DATA MINING

Project Report

Submitted by,

Sindhu R Udupa.

BATCH: PGPDSBA.O. NOV22.B

Contents

	Part – 2 **Principal Component Analysis**	
2.1	Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.	17
2.2	Perform detailed Exploratory analysis by creating certain questions like (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio? (Example Questions). Pick 5 variables out of the given 24 variables below for EDA: No_HH, TOT_M, TOT_F, M_06, F_06, M_SC, F_SC, M_ST, F_ST, M_LIT, F_LIT, M_ILL, F_ILL, TOT_WORK_M, TOT_WORK_F, MAINWORK_M, MAINWORK_F, MAIN_CL_M, MAIN_CL_F, MAIN_AL_M, MAIN_AL_F, MAIN_HH_M, MAIN_HH_F, MAIN_OT_M, MAIN_OT_F	20
2.3	We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?	24
2.4	Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.	25
2.5	Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.	27
2.6	Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.	30
2.7	Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the principal components in terms of actual variables.	31
2.8	Write linear equation for first PC.	35

List of Figures

List of Tables

Part -1

Clustering

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

The given dataset has been read. We see that

a. it contains 23066 records with 19 attributes.

b. We have 6 columns of object type, 6 columns of float type and 7 columns of integer type.

c. There are no duplicates in the data.

d. There are 4736 missing values in the columns CTR, CRM and CPC each.

	Timestamp	Inventory Type	Ad - Length	Ad-Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0	0.35	0.0	0.0031	0.0	0.0
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.0	0.35	0.0	0.0035	0.0	0.0
2	2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0	0.35	0.0	0.0028	0.0	0.0
3	2020-9-3-20	Format1	300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.0	0.35	0.0	0.0020	0.0	0.0
4	2020-9-4-15	Format1	300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.0	0.35	0.0	0.0041	0.0	0.0

Table 1: First five records of the dataset.

	Timestamp	Inventory Type	Ad - Length	Ad-Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
23061	2020-9-13-7	Format5	720	300	216000	Inter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
23062	2020-11-2-7	Format5	720	300	216000	Inter224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.0260	NaN	NaN	NaN
23063	2020-9-14-22	Format5	720	300	216000	Inter218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	NaN	NaN	NaN
23064	2020-11-18-2	Format4	120	600	72000	inter230	Video	Mobile	Video	7	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
23065	2020-9-14-0	Format5	720	300	216000	Inter221	App	Mobile	Video	2	2	2	1	0.09	0.35	0.0585	NaN	NaN	NaN

Table 2: Last five records of the dataset.

```
<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
```

Table 3: Information of the dataset.

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	3.851631e+02	2.336514e+02	120.0000	120.000000	300.00000	7.200000e+02	728.00
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.0000	250.000000	300.00000	6.000000e+02	600.00
Ad Size	23066.0	9.667447e+04	6.153833e+04	33600.0000	72000.000000	72000.00000	8.400000e+04	216000.00
Available_Impressions	23066.0	2.432044e+06	4.742888e+06	1.0000	33672.250000	483771.00000	2.527712e+06	27592861.00
Matched_Queries	23066.0	1.295099e+06	2.512970e+06	1.0000	18282.500000	258087.50000	1.180700e+06	14702025.00
Impressions	23066.0	1.241520e+06	2.429400e+06	1.0000	7990.500000	225290.00000	1.112428e+06	14194774.00
Clicks	23066.0	1.067852e+04	1.735341e+04	1.0000	710.000000	4425.00000	1.279375e+04	143049.00
Spend	23066.0	2.706626e+03	4.067927e+03	0.0000	85.180000	1425.12500	3.121400e+03	26931.87
Fee	23066.0	3.351231e-01	3.196322e-02	0.2100	0.330000	0.35000	3.500000e-01	0.35
Revenue	23066.0	1.924252e+03	3.105238e+03	0.0000	55.365375	926.33500	2.091338e+03	21276.18
CTR	18330.0	7.366054e-02	7.515992e-02	0.0001	0.002600	0.08255	1.300000e-01	1.00
CPM	18330.0	7.672045e+00	6.481391e+00	0.0000	1.710000	7.66000	1.251000e+01	81.56
CPC	18330.0	3.510606e-01	3.433338e-01	0.0000	0.090000	0.16000	5.700000e-01	7.26

Table 4: Statistical description of the data.

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

The formula for the CPC, CTR and CPM are as follows.

a. **CPC = Total Cost (spend) / Number of Clicks.**
b. **CTR = (Total Measured Clicks / Total Measured Ad Impressions) x 100.**
c. **CPM = (Total Campaign Spend / Number of Impressions) x 1000.**

We have used the above formulae for missing value imputation.

```
The total number of null values in each column is as follows:

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
KMeans_Cluster          0
dtype: int64
```

Table 5: Null values in the data after treatment.

We see from the above table that there are no missing values in the dataset now.

1.3. Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ.

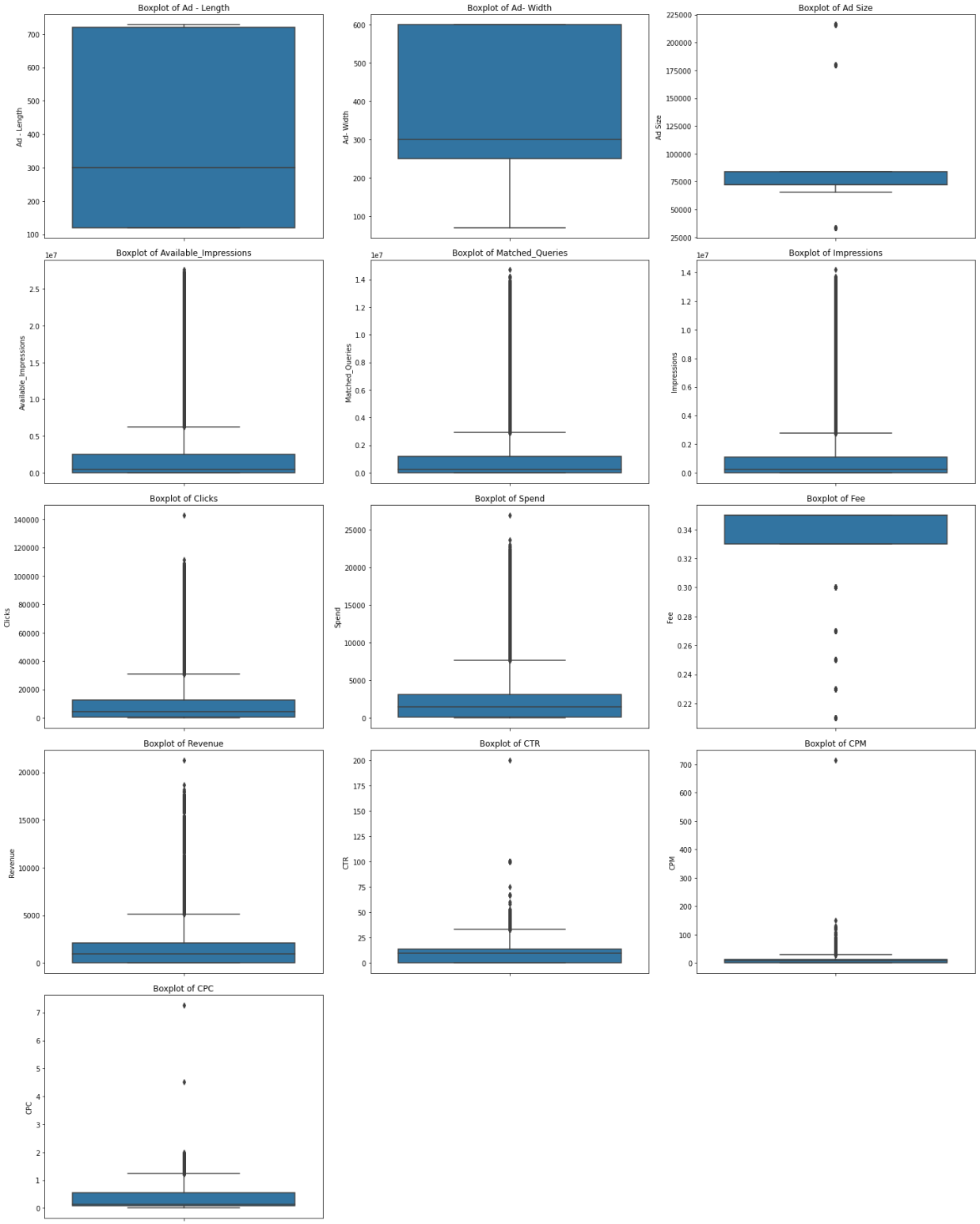


Fig. 1: Checking for outliers in all numerical columns.

From the boxplots above, we can see there are outliers in all the columns except the ‘Ad length’ and ‘Ad width’ columns.

As the name suggests, K-Means clustering technique makes use of the means to form clusters, and the means are highly impacted by the outliers in the particular column. Hence, it is necessary to treat the outliers.

We are using the box plot method to treat the outliers. Basically, in this method, we are going to impute the values that are greater than Q_3 + 1.5 times the IQR and bring them down to the exact mentioned value. Similarly, any values lesser than Q_1 - 1.5 x IQR is also treated. We may note that Q_1 is the first quartile, Q_3 is the third quartile and IQR = $Q_3 – Q_1$.

