Data Mining (DM) Project Report

By

Name: Abhishek Pradhan

BATCH: PGPDSBA.O.FEB23.B

Contents

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.

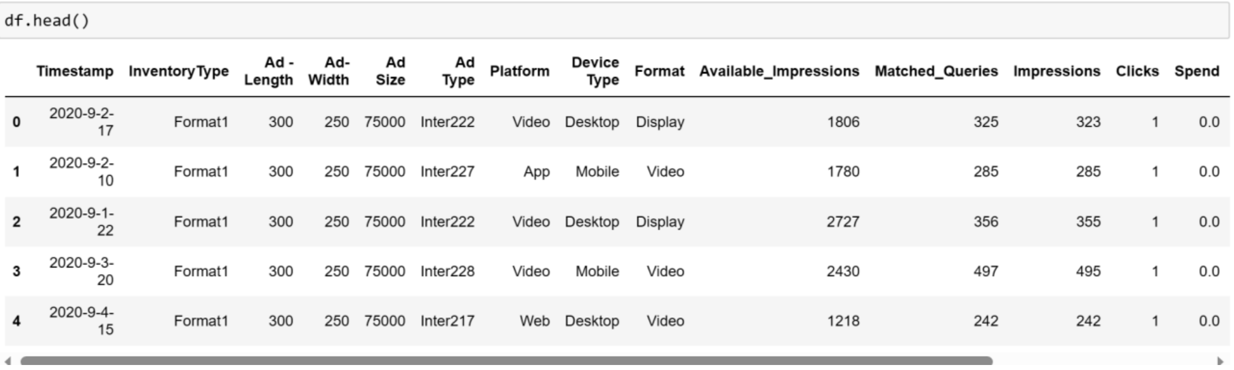
The Data Dictionary and the detailed description of the formulas for CPM, CPC and CTR are given in the sheet 2 of the Clustering Clean ads_data ExcelFile.

Perform the following in given order:

1.1. Read the data and perform basic analysis such as printing a few rows (head andtail), info, data summary, null values duplicate values, etc. (4 marks)

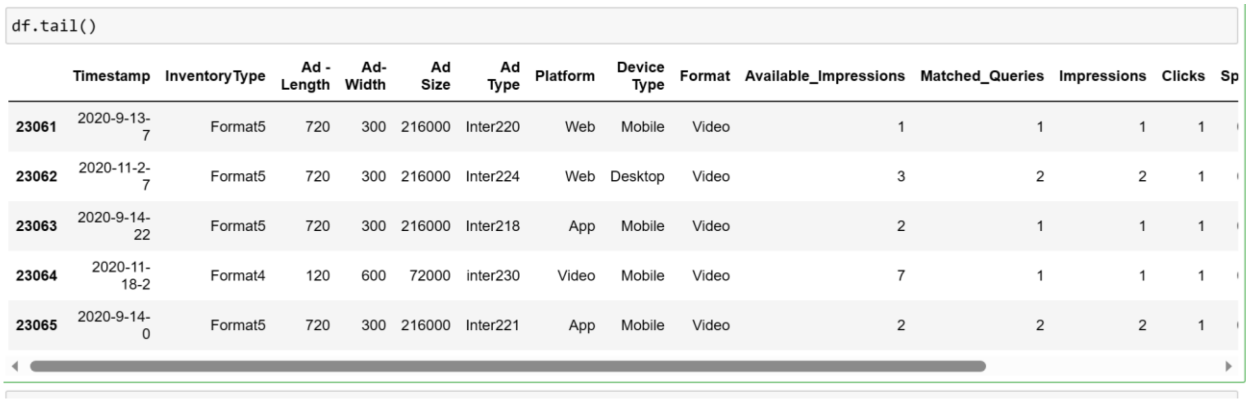
Solution:

ads 24X7 data has (23066, 19)rows and columns respectively.



```
df.head()
```

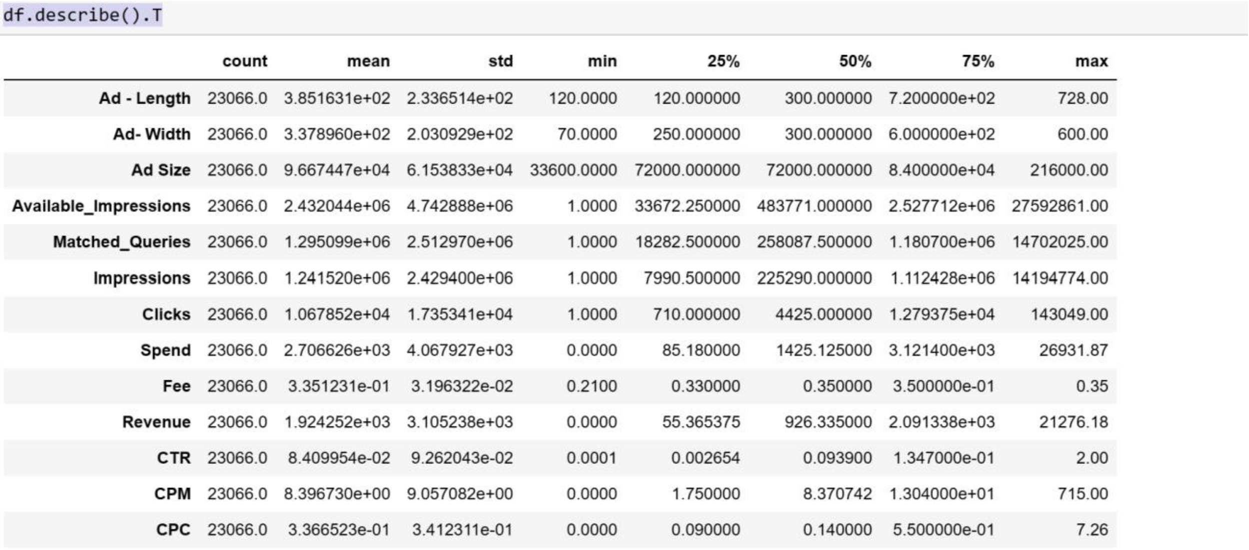
	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.0
2	2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0
3	2020-9-3-20	Format1	300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.0
4	2020-9-4-15	Format1	300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.0



```
df.tail()
```

	Timestamp	InventoryType	Ad - Length	Ad-Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Sp
23061	2020-9-13-7	Format5	720	300	216000	Inter220	Web	Mobile	Video	1	1	1	1	
23062	2020-11-2-7	Format5	720	300	216000	Inter224	Web	Desktop	Video	3	2	2	1	
23063	2020-9-14-22	Format5	720	300	216000	Inter218	App	Mobile	Video	2	1	1	1	
23064	2020-11-18-2	Format4	120	600	72000	inter230	Video	Mobile	Video	7	1	1	1	
23065	2020-9-14-0	Format5	720	300	216000	Inter221	App	Mobile	Video	2	2	2	1	

