# Business Report

# Data Mining

**Dhruv Dosad**

**Contents**

**Problem 1**

**Clustering: Digital Ads Data:**

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

The following three features are commonly used in digital marketing:

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

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

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

**The Data Dictionary**

| Sl. No | Column Name | Column Description |
|---|---|---|
| 1 | Timestamp | The Timestamp of the particular Advertisement. |
| 2 | InventoryType | The Inventory Type of the particular Advertisement. Format 1 to 7. This is a Categorical Variable. |
| 3 | Ad - Length | The Length Dimension of the particular Adverstisement. |
| 4 | Ad- Width | The Width Dimension of the particular Advertisement. |
| 5 | Ad Size | The Overall Size of the particular Advertisement. Length*Width. |
| 6 | Ad Type | The type of the particular Advertisement. This is a Categorical Variable. |
| 7 | Platform | The platform in which the particular Advertisement is displayed. Web, Video or App. This is a Categorical Variable. |
| 8 | Device Type | The type of the device which supports the partciular Advertisement. This is a Categorical Variable. |

| | | |
|---|---|---|
| 9 | Format | The Format in which the Advertisement is displayed. This is a Categorical Variable. |
| 10 | Available_Impressions | How often the particular Advertisement is shown. An impression is counted each time an Advertisement is shown on a search result page or other site on a Network. |
| 11 | Matched_Queries | Matched search queries data is pulled from Advertising Platform and consists of the exact searches typed into the search Engine that generated clicks for the particular Advertisement. |
| 12 | Impressions | The impression count of the particular Advertisement out of the total available impressions. |
| 13 | Clicks | It is a marketing metric that counts the number of times users have clicked on the particular advertisement to reach an online property. |
| 14 | Spend | It is the amount of money spent on specific ad variations within a specific campaign or ad set. This metric helps regulate ad performance. |
| 15 | Fee | The percentage of the Advertising Fees payable by Franchise Entities. |
| 16 | Revenue | It is the income that has been earned from the particular advertisement. |
| 17 | CTR | CTR stands for "Click through rate". CTR is the number of clicks that your ad receives divided by the number of times your ad is shown. Formula used here is CTR = Total Measured Clicks / Total Measured Ad Impressions x 100. Note that the Total Measured Clicks refers to the 'Clicks' Column and the Total Measured Ad Impressions refers to the 'Impressions' Column. |
| 18 | CPM | CPM stands for "cost per 1000 impressions." Formula used here is CPM = (Total Campaign Spend / Number of Impressions) * 1,000. Note that the Total Campaign Spend refers to the 'Spend' Column and the Number of Impressions refers to the 'Impressions' Column. |
| 19 | CPC | CPC stands for "Cost-per-click". Cost-per-click (CPC) bidding means that you pay for each click on your ads. The Formula used here is CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column and the Number of Clicks refers to the 'Clicks' Column. |

Perform the following in given order:

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

**Answer:** Reading the head & tail of the dataset

**Head & Tail of data** - Top & bottom 5 values

| index | Timestamp | InventoryType | Ad - Length | Ad- Width | Ad Size | Ad Type | Platform | Device Type | Format | Available_Impressions | Matched_Queries | Impressions | Clicks | Spend | Fee | Revenue | CTR | CPM | CPC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2020-9-2-17 | Format1 | 300 | 250 | 75000 | Inter222 | Video | Desktop | Display | 1806 | 325 | 323 | 1 | 0.0 | 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 | 0.002 | 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 |

| index | Timestamp | InventoryType | Ad - Length | Ad- Width | Ad Size | Ad Type | Platform | Device Type | Format | Available_Impressions | Matched_Queries | Impressions | Clicks | Spend | Fee | Revenue | CTR | CPM | CPC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 23061 | 2020-9-13-7 | Format5 | 720 | 300 | 216000 | Inter220 | Web | Mobile | Video | 1 | 1 | 1 | 1 | 0.07 | 0.35 | 0.045500000000000006 | NaN | NaN | NaN |
| 23062 | 2020-11-2-7 | Format5 | 720 | 300 | 216000 | Inter224 | Web | Desktop | Video | 3 | 2 | 2 | 1 | 0.04 | 0.35 | 0.026000000000000002 | 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 | 0.045500000000000006 | NaN | NaN | NaN |
| 23065 | 2020-9-14-0 | Format5 | 720 | 300 | 216000 | Inter221 | App | Mobile | Video | 2 | 2 | 2 | 1 | 0.09 | 0.35 | 0.058499999999999996 | NaN | NaN | NaN |

**Tail - bottom 5 values**

**Info of the dataset**

```
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)
```

**Shape of dataset:**

- The dataset has **23066 rows & 19 columns**

**Checking for Null Values**

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

We can see that there are **3 variables** wherein **we have Null values**

- **CTR, CPM, CPC** having **4736 Null values** each

**Checking for Duplicate values**

There are **no duplicate** values in the dataset

**Five Point Summary of the dataset**

| index | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Ad - Length | 23066.0 | 385.163097199341 | 233.65143380263876 | 120.0 | 120.0 | 300.0 | 720.0 | 728.0 |
| Ad- Width | 23066.0 | 337.89603745773 | 203.09288492918358 | 70.0 | 250.0 | 300.0 | 600.0 | 600.0 |
| Ad Size | 23066.0 | 96674.46804820948 | 61538.32955689539 | 33600.0 | 72000.0 | 72000.0 | 84000.0 | 216000.0 |
| Available_Impressions | 23066.0 | 2432043.665871846 | 4742887.764666151 | 1.0 | 33672.25 | 483771.0 | 2527711.75 | 27592861.0 |
| Matched_Queries | 23066.0 | 1295099.1432411342 | 2512969.8612578264 | 1.0 | 18282.5 | 258087.5 | 1180700.0 | 14702025.0 |
| Impressions | 23066.0 | 1241519.5188589266 | 2429399.9610913517 | 1.0 | 7990.5 | 225290.0 | 1112428.5 | 14194774.0 |
| Clicks | 23066.0 | 10678.518815572705 | 17353.409362737486 | 1.0 | 710.0 | 4425.0 | 12793.75 | 143049.0 |
| Spend | 23066.0 | 2706.625688892743 | 4067.9272729126783 | 0.0 | 85.18 | 1425.125 | 3121.4 | 26931.87 |
| Fee | 23066.0 | 0.33512312494580765 | 0.03196322155198887 | 0.21 | 0.33 | 0.35 | 0.35 | 0.35 |
| Revenue | 23066.0 | 1924.25233071187 | 3105.2384103287695 | 0.0 | 55.365375 | 926.335 | 2091.3381499999996 | 21276.18 |
| CTR | 18330.0 | 0.07366054009819965 | 0.07515992494457631 | 0.0001 | 0.0026 | 0.08255000000000001 | 0.13 | 1.0 |
| CPM | 18330.0 | 7.672045280960174 | 6.481390870689706 | 0.0 | 1.71 | 7.66 | 12.51 | 81.56 |
| CPC | 18330.0 | 0.3510605564648118 | 0.34333379368050065 | 0.0 | 0.09 | 0.16 | 0.57 | 7.26 |

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

**Answer:** Treating the Null values using a user defined function.

We created a function given in the problem statement as below. The function successfully added the relevant data in the respective CPM, CPC & CTR variables.

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

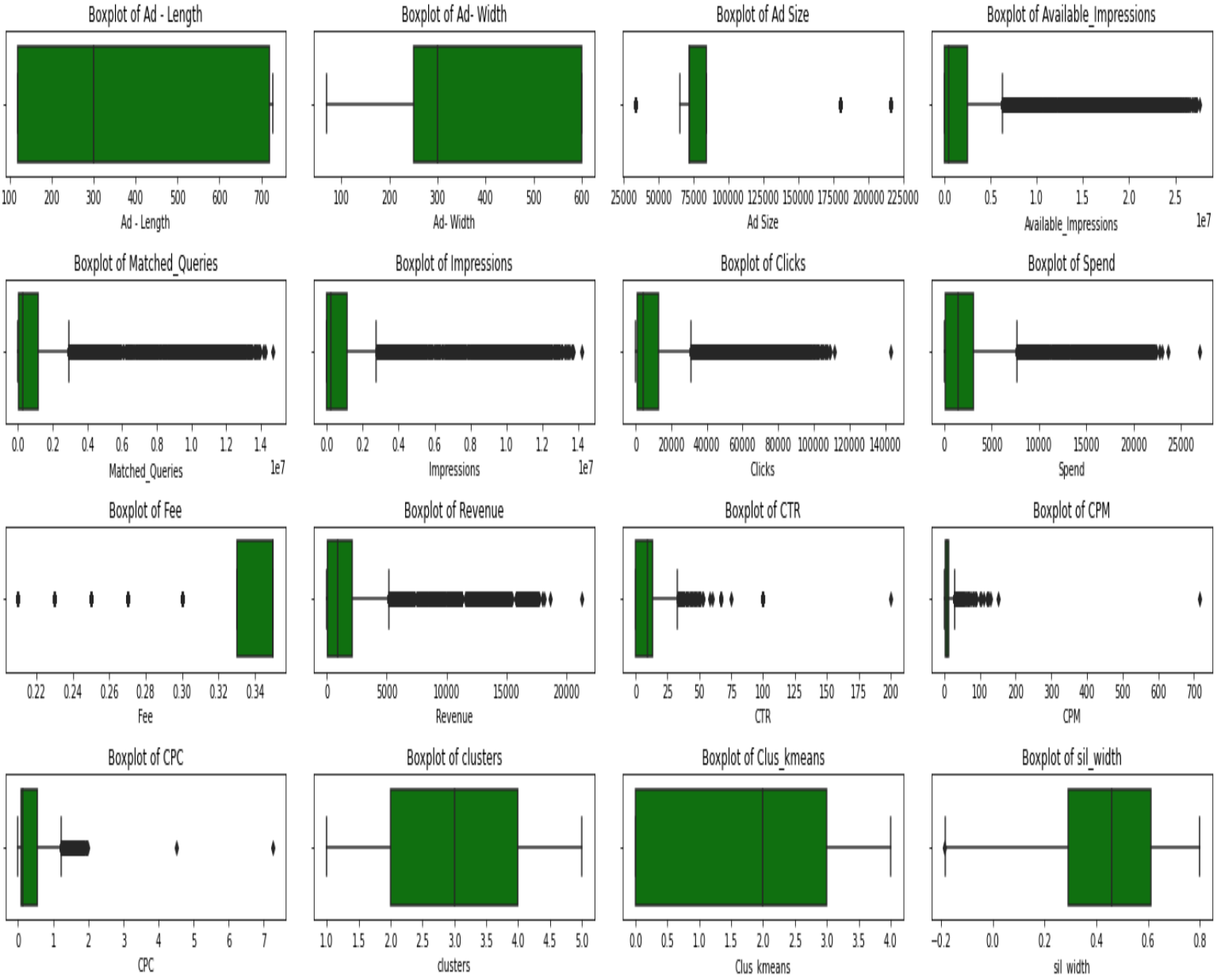
**CPC = Total Cost (spend) / Number of Clicks**.

