Data Mining- Clustering clean Ads Project

by Anisha Sharma

PGPDSBA.O.Mar23.A

Great Learning

CONTENT:

Problem 1

Clustering:

Digital Ads Data:

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

The following three features are commonly used in digital marketing:

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

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

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

Problem 2:

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.

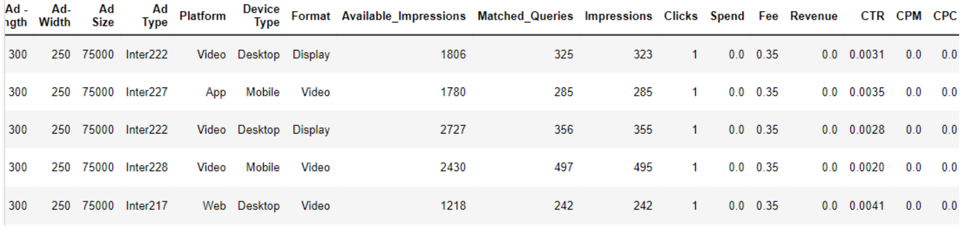
The data collected has so many variables thus making it difficult to find useful details without using Data Science Techniques. You are tasked to perform detailed EDA and identify Optimum Principal Components that explains the most variance in data. Use Sklearn only.

Problem1:

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

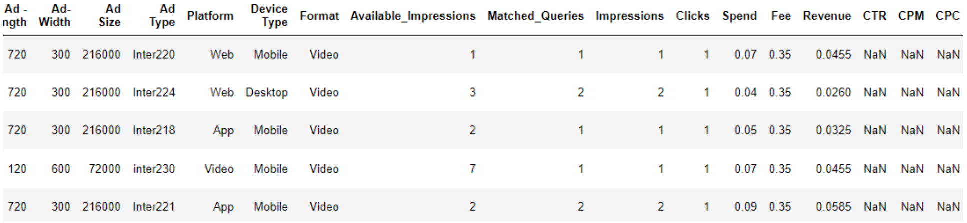
Solution:

- Below are the first five rows of dataset.



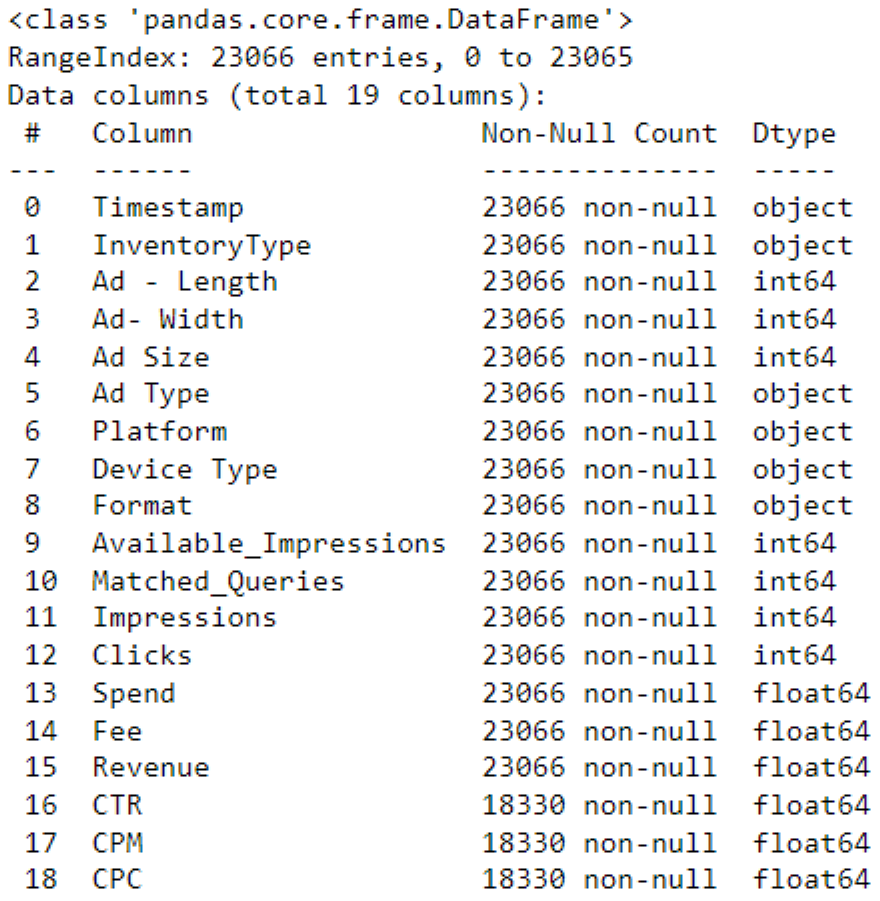
Ad - ngth	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0	0.35	0.0	0.0031	0.0	0.0
300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.0	0.35	0.0	0.0035	0.0	0.0
300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0	0.35	0.0	0.0028	0.0	0.0
300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.0	0.35	0.0	0.0020	0.0	0.0
300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.0	0.35	0.0	0.0041	0.0	0.0

- The last five rows of the dataset.



Ad - ngth	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
720	300	216000	Inter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
720	300	216000	Inter224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.0260	NaN	NaN	NaN
720	300	216000	Inter218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	NaN	NaN	NaN
120	600	72000	inter230	Video	Mobile	Video	7	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
720	300	216000	Inter221	App	Mobile	Video	2	2	2	1	0.09	0.35	0.0585	NaN	NaN	NaN

- Data summary.



```
<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
```

Insights:

- There are 23006 rows and 19 columns in the dataset.
- There is no duplicate values in the dataset.
- Dataset has these datatypes- 6 (float64), 7(int64), 6(objects)
- There are 4736 null values in each column of CTR, CPM and CPC.
- There are some columns in the dataset which is not required for the clustering that is Timestamp, Inventory type, Ad type, Platform, Device type, Format.

- Maximum number of revenue generated by ad is about 21276 million which is much higher then the amount was funding on the project and on average is 926 million.