The data has 19 attributes, 6 of object type and 13 floats.



```
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	3.851631e+02	2.336514e+02	120.0000	120.000000	300.000000	7.200000e+02	728.00
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.0000	250.000000	300.000000	6.000000e+02	600.00
Ad Size	23066.0	9.667447e+04	6.153833e+04	33600.0000	72000.000000	72000.000000	8.400000e+04	216000.00
Available_Impressions	23066.0	2.432044e+06	4.742888e+06	1.0000	33672.250000	483771.000000	2.527712e+06	27592861.00
Matched_Queries	23066.0	1.295099e+06	2.512970e+06	1.0000	18282.500000	258087.500000	1.180700e+06	14702025.00
Impressions	23066.0	1.241520e+06	2.429400e+06	1.0000	7990.500000	225290.000000	1.112428e+06	14194774.00
Clicks	23066.0	1.067852e+04	1.735341e+04	1.0000	710.000000	4425.000000	1.279375e+04	143049.00
Spend	23066.0	2.706626e+03	4.067927e+03	0.0000	85.180000	1425.125000	3.121400e+03	26931.87
Fee	23066.0	3.351231e-01	3.196322e-02	0.2100	0.330000	0.350000	3.500000e-01	0.35
Revenue	23066.0	1.924252e+03	3.105238e+03	0.0000	55.365375	926.335000	2.091338e+03	21276.18
CTR	23066.0	8.409954e-02	9.262043e-02	0.0001	0.002654	0.093900	1.347000e-01	2.00
CPM	23066.0	8.396730e+00	9.057082e+00	0.0000	1.750000	8.370742	1.304000e+01	715.00
CPC	23066.0	3.366523e-01	3.412311e-01	0.0000	0.090000	0.140000	5.500000e-01	7.26

CTR, CPM and CPC have 4736 null-values, remaining variables do not have any null-values.
There are no duplicate values in the dataset.

The missing values were treated using the 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

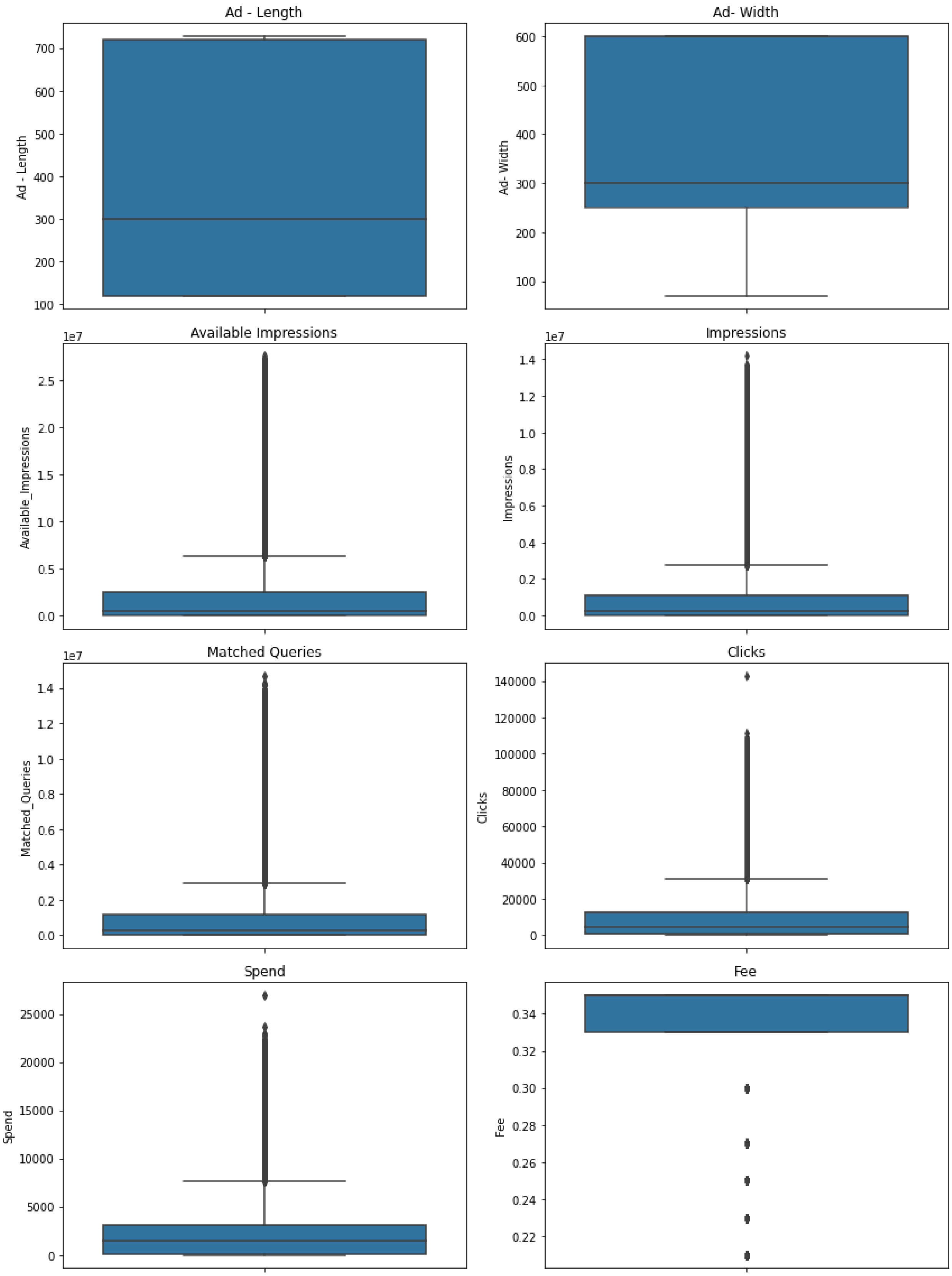
'Clicks' Column in the dataset.

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 other method that could have been used was mean method of imputation.

1.2. 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 andif yes, which method to employ. (As an analyst your judgement may be different from another analyst). (3 marks)

Solution:

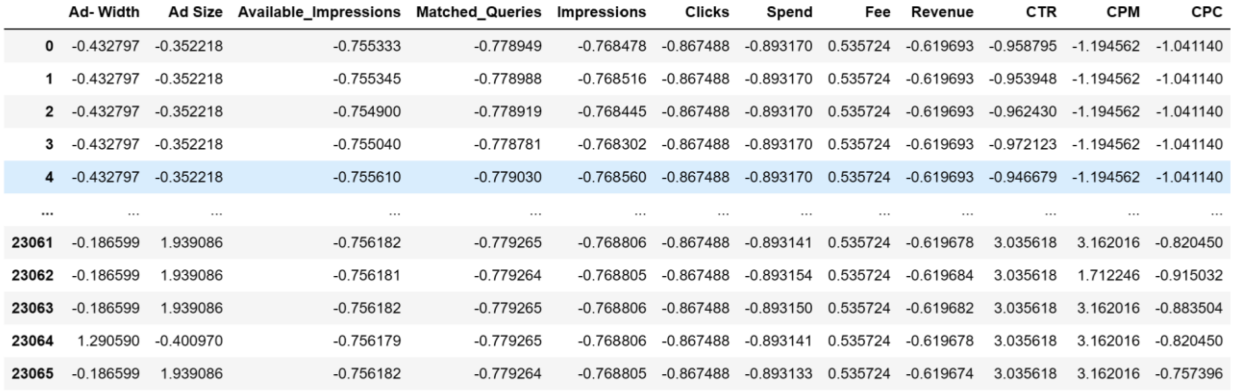

Ad - Length
Ad - Length
100
200
300
400
500
600
700
Ad- Width
Ad- Width
100
200
300
400
500
600
Available Impressions
1e7