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

Dataset info after the null 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                      0
CPM                      0
CPC                      0
dtype: int64
```

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

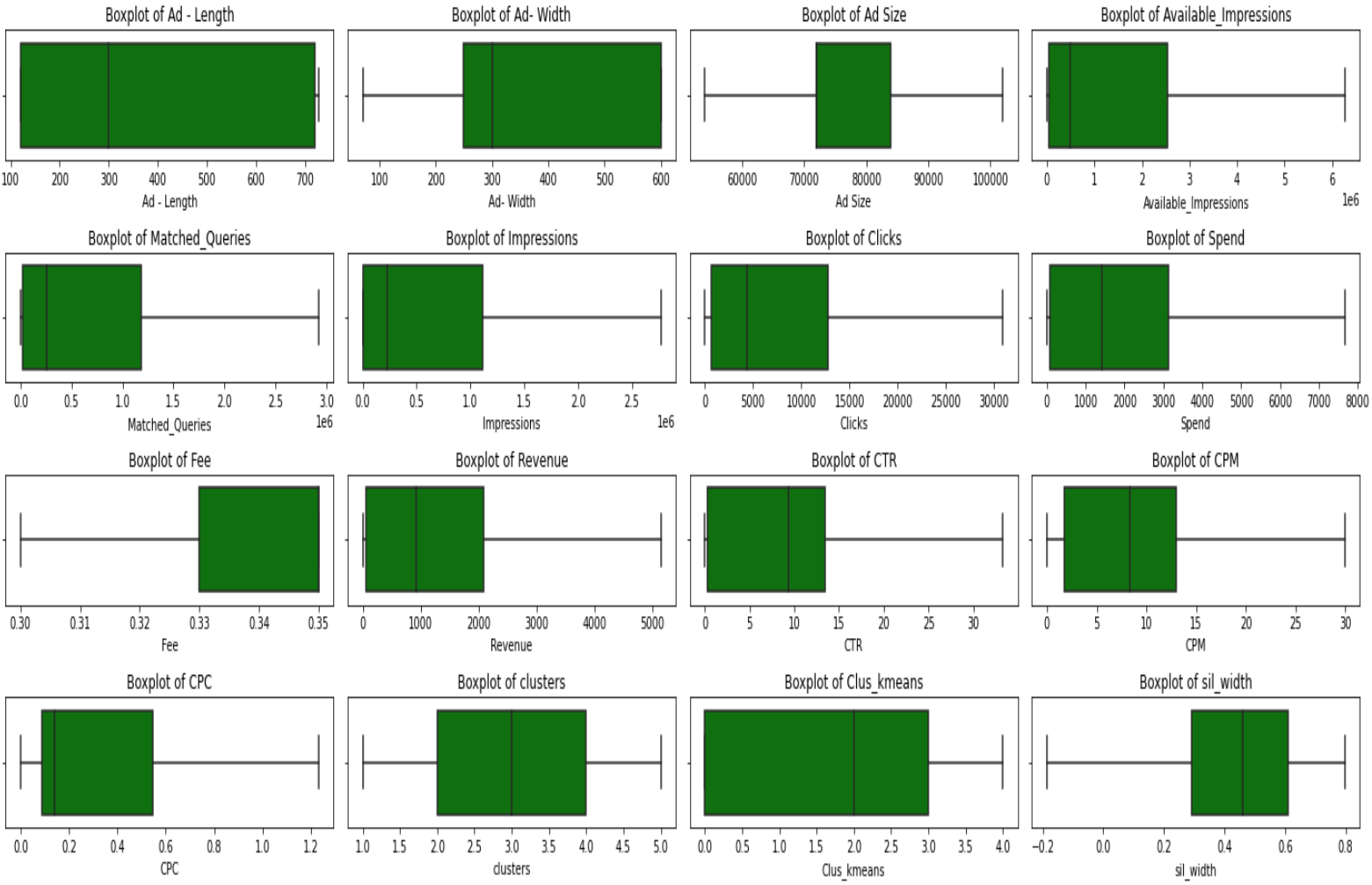



**Answer:** Outliers before treatments

**Outlier treatment :** We will treat the outliers basis the 1.5*IQR on Q1 & Q3 to set upper & lower range i.e.

Upper Range = Q3 + 1.5*IQR.

Lower Range = Q1 -1.5*IQR




**Boxplot after the treatments**

From above boxplot, we can see that the outliers are treated & ready for further analysis

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

We applied z-score scaling on the dataset

**Head & tail of the scaled dataset as shown below:**

| | 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 |

23066 rows × 13 columns

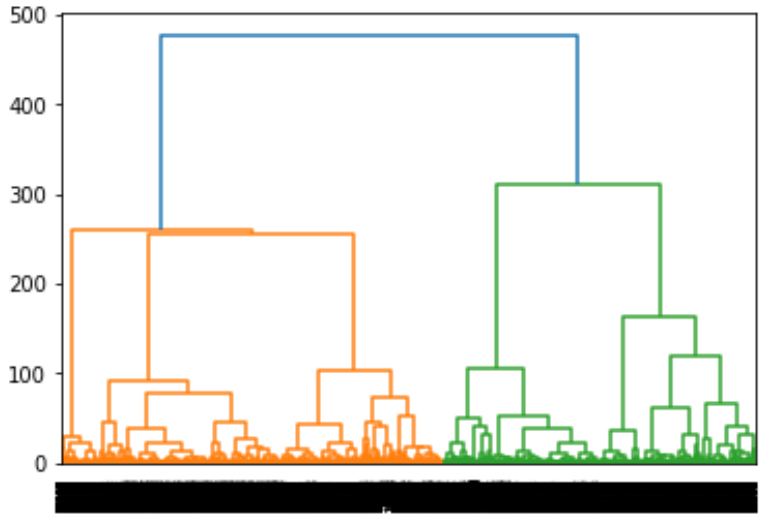
**Summary of the z-score scaled dataset**

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Ad - Length | 23066.0 | 1.281478e-16 | 1.000022 | -1.134891 | -1.134891 | -0.364496 | 1.433093 | 1.467332 |
| Ad- Width | 23066.0 | -1.182903e-16 | 1.000022 | -1.319110 | -0.432797 | -0.186599 | 1.290590 | 1.290590 |
| Ad Size | 23066.0 | 3.055833e-16 | 1.000022 | -1.467840 | -0.297564 | -0.297564 | 0.482620 | 1.652896 |
| Available_Impressions | 23066.0 | 9.857525e-18 | 1.000022 | -0.756182 | -0.740341 | -0.528577 | 0.433059 | 2.193158 |
| Matched_Queries | 23066.0 | 1.971505e-17 | 1.000022 | -0.779265 | -0.761447 | -0.527722 | 0.371498 | 2.070914 |
| Impressions | 23066.0 | 0.000000e+00 | 1.000022 | -0.768806 | -0.760655 | -0.538975 | 0.366051 | 2.056111 |
| Clicks | 23066.0 | -1.182903e-16 | 1.000022 | -0.867488 | -0.793438 | -0.405431 | 0.468629 | 2.361729 |
| Spend | 23066.0 | -9.857525e-17 | 1.000022 | -0.893170 | -0.858046 | -0.305523 | 0.393932 | 2.271900 |
| Fee | 23066.0 | 1.143473e-15 | 1.000022 | -2.222416 | -0.567532 | 0.535724 | 0.535724 | 0.535724 |
| Revenue | 23066.0 | 3.943010e-17 | 1.000022 | -0.880093 | -0.846474 | -0.317607 | 0.389803 | 2.244218 |
| CTR | 23066.0 | 1.380054e-16 | 1.000022 | -0.995031 | -0.964227 | 0.141524 | 0.635787 | 3.035808 |
| CPM | 23066.0 | 2.464381e-17 | 1.000022 | -1.194498 | -0.940303 | 0.022146 | 0.700905 | 3.162718 |
| CPC | 23066.0 | 3.943010e-17 | 1.000022 | -1.042561 | -0.759091 | -0.602371 | 0.682987 | 2.846105 |

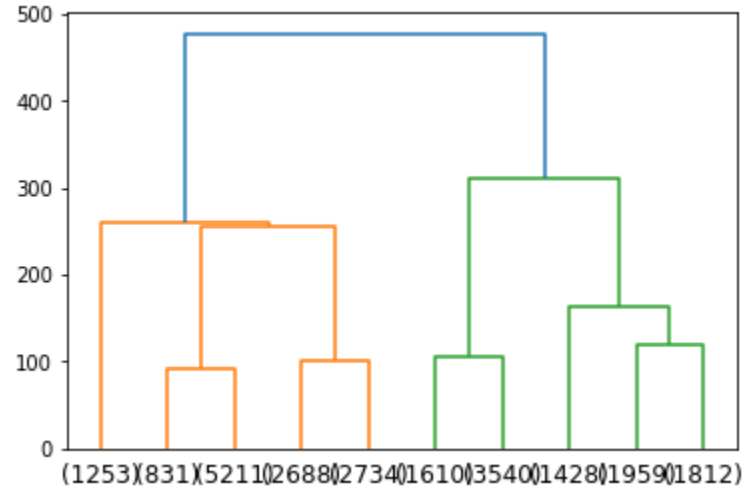
- **Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.**

**Answer:**

Performing Hierarchical clustering by plotting Dendrogram using Ward & Euclidean distance




Dendrogram with p=10




**Cluster Frequency**

| | |
|---|---|
| 1 | 1253 |
| 2 | 6042 |
| 3 | 5422 |
| 4 | 5150 |
| 5 | 5199 |

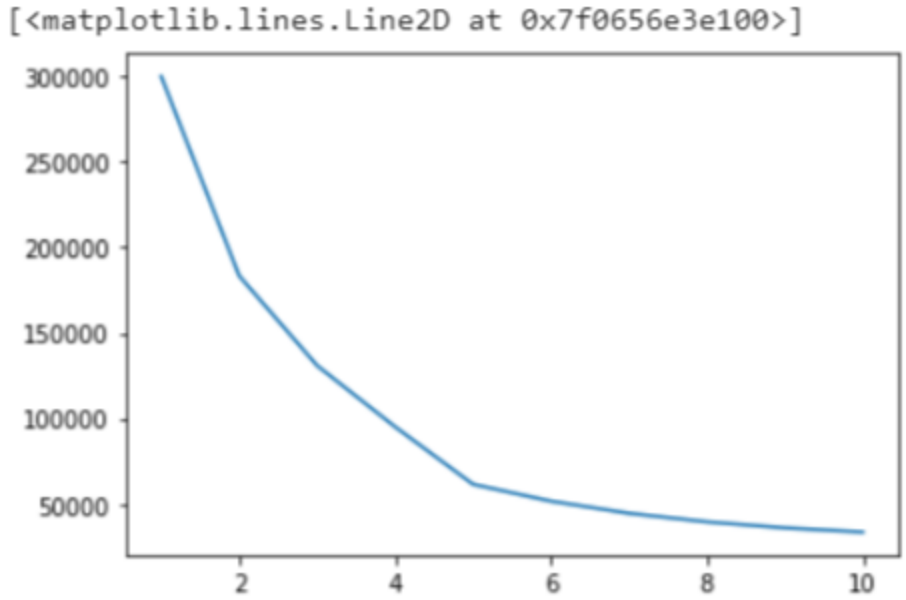
**Head of the dataset after Clustering**

| | Timestamp | InventoryType | Ad - Length | Ad- Width | Ad Size | Ad Type | Platform | Device Type | Format | Available_Impressions | ... | Clicks | Spend | Fee | Revenue | CTR | CPM | CPC | clusters |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2020-9-2-17 | Format1 | 300 | 250 | 75000 | Inter222 | Video | Desktop | Display | 1806 | ... | 1 | 0.0 | 0.35 | 0.0 | 0.309598 | 0.0 | 0.0 | 4 |
| 1 | 2020-9-2-10 | Format1 | 300 | 250 | 75000 | Inter227 | App | Mobile | Video | 1780 | ... | 1 | 0.0 | 0.35 | 0.0 | 0.350877 | 0.0 | 0.0 | 4 |
| 2 | 2020-9-1-22 | Format1 | 300 | 250 | 75000 | Inter222 | Video | Desktop | Display | 2727 | ... | 1 | 0.0 | 0.35 | 0.0 | 0.281690 | 0.0 | 0.0 | 4 |
| 3 | 2020-9-3-20 | Format1 | 300 | 250 | 75000 | Inter228 | Video | Mobile | Video | 2430 | ... | 1 | 0.0 | 0.35 | 0.0 | 0.202020 | 0.0 | 0.0 | 4 |
| 4 | 2020-9-4-15 | Format1 | 300 | 250 | 75000 | Inter217 | Web | Desktop | Video | 1218 | ... | 1 | 0.0 | 0.35 | 0.0 | 0.413223 | 0.0 | 0.0 | 4 |

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

**Answer:**

**Elbow Plot for n=10**




The optimum number of clusters for k-means algorithm are 5 as the drop become significant

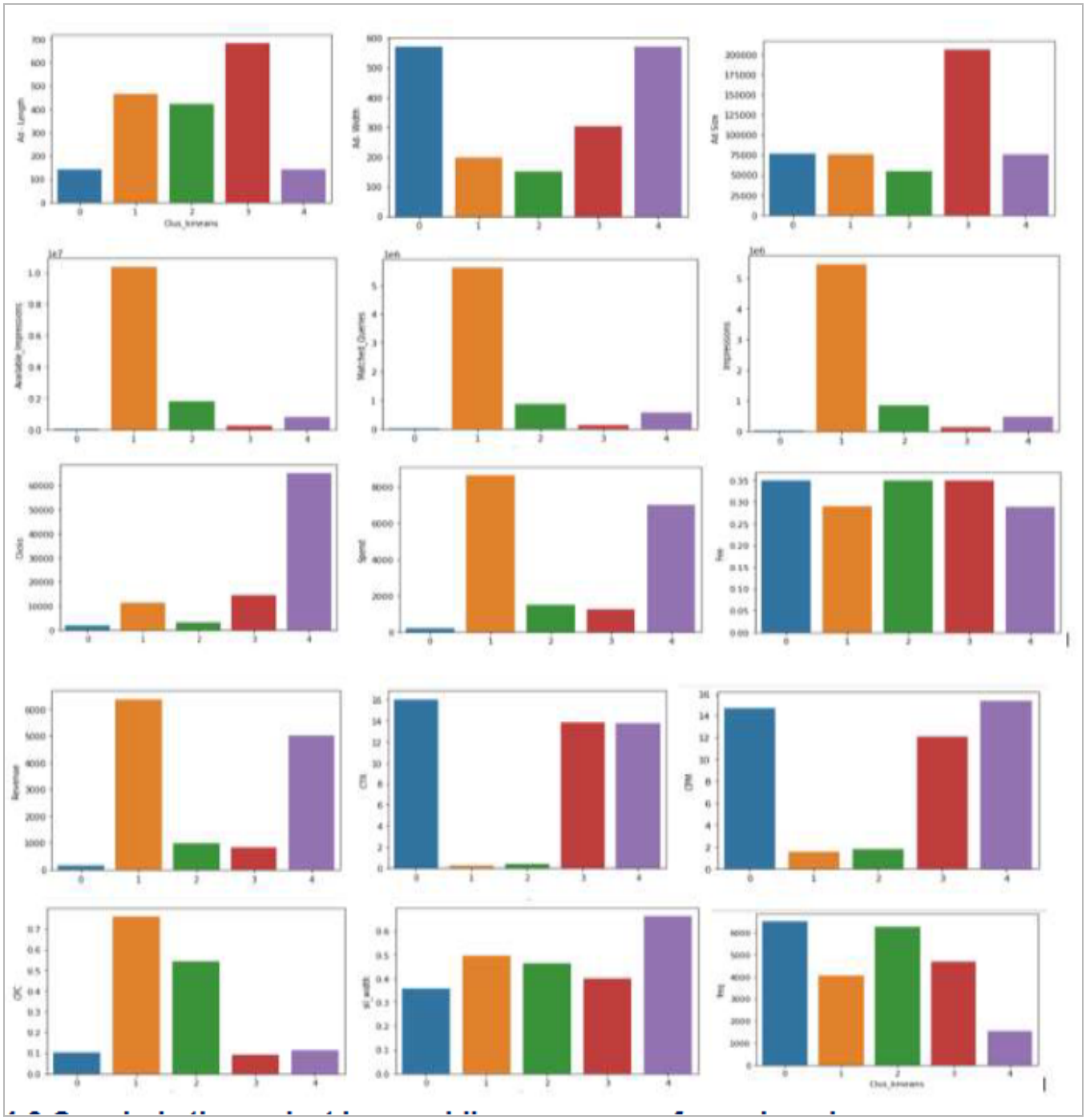
- **Print silhouette scores for up to 10 clusters and identify optimum number of clusters.**

**Answer:** The silhouette score here is 0.5240956940501831.
The silhouette score for rest clusters upto 10.

