GRADED PROJECT ON DATA MINING

Clustering and Principal Component Analysis

Abstract

Use of Python Library Sklearn and Ward linkage for Dendrogram Nov 22

Vikash Kumar

DSBA 2022

Oct 28, 8:00 AM

Description

Dear Participants,

Please find below the Data Mining Project instructions:

- Submissions: 2 separate files

1. **Business Report: Submit answers to all the questions in a sequential manner.** Your report must **include a detailed explanation of the approach taken, inferences, and insights. Include outputs such as graphs, tables, and all other relevant information.** Business Report must not include any codes. **You will be evaluated based on Business Report only**. Hence please ensure that your Business Report is logical and detailed enough (without any code) for a reader somewhat conversant in analytics to understand the solution mechanism. 6 Marks are allotted for the "Quality of Business Report".
2. **Jupyter Notebook File**: This is a must and will be used for reference while evaluating

- Any assignment found copied/ plagiarized by another person will not be graded and marked as zero.
- Please ensure timely submission as a post-deadline assignment will not be accepted.

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 Excel File.

Perform the following in given order:

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

Ans:

1.The showing right skew distribution for available Impressions, Matched Queries, Impressions Clicks, Clicks, Revenue, CPM.

2.On the basis of above description the most of data would have the outliers present.

3.Data is unscaled.

4.Data has 23066 rows and 19 columns

5.It indicates values either heavily tailed or highly skewed.

6.The Data info of Data set to check the Variables, nulls, Data Types, Total Columns and rows.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23066 entries, 0 to 23065
Data columns (total 19 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Timestamp              23066 non-null  object
 1   InventoryType          23066 non-null  object
 2   Ad - Length            23066 non-null  int64
 3   Ad- Width              23066 non-null  int64
 4   Ad Size                23066 non-null  int64
 5   Ad Type                23066 non-null  object
 6   Platform               23066 non-null  object
 7   Device Type            23066 non-null  object
 8   Format                 23066 non-null  object
 9   Available_Impressions  23066 non-null  int64
 10  Matched_Queries        23066 non-null  int64
 11  Impressions            23066 non-null  int64
 12  Clicks                 23066 non-null  int64
 13  Spend                  23066 non-null  float64
 14  Fee                    23066 non-null  float64
 15  Revenue                23066 non-null  float64
 16  CTR                    18330 non-null  float64
 17  CPM                    18330 non-null  float64
 18  CPC                    18330 non-null  float64
dtypes: float64(6), int64(7), object(6)
memory usage: 3.3+ MB
```

7.The last three columns of missing values by 18330 which is treated as per instruction.

8.The data set having 19 columns which many find as not usefull for the analysis, hence dropping the mentioned columns 'Ad - Length','Ad- Width','Ad Size','Timestamp','InventoryType','Ad Type','Platform','Matched_Queries','Fee','Format' from axis=1

The details has been checked with python and inynb file attached for su pport understandings.

- Treat missing values in CPC, CTR and CPM using the formula given. You may refer to the Bank_KMeans Solution File 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.

Ans:

Treating the missing values as per instruction as

CTR = (Clicks/(Impressions)*100)

CPM = (Spend/(Impressions)*1000)

CPC = (Spend/(Clicks))

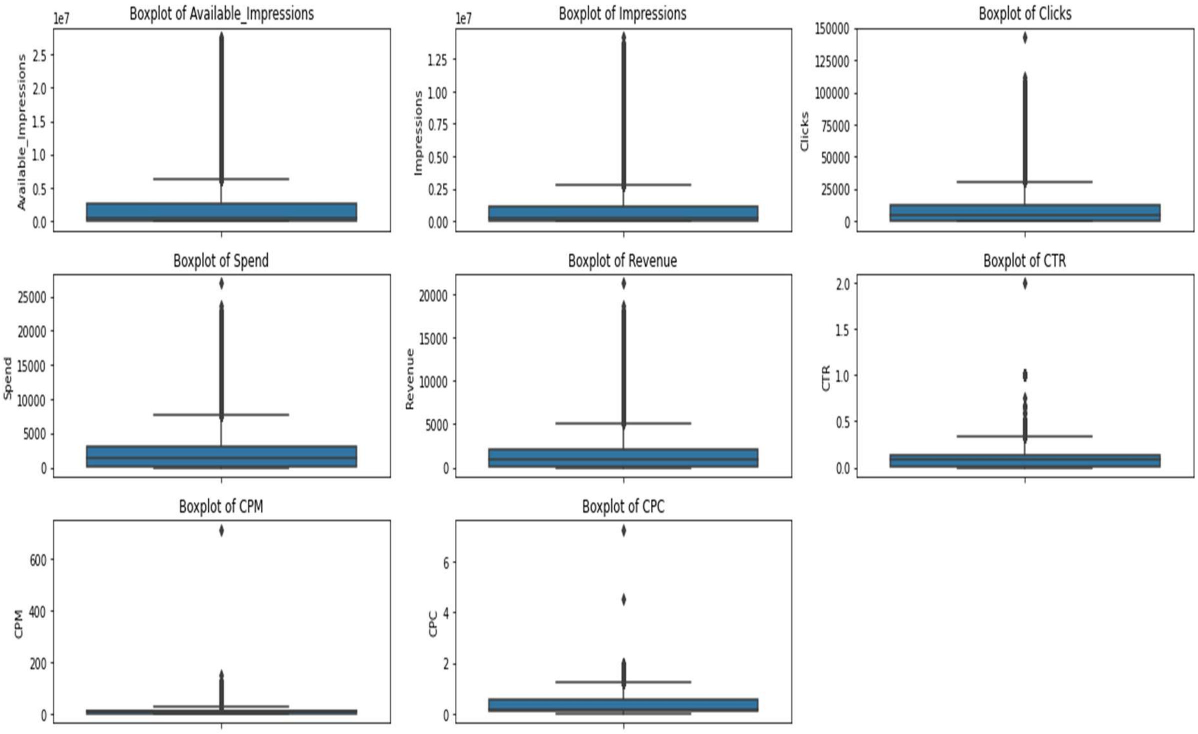
Then the info of data area

```
<class 'pandas. core. frame. DataFrame'>
Range Index: 23066 entries, 0 to 23065
Data columns (total 9 columns):
 #   Column                 Non-Null Count Dtype
---  ------                 -------------- -----
 0   Device Type            23066 non-null object
 1   Available Impressions  23066 non-null  int64
 2   Impressions            23066 non-null int64
 3   Clicks                 23066 non-null int64
 4   Spend                  23066 non-null float64
 5   Revenue                23066 non-null float64
 6   CTR                    23066 non-null float64
 7   CPM                    23066 non-null float64
 8   CPC                    23066 non-null float64

