DATA MINING

**CONTENTS PAGE**

**CLUSTERING**

**1.1 Read the data and perform basic analysis 1-4  
1.2 Treat Missing values 4  
1.3 Checking outliers 5  
1.4 Perform Z-scaling 6  
1.5 Dendrogram 6-8  
1.6 Elbow Plot 8-9  
1.7 Print Silhouette 9-11  
1.8 Profiling the Ads 11-17  
1.9 Conclusion 17**

**PCA**

**2.1 Reading the data and perform basic checks 18-23  
2.2 Perform detailed Exploratory Analysis 23-37  
2.3 Outliers 37-38  
2.4 Scale the data using Z-score 39-41  
2.5 SKlearn , Covariance matrix , eigen value and vector 42  
2.6 90% explained variance , scree plot 43-47  
2.7 Most variance, actual variables 48-49  
2.8 Linear equation for first PC 49**

**TABLE CONTENTS**

**1.1 Top 5 rows 1  
1.2 Bottom 5 rows 2  
1.3 Info 2  
1.4 Data Summary 3  
1.7 Null values 4  
2.1 Treating missing values 4  
4.1 Z-score 5  
7.11 Silhouette score 11  
8.2 Label values 11  
8.3 Clutser4 12  
8.4 Optimal number of clusters 12  
9.1 Top 5 rows 19  
9.2 Last 5 rows 19  
9.3 Info 20-21  
9.4 Describe 21-22  
9.5 Null values 22**

**10.3 Cross table – State wise – Maximum 25  
10.6 Lowest Gender ratio 27  
10.8 District wise – maximum 28  
10.9 District wise – Gender wise gender ratio 29  
10.10 District wise – lowest gender ratio 29**

**Figures**

**3.1 Outliers present 5  
3.2 no outliers present 5  
5.1 to 5.3 WARD 6-7  
5.4 Euclidean 8  
6.2 Elbow method 9  
8.6 Trends of Clicks 13  
8.8 Trends of Spend 13  
8.10 Trends of Revenue 14  
8.12 Trends of CTR 15  
8.14 Trends of CPM 16  
8.16 Trends of CPC 17  
10.1 Male population 23  
10.2 Female population 24  
10.12 Male Literates 30  
10.13 Female literates 30  
10.15 Total Working Male 31  
10.17 Total Working Female 31  
10.19 Main working Male 32  
10.21 Main working Female 32  
10.22 – 10.27 – EDA – Univariate 33-35  
10.28 – 10.30 Bivariate Analysis 36-37   
11.3 Outliers 38  
11.4 No outliers 38  
12.4 Before Scaling outliers 40  
12.5 After Scaling outliers 41  
14.5 Scree Plot 45  
15.1 Most variance 48  
15.2 Heatmap 49**

**OUTPUT**

**1.5 Shape of the dataset 3  
1.6 Duplicate values 3  
2.2 Before treating outliers 4  
6.1 WSS 8  
7.1 – 7.10 Silhouette values 9-10  
8.1 Label value of cluster 4 11  
8.5 Clus-kmeans mean with Clicks 12  
8.7 Clus-kmeans mean with Spend 13  
8.9 Clus-kmeans mean with Revenue 14  
8.11 Clus-kmeans mean with CTR 14  
8.13 Clus-kmeans mean with CPM 15  
8.15 Clus-kmeans mean with CPC 16  
9.6 Duplicate values 23  
10.4 Maximum Males 26  
10.5 Maximum Females 26  
10.7 Minimum Males 28  
10.11 Literacy rate 29  
10.14 Maximum Total working population of male 30  
10.16 Maximum Total working population of females 31  
10.18 Maximum working population of males 32  
10.20 Maximum working population of females 32  
11.1 Removing 2 columns 37  
11.2 After removing 2 columns 37  
12.1 Scaled data 39  
12.2 Values before Scaling 39  
12.3 After scaling 39  
13.1 Covariance matrix 42  
13.2 Eigen vector 42  
13.3 Eigen values 42  
14.1 Variance 43  
14.2 Cumulative variance 43  
14.3 PCA explained variance 43  
14.4 Extracted loadings for 10 PC’s 43-44  
14.6 Cumulative variance explained values 45  
14.7 PCA explained variance ratio 45  
14.8 Optimal number of PC’s 45  
14.9 Extracted loadings for 5 PC components 46-47**

**Problem Statement:**

**Clustering:**

**Digital Ads Data:**

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

The following three features are commonly used in digital marketing:

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

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

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

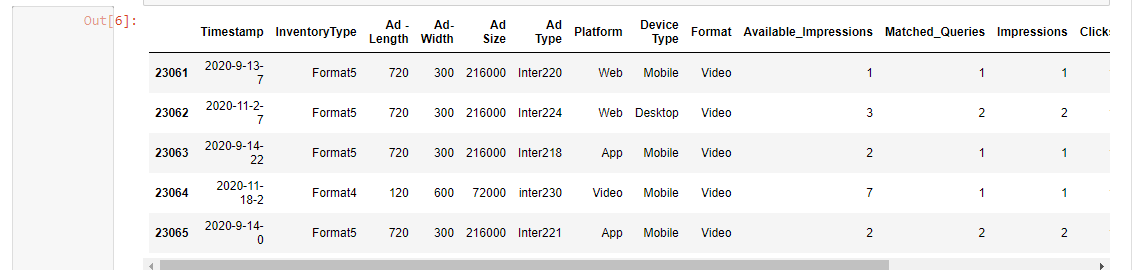
**The Data Dictionary and the detailed description of the formulas for CPM, CPC and CTR are given in the sheet 2 of the**[**Clustering Clean ads\_data**](https://olympus.mygreatlearning.com/courses/81614/files/6807279/download?verifier=9cHqf0pTGTKvHZ3lCWFHuL65aqhpCayisZp1Mpr9&wrap=1)**Excel File.**

**Part 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.   
  
Head :**

Graphical user interface, text, application

Description automatically generated  
**Table 1.1** **Displaying Top 5 rows.**

**Tail :**



**Table 1.2 Displaying last 5 rows.**

**Info :**

Table

Description automatically generated with medium confidence  
**Table 1.3 We can observe that 19 columns are present : 6 object datatype columns , 7 integer datatype columns , 6 float datatype columns.**

**Data Summary :**

Graphical user interface, application

Description automatically generated

Table 1.4 The above screenshot describes the statistical detailed data summary of the given dataset.

