Business Report

Data Mining

Ayush Sharma

Table of Contents

Part 1 - Clustering

Part 2 – PCA

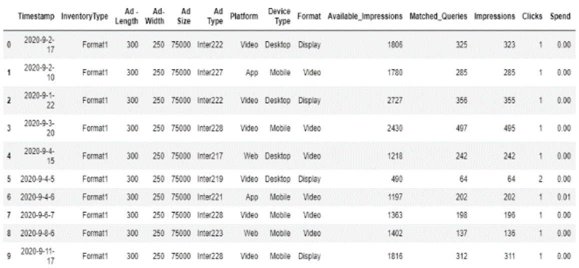
A. Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.	14
B. Perform detailed Exploratory analysis by creating certain questions like (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio?	15-16
C. We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?	17
D. Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.	18
E. Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.	19
F. Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.	20
G. Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.	-
H. Write linear equation for first PC.	-

Part 1 - Clustering

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

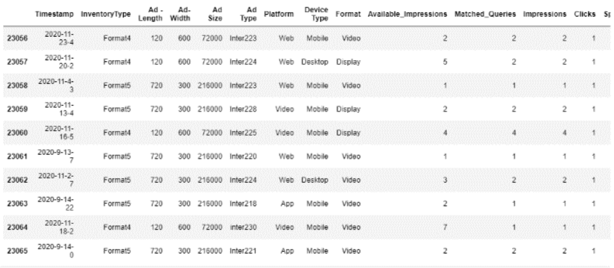
Ans:

Data Head



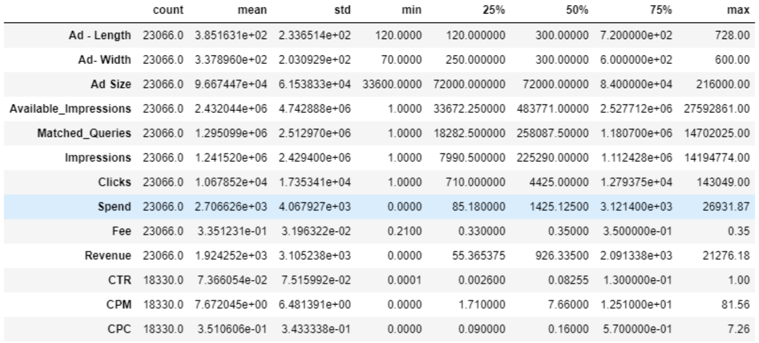
	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend
0	2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.00
1	2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.00
2	2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.00
3	2020-9-3-20	Format1	300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.00
4	2020-9-4-15	Format1	300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.00
5	2020-9-4-5	Format1	300	250	75000	Inter219	Video	Desktop	Display	490	64	64	2	0.00
6	2020-9-4-6	Format1	300	250	75000	Inter221	App	Mobile	Video	1197	202	202	1	0.01
7	2020-9-6-7	Format1	300	250	75000	Inter228	Video	Mobile	Video	1363	198	196	1	0.00
8	2020-9-8-6	Format1	300	250	75000	Inter223	Web	Mobile	Video	1402	137	136	1	0.00
9	2020-9-11-17	Format1	300	250	75000	Inter228	Video	Mobile	Display	1816	312	311	1	0.00

Data Tail



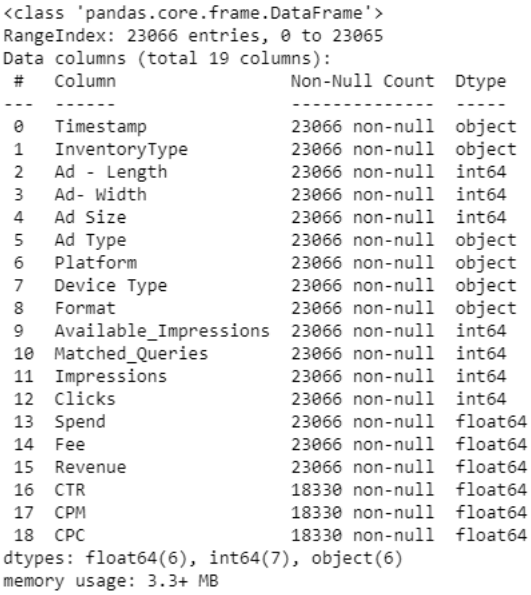
	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Sp
23056	2020-11-23-4	Format4	120	600	72000	Inter223	Web	Mobile	Video	2	2	2	1	
23057	2020-11-20-2	Format4	120	600	72000	Inter224	Web	Desktop	Display	5	2	2	1	
23058	2020-11-4-3	Format5	720	300	216000	Inter223	Web	Mobile	Video	1	1	1	1	
23059	2020-11-13-4	Format5	720	300	216000	Inter228	Video	Mobile	Display	2	2	2	1	
23060	2020-11-16-5	Format4	120	600	72000	Inter225	Video	Mobile	Display	4	4	4	1	
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	

Data Summary



	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	3.851631e+02	2.336514e+02	120.0000	120.000000	300.00000	7.200000e+02	728.00
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.0000	250.000000	300.00000	6.000000e+02	600.00
Ad Size	23066.0	9.667447e+04	6.153833e+04	33600.0000	72000.000000	72000.00000	8.400000e+04	216000.00
Available_Impressions	23066.0	2.432044e+06	4.742888e+06	1.0000	33672.250000	483771.00000	2.527712e+06	27592861.00
Matched_Queries	23066.0	1.295099e+06	2.512970e+06	1.0000	18282.500000	258087.50000	1.180700e+06	14702025.00
Impressions	23066.0	1.241520e+06	2.429400e+06	1.0000	7990.500000	225290.00000	1.112428e+06	14194774.00
Clicks	23066.0	1.067852e+04	1.735341e+04	1.0000	710.000000	4425.00000	1.279375e+04	143049.00
Spend	23066.0	2.706626e+03	4.067927e+03	0.0000	85.180000	1425.12500	3.121400e+03	26931.87
Fee	23066.0	3.351231e-01	3.196322e-02	0.2100	0.330000	0.35000	3.500000e-01	0.35
Revenue	23066.0	1.924252e+03	3.105238e+03	0.0000	55.365375	926.33500	2.091338e+03	21276.18
CTR	18330.0	7.366054e-02	7.515992e-02	0.0001	0.002600	0.08255	1.300000e-01	1.00
CPM	18330.0	7.672045e+00	6.481391e+00	0.0000	1.710000	7.66000	1.251000e+01	81.56
CPC	18330.0	3.510606e-01	3.433338e-01	0.0000	0.090000	0.16000	5.700000e-01	7.26