dtypes: float64(5), int64(3), object (1)
Which seems the all missing values are treated.
```

- Check if there are any outliers.

  Checked the outliers with proper code and found, There is presence of large number of extreme data are in almost all variables except ad length and ad-width. The outliers are seems as justified.

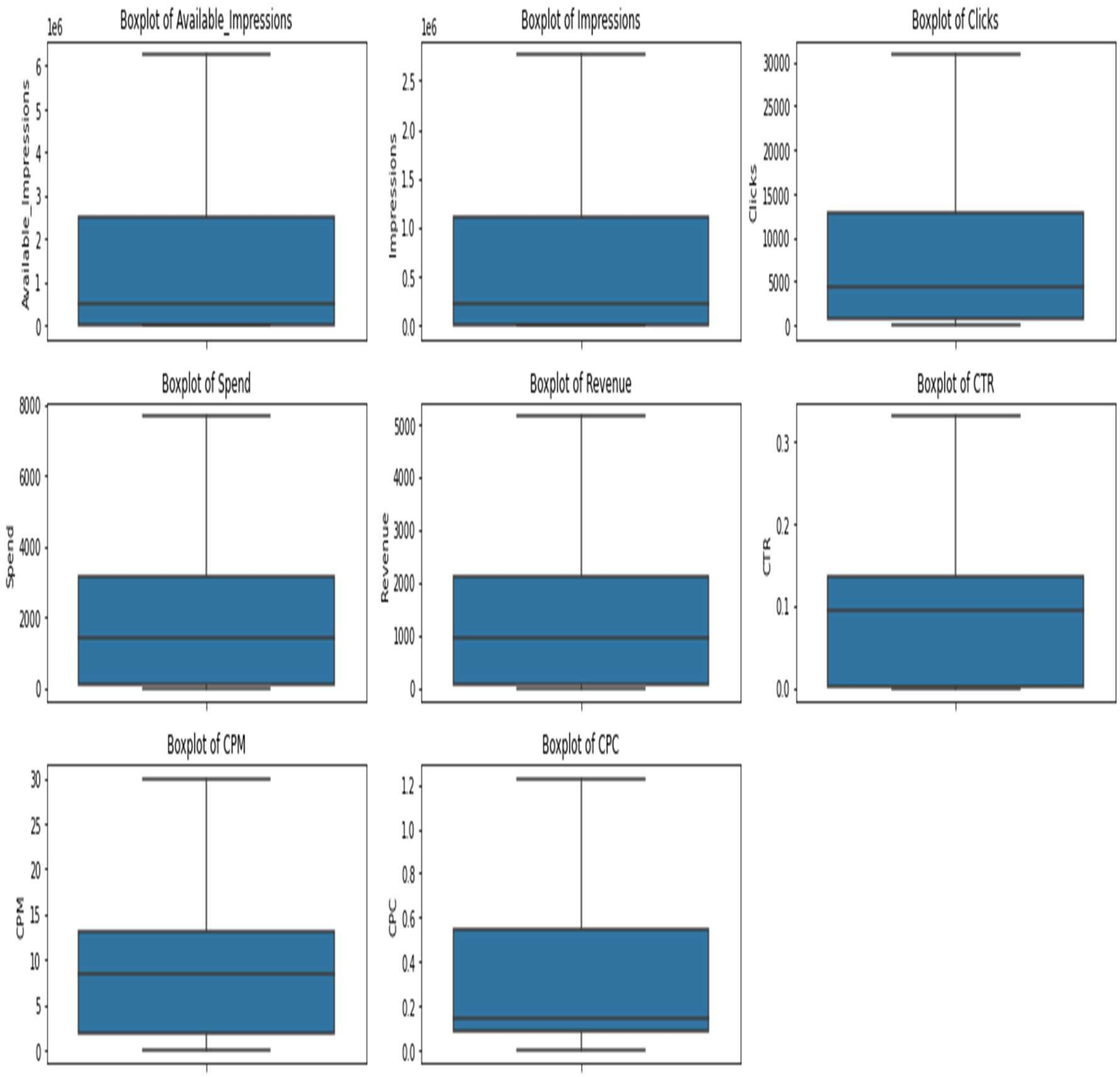


- Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an analyst your judgement may be different from another analyst).

  Ans: **Outlier treatment required**

  1.Since K-Means algorithm is about finding mean of clusters, the algorithm is sensitive to Outliers. The centroids will not be a true representation of a cluster in the presence of outliers.
  2.The sum of squared errors (SSE) will also be very high in the case of outliers. Small Clusters will bond with outliers, which may not be the true representation of the natural patterns of clusters in data.
  3.Hence it is better to identify and remove outliers before applying K-means clustering algorithm.

  After Treatment of Outliers:



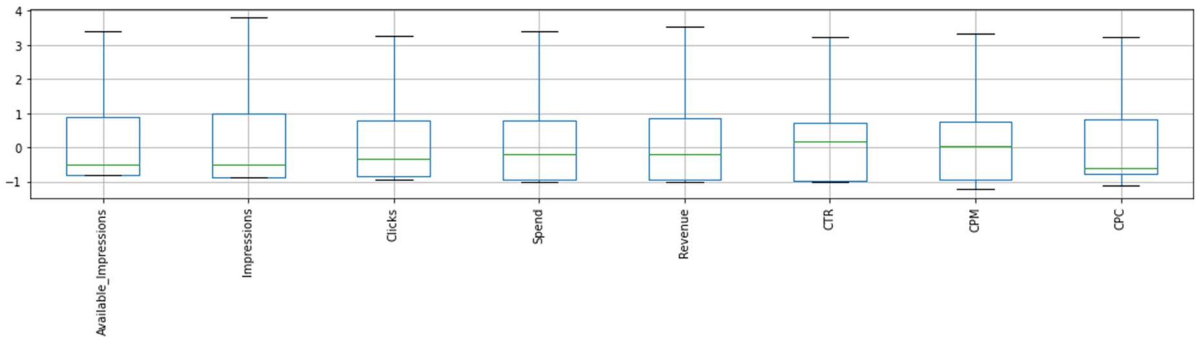
The Tukey's method defines an outlier as those values of the data set that fall far from the central point, the median

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

The data statistics summary after scaling of data

	count	mean	std	min	25%	50%	75%	max
Available Impressions	23066.000	0.000	1.000	-0.821	-0.798	-0.496	0.877	3.390
Impressions	23066.000	0.000	1.000	-0.879	-0.866	-0.497	1.006	3.809
Clicks	23066.000	-0.000	1.000	-0.950	-0.854	-0.348	0.792	3.259
Spend	23066.000	0.000	1.000	-1.004	-0.955	-0.189	0.781	3.385
Revenue	23066.000	-0.000	1.000	-1.008	-0.959	-0.190	0.839	3.535
CTR	23066.000	-0.000	1.000	-1.022	-0.989	0.180	0.703	3.216
CPM	23066.000	-0.000	1.000	-1.226	-0.960	0.046	0.756	3.331
