Machine Learning 1 Project - Business Report

By: Aaryani Kadiyala

PGP-Data Science and Business Analytics

(PGPDSBA.O.JAN24.A)

Table of Contents

1. Problem 1 ---------- 4
2. Problem 2 ---------- 18

List of Tables:

Table 1: Description of all the chosen variables ---------- 21

List of Figures:

Figure 1: First 5 rows of the data frame ---------- 4
Figure 2: Shape of the data frame ---------- 4
Figure 3: Data types of data frame ---------- 5
Figure 4: Duplicates of the data frame ---------- 5
Figure 5: Statistical summary of the data frame ---------- 5
Figure 6: Bar plot for all the columns ---------- 7
Figure 7: Total no. of Clicks, Total Revenue & Ratio of revenue and Clicks for each platform ---------- 7
Figure 8: Total no. of Clicks, Total Revenue & Ratio of revenue and Clicks for each device type ---------- 8
Figure 9: Total no. of Clicks, Total Revenue & Ratio of revenue and Clicks for each format ---------- 8
Figure 10: Ratio of Clicks to total no. of clicks ---------- 8
Figure 12: Barplot with Ad Type & Platform ---------- 8
Figure 13: Correlation Matrix of the data frame ---------- 9
Figure 14: Data frame before treating Null values ---------- 10
Figure 15: Data frame after treating Null values ---------- 10
Figure 16: Box plot before treating outliers ---------- 10
Figure 17: Box plot after treating outliers ---------- 11
Figure 18: Data before Z-score Scaling ---------- 11
Figure 19: Data after Z-score Scaling ---------- 12
Figure 20: Dendogram of complete dataset ---------- 12
Figure 21: Dendogram using Ward linkage and Euclidean distance ---------- 13
Figure 22: Dataset created with 3 clusters ---------- 13
Figure 23: Elbow plot upto n=10 ---------- 14
Figure 24: Silhouette scores for up to 10 clusters ---------- 14

Figure 25: Total no. of point in each KMeans cluster ---------- 15
Figure 26: KMeans Cluster wise frequency ---------- 15
Figure 27: Mean & Median values of all columns ---------- 16
Figure 28: First 5 rows of the Dataset ---------- 19
Figure 29: Shape of the Dataset ---------- 19
Figure 30: Data types of the complete data set ---------- 20
Figure 31: Statistical summary of the data set ---------- 20
Figure 32: Barplot between State & TOT_M ---------- 21
Figure 33: Barplot between State & TOT_F ---------- 21
Figure 34: Barplot between State & M_LIT ---------- 22
Figure 35: Barplot between State & F_LIT ---------- 22
Figure 36: Barplot between State & TOT_WORK_M ---------- 22
Figure 37: Barplot between State & TOT_WORK_F ---------- 23
Figure 38: Boxplot before treating outliers using IQR method ---------- 23
Figure 39: Boxplot after treating outliers using IQR Method ---------- 24
Figure 40: Box plot before scaling the data set with z-score method ---------- 25
Figure 41: Statistical summary before scaling the data set with z-score method ---------- 25
Figure 42: Box plot after scaling the data set with z-score method ---------- 26
Figure 43: Statistical summary after scaling the data set with z-score method ---------- 26
Figure 44: Covariance Matrix of the dataset ---------- 27
Figure 45: Eigen Vector of the dataset ---------- 27
Figure 46: Eigen Values of the dataset ---------- 28
Figure 47: Scree Plot of the dataset ---------- 28
Figure 48: Comparison of PCs with Actual Variables ---------- 29

Problem 1: Clustering

Digital Ads Data:

The ads 24x7 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.

1.1 Problem definition - Check shape, Data types, statistical summary - Univariate analysis - Bivariate analysis - Key meaningful observations on individual variables and the relationship between variables

Step 1: Loaded the Clustering Clean ads data excel file.

]:

	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

Figure 1: First 5 rows of the data frame

Step 2: Checking the dataset shape, Data types & Statistical Summary

```
(23066, 19)    There are 23066 rows & 19 columns in the dataset
```

Figure 2: Shape of the data frame

```
<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
```

Figure 3: Data types of data frame

```
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                  0.0
Fee                    0.0
Revenue                0.0
CTR                    0.0
CPM                    0.0
CPC                    0.0
dtype: object
```

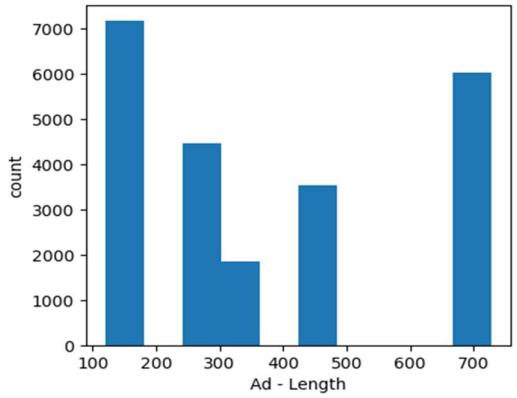
Figure 4: Duplicates of the data frame

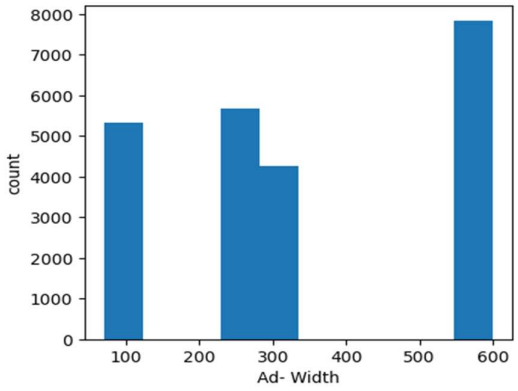
There are no duplicates of the data frame.

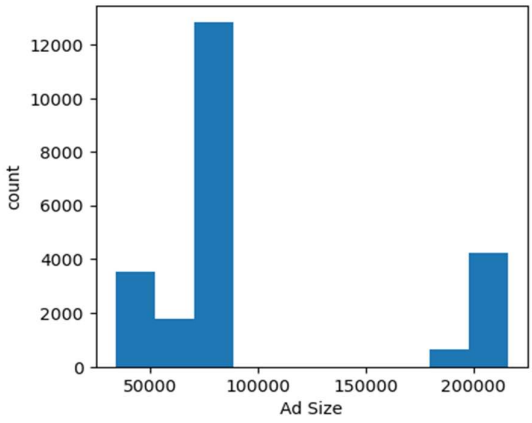
	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	3.851631e+02	2.336514e+02	120.0000	120.000000	300.00000	7.200000e+02	728.00
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.0000	250.000000	300.00000	6.000000e+02	600.00
Ad Size	23066.0	9.667447e+04	6.153833e+04	33600.0000	72000.000000	72000.00000	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	3.198322e-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

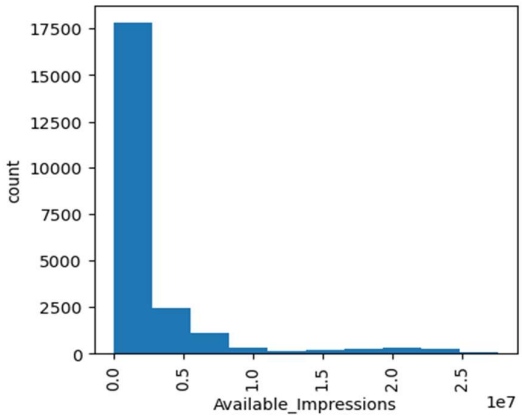
Figure 5: Statistical summary of the data frame

