

|  |
| --- |
| MACHINE LEARNING - 1  2018 |
|  |
| PGP DSBA PROGRAM  by: ABHISHEK K HIREMATH |



|  |  |  |
| --- | --- | --- |
| S NO | Clustering and Cleaning Ads | Page No. |
| 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. | *3 - 4* |
| 1.2 | Treat missing values in CPC, CTR and CPM using the formula given. | *5* |
| 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. (As an analyst your judgement may be different from another analyst). | *5* |
| 1.4 | Perform z-score scaling and discuss how it affects the speed of the algorithm. | *6* |
| 1.5 | Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance. | *7* |
| 1.6 | Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm. | *7* |
| 1.7 | Print silhouette scores for up to 10 clusters and identify optimum number of clusters. | *8* |
| 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]. | *8 – 9* |
| 1.9 | Conclude the project by providing summary of your learnings | *9* |
| S NO | PCA | Page No. |
| 2.1 | Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc. | *10 – 11* |
| 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 | *12* |
| 2.3 | We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary? | *13 – 14* |
| 2.4 | Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment. | *15* |
| 2.5 | Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector. | *15* |
| 2.6 | Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot. | *16* |
| 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. | *16* |
| 2.8 | Write linear equation for first PC. | *17* |

|  |  |  |
| --- | --- | --- |
| S No | Figures and Tables | Page No. |
| 1 | Fig-1: Dataset head | 7 |
| 2 | Fig-2: Dataset tail | 7 |
| 3 | Fig-3: Data info | 7 |
| 4 | Fig-4: Null values of Dataset | 8 |
| 5 | Fig-5: Null values after treating | 9 |
| 6 | Fig-7: Data describe | 9 |
| 8 | Fig-8: Boxplot before Treating outliers | 10 |
| 9 | Fig-9: Boxplot After treating outlier | 11 |
| 10 | Fig-10: After applying Z-score | 11 |
| 11 | Fig-11: Data After scaling with Z-score | 12 |
| 12 | Fig-12: Dendrogram using WARD and Euclidean distance | 13 |
| 13 | Fig-13: Value count | 13 |
| 14 | Fig-14: Data sample | 14 |
| 15 | Fig-15: Elbow plot for k-means algorithm | 14 |
| 16 | Fig-16: | 15 |
| 17 | Fig-17: Cluster k means | 15 |
| 18 | Fig-18: Bar plot Comparisons | 16 |
| 19 | Fig-19: Mean Clicks based on Device Type | 17 |
| 20 | Fig-20: Mean Spend based on Device Type | 17 |
| 21 | Fig-21: Mean revenue based on Device Type | 18 |
| 22 | **Fig-22: Mean CPM (Cost per 1000 impressions) based on Device type** | 18 |
| 23 | **Fig-23: Mean CTR (Click through Rate) based on Device type** | 19 |
| 24 | **Fig-24: Mean CPC (Cost per Click) based on Device type** | 19 |
| 25 | **Fig-25: Head of the Dataset** | 23 |
| 26 | **Fig-26: Tail of the dataset** | 23 |
| 27 | **Fig-27: Data information** | 24 |
| 28 | **Fig-28: Data Describe** | 24 |
| 29 | **Fig-29: Null values** | 25 |
| 30 | **Fig-30: Chosen Variables** | 25 |
| 31 | **Fig-31: Calculated gender Ratio** | 26 |
| 32 | **Fig-32: Gender Ratio** | 26 |
| 33 | **Fig-33: District wise Gender Ratio** | 27 |
| 34 | **Fig-34: Boxplot before scaling** | 28 |
| 35 | **Fig-35: Dataset after applying Z-score** | 28 |
| 36 | **Fig-36: Boxplot After Scaling** | 29 |
| 37 | **Fig-37: Heatmap** | 30 |
| 38 | **Fig-38: Covariance matrix** | 31 |
| 39 | **Fig-39: Eigen value** | 31 |
| 40 | **Fig-40: Eigen vector** | 32 |
| 41 | **Fig-41: Explained variance** | 32 |
| 42 | **Fig-42: Cumulative Explained variance** | 33 |
| 43 | **Fig-43: Selected components** | 33 |
| 44 | Fig-44: Heatmap | 34 |
| 45 | Fig-45: Comparision | 34 |
| 1 | Table – 1 | 5-7 |
| 2 | Table - 2 | 22-24 |

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.

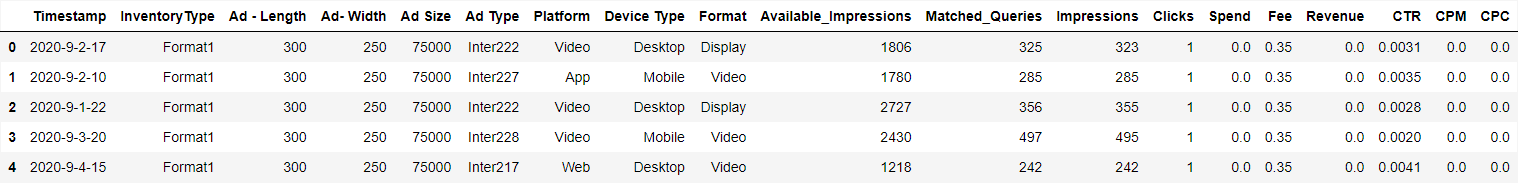
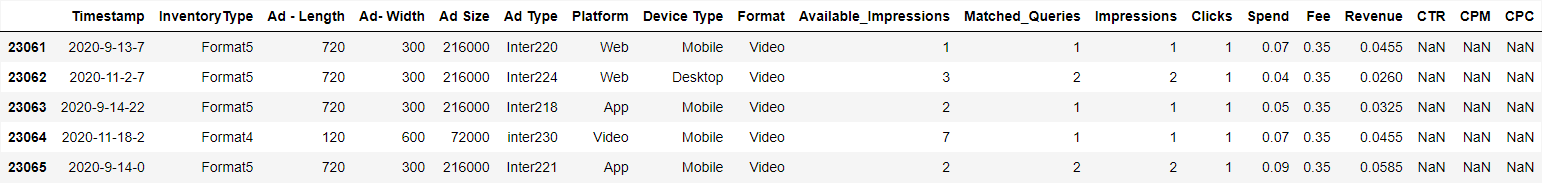
CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column in the dataset and the Number of Clicks refers to the 'Clicks' Column in the dataset.

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.

|  |  |  |
| --- | --- | --- |
| Sl. No | Column Name | Column Description |
| 1 | Timestamp | The Timestamp of the particular Advertisement. |
| 2 | Inventory Type | 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 Advertisement. |
| 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 particular 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 counts 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. |

Table: -1

Q1.1) Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.

● Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.Fig: -1: Dataset HeadFig: -2: Dataset Tail

# 

# Fig: -3: Data info

# Shape of the dataset:

* The dataset has 23066 rows & 19 columns Checking for Null Values

# Above these are the head and tail of the dataset, and there are 6 float, 7 integer and 6 object columns

Q1.2) Treat missing values in CPC, CTR and CPM using the formula given.

# 

# Fig: -4: Null values of Dataset

* 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

We created three functions such as ‘calculate CPC’, ‘calculate CTR’, and ‘calculate CPM’ to treat missing values in CPC, CTR, and CPM columns using the following formula.

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.

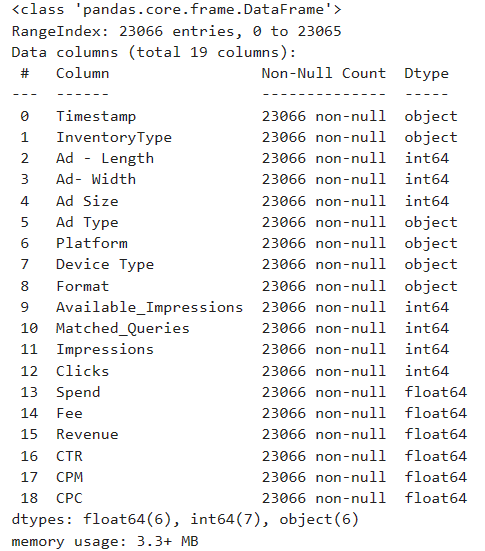
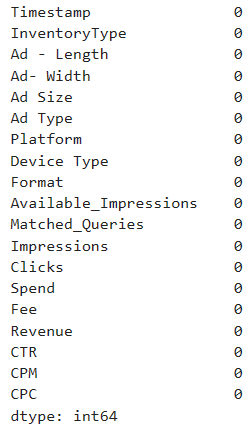


Fig: -6: Dataset after treating Null values

Five Point Summary of the Dataset:

# 

# Fig: -7: Data describe

# As we can see there are no null values now

Q1.3) Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an analyst your judgement may be different from another analyst).

I have checked with the data and it seems that there are Outliers. Below is the Boxplot figure of Features before Treating Outliers.

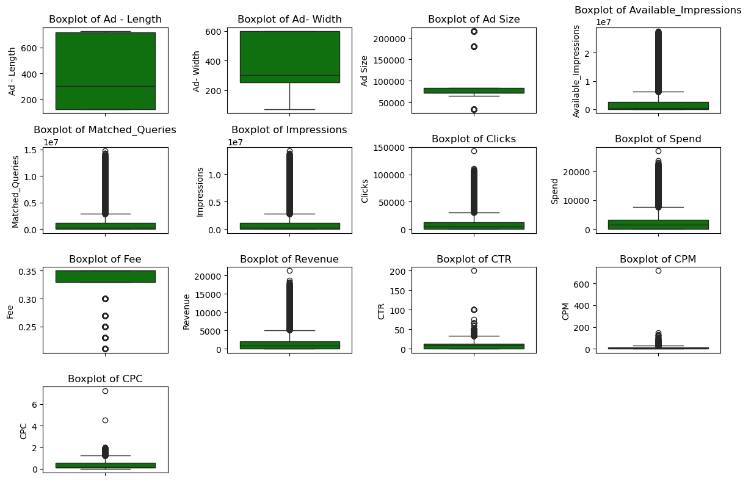


Fig: -8: Boxplot before Treating outliers

Yes. Treating Outliers is necessary for K Means Clustering. We are going to treat outliers by IQR method (Inter Quartile Range).

I have created a ‘remove outlier’ function using IQR formulas. We can’t perform Outlier Treatment on Categorical features. Hence, I have created new dataset of Int64 and Float64 datatypes with the name of df\_Num. And, applied ‘remove outlier’ function on it. And removed outliers.

To treat outliers, we defined a function 'treat outlier' where:

* The larger values (>upper whisker) are all equated to the 95th percentile value of the distribution.
* The smaller values (<lower whisker) are all equated to the 5th percentile value of the distribution.
* Upper Range = Q3 + 1.5\*IQR.
* Lower Range = Q1 -1.5\*IQR

Find below Boxplot diagram after treating Outliers.