**Shape :**

  
Output 1.5 Shape of the dataset

**Duplicate values :**   
  
**Output 1.6 So no duplicate values present in the given dataset.**

**Null values :**Table

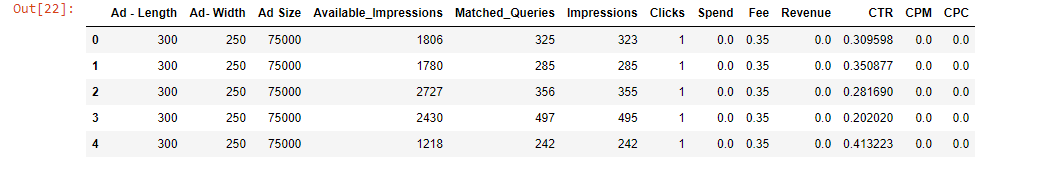
Description automatically generated with medium confidence  
**Table 1.7 We can see that there are 4736 Null values present in CTR , CPM , CPC so we shall treat it further.**

**Part 1 - Clustering: Treat missing values in CPC, CTR and CPM using the formula given.** You may refer to the [Bank\_KMeans Solution File](https://olympus.mygreatlearning.com/courses/81614/files/6807289/download?verifier=k5pRaBGUtrXqimGXASOQc2aGlbqBtTbZ6hwQ43Vh&wrap=1" \t "_blank" \o "Bank_KMeans_Solution_File-1.ipynb) to understand the coding behind treating the missing values using a specific formula. You have to basically create an user defined function and then call the function for imputing.

A picture containing text

Description automatically generated

**Table 2.1 So after treating the missing values we can see now that there is no Null values present.**



Output 2.2 So the above screenshot, shows the top 5 rows values after removing **'Timestamp', 'InventoryType', 'Platform', 'Device Type' , 'Format', 'Ad Type' c**olumns from the dataset before treating outliers as those are categorical column values.

**Part 1 - 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. (As an analyst your judgement may be different from another analyst).**

**Ans : Yes. Outliers are present. Outliers have an significant impact on clustering. So it is better to identify and remove outliers before we apply K-means clustering algorithm. I have used boxplot method IQR value to remove outliers.**

Chart

Description automatically generated

**Fig 3.1 We can see that there are outliers present in the data.**

Chart, box and whisker chart

Description automatically generated  
**Fig 3.2 Now we can see there are no outliers present.**

**Part 1 - Clustering: Perform z-score scaling and discuss how it affects the speed of the algorithm.**

Graphical user interface, application

Description automatically generated

**Table 4.1**

**Ans : Z-score speeds up its execution. As we can see that there were multiple values in single digit and double digits so with the help of z-score scaling, we have scaled the dataset in normal form for further analysis.**

**Part 1 - Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.**

**WARD : The following screenshots shows the dendrogram with the help of ward linkage function.** Chart, box and whisker chart

Description automatically generated

**Fig 5.1**

Chart, histogram

Description automatically generated

Fig 5.2

We can observe the last 10 merge clusters using the truncate with p value as 10 and it shows as :   
Chart, histogram

Description automatically generated  
Fig 5.3

**Euclidean : The following screenshot shows the dendrogram with the help of Euclidean function.**

Chart, histogram

Description automatically generated **Fig 5.4**

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

A screenshot of a computer

Description automatically generated with medium confidence  
Output 6.1 The screenshot displays the output of WSS values for n=10.

Chart, line chart

Description automatically generated  
Fig 6.2 Graph displaying the Elbow method to identify optimum number of clusters. From the above plot, we can see that from K=1 to K=2 , it shows a significant reduce in the value and again from K=2 to K=3 and again from K=3 to K=4 we can see again significant reduce in the values.

So we can conclude that WSS value is not significantly reducing beyond 4 so hence the optimal number of clusters is 4.

**Part 1 - Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters.**

  
  
Output 7.1 The above screenshot shows the **silhouette value for cluster = 1**

  
  
Output 7.2 The above screenshot shows the **silhouette value for cluster = 2**

  
  
Output 7.3 The above screenshot shows the **silhouette value for cluster = 3**

  
  
Output 7.4 The above screenshot shows the **silhouette value for cluster = 4**

  
  
Output 7.5 The above screenshot shows the **silhouette value for cluster = 5**

Output 7.6 The above screenshot shows the **silhouette value for cluster = 6**

Output 7.7 The above screenshot shows the **silhouette value for cluster = 7**



Output 7.8 The above screenshot shows the **silhouette value for cluster = 8**

Output 7.9 The above screenshot shows the **silhouette value for cluster = 9**

Output 7.10 The above screenshot shows the **silhouette value for cluster = 10**

Graphical user interface, application

Description automatically generated

**Table 7.11**

**Ans : I see silhouette score is 0.5 and we can conclude that it’s a nicely distinguished set of clusters and the 4 clusters which are created have a silhouette score of 0.5 hence silhouette score is better for 4 clusters than for 3 clusters. So 4 is the optimum number of clusters.**

**Part 1 - 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].**



**Output 8.1 The output screenshot is the label value of clusters 4.**

Graphical user interface

Description automatically generated with medium confidence **Table 8.2**

A screenshot of a computer

Description automatically generated with medium confidence **Table 8.3**

**Based on the optimum number of clusters details, we have added the column to the existing data frame by clustering the profile using silhouette score.**

Graphical user interface

Description automatically generated with low confidence

**Table 8.4 In the above screenshot, we can see that we have grouped the data based on the optimal number of clusters to identify the trends with respect to** Clicks, spend, revenue, CPM, CTR, & CPC.