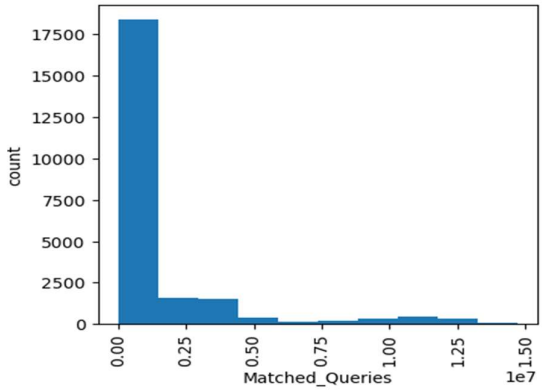
Step 3: Univariant Analysis

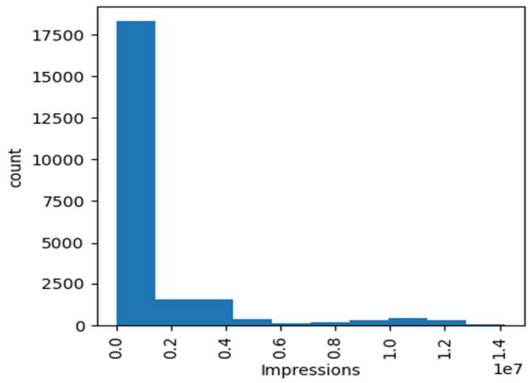


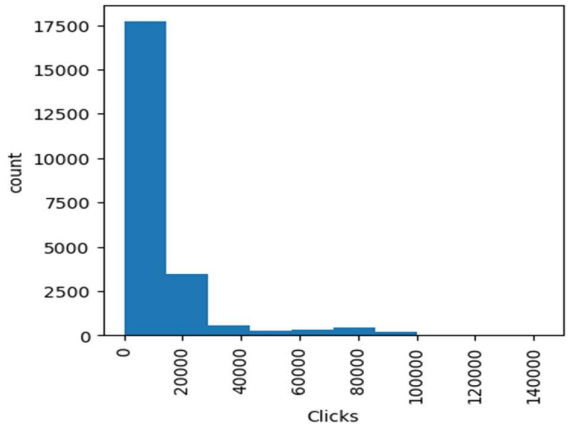


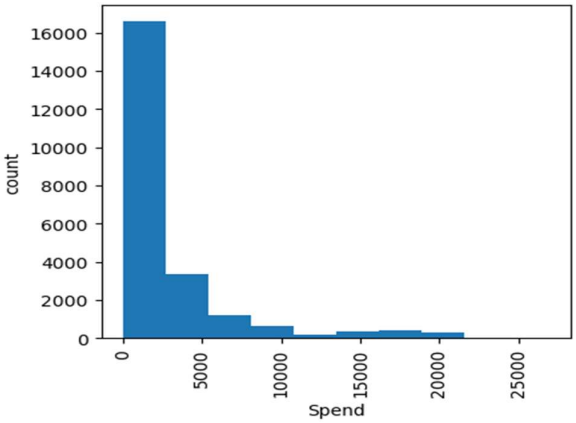












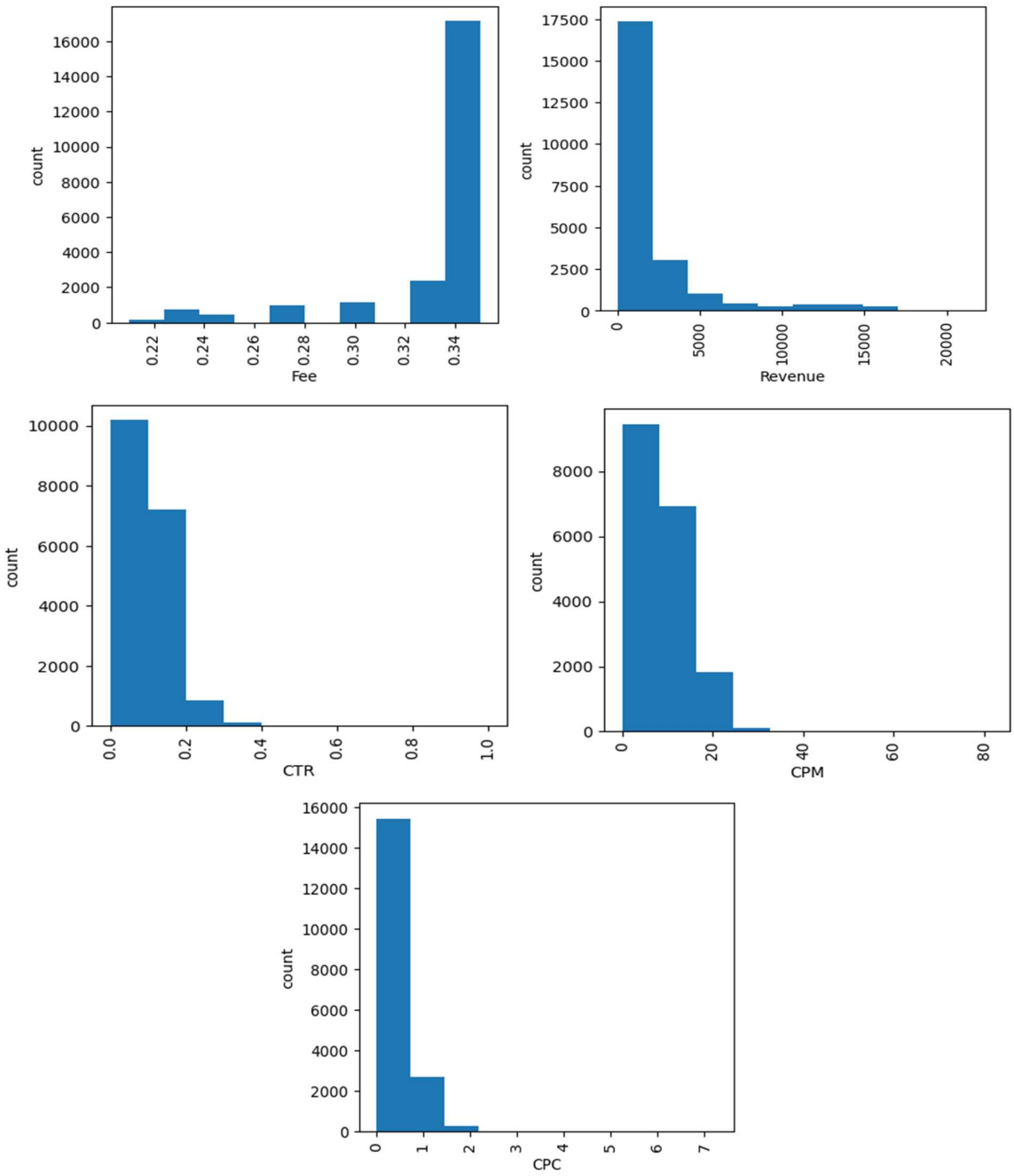


Figure 6: Bar plot for all the columns

Step 4: Bivariant Analysis

```
Platform
App       53110133
Video    106162400
Web       87038182
Name: Clicks, dtype: int64
```

```
Platform
App        9506077.0
Video     19058208.0
Web       15820519.0
Name: Revenue, dtype: float64
```

```
Platform
App      17.898801
Video    17.951938
Web      18.176527
dtype: float64
```

Figure 7: Total no. of Clicks, Total Revenue & Ratio of revenue and Clicks for each platform

```
Device Type
Desktop     89065672
Mobile     157245043
Name: Clicks, dtype: int64
```

```
Device Type
Desktop    15963612.0
Mobile     28421192.0
Name: Revenue, dtype: float64
```

```
Device Type
Desktop    17.923417
Mobile     18.074460
dtype: float64
```

Figure 8: Total no. of Clicks, Total Revenue & Ratio of revenue and Clicks for each device type

```
Format
Display    124114357
Video      122196358
Name: Clicks, dtype: int64
```

```
Format
Display    22229981.0
Video      22154824.0
Name: Revenue, dtype: float64
```

```
Format
Display    17.910886
Video      18.130511
dtype: float64
```

Figure 9: Total no. of Clicks, Total Revenue & Ratio of revenue and Clicks for each format

```
Platform
App      21.562250
Video    43.101008
Web      35.336742
Name: Clicks, dtype: float64
```

Figure 10: Ratio of Clicks to total no. of clicks

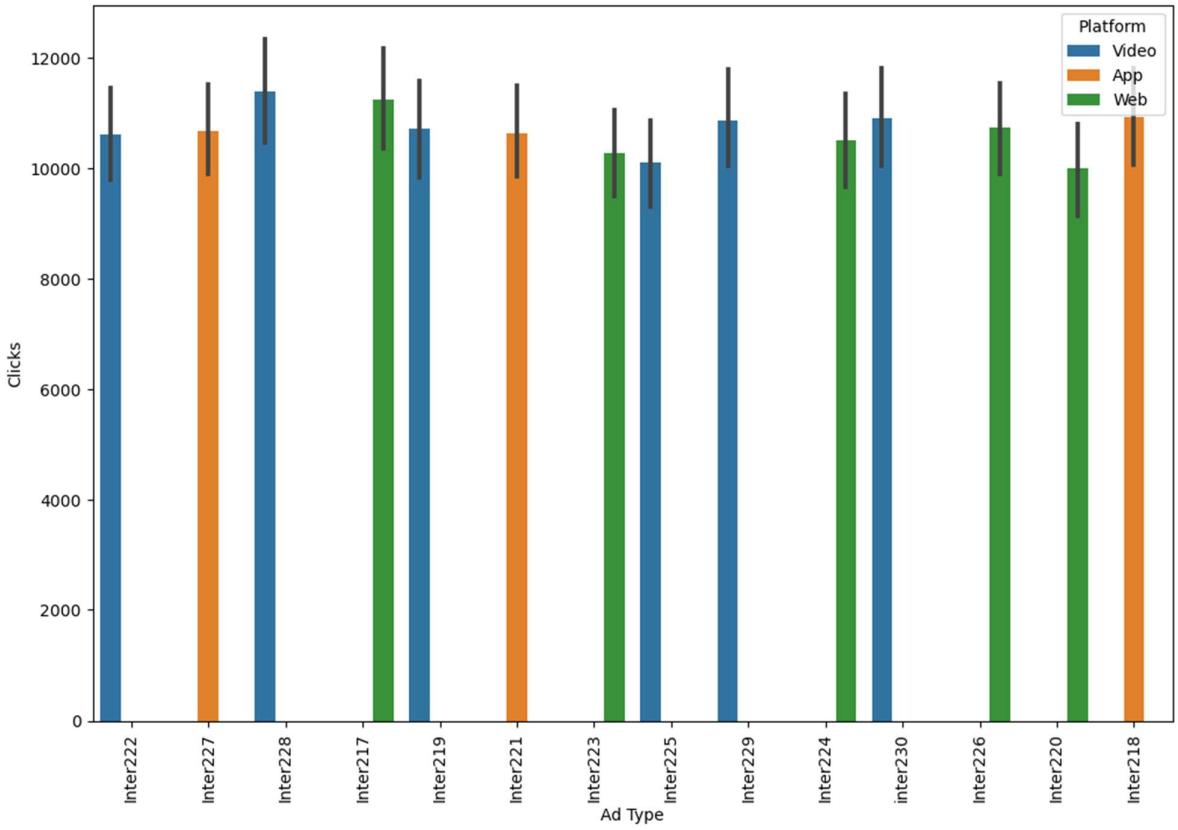


Figure 12: Barplot with Ad Type & Platform

Figure 13: Correlation Matrix of the data frame

Step 5: Key Observations

- As observed earlier, most of the variables have skewed distributions.
- The distribution for the Ad size, length & width are relatively less skewed with fewer to no outliers.
- The Fee is the only variable which is skewed to the left meaning most of the companies have been successful by achieved high revenue.
- The distribution for all other variables is highly skewed to the right. All these variables have outliers to the right end.
- There is a strong positive correlation between CTR and CPM.
- The exposure of the advertisement is positively correlated with CTR. This indicates that ad is generating lot of clicks.
- There is a strong negative correlation between Fee & spend and Fee & Revenue.

1.2 Missing value check and treatment - Outlier Treatment - z-score scaling Note: Treat missing values in CPC, CTR and CPM using the formula given.

Step 1 : Checking for Null Values & treating the null values with given formula

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	4736
CPM	4736
CPC	4736
dtype: int64	