Data Info

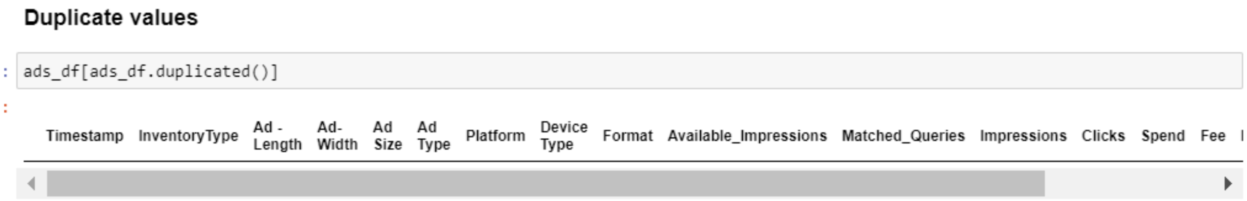


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23066 entries, 0 to 23065
Data columns (total 19 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Timestamp              23066 non-null  object
 1   InventoryType          23066 non-null  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
```

- The data consists of **23,066 rows** and **19 columns**
- There is a total of **13 numeric columns** and **6 categoric columns**
- It can be observed from the data info that null values exist in the **CTR, CPM and CPC columns** of the dataset

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

Ans:



Duplicate values

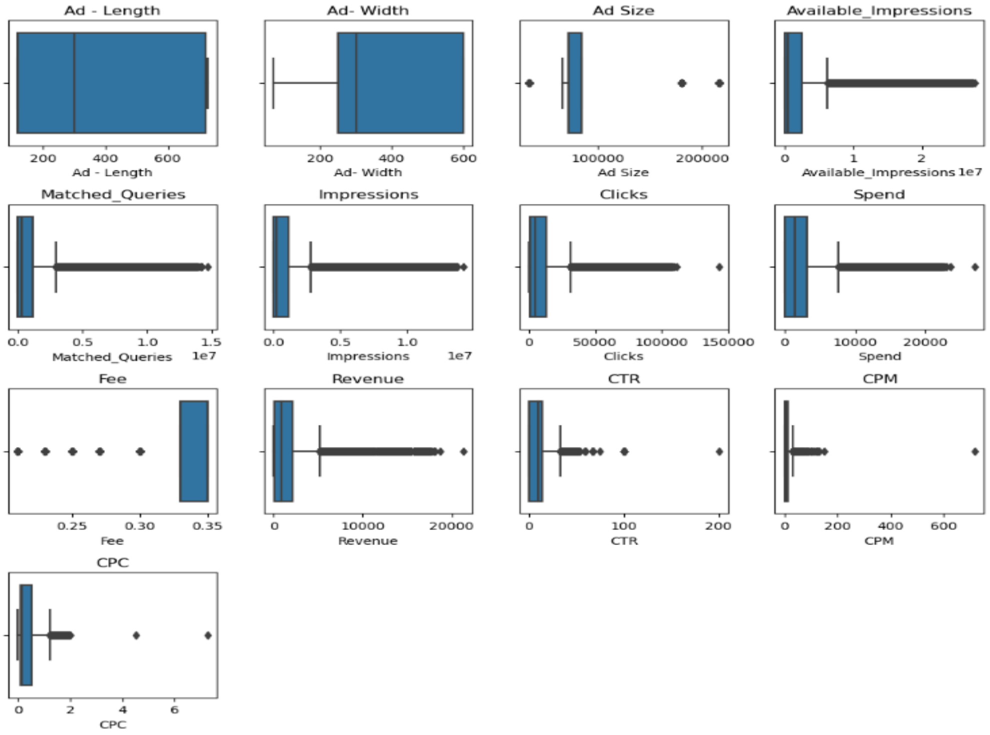
```
ads_df[ads_df.duplicated()]
```

	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee

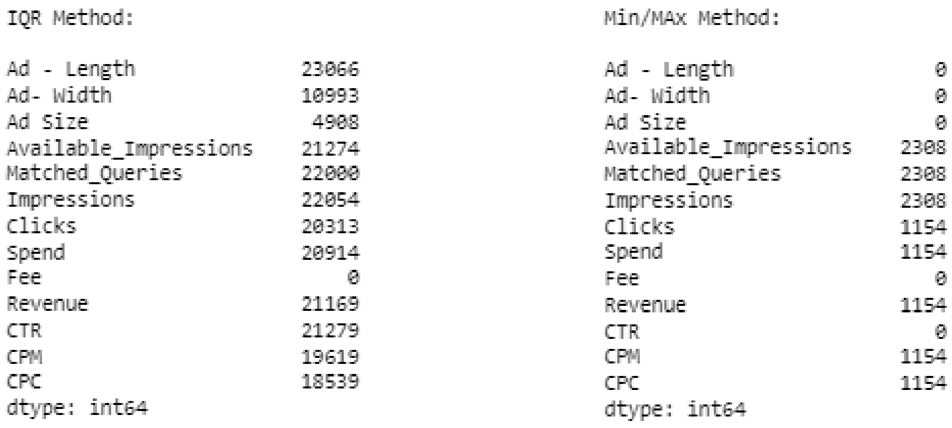
- There are **no duplicate entries** in the dataset.
- The missing values for CPC, CTR and CPM can be treated by using the formulae provided.

C. 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.

Ans:

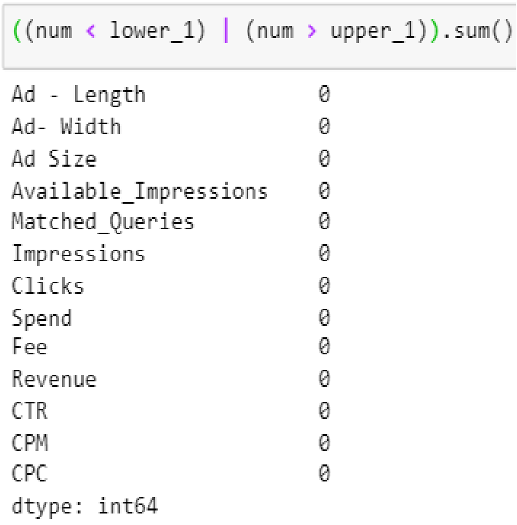


- It can be observed that there are various outliers in the columns.
- K-Means clustering is sensitive to outliers as they can significantly affect the centroids and hence distort the clusters.
- Outliers tend to pull the cluster centres towards them which causes the clusters to be improperly defined.
- Hence, it becomes important for us to treat such outliers before proceeding with K-Means clustering.



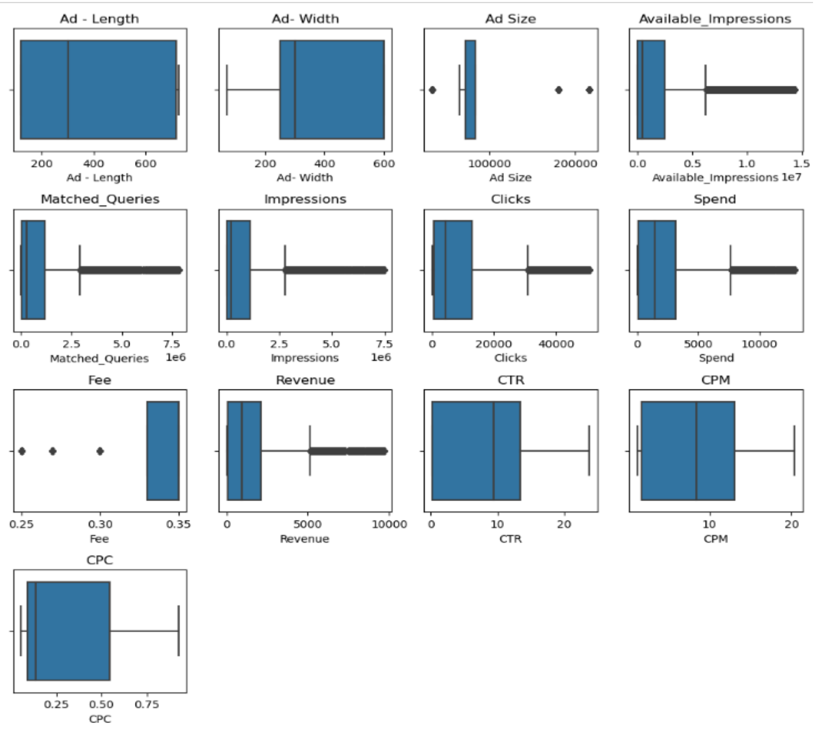
```
IQR Method:                                  Min/MAx Method:

Ad - Length                 23066            Ad - Length                 0
Ad- Width                   10993            Ad- Width                   0
Ad Size                      4908            Ad Size                     0
Available_Impressions       21274            Available_Impressions    2308
Matched_Queries             22000            Matched_Queries          2308
Impressions                 22054            Impressions              2308
Clicks                      20313            Clicks                   1154
Spend                       20914            Spend                    1154
Fee                             0            Fee                         0
Revenue                     21169            Revenue                  1154
CTR                         21279            CTR                         0
CPM                         19619            CPM                      1154
CPC                         18539            CPC                      1154
dtype: int64                                 dtype: int64
```

- It can be observed that the **number of outliers is comparatively higher when employing the IQR method over the Min/Max method** for the outlier calculation.
- Treating a greater number of outliers also results into decreasing the data variability which might not produce accurate results.
- Hence, we can proceed by treating the outliers using the Min/Max method



```
((num < lower_1) | (num > upper_1)).sum()
```

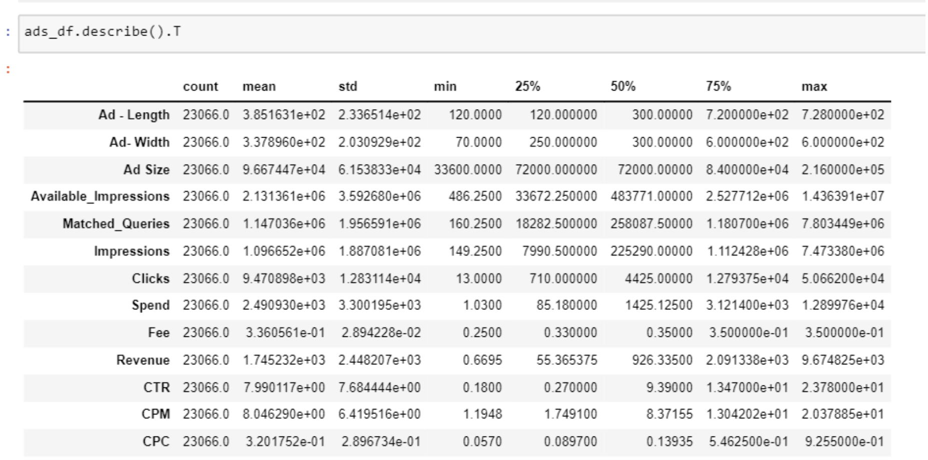
```
Ad - Length              0
Ad- Width                0
Ad Size                  0
Available_Impressions    0
Matched_Queries          0
Impressions              0
Clicks                   0
Spend                    0
Fee                      0
Revenue                  0
CTR                      0
CPM                      0
CPC                      0
dtype: int64
```



From the boxplots it can be visualized that the outliers have now been treated for the numeric variables.

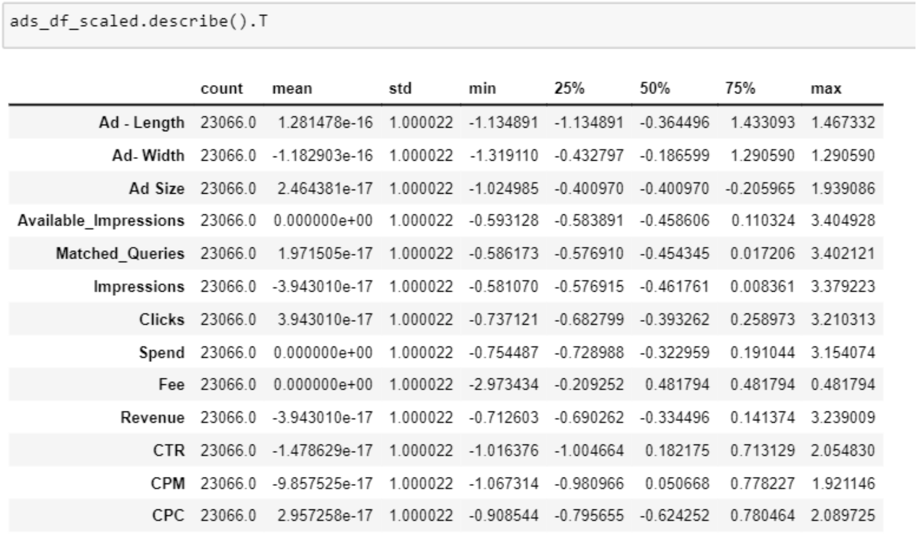
D. Perform z-score scaling and discuss how it affects the speed of the algorithm.

Ans:



```
ads_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	300.00000	7.200000e+02	7.280000e+02
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.0000	250.000000	300.00000	6.000000e+02	6.000000e+02
Ad Size	23066.0	9.667447e+04	6.153833e+04	33600.0000	72000.000000	72000.00000	8.400000e+04	2.160000e+05
Available_Impressions	23066.0	2.131361e+06	3.592680e+06	486.2500	33672.250000	483771.00000	2.527712e+06	1.436391e+07
Matched_Queries	23066.0	1.147036e+06	1.956591e+06	160.2500	18282.500000	258087.50000	1.180700e+06	7.803449e+06
Impressions	23066.0	1.096652e+06	1.887081e+06	149.2500	7990.500000	225290.00000	1.112428e+06	7.473380e+06
Clicks	23066.0	9.470898e+03	1.283114e+04	13.0000	710.000000	4425.00000	1.279375e+04	5.066200e+04
Spend	23066.0	2.490930e+03	3.300195e+03	1.0300	85.180000	1425.12500	3.121400e+03	1.289976e+04
Fee	23066.0	3.360561e-01	2.894228e-02	0.2500	0.330000	0.35000	3.500000e-01	3.500000e-01
Revenue	23066.0	1.745232e+03	2.448207e+03	0.6695	55.365375	926.33500	2.091338e+03	9.674825e+03
CTR	23066.0	7.990117e+00	7.684444e+00	0.1800	0.270000	9.39000	1.347000e+01	2.378000e+01
CPM	23066.0	8.046290e+00	6.419516e+00	1.1948	1.749100	8.37155	1.304202e+01	2.037885e+01
CPC	23066.0	3.201752e-01	2.896734e-01	0.0570	0.089700	0.13935	5.462500e-01	9.255000e-01