CPC	23066.000	-0.000	1.000	-1.102	-0.786	-0.611	0.823	3.231

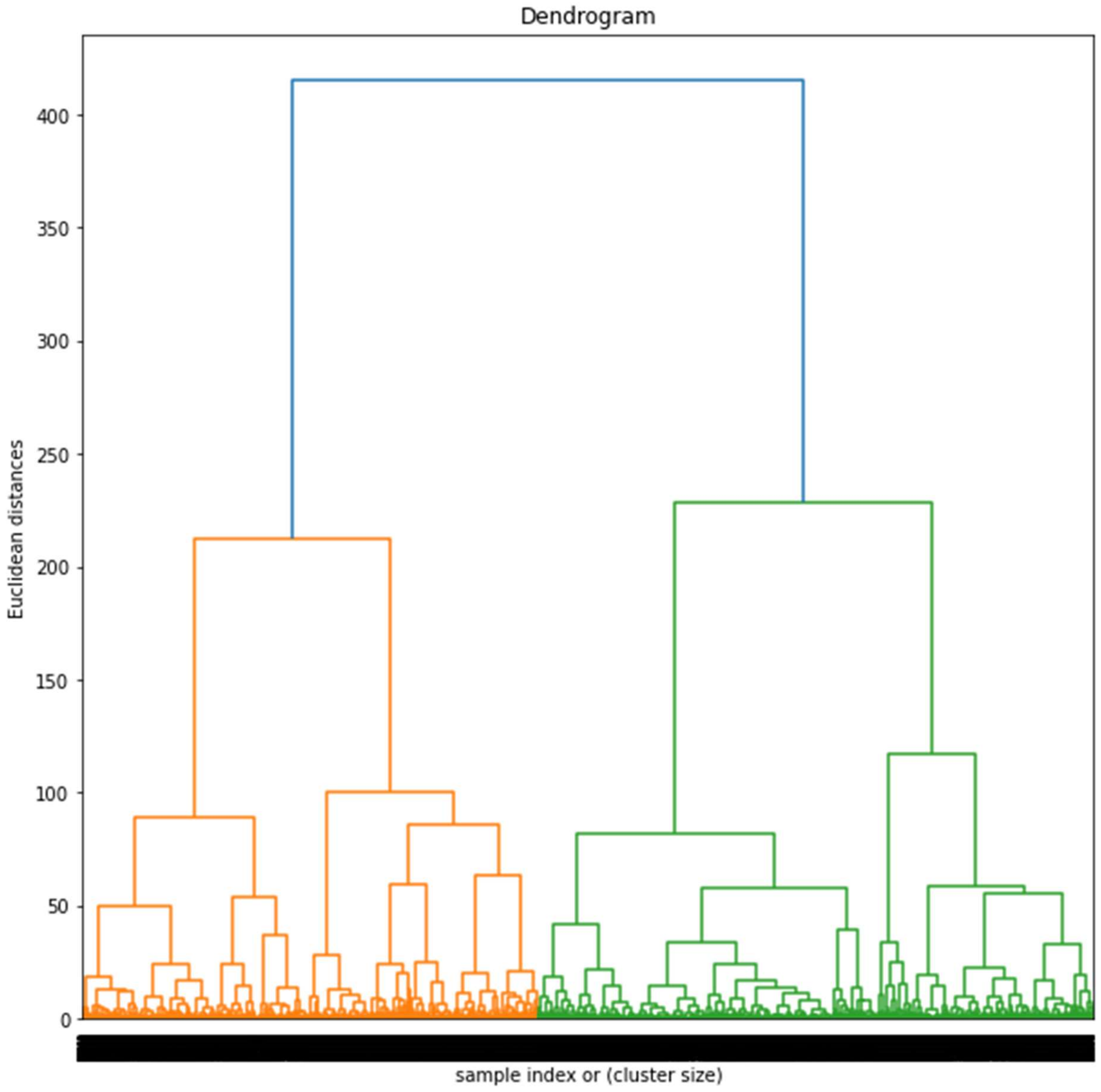
Box plot of scaled Data



Discussion on choosing Z score:
1.The benefit of performing the Z score normalization is that the clear outlier in the dataset has been transformed in such a way that it's no longer a massive outlier.
2.The Z-score of an observation is defined as the number of standard deviations it falls above or below the mean, in other words it computes the variance (i.e. distance). Clustering data modelling technique needs normalization, in the sense that it requires to compute the Euclidean distance. The Z-score is suited well and is essential to compare similarities between attributes based on certain distance measure. The same applies to Principal Components Regression (PCR); in it we are interested in the components that maximize the variance.

3.These speeds up the algorithm to a great extent as many elements are skipped. If this technique is not used, the algorithm would perform computations for all the elements, and thus get reduced to a quadratic [O(n^2n2)] algorithm, equivalent to naive pattern searching.
4.Better computation. Computationally, it can speed up the calculation due to rounding errors. Computers, like humans, work better with numbers that are on a similar scale. I think modern software does this on its own, but not always.
5.Interpretation. It can improve interpretation, particularly when you are comparing across variables with very different scales and means. You can even sometimes compare variables in different experiments.

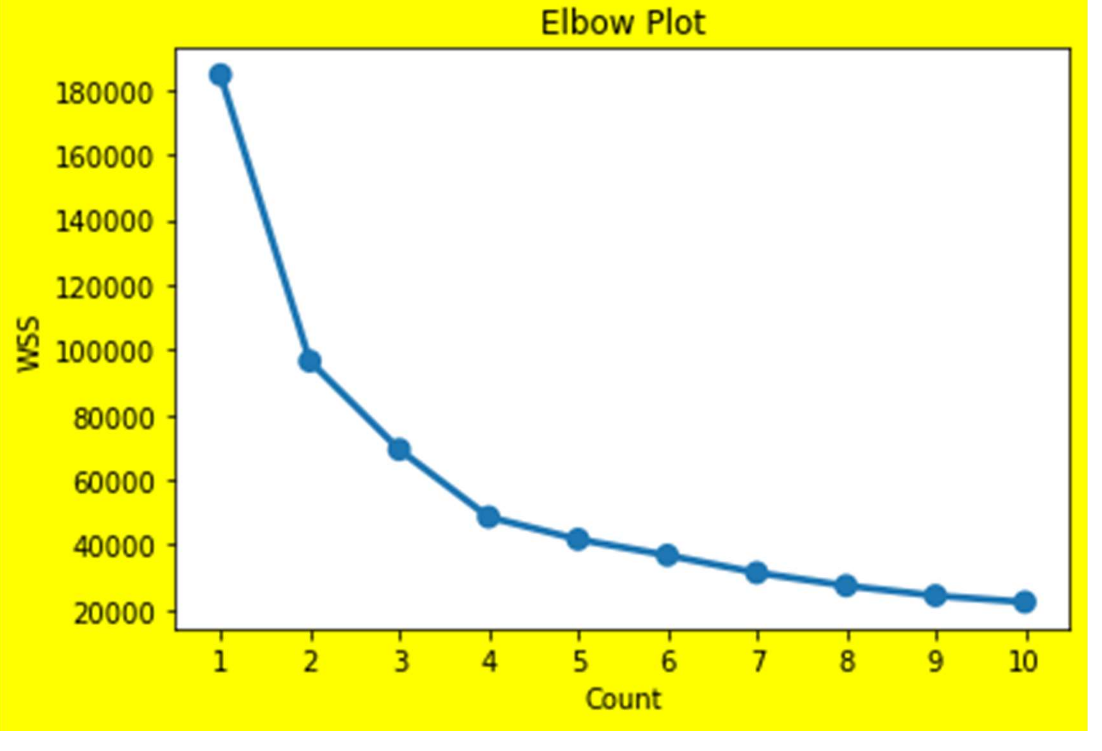
- Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.



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

Ans: The Elbow plot for the n=10 . For this the within Sum Square Values (WSS) are