Available_Impressions
0.0
0.5
1.0
1.5
2.0
2.5
Impressions
1e7
Impressions
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Matched Queries
1e7
Matched_Queries
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Clicks
Clicks
0
20000
40000
60000
80000
100000
120000
140000
Spend
Spend
0
5000
10000
15000
20000
25000
Fee
Fee
0.22
0.24
0.26
0.28
0.30
0.32
0.34

K-means clustering is sensitive to outliers so outlier treatment is a must and hence done using lower and upper nod method using lower_range= Q1-(1.5 * IQR) andupper_range= Q3+(1.5 * IQR) as these.

1.3. Perform z-score scaling and discuss how it affects the speed of the algorithm. (3marks)

Solution:

Scaling (i.e. z=x-u/s) calculation is required as some variables are in hundred andthousands ranges and others are in unit digits. Below is the scaled data:



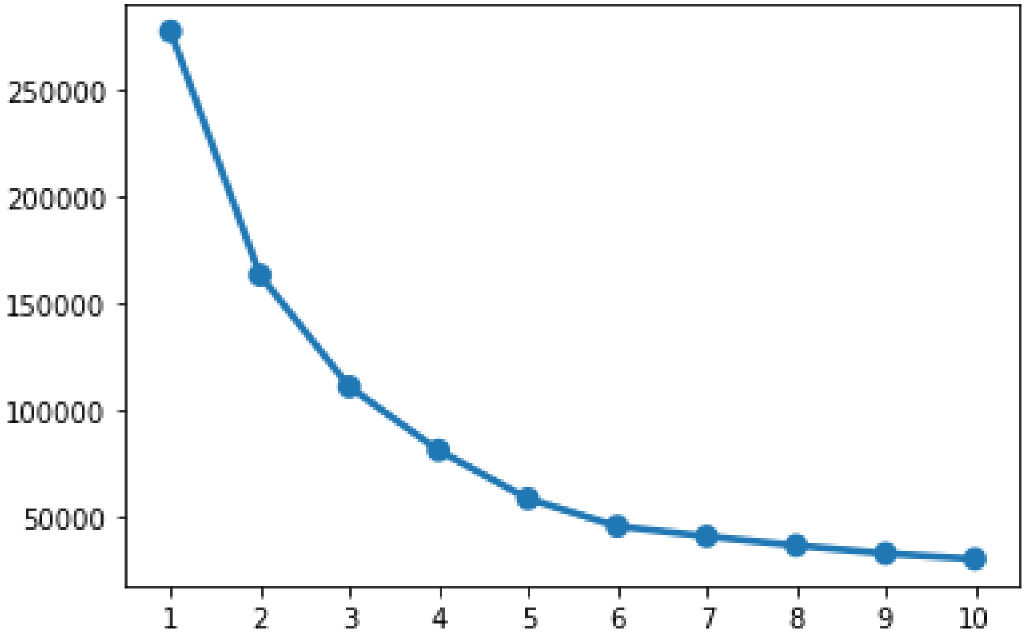
	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	-0.432797	-0.352218	-0.755333	-0.778949	-0.768478	-0.867488	-0.893170	0.535724	-0.619693	-0.958795	-1.194562	-1.041140
1	-0.432797	-0.352218	-0.755345	-0.778988	-0.768516	-0.867488	-0.893170	0.535724	-0.619693	-0.953948	-1.194562	-1.041140
2	-0.432797	-0.352218	-0.754900	-0.778919	-0.768445	-0.867488	-0.893170	0.535724	-0.619693	-0.962430	-1.194562	-1.041140
3	-0.432797	-0.352218	-0.755040	-0.778781	-0.768302	-0.867488	-0.893170	0.535724	-0.619693	-0.972123	-1.194562	-1.041140
4	-0.432797	-0.352218	-0.755610	-0.779030	-0.768560	-0.867488	-0.893170	0.535724	-0.619693	-0.946679	-1.194562	-1.041140
...	...	...	...	...	...	...	...	...	...	...	...	...
23061	-0.186599	1.939086	-0.756182	-0.779265	-0.768806	-0.867488	-0.893141	0.535724	-0.619678	3.035618	3.162016	-0.820450
23062	-0.186599	1.939086	-0.756181	-0.779264	-0.768805	-0.867488	-0.893154	0.535724	-0.619684	3.035618	1.712246	-0.915032
23063	-0.186599	1.939086	-0.756182	-0.779265	-0.768806	-0.867488	-0.893150	0.535724	-0.619682	3.035618	3.162016	-0.883504
23064	1.290590	-0.400970	-0.756179	-0.779265	-0.768806	-0.867488	-0.893141	0.535724	-0.619678	3.035618	3.162016	-0.820450
23065	-0.186599	1.939086	-0.756182	-0.779264	-0.768805	-0.867488	-0.893133	0.535724	-0.619674	3.035618	3.162016	-0.757396

