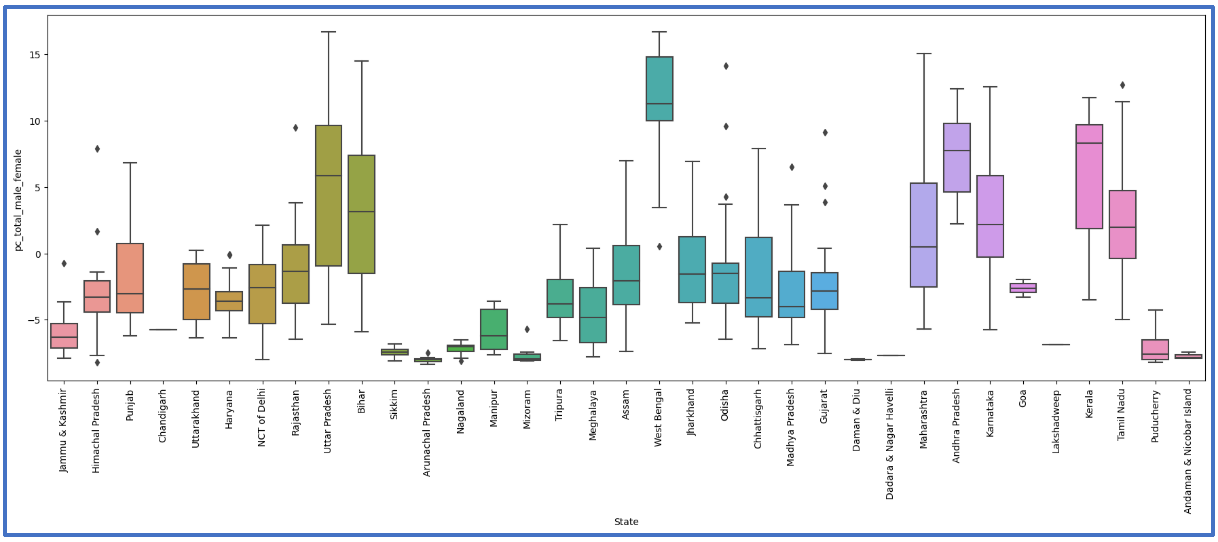
Table 2.13: Adding the Identified variables with the data

2.32 EDA (Categorical Fields & Principal Components)

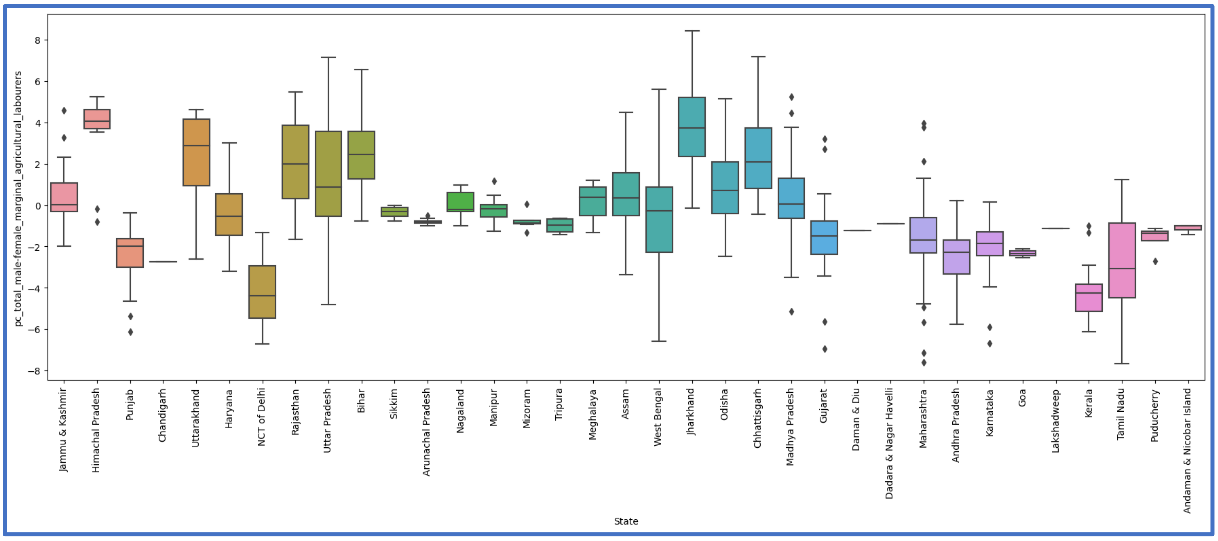
	State	Area Name	pc_total_male_female	pc_total_male-female_marginal_agricultural_labourers	pc_main_agricultural_labourers_female_population	pc_female_population_cu
0	Jammu & Kashmir	Kupwara	-5.528161	0.430378	-1.473827	
1	Jammu & Kashmir	Badgam	-5.492016	-0.106110	-2.015641	
2	Jammu & Kashmir	Leh(Ladakh)	-7.474643	-0.217194	-0.247428	
3	Jammu & Kashmir	Kargil	-7.919737	-0.652311	-0.659220	
4	Jammu & Kashmir	Punch	-5.175695	2.304059	-1.157327	