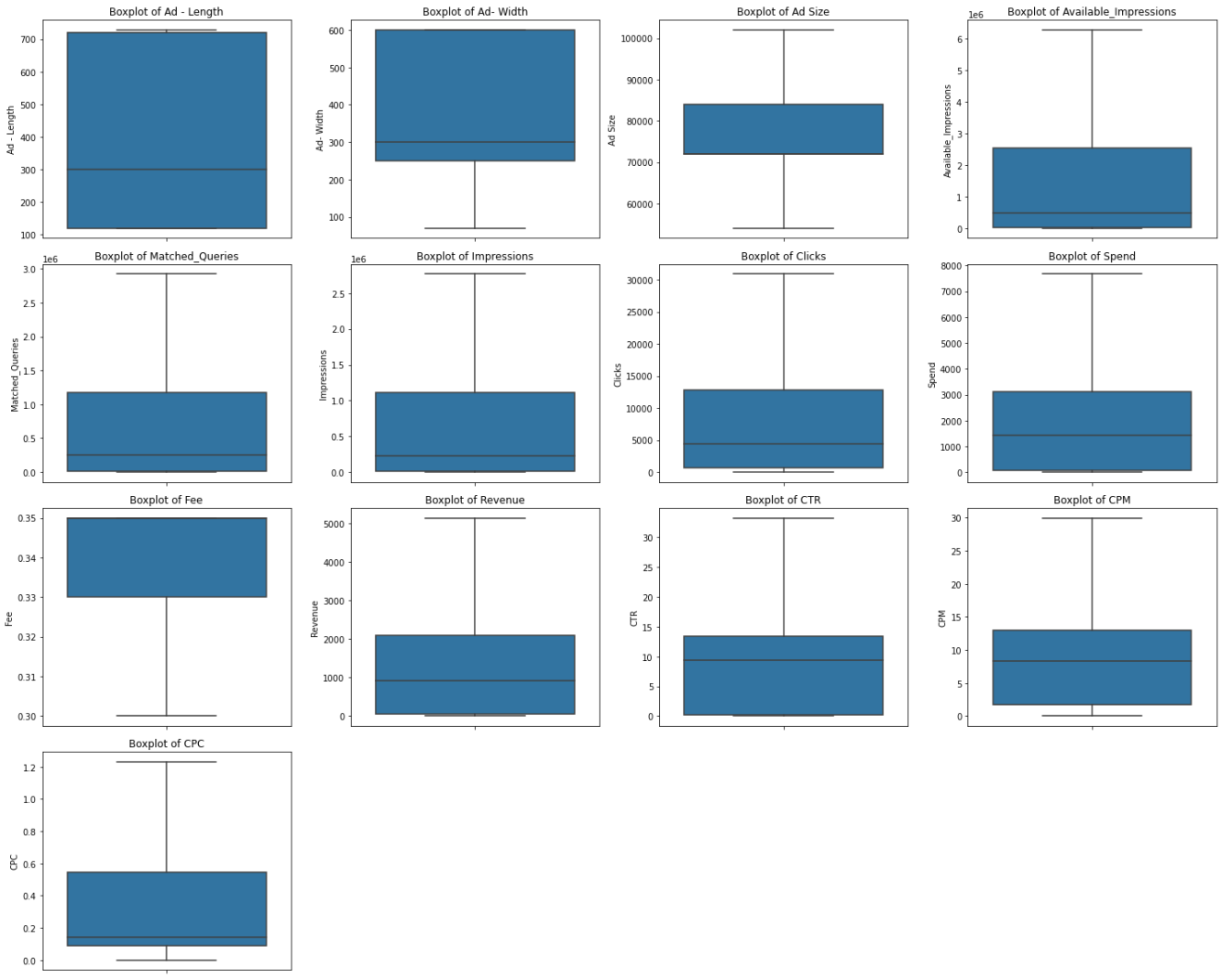


Fig. 2: Boxplots of all numerical columns after outlier treatment.

From the graphs above, we confirm that all the outliers in all the numerical columns have been treated.

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

We have applied z-scores to all the numerical columns of the original data. Z-score scaling helps us bring the mean of all columns close to 0 and standard deviation close to 1.

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	-0.364496	-0.432797	-0.102518	-0.755333	-0.778949	-0.768478	-0.867488	-0.893170	0.535724	-0.880093	-0.958836	-1.194498	-1.042561
1	-0.364496	-0.432797	-0.102518	-0.755345	-0.778988	-0.768516	-0.867488	-0.893170	0.535724	-0.880093	-0.953835	-1.194498	-1.042561
2	-0.364496	-0.432797	-0.102518	-0.754900	-0.778919	-0.768445	-0.867488	-0.893170	0.535724	-0.880093	-0.962218	-1.194498	-1.042561
3	-0.364496	-0.432797	-0.102518	-0.755040	-0.778781	-0.768302	-0.867488	-0.893170	0.535724	-0.880093	-0.971871	-1.194498	-1.042561
4	-0.364496	-0.432797	-0.102518	-0.755610	-0.779030	-0.768560	-0.867488	-0.893170	0.535724	-0.880093	-0.946281	-1.194498	-1.042561
...	...	...	...	...	...	...	...	...	...	...	...	...	...
23061	1.433093	-0.186599	1.652896	-0.756182	-0.779265	-0.768806	-0.867488	-0.893141	0.535724	-0.880066	3.035808	3.162718	-0.821435
23062	1.433093	-0.186599	1.652896	-0.756181	-0.779264	-0.768805	-0.867488	-0.893154	0.535724	-0.880078	3.035808	1.712113	-0.916204
23063	1.433093	-0.186599	1.652896	-0.756182	-0.779265	-0.768806	-0.867488	-0.893150	0.535724	-0.880074	3.035808	3.162718	-0.884614
23064	-1.134891	1.290590	-0.297564	-0.756179	-0.779265	-0.768806	-0.867488	-0.893141	0.535724	-0.880066	3.035808	3.162718	-0.821435
23065	1.433093	-0.186599	1.652896	-0.756182	-0.779264	-0.768805	-0.867488	-0.893133	0.535724	-0.880058	3.035808	3.162718	-0.758256

Table 6: Scaled data.

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	-4.030447e-15	1.000022	-1.134891	-1.134891	-0.364496	1.433093	1.467332
Ad- Width	23066.0	5.390161e-15	1.000022	-1.319110	-0.432797	-0.186599	1.290590	1.290590
Ad Size	23066.0	-4.156304e-15	1.000022	-1.467840	-0.297564	-0.297564	0.482620	1.652896
Available_Impressions	23066.0	-3.617510e-15	1.000022	-0.756182	-0.740341	-0.528577	0.433059	2.193158
Matched_Queries	23066.0	1.341008e-15	1.000022	-0.779265	-0.761447	-0.527722	0.371498	2.070914
Impressions	23066.0	-1.224345e-15	1.000022	-0.768806	-0.760655	-0.538975	0.366051	2.056111
Clicks	23066.0	1.960656e-15	1.000022	-0.867488	-0.793438	-0.405431	0.468629	2.361729
Spend	23066.0	1.250852e-15	1.000022	-0.893170	-0.858046	-0.305523	0.393932	2.271900
Fee	23066.0	-2.322121e-14	1.000022	-2.222416	-0.567532	0.535724	0.535724	0.535724
Revenue	23066.0	3.136228e-15	1.000022	-0.880093	-0.846474	-0.317607	0.389803	2.244218
CTR	23066.0	1.329072e-15	1.000022	-0.995031	-0.964227	0.141524	0.635787	3.035808
CPM	23066.0	5.791296e-17	1.000022	-1.194498	-0.940303	0.022146	0.700905	3.162718
CPC	23066.0	1.987283e-15	1.000022	-1.042561	-0.759091	-0.602371	0.682987	2.846105

Table 7: Statistical description of the scaled data.

From the above table, the mean is very close to 0 and the standard deviation is almost equal to 1. Scaling the data reduces the computational time and also the memory usage. For example, let us consider a column whose values are in the range of millions. If we scale the mentioned column, every value will be reduced to the units. Carrying out calculations in terms of units will save much memory and also time compared to carrying out calculations in terms of millions.

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

We are constructing a dendrogram by performing hierarchical clustering. Here, we are measuring the distance by Euclidean distance and the linkage by Ward's method.

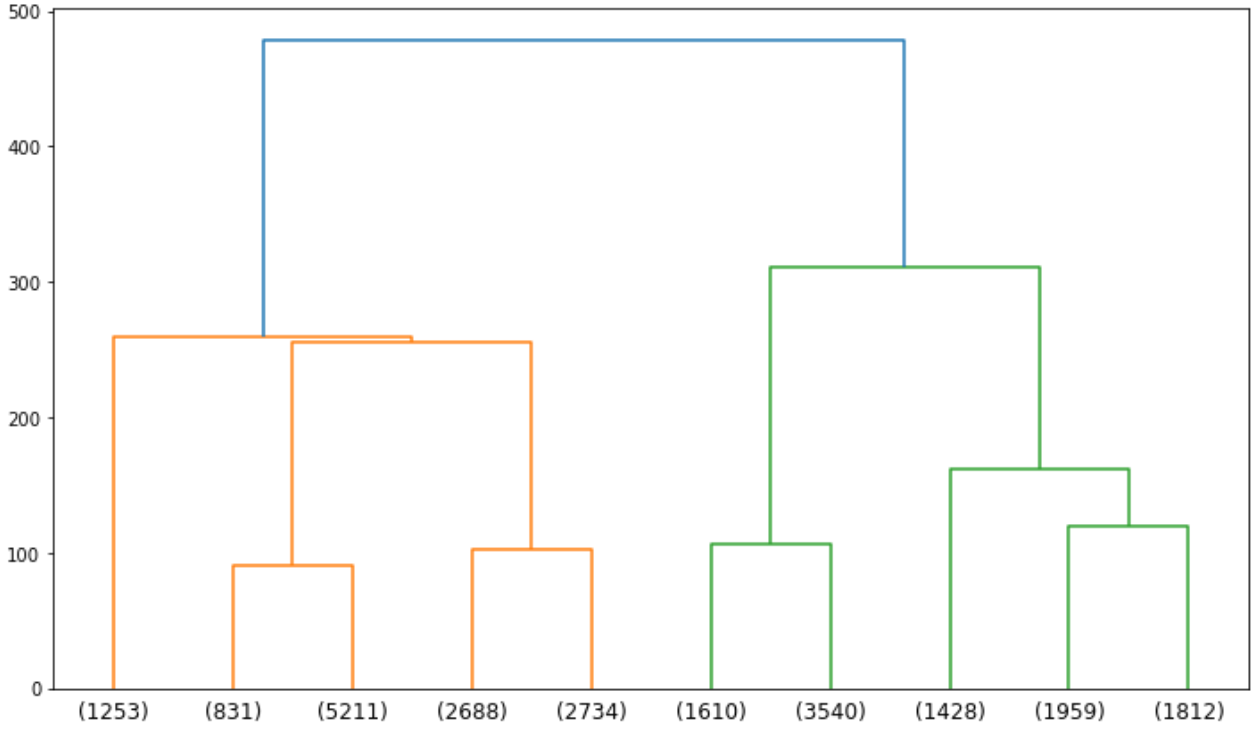


Fig. 3: Dendrogram

In the beginning of the hierarchical clustering, each observation would be considered as an individual cluster. And in each step, the closest observations are progressively merged. In the dendrogram above, the last 10 steps of the hierarchical clustering is shown, where we can see that in the last step all clusters are merged into one single cluster.

The above dendrogram gives us 2 clusters. The first cluster (colored in orange) has 12717 observations and the other cluster (colored in green) has 10349.

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

```
[299858.0000000003,
 183349.10202886089,
 130878.34788742856,
 95573.82185892039,
 61539.18919785387,
 51676.892307099595,
 44598.25849746795,
 39597.84594043502,
 36061.70346592873,
 33017.712783916766]
```

Table 8: Within sum of square values

The above table gives us the within sums of square values for each k, starting from 1 to 10. We are now plotting these values against the values of k in the elbow plot.