Figure 14: Data frame before treating Null values

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	

Figure 15: Data frame after treating Null values

Step 2 : Outlier Treatment

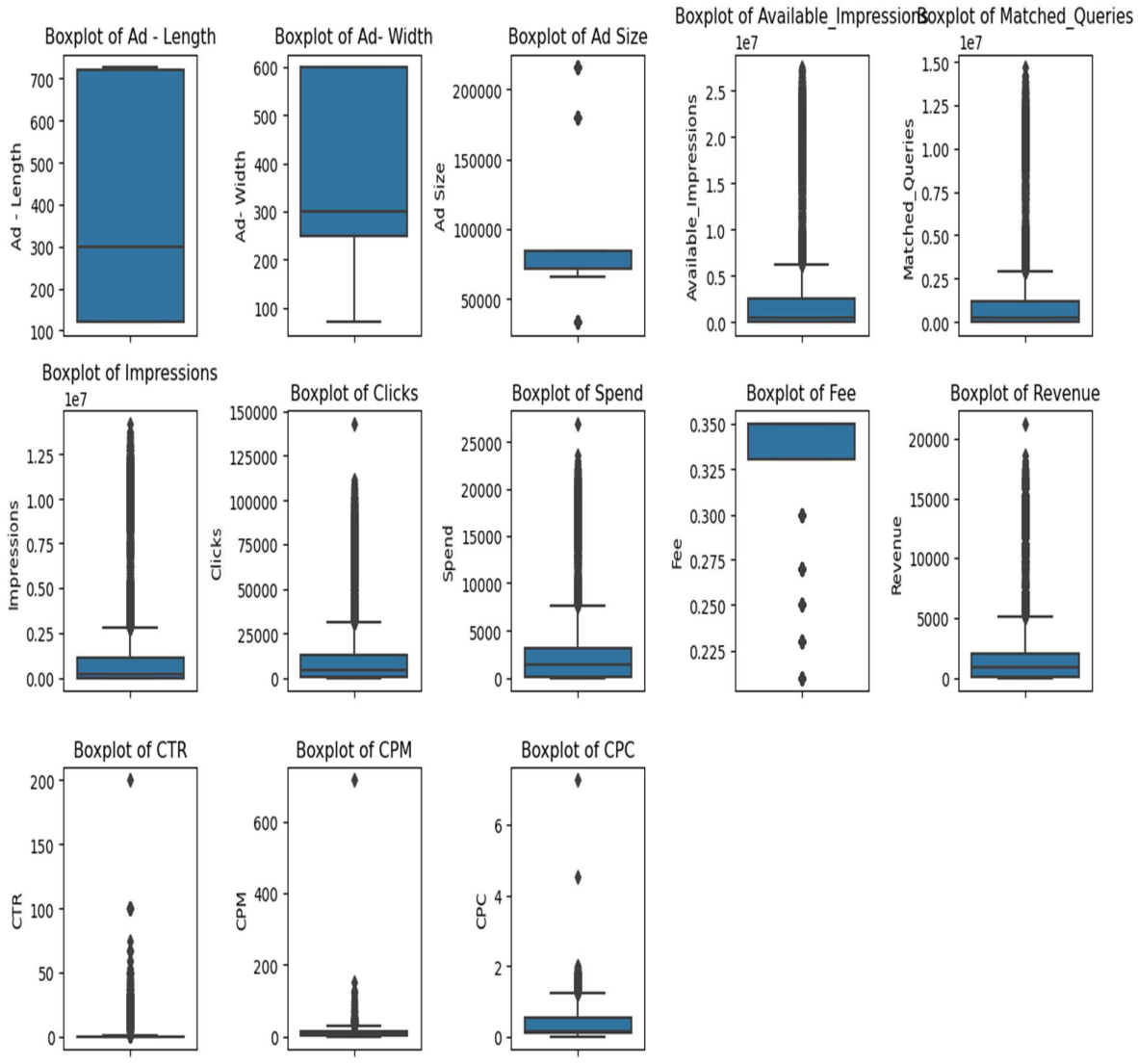


Figure 16: Box plot before treating outliers

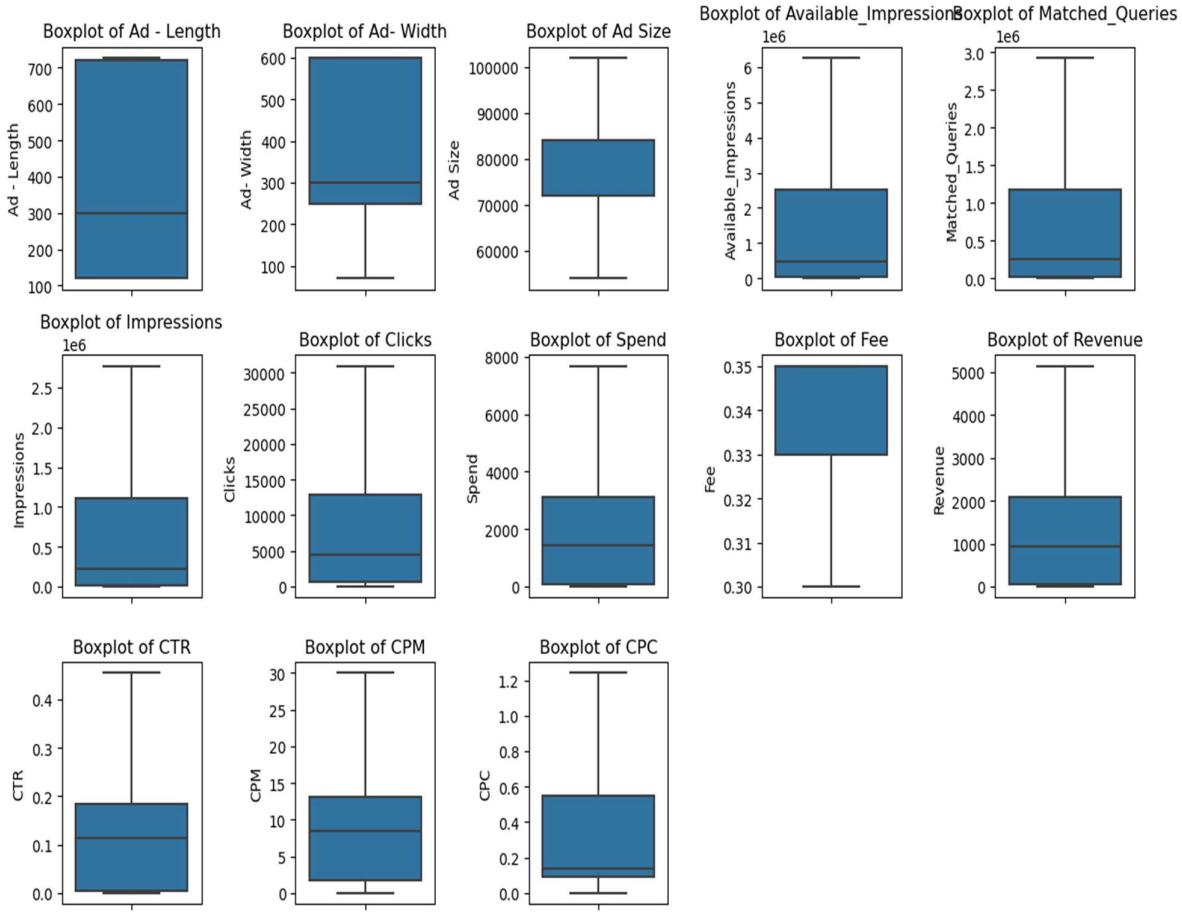


Figure 17: Box plot after treating outliers

Step 3: Z-score Scaling Note

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	3.851631e+02	2.336514e+02	120.000000	120.000000	300.000000	7.200000e+02	7.280000e+02
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.000000	250.000000	300.000000	6.000000e+02	6.000000e+02
Ad Size	23066.0	7.601188e+04	5.252055e+03	72000.000000	72000.000000	72000.000000	8.400000e+04	8.400000e+04
Available_Impressions	23066.0	1.012021e+06	1.046469e+06	33672.250000	33672.437500	483771.000000	2.527583e+06	2.527712e+06
Matched_Queries	23066.0	4.961223e+05	4.846397e+05	18282.500000	18285.875000	258087.500000	1.180566e+06	1.180700e+06
Impressions	23066.0	4.588169e+05	4.607521e+05	7990.500000	7990.875000	225290.000000	1.112413e+06	1.112428e+06
Clicks	23066.0	5.836562e+03	4.910674e+03	710.000000	710.000000	4425.000000	1.279281e+04	1.279375e+04
Spend	23066.0	3.121400e+03	5.752679e-10	3121.400000	3121.400000	3121.400000	3.121400e+03	3.121400e+03
Fee	23066.0	3.448626e-01	8.738349e-03	0.330000	0.330000	0.350000	3.500000e-01	3.500000e-01
Revenue	23066.0	9.842672e+02	8.110098e+02	55.365375	55.369031	926.335000	2.091067e+03	2.091338e+03
CTR	23066.0	9.356938e-02	7.508412e-02	0.003400	0.003400	0.112650	1.837610e-01	1.837777e-01
CPM	23066.0	7.178427e+00	4.981789e+00	1.750000	1.750000	8.370742	1.304000e+01	1.304000e+01
CPC	23066.0	2.711942e-01	1.988347e-01	0.090000	0.090000	0.140000	5.500000e-01	5.500000e-01

Figure 18: Data before Z-score Scaling

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	1.281478e-16	1.000022	-1.134891e+00	-1.134891e+00	-3.644957e-01	1.433093e+00	1.467332e+00
Ad- Width	23066.0	-1.182903e-16	1.000022	-1.319110e+00	-4.327968e-01	-1.865987e-01	1.290590e+00	1.290590e+00
Ad Size	23066.0	5.520214e-16	1.000022	-7.638850e-01	-7.638850e-01	-7.638850e-01	1.520985e+00	1.520985e+00
Available_Impressions	23066.0	-7.886020e-17	1.000022	-9.349252e-01	-9.349250e-01	-5.048040e-01	1.448293e+00	1.448417e+00
Matched_Queries	23066.0	-6.900268e-17	1.000022	-9.859905e-01	-9.859835e-01	-4.911689e-01	1.412305e+00	1.412580e+00
Impressions	23066.0	-1.084328e-16	1.000022	-9.784789e-01	-9.784781e-01	-5.068495e-01	1.418573e+00	1.418606e+00
Clicks	23066.0	8.871773e-17	1.000022	-1.043986e+00	-1.043986e+00	-2.874539e-01	1.416588e+00	1.416779e+00
Spend	23066.0	-9.094947e-13	0.000000	-9.094947e-13	-9.094947e-13	-9.094947e-13	-9.094947e-13	-9.094947e-13
Fee	23066.0	3.233268e-15	1.000022	-1.700881e+00	-1.700881e+00	5.879306e-01	5.879306e-01	5.879306e-01
Revenue	23066.0	9.857525e-18	1.000022	-1.145389e+00	-1.145385e+00	-7.143378e-02	1.364747e+00	1.365082e+00
CTR	23066.0	-7.886020e-17	1.000022	-1.200938e+00	-1.200938e+00	2.541288e-01	1.201234e+00	1.201457e+00
CPM	23066.0	-7.886020e-17	1.000022	-1.089678e+00	-1.089678e+00	2.393400e-01	1.176626e+00	1.176626e+00
CPC	23066.0	1.626492e-16	1.000022	-9.113004e-01	-9.113004e-01	-6.598297e-01	1.402229e+00	1.402229e+00

Figure 19: Data after Z-score Scaling

Scaling can increase the computational complexity of algorithms, as it involves additional computations to transform the data.

1.3 Hierarchical Clustering - Construct a dendrogram using Ward linkage and Euclidean distance - Identify the optimum number of Clusters

