**DATA MINING PROJECT**

**Clustering and pca**

BUSINESS

REPORT

**Prepared by: AMAYIKA**

**Great Learning.**

## Email- Kashyap.amayika009@gmail.com

October 30th, 2022

Table of Contents

[Problem -1 – For The ads24x7 which is a Digital Marketing company, segment type of ads based on the features provided. Use Clustering procedure to segment ads into homogeneous groups. 5](#_bookmark0)

[1.1  Clustering: Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc. 6](#_bookmark1)

[1.2 Clustering: Treat missing values in CPC, CTR and CPM using the formula given. 7](#_bookmark2)

[1.3 Clustering: 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. 11](#_bookmark3)

[1.4 Clustering: Perform z-score scaling and discuss how it affects the speed of the algorithm. 14](#_bookmark4)

[1.5 Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance. 15](#_bookmark5)

[1.6 Clustering: Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm 18](#_bookmark6)

[1.7 Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters. 21](#_bookmark11)

[1.8 Clustering: 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]. 22](#_bookmark12)

[1.9 Clustering: Conclude the project by providing summary of your learnings. 29](#_bookmark14)

[Problem -2 Primary census abstract for female headed households. perform detailed EDA and identify Optimum Principal Components that explains the most variance in data.  31](#_bookmark18)

[2.1  PCA: Read the data and perform basic checks like checking head, info, summary, nulls, & duplicates, etc. 33](#_bookmark21)

[2.2 PCA: Perform detailed Exploratory analysis by creating certain questions like (i) Which state has highest gender ratio and which has the lowest?  37](#_bookmark23)

[2.2 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 39](#_bookmark26)

[2.3  PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary? 41](#_bookmark28)

[2.4  PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment. 45](#_bookmark18)

2.5 PCA: Perform all the required steps for PCA. Create the covariance Matrix Get eigen values and eigen vector. [48](#_bookmark23)

2.6 PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot. [54](#_bookmark26)

[2.7 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. 56](#_bookmark27)

2.8 PCA: Write linear equation for first PC. [62](#_bookmark23)

List of Tables

[Table 1. Data set of Digital marketing Company – first -5 Rows 6](#_bookmark7)

[Table 2. Data set of Digital marketing Company – last -5 Rows 7](#_bookmark8)

[Table 3. Information of datatype – Digital marketing company 8](#_bookmark9)

[Table 4. Description of datatype – Digital Marketing Company 8](#_bookmark10)

[Table 5. Missing values Table for digital marketing company 9](#_bookmark13)

[Table 6. Null values table after correction for Digital marketing company 10](#_bookmark15)

[Table 7. Description of data set of Company after null values computation 11](#_bookmark16)

[Table 8. Only continuous variables in the dataset 13](#_bookmark17)

[Table 9. Scaled Dataset 14](#_bookmark19)

[Table 10. WSS values for k=1 to 10 15](#_bookmark20)

[Table 11. Silhouette scores for 10 clusters 17](#_bookmark22)

[Table 12. Adding Cluster’s column in the original dataset (First Rows) 18](#_bookmark25)

[Table 13. Adding Cluster’s column in the original dataset (Last Rows) 18](#_bookmark7)

[Table 14. Frequency of Clusters 20](#_bookmark8)

[Table 15. Descriptions of data set after Clusters added 21](#_bookmark9)

[Table 16. Cluster Profile – means of variables included 21](#_bookmark10)

[Table 17. Census Data set 31](#_bookmark13)

[Table 18. Census data set continued 31](#_bookmark15)

[Table 19. Transposed data set of census 32](#_bookmark16)

[Table 20. Description of data set 34](#_bookmark17)

[Table 21. Information of data set 35](#_bookmark19)

[Table 22. Null values of the data set 35](#_bookmark20)

[Table 23. Maximum Gender ratio 38](#_bookmark22)

[Table 24. Minimum Gender ratio 38](#_bookmark25)

[Table 25. Value counts of States and Areas 45](#_bookmark7)

[Table 26. Transformed data set 53](#_bookmark8)

[Table 27. Covariance matrix 57](#_bookmark9)

[Table 28. PCA arrays 57](#_bookmark9)

[Table 29. Eigen vectors 55](#_bookmark10)

[Table 30. Eigen values 56](#_bookmark13)

[Table 31. Explained variance ratio 56](#_bookmark15)

[Table 32. Coefficients of all PCs along with the upper 5 row variables 57](#_bookmark16)

[Table 33. Selected PCs Data frame 58](#_bookmark17)

[Table 34. Data frame of fit transformed scaled data of principal components 62](#_bookmark19)

List of Equations and Formulas

[Equation 1. CPM Formula 10](#_bookmark8)

[Equation 2. CPC Formula 10](#_bookmark9)

[Equation 3. CTC Formula 10](#_bookmark10)

[Equation 4. Gender Raatio Formula 38](#_bookmark13)

[Equation 5. Total Variance in the data 55](#_bookmark15)

[Equation 6. Explained variance 56](#_bookmark16)

[Equation 7. Linear Equation of PCs 63](#_bookmark17)

List of Figures

[Figure 1. Pie Chart of null values in the data set 9](#_bookmark9)

[Figure 2. Boxplots of all the variables together. 12](#_bookmark10)

[Figure 3. Boxplots of all the continuous variables separately 13](#_bookmark13)

[Figure 4. Boxplots of all the continuous variables separately after outliers’ treatment 14](#_bookmark15)

[Figure 5. Dendrogram for the scaled data 15](#_bookmark16)

[Figure 6. Dendrogram for the data set for p =10 15](#_bookmark17)

[Figure 7. Elbow plot for K - means 16](#_bookmark19)

[Figure 8. Silhouette Score Plot 16](#_bookmark20)

[Figure 9. Scatter Plot of 6 Clusters 17](#_bookmark20)

[Figure 10. Clusters with CTR, CPM and CPC 23](#_bookmark16)

[Figure 11. Clusters with CTR, CPM and CPC 23](#_bookmark17)

[Figure 12. Bar chart of Clusters and Revenue with Device type 24](#_bookmark19)

[Figure 13. Pie chart of clusters with revenue 25](#_bookmark20)

[Figure 14. Bar graph of clusters & device type with clicks 26](#_bookmark22)

[Figure 15. Pie chart of clusters & device type with clicks 26](#_bookmark17)

[Figure 16. Bar graph of clusters & device type with spending 27](#_bookmark19)

[Figure 17. Bar Graph of Ad size with clusters and device type 28](#_bookmark20)

[Figure 18. Bar Graph of Impressions with clusters and device type 29](#_bookmark22)

[Figure 19. Boxplots and Density plots of variables 38](#_bookmark22)

[Figure 20. State wise distribution of male & female population sorted by male 39](#_bookmark17)

[Figure 21. State wise distribution of male & female population sorted by female 39](#_bookmark19)

[Figure 22. Male and Female literacy rate Different analysis 39](#_bookmark20)

[Figure 23. Male and Female literacy rate together 40](#_bookmark22)

[Figure 24. Total male SC population with States 40](#_bookmark17)

[Figure 25. Total female SC population with States 41](#_bookmark19)

[Figure 26. Total male ST population with States 41](#_bookmark20)

[Figure 27. Total female ST population with States 42](#_bookmark22)

[Figure 28. Total male Working population with States 43](#_bookmark22)

[Figure 29. Total female Working population with States 44](#_bookmark25)

[Figure 30. Frequency distribution of States and Areas (top 30) 47](#_bookmark22)

[Figure 31. Correlation plot for the census variables 48](#_bookmark25)

[Figure 32. Boxplots of all variables together – outliers check 49](#_bookmark22)

[Figure 33. Boxplots of all variables separately – outliers check 50](#_bookmark25)

[Figure 34. Boxplots after outliers’ treatment 51](#_bookmark22)

[Figure 35. Boxplots after outliers’ treatment separately 52](#_bookmark25)

[Figure 36. Boxplots after transformation of the data 54](#_bookmark25)

[Figure 37. Scree plot for PCA 57](#_bookmark25)

[Figure 38. Eigen values vs PCA 57](#_bookmark22)

[Figure 39. PCA vs Explained variance Ratio 59](#_bookmark25)

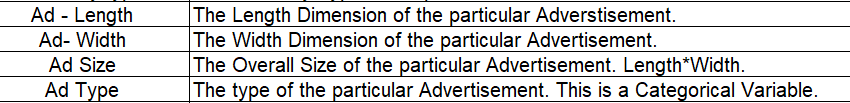
[Figure 40. Correlation Heat map with rectangular box 60](#_bookmark25)

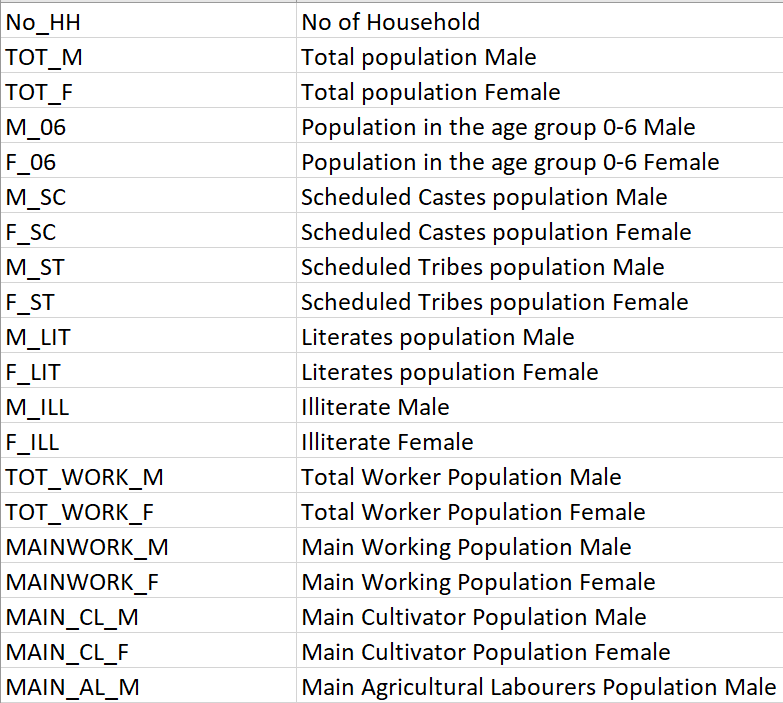
[Figure 40. Bar graphs of PCs with variables 61](#_bookmark25)

[Figure 40. Correlation Heat map with all variables and PCs 62](#_bookmark25)

[Figure 41. Correlation Heat map of final PCs 63](#_bookmark25)

List of Dictionary





# PROBLEM – 1

# Clustering – ****Digital Ads Data****

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.

We have transposed the rows as columns in most of the tables.

Table 1 – Data set of Digital Marketing Company (1st – 5 Rows)

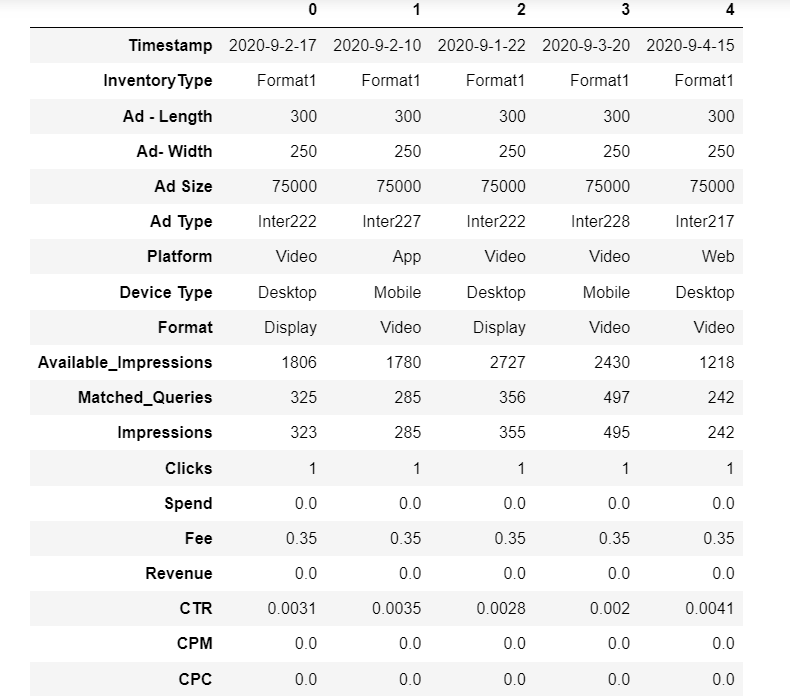
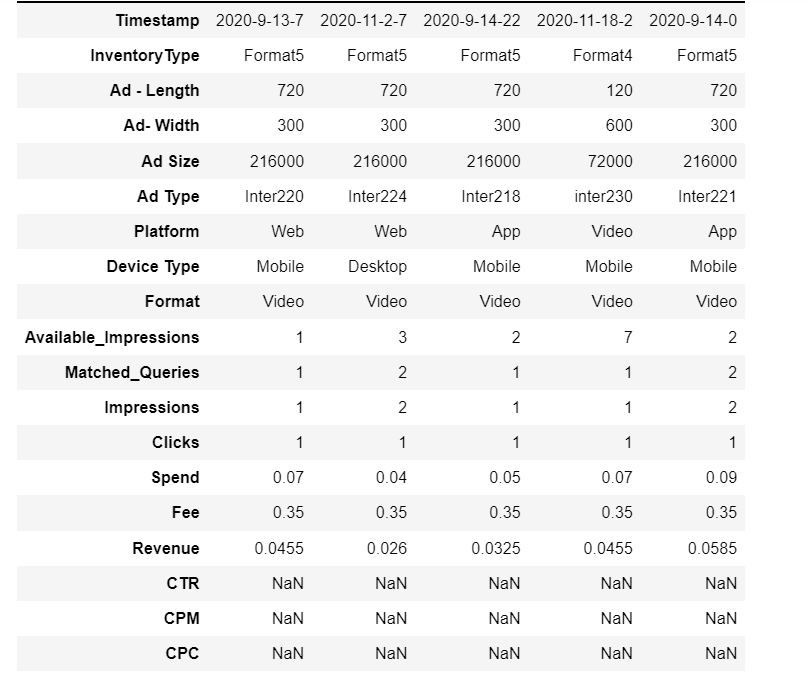


Table 2 – Data set of Digital Marketing Company (Last – 5 Rows)



**Part 1(1) – Clustering: Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.?**

#### After reading the data set, we observed that the data set contains 23066 rows and 19 columns. We used shape function to determine the rows and columns in the data set. Now we will find out about the information of the data set, their summary and description of various columns and their importance along with the null and duplicate values in our Data set. The info function shows the following information regarding the data set.

Our data set contains 6 Float data type, 6 object data type and 7 integer data type as mentioned in the below table. All the rows contains all 23066 values except for the columns CPC, CTR and CPM, which contains 18330 inputs.

Table 3 – Information of Digital Marketing Company

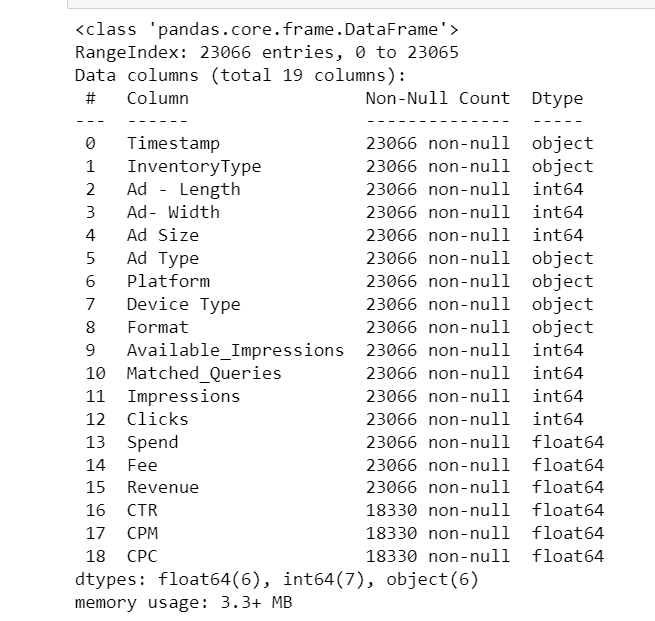


Table 4 – Description of Digital Marketing Company

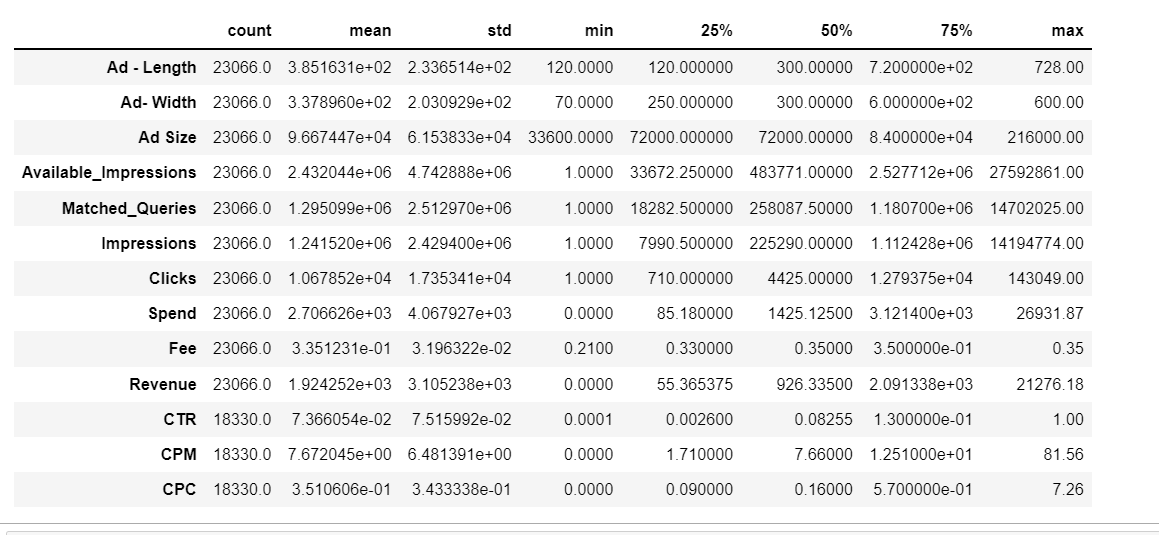
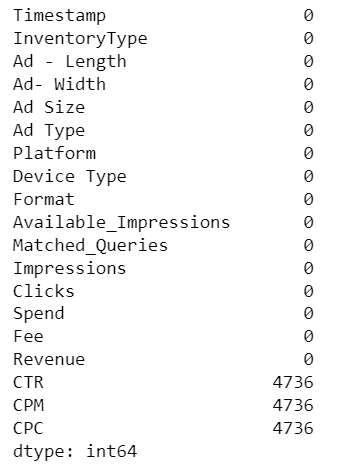
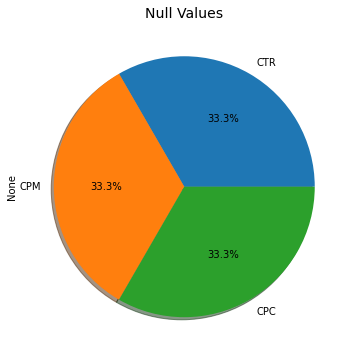


Table 5 – Missing values in the data set of Digital Marketing Company



So, when we checked the null values in the data set using “isnull” function it showed that all the three columns named CTR, CPM and CPC contains some missing values which is equal to 4736 for all the three columns. We can see a graphical representation of the null values as well using pie chart distribution.

 fig -1

Now checking the duplicate rows in the dataset is the next step and if we find any duplicate values, we will drop those columns else we will move forward with our further analysis. So, we used the corresponding duplicated function and it gives out that we have zero (0) duplicate rows in our dataset.

**Part 1 (2) -Treat missing values in CPC, CTR and CPM using the formula given?**

#### As we can see we have missing values in the dataset in the columns named CTR, CPM and CPC, which are already defined in the data dictionary. All the missing values are 4736 in each column mentioned.