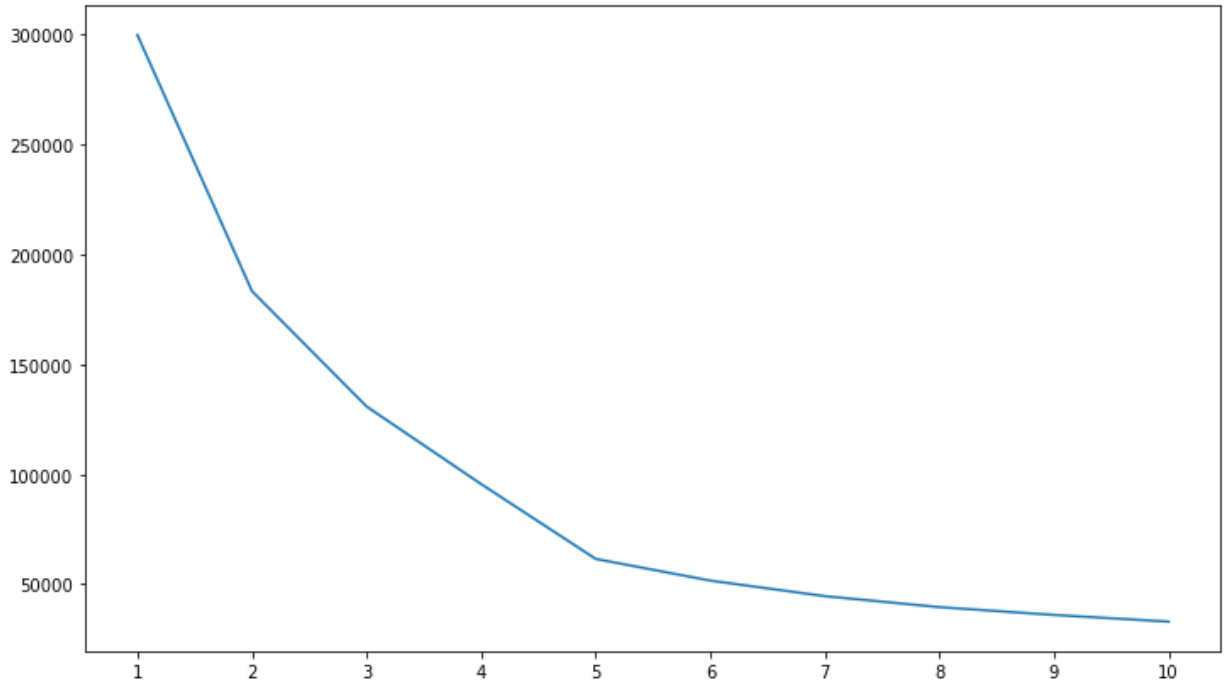


Fig. 4: Elbow plot.

In the above plot, the x-axis is the number of clusters we are considering for the k-means algorithm. The y-axis is 'with in sum of square' values. The number of clusters after which there is no significant drop in the elbow plot/WSS plot is the optimal number of clusters for the k-means. In this case, **the optimal value of k would be 5**, since we see no significant drop in the plot after k=5.

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

Silhouette score helps us analyze how tightly the observations are clustered. The maximum value of the statistic indicates the optimum value of k. The silhouette scores for up to 10 clusters are given below.

```
Silhouette score for 2 clusters = 0.38572769619101077
Silhouette score for 3 clusters = 0.3825476915535516
Silhouette score for 4 clusters = 0.44534519247649795
Silhouette score for 5 clusters = 0.5240956940501831
Silhouette score for 6 clusters = 0.5221533662938636
Silhouette score for 7 clusters = 0.5165635029478517
Silhouette score for 8 clusters = 0.47972249893837277
Silhouette score for 9 clusters = 0.43190129800502564
Silhouette score for 10 clusters = 0.43629905902082944
```

Table 9: Silhouette scores.

From the above table we see that maximum value of the silhouette score is for 5 clusters. This indicates the optimum number of clusters is 5. We may note that we had considered the same value, k=5, by looking at the elbow plot as well.

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

We have grouped the data based on the clusters and the device type to find out how the ads are performing. The graph in the next page gives a clarity on the same.

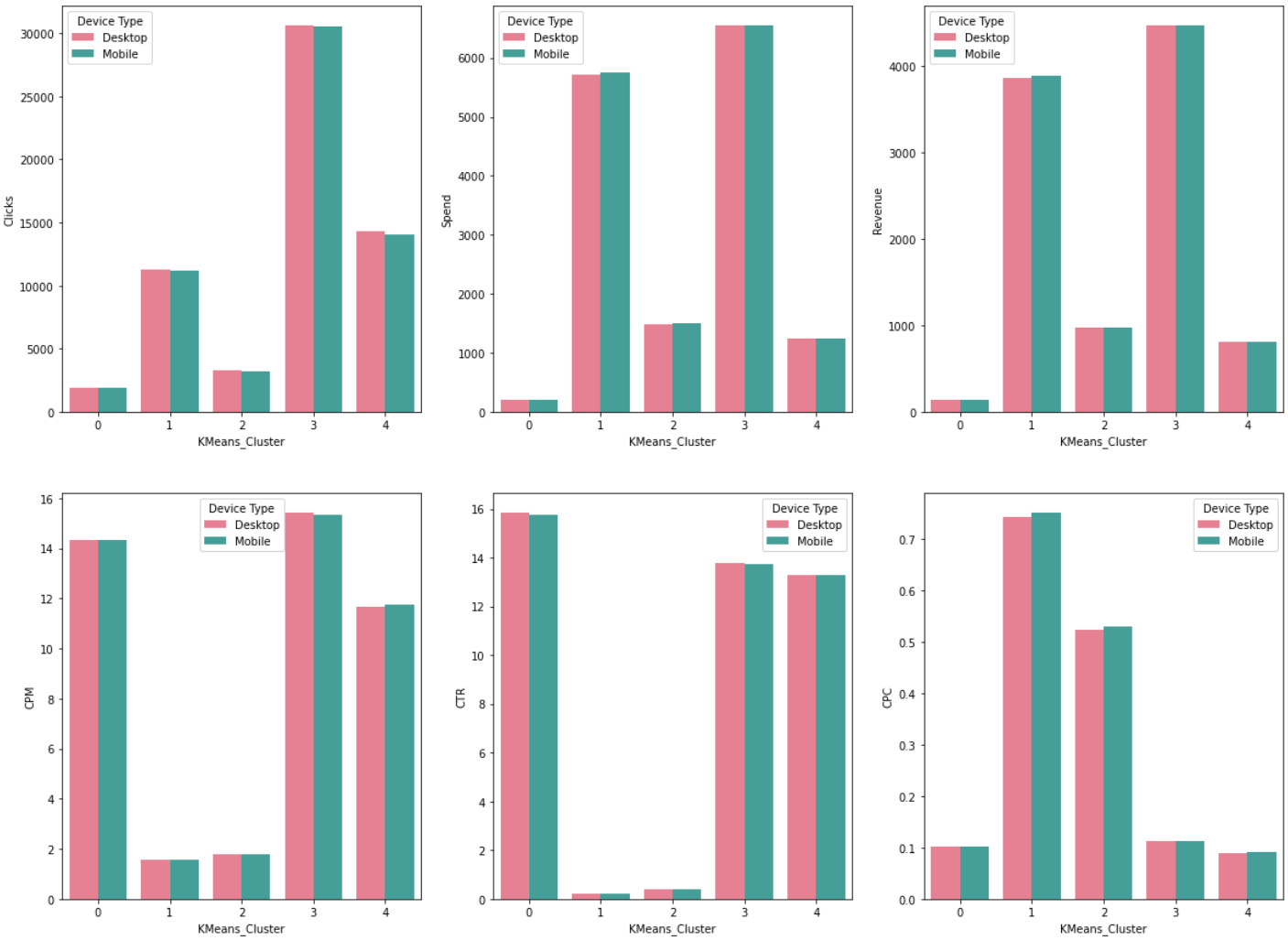


Fig. 5: Cluster Profiling based on device type.

From the figure above, we see the device type may not be playing an important role in determining the clicks, spend, revenue, etc. since we see the graph is same for both desktop and mobile.

- Click through rate is better for clusters 0,3 and 4, which tells us that these ads are reaching the customers.
- Cost per Click is high for cluster 1, which tells us that the company is spending a lot of money on these ads. And we also see good revenue from them.
- CPM is more for cluster 3. But CTR for cluster 3 is quite good and hence those ads are generating the highest revenue.
- Cluster 0 is generating the least revenue. Even though the spend on these ads are less, the cost per mille is high.
- Revenue generated by cluster 2 and 4 are moderate. But we see CPM for cluster 4 is high and CPC for cluster 2 is also high. CTR for cluster 4 is more. Cluster 4 ads are reaching the customers better than cluster 2 ads.

1.9 Conclude the project by providing summary of your learnings.

Upon cluster profiling, we see there are 6524 observations in the first cluster, 4054 in the second, 6275 in the third, 1537 in the fourth and 4676 in the fifth cluster. We shall group the observation by these clusters and take a look at the mean values of all other attributes.

Cluster	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	143.280809	572.103004	73966.738197	3.209356e+04	1.962406e+04	1.349204e+04	1914.448804	209.162609	0.349988	135.993379	15.784443	14.330063	0.102764
1	465.781944	199.148989	72963.936852	5.695405e+06	2.806219e+06	2.671268e+06	11245.754810	5739.327617	0.313281	3878.748366	0.217242	1.573280	0.748699
2	421.696255	152.001594	64299.996813	1.810314e+06	8.642623e+05	8.262209e+05	3263.131952	1500.090563	0.349264	977.424163	0.404392	1.788731	0.528129
3	141.454782	572.446324	73686.402082	8.063284e+05	5.668641e+05	4.781485e+05	30572.439330	6546.373195	0.305569	4471.776116	13.752664	15.385753	0.111918
4	683.825492	303.785287	100775.876818	2.513465e+05	1.375509e+05	1.167714e+05	14127.278203	1252.285569	0.349538	815.541831	13.289690	11.728833	0.090012

Table 10: Cluster profiling.

From the table above we observe the following.

Cluster 0: This cluster comprises of moderate sized, less shown(impressions) and less clicked ads. The spending on these ads and the revenue generated by these ads are very less. The click through rate is quite better compared to other clusters. The average cost per mille (CPM) is around 14.33, which seems to be on the higher end. The average Cost per Click is 0.1, which is on the lower end.

Cluster 1: This cluster has moderate sized, most frequently shown(impressions) and the moderately clicked ads. The spending on these ads and the revenue generated from these ads are high. The click through rate is very less. The average cost per mille (CPM) is around 1.57, which is very less as well. The average Cost per Click is 0.7, which is on the higher end.

Cluster 2: This cluster has relatively small sized, most frequently shown(impressions) and less clicked ads. The spending on these ads and the revenue generated from these ads are moderate. The click through rate is very less. The average cost per mille (CPM) is around 1.78, which is again very less. The average Cost per Click is 0.52, which is moderate.

Cluster 3: This cluster has moderate sized, most frequently shown(impressions) and the most clicked ads. The spending on these ads and the revenue generated from these ads are high. The click through rate is high. The average cost per mille (CPM) is around 15.38, which is again very high. The average Cost per Click is 0.11, which is less.

Cluster 4: This cluster has large sized, moderately shown(impressions) and the moderately clicked ads. The spending on these ads and the revenue generated from these ads are moderate. The click through rate is high. The average cost per mille (CPM) is around 11.72, which is again high. The average Cost per Click is 0.09, which is less.