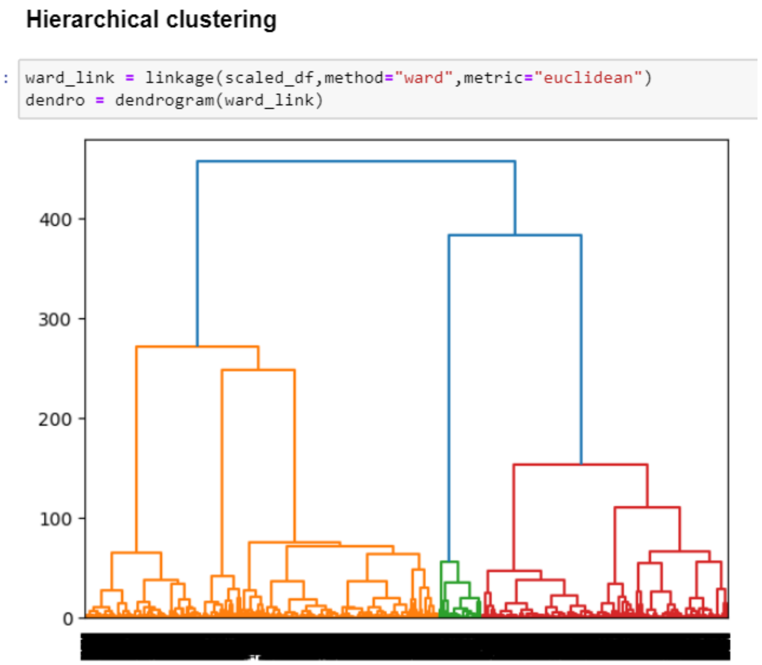
```
ads_df_scaled.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	1.281478e-16	1.000022	-1.134891	-1.134891	-0.364496	1.433093	1.467332
Ad- Width	23066.0	-1.182903e-16	1.000022	-1.319110	-0.432797	-0.186599	1.290590	1.290590
Ad Size	23066.0	2.464381e-17	1.000022	-1.024985	-0.400970	-0.400970	-0.205965	1.939086
Available_Impressions	23066.0	0.000000e+00	1.000022	-0.593128	-0.583891	-0.458606	0.110324	3.404928
Matched_Queries	23066.0	1.971505e-17	1.000022	-0.586173	-0.576910	-0.454345	0.017206	3.402121
Impressions	23066.0	-3.943010e-17	1.000022	-0.581070	-0.576915	-0.461761	0.008361	3.379223
Clicks	23066.0	3.943010e-17	1.000022	-0.737121	-0.682799	-0.393262	0.258973	3.210313
Spend	23066.0	0.000000e+00	1.000022	-0.754487	-0.728988	-0.322959	0.191044	3.154074
Fee	23066.0	0.000000e+00	1.000022	-2.973434	-0.209252	0.481794	0.481794	0.481794
Revenue	23066.0	-3.943010e-17	1.000022	-0.712603	-0.690262	-0.334496	0.141374	3.239009
CTR	23066.0	-1.478629e-17	1.000022	-1.016376	-1.004664	0.182175	0.713129	2.054830
CPM	23066.0	-9.857525e-17	1.000022	-1.067314	-0.980966	0.050668	0.778227	1.921146
CPC	23066.0	2.957258e-17	1.000022	-0.908544	-0.795655	-0.624252	0.780464	2.089725

- Scaling the data by converting it into its respective Z-scores **helps in standardization** and is an important aspect of data pre-processing.
- It ensures that **each feature contributes equally to the distance calculation** and hence helps in the smooth functioning of the algorithms.
- It can be observed from the above summary of the data that **prior to scaling, the data ranges were very varied** however **after it has been scaled, the data has become standardized with similar data ranges**.

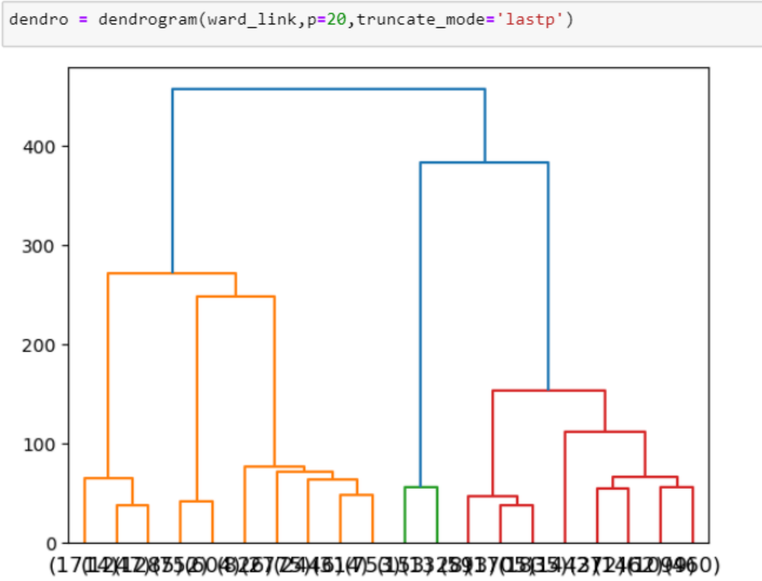
E. **Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.**

Ans:



Hierarchical clustering

```
ward_link = linkage(scaled_df,method="ward",metric="euclidean")
dendro = dendrogram(ward_link)
```

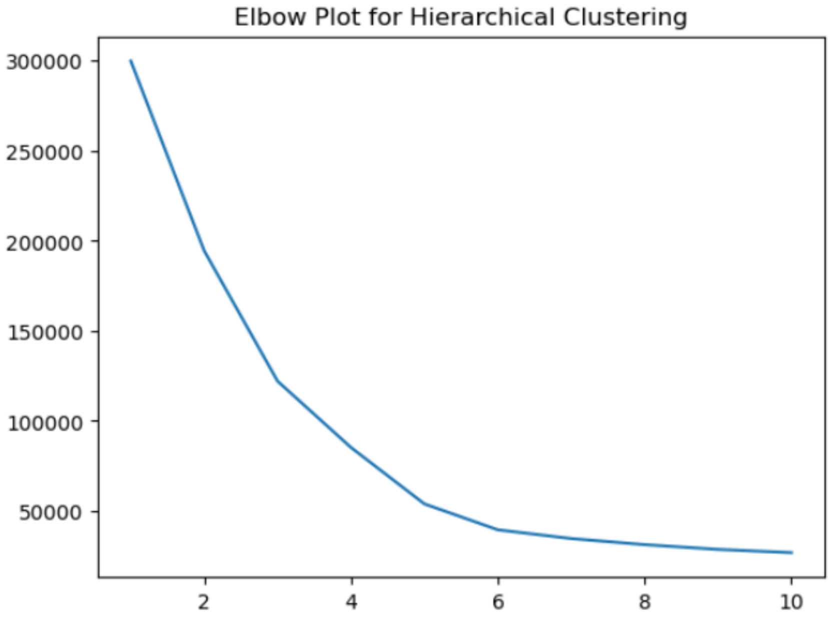


```
dendro = dendrogram(ward_link,p=20,truncate_mode='lastp')
```

As per the dendrogram, it can be observed that the ideal number of clusters should be 3.

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

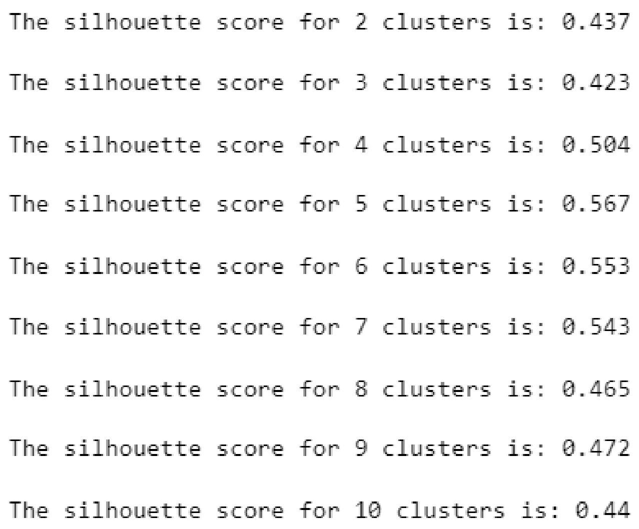
Ans:



As per the elbow plot and the WSS for different numbers of clusters, it seems like **5 clusters are ideal for the K-Means algorithm** as the drop in the WSS values after n=5 isn't as steep as it was for the previous values of n.

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

Ans:



```
The silhouette score for 2 clusters is: 0.437

The silhouette score for 3 clusters is: 0.423

The silhouette score for 4 clusters is: 0.504

The silhouette score for 5 clusters is: 0.567

The silhouette score for 6 clusters is: 0.553

The silhouette score for 7 clusters is: 0.543

The silhouette score for 8 clusters is: 0.465

The silhouette score for 9 clusters is: 0.472

The silhouette score for 10 clusters is: 0.44
```

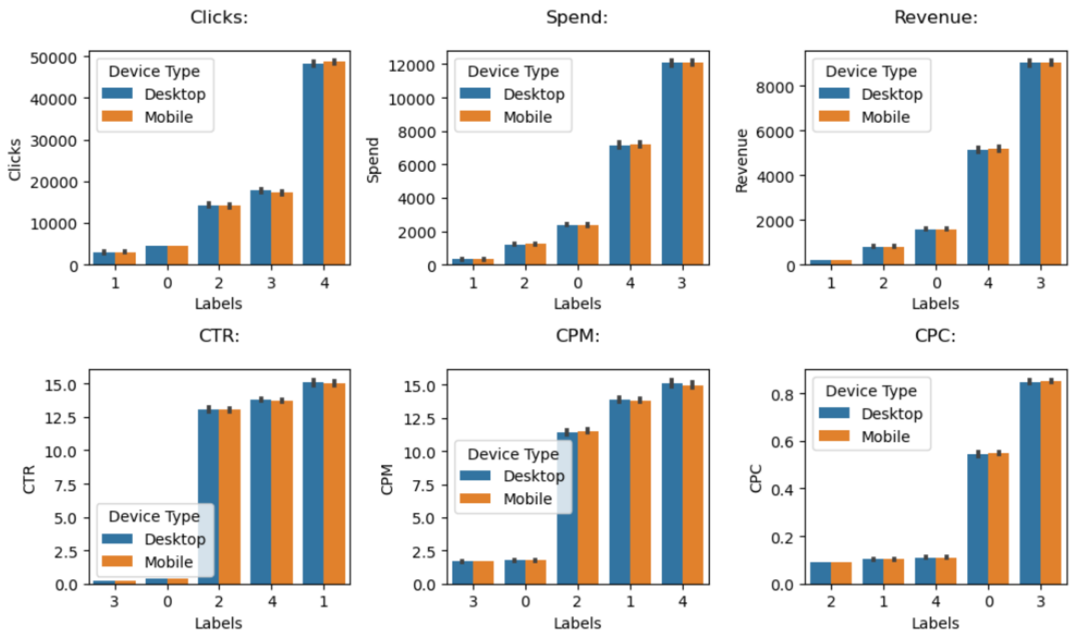
Similar to our previous conclusion derived from the elbow plot and WSS values, it can be observed that **5 clusters are ideal for the K-Means algorithm** as per the silhouette scores.

H. Profile the ads based on optimum number of clusters using silhouette score and your domain understanding.

Ans: As per the conclusions drawn from the scree plot, WSS values and silhouette scores, we can proceed by creating 5 clusters for the dataset. We will input the value of n as 5 and thereafter assign the corresponding cluster labels to our original dataset.

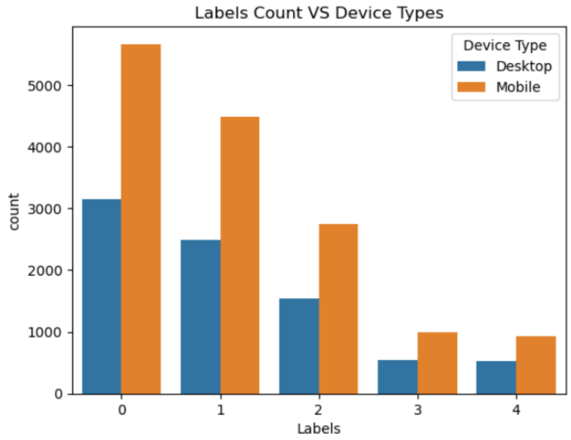
Following inferences can be derived from the silhouette sample scores:

- A negative silhouette width indicates that an observation has been placed incorrectly in a cluster as it is closer to another cluster.
- A total of **33 silhouette width values** out of **approximately 23,000 data entries** have negative values.
- This is a very negligible number which indicates that **our observations have been correctly allocated within the clusters**.



- Cluster 0: The ads category generating **average number of Clicks, Spend and Revenue values with low values of CTR, CPM and CPC** for both desktop and mobile devices.
- Cluster 1: The ads category generating **the lowest number of Clicks, Spend and Revenue values however consisting of the highest values of CTR and high values of CPM.**
- Cluster 2: The ads category generating **lowest values of CPC and average values for Spend, Revenue, Clicks and CPM.**
- Cluster 3: The ads category **generating highest Spend, Revenue and CPC values however consisting of lowest values of CTR and CPM.**
- Cluster 4: The ads category **generating highest values of Clicks, CTR and CPM and high values of Spend and Revenue.**

- Mobiles dominate desktops in all the categories
- The ads category with the 0th label has the most count for both the devices
- The ads category with the 4th label has the least count for both the devices



I. Conclude the project by providing summary of your learnings.

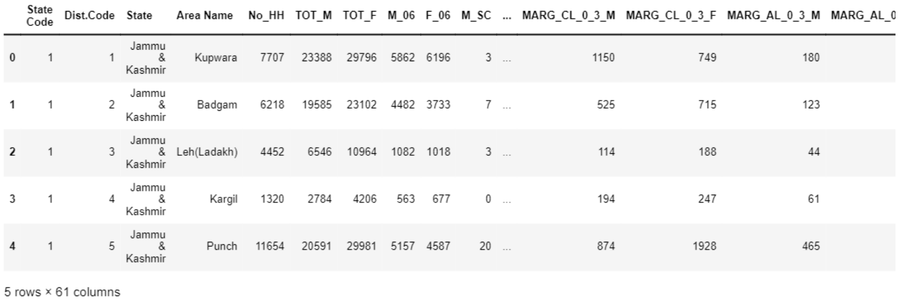
Ans: The following summary can be drawn from the clustering analysis:

- The ads pertaining to clusters 0 and 2 lie in the low to average range when compared to the rest of the clusters. They lie in the middle of almost all the metrics and the ad agency can device new strategies and planning to increase the promotion of such ads.
- The ads pertaining to cluster 1 lie in the low yielding range with the lowest values for Clicks, Spend and Revenue. The CTR values are the highest for this cluster which means that despite of being viewed, the revenue generation for such ads is low. The ad agency can either undertake certain drastic measures to promote or upsell this category of ads to ensure greater revenue generation or it can replace it with more featuring ads.
- The ads pertaining to cluster 3 are responsible for the most revenue generation along with the Spend and CPC costs. The CTR values however are the lowest for such ads and the ads agency can resort to new lucrative strategies in order to promote them.
- The ads pertaining to cluster 4 have high values of Clicks, CTR and CPM. The revenue generation for such ads can be increased by investing more resources in such ad categories.

Part 2 - PCA

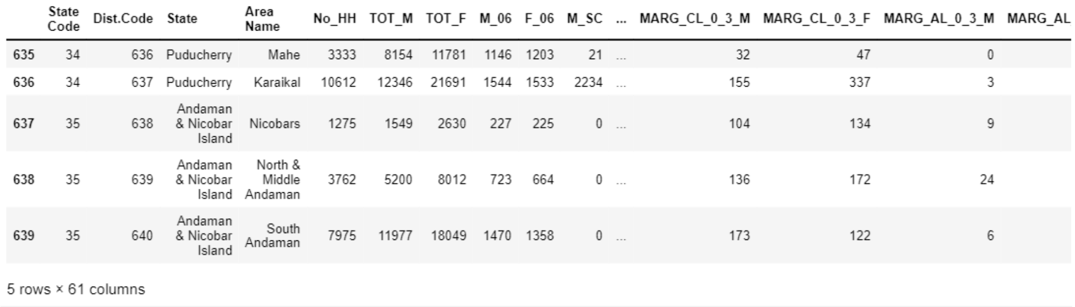
A. Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.

Ans:



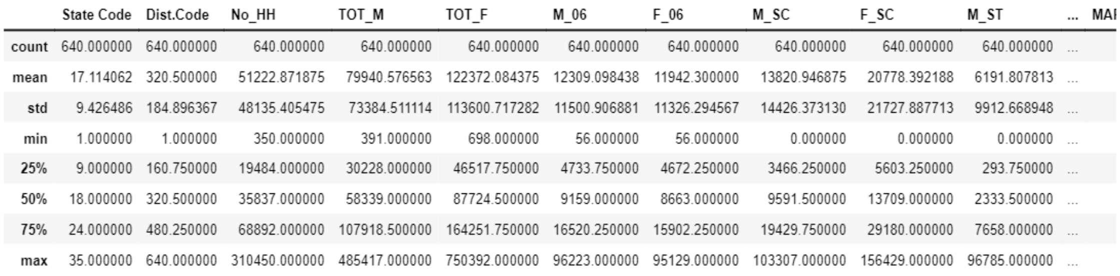
	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180	
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123	
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	...	114	188	44	
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	...	194	247	61	
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	...	874	1928	465	

5 rows × 61 columns

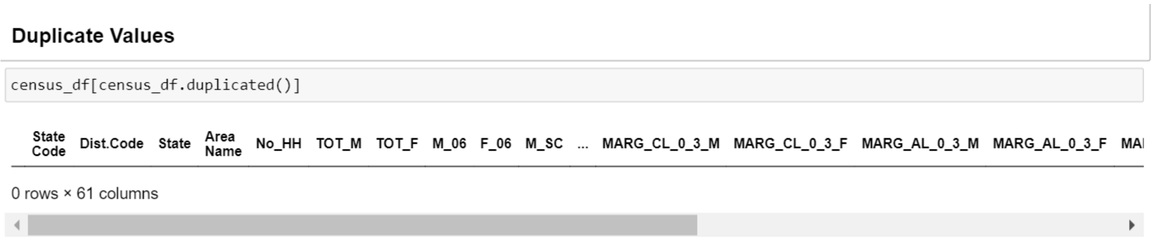


	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL
635	34	636	Puducherry	Mahe	3333	8154	11781	1146	1203	21	...	32	47	0	
636	34	637	Puducherry	Karaikal	10612	12346	21691	1544	1533	2234	...	155	337	3	
637	35	638	Andaman & Nicobar Island	Nicobars	1275	1549	2630	227	225	0	...	104	134	9	
638	35	639	Andaman & Nicobar Island	North & Middle Andaman	3762	5200	8012	723	664	0	...	136	172	24	
639	35	640	Andaman & Nicobar Island	South Andaman	7975	11977	18049	1470	1358	0	...	173	122	6	

5 rows × 61 columns



	State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MAI
count	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	...	
mean	17.114062	320.500000	51222.871875	79940.576563	122372.084375	12309.098438	11942.300000	13820.946875	20778.392188	6191.807813	...	
std	9.426486	184.896367	48135.405475	73384.511114	113600.717282	11500.906881	11326.294567	14426.373130	21727.887713	9912.668948	...	
min	1.000000	1.000000	350.000000	391.000000	698.000000	56.000000	56.000000	0.000000	0.000000	0.000000	...	
25%	9.000000	160.750000	19484.000000	30228.000000	46517.750000	4733.750000	4672.250000	3466.250000	5603.250000	293.750000	...	
50%	18.000000	320.500000	35837.000000	58339.000000	87724.500000	9159.000000	8663.000000	9591.500000	13709.000000	2333.500000	...	
75%	24.000000	480.250000	68892.000000	107918.500000	164251.750000	16520.250000	15902.250000	19429.750000	29180.000000	7658.000000	...	
max	35.000000	640.000000	310450.000000	485417.000000	750392.000000	96223.000000	95129.000000	103307.000000	156429.000000	96785.000000	...	



Duplicate Values