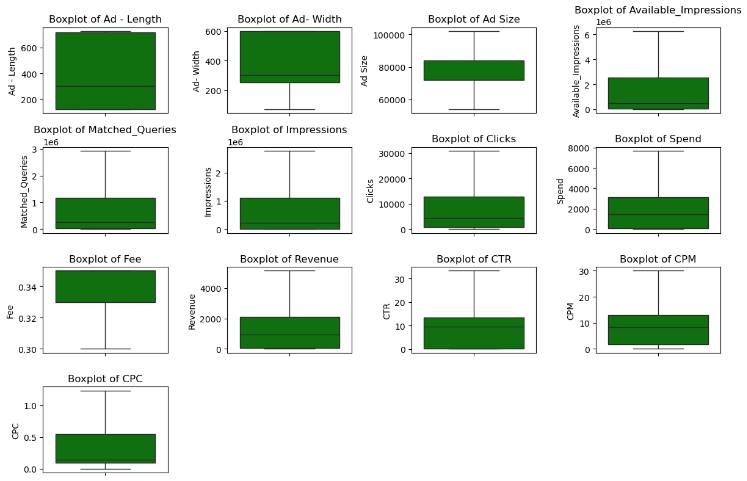


Fig: -9: Boxplot After treating outliers

Q1.4) Perform z-score scaling and discuss how it affects the speed of the algorithm?

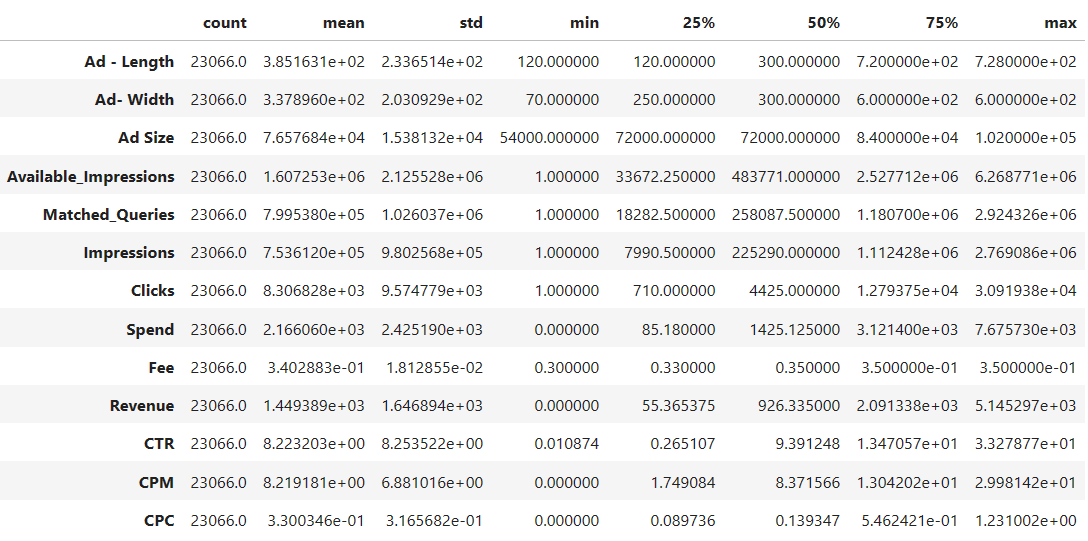
Data before Z-score Scaling is as below 

Fig: -10 After applying Z-score

Here, I have applied z-score method on the ‘df\_Num’. And, I got the below output.



Fig: -11

Scaling of variables is important for clustering to stabilize the weights of the different variables. If there is wide discrepancy in the range of variables (refer to Table 3) cluster formation may be affected by weight differential.

*The features contained in a data set may have different units (e.g. feet, kilometers, and hours) that, in turn, may mean that the variables have different scales. All machine learning algorithms are dependent on the scaling of data. If there is wide discrepancy among the input values, the unscaled model may be unstable, meaning that it may suffer from poor performance during learning and sensitivity to input values resulting in higher generalization error. [2]*

*One of the most common forms of pre-processing consists of a simple linear rescaling of the input variables.*

— Page 298, Neural Networks for Pattern Recognition, 1995.

Q1.5) Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.

Using SciPy’s cluster hierarchy function, we created the below dendrogram.

Please find below Dendrogram performed for Hierarchical using WARD and Euclidean Distance on the Scaled Data such as “df\_scaled”.

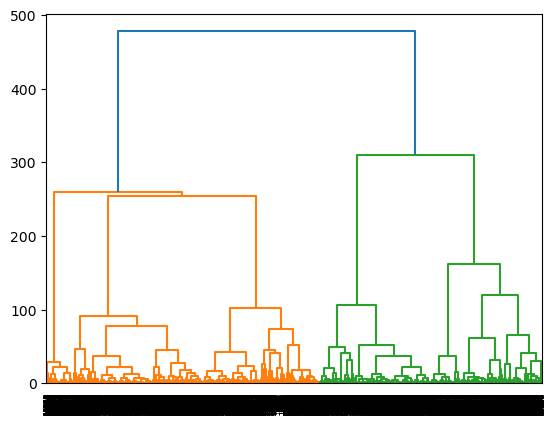


Fig: -12

In a Dendrogram, each branch is called a clade. The terminal end of each clade is called a leaf. The arrangement of the clades tells us which leaves are most similar to each other. The height of the branching points indicates how similar or different they are from each other: the greater the height, the greater the difference.

[reference - [https://wheatoncollege.edu/wp-content/uploads/2012/08/How-to-Read-a-Dendrogram-Web-Ready.pdf ]](https://wheatoncollege.edu/wp-content/uploads/2012/08/How-to-Read-a-Dendrogram-Web-Ready.pdf)

Keeping the above reference as base, we can see the longest branch (tallest branch) is in blue. If we see that only blue, it will result in only 2 clusters which is not acceptable in business. If the segmentation is at the tallest green branches, separated by the yellow horizontal line, 5 clusters are identified. Alternatively, there may be 3 clusters as well, designated by the orange horizontal line. But we choose 5 Clusters using Dendrogram for this project.

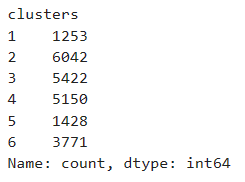
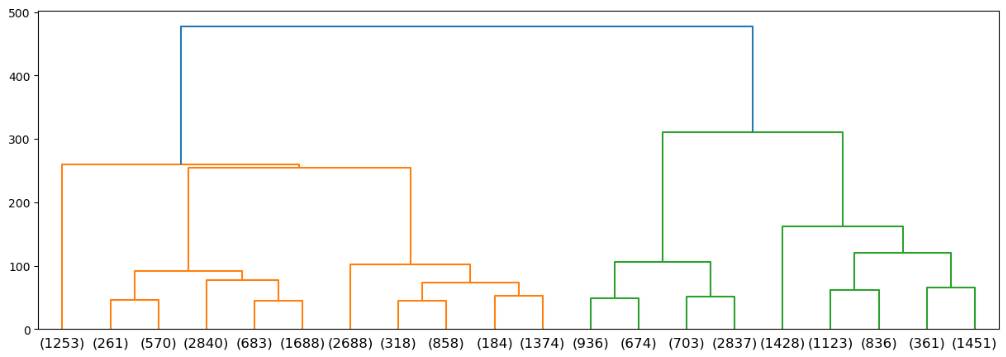


Fig: -13: Value counts

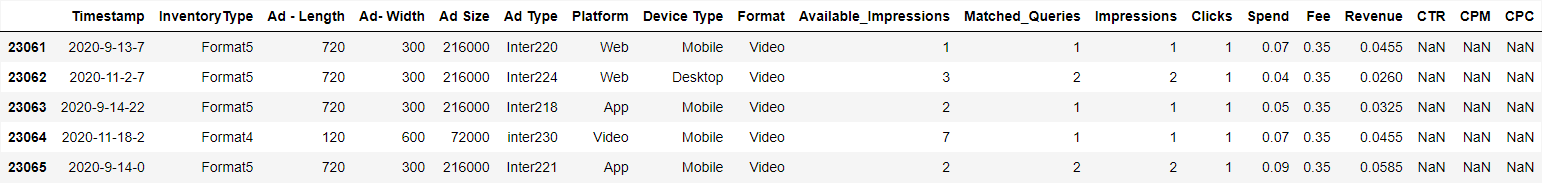


Fig: -14

Q1.6) Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.

For checking the Optimal number of clusters, we use WSS (Within Sum of Square)

Elbow Plot for n=10 The optimum number of clusters for k-means algorithm are 5 as the drop become significant

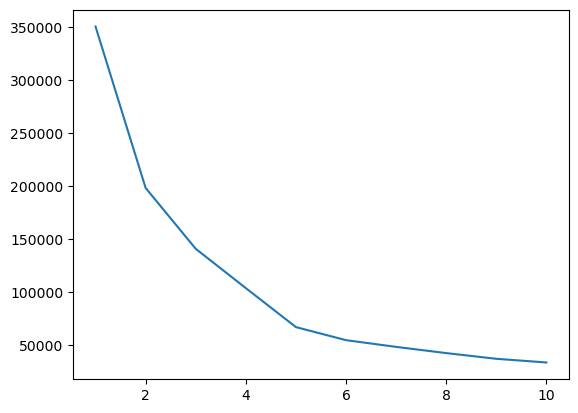


Fig: -15: Line plot

As per the check, when we move from K=1 to K=2, We see that there is a significant drop in the value. Also, when we move from k=2 to k-3, k=3 to k=4, k=4 to k=5 there is a significant drop as well. k=5 to k=6, the drop in values reduces significantly. Hence In this case, the WSS is not significantly dropping beyond 5, so 5 is optimal number of clusters

Q1.7) Print silhouette scores for up to 10 clusters and identify optimum number of clusters.

**The silhouette score here is** 0.5884606993260426**.**

The silhouette score for rest clusters up to 10.

Hierarchical Clustering as well as K-Means Clustering were performed. We used Elbow plot and Silhouette Score to identify optimum number of clusters in K-Means whereas in Hierarchical Clustering dendrogram was drawn. In Hierarchical method, we got 5 clusters while in K-Means, we got 5 (using elbow plot) and 6 clusters (using silhouette score). We can always try alternative approaches to clustering using other linkage types and distance metrics for an exhaustive study of the data. Please refer to the Monograph for details. We observe that the methods used in this project yielded similar results i.e. with 10 clusters. I have calculated Silhouette Score for scaled data using the silhouette\_score () function.

The Silhouette Score is a measure of how similar an object is to its own cluster compared to other clusters, and it ranges from -1 to 1, with higher values indicating better clustering.