Scaling has a positive and synchronizing impact on analysis enhancing speed byreducing errors.

1.4. Perform Hierarchical by constructing a Dendrogram using WARD and Euclideandistance. (4 marks)

Solution:

1.5. Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm. (4 marks)



1.1. Print silhouette scores for up to 10 clusters and identify optimum number ofclusters. (4 marks)

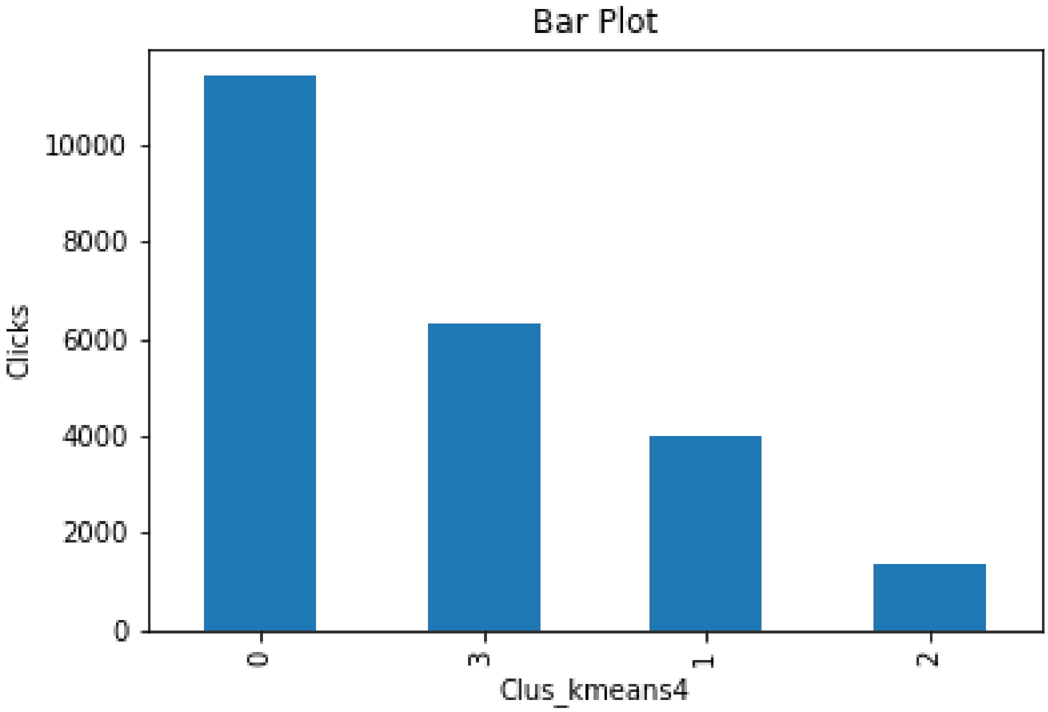
Solution:

silhouette scores = 0.40603313613720726

Optimum No of clusters – 3, because after that the elbow plot seems to flatten.

1.6. Profile the ads based on optimum number of clusters using silhouette score andyour 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.](4 marks)

Solution:



1.7. Conclude the project by providing summary of your learnings. (3 marks)

Solution:

□ Important cluster: 0 12839 1 8866 2 1361 Name: Clus_kmeans, dtype: int64

□ The optimum no of clusters is 3.
CPM is a better differentiator of cluster below of its dispersion among clusters.

PART 2

Problem 2: PCA

PCA FH (FT): Primary census abstract for female headed households excluding institutional households (India & States/UTs - District Level), Scheduled tribes - 2011 PCA for Female Headed Household Excluding Institutional Household.

The Indian Census has the reputation of being one of the best in the world. The first Census in India was conducted in the year 1872. This was conducted at different points of time in different parts of the country. In 1881 a Census was taken for the entire country simultaneously. Since then, Census has been conducted every ten years, withouta 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 andtwenty 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 andare 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

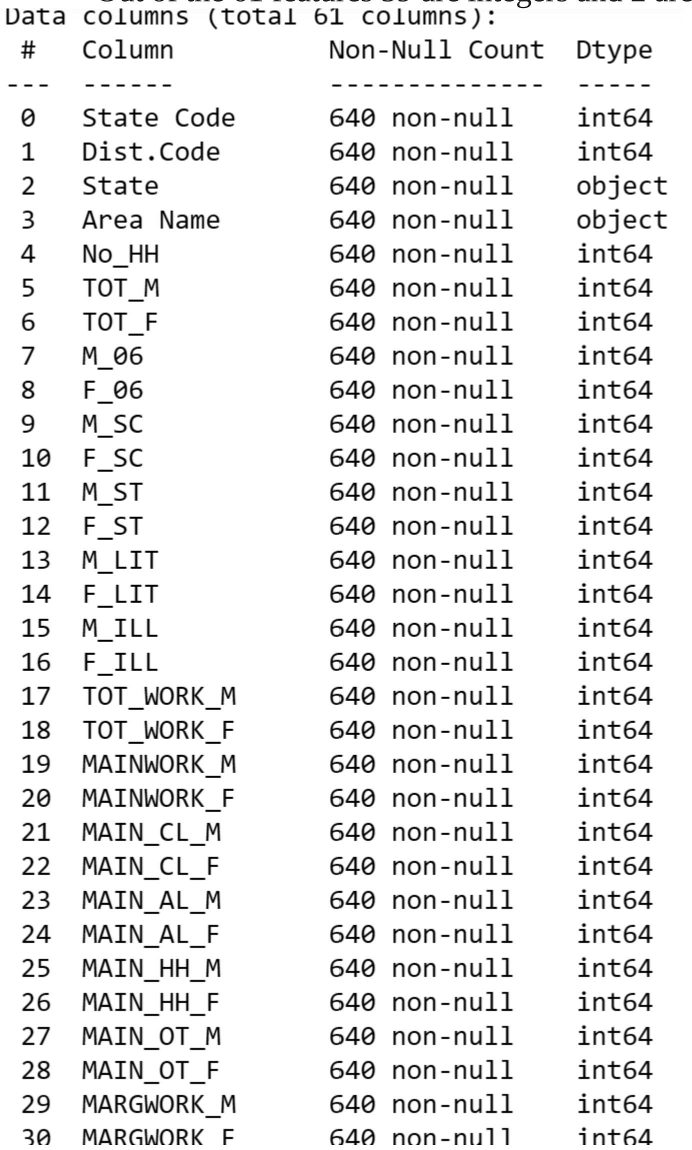
□

2.1. Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc. (4 marks)

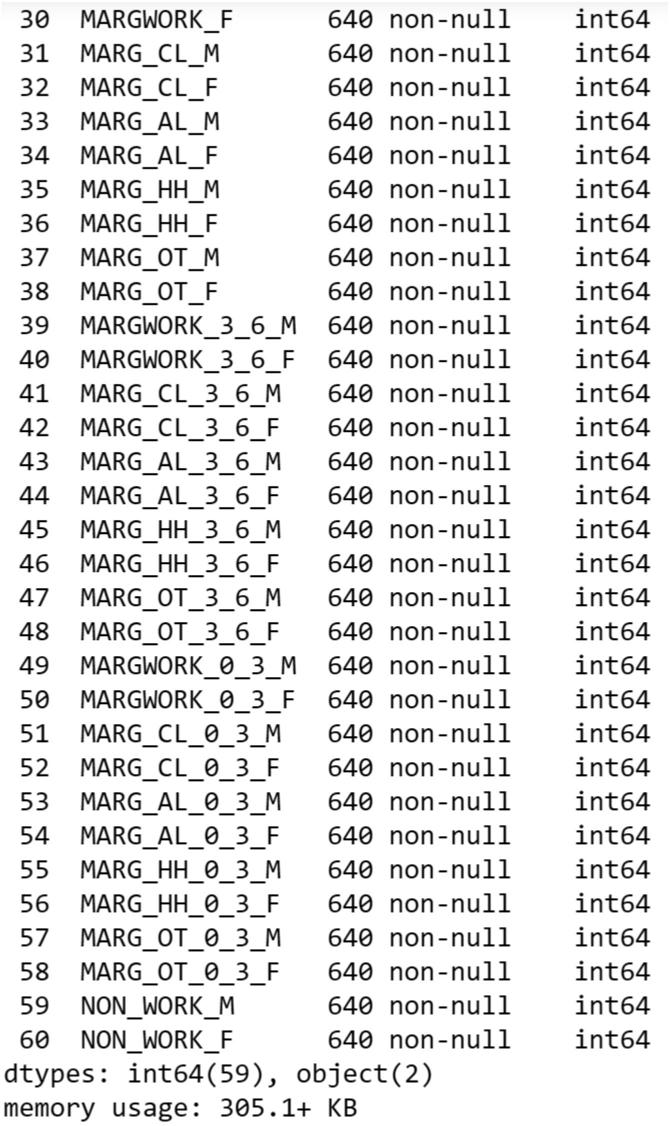
Solution:

The census data set has [640 rows x 61 columns].

Out of the 61 features 59 are integers and 2 are of object type.