#### So now we will treat those missing values according to the given formulas which are as follows:

#### CPM = (Total Campaign Spend / Number of Impressions) \* 1,000 – **Equation 1**

CPC = Total Cost / Number of Clicks *–* **Equation 2**

#### CTR = Total Measured Clicks / Total Measured Ad Impressions x 100. – **Equation 3**

Now using these definitions, we can replace the values which are missing in the given columns. We will use the lambda function and then apply those values to the corresponding columns and after computing those values we will check them by calling the last 5 rows.

Table 6 – Missing values in the data set of Digital Marketing Company

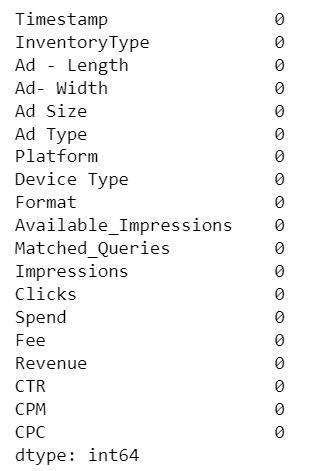
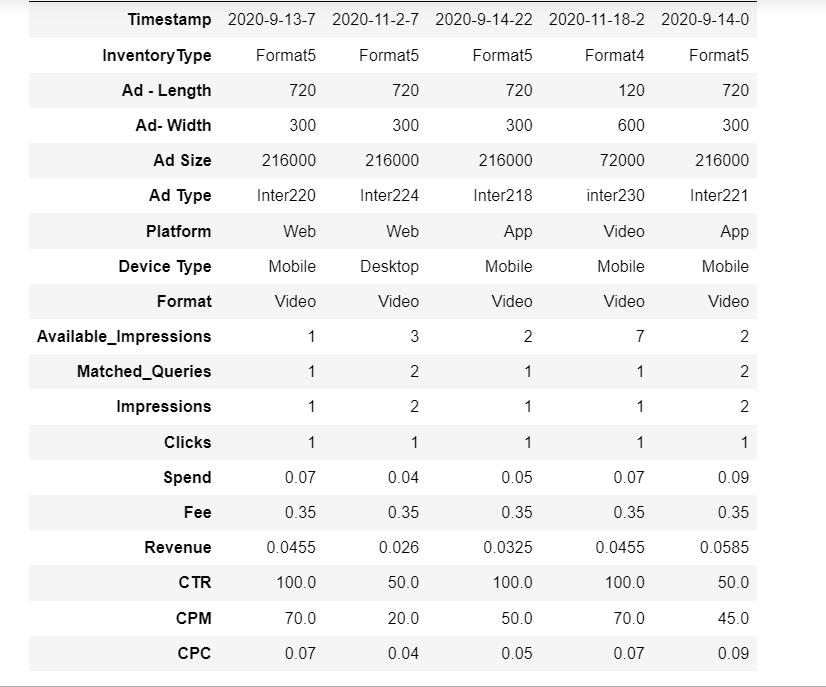


Table 7 – Description of data set of Company after null values computation

****

**Part 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?**

### Outliers checking and their Treatment- Now our next step is to check the outliers present in the data set as they are critical for modelling and may affect our analysis. So, we will treat the outliers or rather make their clusters differently in the dataset. For checking the outliers, we are plotting the boxplots for our continuous variables in the data set.

### fig -2

### Let’s make it more clear for each continuous variables, so that we can figure out the no of outliers, and treat them.

### fig -3

### Why outliers’ treatment is necessary and what are their impact on K-means clustering?

### Listing some of the major concerns related to outliers:

### 1. Outliers increases the mean and median of the data and hence, effects the central tendency. So, this is significant change as the K-mean algorithm is based on centroid and the distance between the observations, it does have an impact

### 2. Sometimes Outliers are large in number and hence can form a separate cluster altogether and can also cause other clusters to merge within their clusters, decreasing the efficiency of the method clustering

### 3. Also, in some cases real data typically has Noise signals within them which occur as outliers in the dataset. This noise can actually affect the clustering process and can vary the model based on that algorithm. This again reduces the efficiency of our modelling system

### Treatment of Outliers

### So, let’s treat the outliers using IQR method that is (Inter Quartile Range) method. To treat outliers, we first make a sub dataset from the original dataset which only contains the continuous datatypes.

### The value that is less than Q1 – 1.5 IQR or more than Q3 + 1.5 IQR is considered an outlier.

### Now we define a function ‘Remove outlier’ where we say that

### 1.The larger values which are greater than upper IQR range are replaced with the 95th percentile value of the given column.

### 2.The smaller values which are lesser than lower IQR range are replaced with the 5th percentile value of the given column.

### Selecting only continuous variables in the dataset, hence, making a new sub data frame from the original data frame which is to be named as “data\_cont”. These are our target variables which we actually need to focus while clustering.

Table 8 – Only continuous variables in the dataset

### 

### fig -4

**Part 1(4) – Perform z-score scaling and discuss how it affects the speed of the algorithm?**

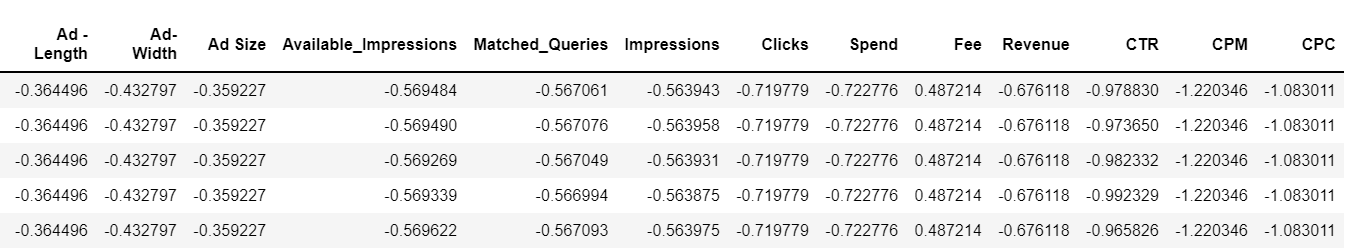
1. Scaling is required because we have different independent variables having vast range of values and having different units such as we can have data containing Distance which can be kilometres and can range from 1 to 1000s, and then we can have weight in Kgs which can range from 1 to 100. So, comparison of these type of values together and their clustering is impossible unless and until they are scaled together having similar range.

​

​2. Also if the data is not scaled it will affect the machine learning algorithm which is highly sensitive to range and scaling. Hence, scaling the data and bringing them in similar range makes the clustering process faster and increases the efficiency of our model

For this we will first import a package Standard Scaler which will help us in scaling the data. We used scikit-learn’s StandardScaler to perform z-score scaling.

Table 9 – Scaled dataset



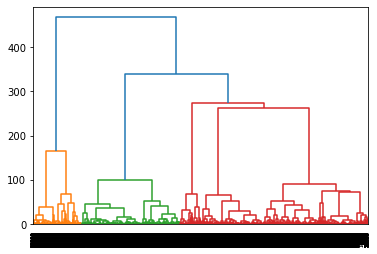
**Part 1(5) -Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance?**

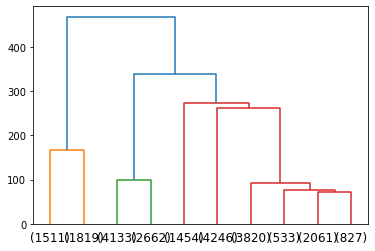
### Let’s import the required package for hierarchical clustering and dendrogram.

Using SciPy’s cluster hierarchy function, we will create the required dendrogram for the scaled data set. This

dendrogram will help us in analysing the number of clusters that are suitable for the given data.

Figure shows Dendrogram using WARD and Euclidean distance

fig -5

 fig -6

A dendrogram is a visual representation of cluster-making. On the x-axis are the item names or item numbers. On the y-axis is the distance or height. The vertical straight lines denote the height where two items or two clusters combine. The higher the level of combining, the distant the individual items or clusters are. By definition of hierarchical clustering, all items must combine to make one cluster.

We observe the dendrogram and we draw a virtual cut-off line through it depending on our range and the number of vertical lines that are cutting the cut-off line, we assume that many clusters to be efficient.

* Now taking the above dendrogram as the reference, 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 feasible for the data set which is this large and it will decrease the efficiency of our model.
* However, if, we draw a cut-off horizontal line at a range somewhere above 200 and below 300 we will have 5 clusters in that case. Alternatively, there may be **3 clusters** as well, based on the colour code in the dendrogram.
* Depending on the size of data set choosing **5 Clusters** using Dendrogram for this project will be more feasible

# Part 1(6) – Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.

# K-means clustering

# k-means clustering is the most used non-hierarchical clustering technique. It aims to partition n observations into k clusters in which each observation belongs to the cluster whose mean (centroid) is nearest to it, serving as a prototype of the cluster. It minimizes within-cluster variances (squared Euclidean distances).

# For this we will first import the required package

# K- mean Algorithm steps:

# • Step 1. Select the first item from the list. This item forms the centroid of the first cluster.

# • Step 2. Search through the subsequent items until an item is found that is at least distance δ away from any previously defined cluster centroid. This item will form the centroid of the next cluster.

# • Step 3: Step 2 is repeated until all k cluster centroids are obtained or no further items can be assigned.

# • Step 4: The initial clusters are obtained by assigning items to the nearest cluster centroids.

# So now we will be Forming clusters with K = 1,2,3,4,5,6,7,8,9,10 and comparing the WSS which is the within sum of squares values of each one of them. WSS reduces as K keeps increasing

# Calculating WSS for other values of K – Elbow Method-

Table 10 – WSS values for k=1 to 10

# 

# Although we see sudden drop till 3 points and again till 5 points it totally depend on our assumption what we chose as optimal number of clusters. So, to be surer let’s use ELBOW method to find out the drop.

# Elbow method:

# There are many methods that are recommended for determination of an optimal number of partitions. Unfortunately, however, there is no closed form solution to the problem of determining k. The choice is somewhat subjective and graphical methods are often employed.

# For a given number of clusters, the total within-cluster sum of squares (WCSS) is computed. That value of k is chosen to be optimum, where addition of one more cluster does not lower the value of total WCSS appreciably.

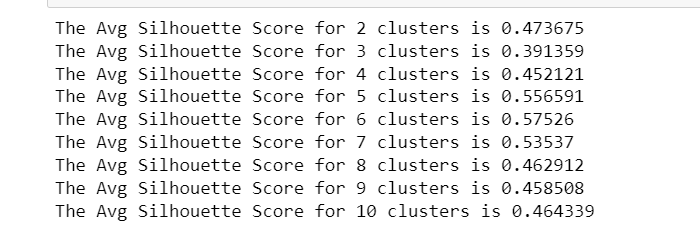
# The Elbow method looks at the total WCSS as a function of the number of clusters.

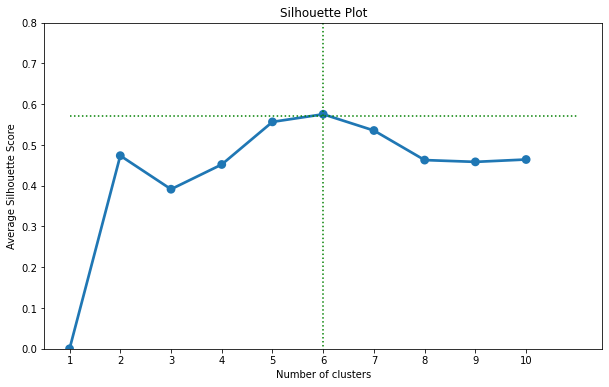
# fig -7

# Hence, we can see a bend till k =5 after that the values and line can be seen as almost constant or not very deflecting. So, we can say that optimal number of clusters can be 5 or 6. We will confirm this using silhouette scores further.

**Part 1(7) -Print silhouette scores for up to 10 clusters and identify optimum number of clusters?**

Table 11 – Silhouette scores for 10 clusters

****

**** fig -8

So, we can see from the ‘Silhouette Plot’ **that k= 6 has the highest no of score as 0.575**, which is slightly higher than that of silhouette score for k = 5 which is around 0.55. so, we can say that the optimal **number of clusters should be 6 according to silhouette scores.**

**Part 1(8) -Profile the ads based on optimum number of clusters using silhouette score and your domain understanding. 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?**

Importing fcluster module to create clusters.

Now since we have done with clustering the observations, we will add these clusters in the original data set as a separate column so that each observation represents its cluster also. We can check these cluster numbers using the upper 5 and lower 5 rows of the data frame using head and tail function.

Table 12 – Adding cluster’s column in Original Dataset (First Rows)

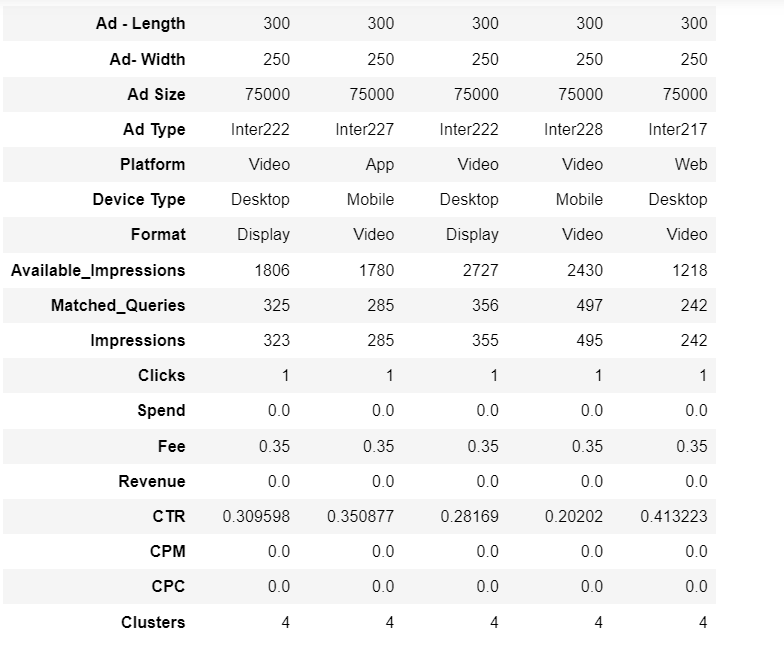
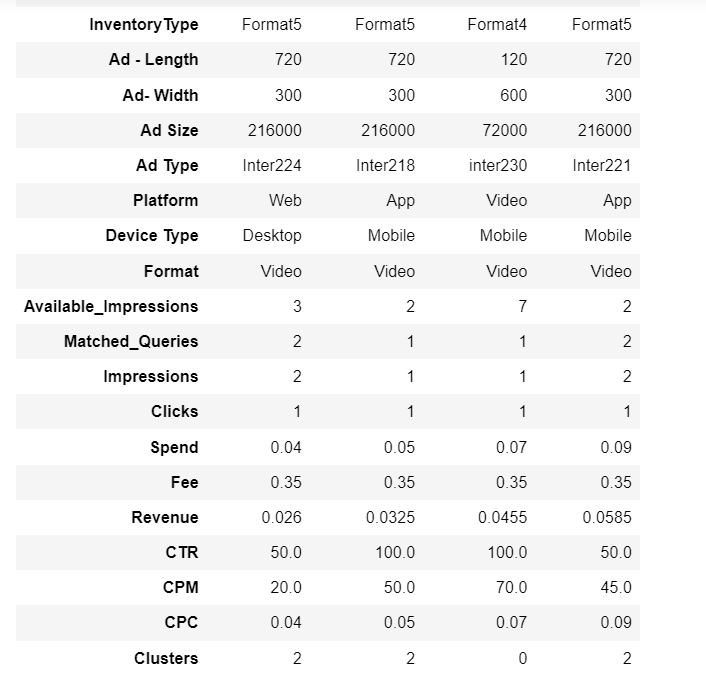
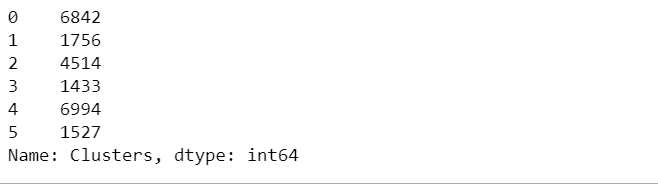


Table 13 – Adding cluster’s column in Original Dataset (Last Rows)



**Cluster Profiling**

Table 14 – Frequency of Clusters

****

After this we will do the profiling of clusters and find out the frequency of each cluster in the given data frame. Here we can observe the following:

1. Cluster 4 has maximum number of observations in it which is equal to 6994 observations.

2. The least number of observations are present in 3rd cluster and is equal to 1433.

Table 15 – Descriptions of data set after Clusters added

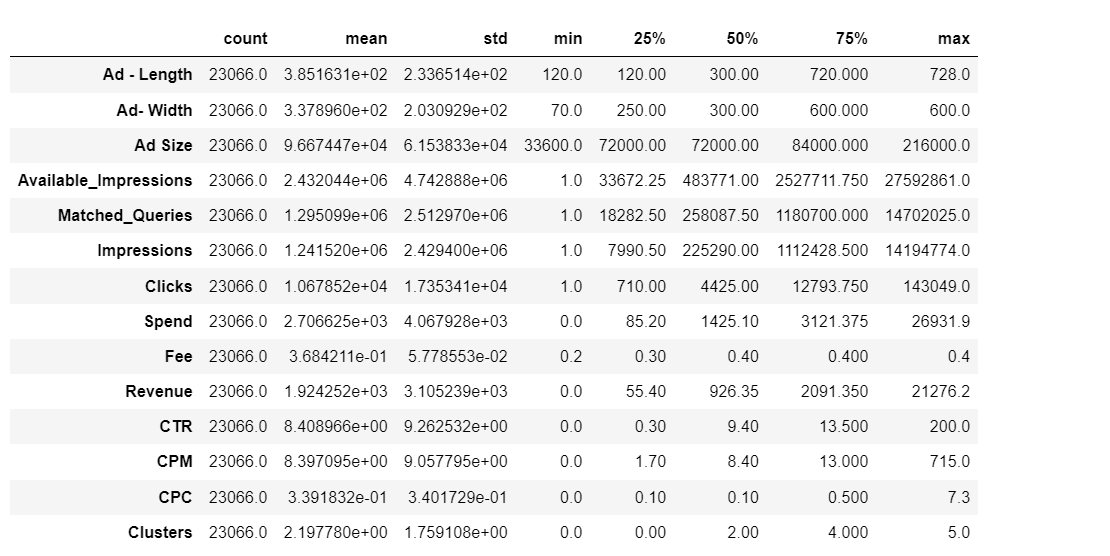
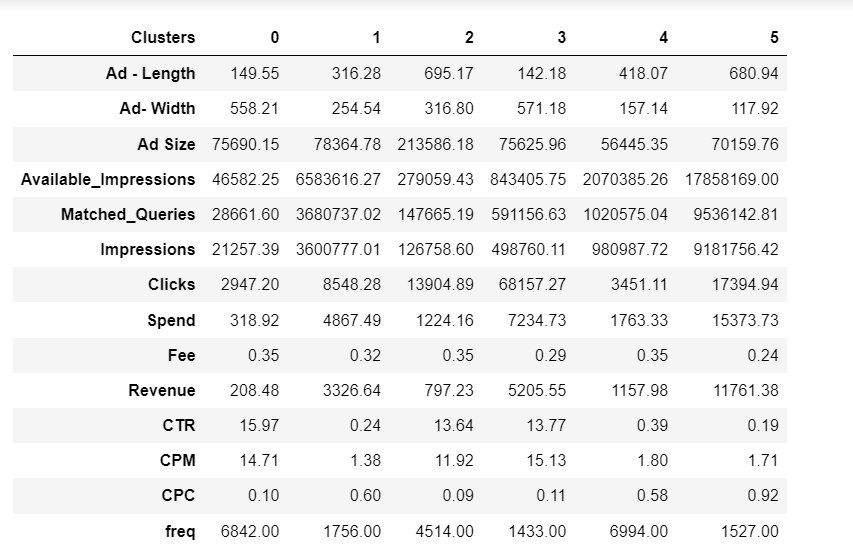


Table 16 – Cluster Profiles: Means of the variables included



### The description after finally adding the clusters can be observed and various kind of inferences can be made from the table above. Let’s see some of the obvious observations and then we will try to plot a scatter plot of the different clusters for each observation.