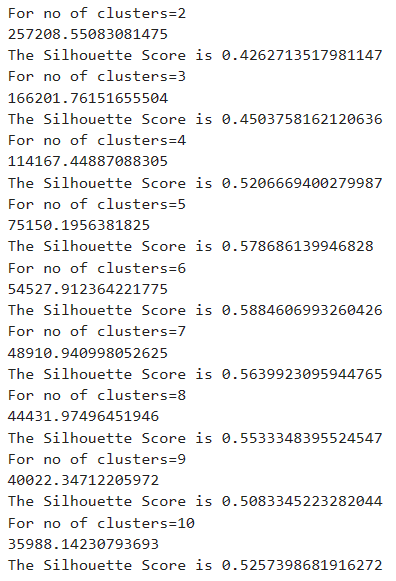


Fig: -16: Clusters

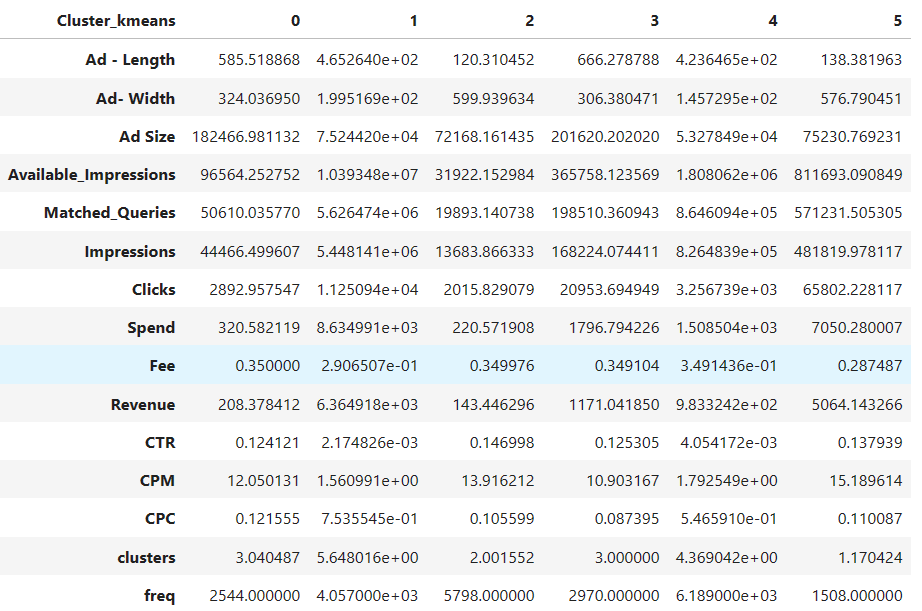


Fig: -17: Cluster k means

Q1.8) 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].

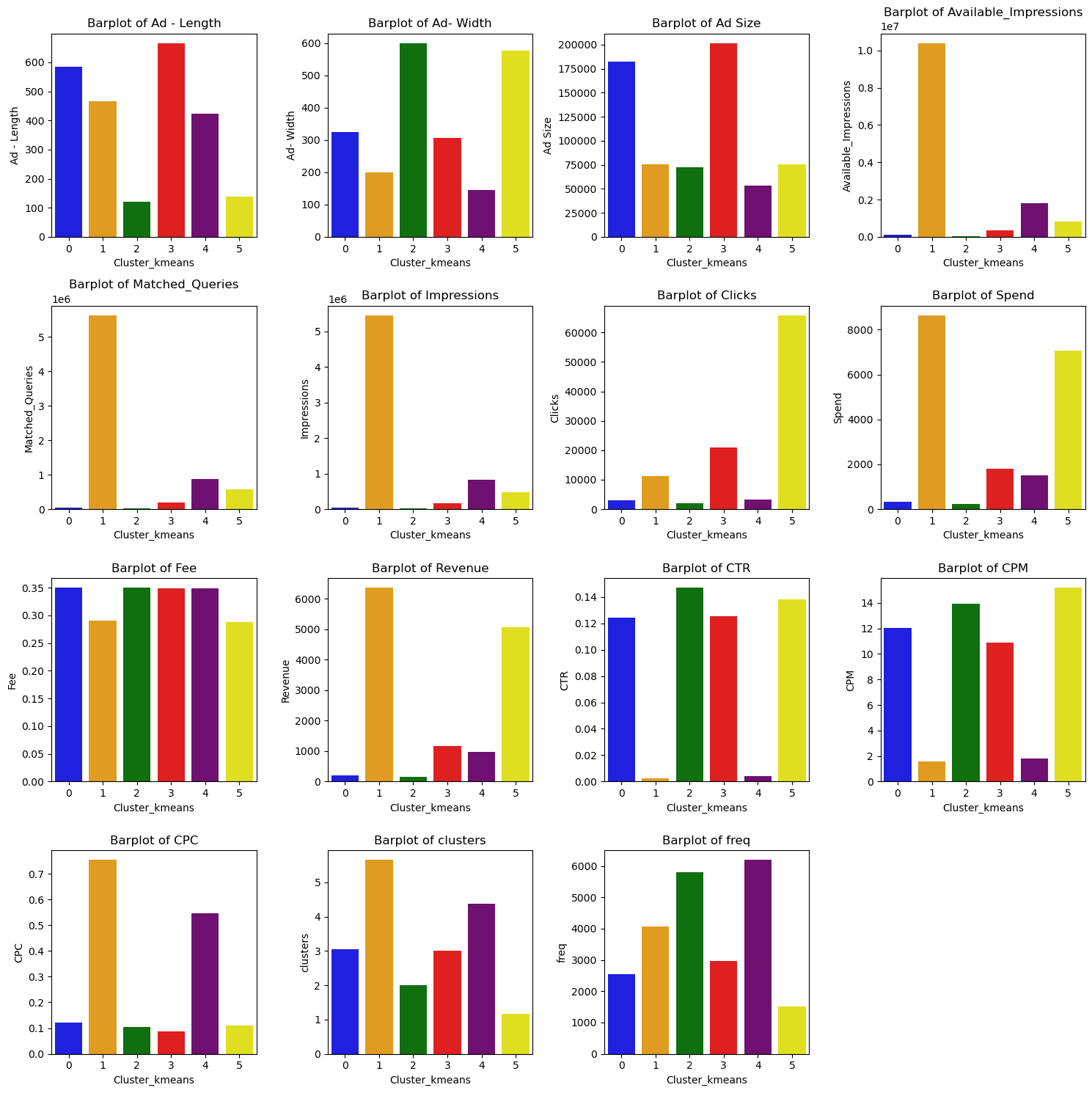


Fig: -18: Bar plot Comparisons

Observations:

* The clusters 3 contain ads that have higher mean length than other clusters.
* The clusters 2 and 5 have ads whose mean width is considerably more than the other clusters
* Cluster 3 has minimum ad size
* Available impressions is highest for cluster 1
* There is not much difference in Fee, but cluster 1 has very high mean spend and mean revenue compared to the others
* Cluster-2 have the most Click through rate (CTR).
* Cluster-1 have the highest cost per 1000 impressions (CPM)
* For cluster 1 and 4 the CPC (Cost per Click) is highest.

The data is grouped/profiled using the optimum number of clusters using silhouette score, which is five, and the mean has been taken to identify the trend in clicks, spends, revenue, CPM, CTR and CPC based on device type.

Means Clicks based on Device Type:

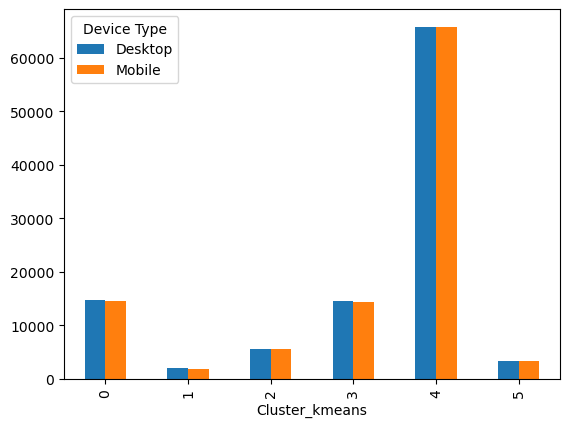
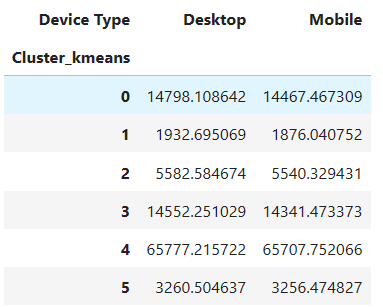


Fig: - 19: Mean Clicks based on Device Type

The above bar chart, fig 19, clearly depicts the mean of cluster 4 at the peak for both the desktop and mobile devices at above 650000 clicks. Followed by Clusters 0 and 3 are at the highest, with means of above 14,798.108 and 14,467.467 clicks, respectively. And cluster 1 and cluster 5 registered with the lowest click.

Mean Spend based on Device type:

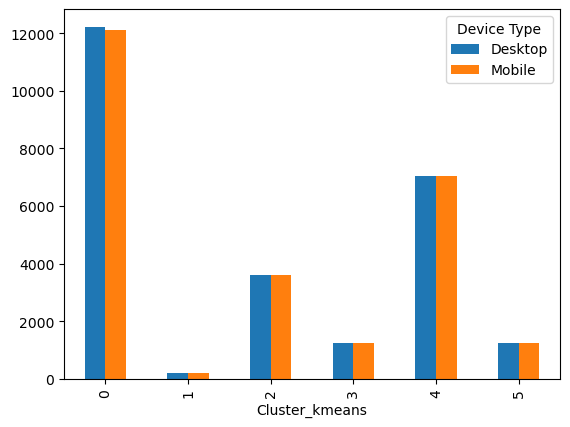
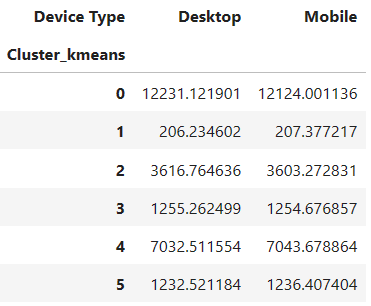


Fig: -20: Mean Spend based on Device type

The above figure 20 bar chart represents clustered data of mean spending based on device type. As we can spend mean, cluster 0 is at the top for both devices and followed the cluster 4 at the second spot. Whereases cluster 1 is at the lowest spend for both the Desktop and mobile at approx.206 and 207.