```
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
```



```
 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)
memory usage: 305.1+ KB
```

The data has no null values and duplicate values in features.

2.2. Perform detailed Exploratory analysis by creating certain questions like

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

Solution: Highest – Lakshadweep
Lowest- Andhra Pradesh

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

Solution: Highest – Lakshadweep
Lowest- Krishna

2.1. We choose not to treat outliers for this case. Do you think that treating outliers forthis case is necessary? (1 marks)

Solution:

Yes, because PCA is sensitive to outliers. For details refer code.

Scale the Data using z-score method. Does scaling have any impact on outliers?Compare boxplots before and after scaling and comment. (3 marks

Yes There is impact.

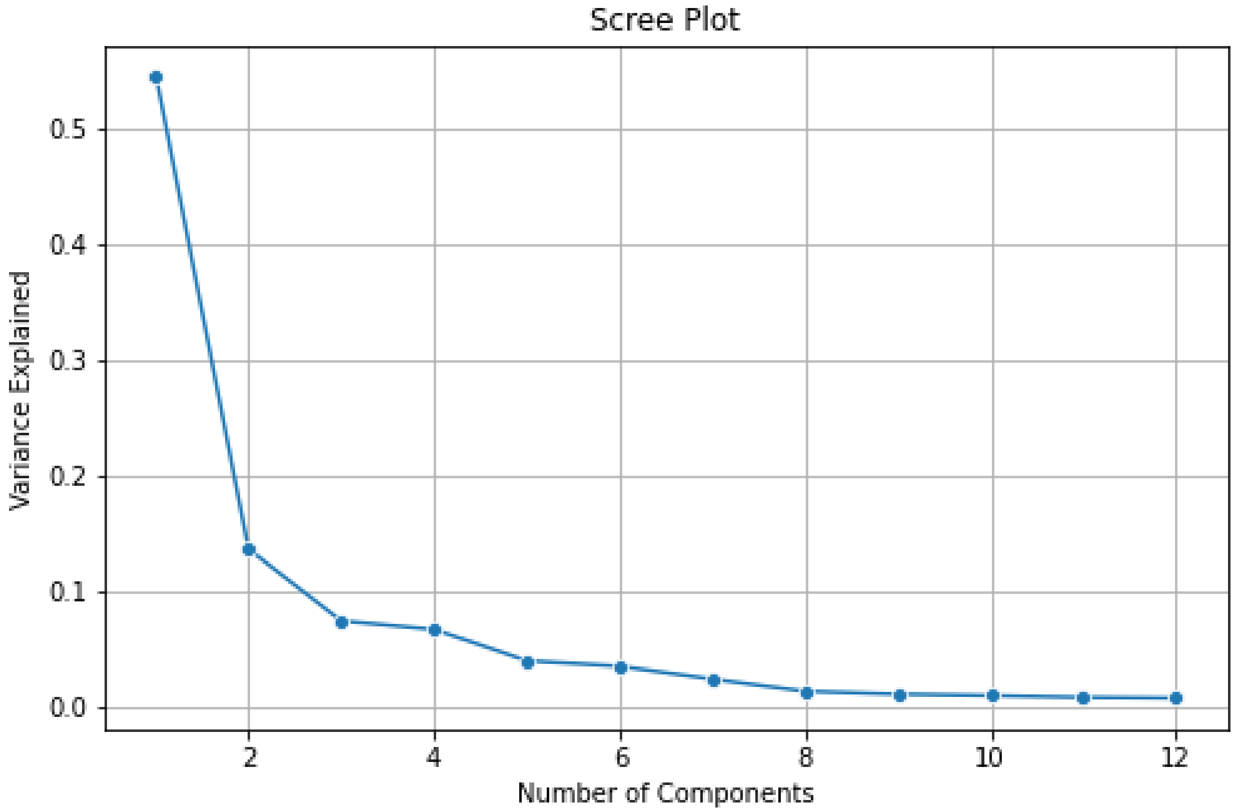
2.1. Perform all the required steps for PCA (use sklearn only) Create the covarianceMatrix Get eigen values and eigen vector. (4 marks)

Solution:

```
array([[ 1.68556868e-01,  1.66572192e-01,  1.64204972e-01,
         1.64599391e-01,  1.53016163e-01,  1.52922221e-01,
         2.74156515e-02,  2.83028895e-02,  1.63097150e-01,
         1.47329455e-01,  1.63962494e-01,  1.66977622e-01,
         1.60830722e-01,  1.46368920e-01,  1.46619648e-01,
         1.23633217e-01,  1.04920245e-01,  7.53781560e-02,
         1.14022386e-01,  7.33464097e-02,  1.33179026e-01,
         8.38377803e-02,  1.23463705e-01,  1.10539932e-01,
         1.67547515e-01,  1.57780024e-01,  8.50761645e-02,
         5.09925494e-02,  1.31506135e-01,  1.16217134e-01,
         1.43797637e-01,  1.29799111e-01,  1.56882634e-01,
         1.48333709e-01,  1.66782687e-01,  1.62454488e-01,
         1.68348123e-01,  1.57863299e-01,  9.58883338e-02,
         5.33417106e-02,  1.31408556e-01,  1.12338148e-01,
         1.42485656e-01,  1.26572235e-01,  1.55883752e-01,
         1.47308659e-01,  1.53209040e-01,  1.42829910e-01,
         5.45756378e-02,  4.34840001e-02,  1.24894532e-01,
         1.18485606e-01,  1.42876279e-01,  1.34545635e-01,
         1.52042271e-01,  1.32060184e-01,  8.46941616e-03],
       [-9.42576150e-02, -1.09280207e-01, -2.62155024e-02,
        -2.44896069e-02, -4.93224160e-02, -5.59902532e-02,
         2.71563578e-02,  2.95976501e-02, -1.19850333e-01,
        -1.57248318e-01, -1.09042343e-02, -1.30470238e-02,
        -1.38182068e-01, -8.82801278e-02, -1.80683042e-01,
        -1.54251752e-01,  6.09236462e-02,  8.68478462e-02,
        -3.32959225e-02, -5.94430570e-02, -8.14513129e-02,
        -8.60907361e-02, -2.17103944e-01, -2.13547005e-01,
         8.84575642e-02,  1.22215370e-01,  2.70509752e-01,
         2.49330271e-01,  1.62595317e-01,  1.38716805e-01,
         6.23819961e-02,  1.88447595e-02, -9.50217291e-02,
        -1.23219307e-01, -4.82649937e-02, -1.10050848e-01,
         7.24707790e-02,  9.99683961e-02,  2.64767353e-01,
         2.46957238e-01,  1.55587694e-01,  1.23192647e-01,
         5.65795782e-02,  9.43843433e-03, -9.87348883e-02,
        -1.30975717e-01,  1.47197064e-01,  1.78378739e-01,
         2.53680155e-01,  2.43575325e-01,  1.82052299e-01,
         1.78089235e-01,  7.93813689e-02,  4.55048849e-02,
        -7.06126558e-02, -7.82294803e-02,  3.49126266e-02],
       [ 5.41119506e-02,  2.04784813e-02,  6.58425721e-02,
  5.86170652e-02,  1.10948476e-02, -2.92712036e-02,
 -1.73216119e-01, -1.93276010e-01,  7.25792379e-02,
  9.55196930e-02, -4.38557339e-03, -1.02749638e-01,
  3.65819253e-02, -1.13242836e-01,  3.93880687e-02,
 -1.12905154e-01, -6.83685490e-02, -7.18678665e-02,
 -2.61762146e-01, -3.08207952e-01,  6.02204333e-02,
 -5.39507392e-02,  1.21300917e-01,  5.96790680e-02,
  1.24580925e-02, -7.50067644e-02,  1.75734440e-01,
  2.17093589e-01, -1.58716163e-01, -2.85278221e-01,
  3.13292804e-02, -4.36581112e-02,  1.23993056e-01,
  9.38244847e-02,  6.75799930e-02,  7.88008508e-02,
 -1.87332605e-03, -1.00164013e-01,  1.38578081e-01,
  2.03016613e-01, -1.70446259e-01, -3.03141572e-01,
  3.21204698e-02, -4.48946412e-02,  1.22883795e-01,
  9.10156990e-02,  6.95448067e-02,  8.36429313e-03,
  2.31576712e-01,  2.38911828e-01, -1.03201057e-01,
 -1.98681199e-01,  2.75477241e-02, -3.85077939e-02,
  1.21757987e-01,  9.14298576e-02,  1.78160787e-01],
[ 3.11939006e-02,  8.17622931e-02, -5.26012811e-03,
 -9.22268392e-03, -1.74783047e-02,  2.13605683e-02,
  1.88003073e-01,  1.94692674e-01,  5.63619267e-02,
  9.63530063e-02, -4.34822587e-02,  4.36815822e-02,
  5.10647739e-02,  2.08266312e-01,  8.00730190e-02,
  2.28255463e-01,  4.26390602e-02,  2.65908305e-01,
  5.42926063e-02,  2.10202062e-01, -1.33695022e-01,
 -4.88278183e-02,  8.21758930e-02,  1.51911059e-01,
 -8.91964770e-02,  8.17215092e-02,  1.07493924e-01,
  2.28218706e-01, -1.42288046e-01,  3.54623678e-02,
 -2.26116677e-01, -1.91996226e-01, -4.41387622e-02,
  1.82411493e-02,  1.09973163e-02,  1.85890292e-02,
 -9.37240726e-02,  9.31315693e-02,  7.20705287e-02,
  2.35003139e-01, -1.35116804e-01,  6.00689736e-02,
 -2.25958834e-01, -1.86896861e-01, -4.44559948e-02,
  1.49044326e-02, -6.49897676e-02,  3.94651197e-02,
  1.67013687e-01,  2.02944335e-01, -1.63482914e-01,
 -5.06433082e-02, -2.18333632e-01, -1.99965139e-01,
 -3.98013230e-02,  2.79785099e-02, -2.34726822e-01],
[-5.32363691e-02, -2.45632746e-02, -8.46825966e-02,
 -7.74052671e-02, -1.78645159e-01, -1.58114141e-01,
  3.90787020e-01,  4.01848297e-01, -2.35245556e-02,
  4.76975144e-02, -1.30712800e-01, -1.35186163e-01,
 -2.98557433e-02, -3.84748280e-02, -4.65299037e-02,
 -8.06246363e-02, -3.19036464e-01, -2.49243219e-01,
 -2.29974629e-01, -1.93894389e-01, -8.95096381e-02,
 -5.69866245e-02,  6.26556181e-02,  8.35147143e-02,
  5.09353965e-02,  8.98132692e-02, -1.56437845e-02,
 -4.41447373e-02, -4.22647878e-03,  5.36236986e-02,
 -1.93923399e-02,  4.93422309e-02,  1.17178725e-01,
  1.75593556e-01, -7.22082130e-02, -1.63845880e-02,
  4.24113699e-02,  7.04764441e-02, -6.42343554e-03,
 -4.93241141e-02, -1.75761664e-02,  3.48994623e-02,
 -2.49852045e-02,  3.98788976e-02,  1.06640786e-01,
  1.49584877e-01,  8.20049597e-02,  1.40158913e-01,
 -3.24267685e-02, -3.06178195e-02,  4.95596323e-02,
  1.11663785e-01, -8.51741525e-05,  7.50801619e-02,
  1.62265224e-01,  2.46983872e-01, -1.86888891e-01],
[-7.26755768e-02, -4.11236812e-02, -1.52495797e-01,
 -1.50023648e-01, -2.53152162e-02,  1.25640152e-03,
  5.54145781e-02,  6.34765829e-02, -5.59022505e-02,
 -5.33914996e-02, -1.11358570e-01, -1.41201716e-02,
 -7.67230274e-03,  1.00940293e-01,  1.32135592e-02,
  1.23580187e-01,  9.12885427e-03,  1.21919988e-01,
 -6.94185130e-03,  4.30402541e-02,  1.99914037e-01,
  4.39252706e-01,  6.19590014e-03,  5.83776306e-02,
 -9.38356778e-02,  4.27563433e-03,  3.43708520e-02,
  1.00241336e-01, -1.40297903e-01, -1.00839711e-01,
  1.14343770e-01,  3.71217933e-01, -7.28194184e-02,
 -1.75486248e-02, -1.29435240e-01, -1.01552926e-01,
 -9.97106348e-02,  9.05487211e-03,  1.12711046e-02,
  9.90589899e-02, -1.42349313e-01, -9.27692656e-02,
  1.22117298e-01,  3.90190075e-01, -7.20203509e-02,
 -8.06248619e-03, -6.38769343e-02, -1.06336894e-02,
  7.69217133e-02,  9.84375398e-02, -1.24596031e-01,
 -1.18641890e-01,  8.42978819e-02,  3.02848512e-01,
 -7.22409805e-02, -4.98625018e-02, -1.77240938e-01],
[ 6.32386561e-02, -1.43000614e-02,  1.02875210e-01,