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

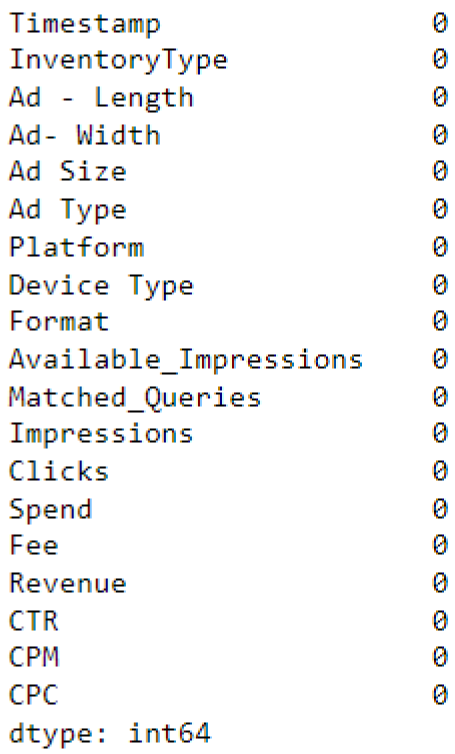
Solution: We have treated the missing values by defining functions for each column of CPC, CTR, CPM having the same argument.

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

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

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

And apply to new column using lambda.