Mean revenue based on Device type:

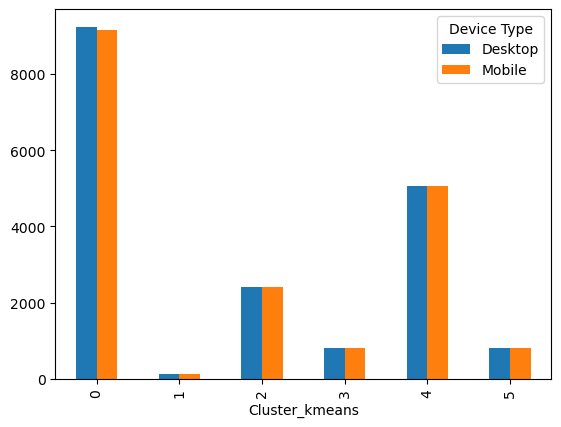
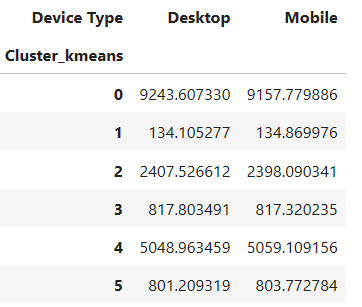


Fig: -21: Mean revenue based on Device type

The above fig 21 chart shows the clustered data of mean revenue based on device type. It is evident that the cluster 0 mean revenue is the highest with approx. 9243 and 9157 and followed by cluster 4 with approx. 5048. Cluster 2 holds the last position with the lowest mean revenue.

Mean CPM (Cost per 1000 impressions) based on Device type:

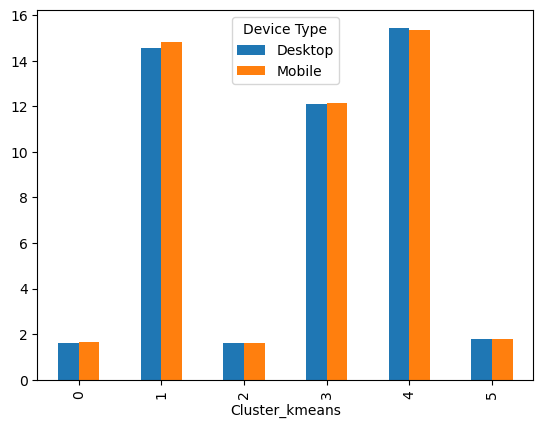
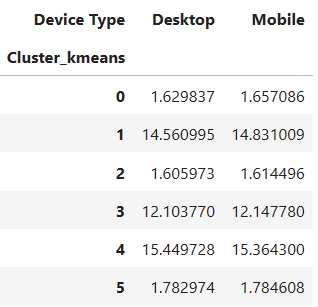


Fig: - 22: Mean CPM (Cost per 1000 impressions) based on Device type

The above chart, fig 22, shows clustered data of the mean of Cost per 1000 Impressions (CPM) based on the device type. As we can see, cluster 4 has the highest mean CPM with close to 16, followed by cluster 1 with approx. 15 for both devices. Clusters 0 and 2 are at the lowest, with approx. 2 mean CPM.

Mean CTR (Click through Rate) based on Device type:

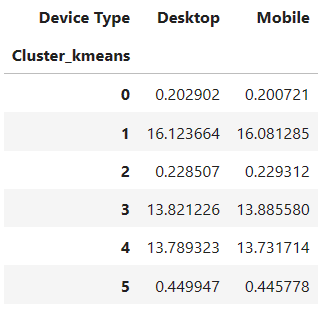
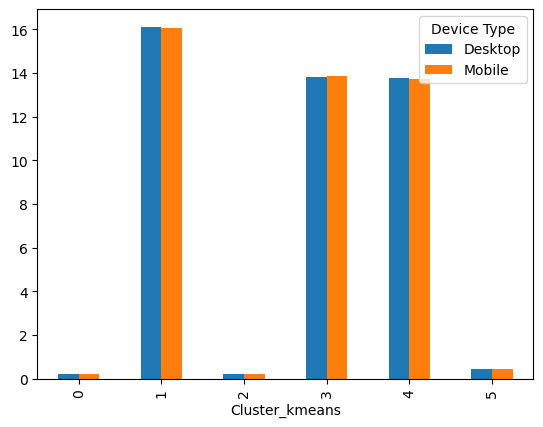
 

Fig: -23: Mean CTR (Click through Rate) based on Device type

The above fig 23 chart shows clustered data of mean of Click Through Rate (CTR) based on the device type. The cluster 1 is at the highest mean CTR for the both mobile and desktop device at 16. Followed by the cluster 3 and cluster 4 at mean CTR of 14 and Cluster 3 at the lowest at cluster 1 and cluster 0.

Mean CPC (Cost per Click) based on Device type:

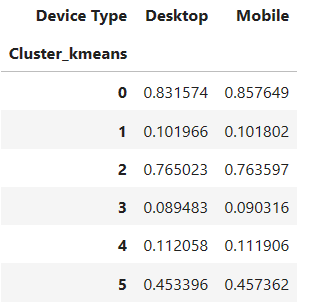
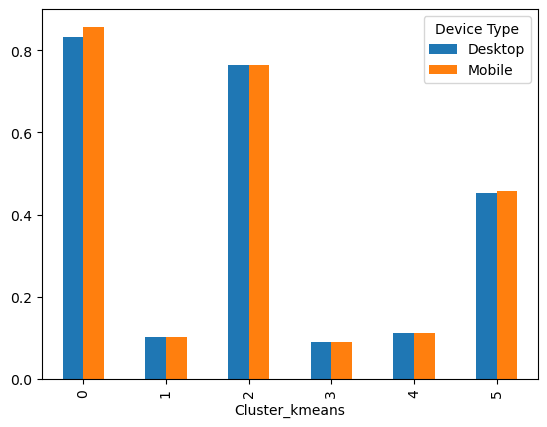
 

Fig: -24: Mean CPC (Cost per Click) based on Device type*:*

The above fig 24 chart shows clustered data of mean Spend per Click (CPC) based on the device type. As we can see, cluster 0 is at the peak for both devices at closer to 0.8 and followed by cluster 2 with a mean of 0.76 CPC. Clusters 1, 3 and 4 follow a similar trend with 0.1.

Q1.9) Clustering: Conclude the project by providing summary of your learnings.

Summary of the dataset:

The ad 24x7 marketing company have collected the data from the marketing intelligence to analysis and segmentalize the ads to target the right set of groups. The following details were found during the assessment of the dataset.

* The dataset has 23066 rows and 19 features. Out of 19 features – 6 are float type, 7 are integer and six are categorical.
* During the project, we also found 4736 missing values; those values were imputed and treated with the given formula using a user-defined function.
* The dataset also had outliers; those outliers were treated using the IQR method since the K-mean clustering is sensitive to outliers and could negatively influence the dataset.
* The dataset also has been scaled. Since the unscaled data could negatively impact the speed of the algorithm and scaling data can make the variable contribute equally to the analysis to take better business decisions.
* As per Elbow plot/scree-plot, we concluded that the optimal number of clusters should be 5.
* Plotted elbow plot and got optimum value is 5
* Hierarchal clustering has been performed with the data to find the optimum number of clusters, which is 5.

Conclusions after Clustering:

* + When Click on Ads gets increases then Revenue is also increases.
  + When amount of money spent on specific ad variations within a specific campaign or ad set is increases then Revenue is also increases.
  + When impression count of the particular Advertisement increases then Revenue is also increases

# Problem 2:

PCA: PCA FH (FT): Primary census abstract for female headed households excluding institutional households (India & States/UTs - District Level), Scheduled tribes - 2011 PCA for Female Headed Household Excluding Institutional Household. The Indian Census has the reputation of being one of the best in the world. The first Census in India was conducted in the year 1872. This was conducted at different points of time in different parts of the country. In 1881 a Census was taken for the entire country simultaneously. Since then, Census has been conducted every ten years, without a break. Thus, the Census of India 2011 was the fifteenth in this unbroken series since 1872, the seventh after independence and the second census of the third millennium and twenty first century. The census has been uninterruptedly continued despite of several adversities like wars, epidemics, natural calamities, political unrest, etc. The Census of India is conducted under the provisions of the Census Act 1948 and the Census Rules, 1990. The Primary Census Abstract which is important publication of 2011 Census gives basic information on Area, Total Number of Households, Total Population, Scheduled Castes, Scheduled Tribes Population, Population in the age group 0-6, Literates, Main Workers and Marginal Workers classified by the four broad industrial categories, namely, (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 | Literate population Male |
| F\_LIT | Literate 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 Laborers Population Male |
| MAIN\_AL\_F | Main Agricultural Laborers 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 Laborers Population Male |
| MARG\_AL\_F | Marginal Agriculture Laborers 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 laborer’s Population 3-6 Male |
| MARG\_AL\_3\_6\_F | Marginal Agriculture Laboure’s 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 Laboure’s Population 0-3 Male |
| MARG\_AL\_0\_3\_F | Marginal Agriculture Laboure’s 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 |

Table: -2

## **Q2.1 PCA: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.**

The below figure shows the first 5 Rows of the dataset.

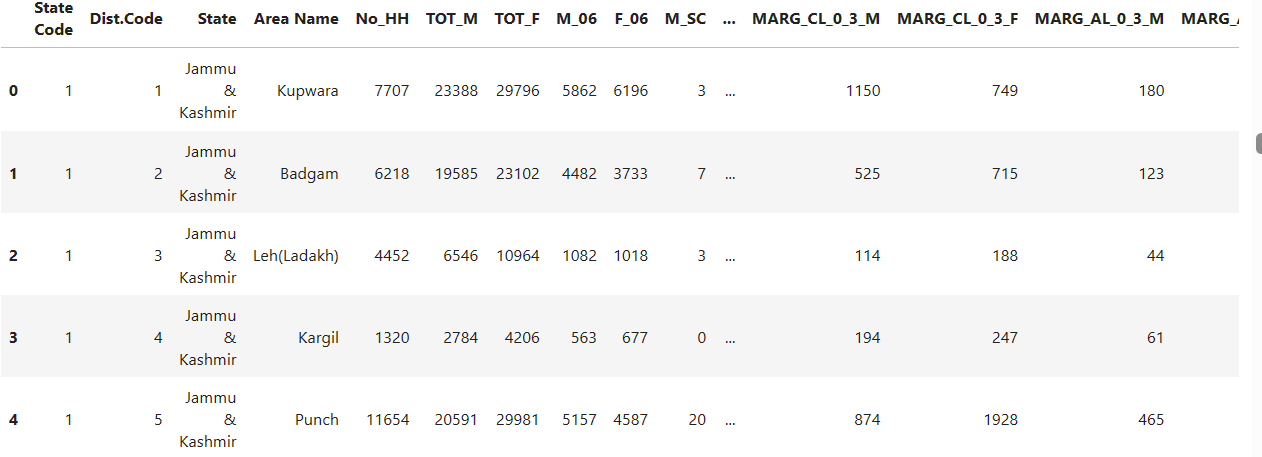


Fig: -25: Head of the Dataset

The below figure shows the last 5 rows of the dataset:

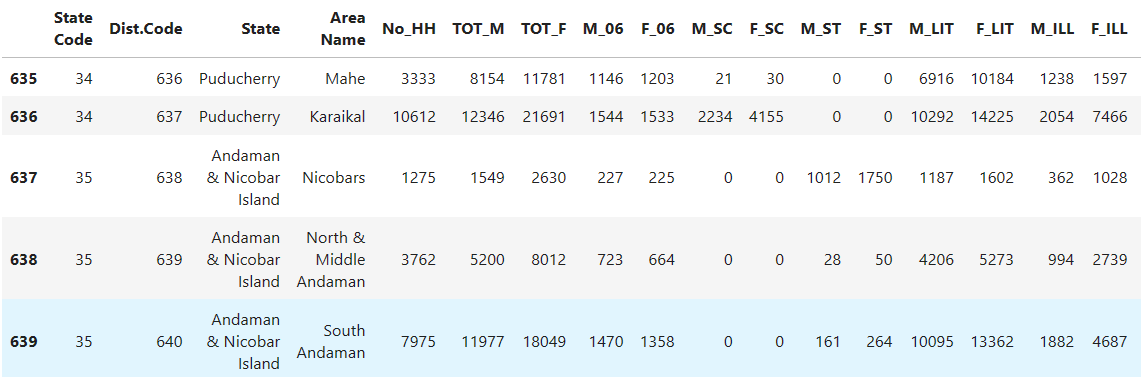


Fig: -26: Tail of the dataset

Data Info:

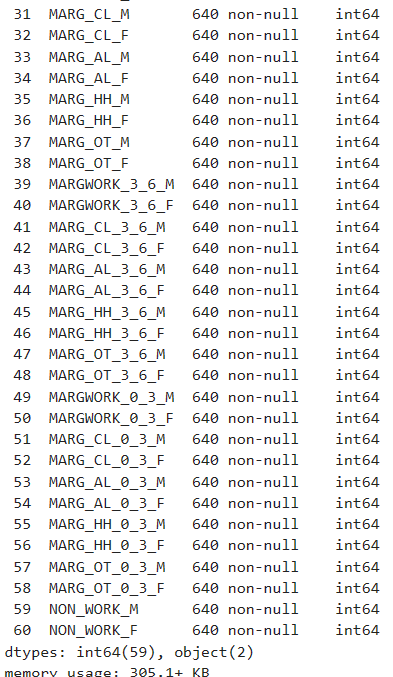
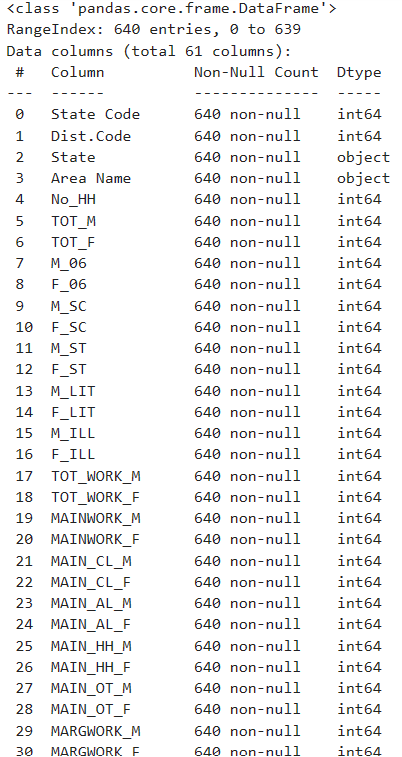


Fig: -27: Data information

From the above data, we can see 640 rows with 61 columns. Out of 61 features – 59 columns belong to the integer data type, and 2 are object (Categorical data type).

The below fig 20 shows the data that depicts the mean, median, min and max values of the dataset. The dataset looks skewed.

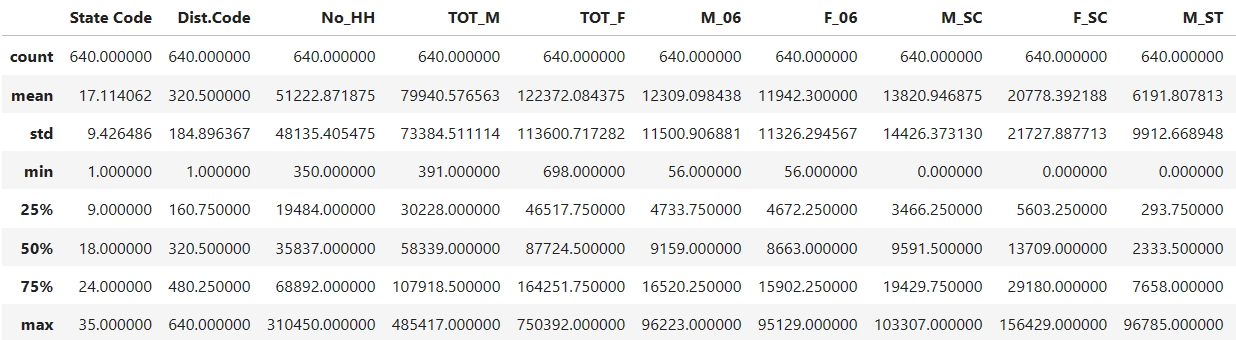


Fig: -28: Data Describe

The below fig 21 proves that there are no duplicates and null values present in the dataset.

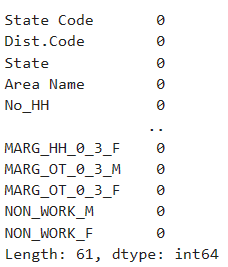


Fig: -29: Null values

Checking for duplicate values:

* + There are no duplicate values in the given dataset Summary of the data

Q2.2 PCA: 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.

(i)Which state has highest gender ratio and which has the lowest?

I have picked 5 Variables such as ’No\_HH’, ‘TOT\_M’, ‘TOT\_F’, ‘M\_06’, and ‘F\_06’. And comparing those 5 variables against 'State' and Area Name.

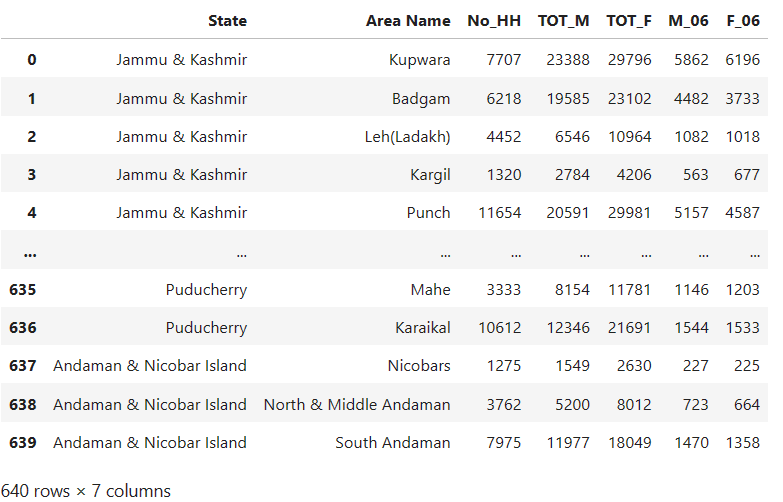


Fig: -30: Chosen Variables

I have created the Gender Ratio Columns by calculating the Male/Female, So the new columns are ‘TOT\_GR’ and ‘GR\_06’.

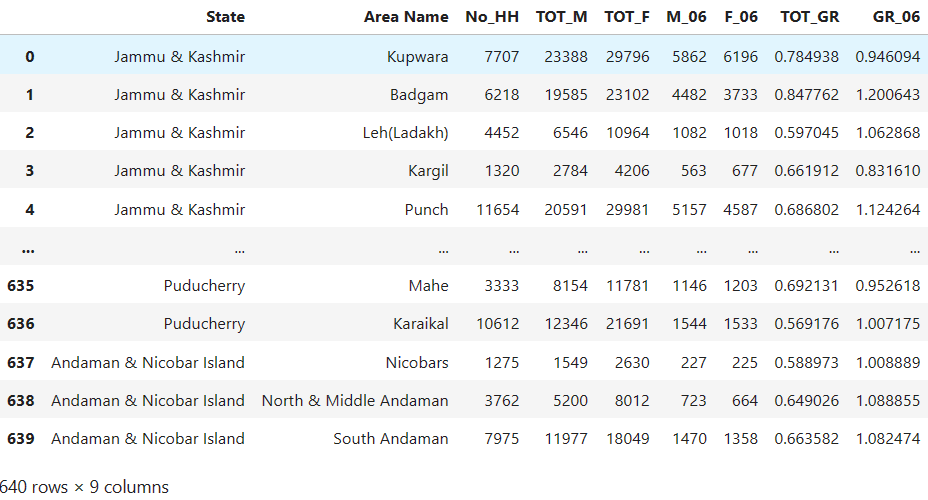


Fig: -31: Calculated gender Ratio

The below bar graph fig 24 represents the gender ratio among the Indian states. We can see that Lakshadweep has the highest gender ratio with over 0.9, and Chhattisgarh and Andra Pradesh have the lowest gender ratio with 0.5.

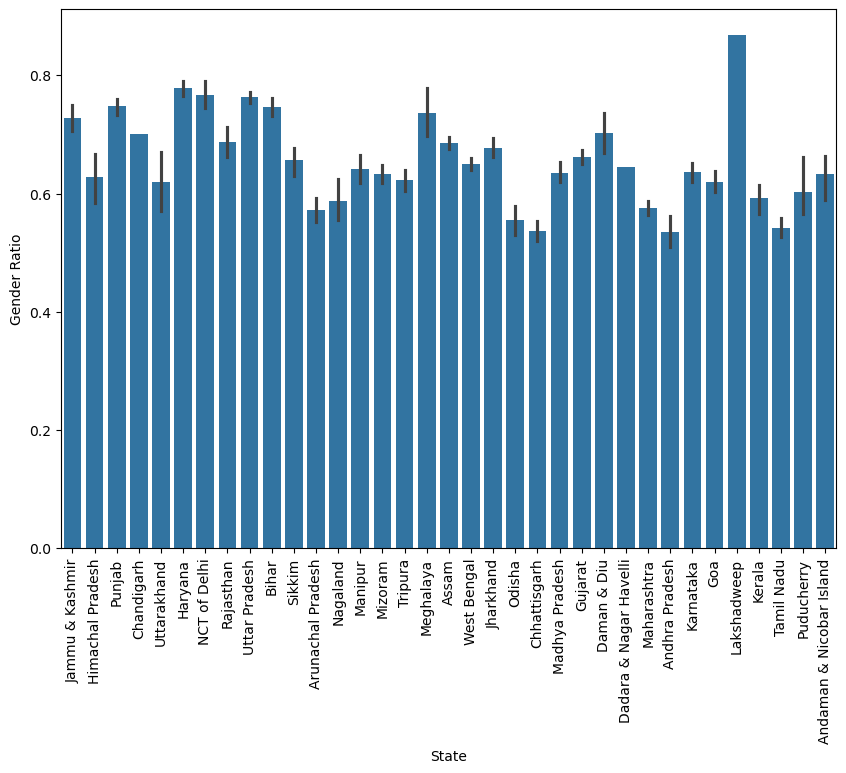
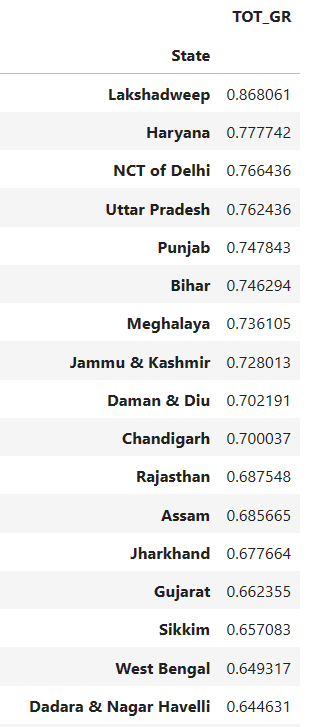


Fig: -32: Gender Ratio

(ii) Which district has the highest & lowest gender ratio?