```
2
183349.10202886112
0.38572769619101077
-0.052971575130041706

3
130878.34240367389
0.3825486036570082
-0.13228597797973302

4
95133.92434119384
0.44534519247649795
-0.1357679674165574

5
61539.18919785395
0.5240956940501831
-0.03709167751991964

6
51676.89230709949
0.5221533662938636
-0.05673927439483464

7
44598.25849746805
0.5165635029478517
-0.04895751557522969

8
39597.84955874646
0.4797334335943954
-0.19992415610651804

9
36061.65051550149
0.4319440318900223
-0.28287351144842826

10
32998.39164098352
0.4406696514831378
-0.18687864301172427
```

- **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].**

| Clus_kmeans | Ad - Length | Ad- Width | Ad Size | Available_Impressions | Matched_Queries | Impressions | Clicks | Spend | Fee | Revenue | CTR | CPM | CPC | clusters | sil_width | freq |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 141.454782 | 572.446324 | 75614.834092 | 8.063284e+05 | 5.668641e+05 | 4.781485e+05 | 65315.176318 | 6990.360898 | 0.288302 | 5017.538285 | 13.752664 | 15.385753 | 0.111918 | 1.202342 | 0.663102 | 1537 |
| 1 | 683.825492 | 303.785287 | 206160.821215 | 2.513465e+05 | 1.375509e+05 | 1.167714e+05 | 14406.540205 | 1252.285569 | 0.349538 | 815.541831 | 13.857220 | 12.098200 | 0.090012 | 3.001069 | 0.402930 | 4676 |
| 2 | 465.781944 | 199.148989 | 75176.566354 | 1.038821e+07 | 5.625808e+06 | 5.447310e+06 | 11245.754810 | 8646.647997 | 0.290439 | 6373.659814 | 0.217242 | 1.573280 | 0.760929 | 5.000000 | 0.496800 | 4054 |
| 3 | 421.696255 | 152.001594 | 55008.841434 | 1.810314e+06 | 8.642623e+05 | 8.262209e+05 | 3263.131952 | 1500.090563 | 0.349264 | 977.424163 | 0.404392 | 1.788731 | 0.544614 | 4.182470 | 0.462916 | 6275 |
| 4 | 143.280809 | 572.103004 | 76597.026364 | 3.209356e+04 | 1.962406e+04 | 1.349204e+04 | 1914.448804 | 209.162609 | 0.349988 | 135.993379 | 16.037897 | 14.693481 | 0.102794 | 2.115573 | 0.343806 | 6524 |



**-Conclusions in next points**

- **Conclude the project by providing a summary of your learnings.**

**We can conclude that**

Cluster-1 have the most Click through rate (CTR).

Cluster-5 have the highest cost per 1000 impressions (CPM)

Cluster-2 have generated the maximum revenue followed by Cluster-5 than Cluster-3

Cluster 5 have the maximum clicks followed by Cluster-4 than Cluster-2.
Maximum spent has been done on Cluster-2 followed by Cluster-5 than Cluster-3.

**PCA:**

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

**Data dictionary:**

| Name | Description |
|---|---|
| State | State Code |
| District | District Code |
| Name | Name |
| TRU1 | Area Name |
| No_HH | No of Household |
| TOT_M | Total population Male |
| TOT_F | Total population Female |

| | |
|---|---|
| M_06 | Population in the age group 0-6 Male |
| F_06 | Population in the age group 0-6 Female |
| M_SC | Scheduled Castes population Male |
| F_SC | Scheduled Castes population Female |
| M_ST | Scheduled Tribes population Male |
| F_ST | Scheduled Tribes population Female |
| M_LIT | Literates population Male |
| F_LIT | Literates population Female |
| M_ILL | Illiterate Male |
| F_ILL | Illiterate Female |
| TOT_WORK_M | Total Worker Population Male |
| TOT_WORK_F | Total Worker Population Female |
| MAINWORK_M | Main Working Population Male |
| MAINWORK_F | Main Working Population Female |
| MAIN_CL_M | Main Cultivator Population Male |
| MAIN_CL_F | Main Cultivator Population Female |
| MAIN_AL_M | Main Agricultural Labourers Population Male |
| MAIN_AL_F | Main Agricultural Labourers Population Female |
| MAIN_HH_M | Main Household Industries Population Male |
| MAIN_HH_F | Main Household Industries Population Female |
| MAIN_OT_M | Main Other Workers Population Male |

| | |
|---|---|
| MAIN_OT_F | Main Other Workers Population Female |
| MARGWORK_M | Marginal Worker Population Male |
| MARGWORK_F | Marginal Worker Population Female |
| MARG_CL_M | Marginal Cultivator Population Male |
| MARG_CL_F | Marginal Cultivator Population Female |
| MARG_AL_M | Marginal Agriculture Labourers Population Male |
| MARG_AL_F | Marginal Agriculture Labourers Population Female |
| MARG_HH_M | Marginal Household Industries Population Male |
| MARG_HH_F | Marginal Household Industries Population Female |
| MARG_OT_M | Marginal Other Workers Population Male |
| MARG_OT_F | Marginal Other Workers Population Female |
| MARGWORK_3_6_M | Marginal Worker Population 3-6 Male |
| MARGWORK_3_6_F | Marginal Worker Population 3-6 Female |
| MARG_CL_3_6_M | Marginal Cultivator Population 3-6 Male |
| MARG_CL_3_6_F | Marginal Cultivator Population 3-6 Female |
| MARG_AL_3_6_M | Marginal Agriculture Labourers Population 3-6 Male |
| MARG_AL_3_6_F | Marginal Agriculture Labourers Population 3-6 Female |
| MARG_HH_3_6_M | Marginal Household Industries Population 3-6 Male |

| | |
|---|---|
| MARG_HH_3_6_F | Marginal Household Industries Population 3-6 Female |
| MARG_OT_3_6_M | Marginal Other Workers Population Person 3-6 Male |
| MARG_OT_3_6_F | Marginal Other Workers Population Person 3-6 Female |
| MARGWORK_0_3_M | Marginal Worker Population 0-3 Male |
| MARGWORK_0_3_F | Marginal Worker Population 0-3 Female |
| MARG_CL_0_3_M | Marginal Cultivator Population 0-3 Male |
| MARG_CL_0_3_F | Marginal Cultivator Population 0-3 Female |
| MARG_AL_0_3_M | Marginal Agriculture Labourers Population 0-3 Male |
| MARG_AL_0_3_F | Marginal Agriculture Labourers Population 0-3 Female |
| MARG_HH_0_3_M | Marginal Household Industries Population 0-3 Male |
| MARG_HH_0_3_F | Marginal Household Industries Population 0-3 Female |
| MARG_OT_0_3_M | Marginal Other Workers Population 0-3 Male |
| MARG_OT_0_3_F | Marginal Other Workers Population 0-3 Female |
| NON_WORK_M | Non Working Population Male |
| NON_WORK_F | Non Working Population Female |

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

**Answer:**

**Checking head & tail of data**

| | 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_3_F | MARG_HH_0_3_M | MARG_HH_0_3_F | MARG_OT_0_3_M | MARG_OT_0_3_F | NON_WORK_M | NON_WORK_F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | Jammu & Kashmir | Kupwara | 7707 | 23388 | 29796 | 5862 | 6196 | 3 | ... | 1150 | 749 | 180 | 237 | 680 | 252 | 32 | 46 | 258 | 214 |
| 1 | 1 | 2 | Jammu & Kashmir | Badgam | 6218 | 19585 | 23102 | 4482 | 3733 | 7 | ... | 525 | 715 | 123 | 229 | 186 | 148 | 76 | 178 | 140 | 160 |
| 2 | 1 | 3 | Jammu & Kashmir | Leh(Ladakh) | 4452 | 6546 | 10964 | 1082 | 1018 | 3 | ... | 114 | 188 | 44 | 89 | 3 | 34 | 0 | 4 | 67 | 61 |
| 3 | 1 | 4 | Jammu & Kashmir | Kargil | 1320 | 2784 | 4206 | 563 | 677 | 0 | ... | 194 | 247 | 61 | 128 | 13 | 50 | 4 | 10 | 116 | 59 |
| 4 | 1 | 5 | Jammu & Kashmir | Punch | 11654 | 20591 | 29981 | 5157 | 4587 | 20 | ... | 874 | 1928 | 465 | 1043 | 205 | 302 | 24 | 105 | 180 | 478 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 635 | 34 | 636 | Puducherry | Mahe | 3333 | 8154 | 11781 | 1146 | 1203 | 21 | ... | 32 | 47 | 0 | 0 | 0 | 0 | 0 | 0 | 32 | 47 |
| 636 | 34 | 637 | Puducherry | Karaikal | 10612 | 12346 | 21691 | 1544 | 1533 | 2234 | ... | 155 | 337 | 3 | 14 | 38 | 130 | 4 | 23 | 110 | 170 |
| 637 | 35 | 638 | Andaman & Nicobar Island | Nicobars | 1275 | 1549 | 2630 | 227 | 225 | 0 | ... | 104 | 134 | 9 | 4 | 2 | 6 | 17 | 47 | 76 | 77 |
| 638 | 35 | 639 | Andaman & Nicobar Island | North & Middle Andaman | 3762 | 5200 | 8012 | 723 | 664 | 0 | ... | 136 | 172 | 24 | 44 | 11 | 21 | 1 | 4 | 100 | 103 |
| 639 | 35 | 640 | Andaman & Nicobar Island | South Andaman | 7975 | 11977 | 18049 | 1470 | 1358 | 0 | ... | 173 | 122 | 6 | 2 | 17 | 17 | 2 | 4 | 148 | 99 |

640 rows × 61 columns

**Shape of the data**

- We have **640 rows** and **61 columns**

RangeIndex: 640 entries, 0 to 639
Data columns (total 61 columns):

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | State Code | 640 non-null | int64 |
| 1 | Dist.Code | 640 non-null | int64 |
| 2 | State | 640 non-null | object |
| 3 | Area Name | 640 non-null | object |
| 4 | No_HH | 640 non-null | int64 |
| 5 | TOT_M | 640 non-null | int64 |
| 6 | TOT_F | 640 non-null | int64 |
| 7 | M_06 | 640 non-null | int64 |
| 8 | F_06 | 640 non-null | int64 |
| 9 | M_SC | 640 non-null | int64 |
| 10 | F_SC | 640 non-null | int64 |
| 11 | M_ST | 640 non-null | int64 |
| 12 | F_ST | 640 non-null | int64 |
| 13 | M_LIT | 640 non-null | int64 |
| 14 | F_LIT | 640 non-null | int64 |
| 15 | M_ILL | 640 non-null | int64 |
| 16 | F_ILL | 640 non-null | int64 |
| 17 | TOT_WORK_M | 640 non-null | int64 |
| 18 | TOT_WORK_F | 640 non-null | int64 |
| 19 | MAINWORK_M | 640 non-null | int64 |
| 20 | MAINWORK_F | 640 non-null | int64 |
| 21 | MAIN_CL_M | 640 non-null | int64 |