Part -2

Principal Component Analysis

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

The given dataset has been read. We see that

a. it contains 640 observations with 61 attributes.

b. We have 2 columns of object type, 59 columns of integer type.

c. There are no duplicates in the data.

d. There are no missing values in the data.

	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	F_LIT	M_ILL	F_ILL	TOT_WORK_M	TOT
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	0	1999	2598	13381	11364	10007	18432	6723	
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	6	427	517	10513	7891	9072	15211	6982	
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	6	5806	9723	4534	5840	2012	5124	2775	
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	0	2666	3968	1842	1962	942	2244	1002	
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	33	7670	10843	13243	13477	7348	16504	5717	

Table 11: Partial display of first five observations.

	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	F_LIT	M_ILL	F_ILL	TOT_WORK_M	TO
635	34	636	Puducherry	Mahe	3333	8154	11781	1146	1203	21	30	0	0	6916	10184	1238	1597	3808	
636	34	637	Puducherry	Karaikal	10612	12346	21691	1544	1533	2234	4155	0	0	10292	14225	2054	7466	6458	
637	35	638	Andaman & Nicobar Island	Nicobars	1275	1549	2630	227	225	0	0	1012	1750	1187	1602	362	1028	715	
638	35	639	Andaman & Nicobar Island	North & Middle Andaman	3762	5200	8012	723	664	0	0	28	50	4206	5273	994	2739	2707	
639	35	640	Andaman & Nicobar Island	South Andaman	7975	11977	18049	1470	1358	0	0	161	264	10095	13362	1882	4687	6345	

Table 12: Partial display of last five observations.

```
<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
```

Table 13: Partial display of data information

	count	mean	std	min	25%	50%	75%	max
State Code	640.0	17.11	9.43	1.0	9.00	18.0	24.00	35.0
Dist.Code	640.0	320.50	184.90	1.0	160.75	320.5	480.25	640.0
No_HH	640.0	51222.87	48135.41	350.0	19484.00	35837.0	68892.00	310450.0
TOT_M	640.0	79940.58	73384.51	391.0	30228.00	58339.0	107918.50	485417.0
TOT_F	640.0	122372.08	113600.72	698.0	46517.75	87724.5	164251.75	750392.0
M_06	640.0	12309.10	11500.91	56.0	4733.75	9159.0	16520.25	96223.0
F_06	640.0	11942.30	11326.29	56.0	4672.25	8663.0	15902.25	95129.0
M_SC	640.0	13820.95	14426.37	0.0	3466.25	9591.5	19429.75	103307.0
F_SC	640.0	20778.39	21727.89	0.0	5603.25	13709.0	29180.00	156429.0
M_ST	640.0	6191.81	9912.67	0.0	293.75	2333.5	7658.00	96785.0
F_ST	640.0	10155.64	15875.70	0.0	429.50	3834.5	12480.25	130119.0
M_LIT	640.0	57967.98	55910.28	286.0	21298.00	42693.5	77989.50	403261.0
F_LIT	640.0	66359.57	75037.86	371.0	20932.00	43796.5	84799.75	571140.0
M_ILL	640.0	21972.60	19825.61	105.0	8590.00	15767.5	29512.50	105961.0
F_ILL	640.0	56012.52	47116.69	327.0	22367.00	42386.0	78471.00	254160.0
TOT_WORK_M	640.0	37992.41	36419.54	100.0	13753.50	27936.5	50226.75	269422.0
TOT_WORK_F	640.0	41295.76	37192.36	357.0	16097.75	30588.5	53234.25	257848.0
MAINWORK_M	640.0	30204.45	31480.92	65.0	9787.00	21250.5	40119.00	247911.0
MAINWORK_F	640.0	28198.85	29998.26	240.0	9502.25	18484.0	35063.25	226166.0
MAIN_CL_M	640.0	5424.34	4739.16	0.0	2023.50	4160.5	7695.00	29113.0
MAIN_CL_F	640.0	5486.04	5326.36	0.0	1920.25	3908.5	7286.25	36193.0

Table 14: Partial display of statistical description of the data.

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

We are considering the variables NO_HH, TOT_F, TOT_M, MAIN_HH_M and MAIN_HH_F, which depicts the number of households in the given area, Total female population, total male population, female workers in household industries and male workers in house hold industries respectively. We are now going to frame a few questions using the above 5 variables to perform EDA.

a. Which state has highest and lowest number of households?

State	No_HH
Uttar Pradesh	4006871
Maharashtra	3136214
Andhra Pradesh	3127287
Tamil Nadu	2964700
West Bengal	2615284
...	...
Sikkim	16690
Andaman & Nicobar Island	13012
Daman & Diu	7455
Lakshadweep	4445
Dadara & Nagar Havelli	4288

Table 15: Number of households in a state.

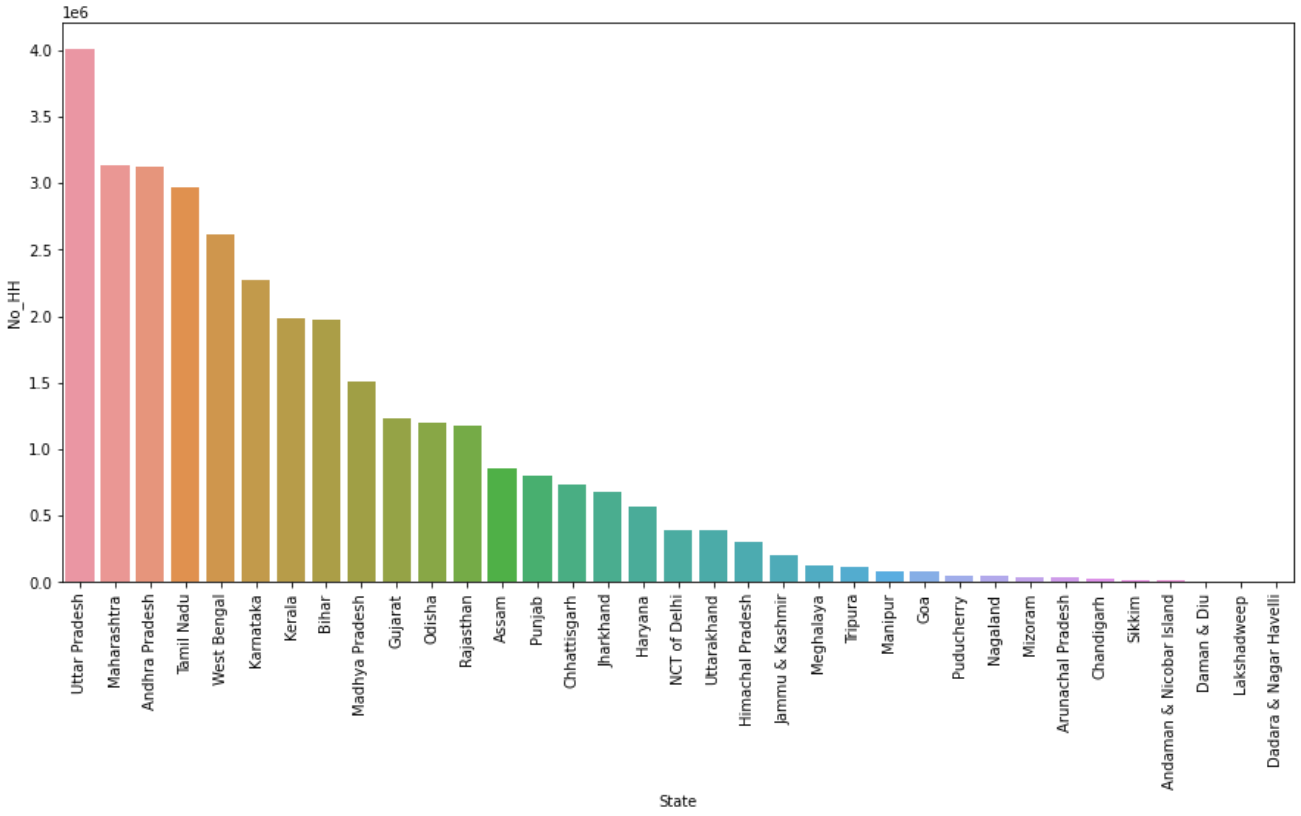


Fig. 6: Count plot – State v/s highest and lowest number of households.

From the table and figure we see that Uttar Pradesh has the maximum number of households and Dadra & Nagar Haveli has the least number of households.

b. Which state has highest and lowest gender ratio?

We have defined the gender ratio to be number of females to the number of males.

State	TOT_M	TOT_F	Gender_ratio
Andhra Pradesh	3274363	6097235	1.862113
Tamil Nadu	3074009	5610310	1.825079
Chhattisgarh	838404	1526592	1.820831
Arunachal Pradesh	50582	88066	1.741054
Odisha	1460031	2536980	1.737621
...	...	...	...
Meghalaya	268036	356355	1.329504
Uttar Pradesh	9043969	12023885	1.329492
NCT of Delhi	833414	1075266	1.290194
Haryana	1167816	1498873	1.283484
Lakshadweep	12823	14772	1.151993

Table 16: Gender ratio in each state.

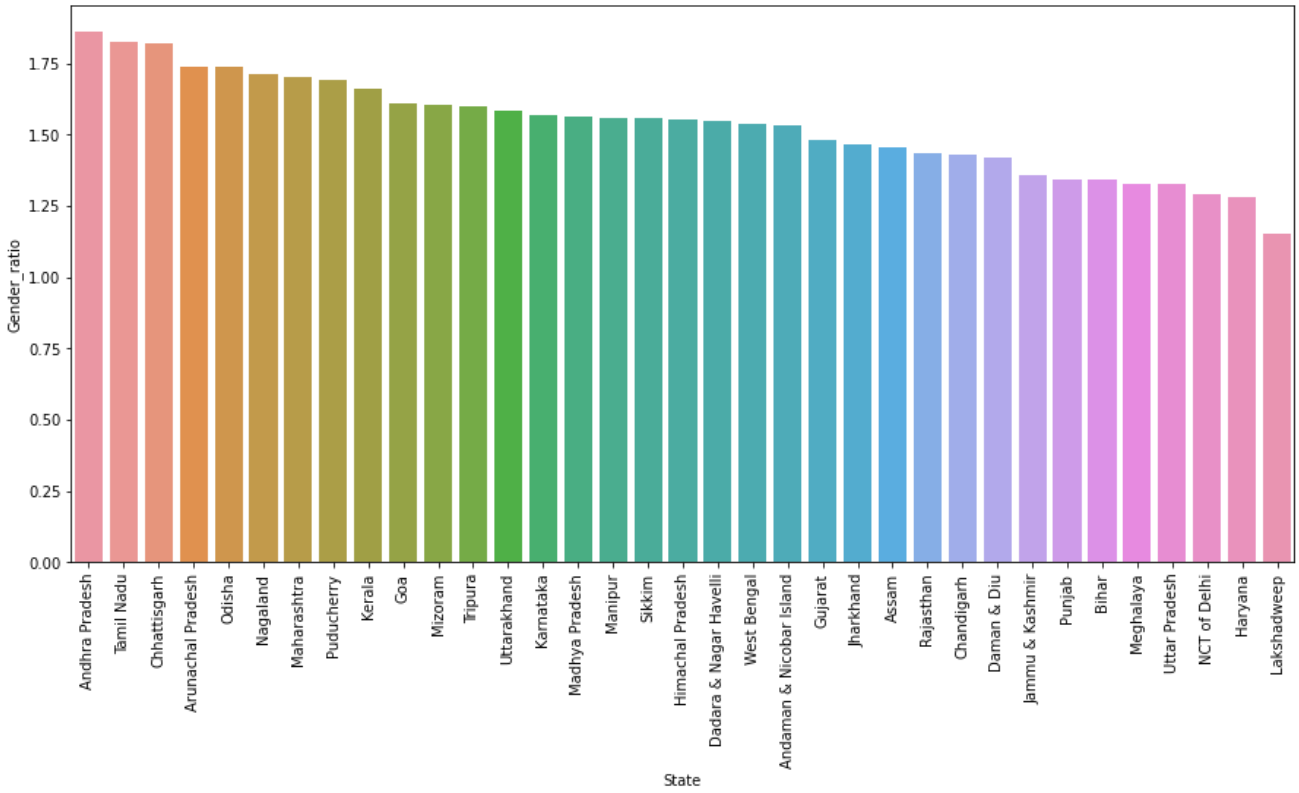


Fig. 7: Count plot – State v/s Gender ratio.