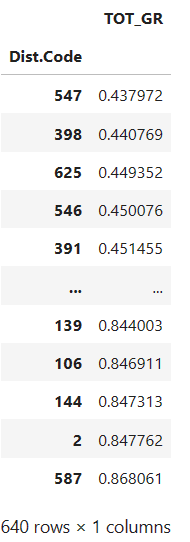


Fig: -33: District wise Gender Ratio

As we can see the highest and lowest Gender Ratio, the above fig-33 represents the gender ratio among the Indian Districts. We can see that 587- has the highest gender ratio with over 0.9, and 547 have the lowest gender ratio with 0.5.

Q2.3 We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?

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.
5. Principal Component Analysis (PCA) is a highly flexible multivariate data dimension reduction method. In the presence of outliers, classical PCA is highly sensitive to them and may draw false conclusions.

## **Q2.4 PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.**

Before scaling and checking the outliers, the Categorical variable has been dropped from the dataset. The below image represents the boxplot of the dataset before scaling.

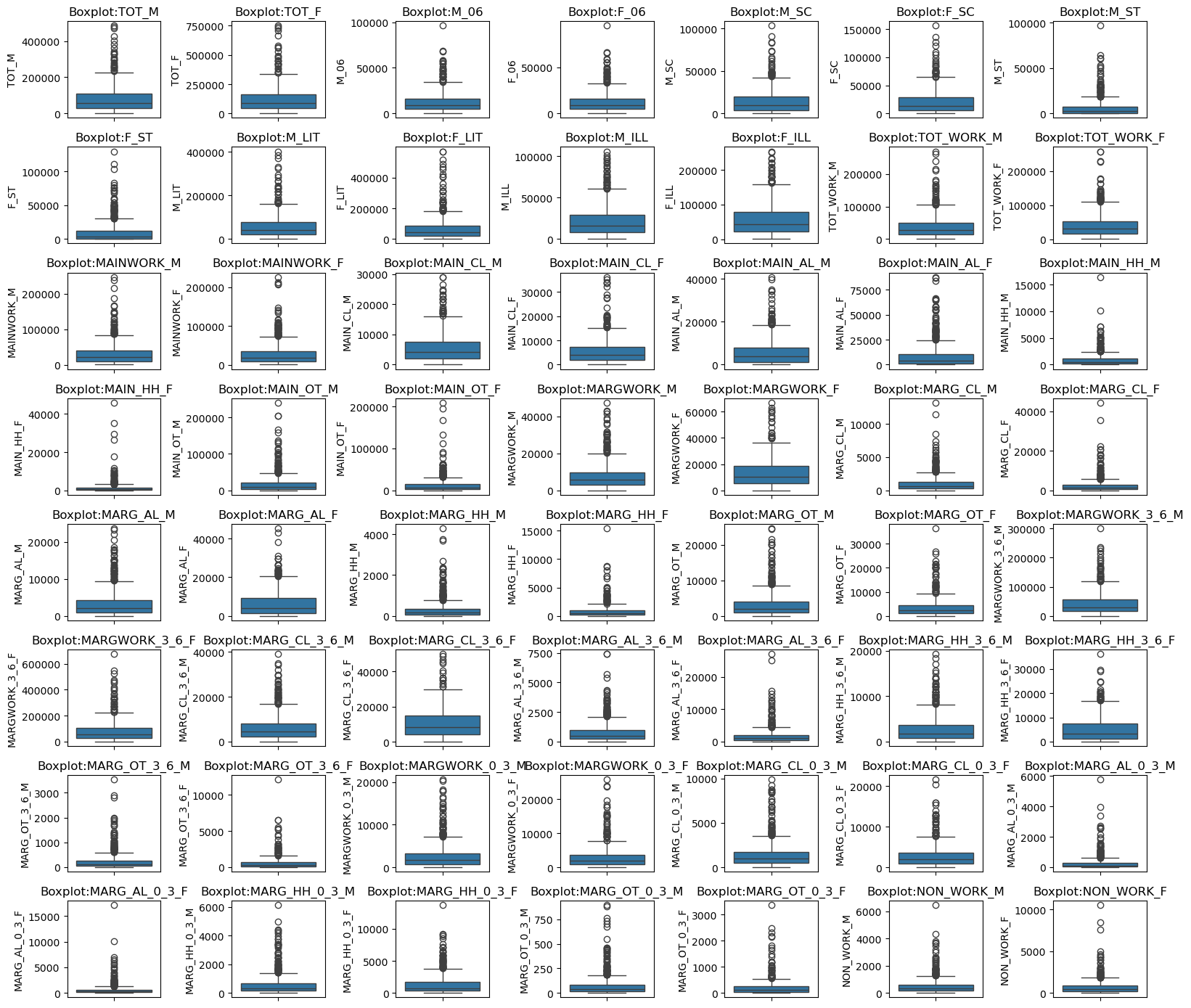


Fig: -34: Boxplot before scaling

As we have applied the Z-score for the Dataset, below we can see the Dataset fig-35 the values tend to be below zero.

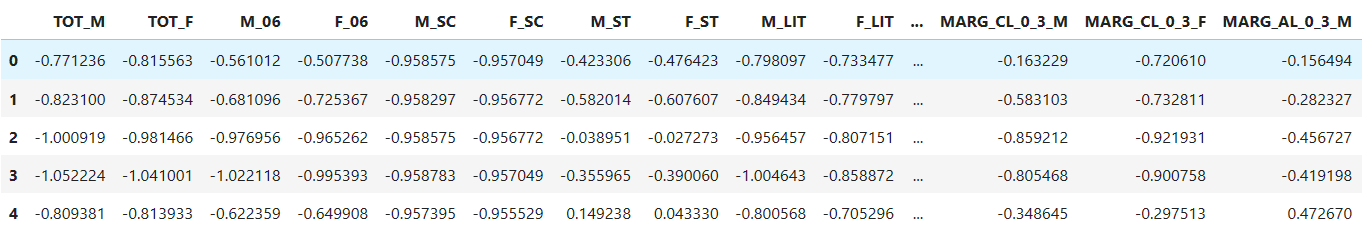


Fig: -35: Dataset after applying Z-score

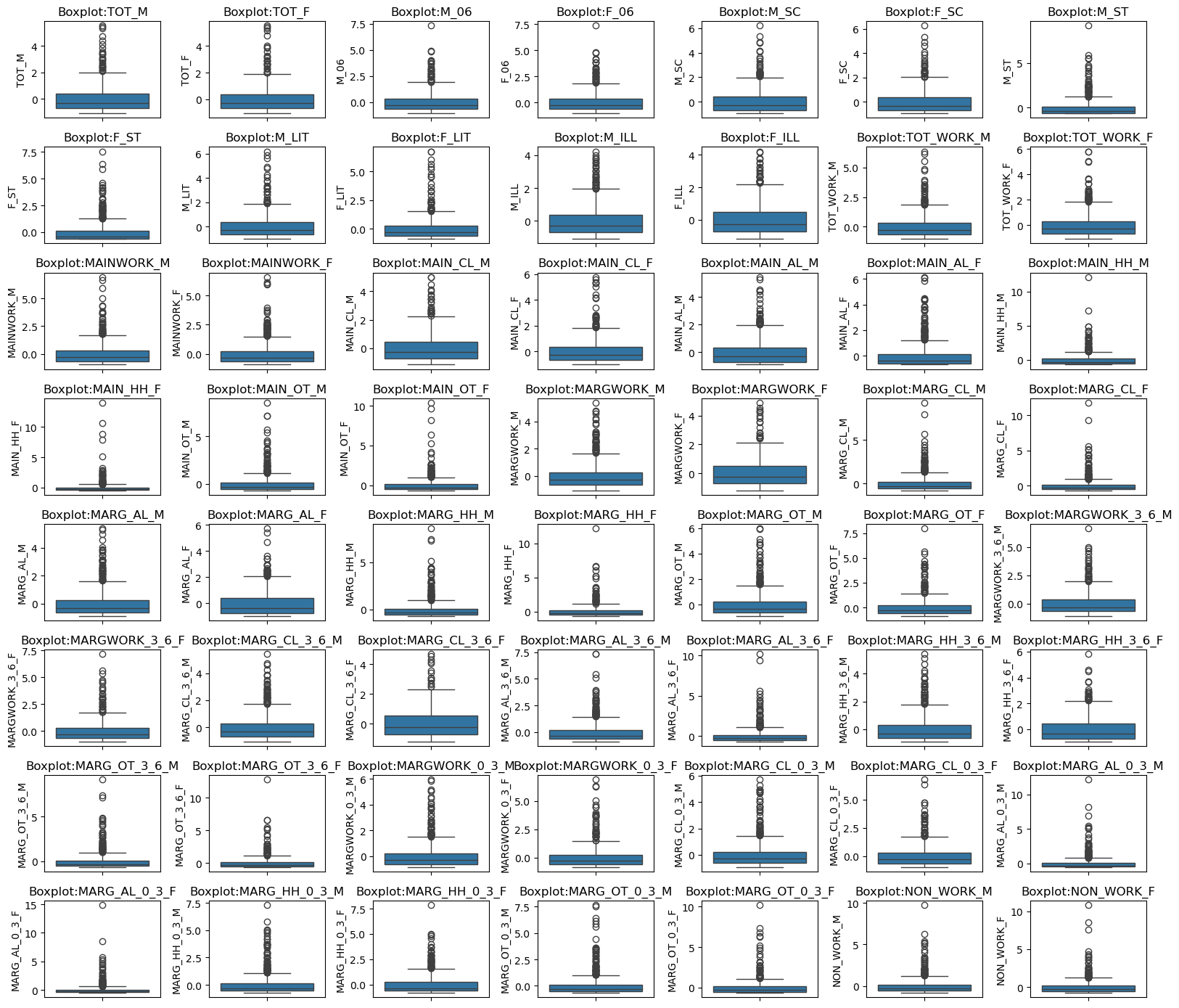


Fig: -36: Boxplot After Scaling

Apart from the scaling adjustment, there are absolutely no changes when we compare the box plot before and after scaling.

## **Q2.5. PCA: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.**

Checking the Correlation.