  1.04884008e-01, -3.15882853e-02, -1.02534542e-01,
  4.97626411e-01,  4.48189702e-01,  2.71437140e-02,
 -3.48417319e-02,  1.57529876e-01,  2.10106375e-02,
  5.13820540e-02, -6.60726696e-02,  6.03772843e-02,
 -4.08428067e-02,  3.16488608e-01,  1.47067348e-01,
  4.50386351e-03, -1.21833480e-01,  8.57401599e-02,
 -2.25495447e-02,  1.00653318e-02, -1.94786040e-02,
 -3.97009595e-03, -1.12052761e-01,  3.02033599e-02,
 -2.05139602e-02, -2.85274346e-05, -1.12401828e-01,
  4.41792833e-02,  1.46865035e-02, -2.47595364e-02,
 -1.00911381e-01,  7.09285375e-02,  1.00385352e-02,
  1.65758563e-03, -9.28563528e-02,  5.64139714e-02,
  2.81214080e-03,  3.65398548e-03, -1.04372996e-01,
  5.09916312e-02,  2.71395500e-02, -2.32416958e-02,
 -7.83718658e-02, -2.64480852e-02, -1.59900446e-01,
 -2.53180244e-02, -7.17091614e-02, -1.48018270e-02,
 -1.28990848e-01,  1.99069807e-02, -2.20226563e-02,
 -3.07608381e-02, -1.69698179e-01,  4.15644959e-01],
[ 8.60259811e-02,  7.52847058e-02,  1.12928703e-01,
  1.09406972e-01, -1.02084142e-01, -1.06422806e-01,
 -1.58670569e-02, -2.21947173e-02,  7.98483638e-02,
  9.05025050e-02,  9.32445674e-02,  3.73812786e-02,
  7.86038340e-02, -2.73661743e-03,  9.13224498e-02,
  1.09254068e-02, -2.64471838e-01, -3.53631983e-01,
  8.33998780e-02, -6.43939421e-02, -1.73000166e-01,
  4.20044266e-01,  1.46375095e-01,  8.98263041e-02,
 -1.64612127e-03, -3.90601993e-02,  6.35139764e-02,
  2.08508059e-02,  1.24937610e-01,  1.45162534e-02,
 -2.54365637e-01,  9.40470574e-02, -1.24750618e-01,
 -1.73726319e-01,  8.83655085e-02,  1.04304393e-01,
 -7.27360086e-03, -6.52678706e-02,  4.97319977e-02,
 -7.30608333e-03,  1.27712791e-01, -6.93058066e-03,
 -2.61331536e-01,  1.05987943e-01, -1.27043357e-01,
 -1.78141786e-01,  2.12032352e-02,  4.41517869e-02,
  8.44015363e-02,  8.28218742e-02,  1.07149837e-01,
  8.53489832e-02, -2.21860297e-01,  5.59953126e-02,
 -1.05547617e-01, -1.34134242e-01,  1.77249357e-01],
[ 5.00954661e-02,  4.35689418e-02,  2.25148636e-02,
  2.47550846e-02, -3.63514049e-02, -3.55870395e-02,
  1.95386855e-02,  4.74881688e-02,  7.08615476e-02,
  6.29200526e-02, -1.44085313e-02,  4.84068191e-03,
  3.77228930e-02,  8.24656082e-02,  5.18163844e-02,
  1.18615632e-01, -3.54320013e-01, -5.11923670e-02,
 -2.03779104e-01, -2.17520585e-02,  3.55113970e-01,
 -1.72714871e-01,  1.59596158e-01,  2.45616655e-01,
 -3.47300158e-02, -4.46663344e-02,  2.31339194e-02,
  8.13374363e-03,  1.32715186e-02,  6.03135427e-02,
  2.36786406e-01, -1.40450756e-01, -1.23970239e-01,
 -1.84769432e-01,  5.89668917e-02,  2.26872744e-02,
 -4.16643468e-02, -3.85965359e-02,  2.18375598e-02,
  1.26214809e-03, -5.76712008e-03,  4.21752472e-02,
  2.37993394e-01, -1.60831068e-01, -1.11400151e-01,
 -1.23277077e-01, -4.40731012e-03, -5.89351841e-02,
  2.33033628e-02,  2.31225494e-02,  8.89507707e-02,
  1.15760726e-01,  2.24073673e-01, -7.62210278e-02,
 -1.78739445e-01, -3.84653925e-01, -1.69737865e-01],
[ 5.61691405e-02,  7.21620479e-02,  2.32834035e-01,
  2.52703464e-01, -4.11153670e-01, -4.06433078e-01,
 -1.10443686e-01, -8.91590244e-02,  3.67801398e-02,
  1.01164251e-01,  1.04186323e-01,  1.28725390e-02,
 -8.63827198e-02, -2.58012415e-02, -1.02650261e-01,
 -3.11488298e-02,  3.57306744e-02,  4.02248972e-01,
 -1.88370132e-01, -1.21431449e-01, -1.18585381e-01,
  1.60211255e-01, -7.83993979e-02, -1.06692073e-01,
  1.15379717e-02, -2.29148958e-03, -2.60957284e-02,
 -3.52541636e-02,  1.66053242e-02, -4.70985295e-03,
  2.03593098e-02,  2.92666333e-02,  1.31650543e-02,
  2.36419433e-02,  1.86141750e-01,  1.10367900e-01,
  3.03779363e-02,  1.61985406e-02,  2.10039580e-02,
  1.10366128e-02,  1.80629470e-02, -4.19307261e-03,
  1.92316355e-02,  4.46706791e-02,  3.34818579e-02,
  2.75780688e-02, -6.53576810e-02, -5.82187238e-02,
 -1.17457493e-01, -1.37093247e-01,  9.87282896e-03,
 -6.01196159e-03,  2.33604746e-02, -1.65374310e-02,
 -8.87079720e-02,  6.05992955e-03, -3.15800240e-01],
[ 3.59616176e-02,  6.62625991e-02,  6.10570364e-02,
  6.37318540e-02,  1.95700987e-01,  2.26684766e-01,
  1.00980462e-01,  1.03682947e-01,  6.81412556e-02,
  1.35941393e-01, -5.90534868e-02, -5.67372669e-02,
 -8.48003316e-03, -1.17655279e-01, -7.70865611e-03,
 -7.01407476e-02,  4.35462656e-02,  7.73751519e-02,
  1.03046423e-01,  2.28478127e-02, -1.07220349e-01,
 -9.14711549e-02, -3.72633254e-02, -1.32790320e-01,
 -8.92797091e-03, -2.06591398e-01,  1.82143993e-02,
 -7.00968095e-02,  4.48168680e-02, -1.91302052e-01,
 -1.88496457e-02,  2.92795182e-02, -6.94878398e-02,
 -1.84960924e-01,  7.54985449e-02,  1.43465085e-01,
 -6.16351112e-02, -2.85702275e-01, -1.95872082e-02,
 -1.11044947e-01,  7.14171970e-03, -2.39959344e-01,
 -4.46821348e-02, -8.28801198e-03, -1.18247833e-01,
 -2.56668757e-01,  2.04649712e-01,  5.28544905e-02,
  9.18257729e-02,  2.44795602e-02,  1.93588830e-01,
 -9.78916790e-03,  6.77285096e-02,  1.37396523e-01,
  1.77374797e-01,  1.02176707e-01, -4.70929783e-01],
[ 1.64877481e-02,  1.93891419e-02,  9.43538468e-02,
  9.57862732e-02, -2.17035936e-01, -1.95181794e-01,
  7.63144263e-02,  6.34166061e-02, -1.98923876e-02,
 -1.10448852e-02,  1.17128043e-01,  6.43382364e-02,
 -3.68397204e-02, -6.36265294e-02, -4.84776676e-02,
 -6.93129127e-02, -2.64724945e-01, -2.75439749e-01,
  4.32463919e-01,  4.14995289e-01,  1.60094538e-01,
 -5.66995635e-02, -1.24434060e-01, -3.04156690e-01,
  2.48874608e-02, -2.61132679e-02,  1.44349410e-02,
  5.80087392e-02, -6.38127840e-02, -7.59938062e-02,
  1.28664372e-01, -2.45280545e-02,  9.62176453e-02,
  1.22087927e-02,  6.53504982e-02,  5.50683248e-02,
  4.08545960e-02, -2.05994141e-02, -3.69585428e-02,
  1.11219768e-02, -3.71954185e-02, -3.75113914e-02,
  1.60563246e-01, -9.76430895e-03,  1.10508984e-01,
  2.02653514e-03, -4.12839427e-02, -4.04223453e-02,
  1.15594484e-01,  1.60170134e-01, -1.67207598e-01,
 -1.98454462e-01,  1.78824934e-02, -6.65527174e-02,
  1.91209820e-02,  4.77885061e-02, -5.57674904e-02]])
```

2.1. Identify the optimum number of PCs (for this project, take at least 90% explainedvariance).

Show Scree plot. (3 marks)



Optimum No. is 5, after that scree plot flattens.

2.3. Compare PCs with Actual Columns and identify which is explaining most variance.Write inferences about all the principal components in terms of actual variables. (4 marks)

Solution:

```
array([0.5438538 , 0.67961021, 0.753649  , 0.82029602, 0.85984845,
       0.89472956, 0.91866361, 0.93209023, 0.94305104, 0.95275747,
       0.96089346, 0.96841283])
```

2.1. Write linear equation for first PC. (2 marks)

Solution:

No_HH + (0.17) *TOT_M + (0.17) * TOT_F + (0.16) * M_06 + (0.16) * F_06 + (0.15) * M_SC + (0.15) * F_SC + (0.03) * M_ST + (0.03) * F_ST + (0.16) * M_LIT + (0.15) * F_LIT + (0.16) * M_ILL + (0.17) * F_ILL + (0.16)

* TOT_WORK_M + (0.15) * TOT_WORK_F + (0.15) * MAINWORK_M + (0.12) * MAINWORK_F + (0.1) * MAIN_CL_M + (0.07) * MAIN_CL_F + (0.11) * MAIN_AL_M + (0.07) * MAIN_AL_F + (0.13) * MAIN_HH_M + (0.08) *

MAIN_HH_F + (0.12) * MAIN_OT_M + (0.11) * MAIN_OT_F + (0.16) * MARGWORK_M + (0.16) * MARGWORK_F + (0.08) * MARG_CL_M + (0.05) * MARG_CL_F + (0.13) * MARG_AL_M + (0.11) * MARG_AL_F + (0.14) * MARG_HH_M + (0.13) * MARG_HH_F + (0.16) * MARG_OT_M + (0.15) * MARG_OT_F + (0.16) * MARGWORK_3_6_M + (0.16) * MARGWORK_3_6_F + (0.17

) * MARG_CL_3_6_M + (0.16) * MARG_CL_3_6_F + (0.09) * MARG_AL_3_6_M + (0.05) * MARG_AL_3_6_F + (0.13) * MARG_HH_3_6_M + (0.11) * MARG_HH_3_6_F + (0.14) * MARG_OT_3_6_M + (0.12) * MARG_OT_3_6_F + (

0.15) * MARGWORK_0_3_M + (0.15) * MARGWORK_0_3_F + (0.15) * MARG_CL_0_3_M + (0.14) * MARG_CL_0_3_F + (0.05) * MARG_AL_0_3_M + (

0.04) * MARG_AL_0_3_F + (0.12) * MARG_HH_0_3_M + (0.12) * MARG_HH_0_3_F + (0.14) * MARG_OT_0_3_M + (0.13) * MARG_OT_0_3_F +