The values of the gender ratio are all greater than 1, which tells us that number of females is greater than the number of males in all states. Gender ratio is greatest in Andhra Pradesh and the least in Lakshadweep.

c. In Karnataka, which district has the most and the least number of female populations?

Area Name	TOT_F
Bangalore	664595
Belgaum	334223
Dakshina Kannada	286313
Mysore	281838
Udupi	262526
...	...
Gadag	105190
Chikmagalur	97951
Chikkaballapura	88696
Bangalore Rural	78317
Kodagu	52937

Table 17: Total female population in Karnataka in different areas.

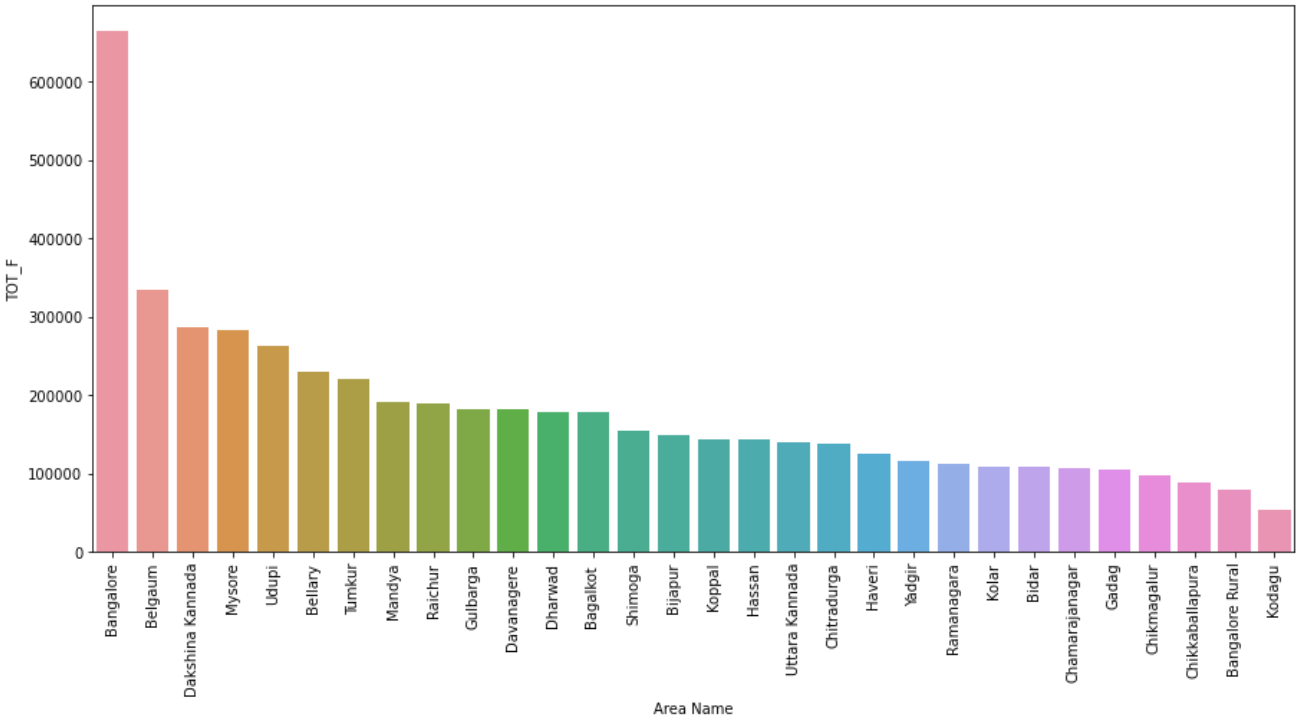


Fig. 8: Count plot – Area v/s female population in Karnataka.

Bangalore has the highest number of females and Kodagu has the least number of females in Karnataka.

d. Which state has the greatest number of house hold industry workers?

State	HH_Workers
Uttar Pradesh	263928
West Bengal	190047
Andhra Pradesh	149919
Tamil Nadu	141510
Karnataka	126662
...	...
Sikkim	320
Andaman & Nicobar Island	145
Dadara & Nagar Havelli	110
Daman & Diu	67
Lakshadweep	48

Table 18: Total household industry workers in different states.

Uttar Pradesh has the greatest number of house hold industry workers, and Lakshadweep has the least number.

e. What percentage of population is working in house hold industries? What is the gender ratio of the workers in the house hold industry?

Only 1.11 % of the total population is working in the house hold industry.

The gender ratio of workers in this industry is 1.56(female is to male), which tells us that there are more female workers in this industry.

f. What is the relation between the number of households and number of house hold industry workers?

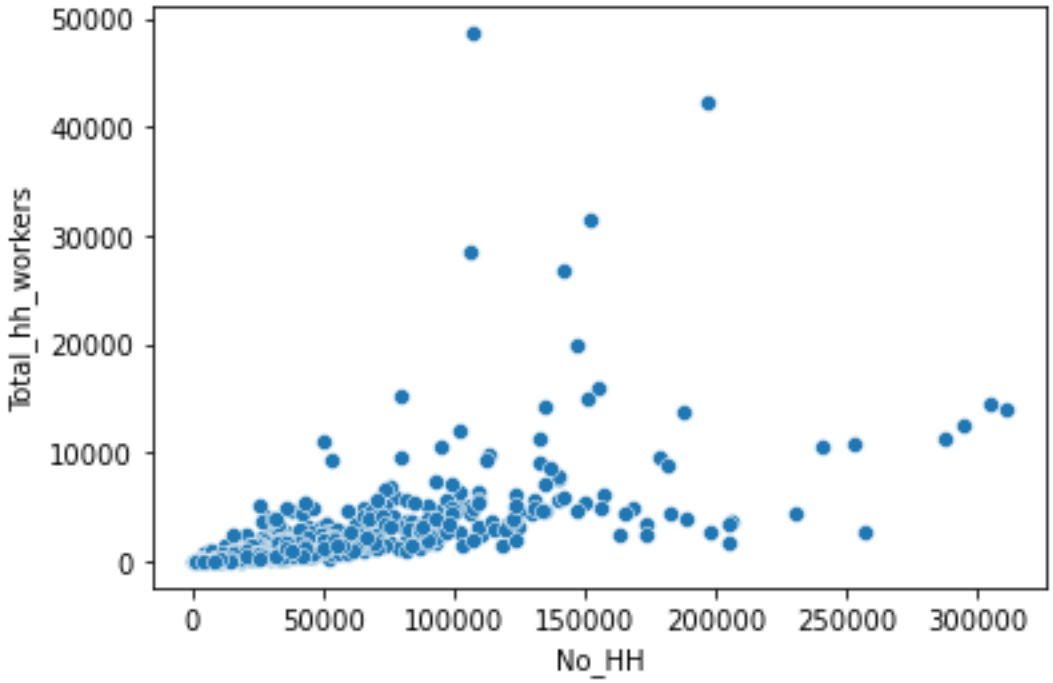


Fig. 9: Scatter plot – Number of households v/s total household industry workers.

From the graph we see that even though the number of house holds is increasing, the values on the y axis is not raising as expected, which tells us that the number of people who prefer working in this industry is less.

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

We see from the graphs below that there are outliers in all columns. Generally, outlier treatment is done before carrying out PCA. But this is a different case since this is the census data, there is a risk of skewing the data, if we treat the outliers. If we carry out the analysis on the skewed data, the conclusions, recommendations that we derive from the analysis may not be of any value.

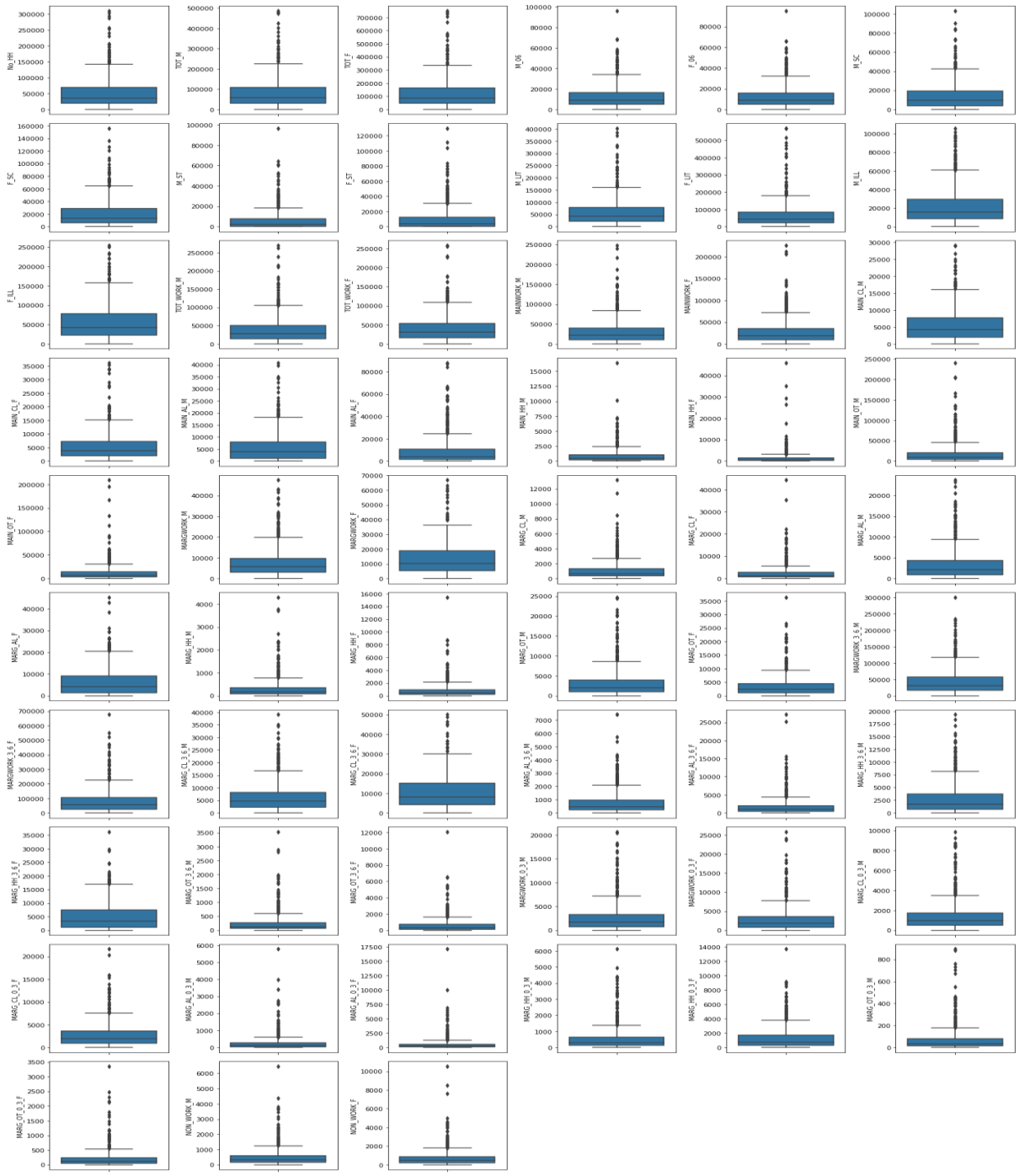


Fig. 10: Box plots depicting the presence of outliers before scaling.