```
[184528.0000000003,
 96470.2073112208,
 69289.05949170691,
 48522.558129064404,
 41640.79126469919,
 36521.95814058762,
 31367.195138734925,
 27406.77378043877,
 24315.29472637937,
 22368.054758388134,
 184528.0000000003,
 96470.5645847809,
 69288.97397747991,
 48522.55825283969,
 41664.45016839325,
 36752.61508590262,
 31367.194592449854,
 27386.331464776988,
 24335.836550185726,
 22316.62098917462]
```



There is sharp declination in WSS of cluster ONE (184528.000) to Cluster TWO (96470.207) while comparatively low difference for Two and THREE(69289.059) and so on.

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

```
The silhouette_score for 1 clusters is 0.05373740761126703
The silhouette_score for 2 clusters is 0.05264193383194291
The silhouette_score for 3 clusters is 0.05454532245189112
The silhouette_score for 4 clusters is 0.056588986821948635
The silhouette_score for 5 clusters is 0.05091745040075091
The silhouette_score for 6 clusters is -0.03372506432855011
The silhouette_score for 7 clusters is 0.05301400790186797
The silhouette_score for 8 clusters is 0.04826677062965853
The silhouette_score for 9 clusters is 0.041923100696678636
The silhouette_score for 10 clusters is 0.053423255185931086
```

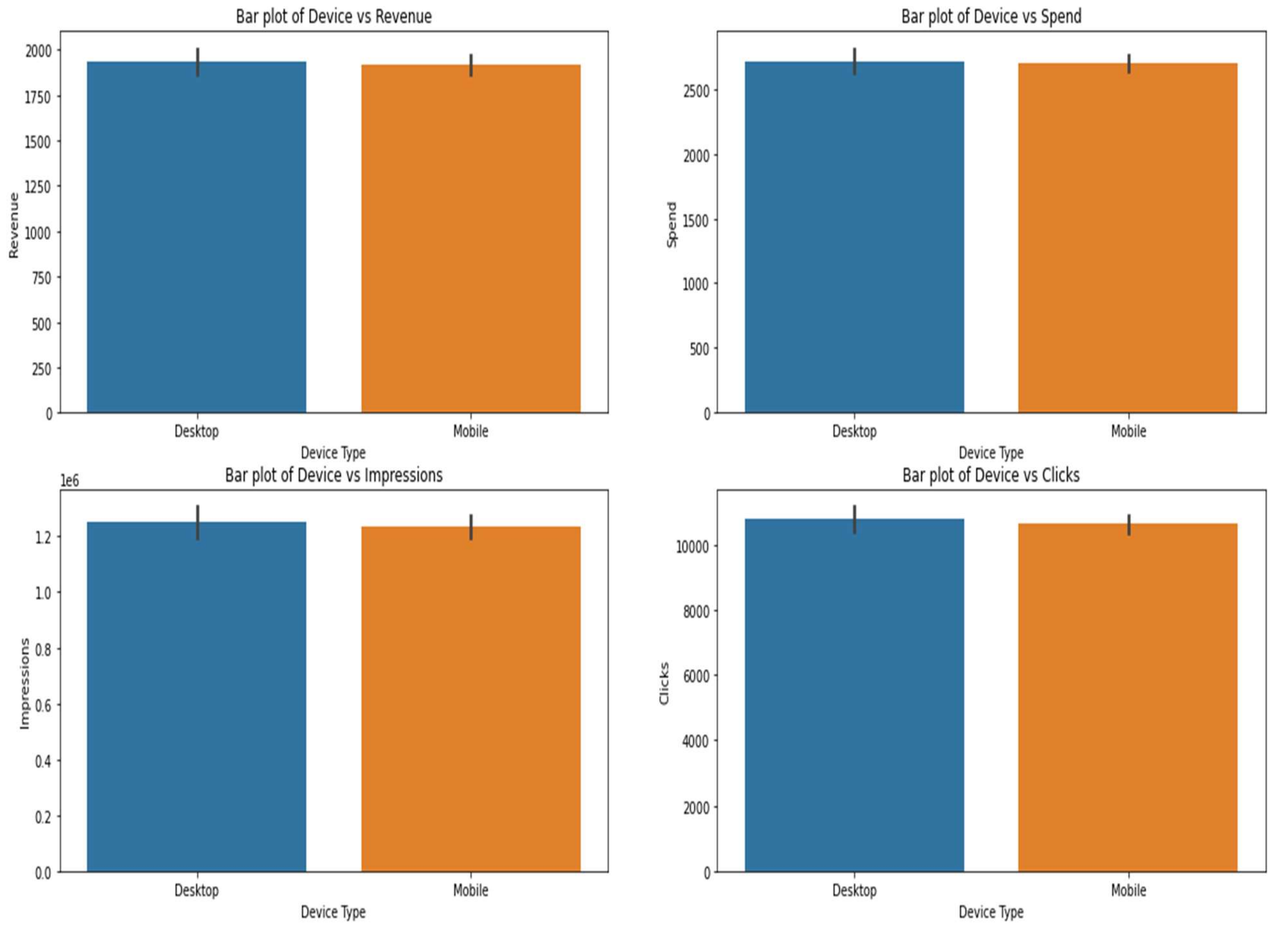
1.Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighbouring clusters.
2.A value of 0 indicates that the sample is on or very close to the decision boundary between two neighbouring clusters.

Clusters	Counts
0	4957
1	5284
2	7753
3	5072
Name: Clus_kmeans	

Deciding the Number of Clusters
1.On The basis of ELBOW plot insight we can observe that the turning point of this curve is at the value of k = 4. Therefore, we can say that the 'right' number of clusters for this data is 5.
2.As we can observe, the value of k = 4(**0.056)** in Silhouette score has the highest value i.e . nearest to +1. So, we can say that the optimal value of 'k' is 4.

- 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.]

		Impression	**Clicks**	**Spend**	**Revenue**	**CTR**	**CPM**	**CPC**	**sil_width**
Clus_kmeans	**Device Type**								
0	Desktop	257299.4	34156.38	3383.062	2346.784	0.132	12.869	0.097	0.362
	Mobile	257865.1	34067.38	3401.626	2359.985	0.131	12.91	0.098	0.363
1	Desktop	678356.1	3110.616	1245.749	809.771	0.004	1.798	0.494	0.458
	Mobile	691818.6	3105.992	1255.563	816.116	0.004	1.783	0.5	0.459
2	Desktop	11137.34	1456.575	144.549	93.957	0.161	14.33	0.099	0.572
	Mobile	11171.82	1453.733	147.34	95.771	0.162	14.524	0.099	0.572
3	Desktop	4695835	9869.9	7475.291	5465.488	0.002	1.613	0.76	0.33
	Mobile	4642330	9752.109	7458.942	5452.519	0.002	1.63	0.769	0.327



1.Both the gadget are (Desktop and Mobile) are supporting the particular advertisement with little difference in Impressions: 1214
Clicks :133
Spend :7
Revenue:5
supports higher on Desktop.That means the desktops having edge over Mobile to some extent as per data.
2.Where as CTR =0, CPM, CPC having edge on Mobile.

- Conclude the project by providing summary of your learnings.

The Dataset containing the variables which has of almost no use for analysis so it has been dropped. The Three variables which have missing values also are found derived and treat with the derivation given.
The Dendrogram and Elbow graph are key component of clustering. On the basis of it we can find the Euclidean distance and linkages among. The treatment of outliers are import feature for better clustering and perform with accuracy.
The K means and Silhouette score has major role in deciding the number of clusters and Frequency. Where as within sum square gives the path for further of clustering.
This analysis allows an object not to be part or strictly part of a cluster, which is called the hard partitioning of this type. However, smooth partitions suggest that each object in the same degree belongs to a cluster. More specific divisions can be created like objects of multiple clusters, a single cluster can be forced to participate, or even hierarchic trees can be constructed in group relations.

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.

- **Note: The 24 variables given in the Rubric is just for performing EDA. You will have to consider the entire dataset, including all the variables for performing PCA.**
  Data file - PCA India Data Census.xlsx

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 640 entries, 0 to 639
Data columns (total 61 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   State Code      640 non-null    int64 
 1   Dist.Code       640 non-null    int64 
 2   State           640 non-null    object
 3   Area Name       640 non-null    object
 4   No_HH           640 non-null    int64 
 5   TOT_M           640 non-null    int64 
 6   TOT_F           640 non-null    int64 
 7   M_06            640 non-null    int64 
 8   F_06            640 non-null    int64 
 9   M_SC            640 non-null    int64 
 10  F_SC            640 non-null    int64 
 11  M_ST            640 non-null    int64 
 12  F_ST            640 non-null    int64 
 13  M_LIT           640 non-null    int64 
 14  F_LIT           640 non-null    int64 
 15  M_ILL           640 non-null    int64 
 16  F_ILL           640 non-null    int64 
 17  TOT_WORK_M      640 non-null    int64 
 18  TOT_WORK_F      640 non-null    int64 
 19  MAINWORK_M      640 non-null    int64 
```

```
 20  MAINWORK_F      640 non-null    int64
 21  MAIN_CL_M       640 non-null    int64
 22  MAIN_CL_F       640 non-null    int64
 23  MAIN_AL_M       640 non-null    int64
 24  MAIN_AL_F       640 non-null    int64
 25  MAIN_HH_M       640 non-null    int64
 26  MAIN_HH_F       640 non-null    int64
 27  MAIN_OT_M       640 non-null    int64
 28  MAIN_OT_F       640 non-null    int64
 29  MARGWORK_M      640 non-null    int64
 30  MARGWORK_F      640 non-null    int64
 31  MARG_CL_M       640 non-null    int64
 32  MARG_CL_F       640 non-null    int64
 33  MARG_AL_M       640 non-null    int64
 34  MARG_AL_F       640 non-null    int64
 35  MARG_HH_M       640 non-null    int64
 36  MARG_HH_F       640 non-null    int64
 37  MARG_OT_M       640 non-null    int64
 38  MARG_OT_F       640 non-null    int64
 39  MARGWORK_3_6_M  640 non-null    int64
 40  MARGWORK_3_6_F  640 non-null    int64
 41  MARG_CL_3_6_M   640 non-null    int64
 42  MARG_CL_3_6_F   640 non-null    int64
 43  MARG_AL_3_6_M   640 non-null    int64
 44  MARG_AL_3_6_F   640 non-null    int64
 45  MARG_HH_3_6_M   640 non-null    int64
 46  MARG_HH_3_6_F   640 non-null    int64
 47  MARG_OT_3_6_M   640 non-null    int64
 48  MARG_OT_3_6_F   640 non-null    int64
 49  MARGWORK_0_3_M  640 non-null    int64
 50  MARGWORK_0_3_F  640 non-null    int64
 51  MARG_CL_0_3_M   640 non-null    int64
 52  MARG_CL_0_3_F   640 non-null    int64
 53  MARG_AL_0_3_M   640 non-null    int64
 54  MARG_AL_0_3_F   640 non-null    int64
 55  MARG_HH_0_3_M   640 non-null    int64
 56  MARG_HH_0_3_F   640 non-null    int64
 57  MARG_OT_0_3_M   640 non-null    int64
 58  MARG_OT_0_3_F   640 non-null    int64
 59  NON_WORK_M      640 non-null    int64
 60  NON_WORK_F      640 non-null    int64
dtypes: int64(59), object(2)
```