### The observations are as follows:

### 1. The mean Ad Length is highest for cluster 2nd which is around 695 followed by cluster 5th and 4th as second and third spot respectively

### 2. The mean Ad Width is highest for observations in cluster 0

### 3. The highest mean Ad size is for observations in cluster 2nd followed by cluster 1st. So, length and width of particular Advertisement together contribute into the size of Advertisement altogether.

### 4. An impression is counted each time an Advertisement is shown on a search result page or other site on a Network. The mean of impressions is highest for cluster 5th as we can see the mean of spend which is the amount of money spent on specific ad variations within a specific campaign or ad set is also highest for cluster 5th followed by cluster 1st.

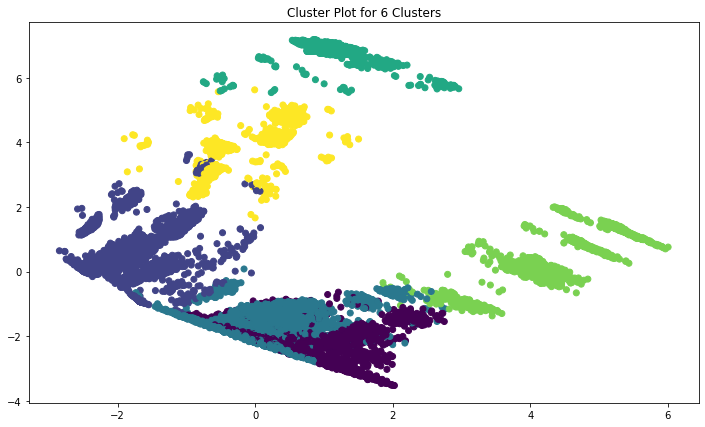
### 5. Although the mean of Clicks which is a marketing metric that counts the number of times users have clicked on the particular advertisement to reach an online property is highest for cluster 3rd, which gives us an idea that click is not only dependent on the impressions and amount spent on campaign. It has other factors to add on too.

### 6. The maximum revenue is generated from cluster 5th as the mean is highest followed by cluster 1st.

### 7. 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. The mean of CTR is highest for cluster 0

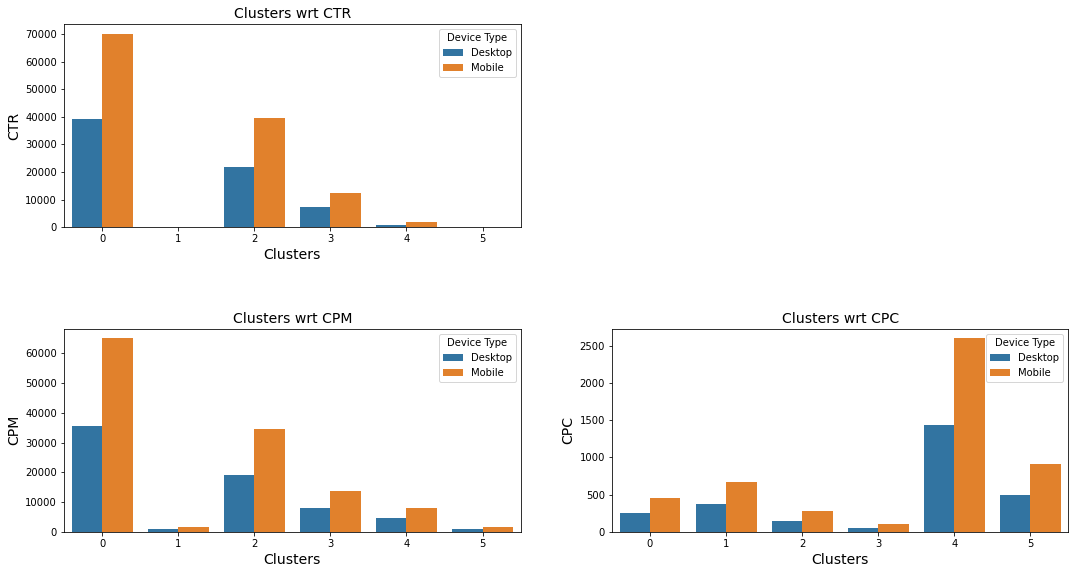
### 8. Similarly, CPM which stands for “cost per 1000 impressions is highest for cluster 3rd and CPC is highest for cluster 5th.

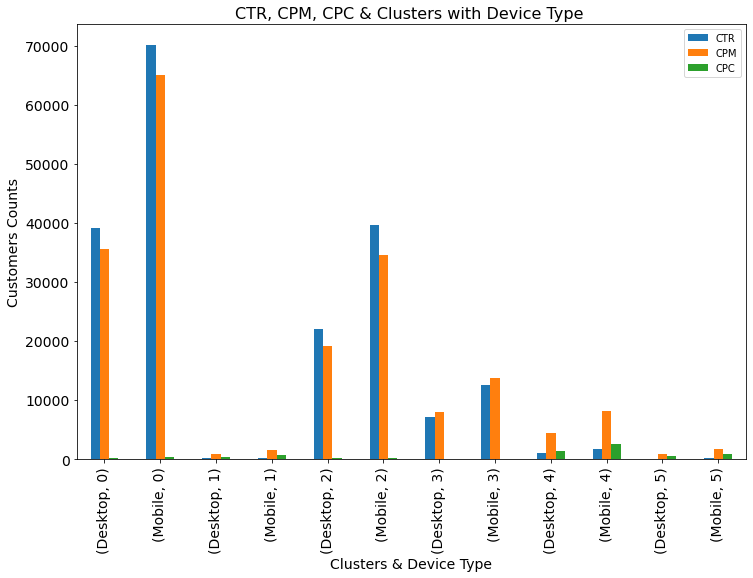
### Scatter plot for 6 clusters

**** fig -9

The scatter plot shown in the graph shows that all the clusters are somewhat different but two of the clusters are a bit overlapping or we can say that are similar to each other.

**CPM, CTR, & CPC based on Device Type and Their Clusters using bar plots and analysis based on that result.**

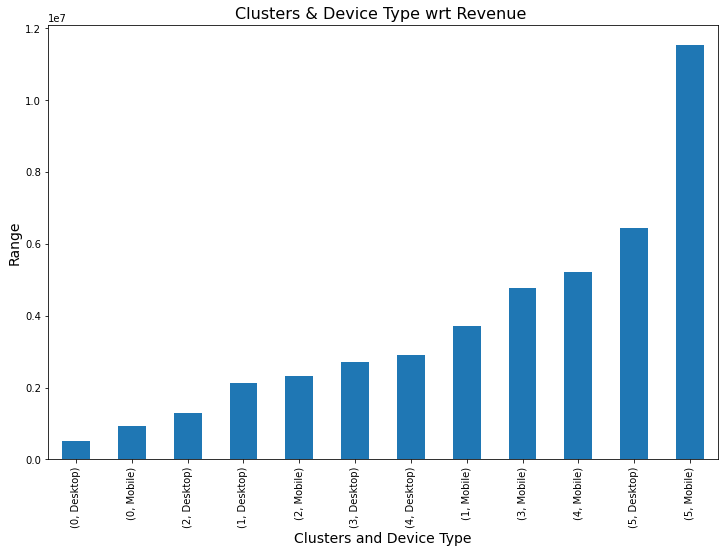
 fig -10

 fig -11

Observations based on CTR, CPM & CPC:

* CPC means Cost Per Click. That implies the highest amount that you’re willing to pay for a click on your Ad. It is highest for Mobile segment of Cluster 4, followed by Desktop segment of cluster 4.
* Cluster 5 has the second highest CPC after cluster 4
* The least CPC is provided by cluster 3
* CPM stands for “cost per 1000 impressions”. A way to bid where you pay per one thousand views (impressions) on the Google Display Network. When it comes to CPM, mobile segment of cluster 0 has the highest value, second being the desktop segment of Cluster 0. Cluster 2 is next to cluster 0 in case of CPM. Cluster 1 has the least value of CPM.
* CPC which means TR Click Through Rate. CTR is the number of clicks that your Ad receives divided by the number of times your Ad is shown: clicks ÷ impressions = CTR. CTR is highest for cluster o in the segment mobile followed by segment desktop in the same cluster.
* The least CTR is shown by cluster 1 and 5 in the desktop segment.

# Revenue based on Device Type and Their Clusters using bar plots and pie charts and analysis based on that result.

 fig -12

# fig -13

# Observations based on revenue generation:

# Cluster 5 has the highest revenue generated from mobile segment which is around 26% of the total, followed by desktop segment of cluster 5 having a percentage equal to 14.5%. This shows that cluster 5 is generating quality Ads and spending much more than any other clusters (as we will observe in the spending graph below). Also, their Ad sizes are not that much bigger, showcasing the fact that their main agenda is Quality over Quantity of Ads.

# Least revenue is generated from cluster 0 in both the segments, where device type “Desktop” is least of the two.

# Cluster 1, 2 and 3 are somewhat close to each other when it comes to revenue generated from the Ads.

# Clicks based on Device Type and Their Clusters using bar plots and pie charts and analysis based on that result.

# fig -14

# fig -15

# Observation based on clicks:

# When we considered Clusters and device types together with respect to clicks, we observed the following from the bar graph and pie chart presented above:

# Mobile segment of Cluster 3 has the highest clicks almost around 25.3% of the total, followed by the mobile segment of cluster 2 with 16% of the total clicks.

# In case of device type “desktop”, cluster 3 holds the highest position with 14.4% amount.

# Desktop overall shows less clicks than the mobile segments, showing that its more user friendly for people to click Ads on mobile. Rather Desktops are mostly used for much professional purposes.

# The least clicks are observed in cluster 0 and 1, in desktop segments

# Clusters and device type with respect to spending:

# fig -16

# Observations:

# The above two graphs tells us about the spending on various clusters with respect to various device types. We can observe the followings:

# In cluster 5, the device type “mobile” shows the highest total spending which means they are spending on some authentic quality of advertising, followed by “desktop” in cluster 5, although their Ad size is not the highest. This explains that cluster 5 is going for quality over quantity.

# In case of Mobile segments clusters 3 and 4 shows the highest spending after cluster 5th respectively, whereas the least spending is done by cluster 0 on both device types “mobile” as well as “desktop”, although their Ad size is quite prominent and is bigger than cluster 5 ad sizes.

# Cluster 2nd is having the highest Ad size for device type “mobile” followed by type “desktop”

# fig -17

# fig -18

### Observations based on Ad size and Impressions in various clusters:

### Highest impressions are observed from the mobile and desktop segment of cluster 5 followed by cluster 4 and 1 respectively, although their Ad size is in between highest and lowest. This shows that Ad size and spending is not the only factors to get impressions or revenue, people tend to go for premium Ads which has quality in itself.

### The least impressions are from cluster 0 although their spendings are good enough, as compared to other clusters.

### Part 1(9) – Conclude the project by providing summary of your learnings.?

# Conclusion:

# In this project we learned so many techniques and understood the idea of clustering the data set, by scaling and treating the outliers in the data sets. We treated the missing values in the data set based on the formulas provided and based on our results we have some observations which are as follows:

# CPC means Cost Per Click. That implies the highest amount that you’re willing to pay for a click on your Ad. It is highest for Mobile segment of Cluster 4, followed by Desktop segment of cluster 4. Cluster 5 has the second highest CPC after cluster 4. The least CPC is provided by cluster 3

# 2. CPM stands for “cost per 1000 impressions”. A way to bid where you pay per one thousand views (impressions) on the Google Display Network. When it comes to CPM, mobile segment of cluster 0 has the highest value, second being the desktop segment of Cluster 0. Cluster 2 is next to cluster 0 in case of CPM. Cluster 1 has the least value of CPM.

# 3. CPC which means TR Click Through Rate. CTR is the number of clicks that your Ad receives divided by the number of times your Ad is shown: clicks ÷ impressions = CTR. CTR is highest for cluster o in the segment mobile followed by segment desktop in the same cluster. The least CTR is shown by cluster 1 and 5 in the desktop segment.

# 4. Cluster 5 has the highest revenue generated from mobile segment which is around 26% of the total, followed by desktop segment of cluster 5 having a percentage equal to 14.5%. This shows that cluster 5 is generating quality Ads and spending much more than any other clusters (as we will observe in the spending graph below). Also, their Ad sizes are not that much bigger, showcasing the fact that their main agenda is Quality over Quantity of Ads.

# 6. Least revenue is generated from cluster 0 in both the segments, where device type “Desktop” is least of the two. Cluster 1 2 and 3 are somewhat close to each other when it comes to revenue generated from the Ads.

# 7. In cluster 5, the device type “mobile” shows the highest total spending which means they are spending on some authentic quality of advertising, followed by “desktop” in cluster 5, although their Ad size is not the highest. This explains that cluster 5 is going for quality over quantity.

# 8. In case of Mobile segments clusters 3 and 4 shows the highest spending after cluster 5th respectively, whereas the least spending is done by cluster 0 on both device types “mobile” as well as “desktop”, although their Ad size is quite prominent and is bigger than cluster 5 ad sizes. Cluster 2nd is having the highest Ad size for device type “mobile” followed by type “desktop”

# PROBLEM STATEMENT– 2

# PCA FH (FT): Primary census abstract for female headed households excluding institutional households (India & States/UT’s – 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.

## **About Data**

# Data file contains information about the census and its history in various areas of country having different states and UT’s.

# All important details regarding female headed households excluding institutional households, scheduled tribes -2011, Female Headed Household Excluding Institutional Household, Marginal Agriculture Laboure’s Population 3-6 Female.

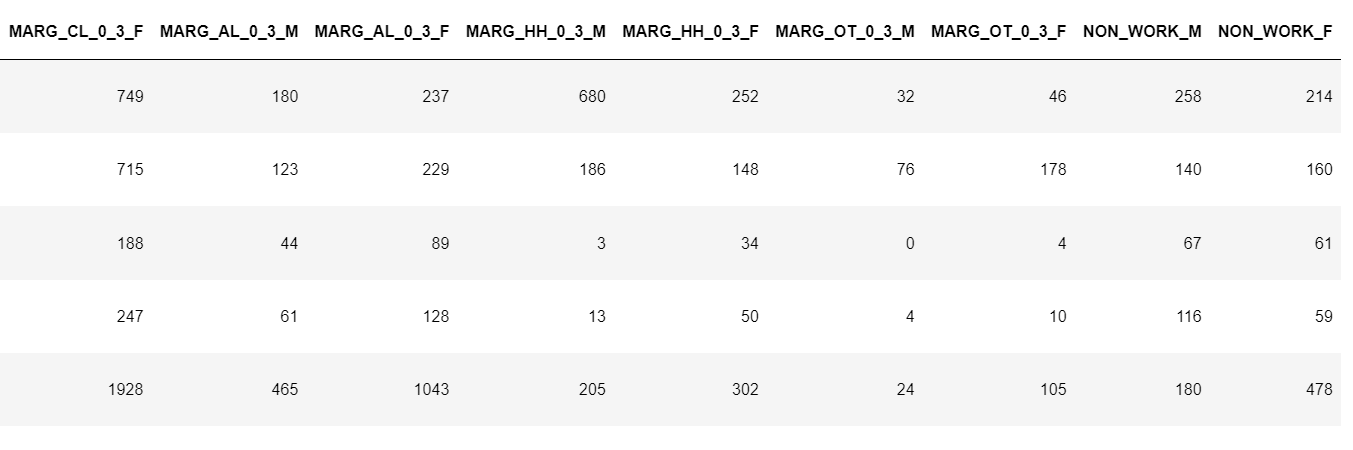
# Marginal Household Industries Population Male & Female, Marginal Other Workers Population Person 3-6 Male & Female, Marginal Worker Population 0-3 Male & Female, Marginal Cultivator Population 0-3 Male & Female, Marginal Agriculture Laboure’s Population 0-3 Male & Female, Marginal Household Industries Population 0-3 Male & Female, Marginal Other Workers Population 0-3 Male & Female, Non-Working Population Male & Female etc. has been provided.

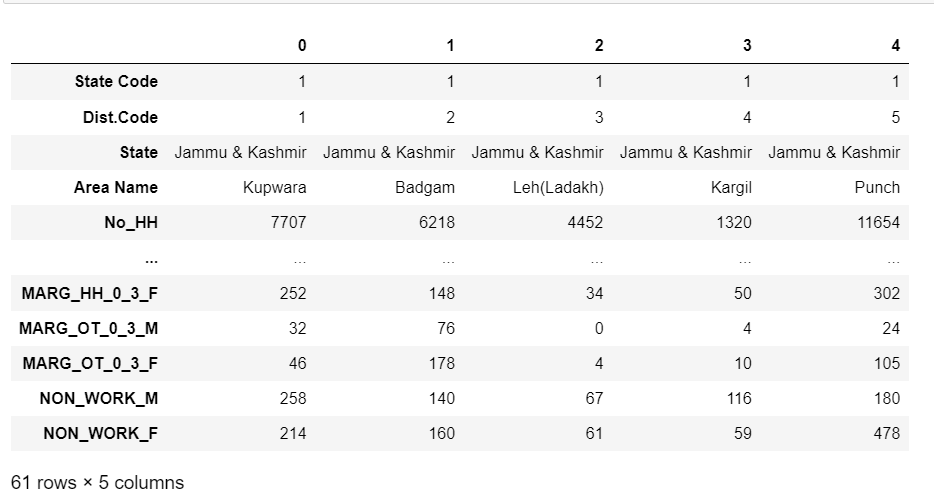
# Using all these features that corresponds as an important element for conducting a successful census in the country, we will perform Principal Component Analysis (PCA) and perform detailed EDA and identify Optimum Principal Components that explains the most variance in data.

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

# Going through the dataset – Reading the top 5 rows

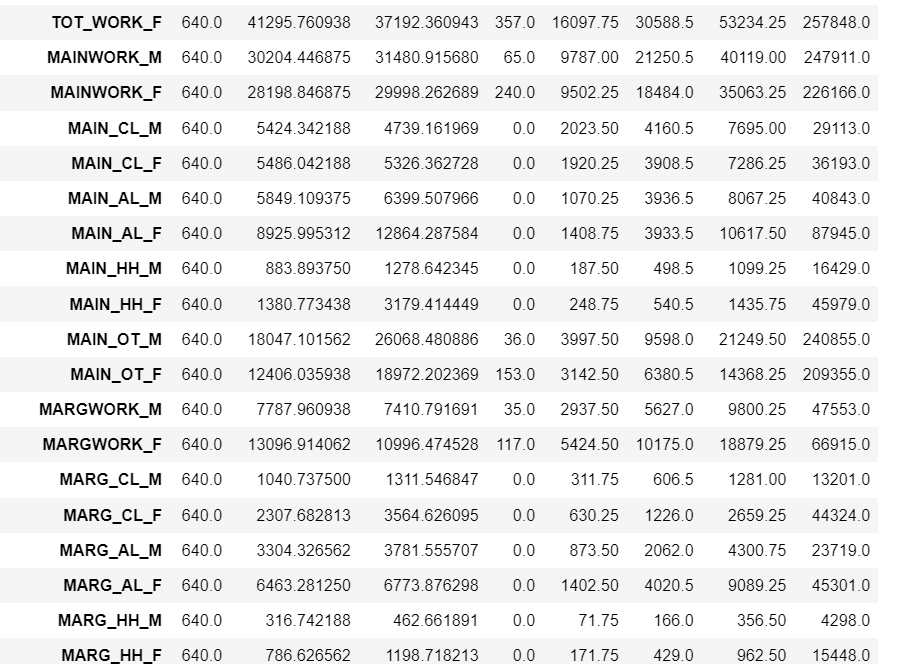
# Table-17-Census Data Set

Table-18, Census Data set continued

Table -19, Transposed data set

**Description of the Data set:**







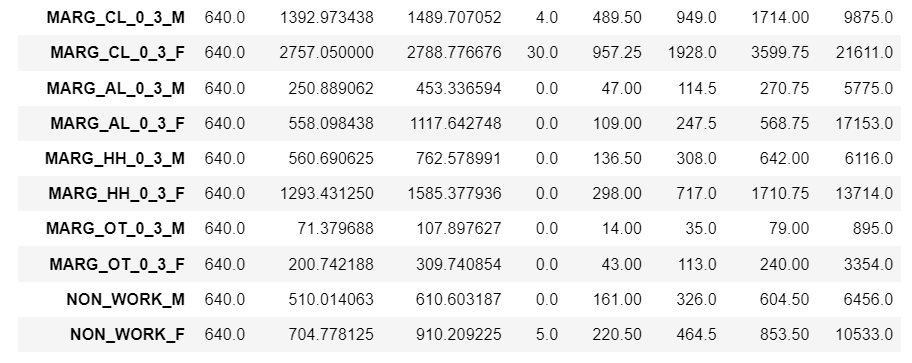


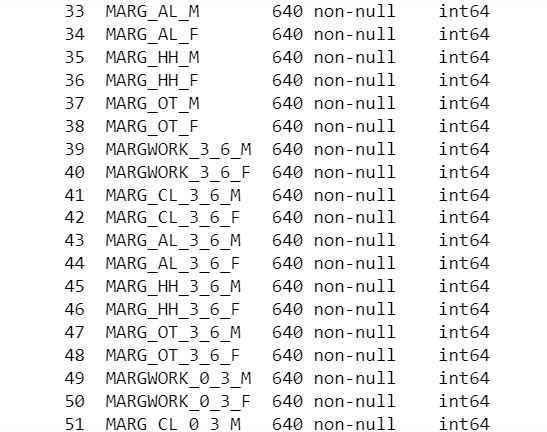
Table -20, Description of data set.