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

We have now applied the z-score method to scale the data. This brings the mean of all columns close to 0 and standard deviation close to 1.

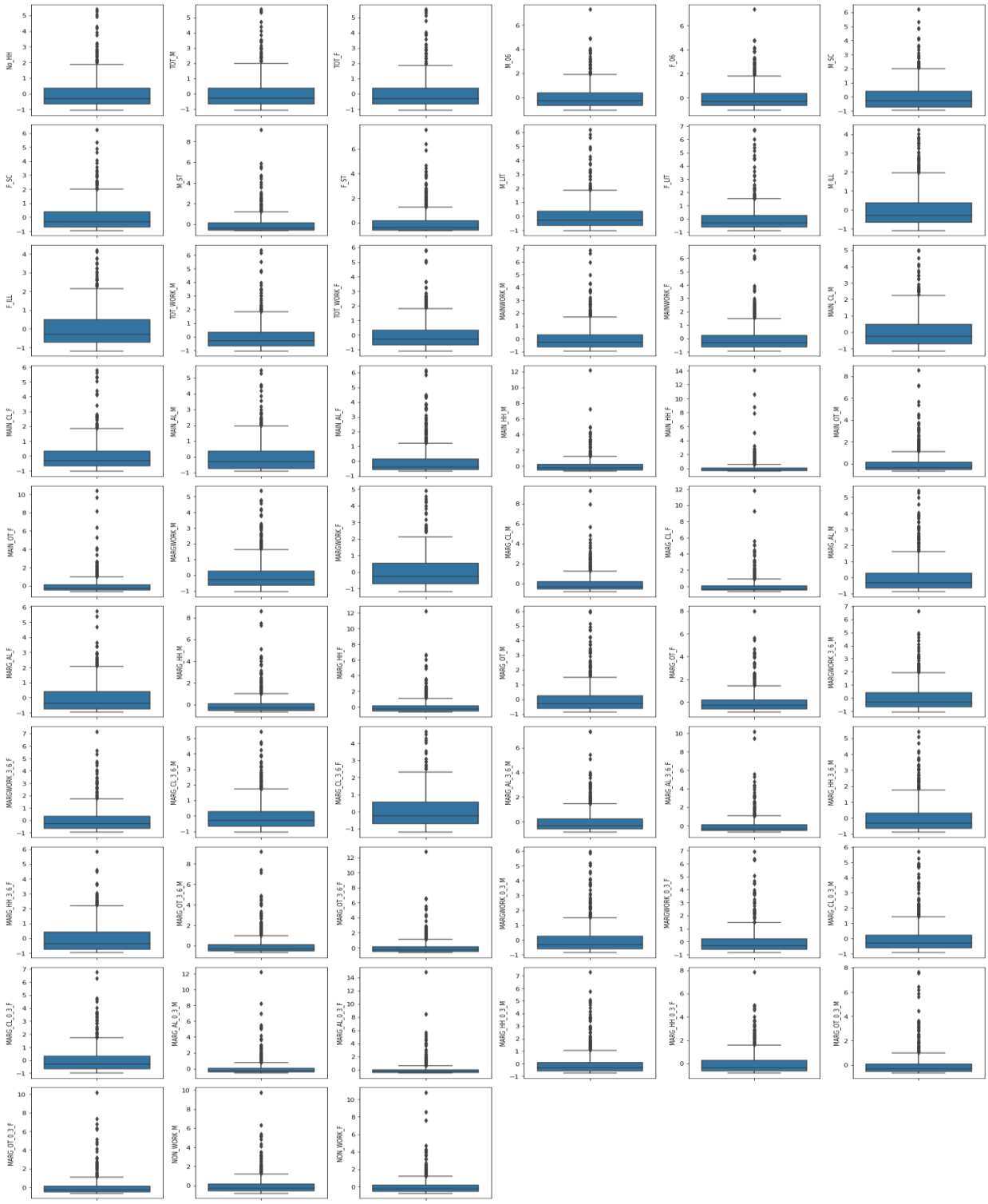


Fig. 11: Box plots depicting the presence of outliers after scaling.

By comparing the figure 6 and 7, we see that outliers are present even after scaling. Hence, we may say that scaling does not have any impact on outliers.

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

Now we have completed scaling, we will proceed to check for correlation among the variables and also check for the adequacy of the data.

a. Presence of correlations: We may check for correlation using heat map and Bartlett test of sphericity.

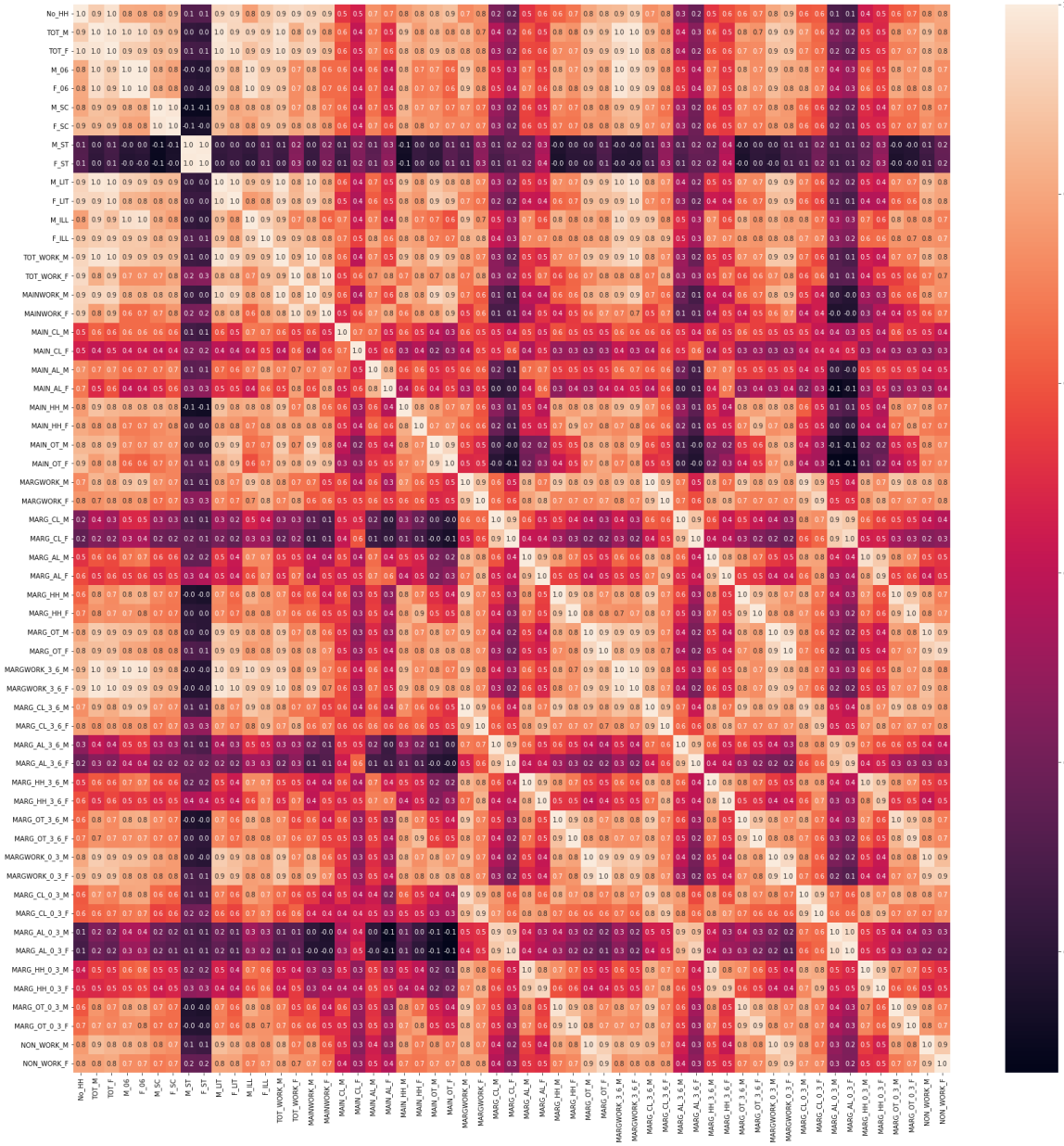


Fig. 12: Heat map depicting the correlation.

Since we have 57 numerical variables, the heatmap looks complicated to come to a conclusion, hence we will test using the Bartlett test of sphericity.

The null hypothesis states that there is no significant correlation and the alternative hypothesis states that there is a significant correlation.

The p-value comes out to be 0.0, which is less than 5%, so we reject the null hypothesis and conclude that there is a significant correlation among the variables and proceed to do PCA.

b. Adequacy of sample: We conduct the Kaiser-Meyer-Olkin Test to determine whether the sample is adequate to conduct PCA. If the p-value is more than 0.7, we say the sample is adequate. In this case we get the p-value to be 0.8039, which is good enough to proceed further.

c. Covariance matrix: This is a 57 x 57 matrix, which gives the covariance of each variable with all the other variables.