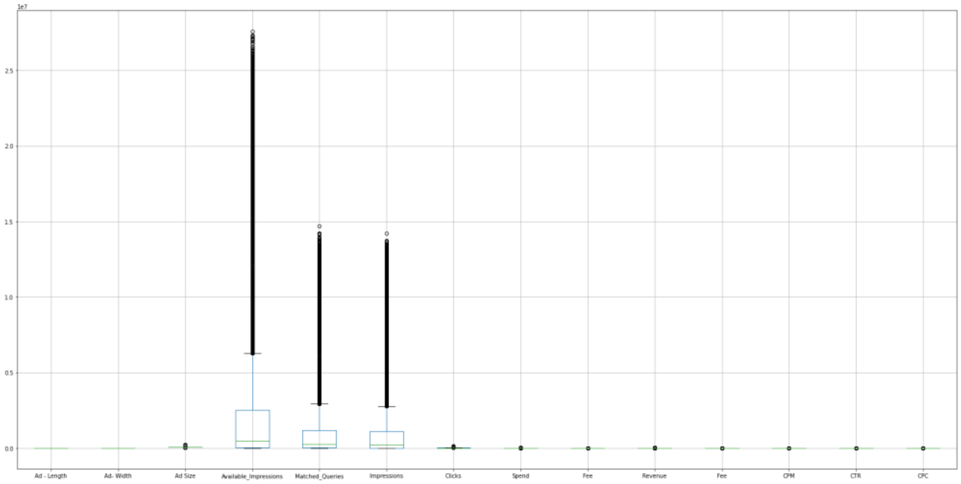
```
Timestamp                 0
InventoryType             0
Ad - Length               0
Ad- Width                 0
Ad Size                   0
Ad Type                   0
Platform                  0
Device Type               0
Format                    0
Available_Impressions     0
Matched_Queries 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

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

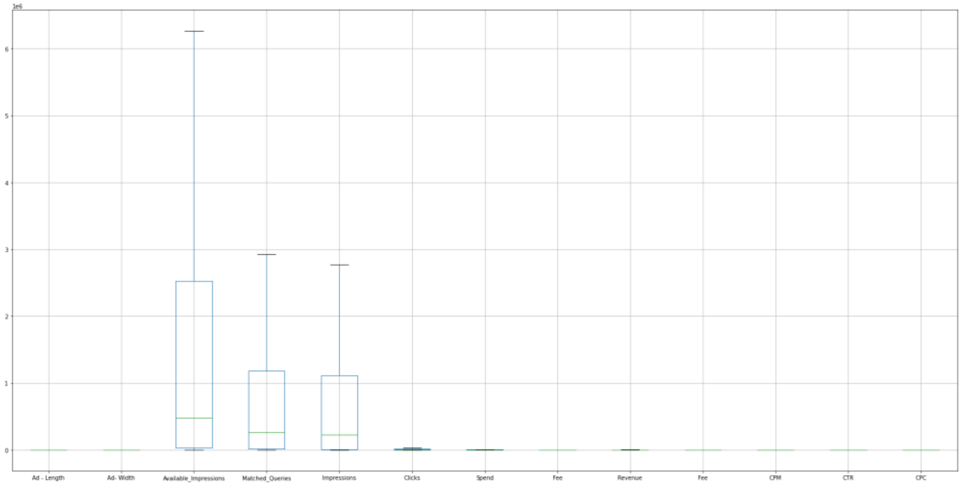
Solution:

- Yes, There are ouliers in every column of the dataset except Ad-length and Ad-width.
- Treating outliers are necessary for k-means clustering as the K-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values
- We have used IQR method to treat outliers.

This below graph is before treating outliers:



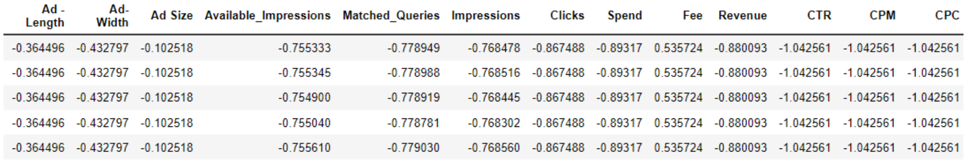
After treating outliers:



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

Solution: We have performed z-score scaling because without scaling data, the algorithm may be biased towards higher value. Scaling can increase the computational complexity of algorithm.

After scaling, the data looks like:

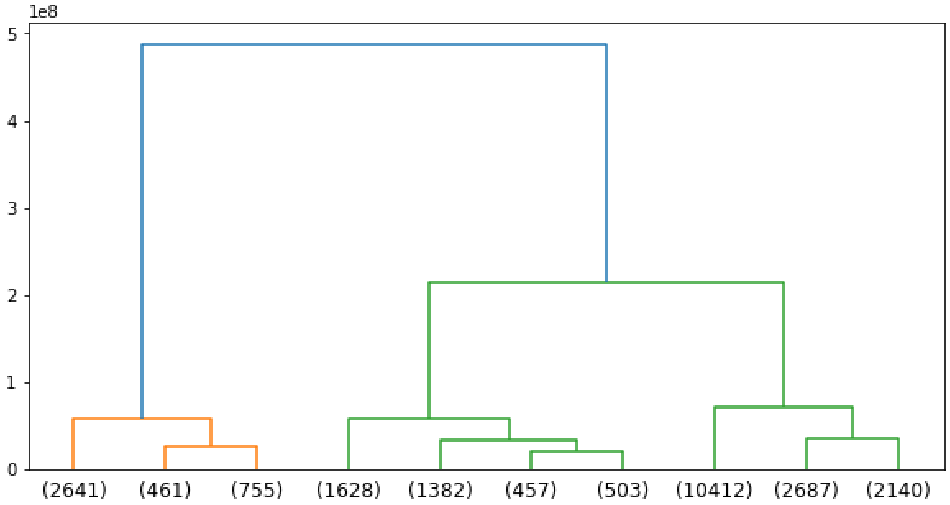


Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
-0.364496	-0.432797	-0.102518	-0.755333	-0.778949	-0.768478	-0.867488	-0.89317	0.535724	-0.880093	-1.042561	-1.042561	-1.042561
-0.364496	-0.432797	-0.102518	-0.755345	-0.778988	-0.768516	-0.867488	-0.89317	0.535724	-0.880093	-1.042561	-1.042561	-1.042561
-0.364496	-0.432797	-0.102518	-0.754900	-0.778919	-0.768445	-0.867488	-0.89317	0.535724	-0.880093	-1.042561	-1.042561	-1.042561
-0.364496	-0.432797	-0.102518	-0.755040	-0.778781	-0.768302	-0.867488	-0.89317	0.535724	-0.880093	-1.042561	-1.042561	-1.042561
-0.364496	-0.432797	-0.102518	-0.755610	-0.779030	-0.768560	-0.867488	-0.89317	0.535724	-0.880093	-1.042561	-1.042561	-1.042561

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

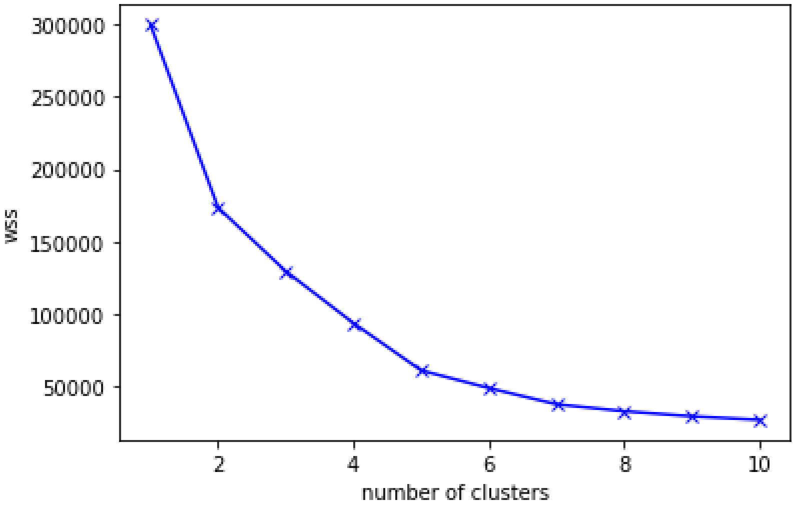
Solution: Please find below Dendrogram performed for Hierarchical using WARD and Euclidean Distance on the Scaled Data.

In this Dendrogram, value of P = 10, which means that only the last 10 merged clusters are shown



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

Solution: To make elbow plot we import kmeans from sklearn.cluster and get the value of wss. This is for 10 number of clusters. The below is the elbow graph:

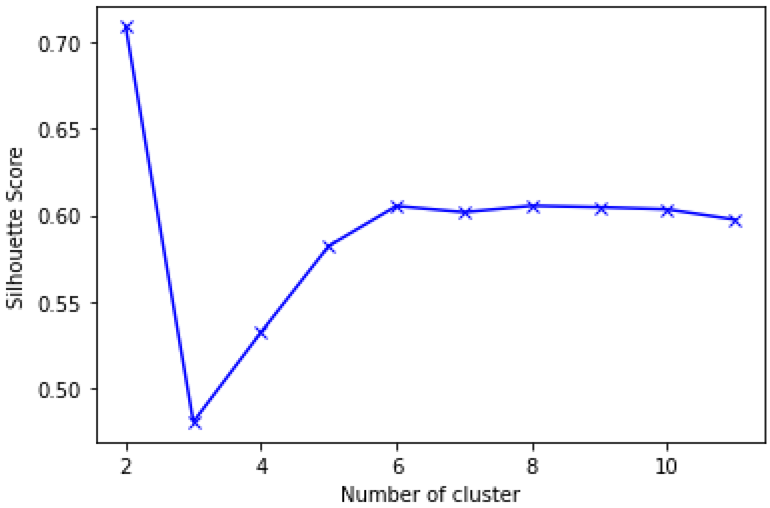


As per the check When we move from K=1 to K=2, We see that there is a significant drop in the value. Also when we move from k=2 to k=3, k=3 to k=4, k=4 to k=5 there is a significant drop as well. k=5 to k=6, the drop in values reduces significantly. Hence In this case, the WSS is not significantly dropping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

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

Solution: To print the silhoutte scores we import silhoutte_score and silhoutte_sample from sklearn.metrics.

Silhoutte score is 0.70. optimum number of clusters are 5 since we can see from the below graph.



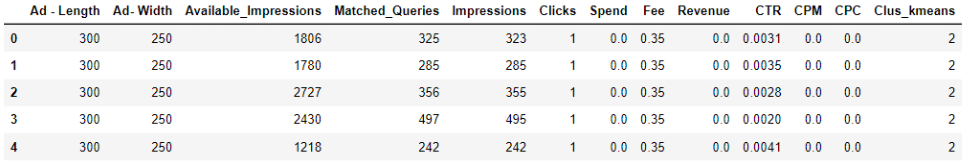
We have calculated Silhouette Score for scaled data using the silhouette_score() function.The Silhouette Score is a measure of how similar an object is to its own cluster compared to otherclusters, and it ranges from -1 to 1, with higher values indicating better clustering.As per Elbow plot/scree-plot, we concluded that the optimal number of clusters should be 5.Because 2 would be very less number of clusters.

1.8 Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots].

Solution:

- We have performed KMeans Clustering on scaled data, and then added the predicted cluster labels to two different data sets: data_df_scaled and data_df.
- The KMeans function from scikit-learn is used to create a KMeans object with n_clusters=5(i.e., 5 clusters).

- We have Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC
- Created clusters for the Ads based on optimum number of clusters using silhouette score.



	Ad - Length	Ad- Width	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	Clus_kmeans
0	300	250	1806	325	323	1	0.0	0.35	0.0	0.0031	0.0	0.0	2
1	300	250	1780	285	285	1	0.0	0.35	0.0	0.0035	0.0	0.0	2
2	300	250	2727	356	355	1	0.0	0.35	0.0	0.0028	0.0	0.0	2
3	300	250	2430	497	495	1	0.0	0.35	0.0	0.0020	0.0	0.0	2
4	300	250	1218	242	242	1	0.0	0.35	0.0	0.0041	0.0	0.0	2

1.9 Conclude the project by providing summary of your learnings.

Solution:

- There are 23066 rows, and 19 columns into the Dataset.
- There are no duplicate values in dataset.
- There are 4636 Null values in CTR, CPM, and CPC Columns.
- We have treated missing values in CPC, CTR, and CPM columns using the given formula
- It seems that there are Outliers into the Dataset
- We treated outliers using IQR method
- We have applied z-score method on the dataset for scaling.
- We have plotted Dendrogram for value of P = 10
- Plotted elbow plot and got optimum value is 5
- As per Elbow plot/scree-plot, we concluded that the optimal number of clusters should be 5.
- We have created 5 clusters for the Dataset.

Conclusion after Clustering :

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

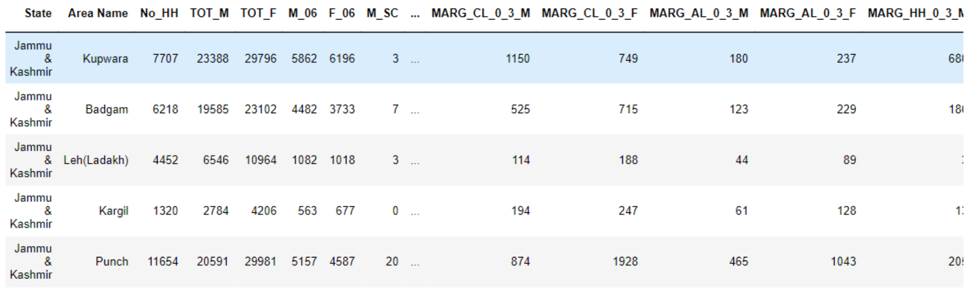
Problem1:

PCA:

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

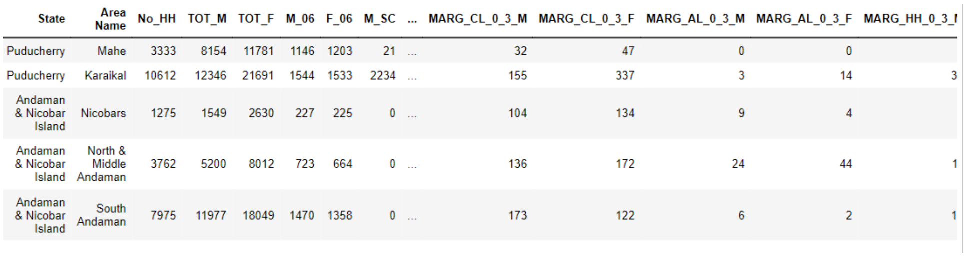
Solution:

- Below are the first five rows of dataset.



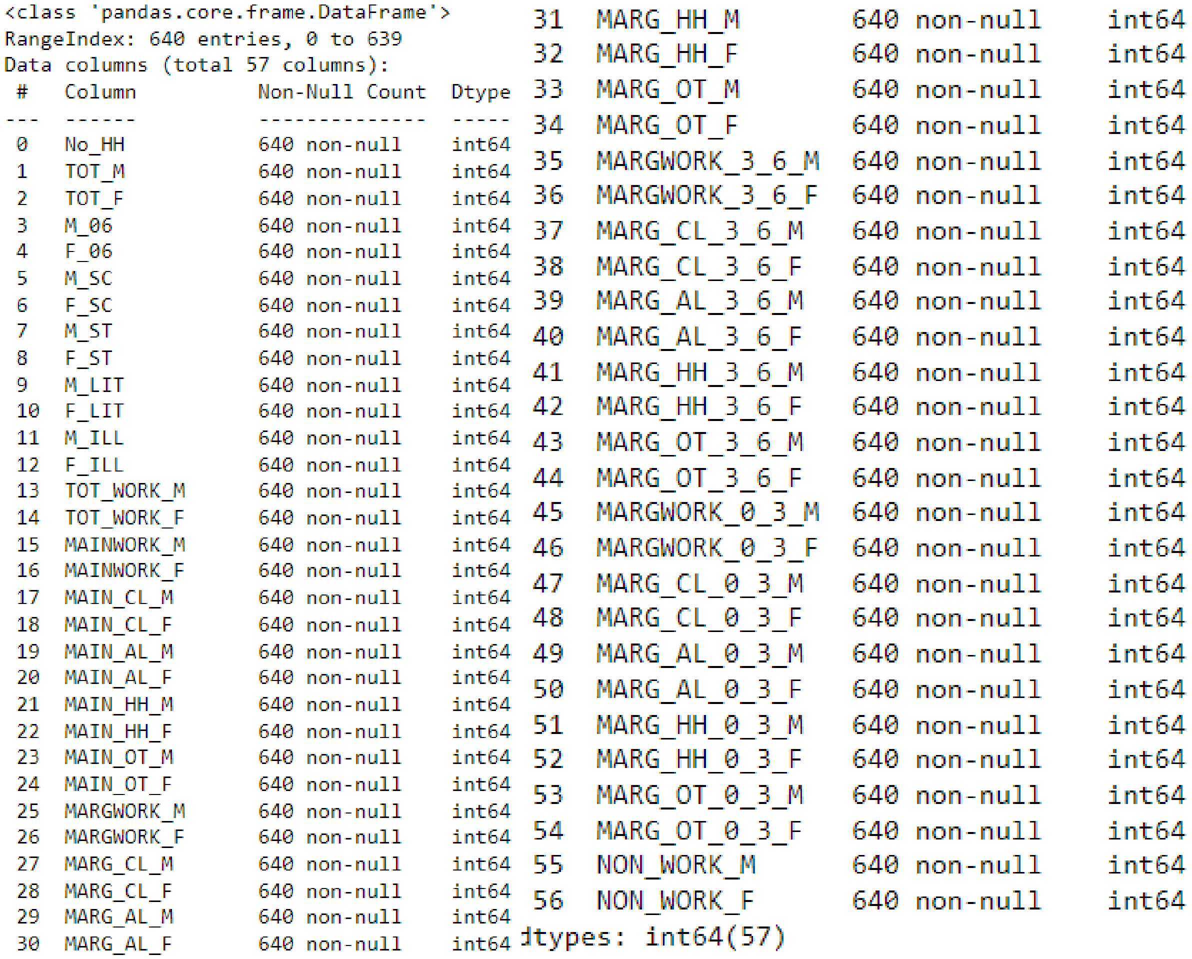
State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M
Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180	237	68
Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123	229	18
Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	...	114	188	44	89	
Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	...	194	247	61	128	1
Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	...	874	1928	465	1043	20

- The last five rows of the dataset.



State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M
Puducherry	Mahe	3333	8154	11781	1146	1203	21	...	32	47	0	0	
Puducherry	Karaikal	10612	12346	21691	1544	1533	2234	...	155	337	3	14	3
Andaman & Nicobar Island	Nicobars	1275	1549	2630	227	225	0	...	104	134	9	4	
Andaman & Nicobar Island	North & Middle Andaman	3762	5200	8012	723	664	0	...	136	172	24	44	1
Andaman & Nicobar Island	South Andaman	7975	11977	18049	1470	1358	0	...	173	122	6	2	1

- Data summary.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 640 entries, 0 to 639
Data columns (total 57 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   No_HH           640 non-null    int64
 1   TOT_M           640 non-null    int64
 2   TOT_F           640 non-null    int64
 3   M_06            640 non-null    int64
 4   F_06            640 non-null    int64
 5   M_SC            640 non-null    int64
 6   F_SC            640 non-null    int64
 7   M_ST            640 non-null    int64
 8   F_ST            640 non-null    int64
 9   M_LIT           640 non-null    int64
 10  F_LIT           640 non-null    int64
 11  M_ILL           640 non-null    int64
 12  F_ILL           640 non-null    int64
 13  TOT_WORK_M      640 non-null    int64
 14  TOT_WORK_F      640 non-null    int64
 15  MAINWORK_M      640 non-null    int64
 16  MAINWORK_F      640 non-null    int64
 17  MAIN_CL_M       640 non-null    int64
 18  MAIN_CL_F       640 non-null    int64
 19  MAIN_AL_M       640 non-null    int64
 20  MAIN_AL_F       640 non-null    int64
 21  MAIN_HH_M       640 non-null    int64
 22  MAIN_HH_F       640 non-null    int64
 23  MAIN_OT_M       640 non-null    int64
 24  MAIN_OT_F       640 non-null    int64
 25  MARGWORK_M      640 non-null    int64
 26  MARGWORK_F      640 non-null    int64
 27  MARG_CL_M       640 non-null    int64
 28  MARG_CL_F       640 non-null    int64
 29  MARG_AL_M       640 non-null    int64
 30  MARG_AL_F       640 non-null    int64
 31  MARG_HH_M       640 non-null    int64
 32  MARG_HH_F       640 non-null    int64
 33  MARG_OT_M       640 non-null    int64
 34  MARG_OT_F       640 non-null    int64
 35  MARGWORK_3_6_M  640 non-null    int64
 36  MARGWORK_3_6_F  640 non-null    int64
 37  MARG_CL_3_6_M   640 non-null    int64
 38  MARG_CL_3_6_F   640 non-null    int64
 39  MARG_AL_3_6_M   640 non-null    int64
 40  MARG_AL_3_6_F   640 non-null    int64
 41  MARG_HH_3_6_M   640 non-null    int64
 42  MARG_HH_3_6_F   640 non-null    int64
 43  MARG_OT_3_6_M   640 non-null    int64
 44  MARG_OT_3_6_F   640 non-null    int64
 45  MARGWORK_0_3_M  640 non-null    int64
 46  MARGWORK_0_3_F  640 non-null    int64
 47  MARG_CL_0_3_M   640 non-null    int64
 48  MARG_CL_0_3_F   640 non-null    int64
 49  MARG_AL_0_3_M   640 non-null    int64
 50  MARG_AL_0_3_F   640 non-null    int64
 51  MARG_HH_0_3_M   640 non-null    int64
 52  MARG_HH_0_3_F   640 non-null    int64
 53  MARG_OT_0_3_M   640 non-null    int64
 54  MARG_OT_0_3_F   640 non-null    int64
 55  NON_WORK_M      640 non-null    int64
 56  NON_WORK_F      640 non-null    int64
dtypes: int64(57)
```