Text

Description automatically generated

Output 8.5

Chart, bar chart

Description automatically generated  
Fig 8.6 The above bar plot represents the trends of Clicks with respect to Optimal number of clusters.

Text

Description automatically generated

Output 8.7

Chart, bar chart

Description automatically generated

Fig 8.8 The above bar plot represents the trends of Spend with respect to Optimal number of clusters.

Text

Description automatically generated

Output 8.9  
Chart, bar chart

Description automatically generated

Fig 8.10 The above bar plot represents the trends of Revenue with respect to Optimal number of clusters.

Text

Description automatically generated  
Output 8.11

Chart, bar chart

Description automatically generated

Fig 8.12 The above bar plot represents the trends of CTR with respect to Optimal number of clusters.

Text

Description automatically generated **Output 8.13**Chart, bar chart, histogram

Description automatically generated

Fig 8.14 The above bar plot represents the trends of CPM with respect to Optimal number of clusters.

Text

Description automatically generated  
Output 8.15   
Chart, bar chart, histogram

Description automatically generated  
Fig 8.16 The above bar plot represents the trends of CPC with respect to Optimal number of clusters.

**Part 1 - Clustering: Conclude the project by providing summary of your learnings.**

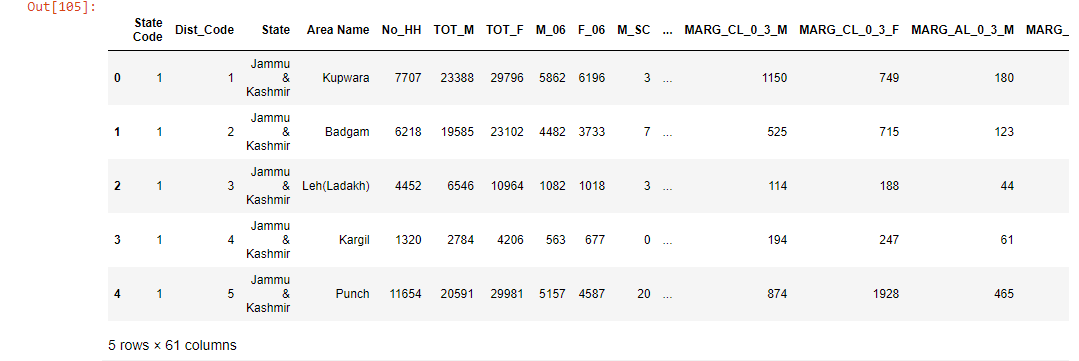
**Ans : From the given dataset, we can see there are 23066 observations with 19 columns. As we have already done the necessary basic data analysis by treating the NULL values in CPM , CPC and CTR columns with the help of formula provided. So then outliers were present and we have treated the outliers. Later, we did the visualization part with the help of dendrogram. We also created linkage using Ward and Euclidean function using dendrogram to visualize the output. Then we have done the analysis by using K-Means algorithm with values of k up to 10. And for each K value, we have also calculated silhouette score and decided the K value which is the optimal number of clusters needed.**

**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 file - [PCA India Data Census.xlsx](https://olympus.mygreatlearning.com/courses/81614/files/6750109/download?verifier=3jH6yoph8B7B2B4v23eVpTawyazso9x4uESkAZG5&wrap=1)

**Part 2 - PCA: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.**

**Head :**



**Table 9.1 We can see the top 5 rows of the dataset.**

**Tail :**

A screenshot of a computer

Description automatically generated with medium confidence

**Table 9.2 So the screenshot displays the last 5 rows.**

**Info :**   
Table

Description automatically generated Text, table

Description automatically generated

**Table 9.3 Here we can see the info details such as datatypes of the columns : 2 object columns and 59 integer columns are present.**

**Describe :**   
Table

Description automatically generated with low confidence  
Table

Description automatically generated  
**Table 9.4 It describes the statistical overall descripton data summary of the given dataset.**

**Null Values :**

Text, table

Description automatically generated with medium confidence  
**Table 9.5 We can see that there is no Null values present in the data.**

Duplicate values :  
**Output 9.6 We can see that there is no duplicate values.**

Part 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?

Chart, histogram

Description automatically generated

Fig 10.1 The graph shows the male population in various states.

Chart

Description automatically generated

Fig 10.2 The graph shows the female population in various states.

Graphical user interface, table

Description automatically generated with medium confidence  
Table 10.3 The above output is the combination of values with the help of crosstab using State , Tot\_M and Tot\_F columns.

  
Output 10.4 So maximum number of Males in the population is 485417.



Output 10.5 So maximum number of Females in the population is 485417.

**Ans : Maharashtra state has highest gender ratio with respect to males and females.**

Which state has the lowest gender ratio.   
Graphical user interface, application

Description automatically generated

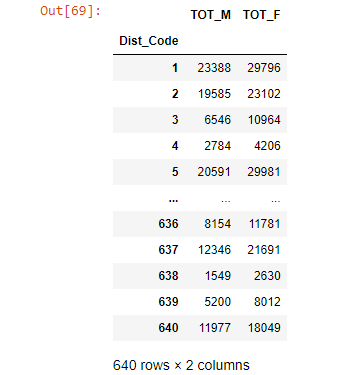


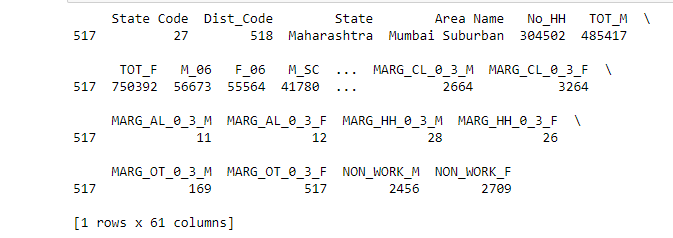
Table 10.6 Minimum number of males in the population is 391.