```
 22  MAIN_CL_F      640 non-null    int64
 23  MAIN_AL_M      640 non-null    int64
 24  MAIN_AL_F      640 non-null    int64
 25  MAIN_HH_M      640 non-null    int64
 26  MAIN_HH_F      640 non-null    int64
 27  MAIN_OT_M      640 non-null    int64
 28  MAIN_OT_F      640 non-null    int64
 29  MARGWORK_M     640 non-null    int64
 30  MARGWORK_F     640 non-null    int64
 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
```

**Checking Null 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
```

- There are **no null values**

**Checking for duplicate values:**

- There are **no duplicate values** in the given dataset

**Summary of the data**

| index | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| State Code | 640 | 17.1140625 | 9.426486295 | 1 | 9 | 18 | 24 | 35 |
| Dist.Code | 640 | 320.5 | 184.8963674 | 1 | 160.75 | 320.5 | 480.25 | 640 |
| No_HH | 640 | 51222.87188 | 48135.40547 | 350 | 19484 | 35837 | 68892 | 310450 |
| TOT_M | 640 | 79940.57656 | 73384.51111 | 391 | 30228 | 58339 | 107918.5 | 485417 |
| TOT_F | 640 | 122372.0844 | 113600.7173 | 698 | 46517.75 | 87724.5 | 164251.75 | 750392 |
| M_06 | 640 | 12309.09844 | 11500.90688 | 56 | 4733.75 | 9159 | 16520.25 | 96223 |
| F_06 | 640 | 11942.3 | 11326.29457 | 56 | 4672.25 | 8663 | 15902.25 | 95129 |
| M_SC | 640 | 13820.94688 | 14426.37313 | 0 | 3466.25 | 9591.5 | 19429.75 | 103307 |
| F_SC | 640 | 20778.39219 | 21727.88771 | 0 | 5603.25 | 13709 | 29180 | 156429 |
| M_ST | 640 | 6191.807813 | 9912.668948 | 0 | 293.75 | 2333.5 | 7658 | 96785 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| F_ST | 640 | 10155.64063 | 15875.70149 | 0 | 429.5 | 3834.5 | 12480.25 | 130119 |
| M_LIT | 640 | 57967.97969 | 55910.28247 | 286 | 21298 | 42693.5 | 77989.5 | 403261 |
| F_LIT | 640 | 66359.56563 | 75037.86021 | 371 | 20932 | 43796.5 | 84799.75 | 571140 |
| M_ILL | 640 | 21972.59688 | 19825.60527 | 105 | 8590 | 15767.5 | 29512.5 | 105961 |
| F_ILL | 640 | 56012.51875 | 47116.69377 | 327 | 22367 | 42386 | 78471 | 254160 |
| TOT_WORK_M | 640 | 37992.40781 | 36419.53749 | 100 | 13753.5 | 27936.5 | 50226.75 | 269422 |
| TOT_WORK_F | 640 | 41295.76094 | 37192.36094 | 357 | 16097.75 | 30588.5 | 53234.25 | 257848 |
| MAINWORK_M | 640 | 30204.44688 | 31480.91568 | 65 | 9787 | 21250.5 | 40119 | 247911 |
| MAINWORK_F | 640 | 28198.84688 | 29998.26269 | 240 | 9502.25 | 18484 | 35063.25 | 226166 |
| MAIN_CL_M | 640 | 5424.342188 | 4739.161969 | 0 | 2023.5 | 4160.5 | 7695 | 29113 |
| MAIN_CL_F | 640 | 5486.042188 | 5326.362728 | 0 | 1920.25 | 3908.5 | 7286.25 | 36193 |
| MAIN_AL_M | 640 | 5849.109375 | 6399.507966 | 0 | 1070.25 | 3936.5 | 8067.25 | 40843 |
| MAIN_AL_F | 640 | 8925.995313 | 12864.28758 | 0 | 1408.75 | 3933.5 | 10617.5 | 87945 |
| MAIN_HH_M | 640 | 883.89375 | 1278.642345 | 0 | 187.5 | 498.5 | 1099.25 | 16429 |
| MAIN_HH_F | 640 | 1380.773438 | 3179.414449 | 0 | 248.75 | 540.5 | 1435.75 | 45979 |
| MAIN_OT_M | 640 | 18047.10156 | 26068.48089 | 36 | 3997.5 | 9598 | 21249.5 | 240855 |
| MAIN_OT_F | 640 | 12406.03594 | 18972.20237 | 153 | 3142.5 | 6380.5 | 14368.25 | 209355 |
| MARGWORK_M | 640 | 7787.960938 | 7410.791691 | 35 | 2937.5 | 5627 | 9800.25 | 47553 |
| MARGWORK_F | 640 | 13096.91406 | 10996.47453 | 117 | 5424.5 | 10175 | 18879.25 | 66915 |
| MARG_CL_M | 640 | 1040.7375 | 1311.546847 | 0 | 311.75 | 606.5 | 1281 | 13201 |
| MARG_CL_F | 640 | 2307.682813 | 3564.626095 | 0 | 630.25 | 1226 | 2659.25 | 44324 |
| MARG_AL_M | 640 | 3304.326563 | 3781.555707 | 0 | 873.5 | 2062 | 4300.75 | 23719 |
| MARG_AL_F | 640 | 6463.28125 | 6773.876298 | 0 | 1402.5 | 4020.5 | 9089.25 | 45301 |
| MARG_HH_M | 640 | 316.7421875 | 462.6618914 | 0 | 71.75 | 166 | 356.5 | 4298 |
| MARG_HH_F | 640 | 786.6265625 | 1198.718213 | 0 | 171.75 | 429 | 962.5 | 15448 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| MARG_OT_M | 640 | 3126.154688 | 3609.391821 | 7 | 935.5 | 2036 | 3985.25 | 24728 |
| MARG_OT_F | 640 | 3539.323438 | 4115.191314 | 19 | 1071.75 | 2349.5 | 4400.5 | 36377 |
| MARGWORK_3_6_M | 640 | 41948.16875 | 39045.31692 | 291 | 16208.25 | 30315 | 57218.75 | 300937 |
| MARGWORK_3_6_F | 640 | 81076.32344 | 82970.40622 | 341 | 26619.5 | 56793 | 107924 | 676450 |
| MARG_CL_3_6_M | 640 | 6394.9875 | 6019.806644 | 27 | 2372 | 4630 | 8167 | 39106 |
| MARG_CL_3_6_F | 640 | 10339.86406 | 8467.473429 | 85 | 4351.5 | 8295 | 15102 | 50065 |
| MARG_AL_3_6_M | 640 | 789.8484375 | 905.6392794 | 0 | 235.5 | 480.5 | 986 | 7426 |
| MARG_AL_3_6_F | 640 | 1749.584375 | 2496.541514 | 0 | 497.25 | 985.5 | 2059 | 27171 |
| MARG_HH_3_6_M | 640 | 2743.635938 | 3059.586387 | 0 | 718.75 | 1714.5 | 3702.25 | 19343 |
| MARG_HH_3_6_F | 640 | 5169.85 | 5335.64096 | 0 | 1113.75 | 3294 | 7502.25 | 36253 |
| MARG_OT_3_6_M | 640 | 245.3625 | 358.7285666 | 0 | 58 | 129.5 | 276 | 3535 |
| MARG_OT_3_6_F | 640 | 585.884375 | 900.0258173 | 0 | 127.75 | 320.5 | 719.25 | 12094 |
| MARGWORK_0_3_M | 640 | 2616.140625 | 3036.964381 | 7 | 755 | 1681.5 | 3320.25 | 20648 |
| MARGWORK_0_3_F | 640 | 2834.545313 | 3327.836932 | 14 | 833.5 | 1834.5 | 3610.5 | 25844 |
| MARG_CL_0_3_M | 640 | 1392.973438 | 1489.707052 | 4 | 489.5 | 949 | 1714 | 9875 |
| MARG_CL_0_3_F | 640 | 2757.05 | 2788.776676 | 30 | 957.25 | 1928 | 3599.75 | 21611 |
| MARG_AL_0_3_M | 640 | 250.8890625 | 453.336594 | 0 | 47 | 114.5 | 270.75 | 5775 |
| MARG_AL_0_3_F | 640 | 558.0984375 | 1117.642748 | 0 | 109 | 247.5 | 568.75 | 17153 |
| MARG_HH_0_3_M | 640 | 560.690625 | 762.5789913 | 0 | 136.5 | 308 | 642 | 6116 |
| MARG_HH_0_3_F | 640 | 1293.43125 | 1585.377936 | 0 | 298 | 717 | 1710.75 | 13714 |
| MARG_OT_0_3_M | 640 | 71.3796875 | 107.8976268 | 0 | 14 | 35 | 79 | 895 |
| MARG_OT_0_3_F | 640 | 200.7421875 | 309.740854 | 0 | 43 | 113 | 240 | 3354 |
| NON_WORK_M | 640 | 510.0140625 | 610.6031868 | 0 | 161 | 326 | 604.5 | 6456 |
| NON_WORK_F | 640 | 704.778125 | 910.209225 | 5 | 220.5 | 464.5 | 853.5 | 10533 |
| NON_WORK_F | 640 | 704.778125 | 910.209225 | 5 | 220.5 | 464.5 | 853.5 | 10533 |

- **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**

**Answer:** Performing the EDA..

- We have picked **No_HH, TOT_M, TOT_F, M_LIT, F_LIT** for this purpose.
- Below is the details of EDA performed:

**#Which state has highest Female gender ratio and which has the lowest?**(Not in terms of per 1000 population)

| State | Female_Gender_Ratio |
|---|---|
| Lakshadweep | 1.151993 |
| Haryana | 1.283484 |
| NCT of Delhi | 1.290194 |
| Uttar Pradesh | 1.329492 |
| Meghalaya | 1.329504 |

| State | Female_Gender_Ratio |
|---|---|
| Andhra Pradesh | 1.862113 |
| Tamil Nadu | 1.825079 |
| Chhattisgarh | 1.820831 |
| Arunachal Pradesh | 1.741054 |
| Odisha | 1.737621 |

- We can see that Andhra has the highest while Lakshadweep has the lowest

**# Which District code has highest Female gender ratio and which has the lowest?** (Not in terms of per 1000 population)

| Dist.Code | Female_Gender_Ratio |
|---|---|
| 587 | 1.151993 |
| 2 | 1.179576 |
| 144 | 1.180202 |
| 106 | 1.180761 |
| 139 | 1.184830 |

| Dist.Code | Female_Gender_Ratio |
|---|---|
| 547 | 2.283250 |
| 398 | 2.268763 |
| 625 | 2.225429 |
| 546 | 2.221849 |
| 391 | 2.215060 |

- We can see that the Dist. 587 has the lowest while Dist code 547 has the highest

**#. Which state has highest No of Literate Male and which has the lowest? (**Based on the absolute nos, not ratio)

| State | |
|---|---|
| Dadara & Nagar Havelli | 5119 |
| Lakshadweep | 10601 |
| Daman & Diu | 10880 |
| Andaman & Nicobar Island | 15488 |
| Sikkim | 21230 |

| State | |
|---|---|
| Uttar Pradesh | 6016402 |
| Maharashtra | 3308633 |
| West Bengal | 2932621 |
| Karnataka | 2554163 |
| Tamil Nadu | 2485404 |