Statistical test is done before performing PCA. Though we have seen few correlations in the dataset. The *Bartletts test of Sphericity* is performed to understand correlation significance in the population.

The Null hypothesis an alternative hypothesis will be defined.

H0: All variables in the data are uncorrelated

Ha: At least one pair of variables in the dataset are correlated.

We Decide the significance level Here we select

α = 0.05. – we got ‘0’

The p-value is 0, which is less than the significant level. Therefore, the null hypothesis will be rejected. Hence, it proves that at least one pair of variables in the dataset is correlated, and PCA will be performed.

*And the next step would KMO test: (Kaiser-Meyer-Olkin)*

The KMO test will be conducted to measure the sample adequacy (MSA) of the dataset.

If MSA is less than 0.5, PCA will not be suggested. Alternatively, if it is greater than 0.7, it gives a substantial reduction in dimension and extracts significant components.

KMO is 0.80, Since it is greater 0.7 – the PCA is recommend.

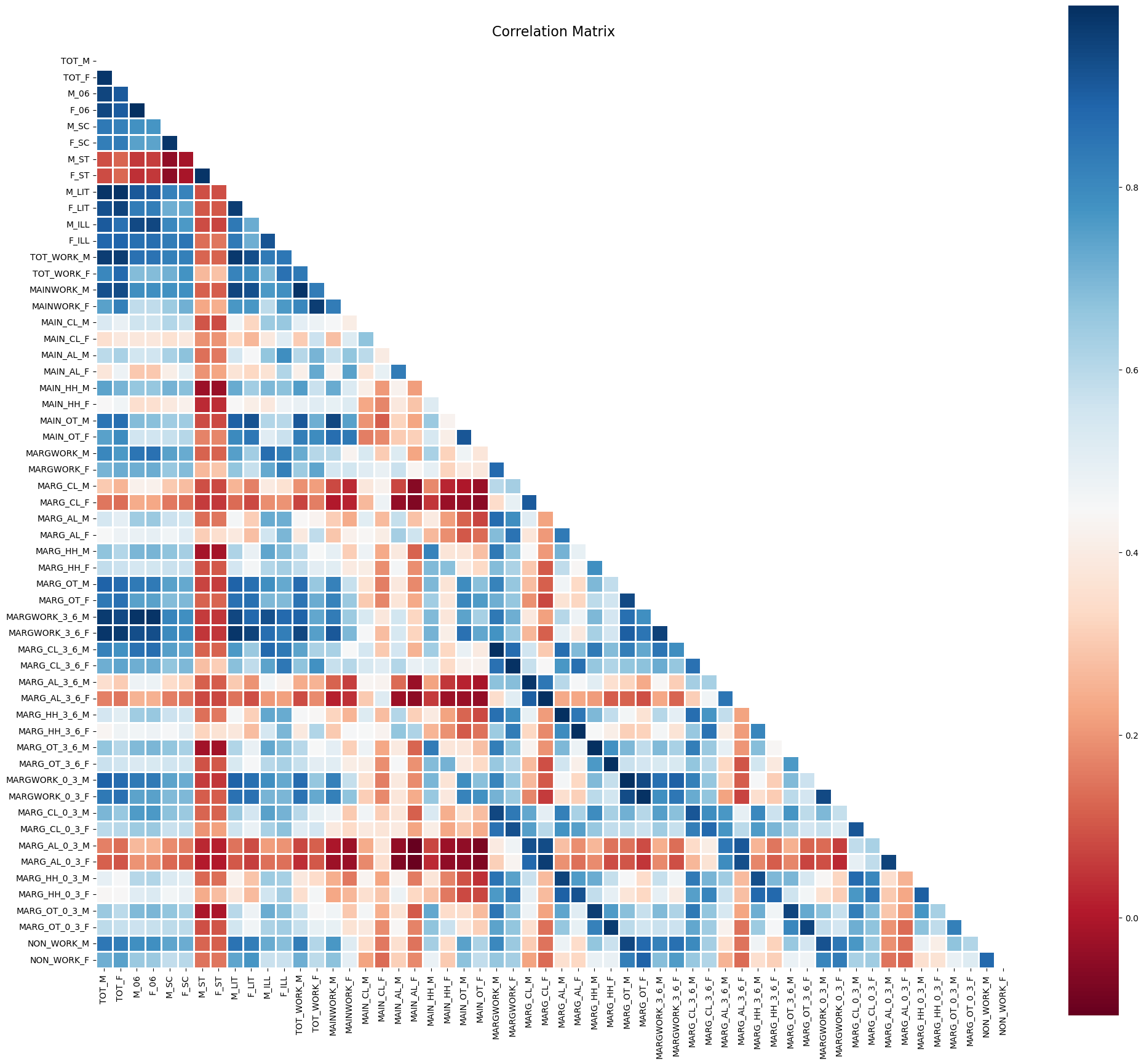
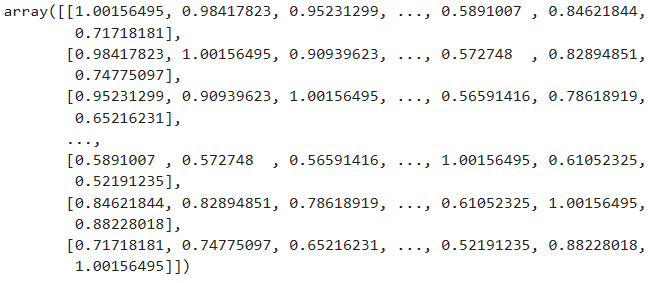


Fig: -37: Heatmap

Covariance Matrix:



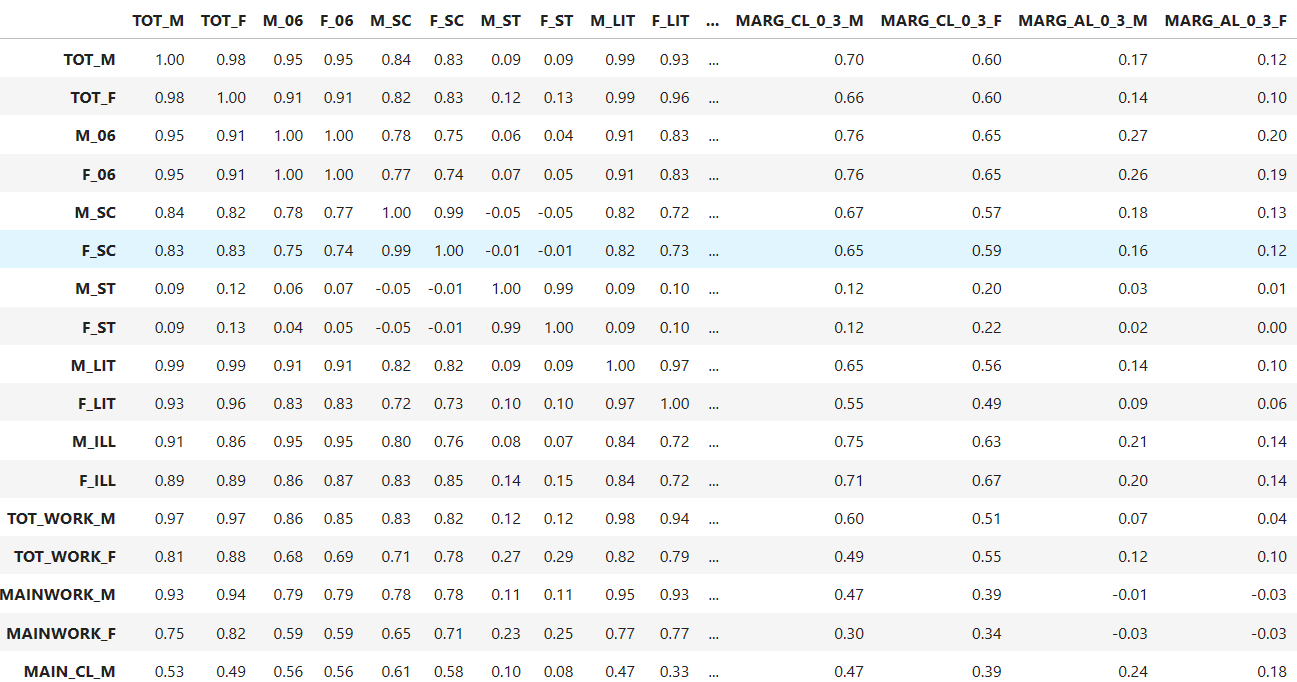


Fig: -38: Covariance matrix

Eigen Values:

The below fig 39 represents the Eigen value of all 12 principal components.

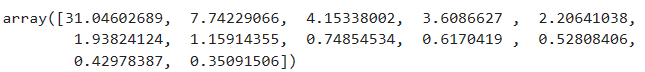


Fig: -39: Eigen value

Maximum variance is explained by PC1 = 31.81.

PC2 explains 7.86

PC3 explains 4.153

PC4 explains 3.66

PC5 explains 2.20 and PC6 explains 1.93.

The below figure shows the Eigen Vector of all the principal components. It derived by using Components

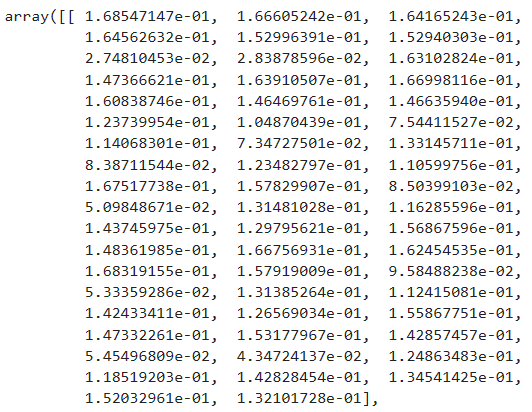


Fig: -40: Eigen vector

## **Q2.6 PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.**

The Explained variance ratio explain the proposition of variance of the principal components.

Let’s check out the percentage of variance explained by each PC that is variance explained by an individual principal component divided by total variance explained by all the PC’s.

In other words – Percentage of explained variance= Eigen value of each PC/sum of Eigen values of all PCs

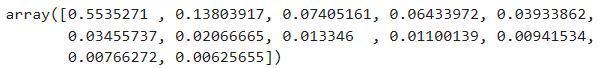


Fig: -41: Explained variance

55% of total variance is explained by PC1. 13.7% of total variance is explained by PC2.

7.2% of total variance is explained by PC3. 6.4% of total variance is explained by PC4.

3.8% of total variance is explained by PC5. 3.4% of total variance is explained by PC6.

Cumulative explained variance ratio:

The cumulative explained variance ratio to find a cut off for selecting the number of Principal components. (PCs)

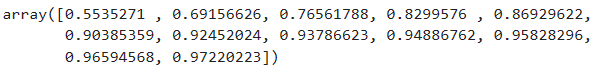


Fig: -42: Cumulative Explained variance ratio

We can see from the above fig.42 that Cumulative explained variance ratio of 6 pcs is more than 90%. Therefore, we can conclude by saying the optimum number of PCs is 6 and the below fig 43 Scree plot supports the same.