**Information and Null Values counts of the complete dataset**:

Now let’s check the null and duplicate values of the given sample. If null values or duplicate data are present in the sample data set, we will treat them and if not, we will further proceed with the outlier’s detection and their treatment.







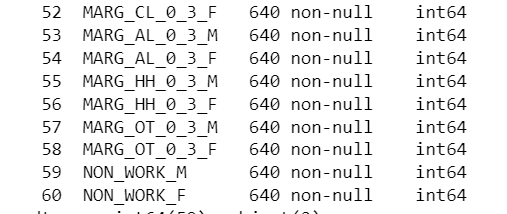
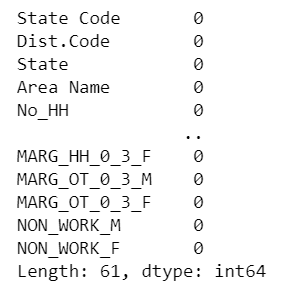


Table-21, Information of census

 Table-22, Null Values count of census

Also, the duplicate value count in the data set is also zero. So, the next step for us to do is checking the outliers in the given variables of census data. This will be done using the analysis of histograms and boxplots of the variables.

**Observations:**

- Data consists of 640 area details with 61 features

- Data has 35 unique States (including UT’s) and 635 unique Areas in these States and UT’s

- Mean of Total population Female is more than mean of Total population Male in the given data set.

- Uttar Pradesh = 71 and Madhya Pradesh =50 has the highest and second highest data representation in the given census data set and Lakshadweep, Chandigarh and Dadar & Nagar Haveli having the least numbers which is equal to 1.

- The census data of Female headed households excluding institutional households shows that the mean literacy of females is more than that of literacy of males in such households.

- “State code” and “district code” are numerical variables but does not hold much of importance for the census analysis, so changing them into categorical variables or we will drop them for further analysis.

- There are no missing values in the data

- We have 2 categorical fields “State” and ‘Area Name’. We are dropping the other two ‘district code’ and ‘state code’ since they are not important for PCA.

- For features like Main Cultivator Population Male ‘Main Cultivator Population Female’, ‘Main Agricultural Labourers Population Male’, ‘Main Agricultural Labourers Population Female’, ‘Main Household Industries Population Male’, ‘Main Household Industries Population Female’ the minimum number in the households is equal to 0.

### Part 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?

**EDA Exploratory Data Analysis**

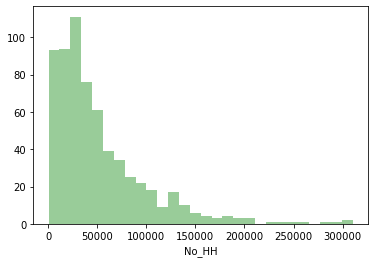
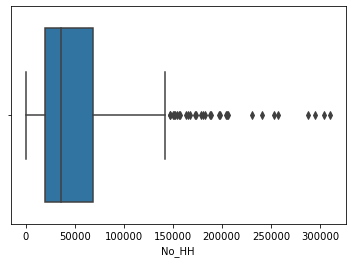
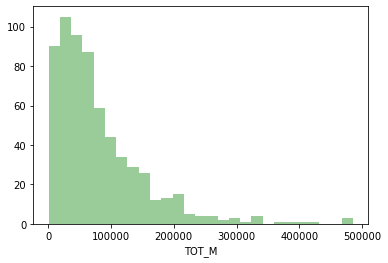
**Univariate Analysis**

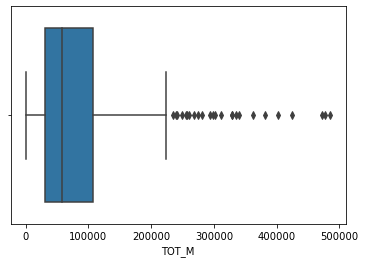
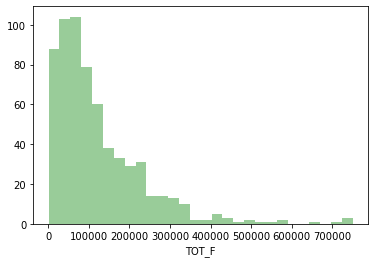
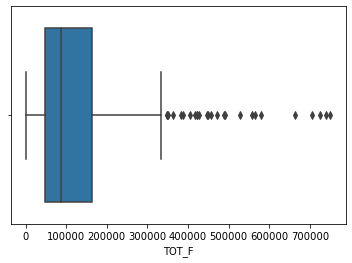
Let us define a function “univariateAnalysis\_numeric” to display information as part of univariate analysis of numeric variables. The function will accept column name and number of bins as arguments.

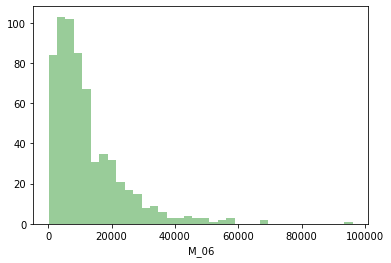
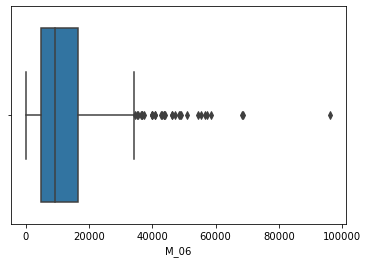
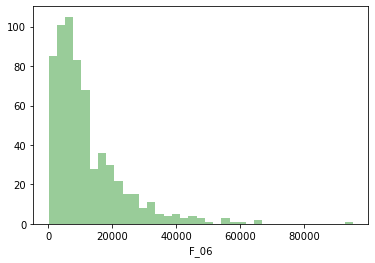
The function will display the statistical description of the numeric variable, histogram or density plot to view the distribution and the box plot to view 5-point summary and outliers if any.

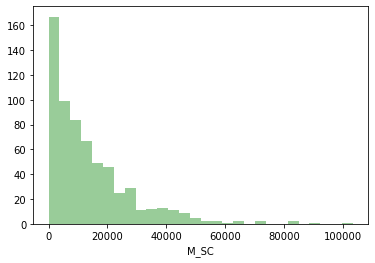
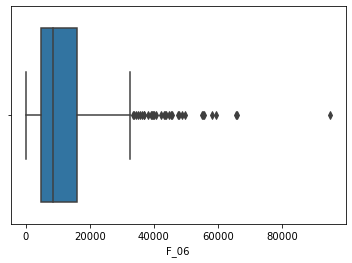
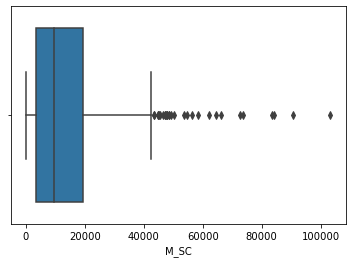
We have created a new subset of the original data frame named as “data numerical”, which only includes continuous variables and from now on we will perform the univariate analysis on this data set.

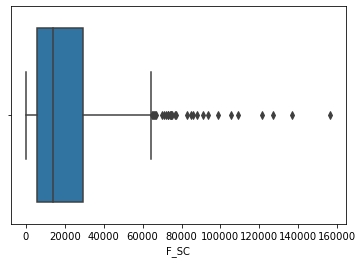
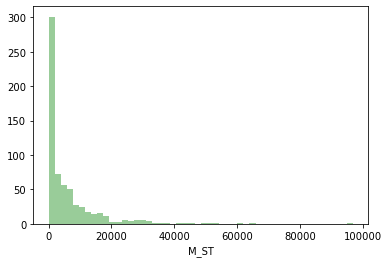
**Plotting the histogram and boxplots of “All the numerical variables together”**

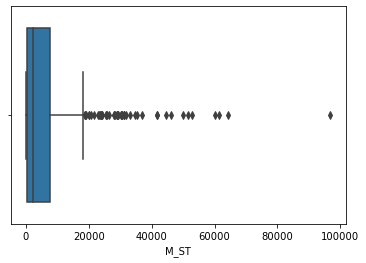
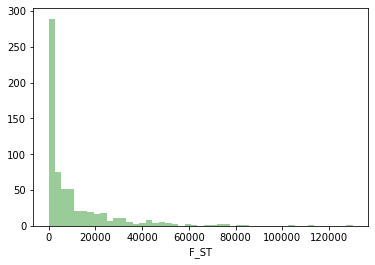
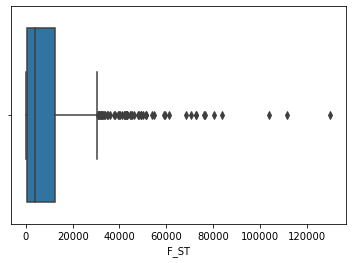
  

fig-19, BOXPLOTS AND DENSITY PLOTS

Similarly, we have the boxplots and density plots of all the variables. Here we are considering any 5 variables to study the census report.

Now let’s do some analysis based on various features with respect to various States and UTs. For this we will first create a new Data frame named “data frame Gender” which is the copy of original data frame. This will also include an extra column named “Gender ratio”

**Gender Ratio = Total Population Male / Total Population Female –** Equation 4

This will give us a new column which will include gender ratio of each area given in the data frame. Since, Gender ratio is always between 0 to 1, so, a ratio closed to 1:1 is considered ideal. We will find out the minimum and maximum values of the gender ratio and will figure out which State or UT and which area has lowest and highest ratios.

**Gender ratios Based on states and areas**

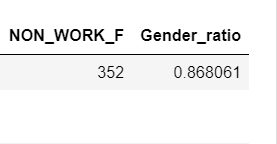
**** ****

Table-23, Maximum Gender ratio

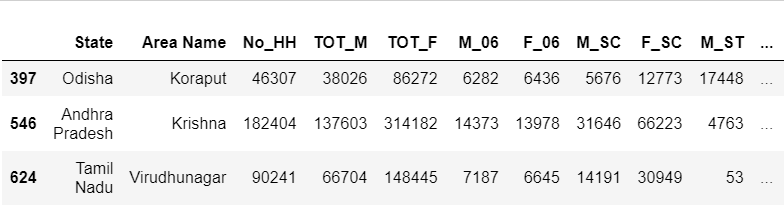
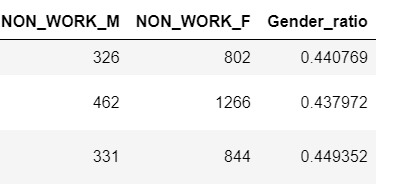
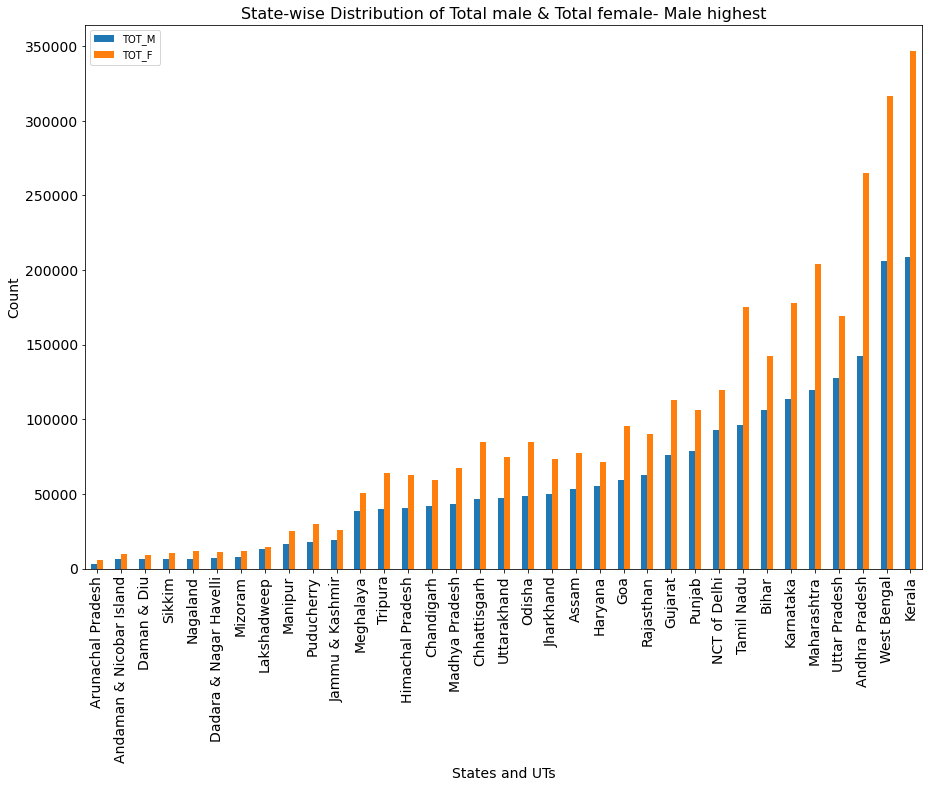
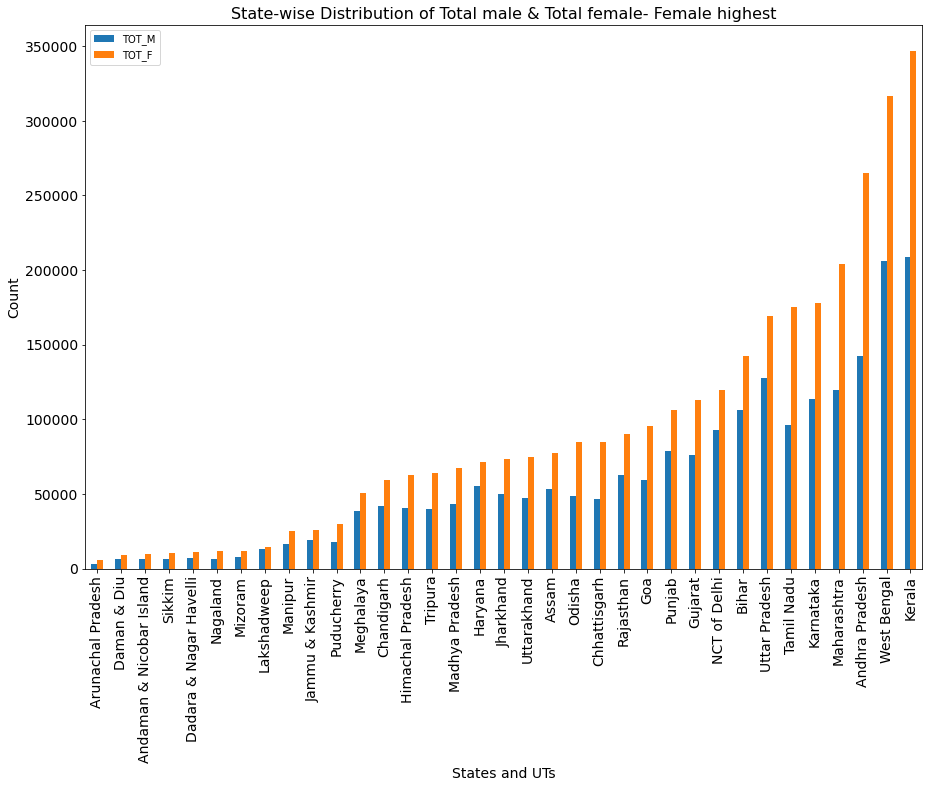
 