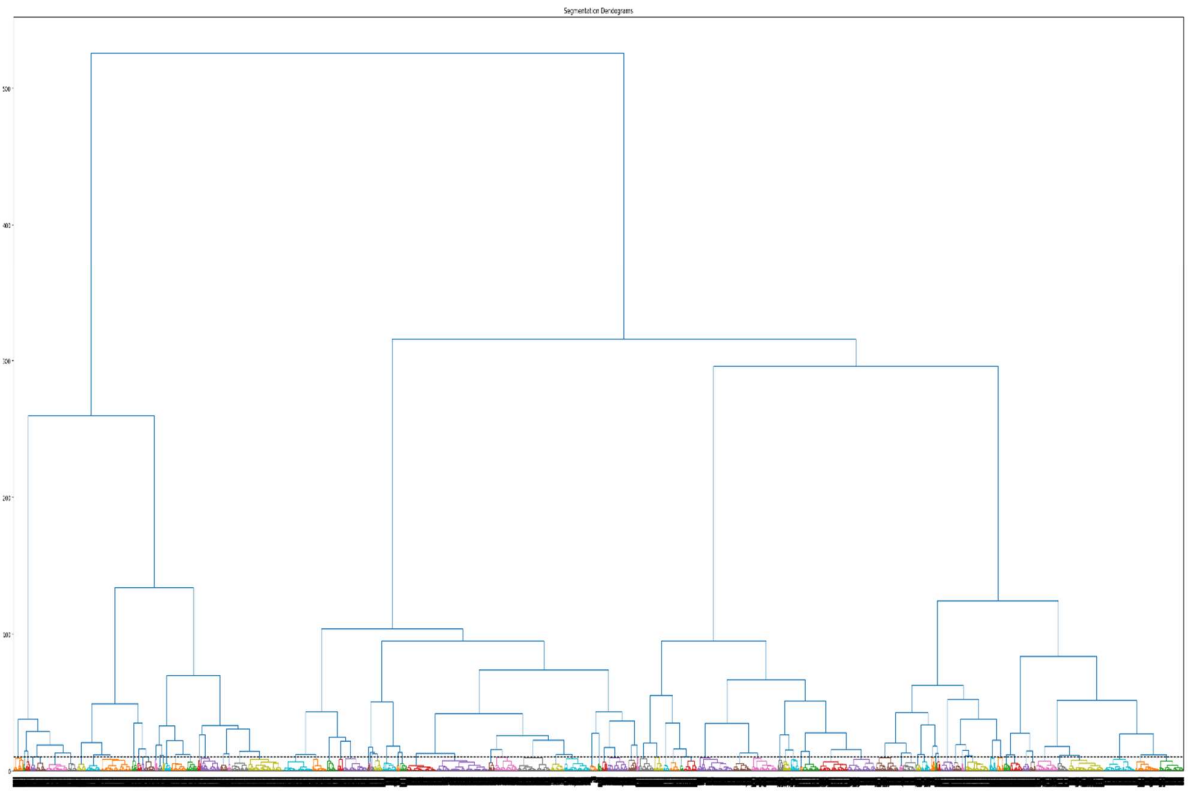


Figure 20: Dendogram of complete dataset

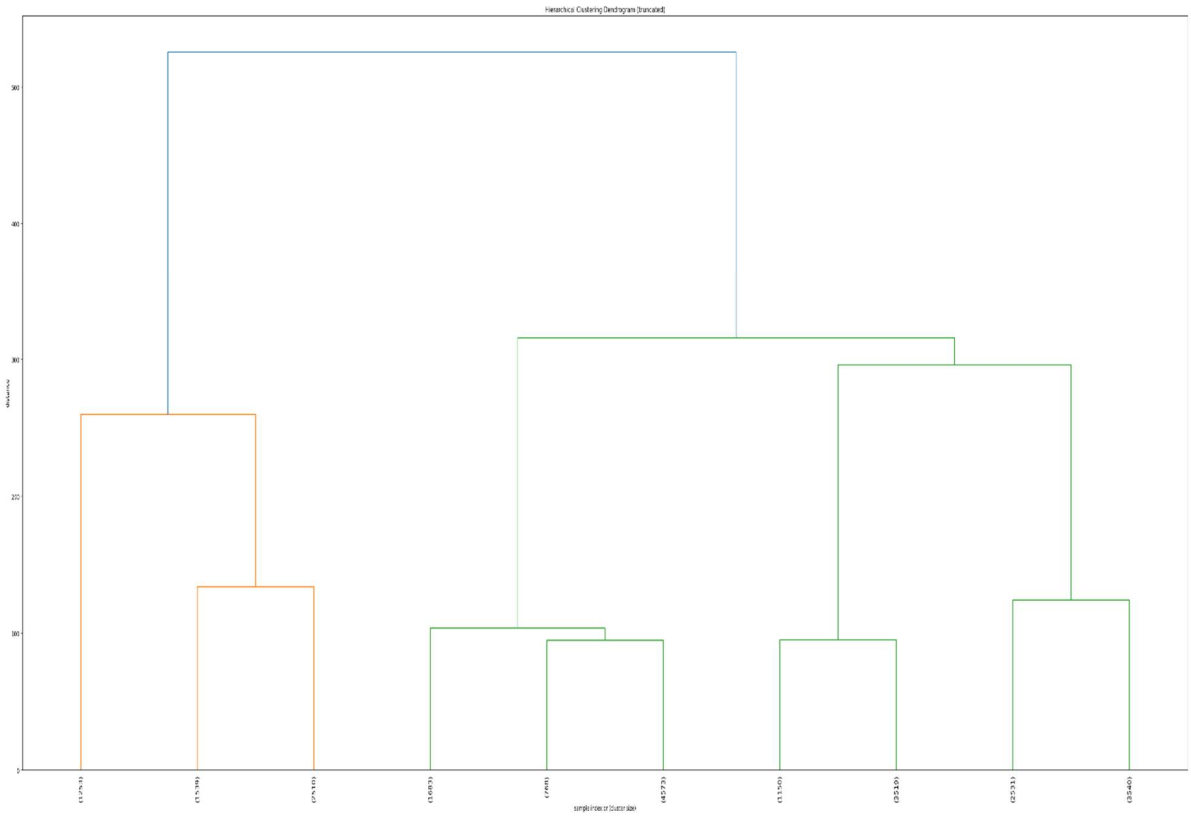


Figure 21: Dendogram using Ward linkage and Euclidean distance

According to above dendograms, we are able to generate with more than 2 clusters, better it is for the business. Hence let's consider 3 clusters and plot the clusters to confirm if the derived clusters are providing the required segmentation details.

	Ad - Length	Ad-Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	cluster_1	cluster_2
0	300	250	75000.0	33672.25	18282.5	7990.5	710.0	3121.4	0.35	55.365375	0.0034	1.75	0.09	3	44
1	300	250	75000.0	33672.25	18282.5	7990.5	710.0	3121.4	0.35	55.365375	0.0035	1.75	0.09	3	44
2	300	250	75000.0	33672.25	18282.5	7990.5	710.0	3121.4	0.35	55.365375	0.0034	1.75	0.09	3	44
3	300	250	75000.0	33672.25	18282.5	7990.5	710.0	3121.4	0.35	55.365375	0.0034	1.75	0.09	3	44
4	300	250	75000.0	33672.25	18282.5	7990.5	710.0	3121.4	0.35	55.365375	0.0041	1.75	0.09	3	44

Figure 22: Dataset created with 3 clusters

1.4 K-means Clustering - Apply K-means Clustering - Plot the Elbow curve - Check Silhouette Scores - figure out the appropriate number of clusters - Cluster Profiling

For checking the Optimal number of clusters, we use WSS (Within Sum of Square)

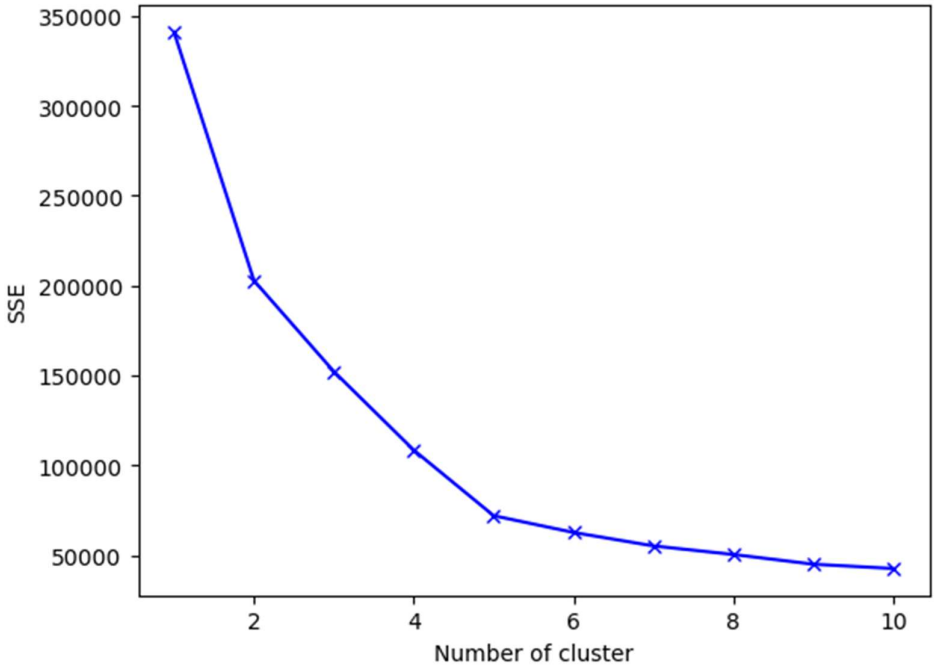


Figure 23: Elbow plot upto n=10

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.

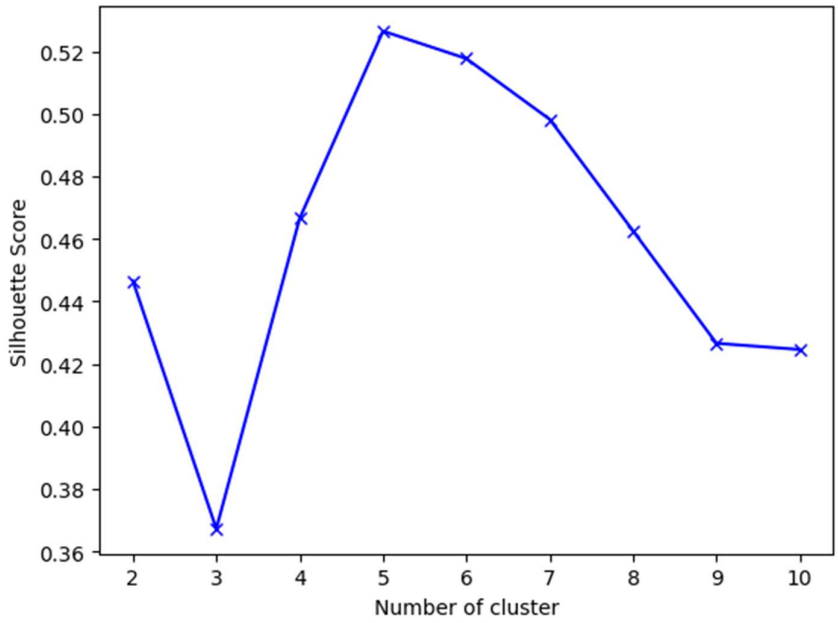


Figure 24: Silhouette scores for up to 10 clusters

Silhouette Score for k=2: 0.446
Silhouette Score for k=3: 0.367
Silhouette Score for k=4: 0.466
Silhouette Score for k=5: 0.526
Silhouette Score for k=6: 0.517
Silhouette Score for k=7: 0.498
Silhouette Score for k=8: 0.465
Silhouette Score for k=9: 0.426
Silhouette Score for k=10: 0.424
Optimal number of clusters: 5

I 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 other clusters, 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.

Cluster Profiling:

```
KMeans_Labels
0    6140
1    4698
2    6640
3    4049
4    1539
Name: count, dtype: int64
```

Figure 25: Total no. of point in each KMeans cluster

KMeans_Labels	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenu
0	424.471010	146.286645	53525.081433	1.838272e+06	8.784161e+05	8.398711e+05	3304.398860	1524.047645	0.349235	993.0950
1	682.020434	305.246914	206058.748404	2.626464e+05	1.416907e+05	1.207011e+05	14363.409110	1254.130773	0.349544	816.7198
2	146.024096	568.373494	77139.759036	3.648993e+04	2.181334e+04	1.566770e+04	1888.464759	210.052837	0.349991	136.5621
3	465.880958	199.212151	75205.058039	1.039627e+07	5.630305e+06	5.451651e+06	11253.998024	8653.044280	0.290385	6378.6766
4	141.543860	572.482131	75680.311891	8.055940e+05	5.663903e+05	4.777502e+05	65260.276803	6985.407472	0.288356	5013.7854

Figure 26: KMeans Cluster wise frequency

- After performing KMeans Clustering on scaled data, and then added the predicted cluster labels to two different data sets: data_scaled_copy and df.
- The KMeans function from scikit-learn is used to create a KMeans object with n_clusters=5 (i.e., 5 clusters).
- Created clusters for the Ads based on optimum number of clusters using silhouette score.