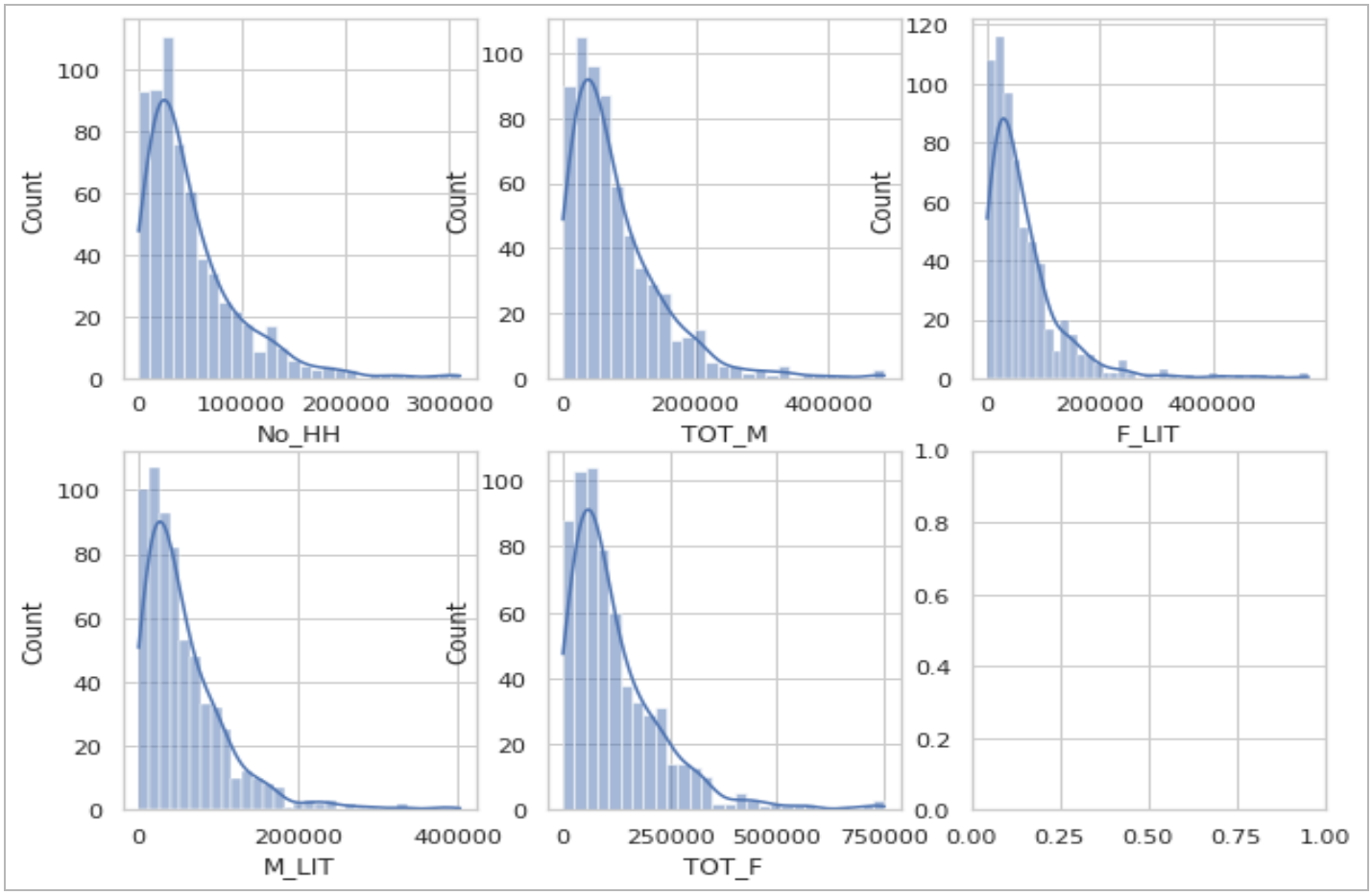
**#. Which state has highest No of Literate Female and which has the lowest? (**Based on the absolute nos, not ratio)

| State | |
|---|---|
| Dadara & Nagar Havelli | 5308 |
| Lakshadweep | 11334 |
| Daman & Diu | 12520 |
| Andaman & Nicobar Island | 20237 |
| Sikkim | 27112 |

| State | |
|---|---|
| Uttar Pradesh | 5574752 |
| Maharashtra | 4619012 |
| Kerala | 3878204 |
| West Bengal | 3479316 |
| Tamil Nadu | 3205093 |

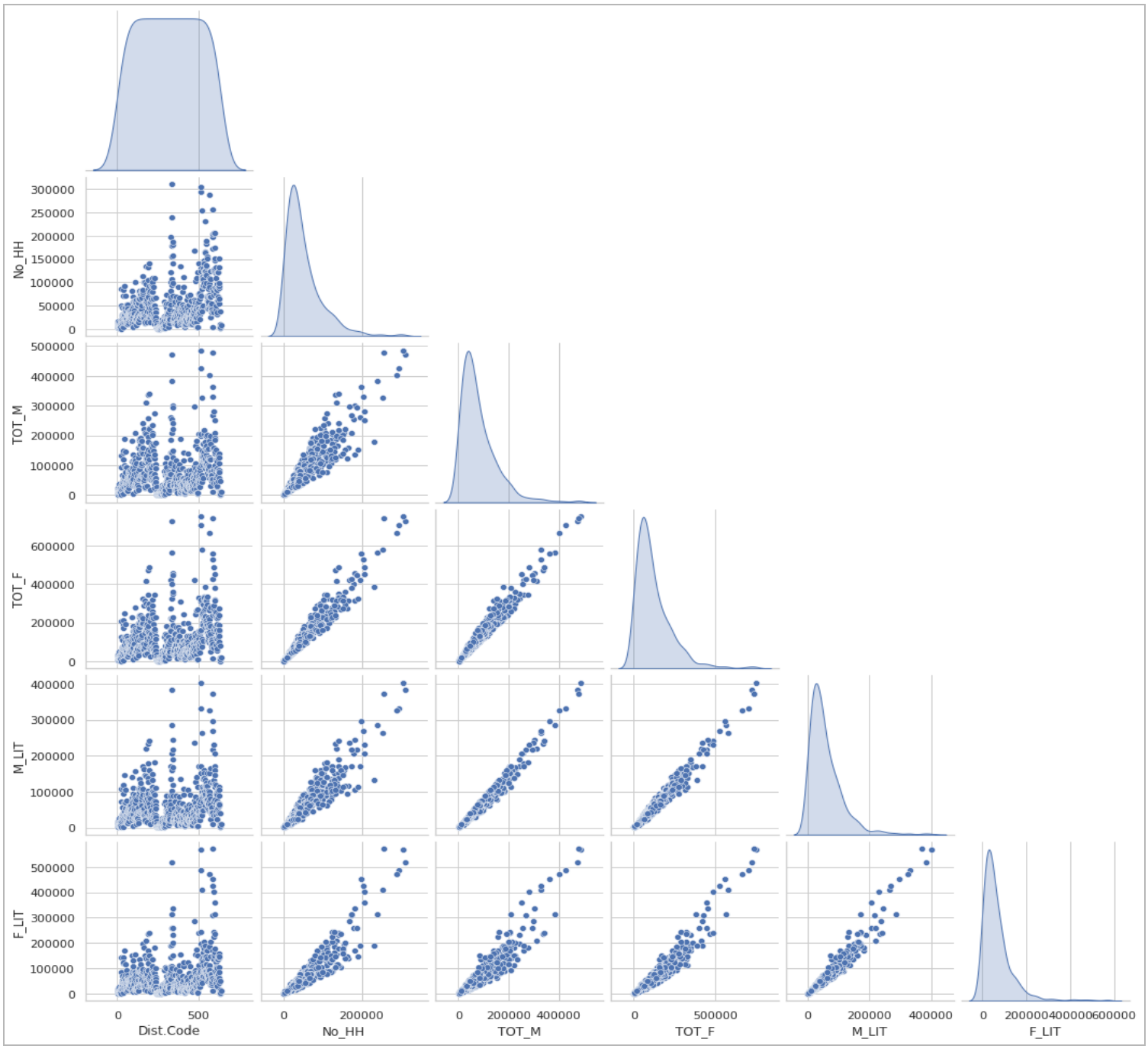
- From Above two, we can see that Dadara & Nagar Havelli has the lowest literate while Uttar Pradesh has the highest.
- However, the population of Uttar pradesh being higher than other states & Dadara & Nagar Haweli being low would not give insightful data, we can look basis the population ratio to further analyze it)

**Plotting the Histogram of these 5 variables selected**:.

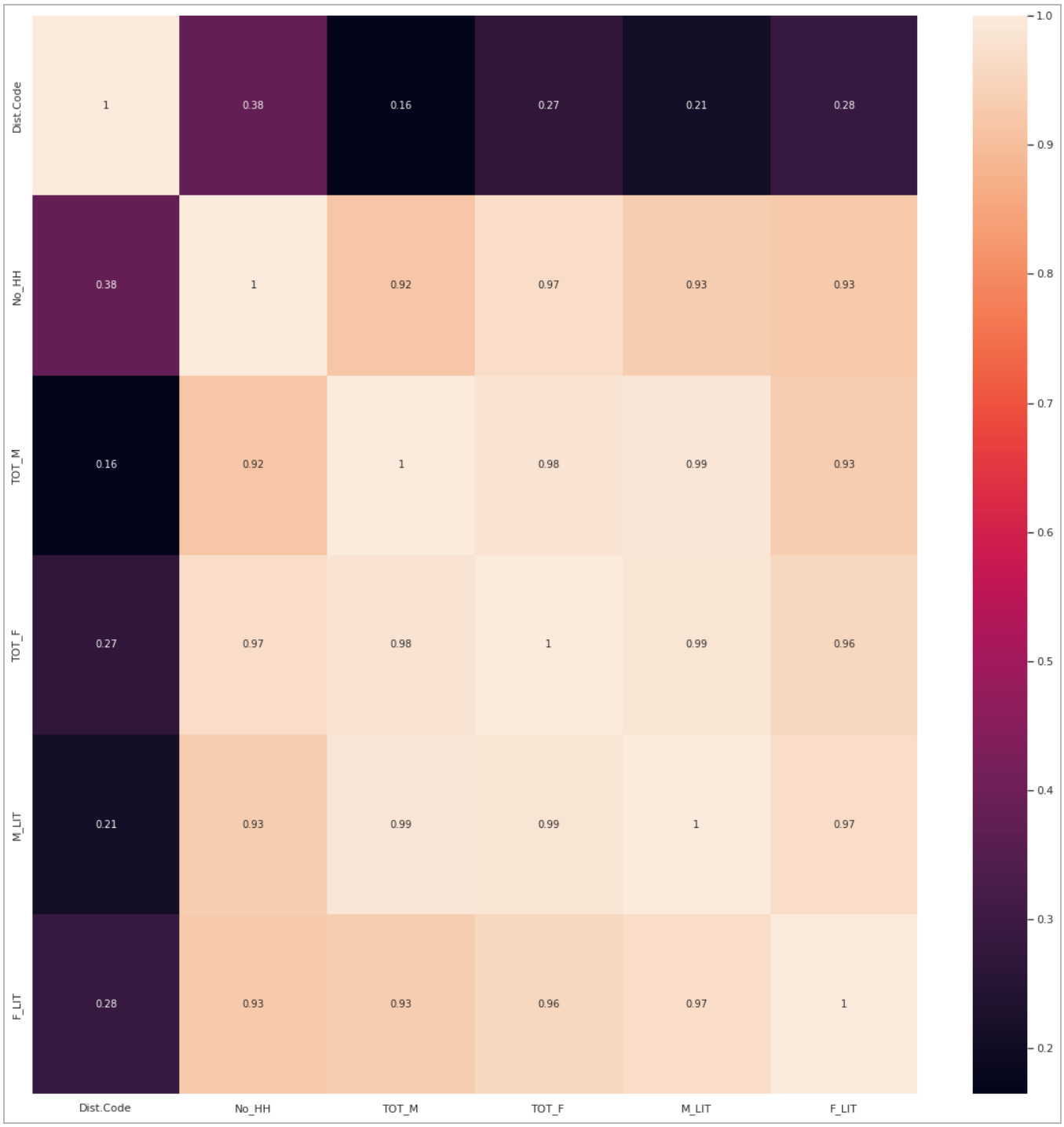



- Since it is population data, tends to be normally distributed

**Plotting Papilot**




**Correlation Matrix**




- From Paiplot & above correlation, we see a **positive correlation** between the selected variables