Table-24, Minimum Gender ratio

**Male and Female Distribution State-wise- Sorted according to Mean of male population**

 fig-20

 fig-21

**Total Literacy rate Males and Females of various States and UTs with respect to Mean of Population Distribution:**

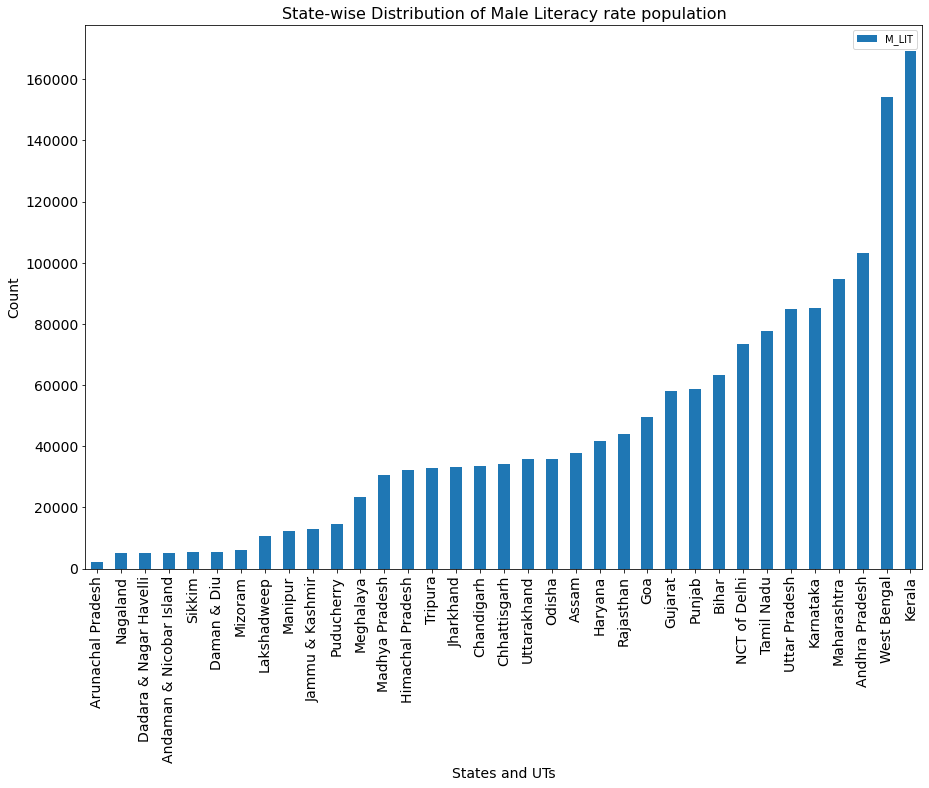
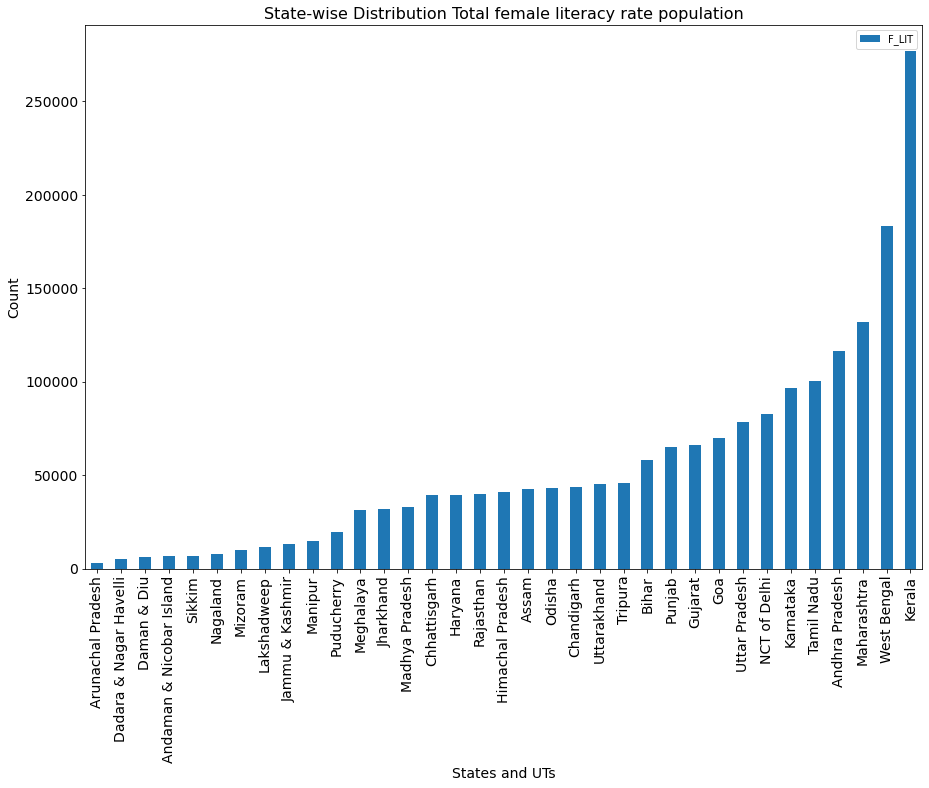
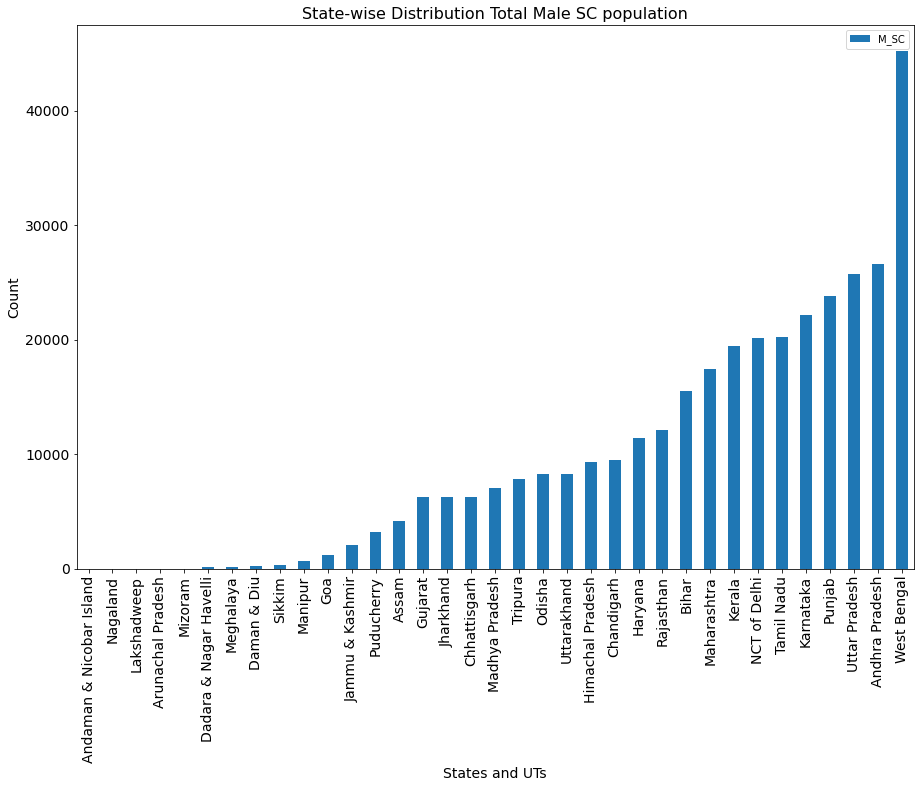
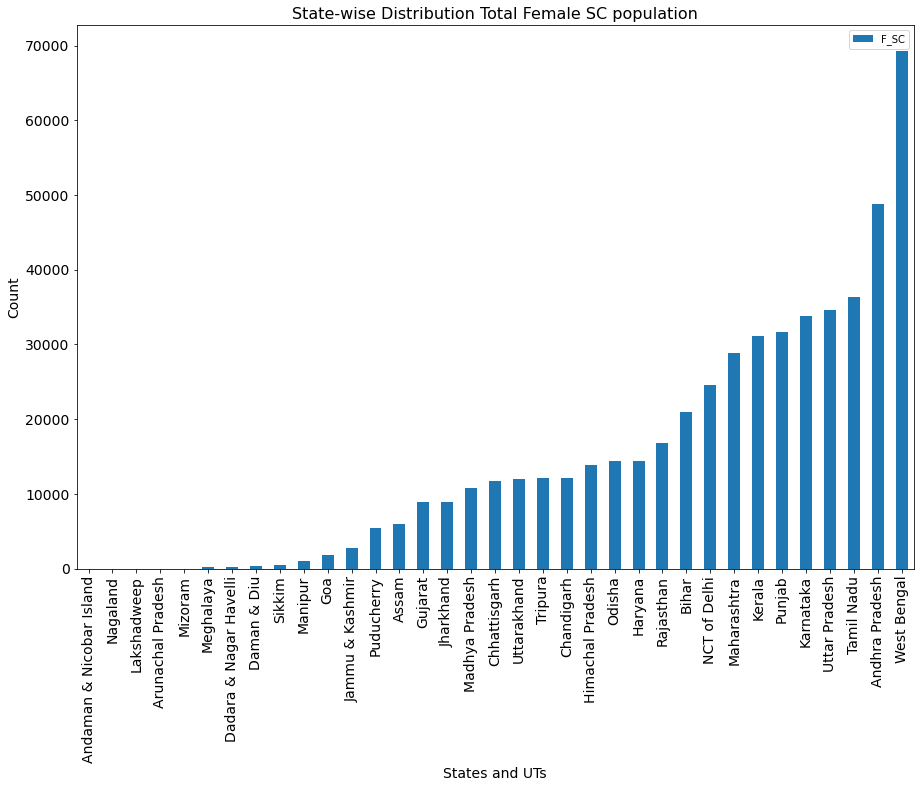
** **

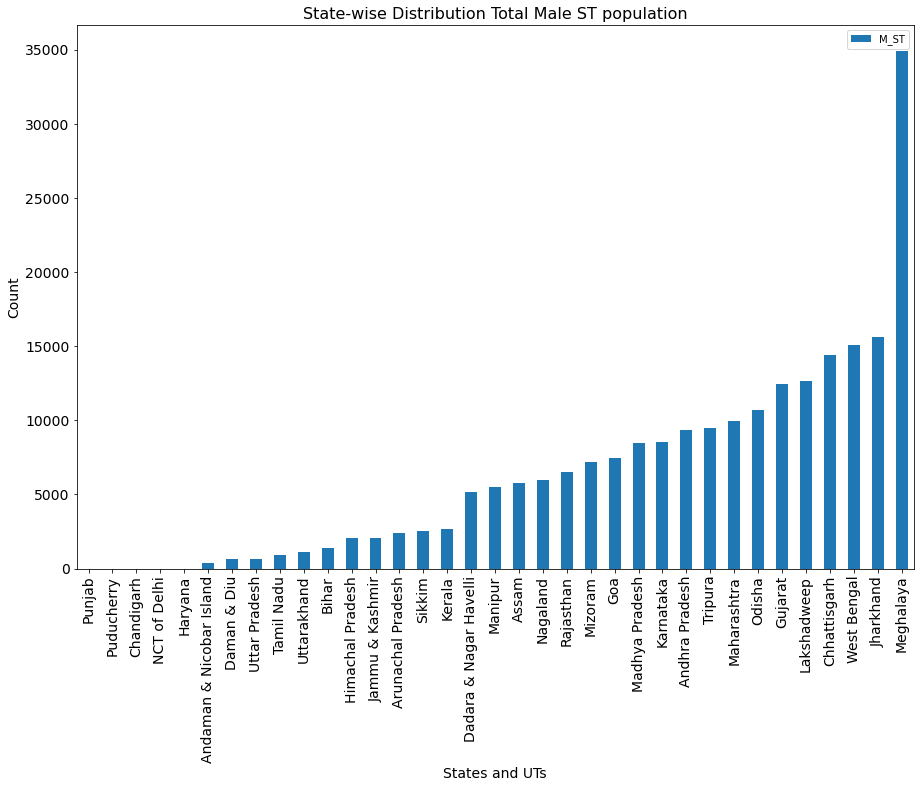
fig-22

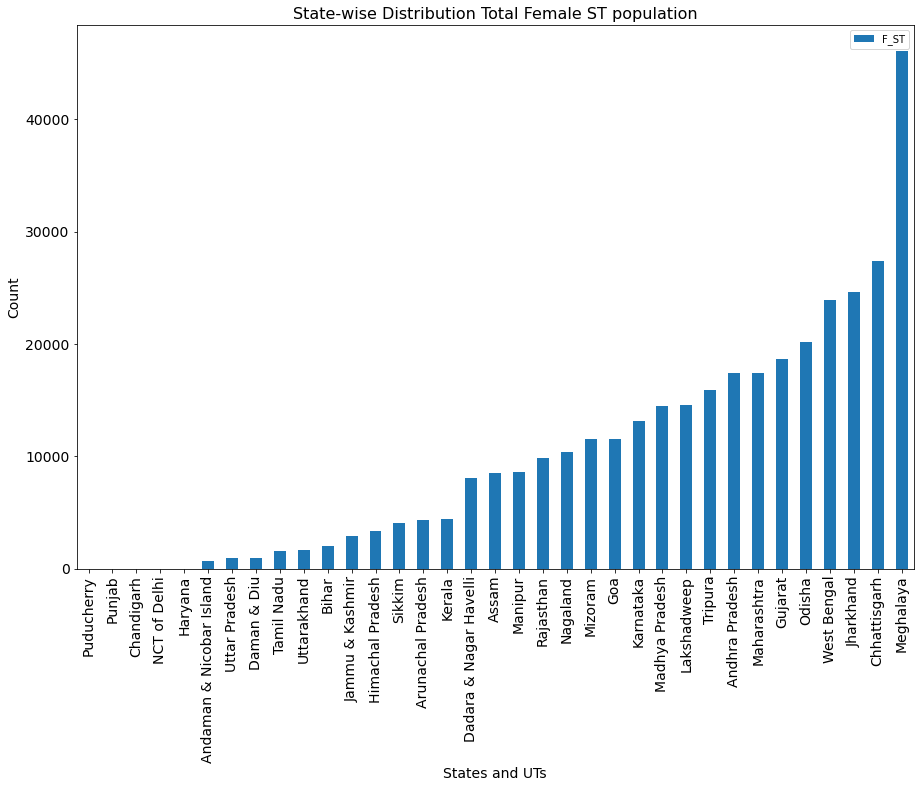
 fig-23

**Total Population Distribution of Scheduled Tribes and Scheduled Castes for both Males and Females of various States and UTs:**

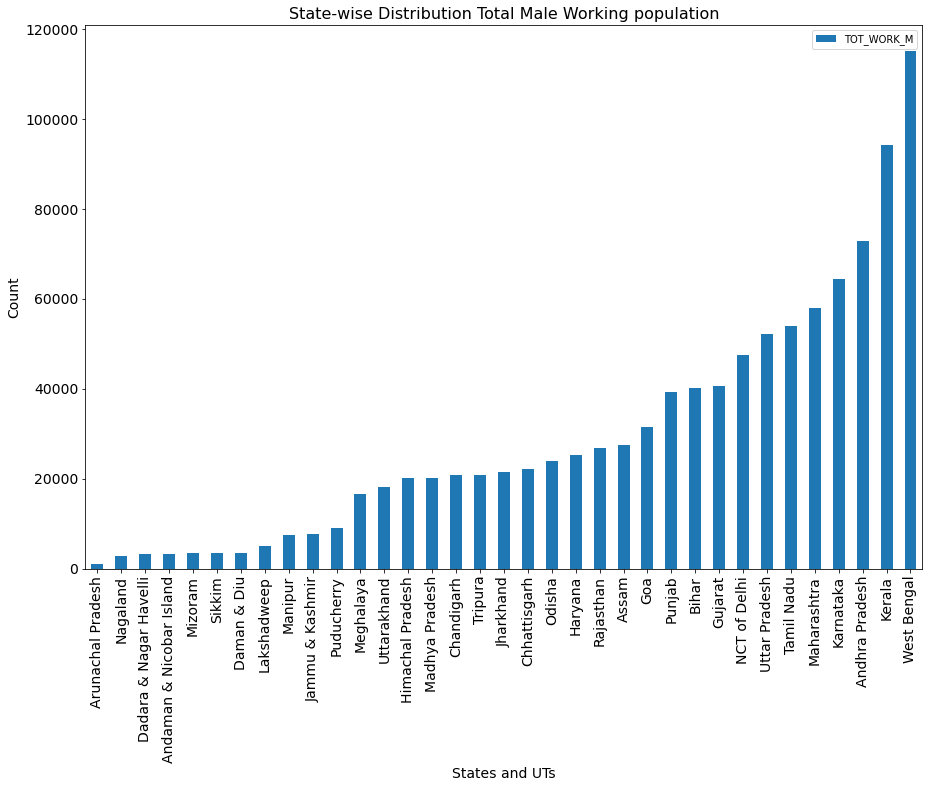
**** fig-24

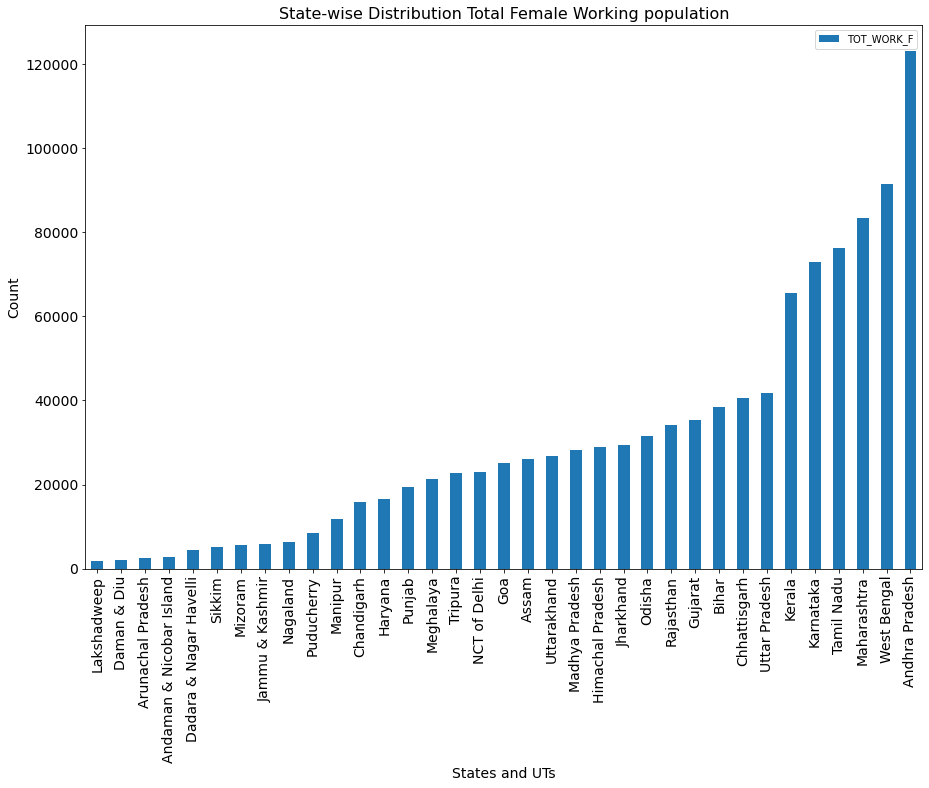
 fig-25

fig-26

 fig-27

**State-Wise Distribution of Total Working population of Male and Female:**

**** fig-28

**** fig-29

**Observations:**

There are 57 numeric fields in the data after dropping "District code" and "State code"

Total population Male ranges from minimum of 391 to maximum of 485417

Total population Female ranges from minimum of 698 to maximum of 750392

The mean population of male between age group of 0 to 6 (12309) is higher than that of mean population of female in the age group of 0 to 6 (11942)

The mean literacy rate of female is higher than that of mean literacy rate of male

Mean of Total working population of male (37992) is lesser than mean of Total working population of female (41295), which shows that female is more work oriented

Range of Scheduled Tribes population of Male is Lesser than Scheduled Tribes population of Female and same is the case for population of Scheduled Castes

Lakshadweep area in the UT of Lakshadweep has the Highest gender ratio of around 0.86

The State of Andhra Pradesh which has an area named Krishna has the lowest Gender ratio which is equal to 0.43

Kerala has the highest mean Literacy rate of both male and female population followed by West Bengal, But the state which holds the third position when it comes to mean of Literacy rate differs for both male and female. For Females its Maharashtra and for males its Andhra Pradesh.

No State or UTs has Gender ratio greater than 0.9

State-wise Distribution for the mean of Total Male SC population - 1st rank goes to West Bengal, 2nd is Andhra Pradesh and 3rd is Uttar Pradesh. For Female Population the 1st and 2nd are same state but the 3rd in the case of Female is Tamil Nadu. For both male and female, the least rank holder is Andaman & Nicobar Islands

State-wise Distribution for the mean of Total Male ST population - 1st rank goes to Meghalaya, 2nd is Jharkhand and 3rd is West Bengal and the least rank goes to State of Punjab. For Female Population the 1st is same, 2nd is Chhattisgarh and the 3rd in the case of Female is Jharkhand. The least in case of female ST population is UT of Puducherry.

When it comes to Total Working Population of male- West Bengal holds 1st position, Kerala is 2nd and Andhra Pradesh is 3rd. The least male working population is in Arunachal Pradesh.

When it comes to Total Working Population of male- Andhra Pradesh holds 1st position, West Bengal is 2nd and Maharashtra is 3rd. The least male working population is in Arunachal Pradesh. Lakshadweep has lowest female working population.

From above results we see that just 'Literacy' may not be a feature driving Working population. Other features like 'SC status', 'marginal workers and their age groups' are also contributing.

It is difficult to consider 57 numeric dimensions to analyse this behaviour.

Hence, PCA can help to reduce dimensions, help in further analysis and derive patterns.

Outliers to be treated

Let us define a function **'univariate Analysis category'** to display information as part of univariate analysis of categorical variables.