- The groupby method from Pandas is used to group the data by the KMeans cluster labels, and the mean and median methods are used to compute the mean and median values of each feature within each cluster. The resulting data frames are stored in the variables mean and median.

	group_0 Mean	group_1 Mean	group_2 Mean	group_3 Mean	group_4 Mean	group_0 Median	group_1 Median	group_2 Median	group_3 Median	grou Me
Ad - Length	4.244710e+02	682.020434	146.024096	4.658810e+02	141.543860	4.800000e+02	720.0000	120.0000	3.000000e+02	120.0
Ad- Width	1.462866e+02	305.246914	568.373494	1.992122e+02	572.482131	7.000000e+01	300.0000	600.0000	2.500000e+02	600.0
Ad Size	5.352508e+04	206058.748404	77139.759036	7.520506e+04	75680.311891	3.360000e+04	216000.0000	72000.0000	7.500000e+04	72000.0
Available_Impressions	1.838272e+06	262646.369519	36489.929217	1.039627e+07	805593.964263	1.863021e+06	214679.5000	13633.5000	7.055688e+06	828944.0
Matched_Queries	8.784161e+05	141690.728608	21813.339006	5.630305e+06	566390.274854	8.722895e+05	138272.0000	8150.0000	3.878848e+06	582970.0
Impressions	8.398711e+05	120701.134951	15667.703916	5.451651e+06	477750.160494	8.263720e+05	116475.5000	3578.0000	3.797137e+06	490231.0
Clicks	3.304399e+03	14363.409110	1888.464759	1.125400e+04	65260.276803	3.320000e+03	14625.5000	449.5000	8.931000e+03	68243.0
Spend	1.524048e+03	1254.130773	210.052837	8.653044e+03	6985.407472	1.561280e+03	1340.8300	46.7500	5.231320e+03	7172.1
Fee	3.492345e-01	0.349544	0.349991	2.903853e-01	0.288356	3.500000e-01	0.3500	0.3500	3.000000e-01	0.2
Revenue	9.930951e+02	816.719858	136.562174	6.378677e+03	5013.785448	1.014834e+03	871.5390	30.3840	3.661920e+03	5235.6
CTR	4.051650e-03	0.124560	0.144317	2.173004e-03	0.137939	4.000000e-03	0.1232	0.1282	2.200000e-03	0.1
CPM	1.788734e+00	11.185018	13.708249	1.567269e+00	15.184727	1.810000e+00	10.9000	12.9400	1.560000e+00	14.6
CPC	5.450473e-01	0.092146	0.109895	7.566183e-01	0.110077	4.600000e-01	0.0900	0.1000	7.100000e-01	0.1
KMeans_Lables	0.000000e+00	1.000000	2.000000	3.000000e+00	4.000000	0.000000e+00	1.0000	2.0000	3.000000e+00	4.0

Figure 27: Mean & Median values of all columns

1.4 **Actionable Insights & Recommendations** - Extract meaningful insights (atleast 3) from the clusters to identify the most effective types of ads, target audiences, or marketing strategies that can be inferred from each segment. - Based on the clustering analysis and key insights, provide actionable recommendations (atleast 3) to Ads 24x7 on how to optimize their digital marketing efforts, allocate budgets efficiently, and tailor ad content to specific audience segments.

Ans:

- There are 23066 rows, and 19 columns into the Dataset.
- There are no duplicate values in data frame.
- There are 4736 Null values in CTR, CPM, and CPC Columns. I 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.
- I have applied z-score method on the data frame for scaling.
- I 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.
- I have created 5 clusters for the Dataset.

Conclusion after Clustering:

Insights:

- When Click on Ads gets increases then Revenue is also increases.

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

Actionable Recommendations:

- The revenue can be improved by getting more no. of clicks in Apps because compared to video & web revenue w.r.t no. of clicks is more.
- The Ad campaign or Ad concept must be changed in the vide & web platforms in order increase the no. of clicks.
- The Impressions on all the platforms should be increased in order to get more no. of clicks.

Project 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, 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

2.1. Define the problem and perform Exploratory Data Analysis- Problem Definition – Check shape, Data types, statistical summary - Perform an EDA on the data to extract useful insights Note: 1. 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 2. Example questions to answer from EDA - (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio?

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

Figure 28: First 5 rows of the Dataset

```
(640, 61)
```

There are 640 rows & 61 columns in the dataset

Figure 29: Shape of the Dataset

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

Figure 30: Data types of the complete data set

There are no duplicates in the data set.

	State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_
count	640.00	640.00	640.00	640.00	640.00	640.00	640.00	640.00	640.00	640.00	...	640.00	640.00	
mean	17.11	320.50	51222.87	79940.58	122372.08	12309.10	11942.30	13820.95	20778.39	6191.81	...	1392.97	2757.05	
std	9.43	184.90	48135.41	73384.51	113600.72	11500.91	11326.29	14426.37	21727.89	9912.67	...	1489.71	2788.78	
min	1.00	1.00	350.00	391.00	698.00	56.00	56.00	0.00	0.00	0.00	...	4.00	30.00	
25%	9.00	160.75	19484.00	30228.00	46517.75	4733.75	4672.25	3466.25	5603.25	293.75	...	489.50	957.25	
50%	18.00	320.50	35837.00	58339.00	87724.50	9159.00	8663.00	9591.50	13709.00	2333.50	...	949.00	1928.00	
75%	24.00	480.25	68892.00	107918.50	164251.75	16520.25	15902.25	19429.75	29180.00	7658.00	...	1714.00	3599.75	
max	35.00	640.00	310450.00	485417.00	750392.00	96223.00	95129.00	103307.00	156429.00	96785.00	...	9875.00	21611.00	

8 rows × 59 columns

Figure 31: Statistical summary of the data set

5 Variables chosen are 'TOT_M', 'TOT_F','M_LIT','F_LIT', 'TOT_WORK_M' & 'TOT_WORK_F'. And comparing those 5 variables against 'State' and 'Area Name'.

Name	Description
State	Name of the State
Area Name	Name of the Area
TOT_M	Total population Male
TOT_F	Total population Female
M_LIT	Literates' population Male
F_LIT	Literates' population Female
TOT_WORK_M	Total Worker Population Male
TOT_WORK_F	Total Worker Population Female