  
Output 10.7 Minimum number of females in the population is 698.

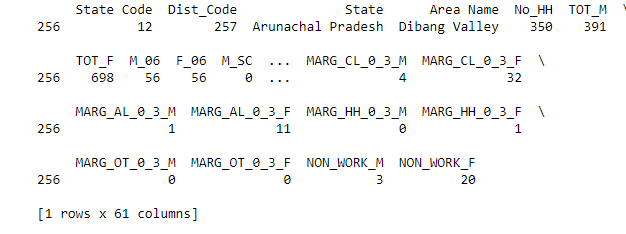
**Ans : Arunachal Pradesh state has the lowest gender ratio.**

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

  
Table 10.8 Using crosstab we can display the output of maximum number of Males and Females with respect to State wise.

  
Table 10.9

**Dist code 518 has highest gender ratio.**

  
Table 10.10

**Dist code 257 has lowest gender ratio.**

**What is the literacy rate in the given data set?**

  
Output 10.11

**So 5.96% is the literacy rate in the given data set.**

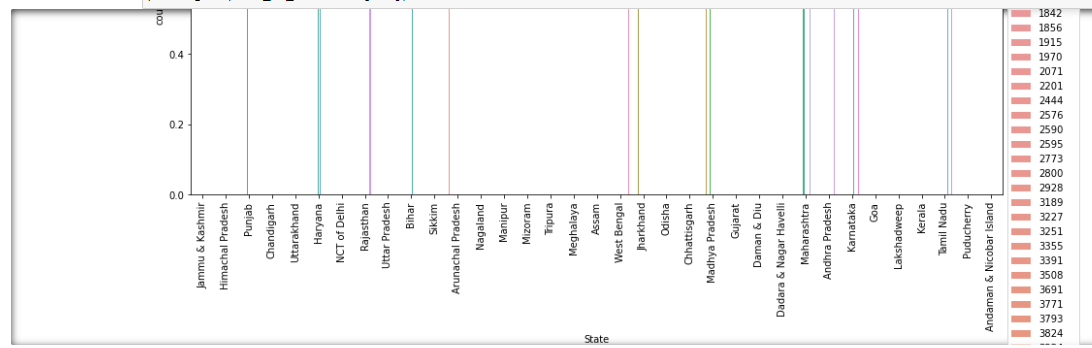


Fig 10.12 The graph shows the M\_LIT Male literates columns values based on State.

Chart, histogram

Description automatically generated

Fig 10.13 The graph shows the F\_LIT Female literates columns values based on State.

**Compare the Total Worker Population of male V/s Female**

  
Output 10.14

Total worker population of Male is 269422.

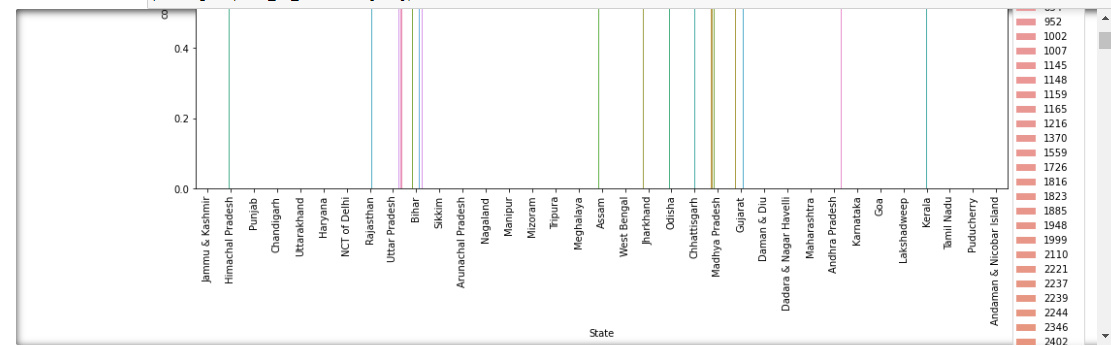


Fig 10.15 The above graph shows TOT\_WORK\_M total working males based on the State.



Output 10.16

Total worker population of Female is 257848.

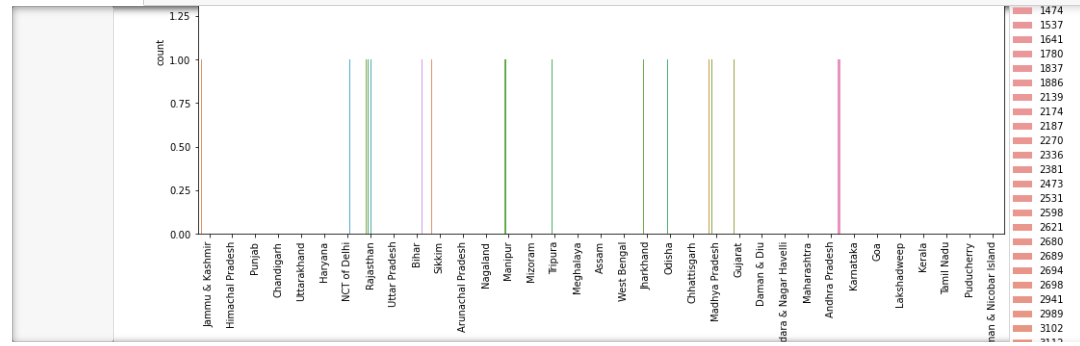
.

Fig 10.17 The above graph shows TOT\_WORK\_F total working females based on the State.

**Ans : So total working population of male is more than female.**

**Compare the on Working Population Male of male V/s Female**



Output 10.18

Maximum working main population of male is 247911.