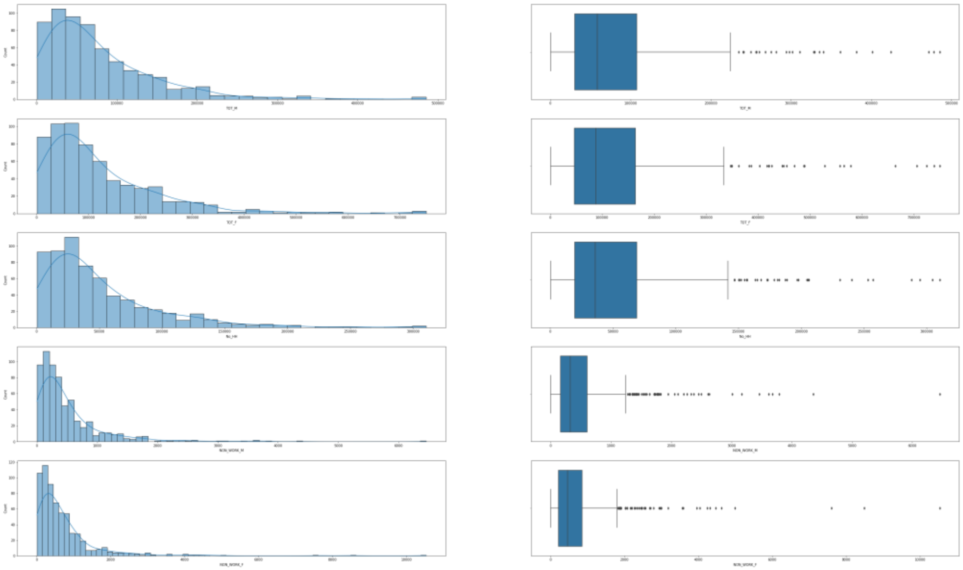
Insights:

- There are 640 number of rows and 61 number of columns in the data.
- There are no duplicate values in the dataset.
- There are no null values in the dataset.

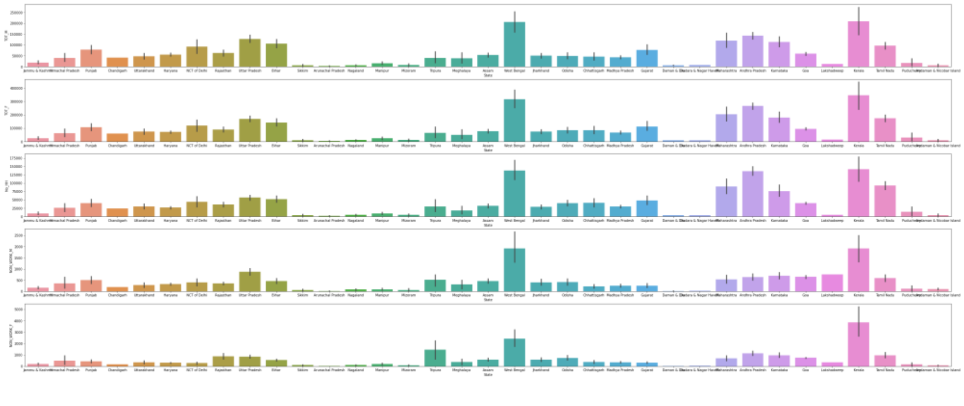
2.1 Perform detailed Exploratory analysis by creating certain questions like (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio? (Example Questions). Pick 5 variables out of the given 24 variables below for EDA: No_HH, TOT_M, TOT_F, 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.

Solution: The five variables we are taking for analyzing are No_HH, TOT_M, TOT_F, NON_WORK_M, NON_WORK_F.

This is the univariate analyses of these variables:-



This is the bivariate analyses of these variable:-

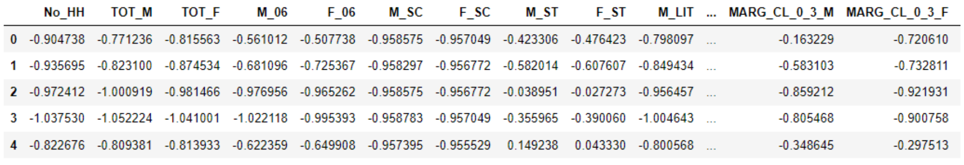


2.3 We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?

Solution: Treating outliers is not necessary unless they are resulting from a processing mistakes or false measurements. We can kept the outliers in the data.

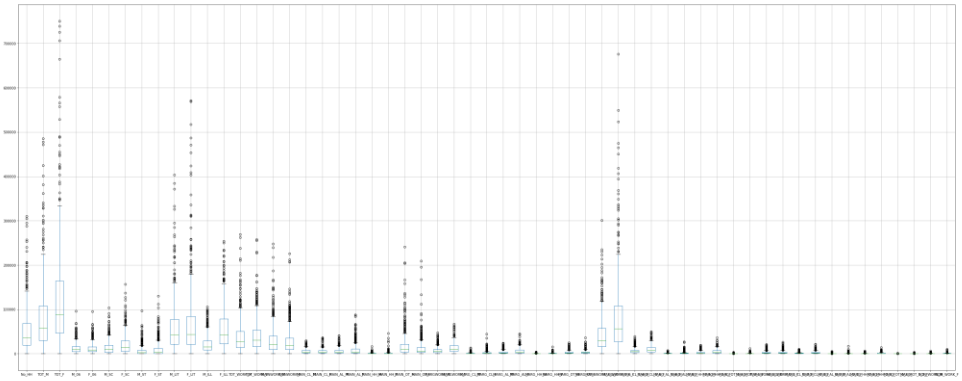
2.4 Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.

Solution: This is the table of after scaling data.

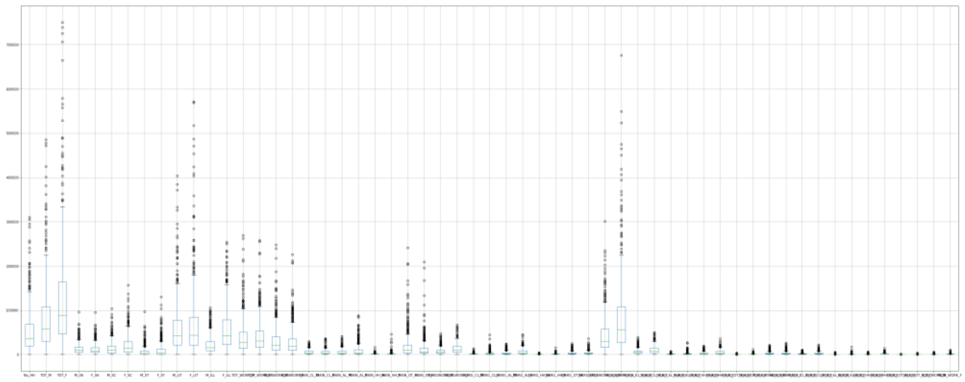


	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_0_3_M	MARG_CL_0_3_F
0	-0.904738	-0.771236	-0.815563	-0.561012	-0.507738	-0.958575	-0.957049	-0.423306	-0.476423	-0.798097	...	-0.163229	-0.720610
1	-0.935695	-0.823100	-0.874534	-0.681096	-0.725367	-0.958297	-0.956772	-0.582014	-0.607607	-0.849434	...	-0.583103	-0.732811
2	-0.972412	-1.000919	-0.981466	-0.976956	-0.965262	-0.958575	-0.956772	-0.038951	-0.027273	-0.956457	...	-0.859212	-0.921931
3	-1.037530	-1.052224	-1.041001	-1.022118	-0.995393	-0.958783	-0.957049	-0.355965	-0.390060	-1.004643	...	-0.805468	-0.900758
4	-0.822676	-0.809381	-0.813933	-0.622359	-0.649908	-0.957395	-0.955529	0.149238	0.043330	-0.800568	...	-0.348645	-0.297513

This is the graph of before scaling the data.