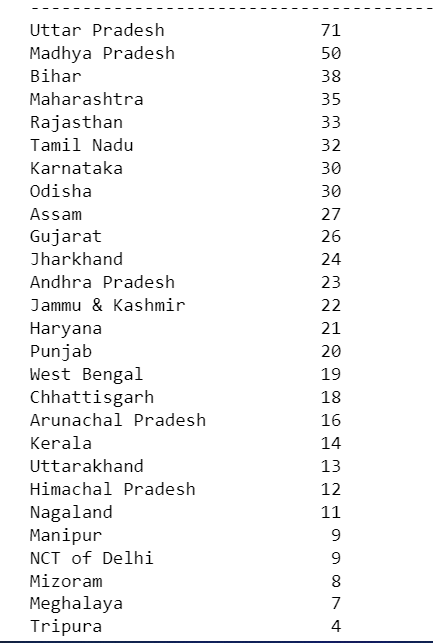
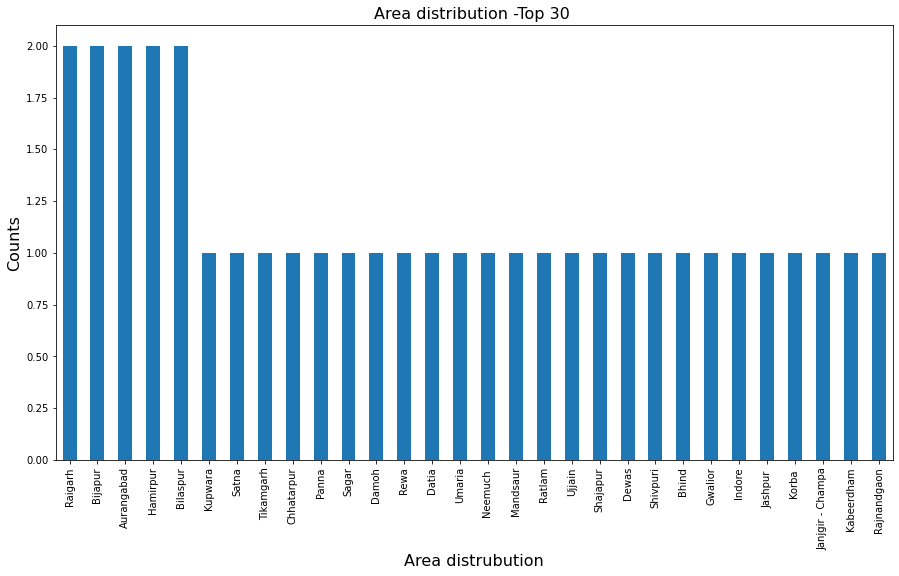
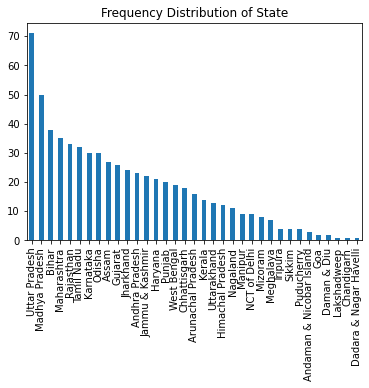




Table-25, value counts of states and areas

The function should display the frequency of all the levels within the field and display a frequency plot

Fig-30

**Bivariate Analysis**

We will draw pair plots and correlation plots for the variables and their correlation with each other.

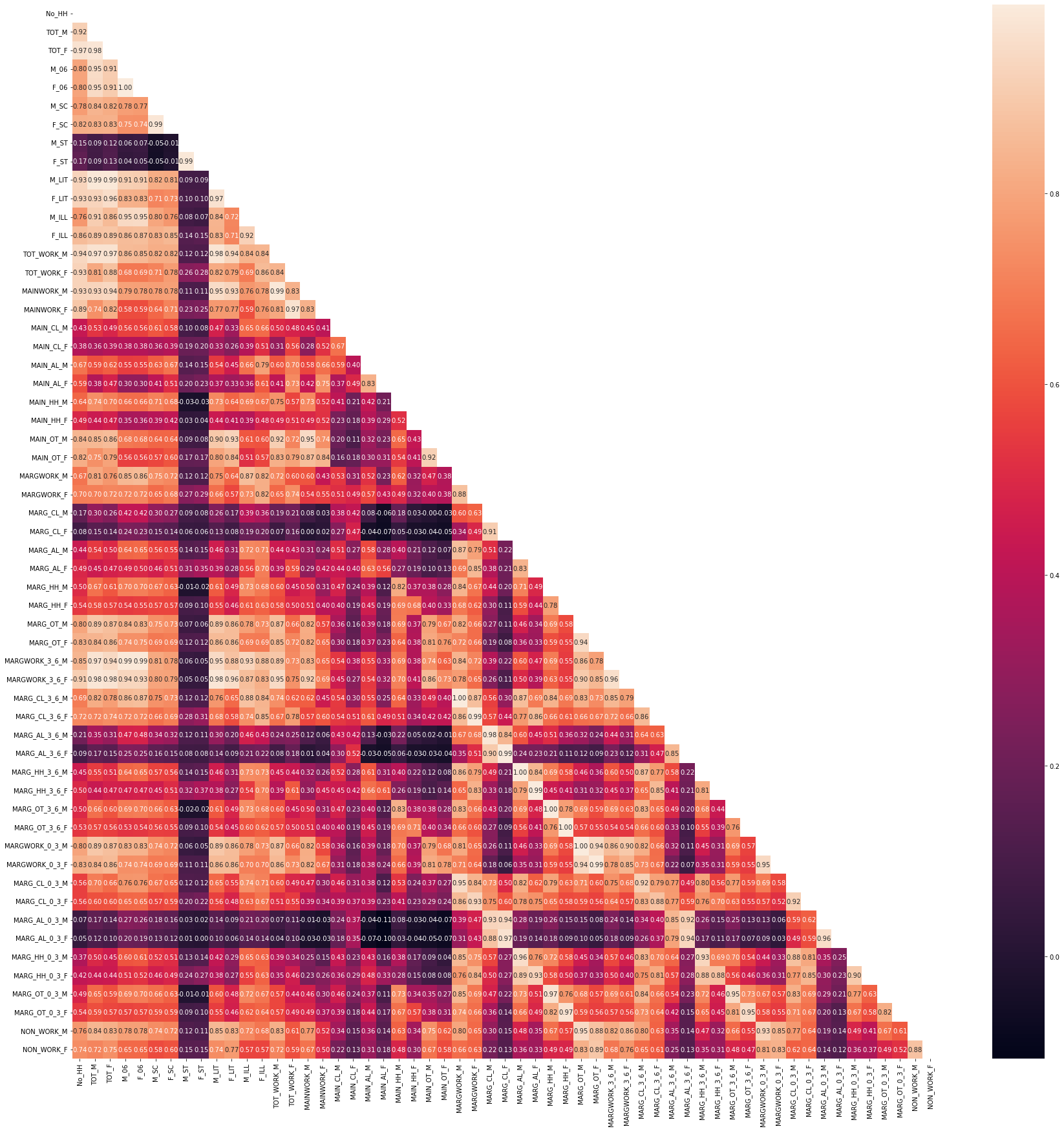


Fig-31

**Observation:**

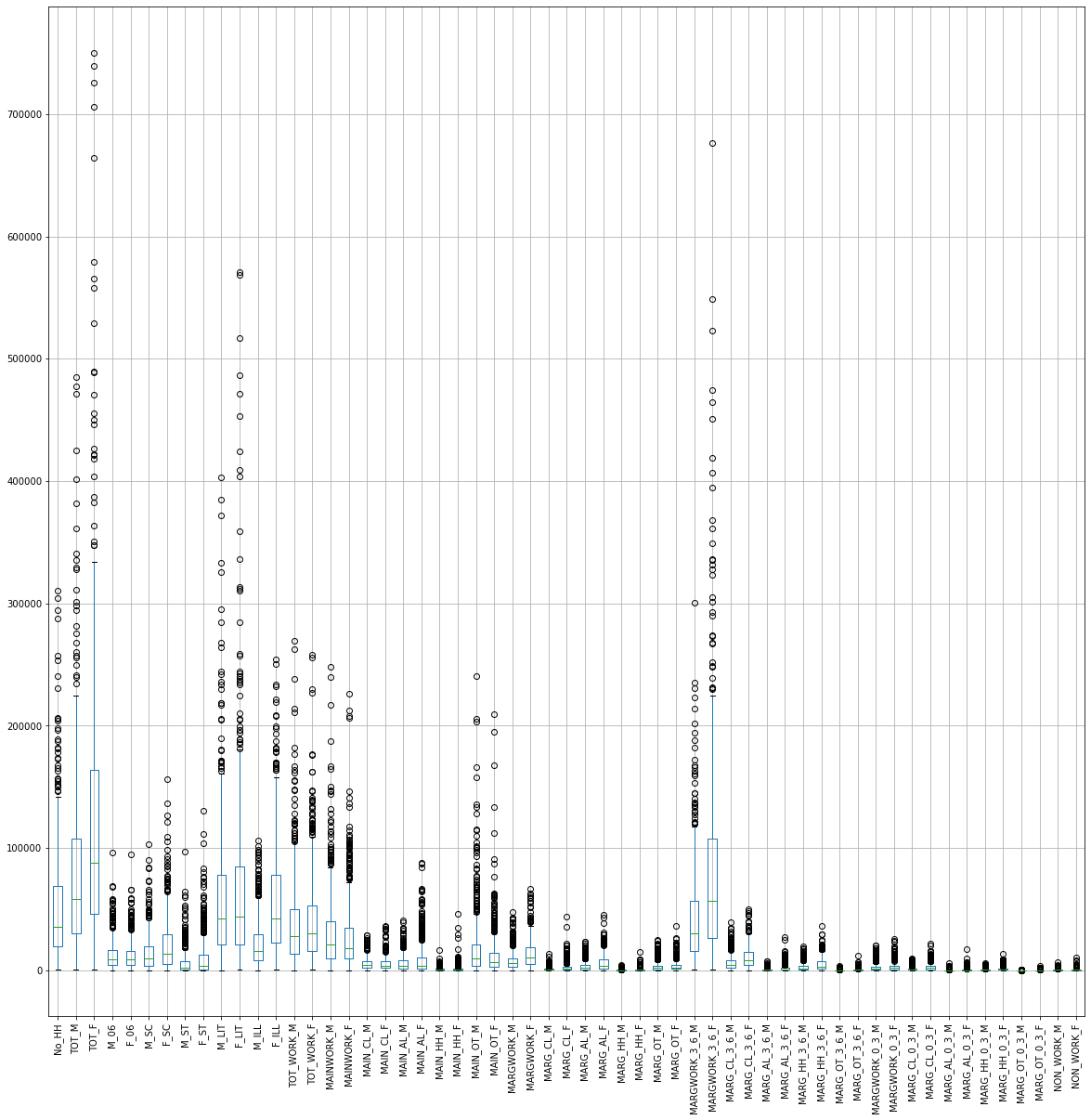
There are considerable number of features that are highly correlated.

'Total Working Male and Female population' shows high correlation with 'No of Households’, ‘Male Literacy’, ‘Female Literacy'. 'Non-Working Population' has high correlation with ‘Illiteracy' etc. There are so many features which are co-dependent on more than one features but the variety of features is very large. So, we need dimensionality reduction using PCA so that we can find out the top 5 to 10 features for the census.

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

We can see the outliers are present in almost all the variables so it will be good to treat these outliers first and then move ahead, because outliers are something which can affect the central tendency like mean & median, and mean and variance being the critical points in analysis of various kinds of PCA, if not treated it may deflect the actual central tendencies and hence the test results.

An outlier may indicate bad data or Noise. Both the variance and the variance–covariance matrix is known to be sensitive to outliers. A single bad outlier may cause that principal components are distorted so as to fit the outlier well, leading to bad interpretation of the results. As, they can negatively affect the statistical analysis and the training process of a machine learning algorithm resulting in lower accuracy. **It’s good to treat outliers.**

****Fig -32

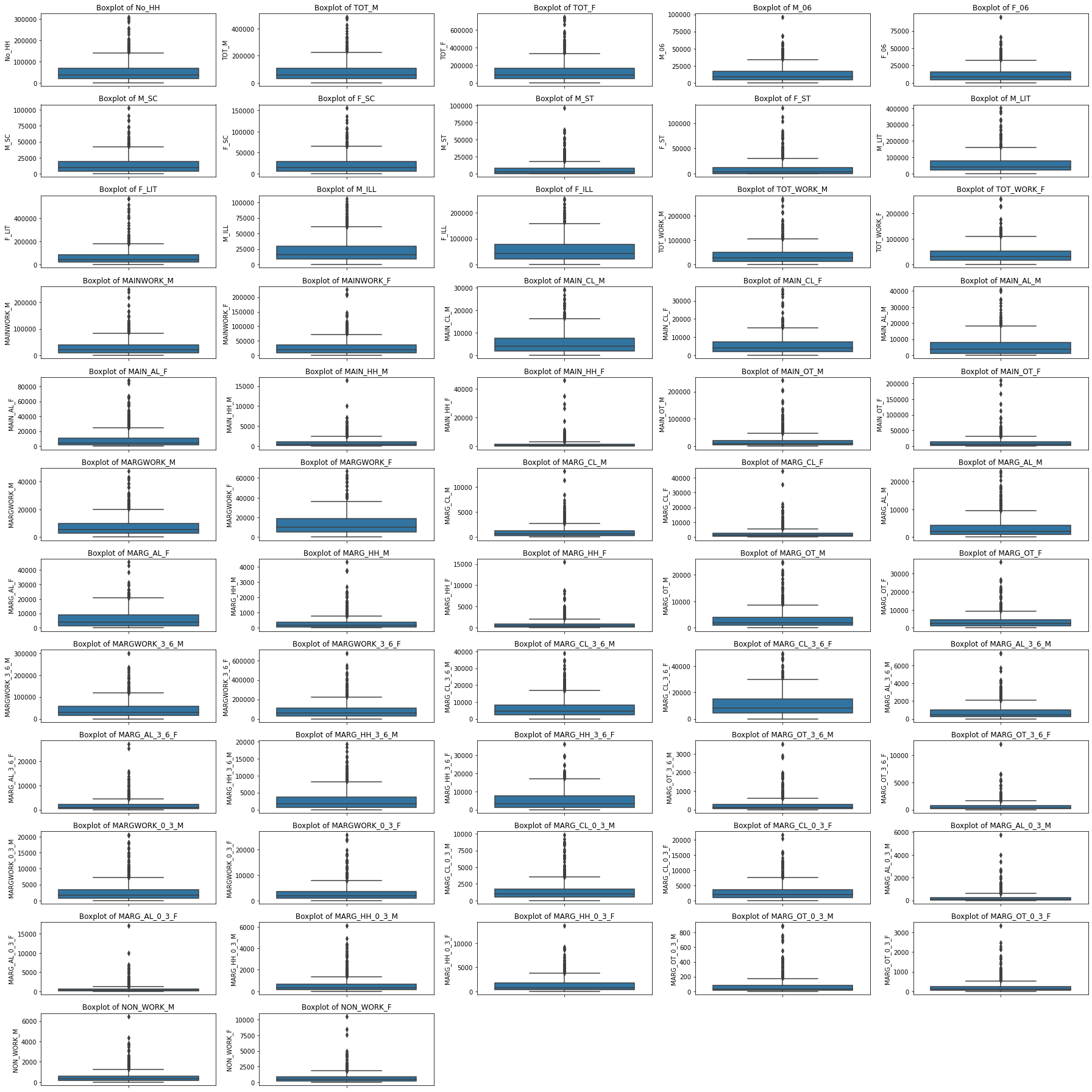


Fig-33

**Outlier Treatment**

To treat outliers lets define a function 'treat outlier'.

- For the higher-level outliers, we will treat it to get it at 95 percentile value.

- Lower-level outliers will be treated to get it at 5 percentile value.

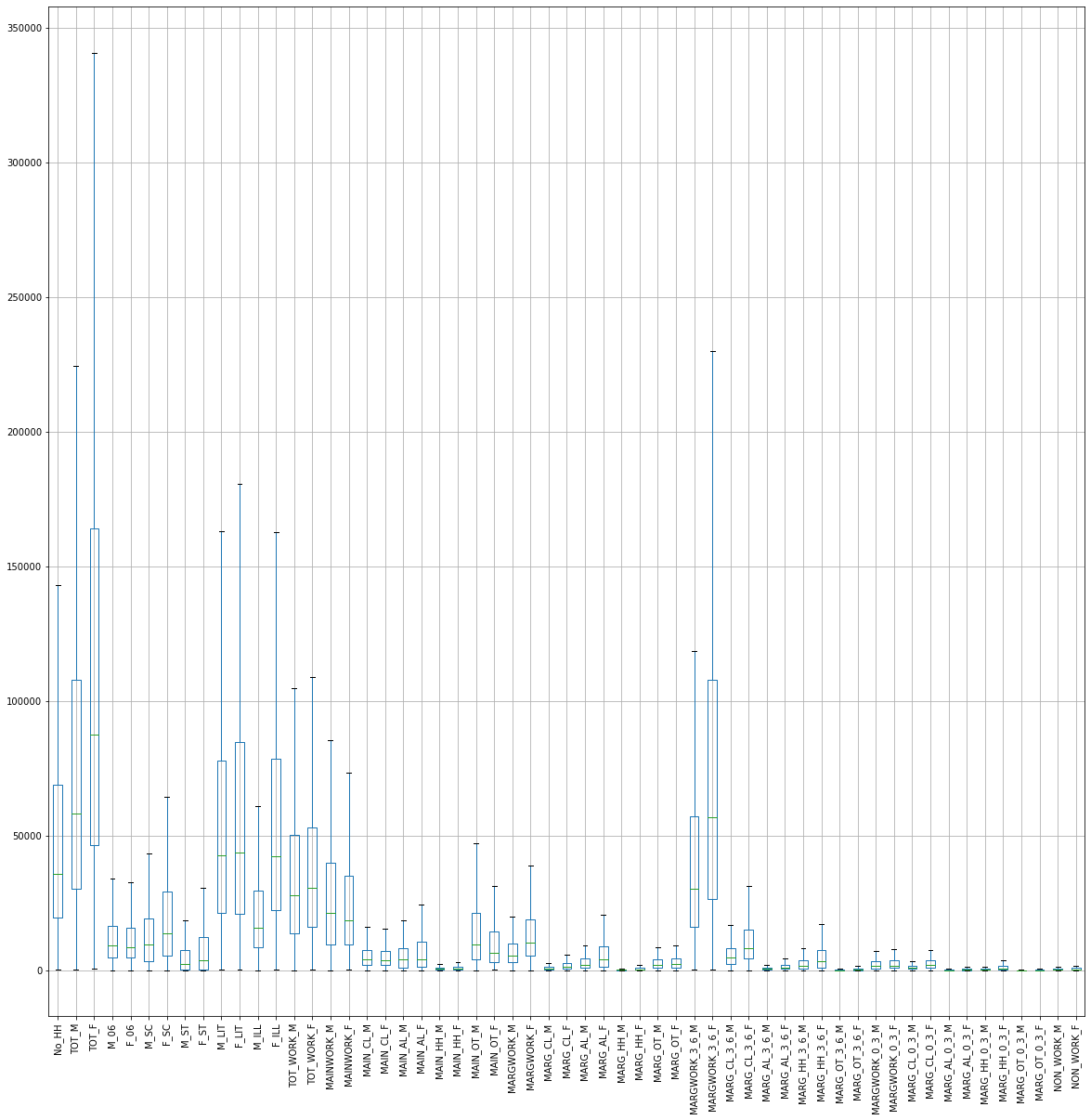


Fig-34

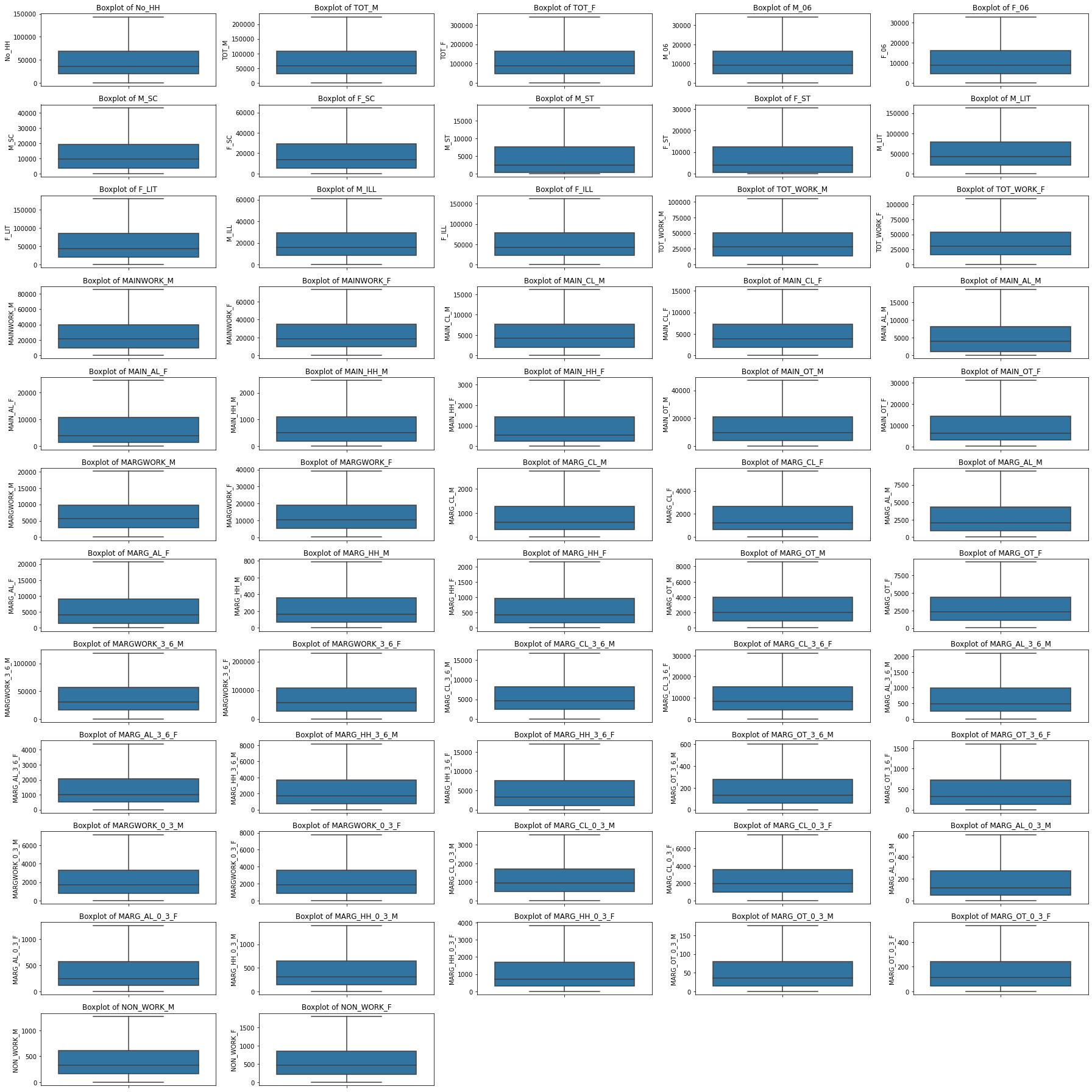


Fig-35

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

Why is PCA important?