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

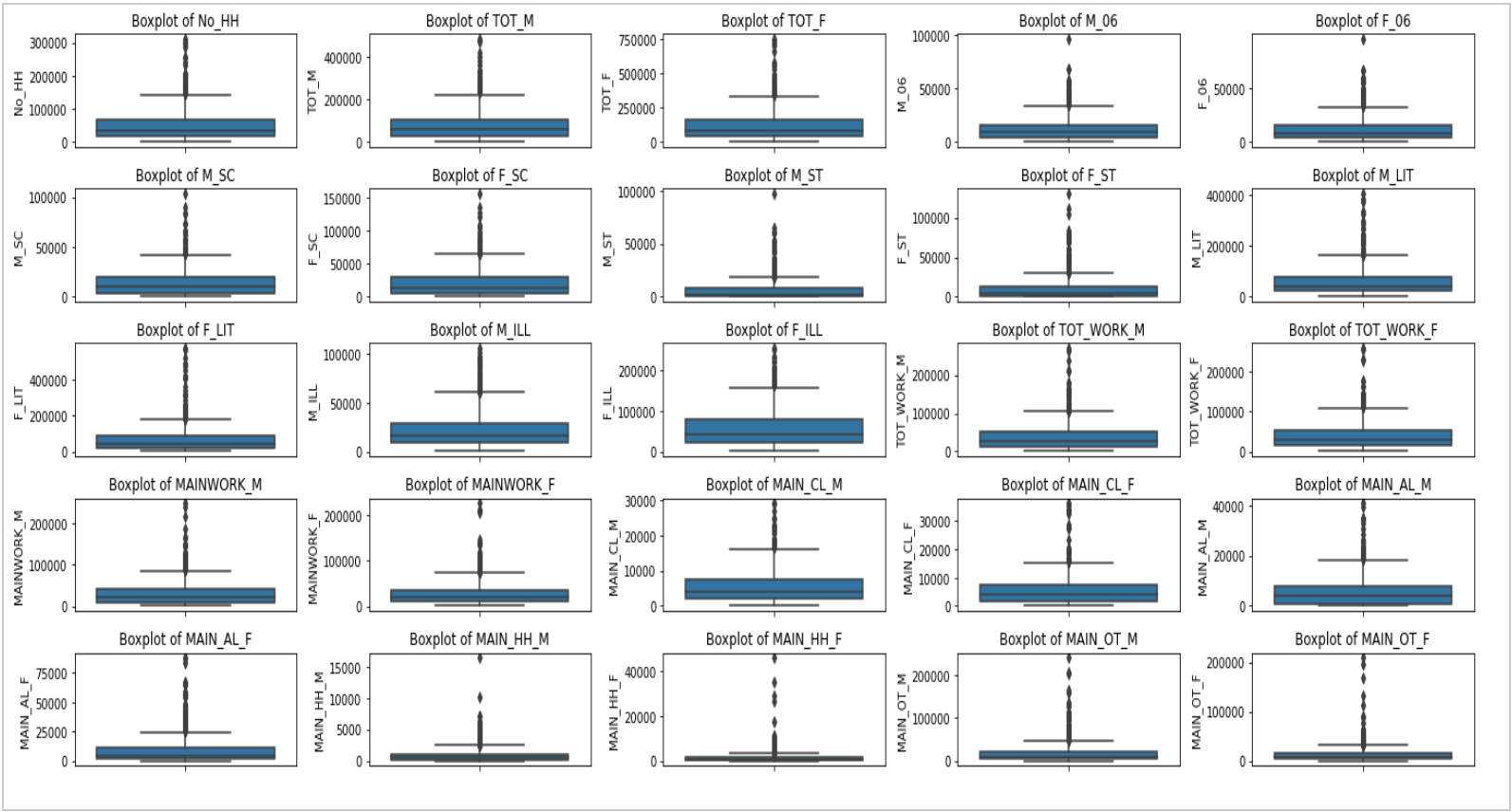
**Answer:** In the case of census data, outlier treatment may not be necessary for several reasons:

1. The data is usually collected from a large and representative sample of the population. This means that the data is likely to be normally distributed, and outliers are less likely to occur
2. Census data is often collected using standardized methods and questionnaires, which reduce the likelihood of errors and outliers
3. The purpose of census data is often to provide an accurate representation of the population as a whole. Outliers, by definition, are not representative of the population and may not provide any useful information
4. Outliers may also be due to errors or anomalies in the data collection process. In the case of census data, the data collection process is typically rigorous and standardized, making it less likely that errors will occur.

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

**Answer:**

**Data before scaling**

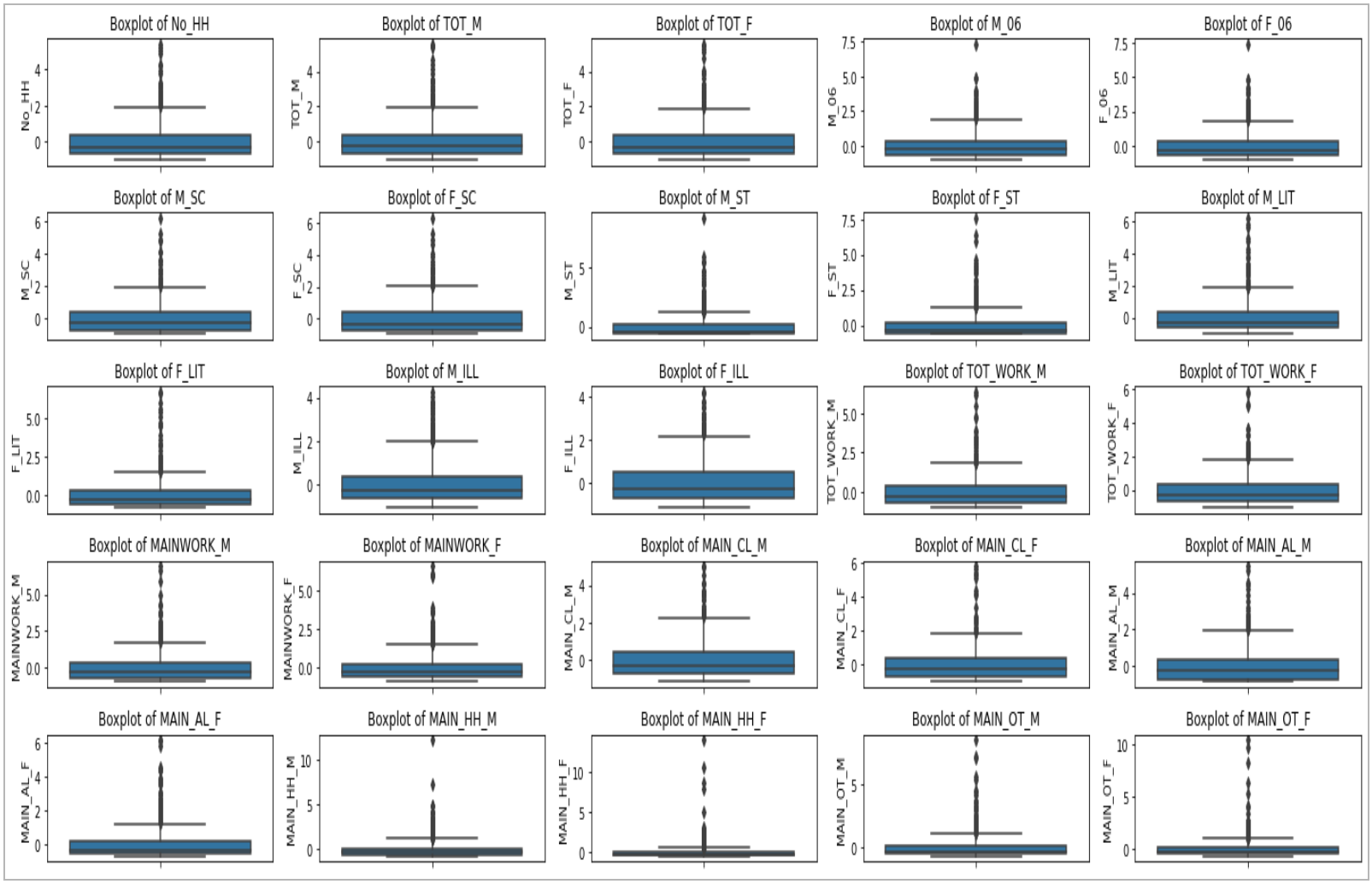



**Scaled data - head & tail**

| | No_HH | TOT_M | TOT_F | M_06 | F_06 | M_SC | F_SC | M_ST | F_ST | M_LIT | ... | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | -0.904738 | -0.771236 | -0.815563 | -0.561012 | -0.507738 | -0.958575 | -0.957049 | -0.423306 | -0.476423 | -0.798097 | ... | -0.872367 | -0.898216 | -1.042844 | -0.986630 | -0.851060 | -0.683276 | -0.630766 | -0.407555 | -0.624042 | -0.611637 |
| 1 | -0.935695 | -0.823100 | -0.874534 | -0.681096 | -0.725367 | -0.958297 | -0.956772 | -0.582014 | -0.607607 | -0.849434 | ... | -0.813078 | -0.882936 | -0.913606 | -0.963707 | -0.845587 | -0.685999 | -0.270728 | -0.326659 | -0.594942 | -0.605624 |
| 2 | -0.972412 | -1.000919 | -0.981466 | -0.976956 | -0.965262 | -0.958575 | -0.956772 | -0.038951 | -0.027273 | -0.956457 | ... | -0.898530 | -0.843236 | -1.035875 | -0.804375 | -0.909079 | -0.688878 | -0.676945 | -0.417313 | -0.640396 | -0.570440 |
| 3 | -1.037530 | -1.052224 | -1.041001 | -1.022118 | -0.995393 | -0.958783 | -0.957049 | -0.355965 | -0.390060 | -1.004643 | ... | -0.944594 | -0.927140 | -1.138083 | -1.011620 | -0.913457 | -0.692534 | -0.684772 | -0.432737 | -0.675984 | -0.639858 |
| 4 | -0.822676 | -0.809381 | -0.813933 | -0.622359 | -0.649908 | -0.957395 | -0.955529 | 0.149238 | 0.043330 | -0.800568 | ... | -0.879997 | -0.865121 | -0.988572 | -0.886859 | -0.874987 | -0.675964 | -0.664422 | -0.414480 | -0.635597 | -0.591118 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 635 | -0.995677 | -0.978990 | -0.974268 | -0.971387 | -0.948916 | -0.957326 | -0.955667 | -0.625124 | -0.640197 | -0.913820 | ... | -0.850241 | -0.907490 | -1.143784 | -1.030221 | -0.910486 | -0.694012 | -0.679293 | -0.433681 | -0.562003 | -0.602406 |
| 636 | -0.844340 | -0.921822 | -0.886965 | -0.936754 | -0.919757 | -0.803806 | -0.765670 | -0.625124 | -0.640197 | -0.853390 | ... | -0.781574 | -0.803836 | -1.117599 | -1.010492 | -0.813841 | -0.624152 | -0.672249 | -0.380170 | -0.507910 | -0.500388 |
| 637 | -1.038465 | -1.069066 | -1.054885 | -1.051356 | -1.035331 | -0.958783 | -0.957049 | -0.522953 | -0.529880 | -1.016367 | ... | -0.949871 | -0.922936 | -1.143784 | -1.029282 | -0.914552 | -0.694323 | -0.679293 | -0.422664 | -0.681320 | -0.628728 |
| 638 | -0.986758 | -1.019276 | -1.007472 | -1.008195 | -0.996541 | -0.958783 | -0.957049 | -0.622297 | -0.637046 | -0.962328 | ... | -0.893507 | -0.885171 | -1.028695 | -0.988509 | -0.904856 | -0.692223 | -0.685554 | -0.432422 | -0.636250 | -0.580251 |
| 639 | -0.899166 | -0.926854 | -0.919050 | -0.943193 | -0.935220 | -0.958783 | -0.957049 | -0.608870 | -0.623555 | -0.856916 | ... | -0.789617 | -0.790158 | -1.091625 | -0.984563 | -0.900947 | -0.689189 | -0.662856 | -0.422349 | -0.501422 | -0.434872 |

640 rows × 25 columns

**Outliers after Scaling**

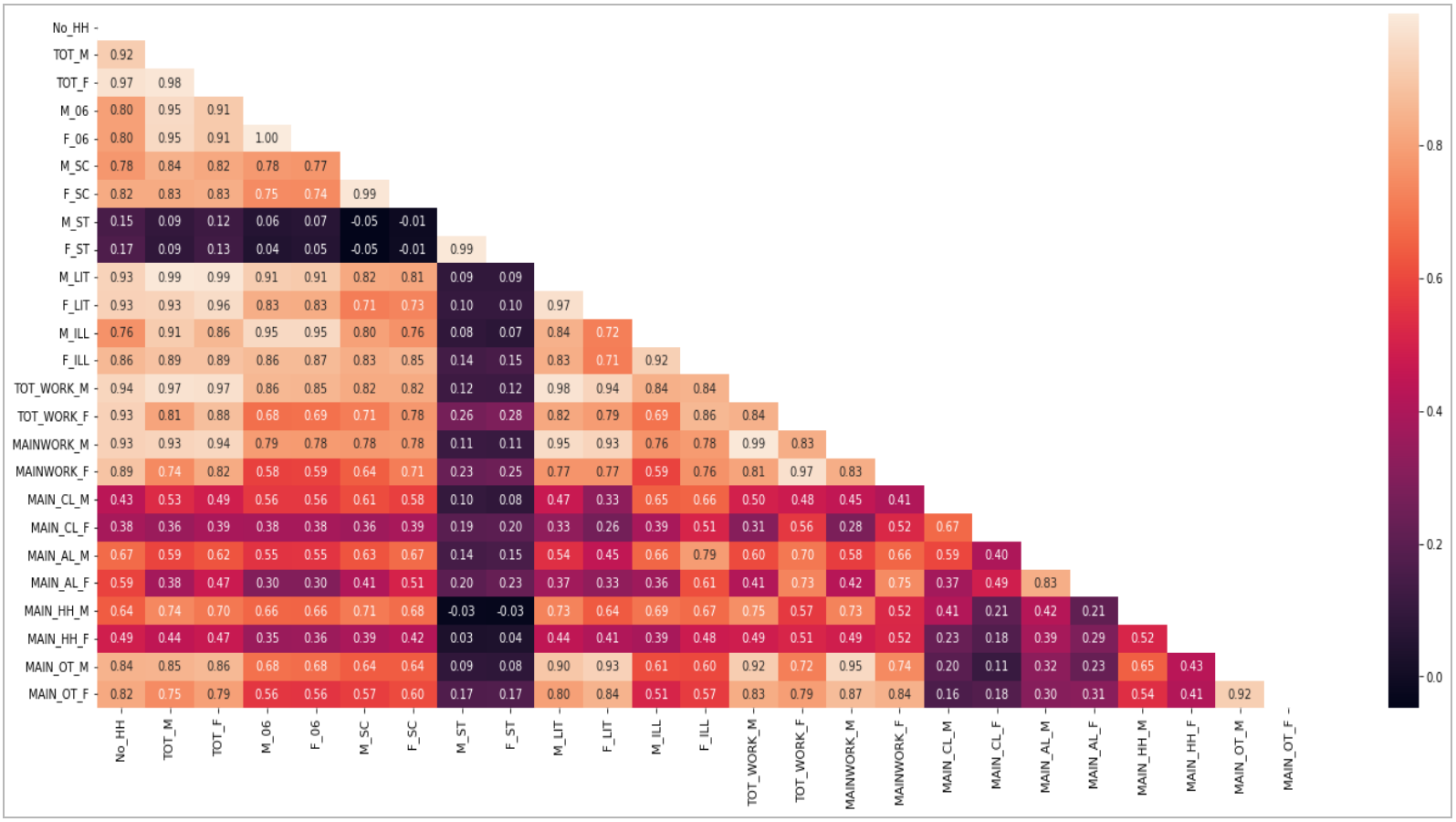



Thus we can see from above that scaling has **no significant impact on Outliers**