Table 1: Description of all the chosen variables

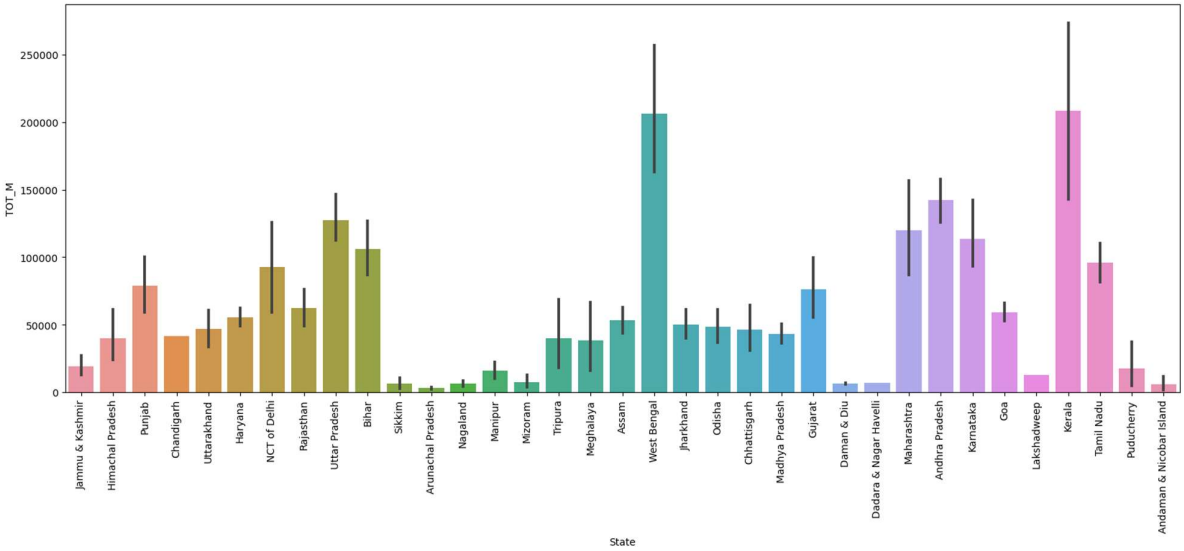


Figure 32: Barplot between State & TOT_M

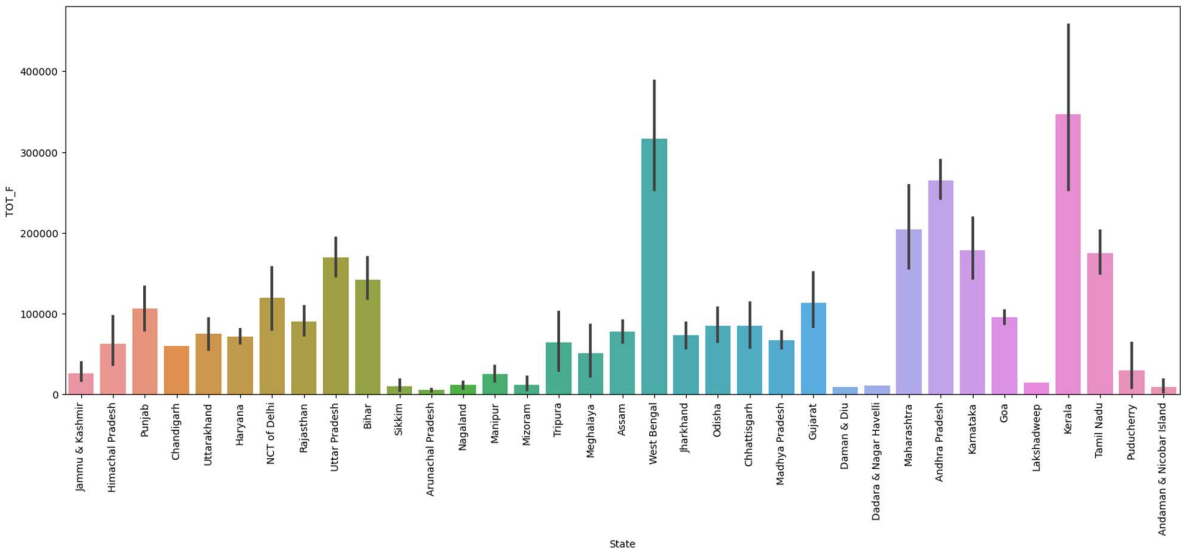


Figure 33: Barplot between State & TOT_F

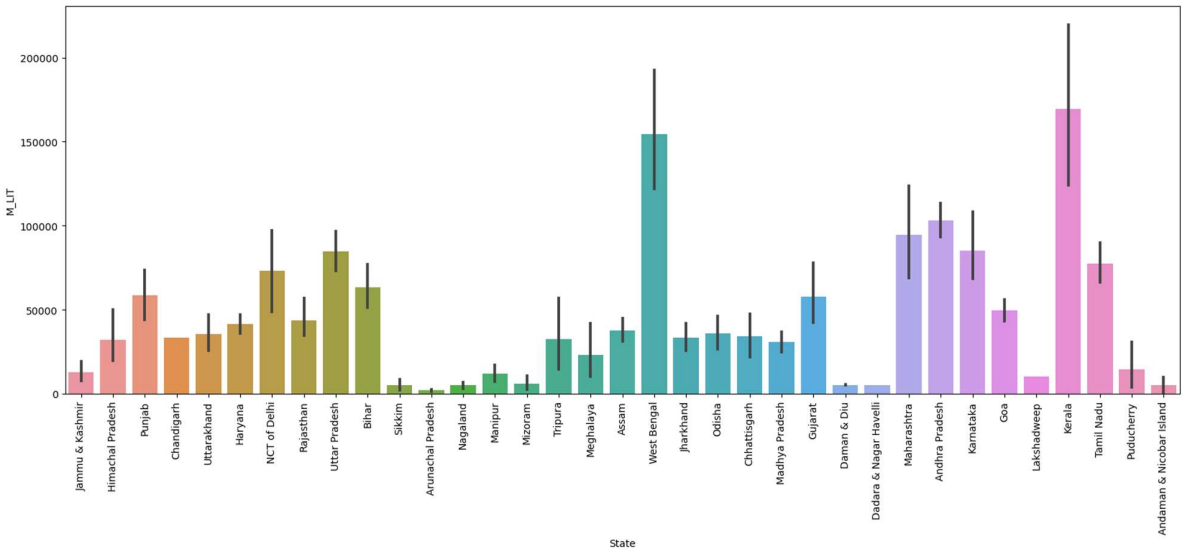


Figure 34: Barplot between State & M_LIT

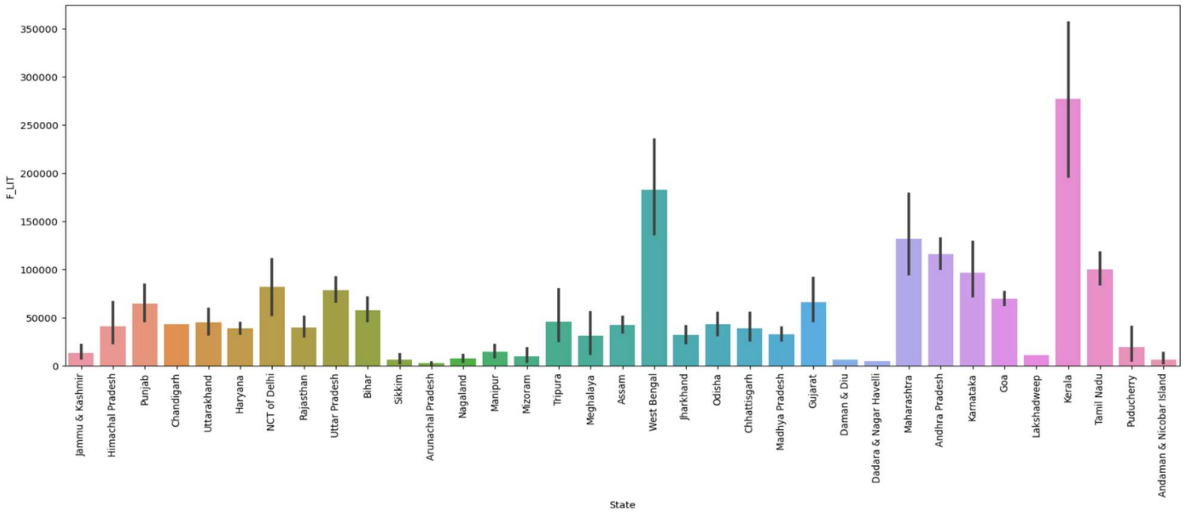


Figure 35: Barplot between State & F_LIT

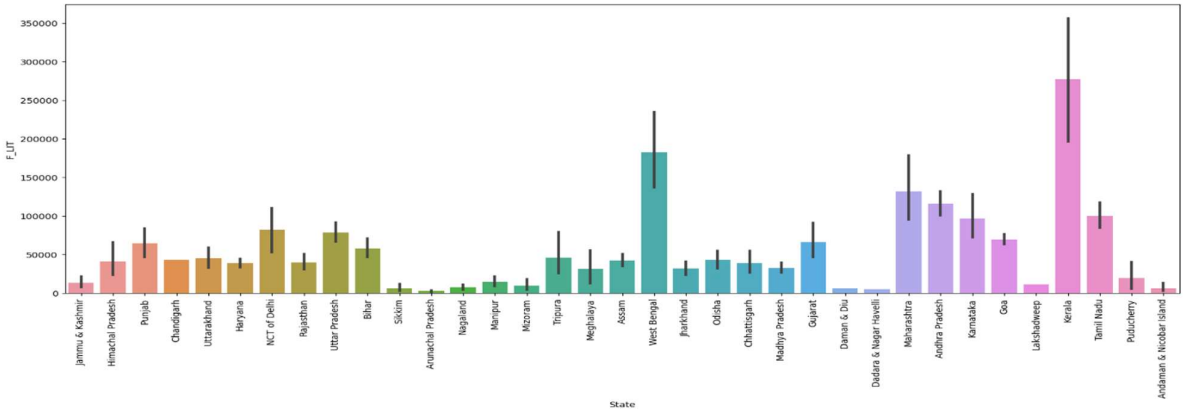


Figure 36: Barplot between State & TOT_WORK_M

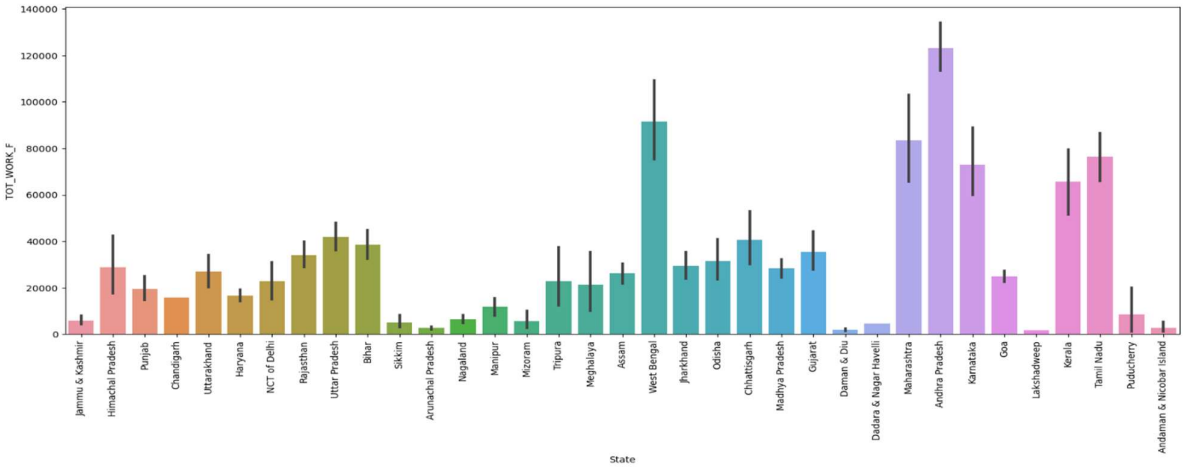


Figure 37: Barplot between State & TOT_WORK_F

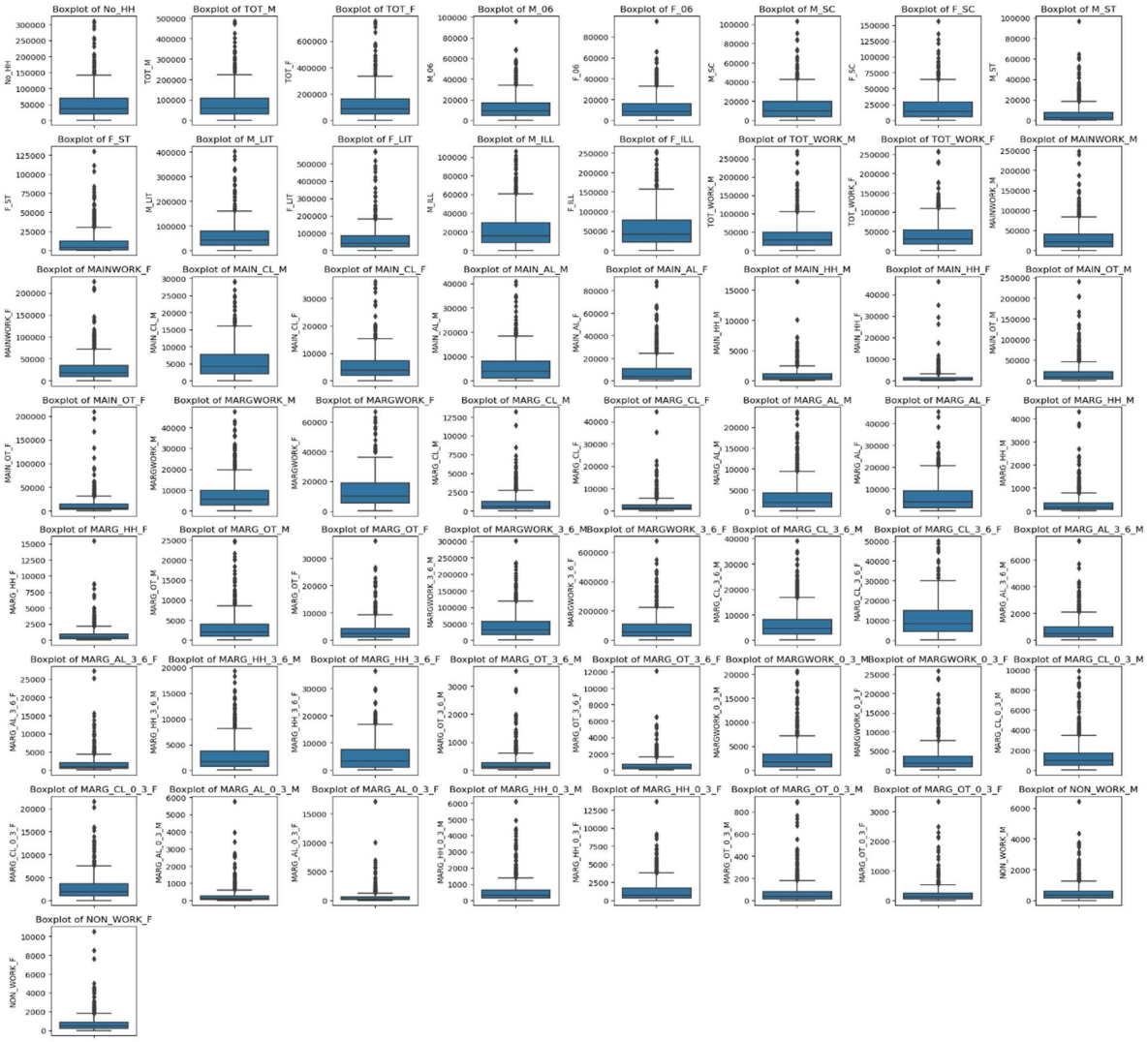


Figure 38: Boxplot before treating outliers using IQR method

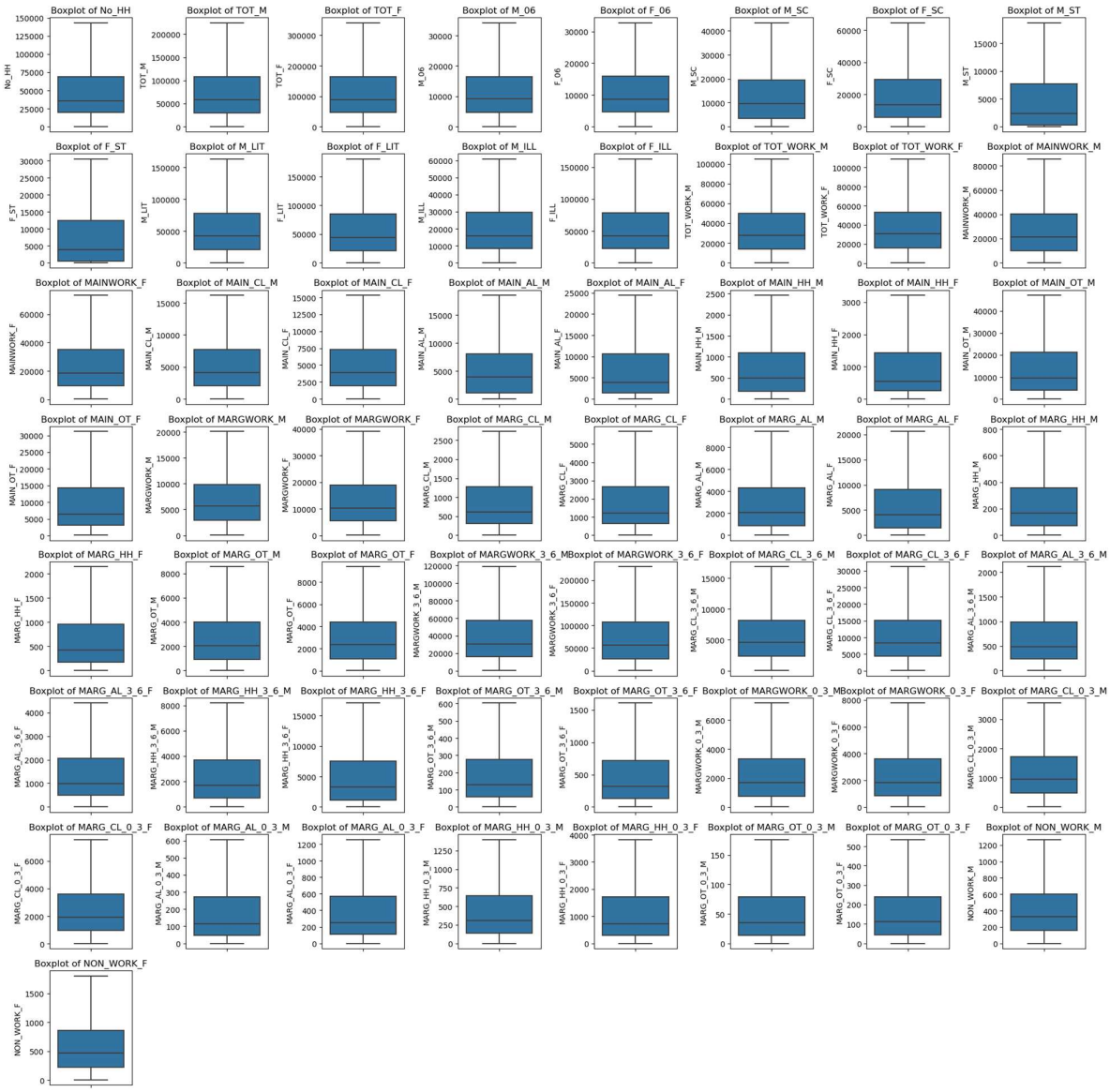


Figure 39: Boxplot after treating outliers using IQR Method

2.2) Data Preprocessing - Check for and treat (if needed) missing values - Check for and treat (if needed) data irregularities - Scale the Data using the z-score method - Visualize the data before and after scaling and comment on the impact on outliers

Ans:

There are no missing values

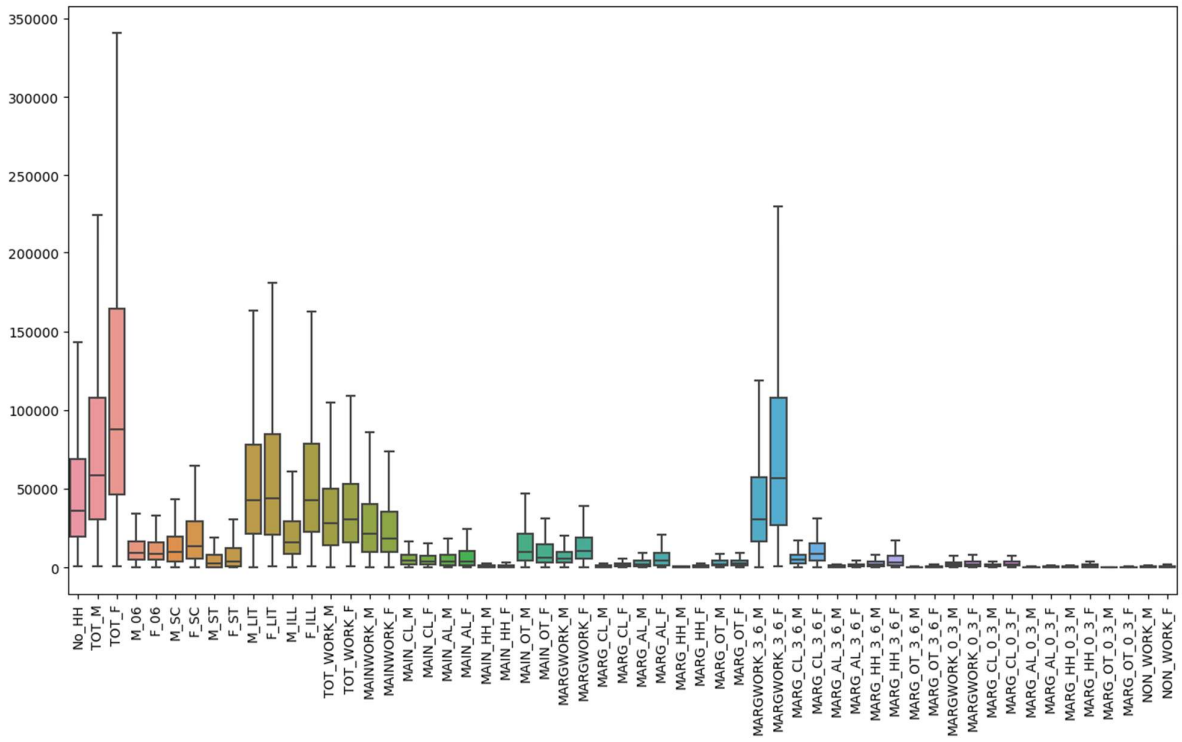


Figure 40: Box plot before scaling the data set with z-score method

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	..
count	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	..
mean	51222.871875	79940.576563	122372.084375	12309.098438	11942.300000	13820.946875	20778.392188	6191.807813	10155.640625	57967.979688	..
std	48135.405475	73384.511114	113600.717282	11500.906881	11326.294567	14426.373130	21727.887713	9912.668948	15875.701488	55910.282466	..
min	350.000000	391.000000	698.000000	56.000000	56.000000	0.000000	0.000000	0.000000	0.000000	286.000000	..
25%	19484.000000	30228.000000	46517.750000	4733.750000	4672.250000	3466.250000	5603.250000	293.750000	429.500000	21298.000000	..