Insights:

1.There is No Missing and Duplicated data in data set.so seems data are good for further analysis.
2.Data have originally total 61 columns and 640 rows in which the dtypes are int64(59) and object(2)..
3. Most of the cases the Mean value and standard deviation are not too far,but data seems as right skewed.
4.Almost 18 variables has "0" Minimum value and large maximum value which indicate towards outliers present in data.
5.Outlier seems due to Variability in the data

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 640 entries, 0 to 639
Data columns (total 39 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   State            640 non-null    object
 1   Area Name        640 non-null    object
 2   No_HH            640 non-null    int64 
 3   TOT_M            640 non-null    int64 
 4   TOT_F            640 non-null    int64 
 5   TOT_WORK_M       640 non-null    int64 
 6   TOT_WORK_F       640 non-null    int64 
 7   MARGWORK_M       640 non-null    int64 
 8   MARGWORK_F       640 non-null    int64 
 9   MARG_CL_M        640 non-null    int64 
 10  MARG_CL_F        640 non-null    int64 
 11  MARG_AL_M        640 non-null    int64 
 12  MARG_AL_F        640 non-null    int64 
 13  MARG_HH_M        640 non-null    int64 
 14  MARG_HH_F        640 non-null    int64 
 15  MARG_OT_M        640 non-null    int64 
 16  MARG_OT_F        640 non-null    int64 
 17  MARGWORK_3_6_M   640 non-null    int64 
 18  MARGWORK_3_6_F   640 non-null    int64 
 19  MARG_CL_3_6_M    640 non-null    int64 
 20  MARG_CL_3_6_F    640 non-null    int64 
 21  MARG_AL_3_6_M    640 non-null    int64 
 22  MARG_AL_3_6_F    640 non-null    int64 
 23  MARG_HH_3_6_M    640 non-null    int64 
 24  MARG_HH_3_6_F    640 non-null    int64 
 25  MARG_OT_3_6_M    640 non-null    int64 
 26  MARG_OT_3_6_F    640 non-null    int64 
 27  MARGWORK_0_3_M   640 non-null    int64 
 28  MARGWORK_0_3_F   640 non-null    int64 
 29  MARG_CL_0_3_M    640 non-null    int64 
 30  MARG_CL_0_3_F    640 non-null    int64 
 31  MARG_AL_0_3_M    640 non-null    int64 
 32  MARG_AL_0_3_F    640 non-null    int64 
 33  MARG_HH_0_3_M    640 non-null    int64 
 34  MARG_HH_0_3_F    640 non-null    int64 
 35  MARG_OT_0_3_M    640 non-null    int64 
 36  MARG_OT_0_3_F    640 non-null    int64 
 37  NON_WORK_M       640 non-null    int64 
 38  NON_WORK_F       640 non-null    int64 
dtypes: int64(37), object(2)
memory usage: 195.1+ KB
```

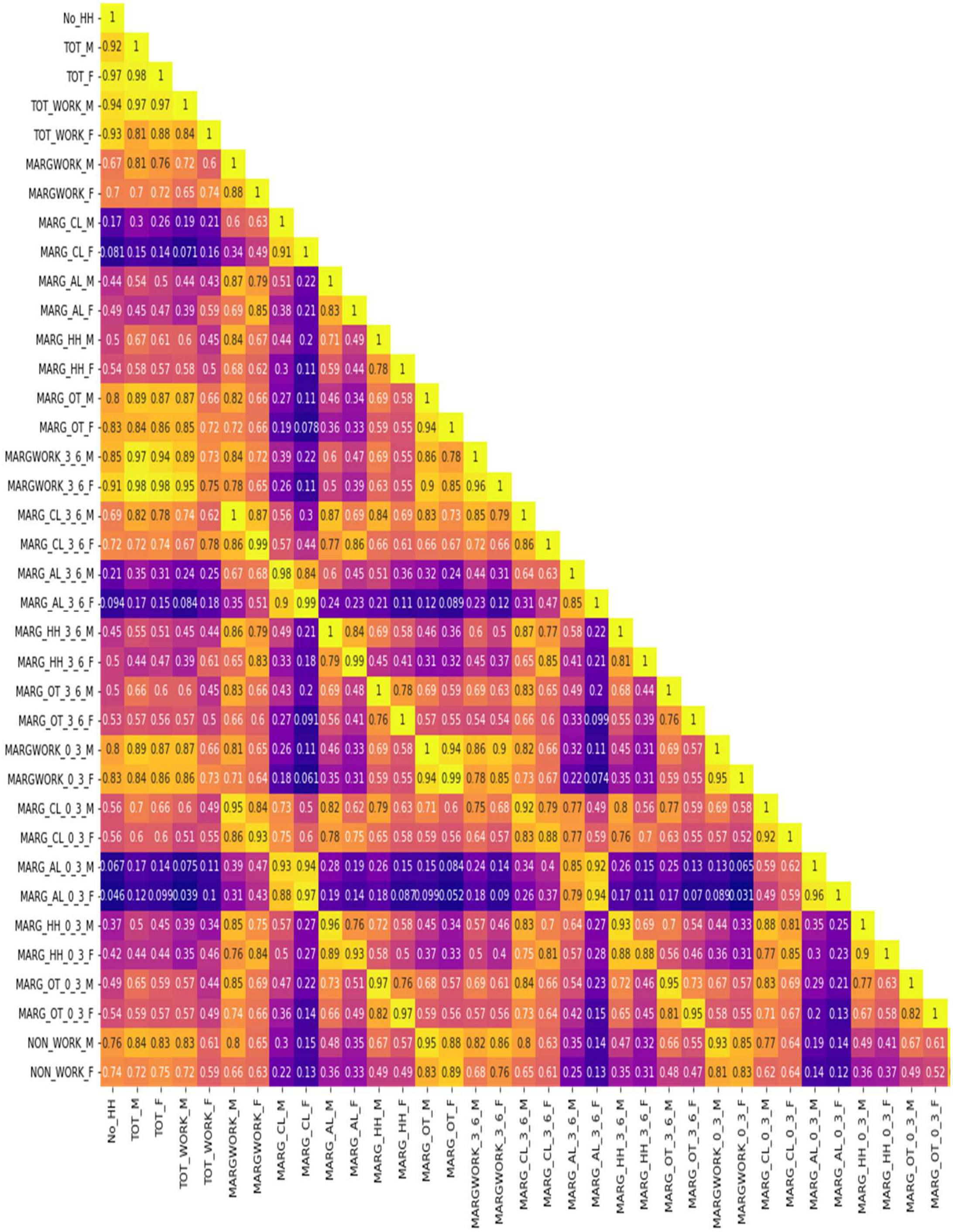
1.We are dropping the 19 more columns as instruction in case study

2. Data have now total 39 (0 to 38) columns and 640 (0 to 639)rows in which the dtypes are int64(37) and object(2).

3.We are dropping the column State code and dist code because for analysis there is State name and dist names are there.

4.Again No missing and duplicated value because it also a part of original Data which was noise free.
5.We are Picking 5 variables for further analysis like : **No_HH (**No of Household) , **TOT_M(**Total population Male), **TOT_F(**Total population Female), **TOT_WORK_M(**Total Worker Population Male) , **TOT_WORK_F(**Total Worker Population Female)

6.The EDA (Both Univariate and Multi variate)has been performed in Attached Jupyter Note book.

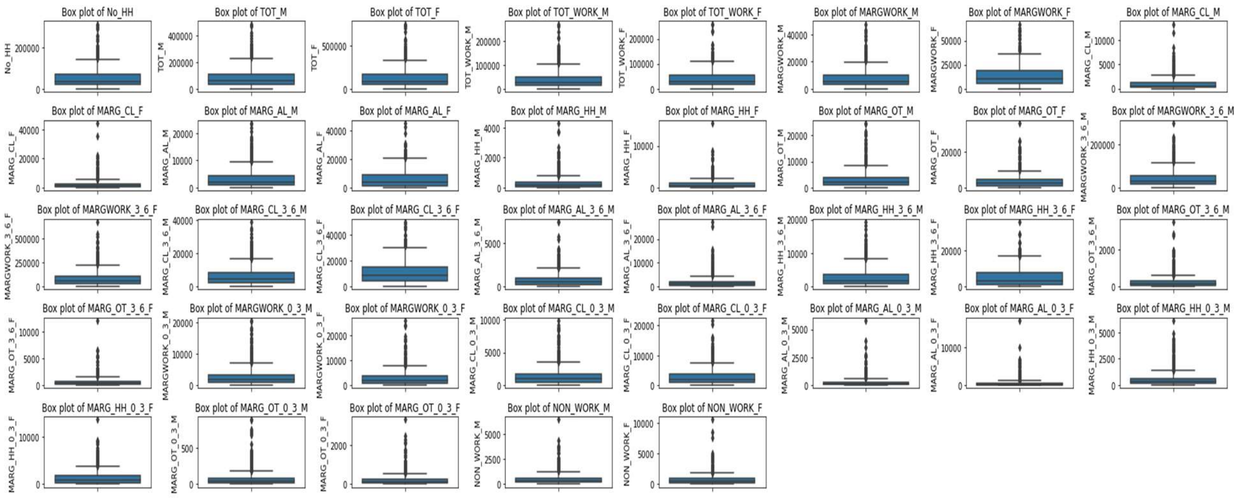


Part 2 - PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?

The outliers are present in data variables, The treating of Outliers are not necessary because it may distort the analysis because the outliers are justified and depicted as per State population density.