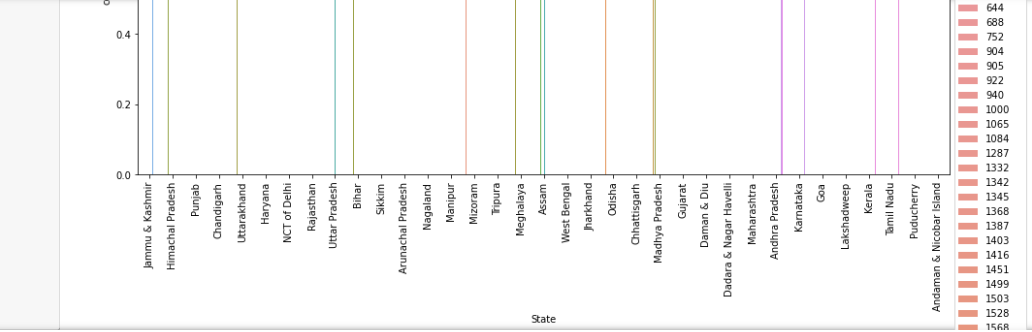


Fig 10.19 The above graph shows as MAINWORK\_M maximum main working of Male columns with respect to State.



Output : 10.20

Maximum working main population of female is 226166.

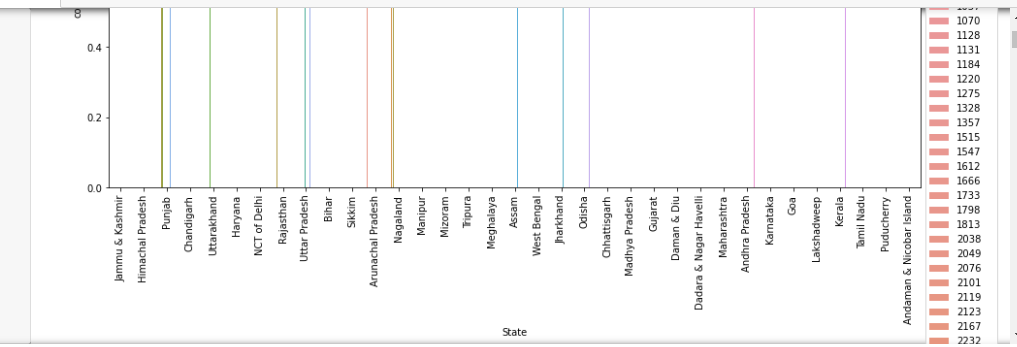


Fig 10.21 The graph shows as MAINWORK\_F maximum main working of Females columns with respect to State.

**So main working population of male is more when compared to female.**

**Exploratory Data Analysis :**

**Univariate Analysis : So we have taken variables as : TOT\_M , TOT\_F , M\_LIT , F\_LIT , TOT\_WORK\_M , TOT\_WORK\_F.**

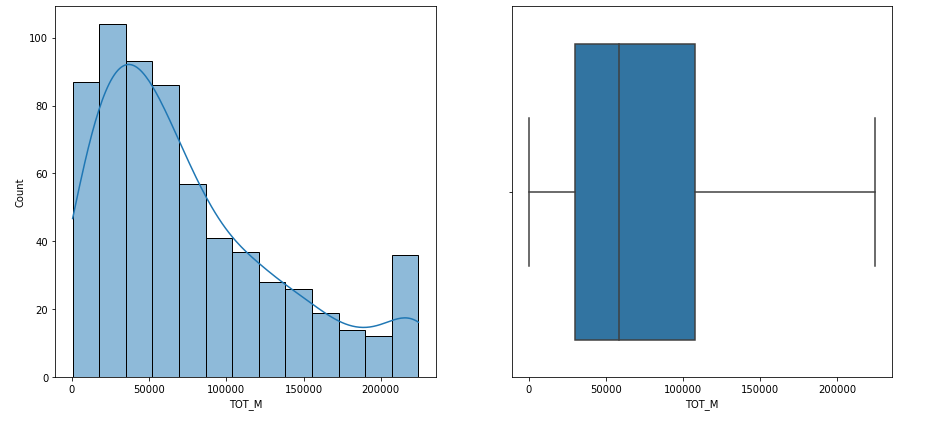


Fig 10.22 The histogram and boxplot visualization of TOT\_M columns.

Chart, histogram

Description automatically generated

Fig 10.23 The histogram and boxplot visualization of TOT\_F columns.

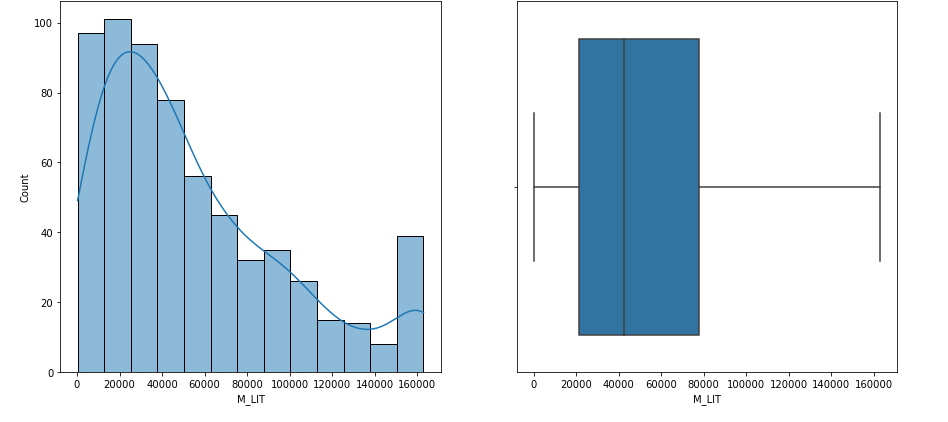


Fig 10.24 The output of Male literate values M\_LIT with the help of histogram and boxplot.