Table 2.14: Adding PC’s to the Categorical columns

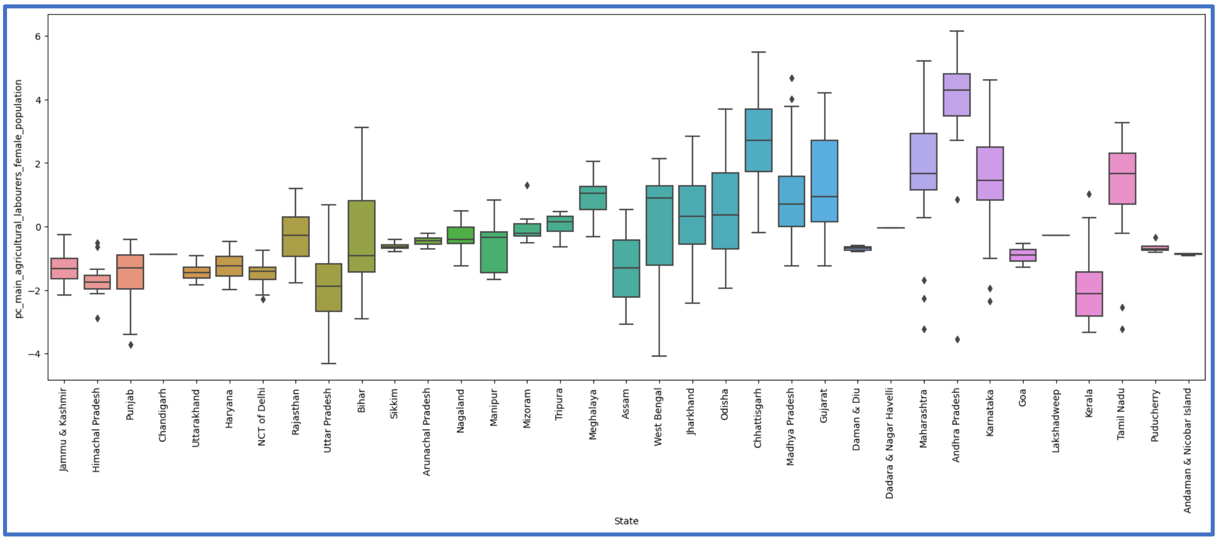
2.33 Inferences



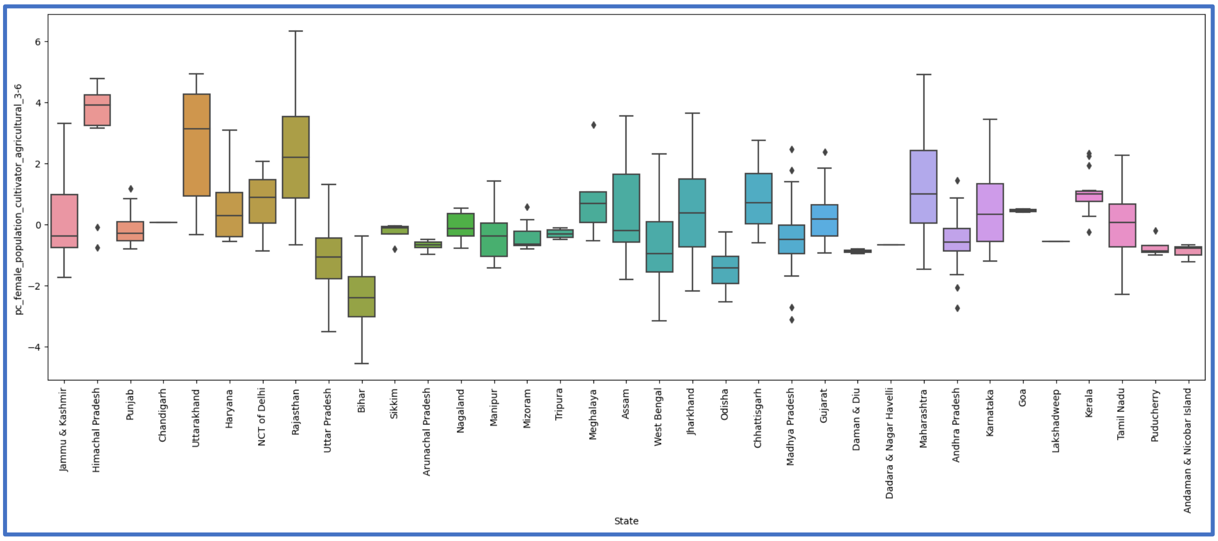
2.14 State Wise : pc_total_male_female



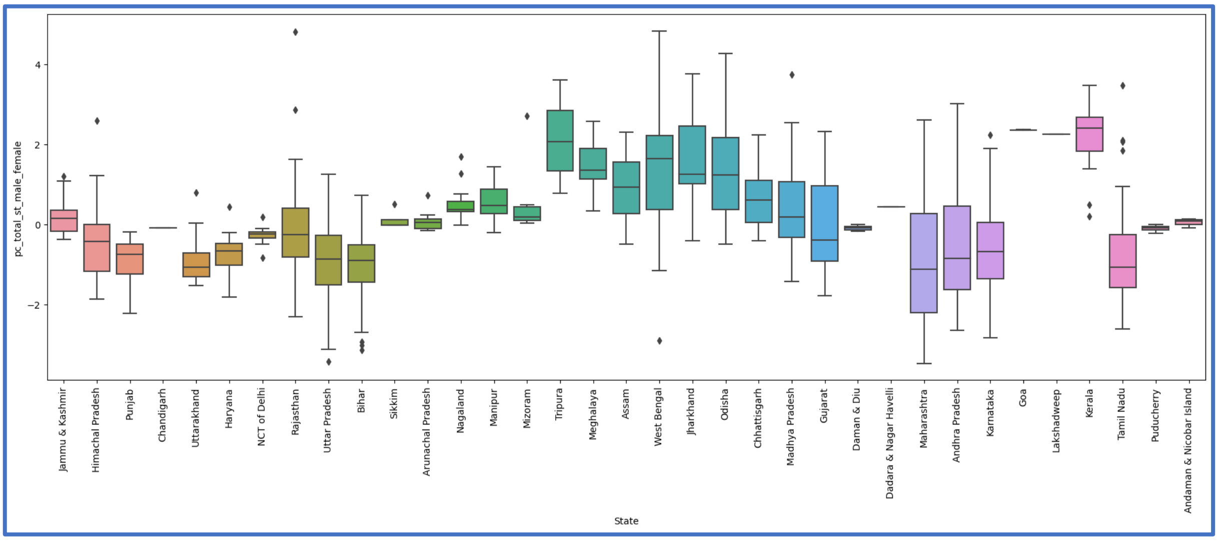
2.15 State Wise : pc_total_male-female_marginal_agricultural_labourers



2.16 State Wise : pc_main_agricultural_labourers_female_population



2.17 State Wise : pc_female_population_cultivator_agricultural_3-6



2.18 State Wise : pc_total_st_male_female

- West Bengal has the highest average male & female total
- Daman & Diu has the lowest average male & female total
- Himachal Pradesh has the highest average of total male & female population of marginal agricultural labourers
- Kerala & NCT of Delhi has the lowest average of total male & female population of marginal agricultural labourers
- Andra Pradesh has the highest average female population of agricultural labourers
- Kerala has the lowest average female population of agricultural labourers
- Himachel Pradesh has the highest average female population of Cultivator & agricultural labourers of age 3-6
- Bihar has the lowest average female population of Cultivator & agricultural labourers of age 3-6
- Kerala has the highest average Total male & female Scheduled Tribe population