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

**Answer:**

Checking for Correlation




**PCA Components**

```
array([[ 2.38709793e-01,  2.41900082e-01,  2.44573206e-01,
         2.20937829e-01,  2.20547952e-01,  2.14817474e-01,
         2.17988628e-01,  3.51272864e-02,  3.56584288e-02,
         2.40051149e-01,  2.25970766e-01,  2.18424186e-01,
         2.29798146e-01,  2.41797250e-01,  2.25614542e-01,
         2.35359144e-01,  2.13711260e-01,  1.42611235e-01,
         1.10946487e-01,  1.72790685e-01,  1.34034463e-01,
         1.84217020e-01,  1.28232746e-01,  2.06845216e-01,
         1.94392948e-01],
        [ 1.62290697e-02, -9.40391173e-02, -4.63346143e-02,
        -1.00523300e-01, -9.33814815e-02, -8.22317092e-02,
        -3.24130673e-02,  4.78461556e-01,  4.88306682e-01,
        -1.09986449e-01, -1.25864235e-01, -3.79126499e-02,
         8.87357981e-02, -7.90420705e-02,  1.77987002e-01,
        -8.16965531e-02,  1.67367642e-01,  1.70396692e-01,
         3.19063353e-01,  2.40585193e-01,  3.63978211e-01,
        -1.61006580e-01, -3.99156365e-03, -1.80799828e-01,
        -7.10691453e-02],
        [ 8.76647456e-02,  1.33624380e-02,  5.32336244e-02,
        -9.64525135e-02, -9.52827253e-02, -1.76882764e-01,
        -1.55443948e-01,  3.49090415e-01,  3.50657157e-01,
         7.76706996e-02,  1.87476114e-01, -1.69578418e-01,
        -1.70224774e-01,  8.95299212e-02,  5.57167231e-02,
         1.38074688e-01,  1.09949099e-01, -3.94896798e-01,
        -2.69405829e-01, -2.56907430e-01, -1.48355950e-01,
        -3.24970967e-02,  3.03101749e-02,  3.03194913e-01,
         3.44997380e-01],
        [-1.18008259e-01,  1.32280714e-01,  3.28898712e-02,
         2.84972993e-01,  2.88517949e-01,  2.67974025e-02,
        -6.18350213e-02,  3.35599493e-01,  3.06504948e-01,
         8.11420277e-02,  2.01677716e-02,  2.60808272e-01,
         4.71800194e-02,  2.30256537e-02, -2.01428618e-01,
        -5.19484729e-02, -3.16826986e-01,  2.19112149e-01,
         2.67476427e-02, -1.81596063e-01, -4.20676847e-01,
         5.40157997e-02, -2.71384805e-01, -6.06377832e-02,
        -1.77742891e-01],
        [-8.76388277e-02, -3.11677026e-02, -7.51896600e-02,
        -4.92306785e-02, -4.27193129e-02,  3.41962968e-02,
         1.04259772e-02,  1.39877719e-01,  1.32026114e-01,
        -6.95207195e-02, -1.57693833e-01,  8.06884038e-02,
         6.98565250e-02,  5.75280609e-03, -7.35201847e-02,
        -2.06298288e-02, -1.01977154e-01,  1.86011693e-02,
```

-2.48559509e-01,  9.36149711e-02, -1.21286896e-01,
4.09028148e-01,  7.79900106e-01, -7.13386533e-02,
-1.39919207e-01],
[-6.21462065e-02, -3.14874634e-02, -2.94719538e-02,
-6.19096038e-02, -6.51777783e-02, -7.91540731e-02,
-1.00183717e-01, -6.28134009e-02, -7.58177727e-02,
6.83852637e-03,  2.65050831e-02, -1.35836259e-01,
-1.13270253e-01, -1.73816018e-02,  8.83568983e-02,
1.97344204e-02,  1.31880849e-01,  2.88436132e-01,
6.61810749e-01, -4.37726183e-01, -2.76204686e-01,
9.38716756e-02,  2.67565417e-01,  7.42473635e-02,
1.65169406e-01],
[ 6.79659636e-02,  8.71613874e-02,  1.12553106e-01,
2.84862949e-01,  2.99099279e-01, -4.76369673e-01,
-4.50805495e-01, -1.04477565e-01, -9.75274163e-02,
5.43394286e-02,  1.15439514e-01,  1.69385144e-01,
8.75226821e-02, -6.46786018e-02,  3.62445204e-02,
-9.11874930e-02,  1.19371039e-02, -2.32930642e-01,
1.01629745e-01,  5.57412156e-02,  9.57699129e-02,
-3.11567095e-01,  3.05301846e-01, -6.61758244e-02,
-1.25758584e-01],
[-6.43699653e-02,  1.01066088e-02, -5.18361136e-02,
-9.98996559e-02, -9.46146237e-02, -3.94286838e-01,
-4.48552947e-01, -3.81513852e-02, -6.27952256e-02,
-5.49152595e-03, -2.86759884e-02,  5.28963074e-02,
-7.93102100e-02,  1.52599101e-01, -9.96060743e-02,
2.49550664e-01,  2.84960681e-02,  3.18076692e-01,
-6.16041808e-02,  2.38725403e-01,  1.31861508e-01,
4.83130764e-01, -2.58489458e-01,  1.61236472e-01,
1.62606062e-02],
[-7.64918910e-03, -6.93858859e-03,  1.93301052e-02,
7.47134132e-02,  7.87682167e-02, -8.84910084e-03,
2.44650725e-02, -1.62273992e-02,  3.38962757e-02,
-2.54220667e-02, -5.98994561e-02,  4.60096922e-02,
1.42001492e-01, -1.34253603e-01,  1.33318720e-01,
-2.06688021e-01,  4.03682070e-02, -5.54065911e-01,
2.49696078e-01, -1.87622687e-01,  1.54250464e-01,
6.15544504e-01, -2.20450555e-01, -1.33006924e-01,
-7.39194336e-02],
[-1.42143998e-01, -3.87168511e-02, -1.17650186e-01,
-2.25354026e-02,  5.63856609e-03, -2.70203331e-02,
-1.02647469e-01, -4.65885208e-02, -4.45930049e-02,

-2.03074693e-01, -4.07601638e-01, 4.29381405e-01,
3.65484245e-01, -3.16676979e-02, 2.01631931e-01,
-6.67349980e-02, 2.09294334e-01, 5.41535912e-02,
-1.82878223e-01, -1.19148943e-01, -1.44185464e-01,
-1.00270489e-01, -7.61942644e-02, -5.62678274e-02,
4.92807235e-01],
[ 4.68119251e-01, -4.94496518e-02, 2.24822988e-01,
-9.08884876e-02, -5.78434692e-02, -2.28867926e-01,
1.47541079e-02, -1.05820175e-01, 6.43370801e-02,
1.25114071e-02, 2.10331628e-01, -2.18321447e-01,
2.07086206e-01, -9.54910810e-02, 2.39639785e-01,
-2.71175782e-01, -7.62154731e-02, 3.22587536e-01,
-2.81843303e-01, -1.63568570e-01, -1.09114861e-01,
1.29432985e-01, -5.23794311e-02, -3.52318061e-01,
4.13810025e-02],
[ 3.26275798e-01, -3.07176442e-02, 5.99477032e-02,
-2.67718922e-01, -2.53379596e-01, -6.97256240e-02,
2.06439000e-04, -6.57225175e-02, 3.78233329e-02,
-8.25836103e-02, -6.25981241e-02, 1.19193016e-01,
2.44230451e-01, 1.69015580e-01, 1.87582990e-02,
9.93626515e-02, -2.17463173e-01, -2.70243438e-01,
3.01694672e-01, 3.88808500e-01, -4.84288957e-01,
-7.11279322e-02, -5.20654304e-02, 7.71628689e-02,
-9.14433977e-02]])

**eigen values**

array([16.21075146, 2.42307151, 2.005783 , 1.3444504 , 0.82011656,
0.72051381, 0.50566632, 0.30728878, 0.30151323, 0.16697367,
0.07479658, 0.06118332])

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

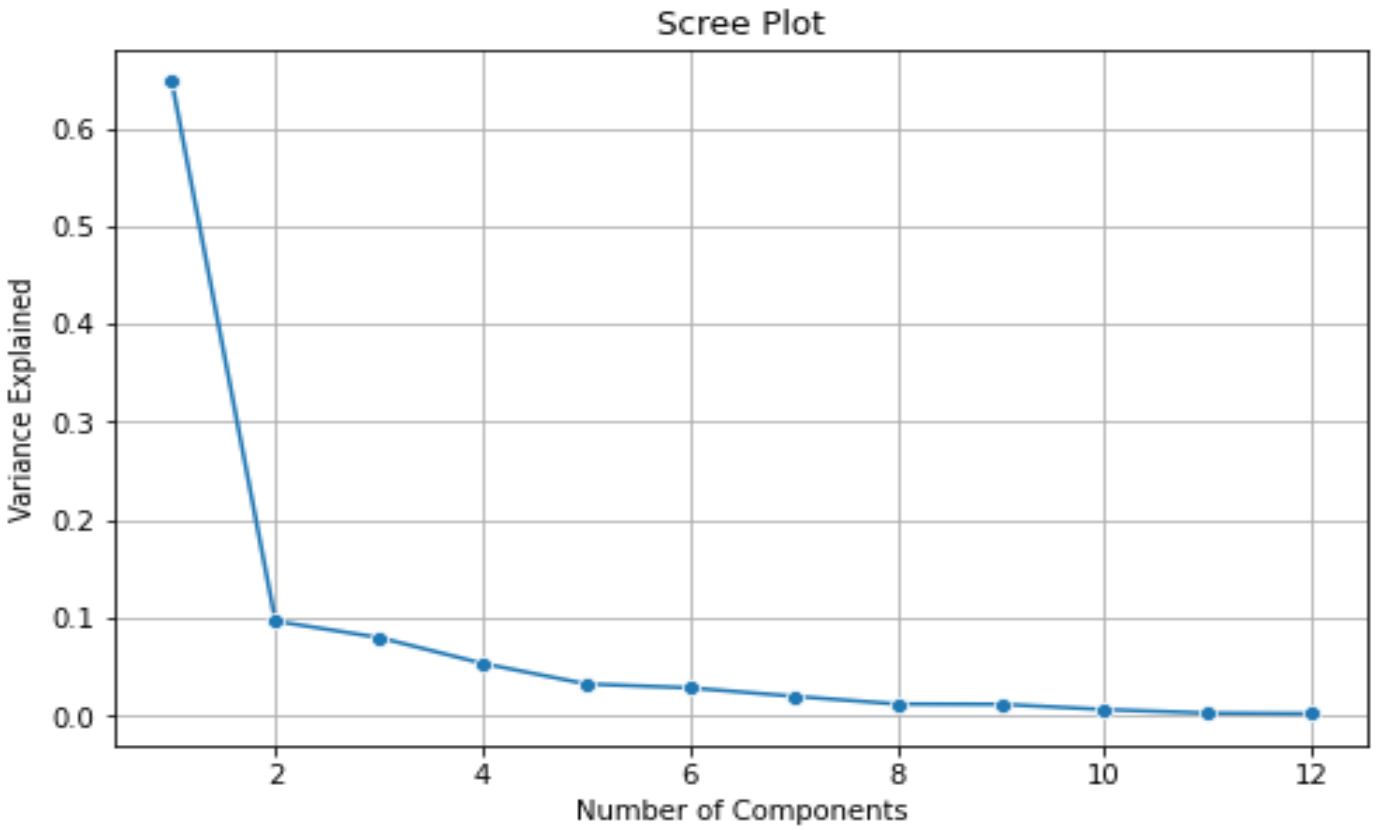
**Answer:**

Explained variance for each PC is as shown below:

array([0.64741689, 0.09677142, 0.08010596, 0.05369399, 0.03275341,
0.02877552, 0.02019505, 0.01227235, 0.01204168, 0.00666851,
0.00298719, 0.00244351])