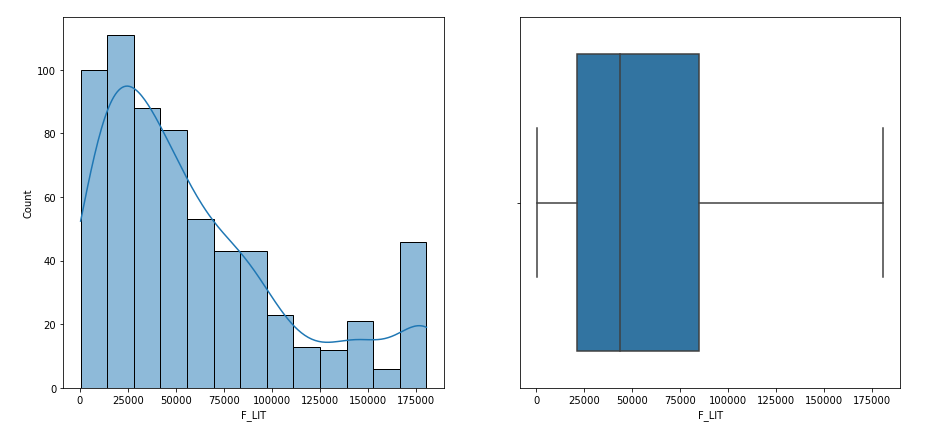


Fig 10.25 The output of Female literate values F\_LIT with the help of histogram and boxplot.

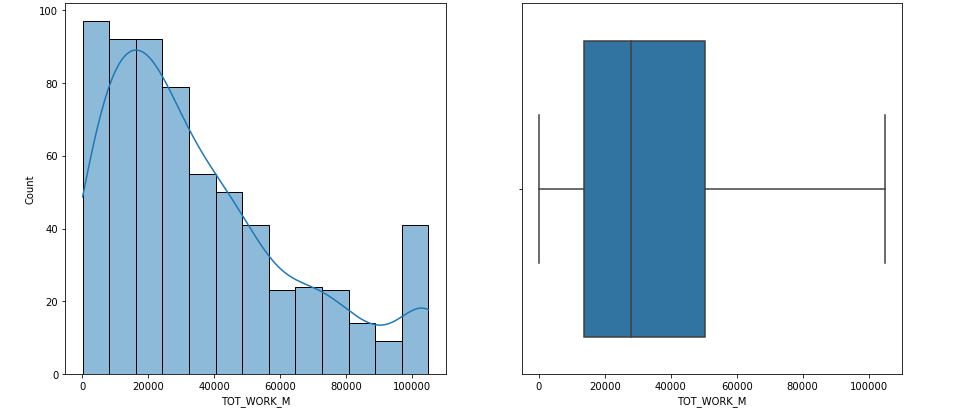


Fig 10.26 The TOT\_WORK\_M columns values being plotted using histogram and boxplot.

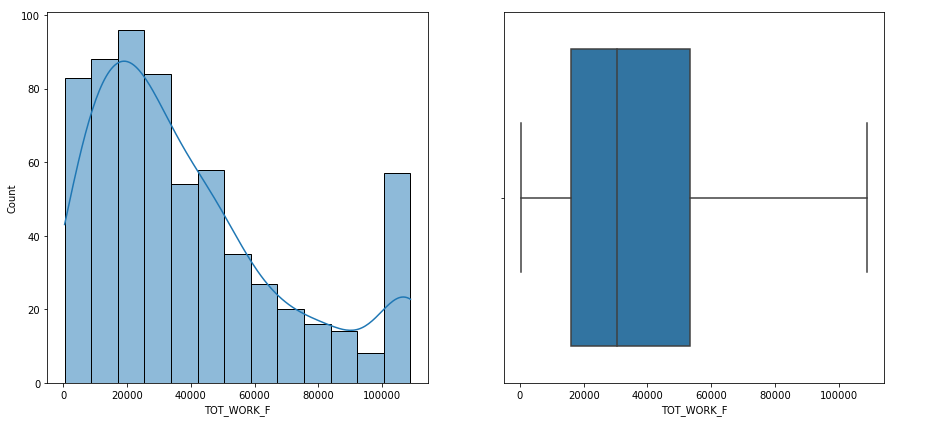


Fig 10.27 The above screenshot displays the histogram and boxplot values of TOT\_WORK\_F total working Females values.

**Bivariate Analysis :**

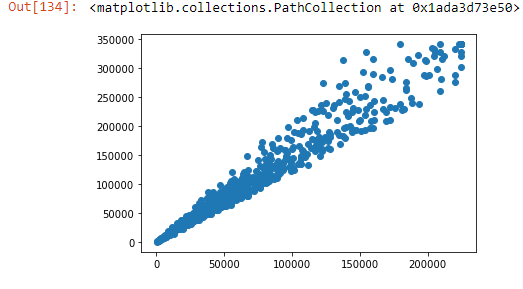


Fig 10.28 The scatterplot shows the plot values of TOT\_M and TOT\_F when bivariate analysis is made.

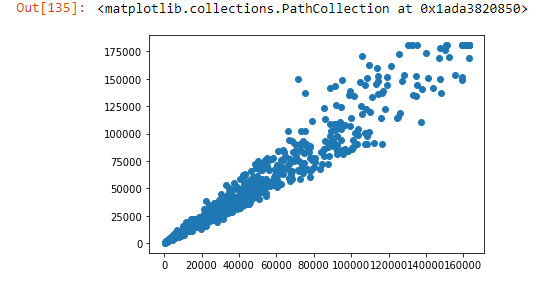


Fig 10.29 With the help of M\_LIT and F\_LIT column values using scatterplot analysis is being displayed.