- Uttarakhand has the lowest average Total male & female Scheduled Tribe population

2.34 Linear Equation

```
0.149 * No_HH + 0.159 * TOT_M + 0.158 * TOT_F + 0.156 * M_06 + 0.157 *
F_06 + 0.143 * M_SC + 0.144 * F_SC + 0.019 * M_ST + 0.018 * F_ST + 0.15
5 * M_LIT + 0.145 * F_LIT + 0.155 * M_ILL + 0.158 * F_ILL + 0.154 * TOT
_WORK_M + 0.143 * TOT_WORK_F + 0.142 * MAINWORK_M + 0.126 * MAINWORK_F
+ 0.112 * MAIN_CL_M + 0.083 * MAIN_CL_F + 0.119 * MAIN_AL_M + 0.09 * MA
IN_AL_F + 0.142 * MAIN_HH_M + 0.134 * MAIN_HH_F + 0.123 * MAIN_OT_M + 0
.117 * MAIN_OT_F + 0.157 * MARGWORK_M + 0.149 * MARGWORK_F + 0.088 * MA
RG_CL_M + 0.065 * MARG_CL_F + 0.127 * MARG_AL_M + 0.116 * MARG_AL_F + 0
.145 * MARG_HH_M + 0.142 * MARG_HH_F + 0.151 * MARG_OT_M + 0.148 * MARG
_OT_F + 0.158 * MARGWORK_3_6_M + 0.156 * MARGWORK_3_6_F + 0.158 * MARG_
CL_3_6_M + 0.15 * MARG_CL_3_6_F + 0.095 * MARG_AL_3_6_M + 0.067 * MARG_
AL_3_6_F + 0.128 * MARG_HH_3_6_M + 0.114 * MARG_HH_3_6_F + 0.145 * MARG
_OT_3_6_M + 0.141 * MARG_OT_3_6_F + 0.151 * MARGWORK_0_3_M + 0.148 * MA
RGWORK_0_3_F + 0.143 * MARG_CL_0_3_M + 0.134 * MARG_CL_0_3_F + 0.063 *
MARG_AL_0_3_M + 0.057 * MARG_AL_0_3_F + 0.119 * MARG_HH_0_3_M + 0.113 *
MARG_HH_0_3_F + 0.142 * MARG_OT_0_3_M + 0.141 * MARG_OT_0_3_F + 0.148 *
NON_WORK_M + 0.142 * NON_WORK_F +
```