This is the graph of after scaling data.

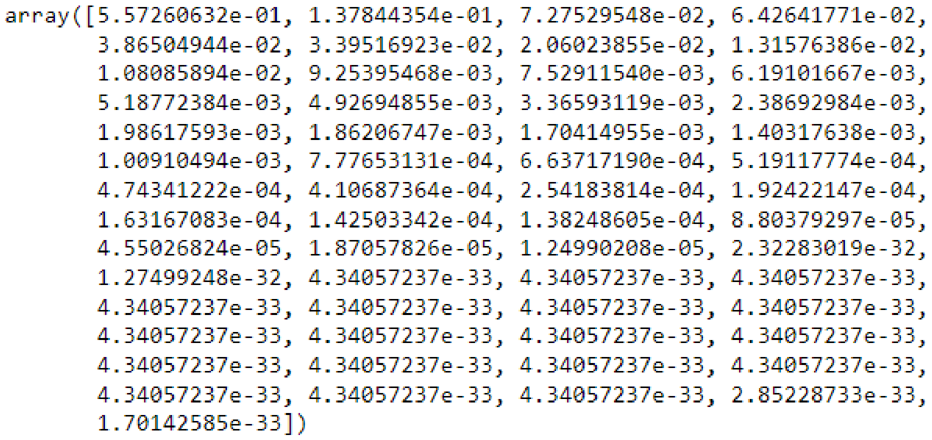


As we can there is no impact on the outliers before and after scaling the dataset.

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

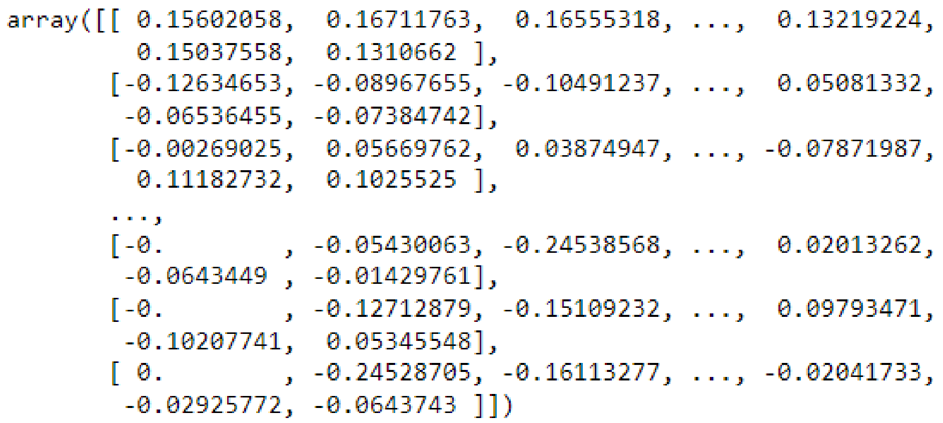
Solution:

Covariance matrix:



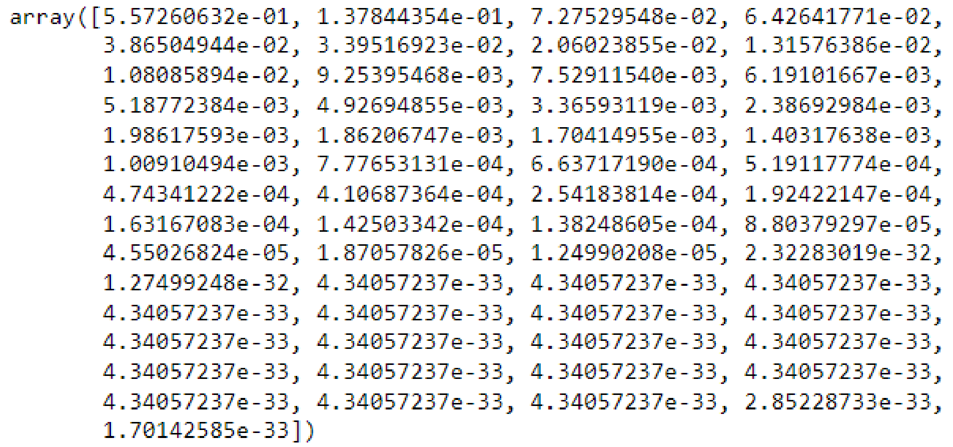
```
array([5.57260632e-01, 1.37844354e-01, 7.27529548e-02, 6.42641771e-02,
       3.86504944e-02, 3.39516923e-02, 2.06023855e-02, 1.31576386e-02,
       1.08085894e-02, 9.25395468e-03, 7.52911540e-03, 6.19101667e-03,
       5.18772384e-03, 4.92694855e-03, 3.36593119e-03, 2.38692984e-03,
       1.98617593e-03, 1.86206747e-03, 1.70414955e-03, 1.40317638e-03,
       1.00910494e-03, 7.77653131e-04, 6.63717190e-04, 5.19117774e-04,
       4.74341222e-04, 4.10687364e-04, 2.54183814e-04, 1.92422147e-04,
       1.63167083e-04, 1.42503342e-04, 1.38248605e-04, 8.80379297e-05,
       4.55026824e-05, 1.87057826e-05, 1.24990208e-05, 2.32283019e-32,
       1.27499248e-32, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 2.85228733e-33,
       1.70142585e-33])
```