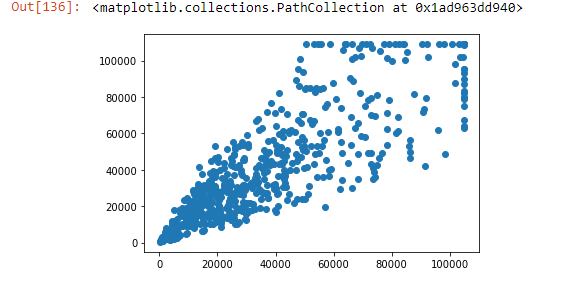


Fig 10.30 The above scatterplot shows the bivariate analysis done between TOT\_WORK\_M and TOT\_WORK\_F columns.

Part 2 - PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?

**Ans : Yes, we need to treat it. It's necessary to treat all the outliers in PCA.**

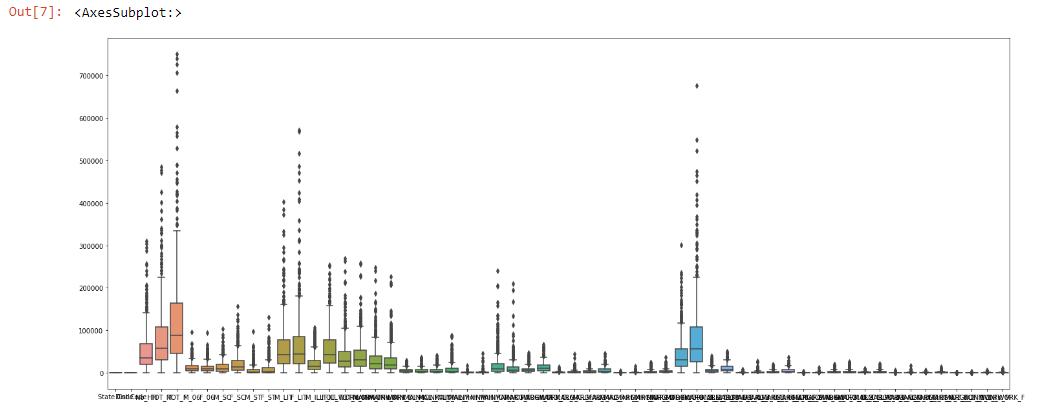
**Before treating outliers, we have removed 2 categorical columns such as 'State', 'Area Name’ from the dataset.**

Text

Description automatically generated  
Output 11.1

  
Output 11.2

**So after removing 2 columns we see now there are 59 values.**



**Fig 11.3 From the above graph, we can see there are outliers present in the dataset.**

**So after treating the outlier values, the following graph shows now that there is no outlier present in the data.**

A picture containing diagram

Description automatically generated

Fig 11.4

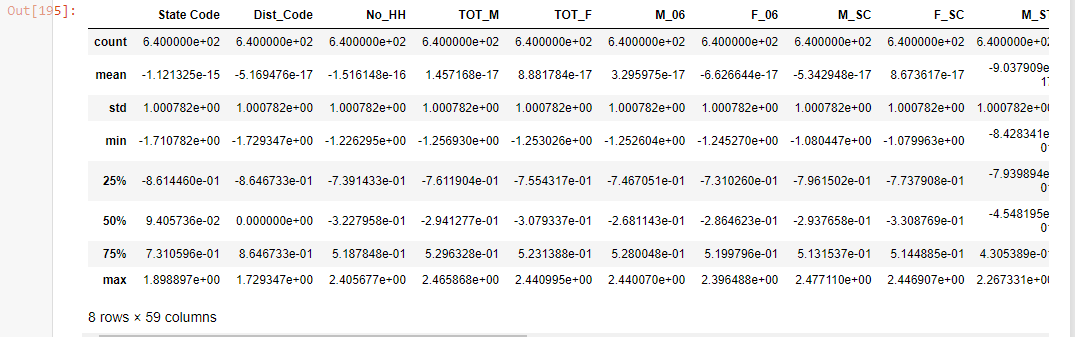
Part 2 - PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.

Graphical user interface, text, application

Description automatically generated

Output 12.1 Output of scaled data using z-score method.

  
Output 12.2 The above screenshot shows the values before scaling.



Output 12.3 After scaling, the values show as depicted in the above screenshot.

**Screenshot of outliers before scaling :**

A picture containing diagram

Description automatically generated

Fig 12.4

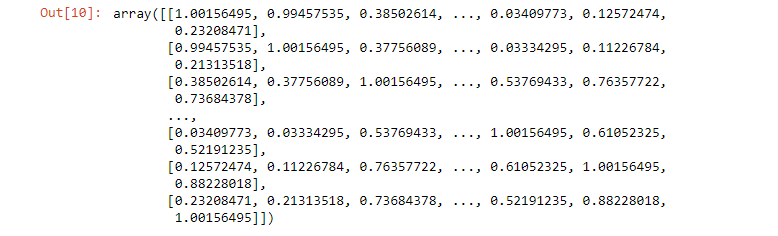
**Screenshot of outliers after scaling :**

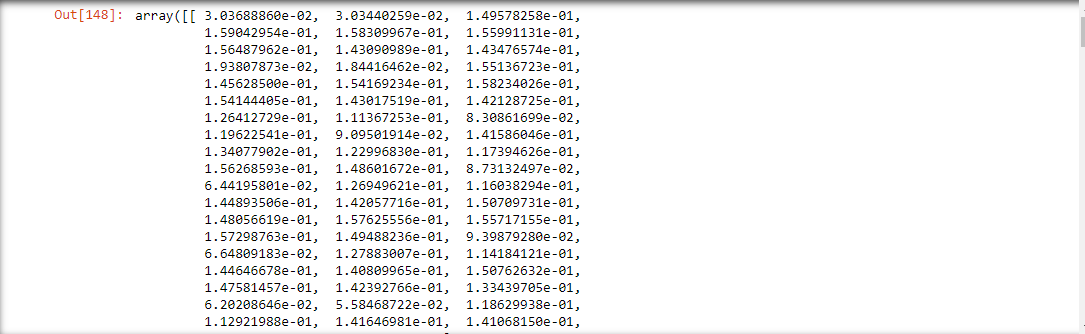


Fig 12.5

**Ans : Though the values of the dataset changes before scaling and after scaling but scaling does not have any impact on the outliers.**

Part 2 - PCA: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.