If a variable of interest is associated with several other variables, then the association will make the standard deviation of the estimates of the associated parameters exceptionally high. This may impact significance of any hypothesis testing. One example of this phenomenon is multicollinearity in regression. Multicollinearity makes the standard deviations of the estimated regression coefficients very high and often reverses the sign of the coefficients. This renders the regression model itself useless.

Before we move forward, let us introduce the notations. Let the observed variables (original attributes) in the data be denoted by 𝑋1, 𝑋2, . . ., 𝑋𝑝 where 𝑝 is a large number and 𝑉𝑎𝑟(𝑋𝑖) = 𝜎𝑖^2. Total variance in the data is defined as

- Equation 5

**Scaling the data**

Scaling is important feature of PCA, because, PCA becomes ineffective in certain circumstances, such as, when attributes are having different scales, presence of discrete data, presence of skew in the data and nonlinear relationship among the variables. Scaling ensures that attribute means are all 0 and variances 1.

When variances are so widely different, it is not a good idea to perform PCA on the unscaled variables. PCA works on the total variance which is the sum of the variances in the data. If one variance (or more) variance(s) is (are) very high compared to the rest, it (they) will dominate the construction of the PCs and all variables will not have proper representation.

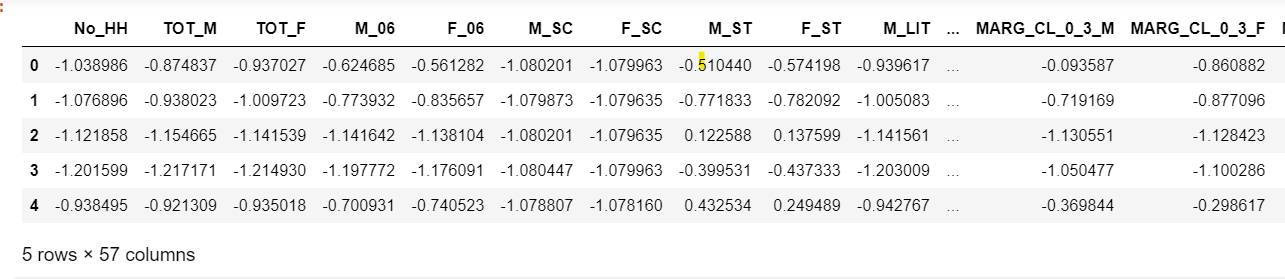
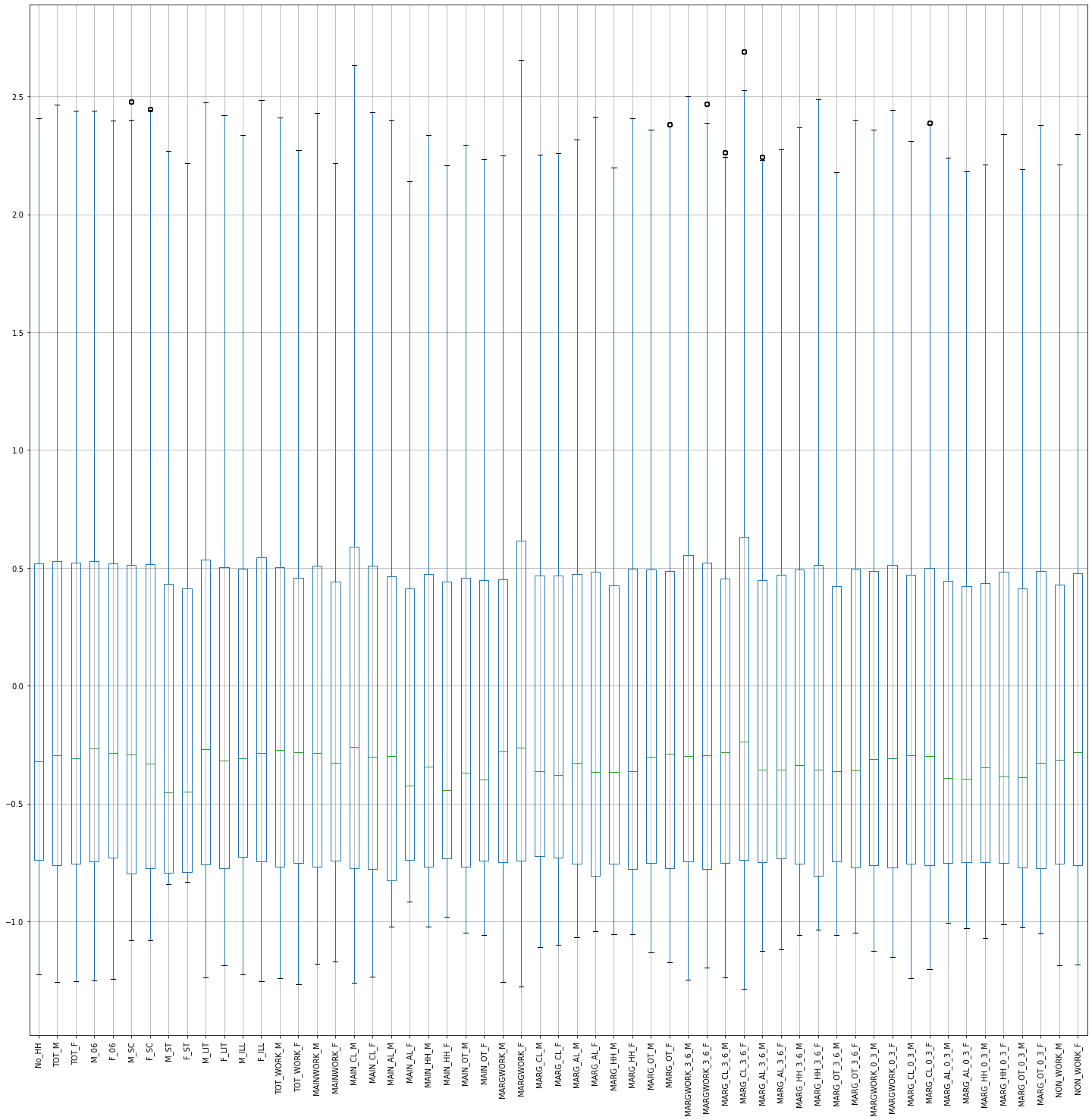


Table-26, Transformed data

 fig-36

This scaling compresses all the inliers in the narrow range. In the presence of outliers, StandardScaler does not guarantee balanced feature scales, due to the influence of the outliers while computing the empirical mean and standard deviation. This Scaler responds well if the standard deviation is small and when a distribution is not Gaussian. This Scaler is sensitive to outliers. Yes, scaling does have a little effect on the outliers but not very effective, although we still can observe some of the outliers after scaling the data, because it’s bringing all the variables in similar range. It just scales the data and bring in the similar range or scale, so that their variances does not vary much with respect to each other. scaling the data makes sure that means are all equals to 0 and the variances equals to 1

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

**Statistical tests to be done before PCA**

Bartletts Test of Sphericity

Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.

- H0: All variables in the data are uncorrelated

- Ha: At least one pair of variables in the data are correlated

If the null hypothesis cannot be rejected, then PCA is not advisable.

If the p-value is small, then we can reject the null hypothesis and agree that there is at least one pair of variables in the data which are correlated hence PCA is recommended.

**Hence, our p value for the test comes out to be 0**.

**KMO Test**

The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.

Generally, if MSA is less than 0.5, PCA is not recommended, since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

**Since MSA (0.93) > 0.7, hence, we are good to go for dimensionality reductions.**

**Covariance matrix, Eigen Values and Eigen vectors:**

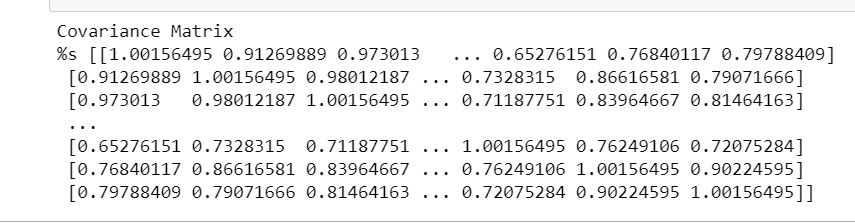
****

Table-27 Covariance matrix

When we generated covariance matrix, we observed that its similar to the correlation matrix which we obtained earlier or the heat map which we have shown above.

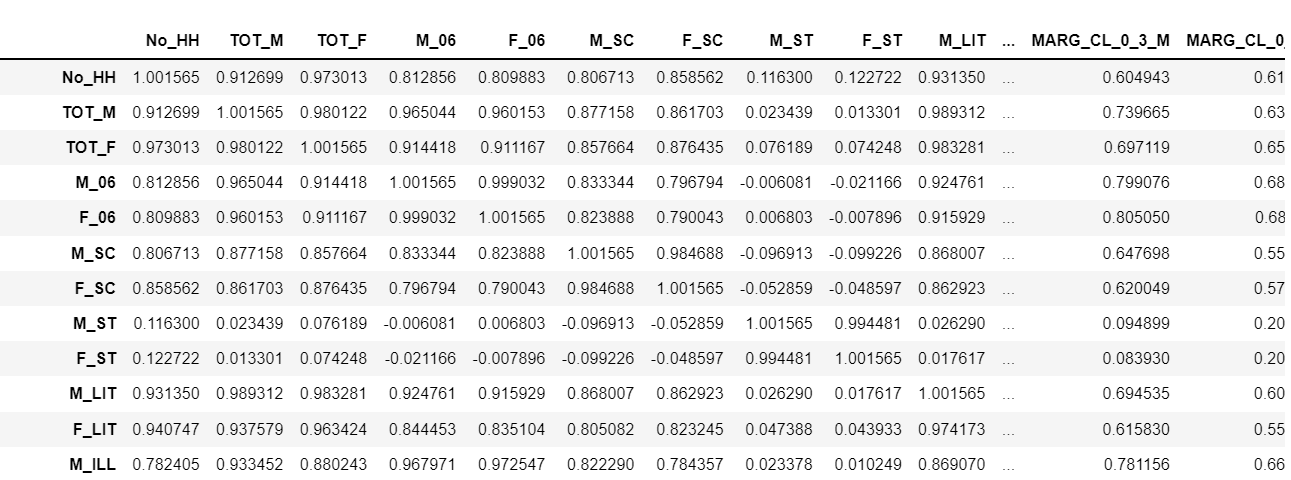


Table-27 Covariance matrix

First, we are generating PCA components which has to be lesser in number than the actual variables. So, PCA transpose comes out to be as follows in the form of array:

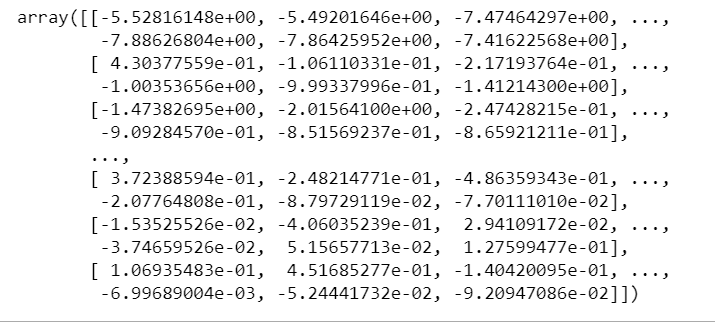
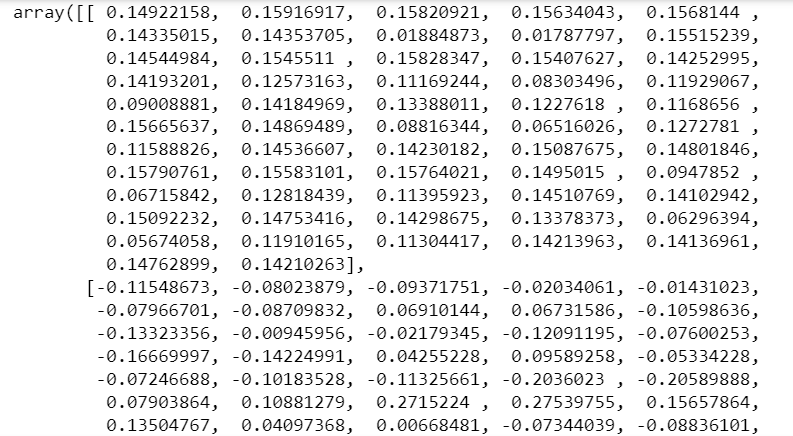


Table-28 PCA arrays

**Eigen Vectors:**

Loading of each feature on the components are actually the method of extracting eigen vectors which are given as follows:

Table-29 Eigen vectors. continued

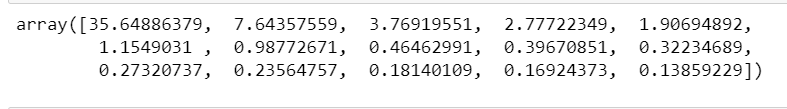
**Eigen values**

For Checking the eigen values we will check the explained variance. This is always returned in descending order.

Now we will Check the explained variance for each PC which returns us our eigen values which is always in descending order.

Explained variance = (eigen value of each PC)/ (sum of eigen values of all PCs)- Equation 6

So, after calculating Eigen values, our explained variance for our 15 PCs comes out in the form of array.

Table-30 Eigen values.

Then we will Check the explained variance ratio for each PC which again comes out in the form of array.

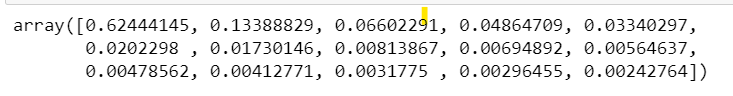
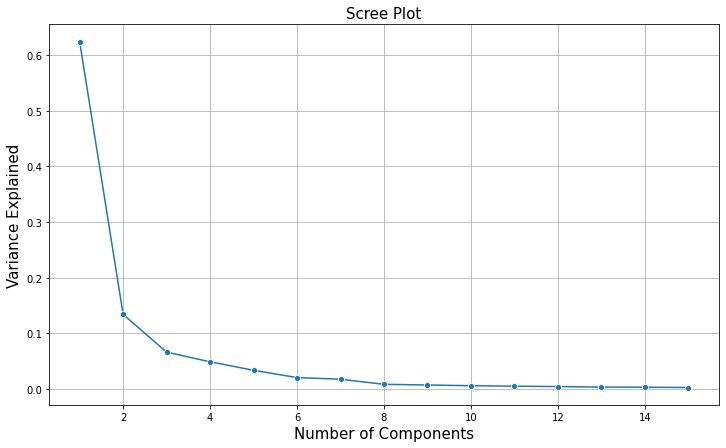


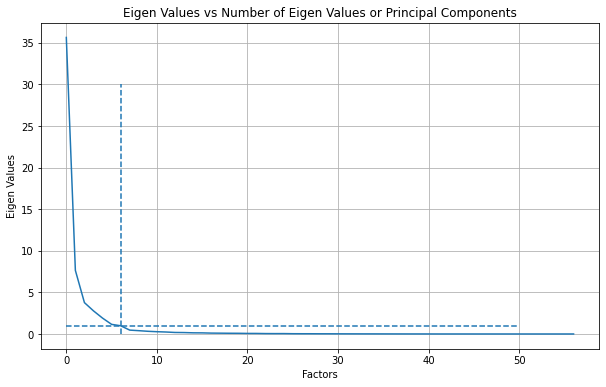
Table-31 explained variance ratio.

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

Choosing the correct number of principal components is pivotal in data analysis and requires a balancing act. On one side, the aim is to reduce the dimension, so keeping too many principal components will not serve the purpose. However, keeping too few components will cause a large proportion of total variation among the original variables to remain unexplained.

Loading the Scree plot for the given PCA components

Fig-37

 Fig-38

The scree plot is a useful visual tool to select 𝑘. On the X-axis are shown the indices of the PCs and on the Y-axis are shown the variances. If there is a distinct break point in the line joining the variances (elbow point) beyond which the line becomes approximately horizontal, then that point may be taken as the value of 𝑘, provided other conditions are also satisfied.

Now we are Creating a data frame containing the loadings or coefficients of all PCs

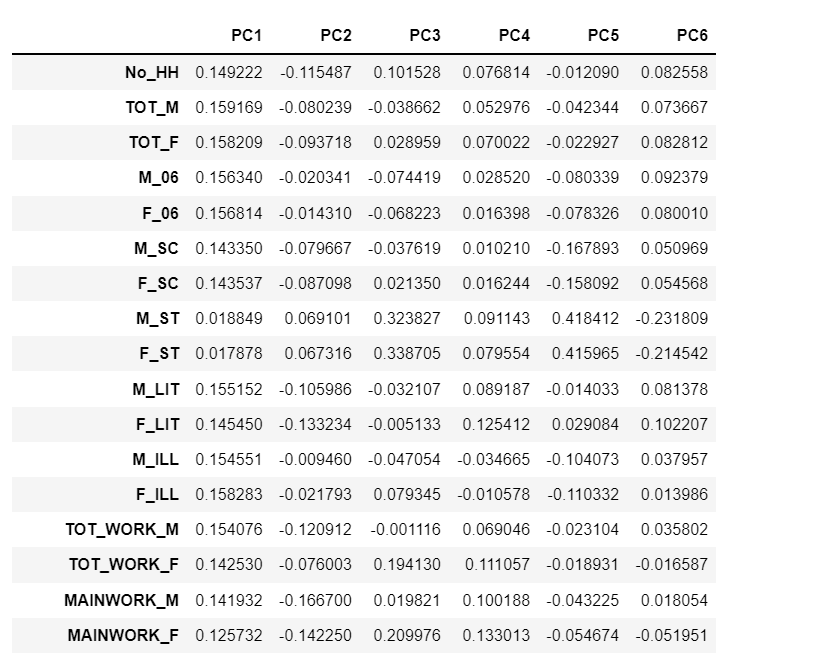
# Description of all the PCs, along with the variables including the top 5 rows only:

# Table- 32 coefficients of all PCs with top 5 variables

For this project, we have to take at least 90% explained variance. So, we will select 6 PCs, which includes almost 92% of explained variance ratio. Cumulative explained variance ratio is calculated to find a cut off for selecting the number of PCs. We can observe from table no 30 that 6 PCs almost contains 92 % of the variances in the data set.