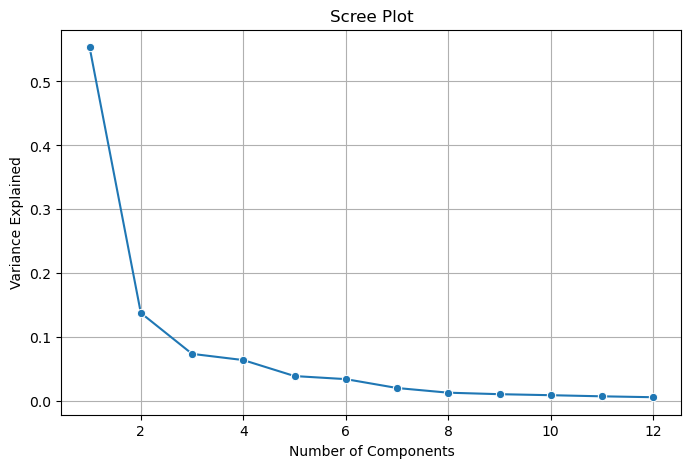


Fig: -42: Scree plot

The dots on the Scree plots are the 12 principal components. We can see a formation of the elbow at 6 PCs. It is evident that post 6 PCs, the drop is not that significant. therefore, the optimum number of PCs is 6.

## **Q2.7 PCA: Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.**

## To categorize the pattern, the components are loaded against each feature in a new data frame. we have 12 principal components and one co-efficient each for all 56 variables.

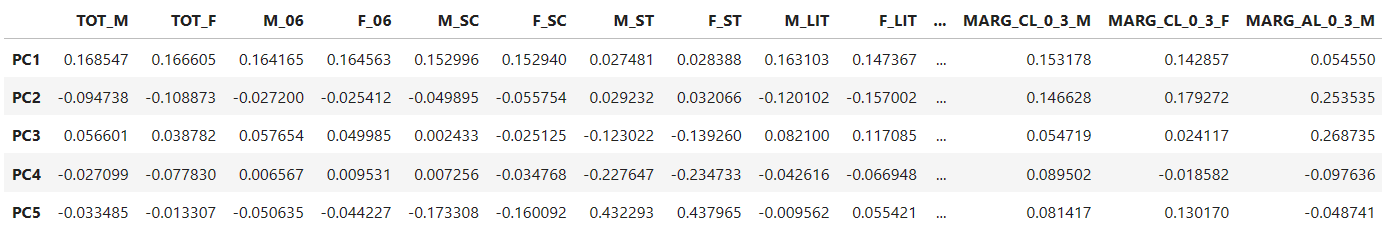


Fig: -43: Selected components

To analysis the variable that has the highest loading among the principal components. The component has to be loaded on a heatmap. For each variable with maximum loading, the heatmap shows a blue rectangle box that is marked across the components. Ref Fig. 44

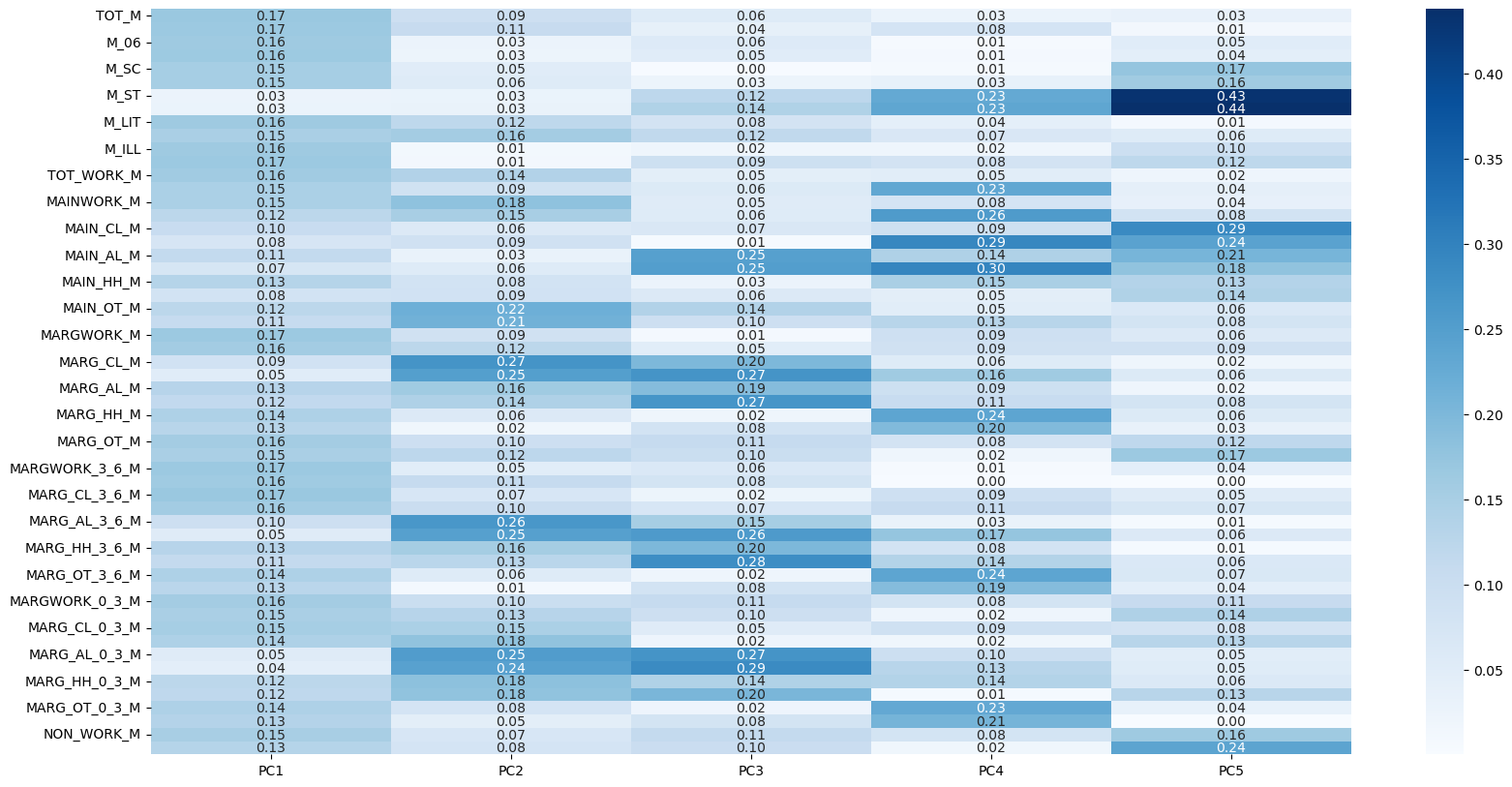


Fig: -44: Heatmap

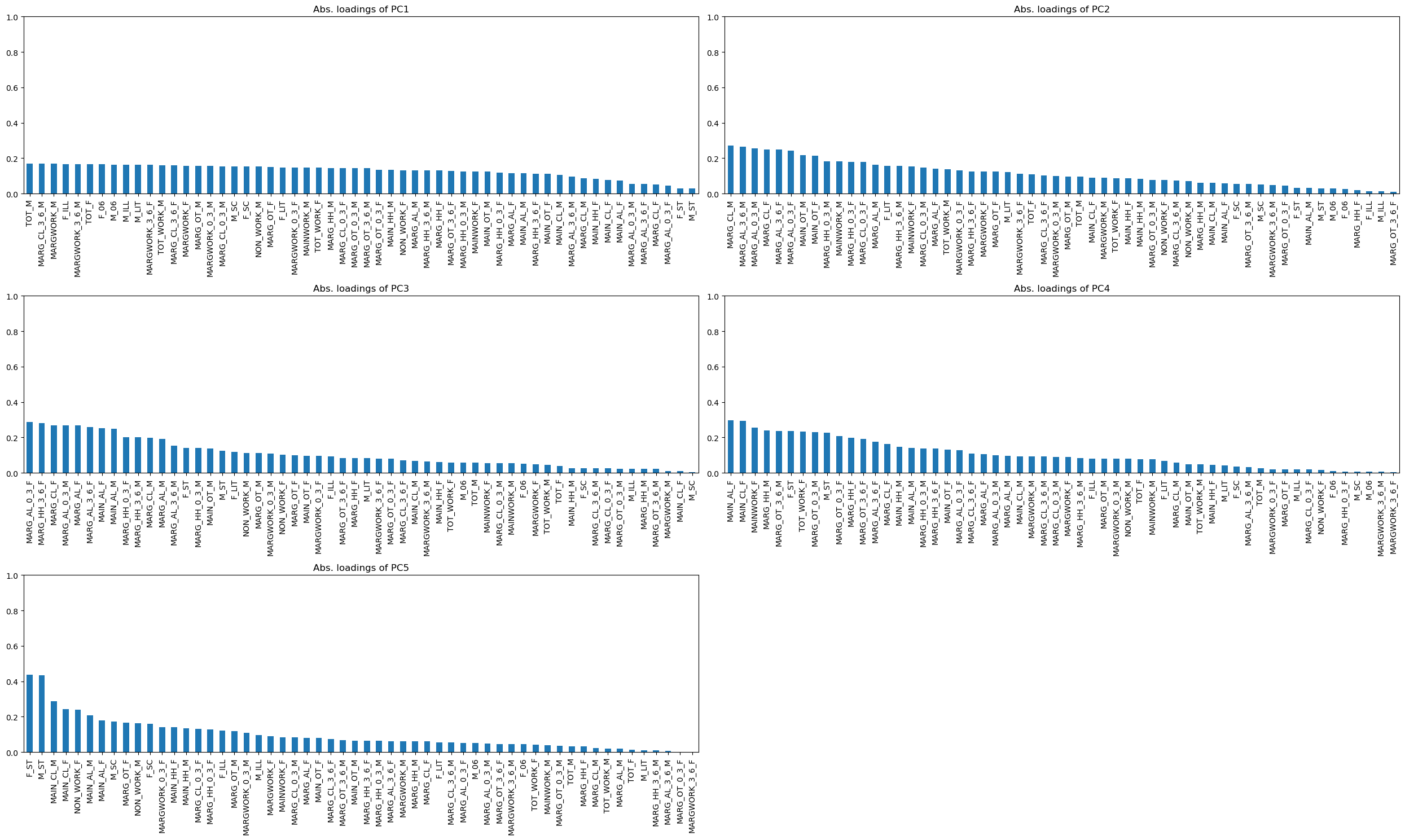


Fig: -45: Comparison

The above table shows the variables that contributes maximum towards the respective PCs:

## **Q2.8 PCA: Write linear equation for first PC.**

The below fig depicts the linear equation for the first PC (PC1):

The value in the Parentheses is the coefficient and those multiplied by variables.