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.

Box plot Before data Scaling by

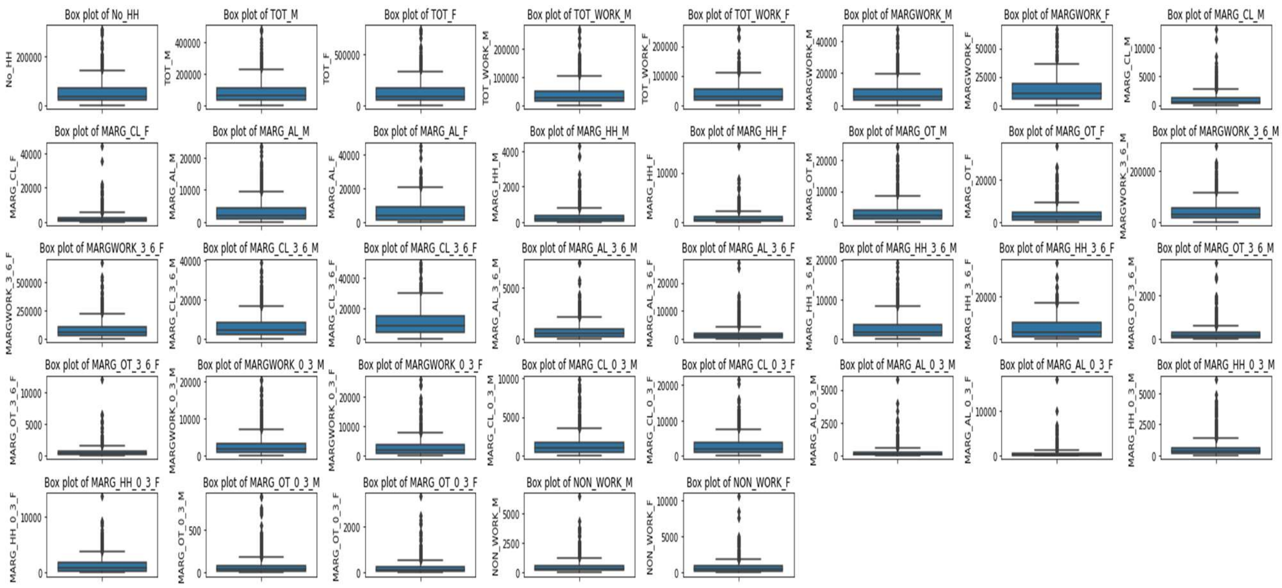


<u>Scaled Data Statistical Summary</u>

	count	mean	std	min	25%	50%	75%	max
No_HH	640.00	0.00	1.00	-1.06	-0.66	-0.32	0.37	5.39
TOT_M	640.00	-0.00	1.00	-1.08	-0.68	-0.29	0.38	5.53
TOT_F	640.00	-0.00	1.00	-1.07	-0.67	-0.31	0.37	5.53
TOT_WORK_M	640.00	-0.00	1.00	-1.04	-0.67	-0.28	0.34	6.36
TOT_WORK_F	640.00	-0.00	1.00	-1.10	-0.68	-0.29	0.32	5.83
MARGWORK_M	640.00	0.00	1.00	-1.05	-0.66	-0.29	0.27	5.37
MARGWORK_F	640.00	-0.00	1.00	-1.18	-0.70	-0.27	0.53	4.90
MARG_CL_M	640.00	-0.00	1.00	-0.79	-0.56	-0.33	0.18	9.28
MARG_CL_F	640.00	-0.00	1.00	-0.65	-0.47	-0.30	0.10	11.80
MARG_AL_M	640.00	0.00	1.00	-0.87	-0.64	-0.33	0.26	5.40
MARG_AL_F	640.00	0.00	1.00	-0.95	-0.75	-0.36	0.39	5.74
MARG_HH_M	640.00	-0.00	1.00	-0.69	-0.53	-0.33	0.09	8.61
MARG_HH_F	640.00	0.00	1.00	-0.66	-0.51	-0.30	0.15	12.24
MARG_OT_M	640.00	0.00	1.00	-0.86	-0.61	-0.30	0.24	5.99
MARG_OT_F	640.00	-0.00	1.00	-0.86	-0.60	-0.29	0.21	7.99
MARGWORK_3_6_M	640.00	0.00	1.00	-1.07	-0.66	-0.30	0.39	6.64
MARGWORK_3_6_F	640.00	-0.00	1.00	-0.97	-0.66	-0.29	0.32	7.18

MARG_CL_3_6_M	640.00	-0.00	1.00	-1.06	-0.67	-0.29	0.29	5.44
MARG_CL_3_6_F	640.00	-0.00	1.00	-1.21	-0.71	-0.24	0.56	4.70
MARG_AL_3_6_M	640.00	0.00	1.00	-0.87	-0.61	-0.34	0.22	7.33
MARG_AL_3_6_F	640.00	0.00	1.00	-0.70	-0.50	-0.31	0.12	10.19
MARG_HH_3_6_M	640.00	-0.00	1.00	-0.90	-0.66	-0.34	0.31	5.43
MARG_HH_3_6_F	640.00	-0.00	1.00	-0.97	-0.76	-0.35	0.44	5.83
MARG_OT_3_6_M	640.00	0.00	1.00	-0.68	-0.52	-0.32	0.09	9.18
MARG_OT_3_6_F	640.00	0.00	1.00	-0.65	-0.51	-0.30	0.15	12.80
MARGWORK_0_3_M	640.00	0.00	1.00	-0.86	-0.61	-0.31	0.23	5.94
MARGWORK_0_3_F	640.00	0.00	1.00	-0.85	-0.60	-0.30	0.23	6.92
MARG_CL_0_3_M	640.00	-0.00	1.00	-0.93	-0.61	-0.30	0.22	5.70
MARG_CL_0_3_F	640.00	-0.00	1.00	-0.98	-0.65	-0.30	0.30	6.77
MARG_AL_0_3_M	640.00	0.00	1.00	-0.55	-0.45	-0.30	0.04	12.19
MARG_AL_0_3_F	640.00	-0.00	1.00	-0.50	-0.40	-0.28	0.01	14.86
MARG_HH_0_3_M	640.00	0.00	1.00	-0.74	-0.56	-0.33	0.11	7.29
MARG_HH_0_3_F	640.00	-0.00	1.00	-0.82	-0.63	-0.36	0.26	7.84
MARG_OT_0_3_M	640.00	-0.00	1.00	-0.66	-0.53	-0.34	0.07	7.64
MARG_OT_0_3_F	640.00	-0.00	1.00	-0.65	-0.51	-0.28	0.13	10.19
NON_WORK_M	640.00	-0.00	1.00	-0.84	-0.57	-0.30	0.15	9.75
NON_WORK_F	640.00	-0.00	1.00	-0.77	-0.53	-0.26	0.16	10.81

Box Plot after Scaling the data



Insight: There is no any deviations in variables are visible in outliers after scaling of data

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

The Step by step analysis has been performed at Jupyter note book for reference

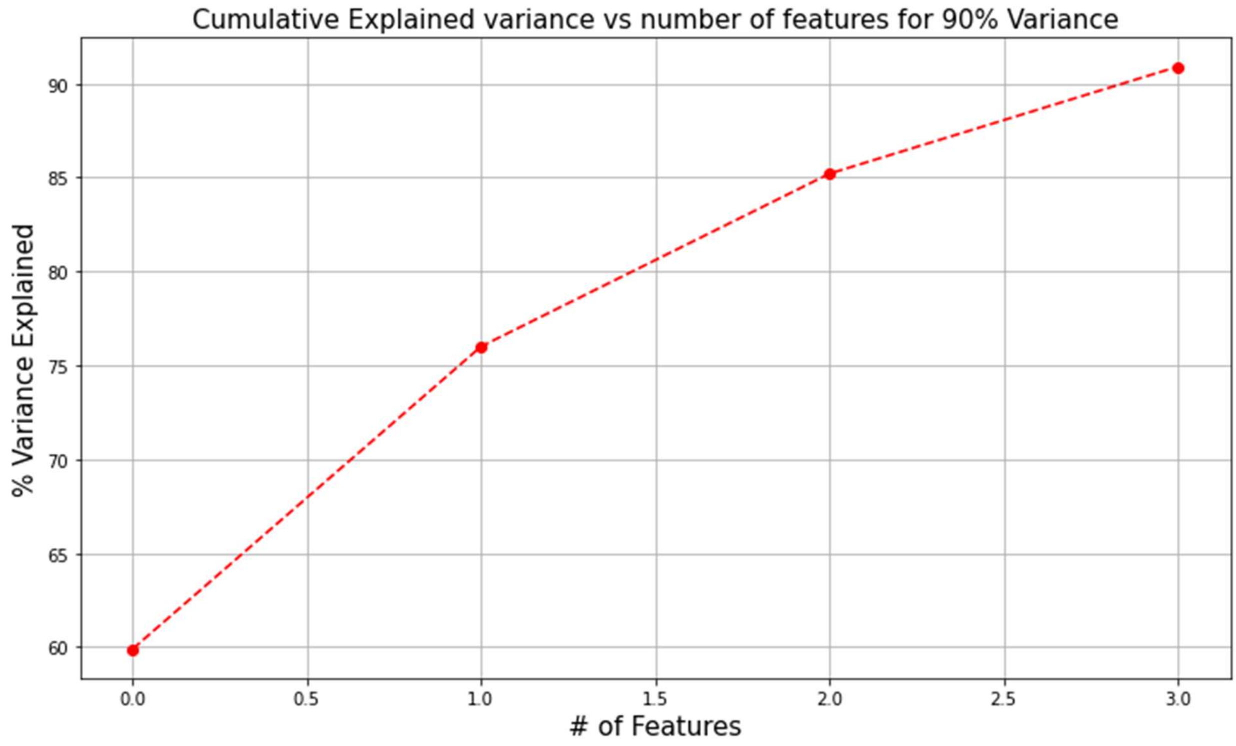
```
Eigen Values
The variance explained by each of eigen values in order is
Out[164]:
array([5.99040664e+01, 1.61124276e+01, 9.21580429e+00, 5.68865484e+00,
       2.99939897e+00, 2.08524780e+00, 1.21369233e+00, 7.34269188e-01,
       3.97714153e-01, 3.38070527e-01, 3.14473045e-01, 2.83914306e-01,
       1.71714347e-01, 1.39805985e-01, 1.08994899e-01, 1.05050348e-01,
       7.20687833e-02, 4.23398754e-02, 3.66354023e-02, 2.93003600e-02,
       6.35649074e-03, 2.80741958e-30, 1.53856214e-30, 6.55403057e-31,
       4.72595970e-31, 3.85939863e-31, 3.85939863e-31, 3.85939863e-31,
       3.85939863e-31, 3.85939863e-31, 3.85939863e-31, 3.85939863e-31,
       3.85939863e-31, 3.85939863e-31, 3.85939863e-31, 3.85939863e-31,
       1.11242758e-31])
```