```
census_df[census_df.duplicated()]
```

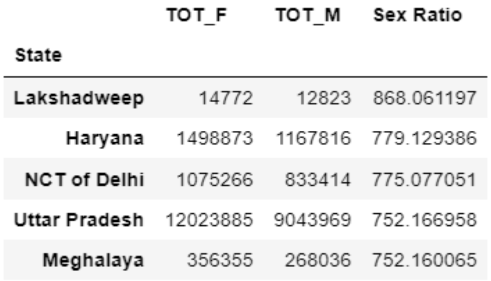
	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MAI

0 rows × 61 columns

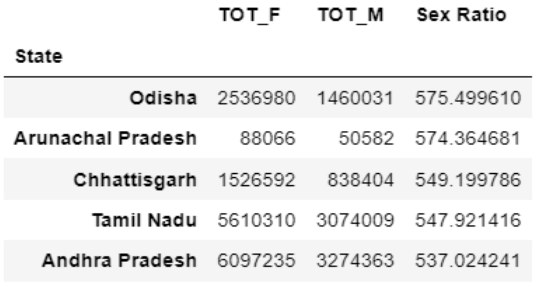
- The data consists of **640 rows** and **61 columns.**
- There is a total of **59 numeric columns** and **2 categoric columns.**
- The dataset has **no null and duplicate values.**

B. **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?**

Ans:

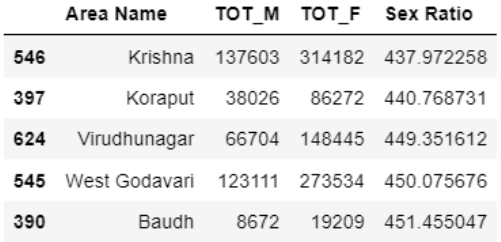


State	TOT_F	TOT_M	Sex Ratio
Lakshadweep	14772	12823	868.061197
Haryana	1498873	1167816	779.129386
NCT of Delhi	1075266	833414	775.077051
Uttar Pradesh	12023885	9043969	752.166958
Meghalaya	356355	268036	752.160065

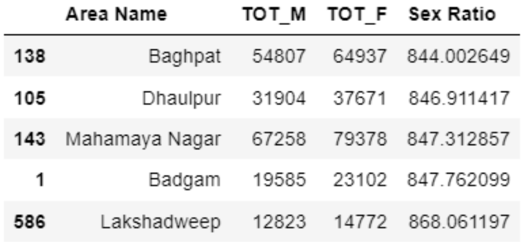


State	TOT_F	TOT_M	Sex Ratio
Odisha	2536980	1460031	575.499610
Arunachal Pradesh	88066	50582	574.364681
Chhattisgarh	1526592	838404	549.199786
Tamil Nadu	5610310	3074009	547.921416
Andhra Pradesh	6097235	3274363	537.024241

i) **Lakshadweep** has the **highest sex ratio** with **868 males for every 1000 females** while **Andhra Pradesh** has the **lowest sex ratio** with **537 males for every 1000 females.**

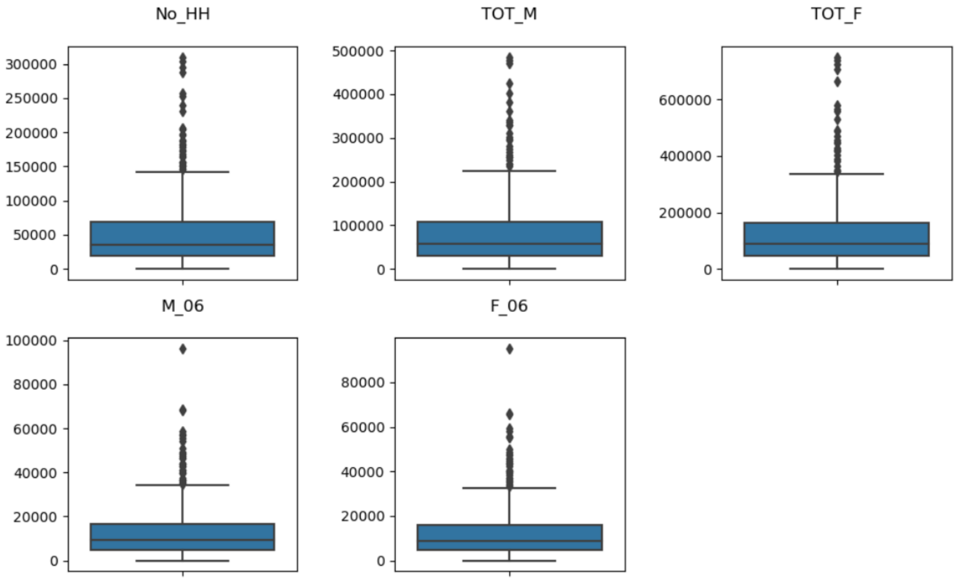


	Area Name	TOT_M	TOT_F	Sex Ratio
546	Krishna	137603	314182	437.972258
397	Koraput	38026	86272	440.768731
624	Virudhunagar	66704	148445	449.351612
545	West Godavari	123111	273534	450.075676
390	Baudh	8672	19209	451.455047



	Area Name	TOT_M	TOT_F	Sex Ratio
138	Baghpat	54807	64937	844.002649
105	Dhaulpur	31904	37671	846.911417
143	Mahamaya Nagar	67258	79378	847.312857
1	Badgam	19585	23102	847.762099
586	Lakshadweep	12823	14772	868.061197

ii) **Lakshadweep** has the **highest sex ratio** followed by the **Bagdam district** with **847 males for every 1000 females** while the **Krishna district** has the **lowest sex ratio with 437 males for every 1000 females.**



The following inferences can be gathered from the dataset:

- There is a total of **59 numeric fields** in the data