**So optimal number of PCs = 6**

So, lets create a data frame including the 6 PCs along with those of the variables in the original data set.

****

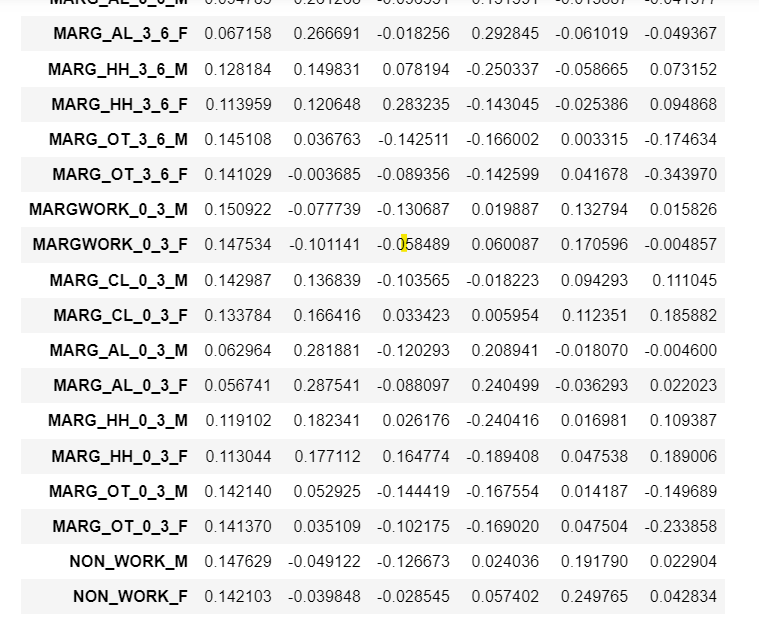


Table-33 Selected PCs

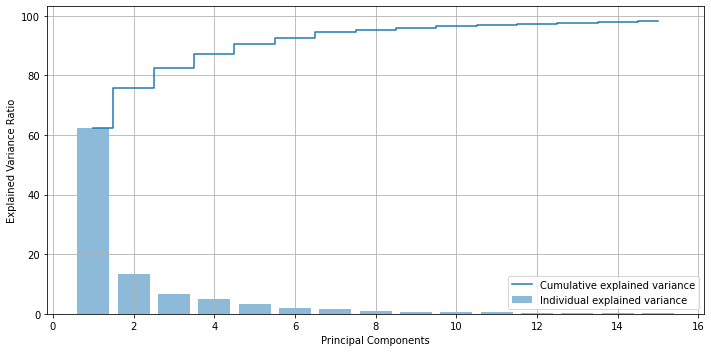
**Part 2.7 -Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the principal components in terms of actual variables.?**

Let's identify which features have maximum loading across the components and which is explaining most variances.

- We will first plot the component loading on a heatmap.

- For each feature, we find the maximum loading value across the components and mark the same with bar and step plot as well which is known as **Pareto chart.**

- We will also consider rectangular box with red markers. Features marked with rectangular red box are the one having maximum loading on the respective component. We consider these marked features to decide the context that the component represents

Fig-39

**Observation:**

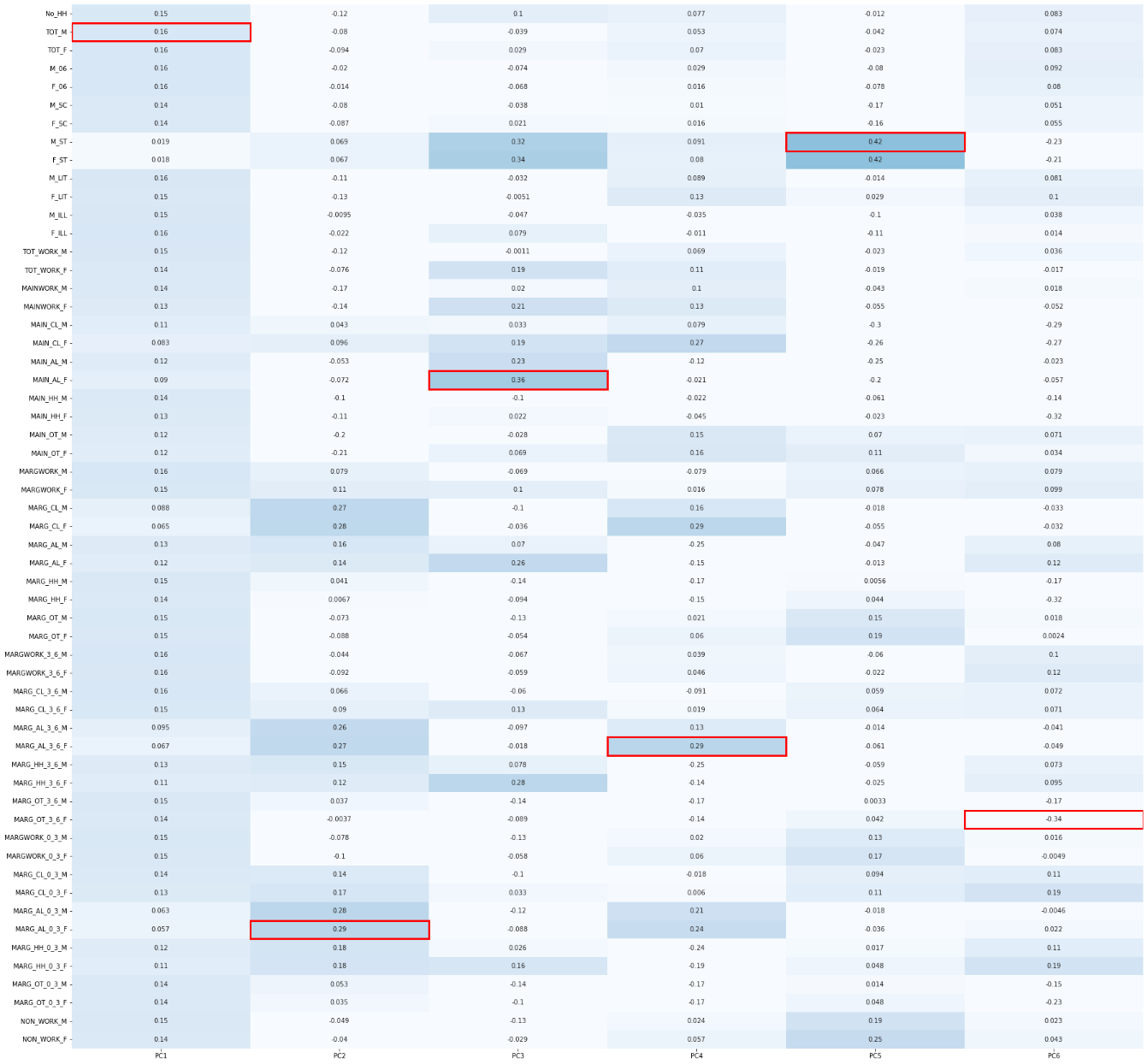
We can see from the above graph of cumulative explained variance and individual explained variance that PC1 has the highest explained variance if we talk about individual case which is almost equal to 35% and we as move to other PCs it decreases to 7% for second PC, 3% for 3rd PC.

But if we talk about cumulative effect of PCs, 1st and 2nd PC together contains 75% of the variable’s characteristics together. If we take 3 PCs combined it has almost 82% of the total variable characteristics and 6 PCs contains almost 92% of the information of the whole census data. We will see the individual effects of the variables on the PCs in the next graph.

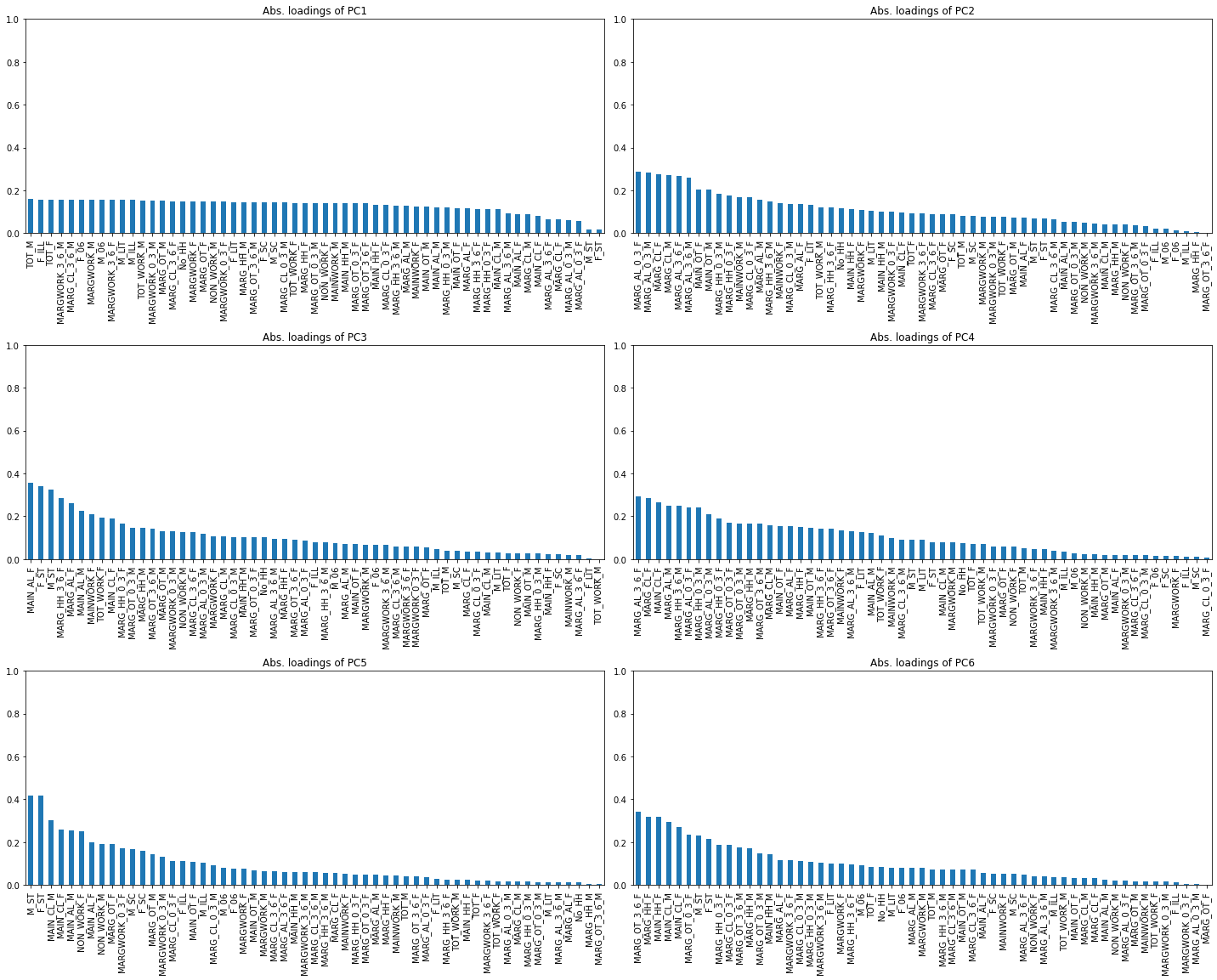
**Correlation heatmap:**

The correlation matrix and pairwise scatterplots indicate high correlation among various variables and the principal components. We will do the further analysis based on the values of the correlation heat map and try to infer the respective relations. Moderate correlation may be detected between several pairs of variables. Existence of such pairs of high and moderate correlations indicate that dimension reduction must be considered for the Places Rated data.

The property that the first principal component 𝑌1 has the largest variance, the second principal component 𝑌2 has the second largest variance and so on, till the 𝑝th, PC 𝑌𝑝 has the smallest variance, ensures that PC is a dimension reduction technique

Fig-40

But since we have selected 6 PCs, we will draw bar graphs for all the 6 PCs along with the variables and find out which PC is explaining the most variance. So, Checking as to how the original features matter to each PC. In this case we will be only considering the absolute values.

Fig-41

**Observations:**

- We can see that in PC1 Total Male, Total Female and Population in Age group of 0 to 6 both male and female and marginal workers population in the age group of 3 to 6 both male and female are having the highest correlation or we can say the highest impact on the principal component 1. So, as we know PC1 has the highest cumulative variance so these factors are the one which are showing the maximum variance.

- In PC2 we can see the Marginal Agriculture Labourers Population 3-6 Female and Male, Marginal Agriculture Labourers Population 0-3 Male and Female, and Main Cultivator Population Male and Female, as the highest impacting variables which are showing the maximum variance in this case. As we can see that PC2 has around 7% variance individually so, these are the variables which is included in this single PC2.

- In PC3 Main Agricultural Labourers Population Female has the highest correlation, followed by Scheduled Castes population of Male and Female.

- In PC4, Marginal Agriculture Labourers Population 3-6 Female contributes the maximum, Followed by Marginal Agriculture Population.

- In PC5 Scheduled Tribes population of Male and Female is dominating with highest correlation, followed by Main Cultivator Population of both Male and Female.

- If we talk about PC6, Marginal Other Workers Population Person 3-6 Male is highest contributing variable. But we can see that the variables and their correlation decrease as we move away from PC1 to PC6.

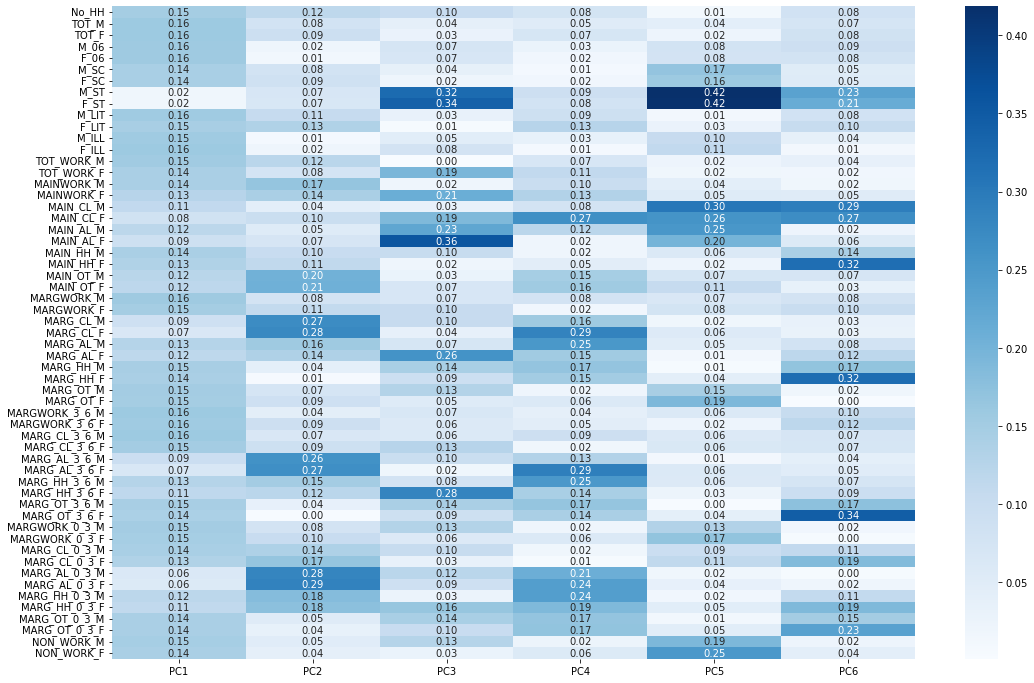
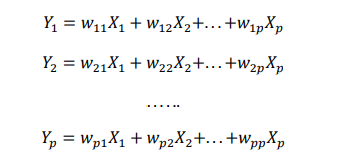


Fig 42

**Part 2.8 - PCA: Write linear equation for first PC**

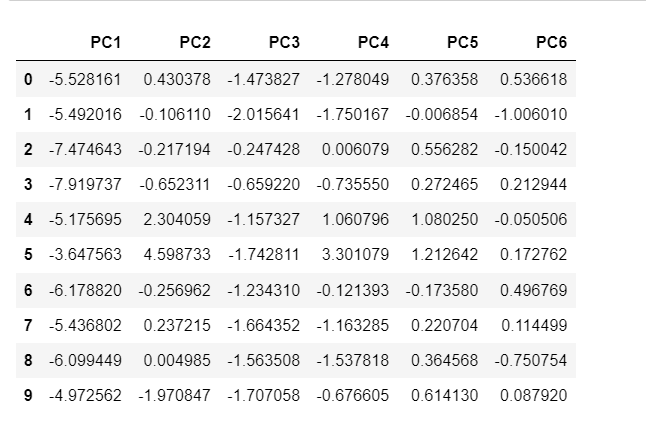
Let the principal components be defined by 𝑌𝑗, 𝑗 = 1,2, . . ., 𝑝. The total number of PCs that can be defined is equal to the number of original attributes in the data. The PCs are linear combinations of the 𝑋’s and may be defined as:

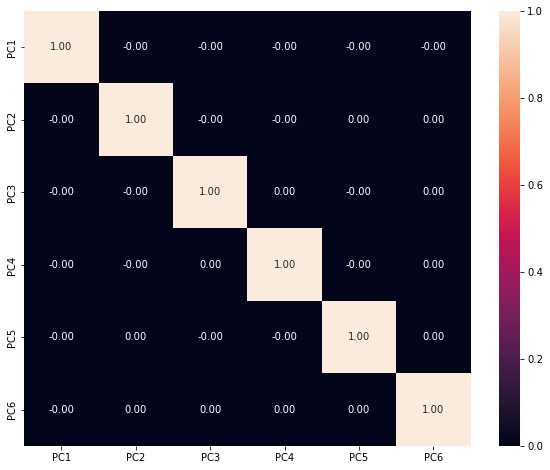
****Equation -7

where the weights 𝑤11, 𝑤12, . . ., 𝑤𝑝𝑝 need to be determined. In fact, the problem of construction of PCs reduces to estimation of 𝑤11, 𝑤𝑝𝑝

Linear Equation = So we have calculated the LINEAR EQUATION by multiplying the eigen vectors or coefficients into 57 variables which are given in scaled form.

**LINEAR EQUATION for PC1 = (0.14922 \* -1.038986) +(0.159169 \* -0.874837) +(0.158209 \* -0.937027) +( 0.156340 \* -0.624685) .......(0.142103 \* -0.774865) = -5.528161476544539**

Table-34 Data frame out of fit transformed scaled data

 Fig 43

So here we can see that we have reduced the dimensions and solved the problem of multi-collinearity and we can see that the principal components have nearly zero correlations with each other. This solves our problem and we have reduced the dimensions or our data set.

**Conclusion:**

* With help of PCA we have been able to reduce 57 numeric features into 6 components which is able to explain 92% of variance in the data
* With help of reduced components, we have been able to observe some patterns. Using some rules around the census data we are able to shortlist some of the features which the government needs to focus on and monitor closely either for advance closure or for possible default case (20 out of 640).
* Using the components additional rules can be derived and analysed.
* After PCA we can actually do Clustering which is a technique of unsupervised learning and segmentation of the data and this will further help the in getting the proper insights from the census data. These data can help government, so that they can work in the direction of required sectors**.**

*Reference – Great Learning lecture videos and Mentors*