```
Cumulative Variance Explained
Out[165]:
array([59.9, 76. , 85.2, 90.9, 93.9, 96. , 97.2, 97.9, 98.3, 98.6, 98.9,
       99.2, 99.4, 99.5, 99.6, 99.7, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8,
       99.8, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8,
       99.8, 99.8, 99.8, 99.8])
```

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

Reduced Data Frame top 5

	0	1	2	3
0	-3.42	0.14	-0.34	1.02
1	-3.60	-0.08	-0.48	1.68
2	-4.77	0.05	0.07	0.65
3	-5.01	-0.19	-0.20	0.76
4	-3.18	1.09	0.65	0.80

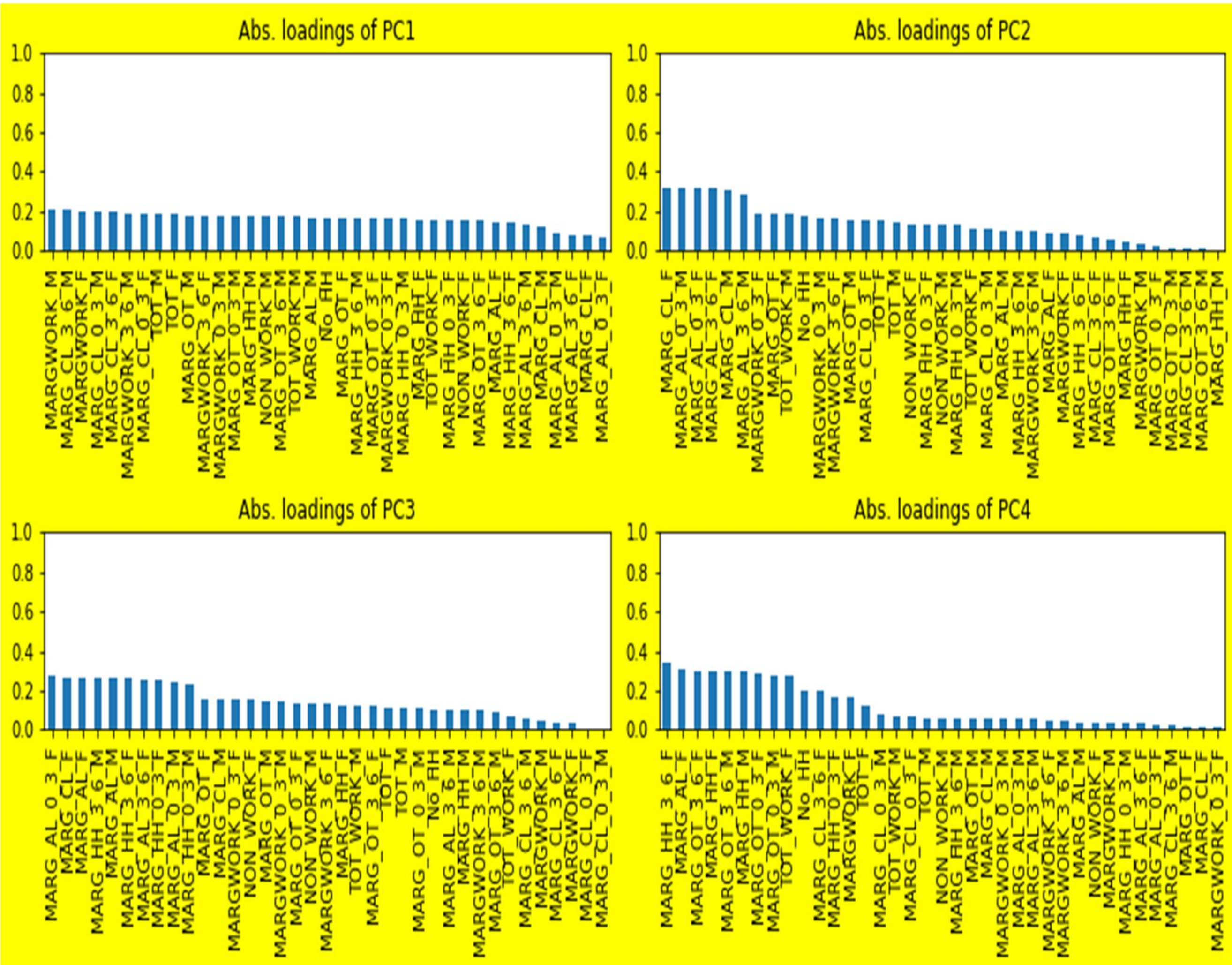


Part 2 - PCA: Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.

Cumulative sum of variance explained 90.9. 85.2, 76.85 59.9

	PC1	**PC2**	**PC3**	**PC4**
No_HH	0.17	-0.17	0.1	-0.2
TOT_M	0.19	-0.14	0.11	-0.06
TOT_F	0.18	-0.15	0.12	-0.13
TOT_WORK_M	0.17	-0.18	0.12	-0.07
TOT_WORK_F	0.16	-0.11	0.07	-0.27
MARGWORK_M	**0.21**	0.03	-0.04	0.03
MARGWORK_F	0.2	0.09	-0.03	-0.17
MARG_CL_M	0.12	**0.31**	0.15	0.05
MARG_CL_F	0.08	**0.32**	**0.27**	-0.01
MARG_AL_M	0.17	0.1	**-0.26**	-0.04
MARG_AL_F	0.15	0.09	**-0.27**	-0.31
MARG_HH_M	0.18	0	-0.1	**0.29**
MARG_HH_F	0.16	-0.04	-0.12	**0.3**
MARG_OT_M	0.18	-0.16	0.14	0.06
MARG_OT_F	0.17	-0.18	0.16	-0.02
MARGWORK_3_6_M	0.19	-0.09	0.1	-0.05
MARGWORK_3_6_F	0.18	-0.16	0.13	-0.05
MARG_CL_3_6_M	**0.21**	0.01	-0.05	0.02

MARG_CL_3_6_F	0.19	0.06	-0.04	-0.2
MARG_AL_3_6_M	0.13	**0.29**	0.1	0.05
MARG_AL_3_6_F	0.08	**0.31**	0.26	-0.03
MARG_HH_3_6_M	0.17	0.09	-0.26	-0.06
MARG_HH_3_6_F	0.14	0.08	-0.26	-0.34
MARG_OT_3_6_M	0.17	-0.01	-0.1	0.3
MARG_OT_3_6_F	0.15	-0.05	-0.12	0.3
MARGWORK_0_3_M	0.18	-0.16	0.14	0.05
MARGWORK_0_3_F	0.17	-0.19	0.15	-0.01
MARG_CL_0_3_M	0.19	0.11	0	0.08
MARG_CL_0_3_F	0.19	0.15	0	-0.07
MARG_AL_0_3_M	0.08	**0.31**	**0.24**	0.05
MARG_AL_0_3_F	0.07	**0.31**	**0.27**	0.02
MARG_HH_0_3_M	0.17	0.13	-0.24	0.03
MARG_HH_0_3_F	0.16	0.13	-0.26	-0.17
MARG_OT_0_3_M	0.18	0.02	-0.11	**0.28**
MARG_OT_0_3_F	0.17	-0.02	-0.14	**0.29**
NON_WORK_M	0.18	-0.13	0.13	0.06
NON_WORK_F	0.16	-0.13	0.15	-0.04



Component Summaries

The graphical representation are describing the Component summary and its relations.

- **First Principal Component Analysis - PC1**

  The first principal component is a measure of the MARGWORK_M and the MARG_CL_3_6_M Majorly while ,Other aspects are very close to it in range of 0.16 to 0.19. They are all positively related to PC1 because they all have positive signs.

- **Second Principal Component Analysis – PC2**

  The second principal component is a measure of MARG_CL_M, MARG_CL_F , MARG_AL_0_3_M, MARG_AL_0_3_M, MARG_AL_3_6_F, MARG_AL_3_6_M

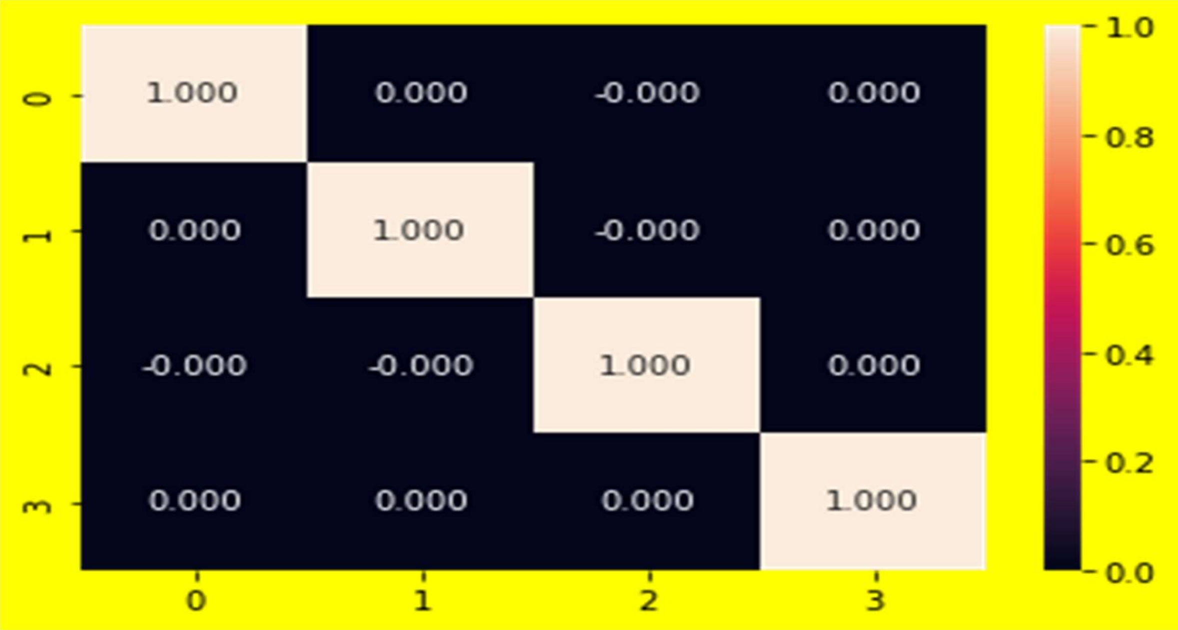
  Here we can see that PC2 distinguishes on the basis of above-mentioned variables.

- **Third Principal Component Analysis - PC3**

  The third principal component is a measure of the MARG_CL_F, MARG_AL_M , MARG_AL_F, MARG_AL_0_3_M , MARG_AL_0_3_F.

- **Fourth Principal Component Analysis - PC4**

  The fourth principal component is a measure of the MARG_HH_M , MARG_HH_F, MARG_OT_0_3_M , MARG_OT_0_3_F.



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

Linear Equation for PC1 for provided data set