Eigen vector:



```
array([[ 0.15602058,  0.16711763,  0.16555318, ...,  0.13219224,
         0.15037558,  0.1310662 ],
       [-0.12634653, -0.08967655, -0.10491237, ...,  0.05081332,
        -0.06536455, -0.07384742],
       [-0.00269025,  0.05669762,  0.03874947, ..., -0.07871987,
         0.11182732,  0.1025525 ],
       ...,
       [-0.        , -0.05430063, -0.24538568, ...,  0.02013262,
        -0.0643449 , -0.01429761],
       [-0.        , -0.12712879, -0.15109232, ...,  0.09793471,
        -0.10207741,  0.05345548],
       [ 0.        , -0.24528705, -0.16113277, ..., -0.02041733,
        -0.02925772, -0.0643743 ]])
```

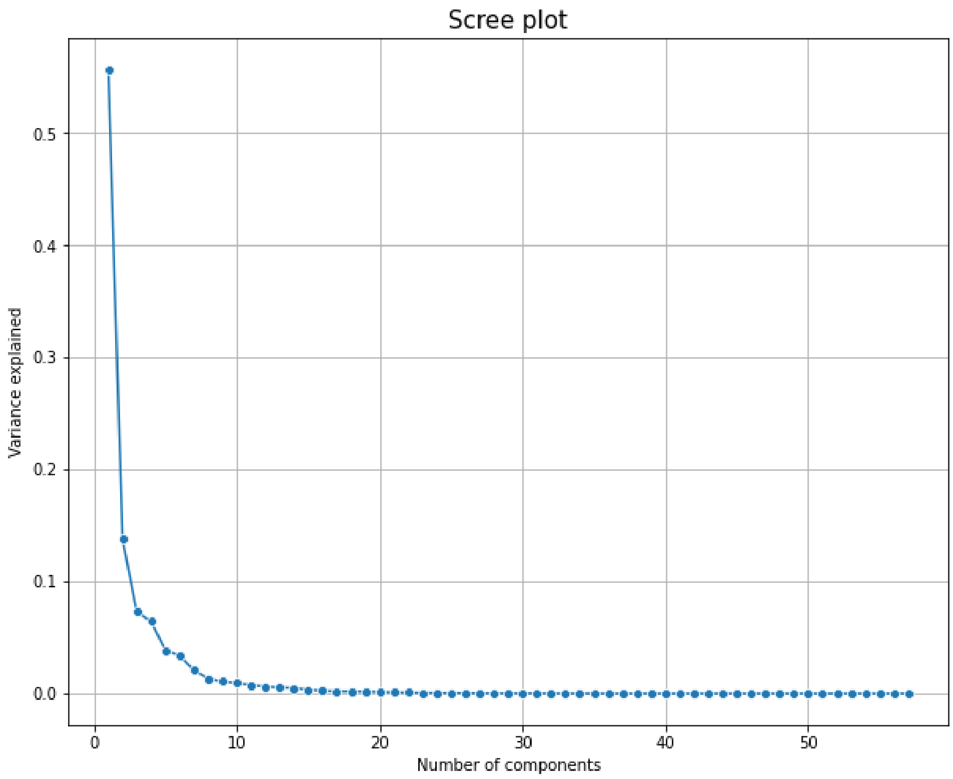
Eigen value:

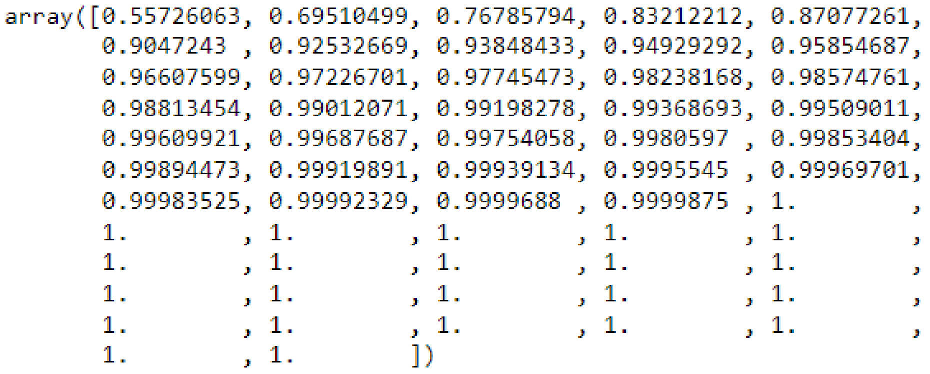


```
array([5.57260632e-01, 1.37844354e-01, 7.27529548e-02, 6.42641771e-02,
       3.86504944e-02, 3.39516923e-02, 2.06023855e-02, 1.31576386e-02,
       1.08085894e-02, 9.25395468e-03, 7.52911540e-03, 6.19101667e-03,
       5.18772384e-03, 4.92694855e-03, 3.36593119e-03, 2.38692984e-03,
       1.98617593e-03, 1.86206747e-03, 1.70414955e-03, 1.40317638e-03,
       1.00910494e-03, 7.77653131e-04, 6.63717190e-04, 5.19117774e-04,
       4.74341222e-04, 4.10687364e-04, 2.54183814e-04, 1.92422147e-04,
       1.63167083e-04, 1.42503342e-04, 1.38248605e-04, 8.80379297e-05,
       4.55026824e-05, 1.87057826e-05, 1.24990208e-05, 2.32283019e-32,
       1.27499248e-32, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 2.85228733e-33,
       1.70142585e-33])
```

2. 5 Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.

Solution:





```
array([0.55726063, 0.69510499, 0.76785794, 0.83212212, 0.87077261,
       0.9047243 , 0.92532669, 0.93848433, 0.94929292, 0.95854687,
       0.96607599, 0.97226701, 0.97745473, 0.98238168, 0.98574761,
       0.98813454, 0.99012071, 0.99198278, 0.99368693, 0.99509011,
       0.99609921, 0.99687687, 0.99754058, 0.9980597 , 0.99853404,
       0.99894473, 0.99919891, 0.99939134, 0.9995545 , 0.99969701,
       0.99983525, 0.99992329, 0.9999688 , 0.9999875 , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        ])
```

For this project, we need to consider at least 90% explained variance, so cut off for selecting the number of PCs is:'6'.

2.6 Write linear equation for first PC.

Solution: PC1=a1x1+a2x2+…….anxn