  
**Output 13.1 covariance Matrix output.**



**Output 13.2 Output of eigen vectors.**

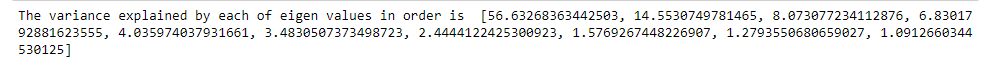


**Output 13.3 Output of eigen values.**



Output 13.4

Part 2 - PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.



Output 14.1 Output of the variance explained by each of eigen values.

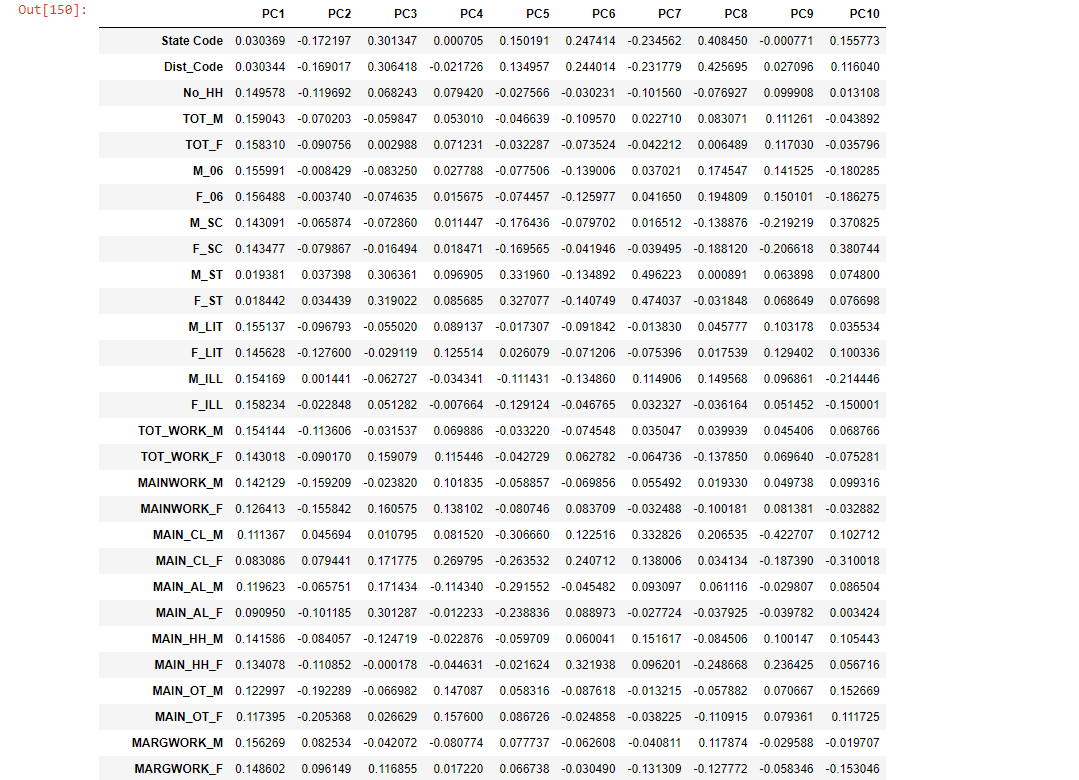
Text

Description automatically generated

Output 14.2 Output values of Cumulative Variance Explained.



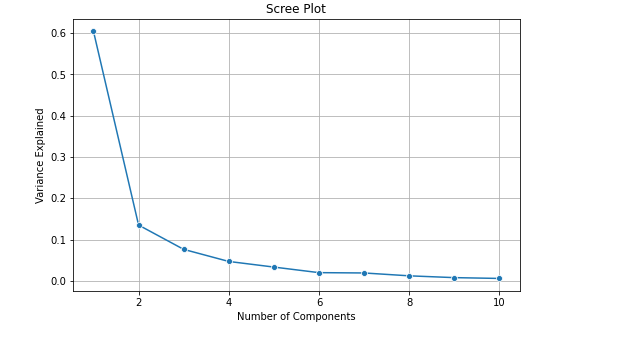
Output 14.3 Output of PCA explained variance ratio.



Table

Description automatically generated

Output 14.4 So we have extracted the loadings for 10 PC components.

  
Fig 14.5



**Output 14.6 The above screenshot shows the output of the cumulative variance explained values.**



Output 14.7 So the above mentioned values is the output of PCA explained variance ratio.



Output 14.8

**Ans : With the help of Scree Plot, we can see that 5 is the optimum number of PCs.**

Graphical user interface, application, table

Description automatically generated  
Output 14.9

Application

Description automatically generated with low confidence

Output 14.20 The above screenshot shows as extracted loadings for 5 PC components.

Part 2 - 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.

Chart, histogram

Description automatically generated

Fig 15.1

**Ans : PC5 Component is the most variance. So from the above screenshot, we can see in PC5 the value of the progress in each variables is linear in status and in PC2 component initially the variables values shows as high and later it was decreased and in PC3 components we can see rapid decrease in the variables values and in PC4 component we can see variables values decreasing in a significantly manner and in PC5 component we can standard progress in terms of actual variables in the given dataset.**



Fig 15.2

**From the above heatmap, we can conclude that in PC1 component, has a proper correlation relationship between each variables. In PC1 component, we can see all the types of correlations in a proper segregated manner which has low correlation , medium and low correlation found in each variables with respect to component PCA1.**

**Part 2 - PCA: Write linear equation for first PC.**

**Ans : PC1 = a1x1 + a2x2 + a3X3 + a4X4 + a5X5 + a6X6 …….+ a59x59**