```
0.168 x No_HH
0.185 x TOT_M
0.182 x TOT_F
0.173 x TOT_WORK_M
0.157 x TOT_WORK_F
0.206 x MARGWORK_M
0.196 x MARGWORK_F
0.121 x MARG_CL_M
0.077 x MARG_CL_F
0.169 x MARG_AL_M
0.147 x MARG_AL_F
0.177 x MARG_HH_M
0.158 x MARG_HH_F
0.18 x MARG_OT_M
0.168 x MARG_OT_F
0.187 x MARGWORK_3_6_M
0.179 x MARGWORK_3_6_F
0.206 x MARG_CL_3_6_M
0.193 x MARG_CL_3_6_F
0.133 x MARG_AL_3_6_M
0.079 x MARG_AL_3_6_F
0.168 x MARG_HH_3_6_M
0.14 x MARG_HH_3_6_F
0.175 x MARG_OT_3_6_M
0.153 x MARG_OT_3_6_F
0.178 x MARGWORK_0_3_M
```

```
0.166 x MARGWORK_0_3_F
0.195 x MARG_CL_0_3_M
0.186 x MARG_CL_0_3_F
0.083 x MARG_AL_0_3_M
0.069 x MARG_AL_0_3_F
0.165 x MARG_HH_0_3_M
0.157 x MARG_HH_0_3_F
0.178 x MARG_OT_0_3_M
0.167 x MARG_OT_0_3_F
0.177 x NON_WORK_M
0.155 x NON_WORK_F
```