50%	35837.000000	58339.000000	87724.500000	9159.000000	8663.000000	9591.500000	13709.000000	2333.500000	3834.500000	42693.500000	..
75%	68892.000000	107918.500000	164251.750000	16520.250000	15902.250000	19429.750000	29180.000000	7658.000000	12480.250000	77989.500000	..
max	310450.000000	485417.000000	750392.000000	96223.000000	95129.000000	103307.000000	156429.000000	96785.000000	130119.000000	403261.000000	..

8 rows × 57 columns

Figure 41: Statistical summary before scaling the data set with z-score method

After applying z score method for data scaling,

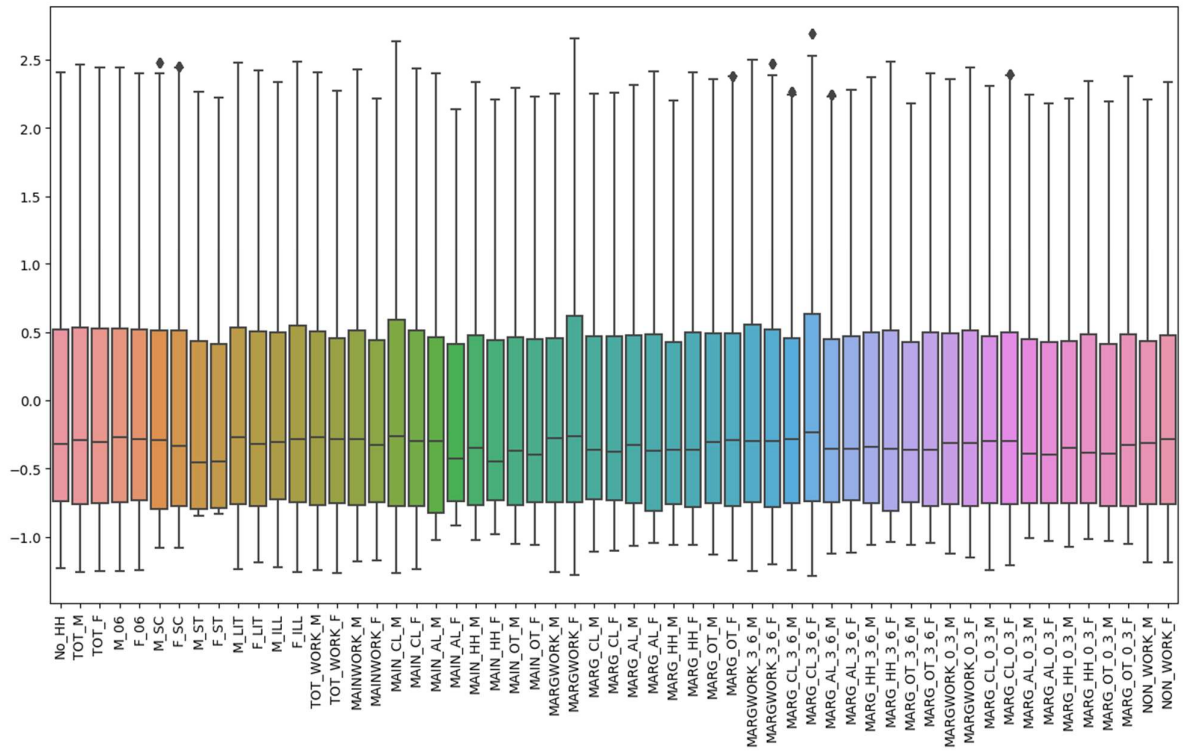


Figure 42: Box plot after scaling the data set with z-score method

	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	MARG_AL_0_3_M	MARG_AL_0_
count	640.00	640.00	640.00	640.00	640.00	640.00	640.00	640.00	640.00	640.00	...	640.00	640.00	640.00	640
mean	0.00	-0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	...	-0.00	-0.00	0.00	-0
std	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	...	1.00	1.00	1.00	1
min	-1.06	-1.08	-1.07	-1.07	-1.05	-0.96	-0.96	-0.63	-0.64	-1.03	...	-0.93	-0.98	-0.55	-0
25%	-0.66	-0.68	-0.67	-0.66	-0.64	-0.72	-0.70	-0.60	-0.61	-0.66	...	-0.61	-0.65	-0.45	-0
50%	-0.32	-0.29	-0.31	-0.27	-0.29	-0.29	-0.33	-0.39	-0.40	-0.27	...	-0.30	-0.30	-0.30	-0
75%	0.37	0.38	0.37	0.37	0.35	0.39	0.39	0.15	0.15	0.36	...	0.22	0.30	0.04	0
max	5.39	5.53	5.53	7.30	7.35	6.21	6.25	9.15	7.56	6.18	...	5.70	6.77	12.19	14

8 rows × 57 columns

Figure 43: Statistical summary after scaling the data set with z-score method

2.3) PCA - Create the covariance matrix - Get eigen values and eigen vectors - Identify the optimum number of PCs - Show Scree plot - Compare PCs with Actual Columns and identify which is explaining most variance - Write inferences about all the PCs in terms of actual variables - Write linear equation for first PC Note: For the scope of this project, take at least 90% explained variance.