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

Table 19: Partial display of the covariance matrix.

d. Principal Component Analysis: We have now applied PCA taking all features into account.

e. Eigen vectors: The eigen vectors tells us the direction of each principal component. Each vector has 57 dimensions.

```
array([[ 1.56020579e-01,  1.67117635e-01,  1.65553179e-01,
         1.62192948e-01,  1.62566396e-01,  1.51357849e-01,
         1.51566500e-01,  2.72341946e-02,  2.81833150e-02,
         1.61992837e-01,  1.46872680e-01,  1.61749445e-01,
         1.65248187e-01,  1.59871988e-01,  1.45935804e-01,
         1.46200730e-01,  1.23970284e-01,  1.03127159e-01,
         7.45397856e-02,  1.13355712e-01,  7.38821590e-02,
         1.31572584e-01,  8.33826397e-02,  1.23526242e-01,
         1.11021264e-01,  1.64615479e-01,  1.55395618e-01,
         8.23885414e-02,  4.91953957e-02,  1.28598563e-01,
         1.14305073e-01,  1.40853227e-01,  1.27669598e-01,
         1.55262872e-01,  1.47286584e-01,  1.64971950e-01,
         1.61253433e-01,  1.65501611e-01,  1.55647049e-01,
         9.30142064e-02,  5.15358640e-02,  1.28576116e-01,
         1.10645843e-01,  1.39592763e-01,  1.24545909e-01,
         1.54293786e-01,  1.46285654e-01,  1.50125706e-01,
         1.40157047e-01,  5.25417829e-02,  4.17859530e-02,
         1.21840354e-01,  1.16011410e-01,  1.39868774e-01,
         1.32192245e-01,  1.50375578e-01,  1.31066203e-01],
       [-1.26346525e-01, -8.96765481e-02, -1.04912371e-01,
        -2.20945086e-02, -2.02705496e-02, -4.51109032e-02,
        -5.19237543e-02,  2.76790387e-02,  3.02225550e-02,
        -1.15354767e-01, -1.53109487e-01, -6.62537326e-03,
        -9.10743690e-03, -1.33529221e-01, -8.50869690e-02,
        -1.76368057e-01, -1.51412544e-01,  6.24149872e-02,
         8.64767271e-02, -3.10403497e-02, -5.86880215e-02,
        -7.60210677e-02, -8.24766376e-02, -2.12984254e-01,
        -2.10071166e-01,  9.29935013e-02,  1.25269967e-01,
         2.69449716e-01,  2.46546811e-01,  1.65830750e-01,
         1.40957749e-01,  6.80679428e-02,  2.42164125e-02,
        -8.94419719e-02, -1.17899307e-01, -4.39949602e-02,
        -1.05501898e-01,  7.71926976e-02,  1.03173976e-01,
         2.64409408e-01,  2.44261318e-01,  1.58782773e-01,
         1.25286970e-01,  6.22623250e-02,  1.47659020e-02,
        -9.31585893e-02, -1.25595577e-01,  1.50680869e-01,
         1.80690375e-01,  2.51328441e-01,  2.40719745e-01,
         1.85277342e-01,  1.80615650e-01,  8.48690450e-02,
         5.08133220e-02, -6.53645528e-02, -7.38474209e-02],
       [-2.69024836e-03,  5.66976187e-02,  3.87494756e-02,
         5.77881520e-02,  5.01255683e-02,  2.56890443e-03,
        -2.51008777e-02, -1.23504453e-01, -1.39768832e-01,
         8.21676681e-02,  1.17097686e-01, -2.18550959e-02,
        -9.30623778e-02,  4.51763683e-02, -5.94495459e-02,
         5.42945280e-02, -5.56090962e-02, -6.73992946e-02,
        -9.23809162e-03, -2.47917056e-01, -2.51932294e-01
```

Table 20: Partial display of the eigen vectors.

f. Eigen Values: The eigen values give the length or magnitude of the corresponding eigen vectors. We have a total of 57 eigen values.

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

Table 21: Eigen values of each principal component.

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

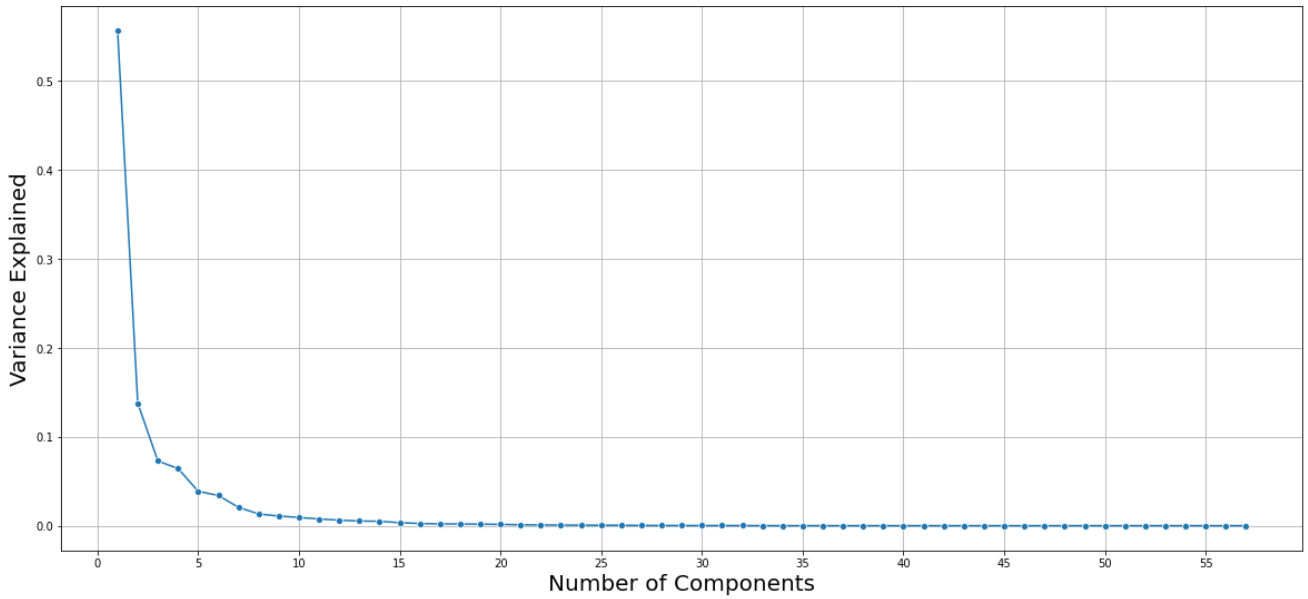


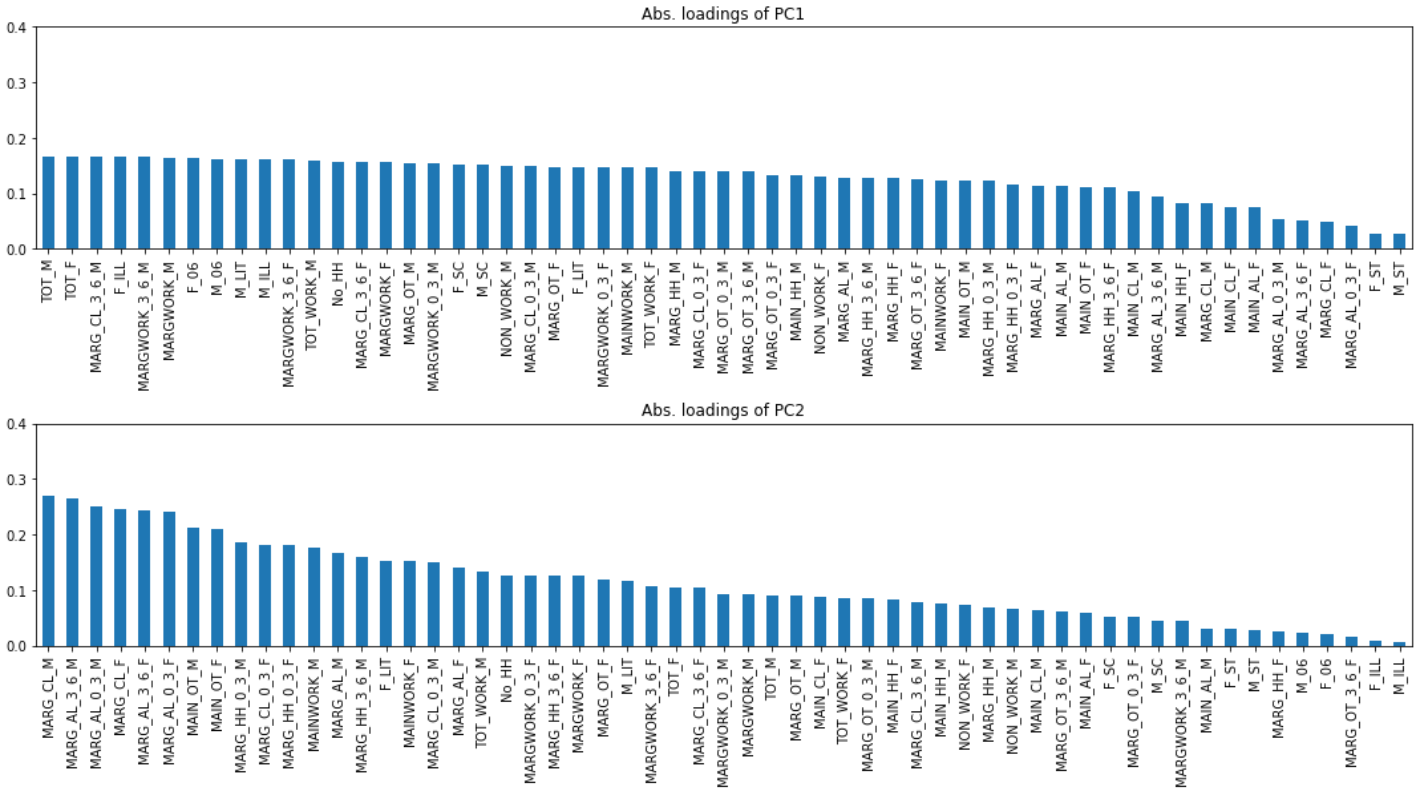
Fig. 13: Scree plot.

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

Table 22: Cumulative sums of the explained variance ratio.

From the scree plot and table above, we see that 6 principal components are capturing about 90.47% of the variance in the data. Hence, the optimum number of principal components would be 6.