- The **average male population is 79,940** while the **average female population is 1,22,372**
- **Uttar Pradesh** has the **highest male and female populations**
- **Dadara and Nagar Havelli** has the **lowest male and female populations**
- The **male population ranges from 391 to 4,85,417** while the **female population ranges from 640 to 7,50,392**
- The **number of households ranges from 350 to 3,10,000**
- The **male population** in the **age group of 0-6 years** lies between **640 to 96,223**
- The **female population** in the **age group of 0-6 years** lies between **640 to 95,129**

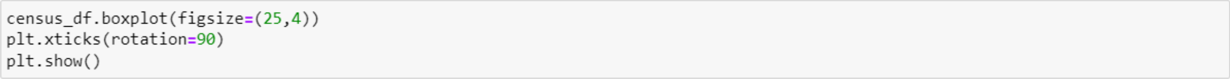
C. We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?

Ans: Outlier treatment is not necessary here as the variation in the population sizes is caused due to a wide variety of factors in the dataset. Treating the outliers may result in inaccuracy when determining the principal components using PCA as the effects of these factors would be nullified causing it to not be accounted for. Hence outlier treatment is not required here.

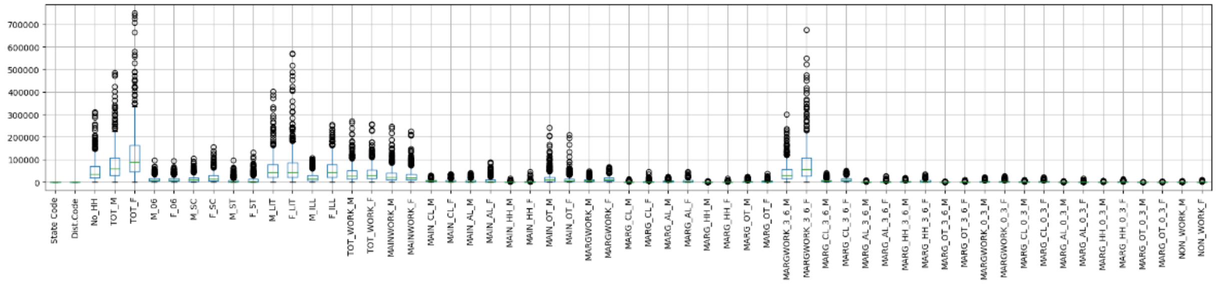
D. **Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.**

Ans:

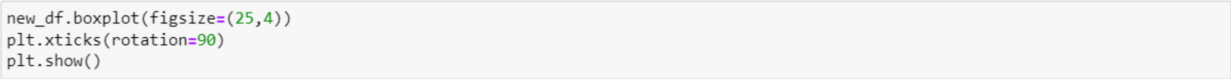
Unscaled Data



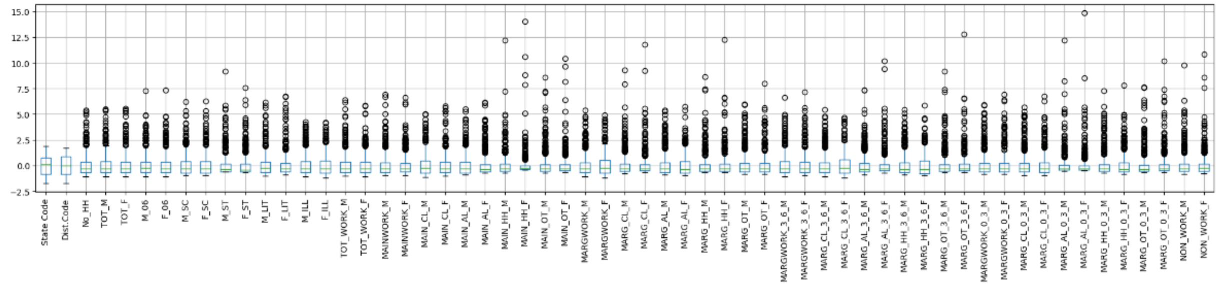
```
census_df.boxplot(figsize=(25,4))
plt.xticks(rotation=90)
plt.show()
```



Scaled data



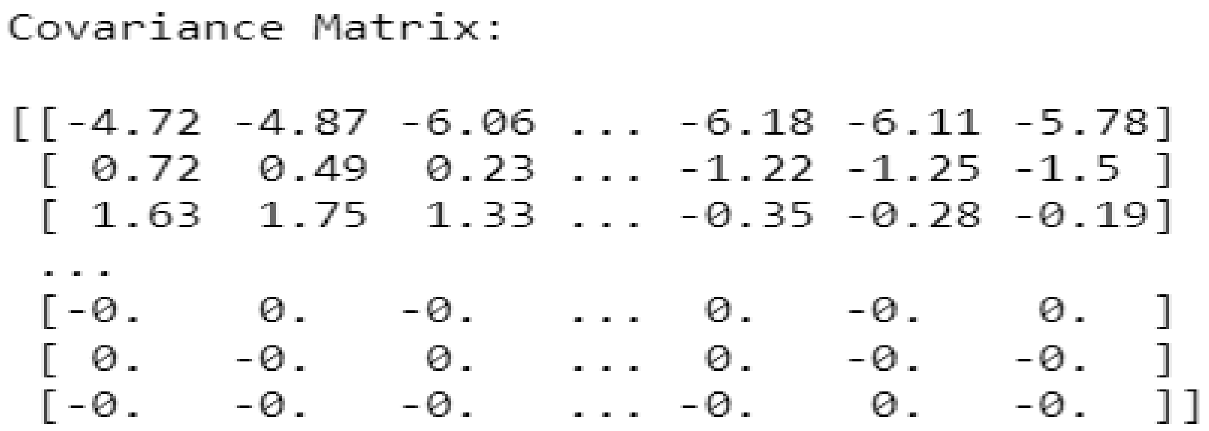
```
new_df.boxplot(figsize=(25,4))
plt.xticks(rotation=90)
plt.show()
```



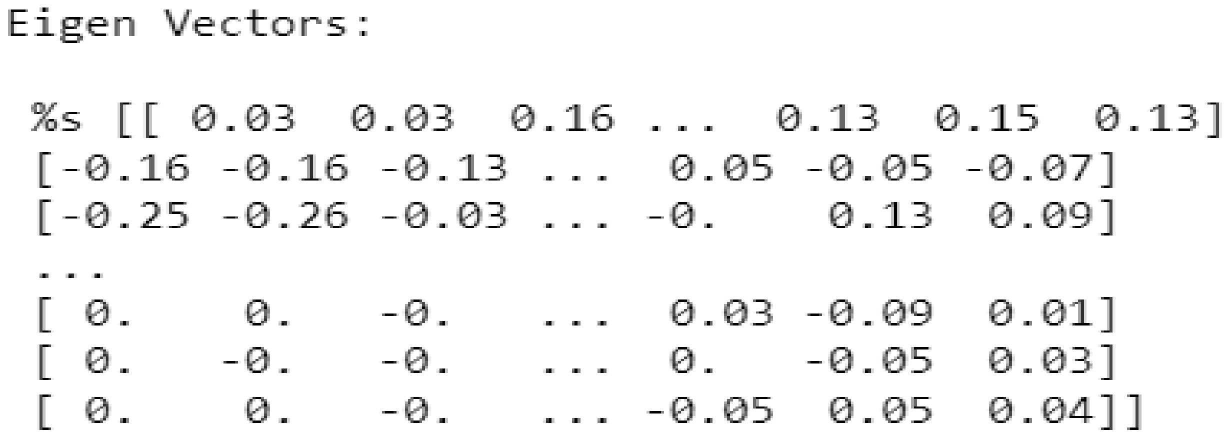
It can be observed that scaling has changed the outlier distribution for the variables. Earlier, the outlier distribution was varied for different variables not to mention the difference in their population ranges. Scaling has standardized both the outlier distribution along with the data ranges.

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

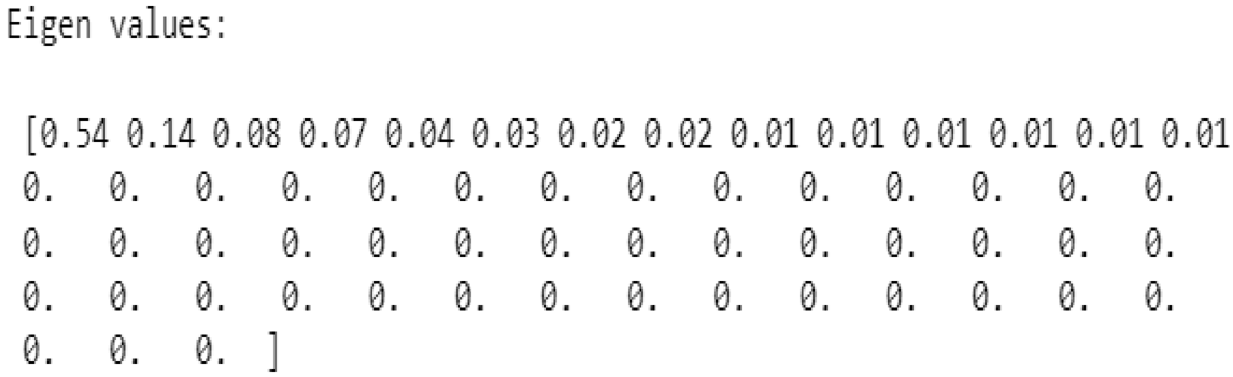
Ans:



```
Covariance Matrix:

[[-4.72 -4.87 -6.06 ... -6.18 -6.11 -5.78]
 [ 0.72  0.49  0.23 ... -1.22 -1.25 -1.5 ]
 [ 1.63  1.75  1.33 ... -0.35 -0.28 -0.19]
 ...
 [-0.    0.   -0.   ...  0.   -0.    0.  ]
 [ 0.   -0.    0.   ...  0.   -0.   -0.  ]
 [-0.   -0.   -0.   ... -0.    0.   -0.  ]]
```



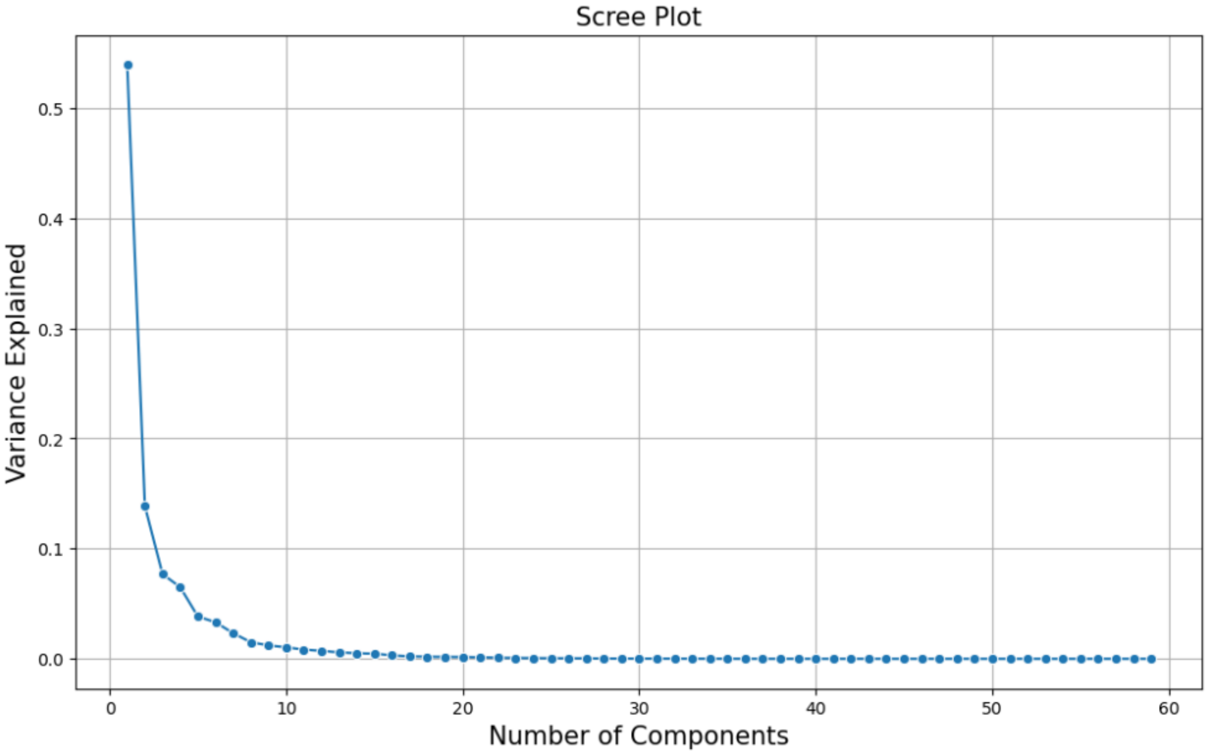
```
Eigen Vectors:

 %s [[ 0.03  0.03  0.16 ...  0.13  0.15  0.13]
 [-0.16 -0.16 -0.13 ...  0.05 -0.05 -0.07]
 [-0.25 -0.26 -0.03 ... -0.    0.13  0.09]
 ...
 [ 0.    0.   -0.   ...  0.03 -0.09  0.01]
 [ 0.   -0.   -0.   ...  0.   -0.05  0.03]
 [ 0.    0.   -0.   ... -0.05  0.05  0.04]]
```



```
Eigen values:

 [0.54 0.14 0.08 0.07 0.04 0.03 0.02 0.02 0.01 0.01 0.01 0.01 0.01 0.01
 0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.
 0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.
 0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.
 0.   0.   0.  ]
```

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

Ans:



The number of components can be decided upon the explained variance. It can be observed from the cumulative variance values and from the scree plot that at least 90% of the explained variance is captured by having 7 principal components.