**From the above, we can see that Optimum number of PCs for 90 % variance is 5.**

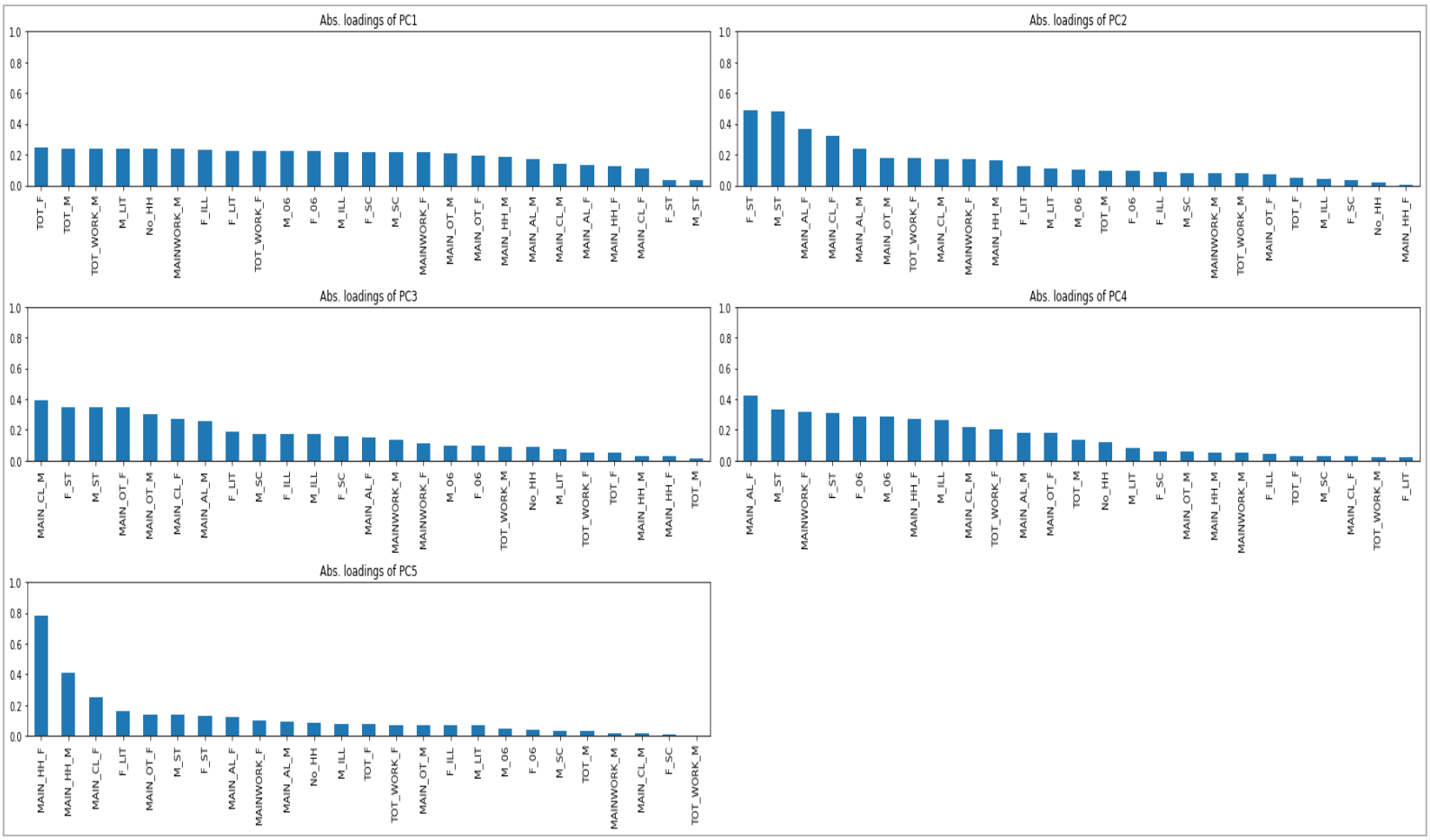
**Scree plot**



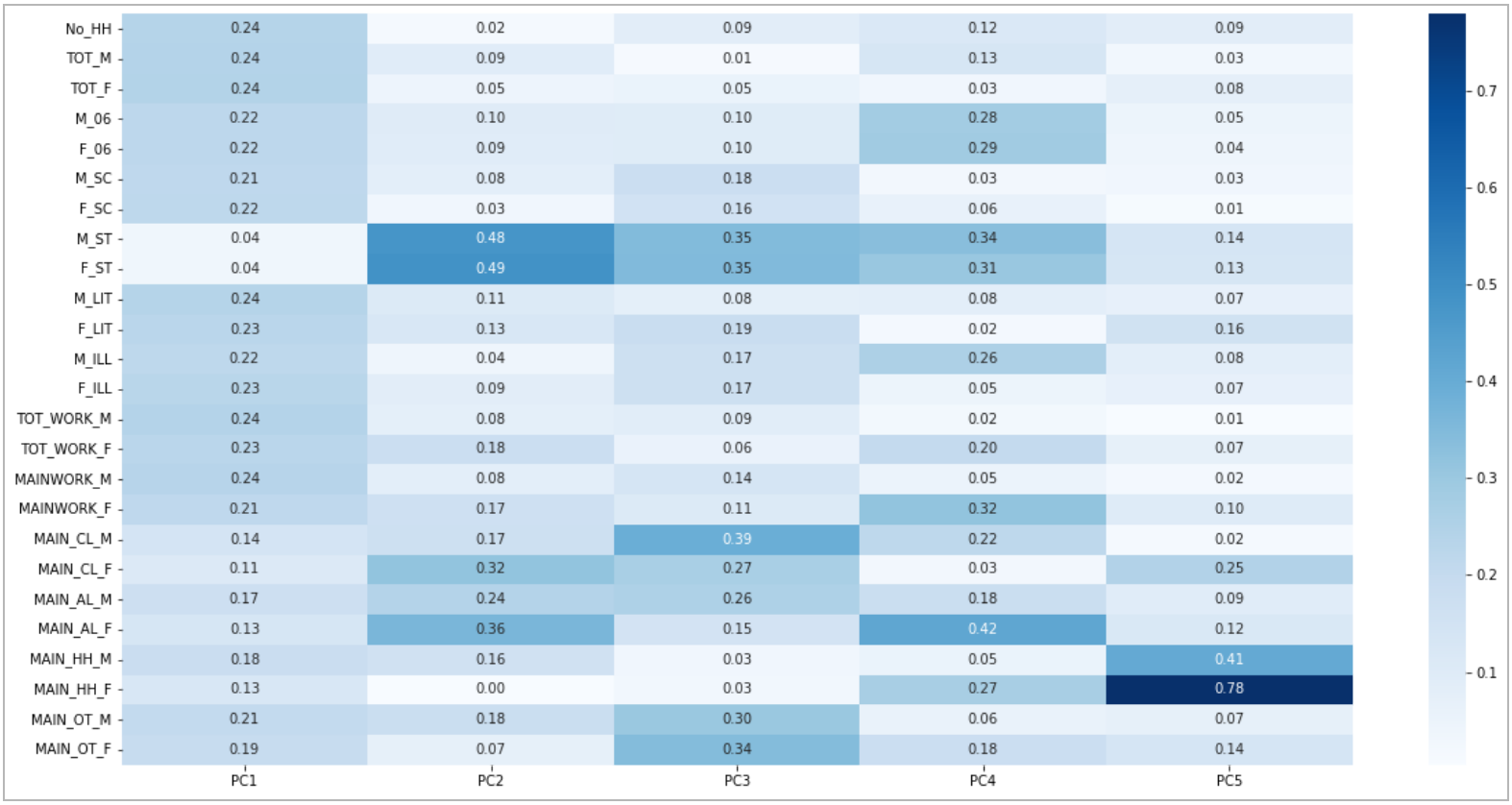

- **Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.**

**Answer:**

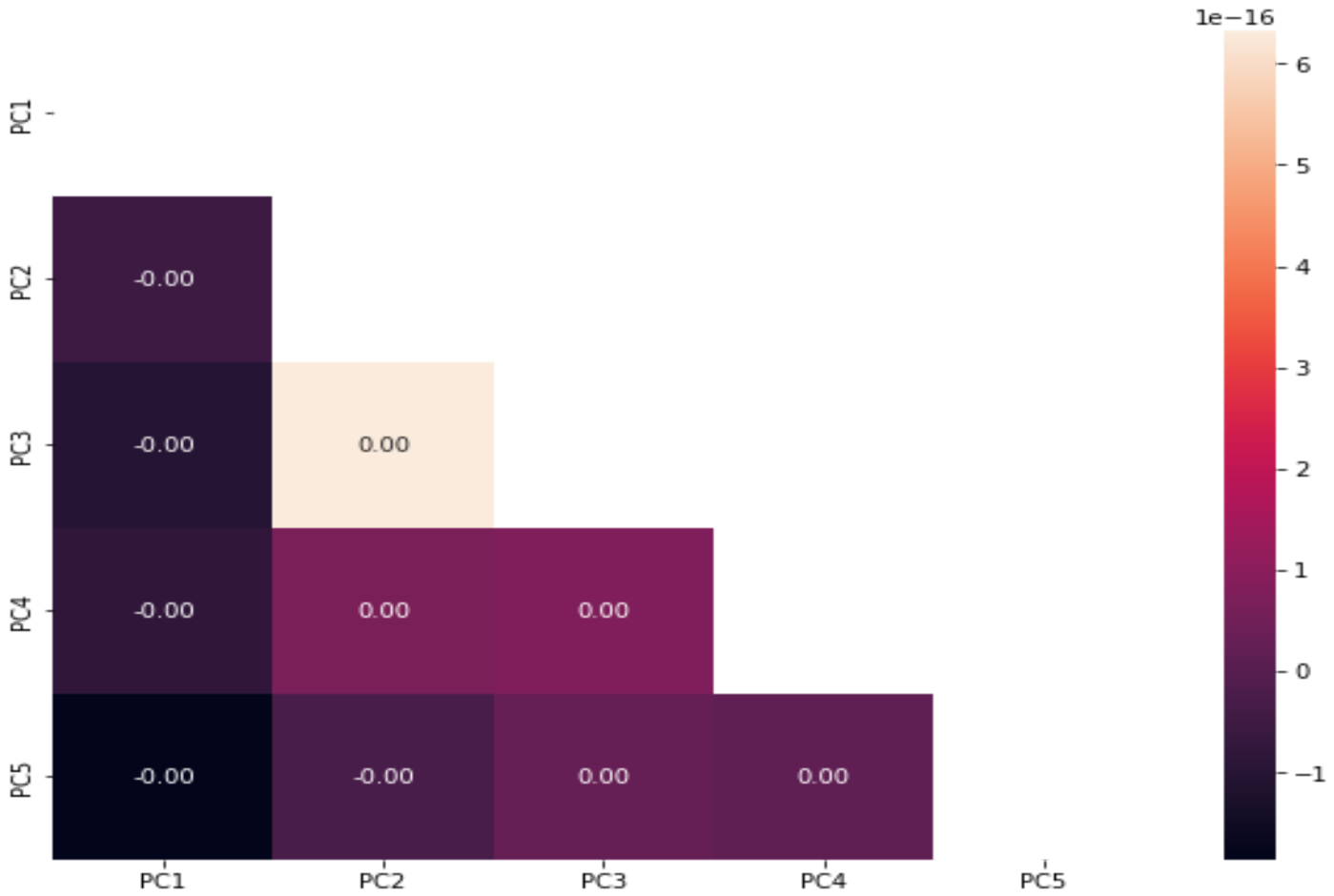
Below graphs shows how the original features matter to each PC




**Comparison how original features influence various PCs**




**Presence of correlations among the PCs**




The above heatmap indicates that the PC's are not correlated with each other which should be the ideal case

- **Write linear equation for first PC.**

The Linear equation for first PC is:

**PC1 = 0.246 * x1 + 0.381 * x2 + 0.599 * x3 + 0.055 * x4 + 0.054 * x5 + 0.063 * x6 + 0.096 * x7 + 0.007 * x8 + 0.011 * x9 + 0.292 * x10 + 0.380 * x11 + 0.089 * x12 + 0.218 * x13 + 0.188 * x14 + 0.172 * x15 + 0.159 * x16 + 0.132 * x17 + 0.012 * x18 + 0.011 * x19 + 0.021 * x20 + 0.032 * x21 + 0.005 * x22 + 0.008 * x23 + 0.121 * x24 + 0.081 * x25**