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



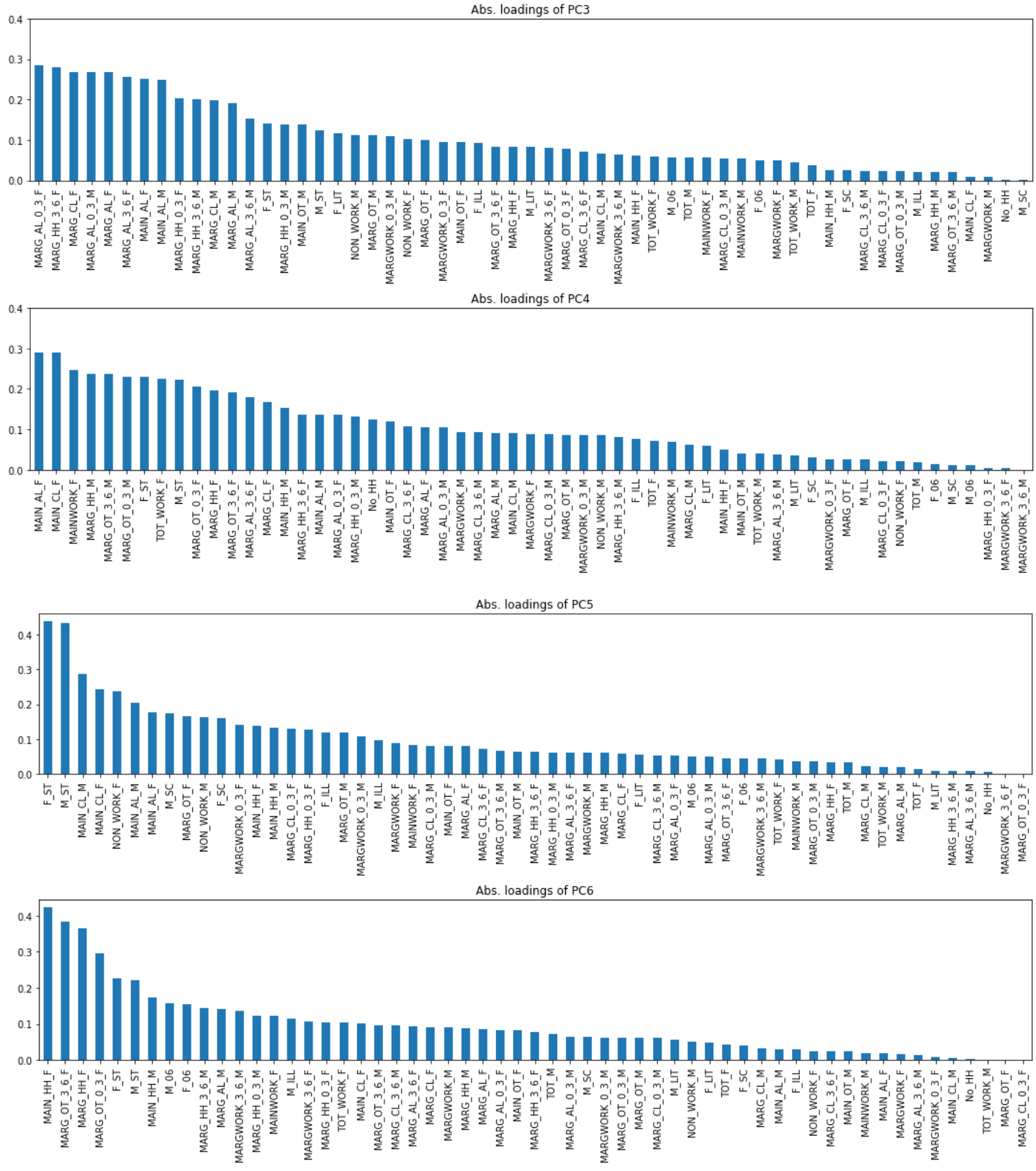


Fig. 14: Absolute loadings of different PCs.

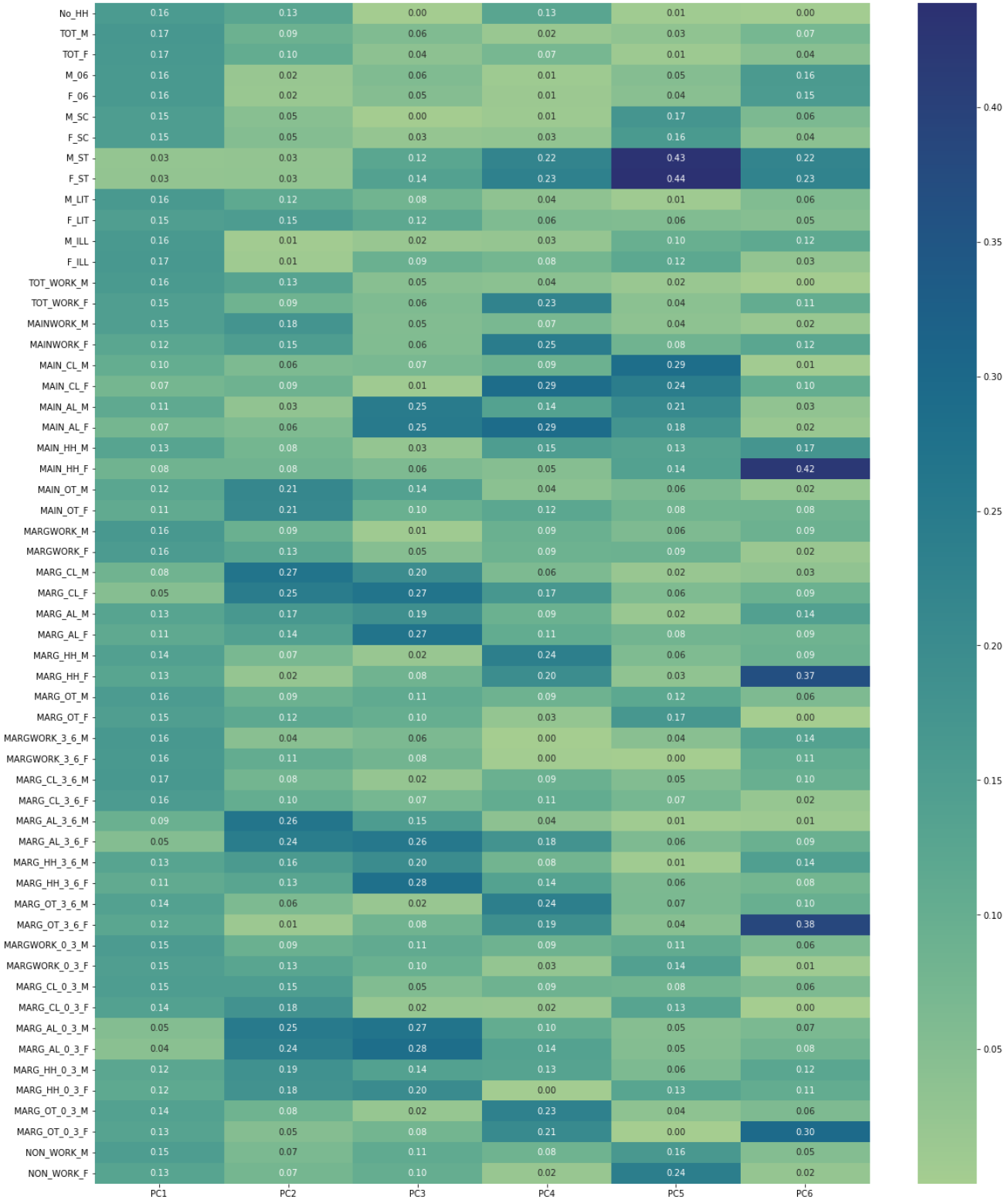


Fig. 15: Heat map depicting the influence of original features on PCs.

From Table 21, we see that the first principal component explains 55.72% of the variance of the data. Hence, it is the most important principal component. From figure 14, we see that the variables such as TOT_M, TOT_F, MARG_CL_3_6_M, F_ILL, MARGWORK _3_6_M etc have more influence on the first principal component, we may say these variables are important.

For the second principal component, MARG_CL_M, MARG_AL_3_6_M, MARG_AL_0_3_M have more influence.

The third principal component carries more weight from the variables such as MARG_AL_0_3_F, MARG_HH_3_6_F, MARG_CL_F.

The fourth principal component is has the highest influence from MAIN_AL_F, MAIN_CL_F,MAINWORK_F.

The fifth principal component is made up of F_ST, M_ST, MAIN_CL_M.

Finally, the last principal component has the highest influence from MAIN_HH_F, MARG_OT_3_6_F, MARG_HH_F.

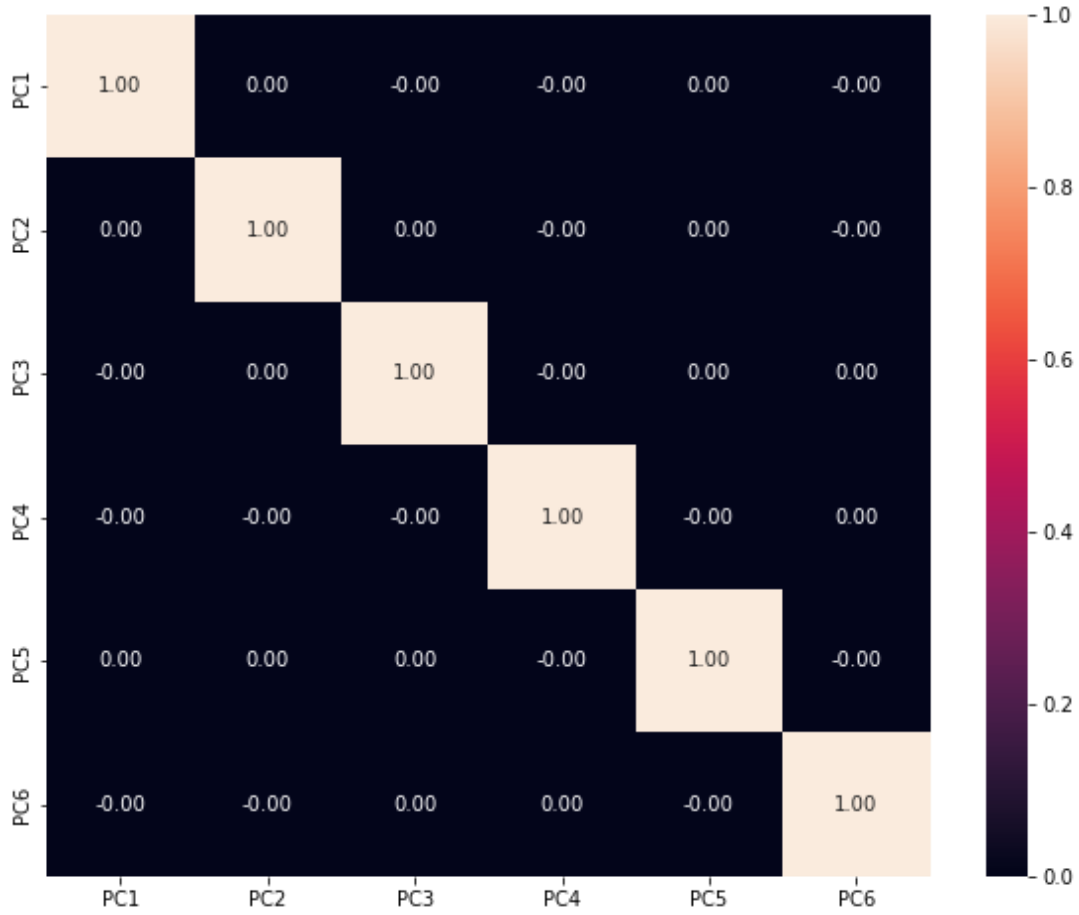


Fig. 16: Heat map showing no correlation among the 6 PCs.

2.8 Write linear equation for first PC.

The linear equation any principal components is of the form:

$PC_j = W_{j1}X_1 + W_{j2}X_2 + \ldots + W_{jn}X_n$

In this case, it is given as below.

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

Table 23: Linear equation of first PC.