Ans:

Find below steps for PCA (Principal Component Analysis

1. Performing Outlier Treatment
2. Scaling of the data
3. Create Covariance Matrix
4. Extract Eigen Vector

5. Find Eigen Value
6. Create WSS Scree plot for variance
7. Find a cut off for selecting the number of PCs

```
Covariance matrix :
 [[2.65384958 2.71559682 3.0810424  ... 3.32705346 3.26012156 3.09943089]
 [2.71559682 2.95818195 3.21927718 ... 3.52898196 3.42192001 3.25382599]
 [3.0810424  3.21927718 3.85909673 ... 4.05405594 3.96445766 3.77161732]
 ...
 [3.32705346 3.52898196 4.05405594 ... 4.36443144 4.25265838 4.05275828]
 [3.26012156 3.42192001 3.96445766 ... 4.25265838 4.15880459 3.96567682]
 [3.09943089 3.25382599 3.77161732 ... 4.05275828 3.96567682 3.79502824]]
```

Figure 44: Covariance Matrix of the dataset

```
Eigen Vector :
 [[ 0.16  0.17  0.17  0.16  0.16  0.15  0.15  0.03  0.03  0.16  0.15  0.16
   0.17  0.16  0.15  0.15  0.12  0.1   0.07  0.11  0.07  0.13  0.08  0.12
   0.11  0.16  0.16  0.08  0.05  0.13  0.11  0.14  0.13  0.16  0.15  0.16
   0.16  0.17  0.16  0.09  0.05  0.13  0.11  0.14  0.12  0.15  0.15  0.15
   0.14  0.05  0.04  0.12  0.12  0.14  0.13  0.15  0.13]
 [-0.13 -0.09 -0.1  -0.02 -0.02 -0.05 -0.05  0.03  0.03 -0.12 -0.15 -0.01
  -0.01 -0.13 -0.09 -0.18 -0.15  0.06  0.09 -0.03 -0.06 -0.08 -0.08 -0.21
  -0.21  0.09  0.13  0.27  0.25  0.17  0.14  0.07  0.02 -0.09 -0.12 -0.04
  -0.11  0.08  0.1   0.26  0.24  0.16  0.13  0.06  0.01 -0.09 -0.13  0.15
   0.18  0.25  0.24  0.19  0.18  0.08  0.05 -0.07 -0.07]
 [-0.    0.06  0.04  0.06  0.05  0.   -0.03 -0.12 -0.14  0.08  0.12 -0.02
  -0.09  0.05 -0.06  0.05 -0.06 -0.07 -0.01 -0.25 -0.25  0.03 -0.06  0.14
   0.1  -0.01 -0.05  0.2   0.27 -0.19 -0.27 -0.02 -0.08  0.11  0.1   0.06
   0.08 -0.02 -0.07  0.15  0.26 -0.2  -0.28 -0.02 -0.08  0.11  0.1   0.05
   0.02  0.27  0.28 -0.14 -0.2  -0.02 -0.08  0.11  0.1 ]
 [-0.13 -0.02 -0.07  0.01  0.01  0.01 -0.03 -0.22 -0.23 -0.04 -0.06  0.03
  -0.08 -0.04 -0.23 -0.07 -0.25 -0.09 -0.29 -0.14 -0.29  0.15  0.05 -0.04
  -0.12  0.09 -0.09 -0.05 -0.17  0.09 -0.11  0.24  0.2   0.09  0.03 -0.
   0.    0.09 -0.11 -0.04 -0.18  0.08 -0.14  0.24  0.19  0.09  0.03  0.09
  -0.02 -0.1  -0.14  0.13  0.    0.23  0.21  0.08  0.02]
 [-0.01 -0.03 -0.01 -0.05 -0.04 -0.17 -0.16  0.43  0.44 -0.01  0.06 -0.1
  -0.12 -0.02 -0.04 -0.04 -0.08 -0.29 -0.24 -0.21 -0.18 -0.13 -0.14  0.06
   0.08  0.06  0.09 -0.02 -0.06  0.02  0.08 -0.06 -0.03  0.12  0.17 -0.04
   0.    0.05  0.07 -0.01 -0.06  0.01  0.06 -0.07 -0.04  0.11  0.14  0.08
   0.13 -0.05 -0.05  0.06  0.13 -0.04  0.    0.16  0.24]
 [ 0.   -0.07 -0.04 -0.15 -0.15 -0.06 -0.04  0.22  0.23 -0.06 -0.05 -0.12
  -0.03 -0.    0.11  0.02  0.12 -0.01  0.1  -0.03  0.02  0.17  0.42  0.02
   0.08 -0.09  0.02  0.03  0.09 -0.14 -0.09  0.09  0.37 -0.06  0.   -0.14
  -0.11 -0.1   0.02  0.01  0.09 -0.14 -0.08  0.1   0.38 -0.06  0.01 -0.06
  -0.    0.07  0.08 -0.12 -0.11  0.06  0.3  -0.05 -0.02]
 [-0.12  0.09 -0.    0.17  0.17 -0.   -0.08  0.41  0.36  0.05 -0.02  0.2
   0.03  0.05 -0.12  0.05 -0.09  0.39  0.21 -0.01 -0.16  0.12 -0.14 -0.02
  -0.07  0.02 -0.15  0.03 -0.05  0.02 -0.15  0.11 -0.05 -0.   -0.12  0.13
   0.05  0.03 -0.14  0.05 -0.02  0.02 -0.15  0.12 -0.04 -0.   -0.09 -0.01
  -0.2  -0.04 -0.1   0.01 -0.15  0.08 -0.07 -0.02 -0.2 ]
 [ 0.06  0.11  0.09  0.17  0.17 -0.13 -0.14  0.02  0.01  0.1   0.11  0.13
   0.04  0.08 -0.04  0.09 -0.02 -0.23 -0.3   0.05 -0.11 -0.14  0.38  0.14
   0.07  0.   -0.09  0.07  0.    0.14 -0.03 -0.21  0.07 -0.14 -0.23  0.14
   0.14 -0.   -0.12  0.07 -0.02  0.14 -0.05 -0.21  0.08 -0.14 -0.22  0.03
   0.    0.08  0.06  0.14  0.07 -0.17  0.04 -0.12 -0.2 ]
 [ 0.    0.02  0.01 -0.06 -0.06  0.04  0.04  0.02  0.04  0.05  0.02 -0.06
   0.    0.05  0.11  0.06  0.14 -0.36 -0.11 -0.18  0.02  0.38 -0.21  0.17
   0.28 -0.04 -0.01  0.02  0.02 -0.01  0.09  0.24 -0.15 -0.11 -0.14 -0.01
  -0.03 -0.05  0.    0.01  0.01 -0.03  0.08  0.24 -0.17  0.1  -0.08  0.01
  -0.03  0.03  0.05  0.05  0.12  0.22 -0.08 -0.16 -0.35]
 [ 0.02 -0.02 -0.04 -0.15 -0.17  0.45  0.45  0.16  0.13 -0.01 -0.04 -0.07
  -0.03  0.08 -0.07  0.09 -0.05  0.07 -0.36  0.26  0.09  0.01 -0.15  0.03
  -0.02  0.   -0.11  0.04 -0.    0.02 -0.11 -0.07  0.01 -0.03 -0.1  -0.12
  -0.02 -0.03 -0.15 -0.01 -0.06  0.01 -0.13 -0.07 -0.02 -0.07 -0.14  0.15
   0.05  0.15  0.13  0.07 -0.04 -0.04  0.07  0.17  0.05]]
```

Figure 45: Eigen Vector of the dataset

```
Eigen values of all 10 pc's :
 [0.56 0.14 0.07 0.06 0.04 0.03 0.02 0.01 0.01 0.01]
```

Figure 46: Eigen Values of the dataset

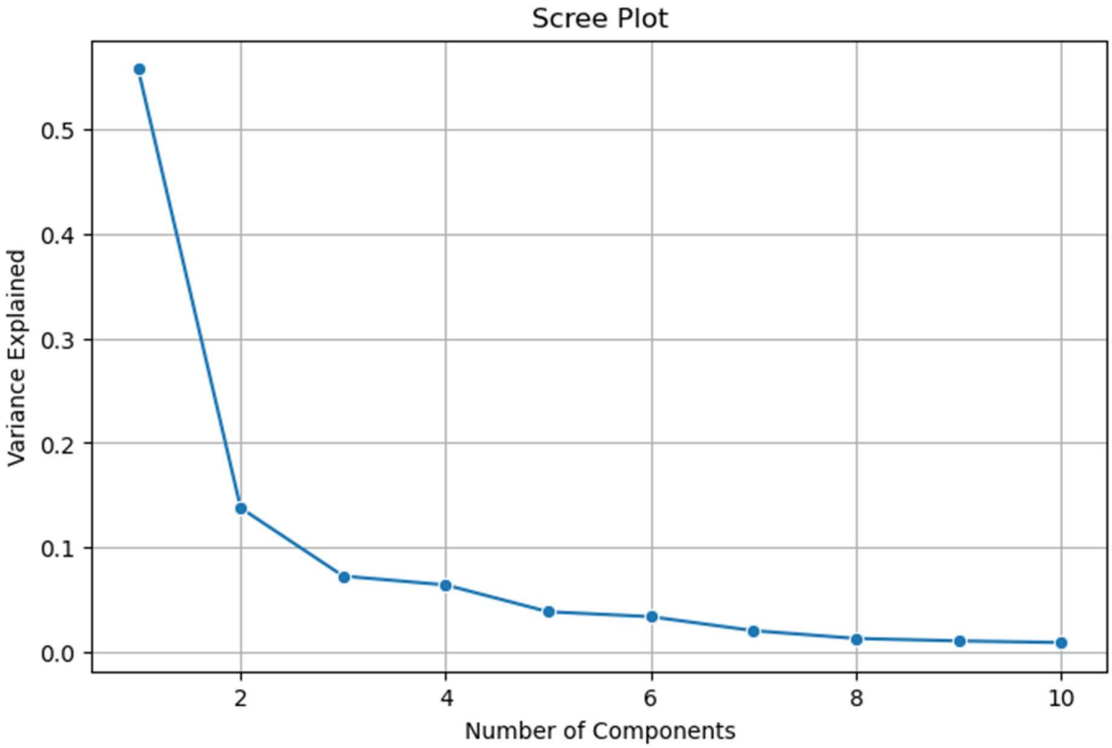


Figure 47: Scree Plot of the dataset

Number of components needed to explain 90% of the variance: 5

Optimum number of PC's is 5.

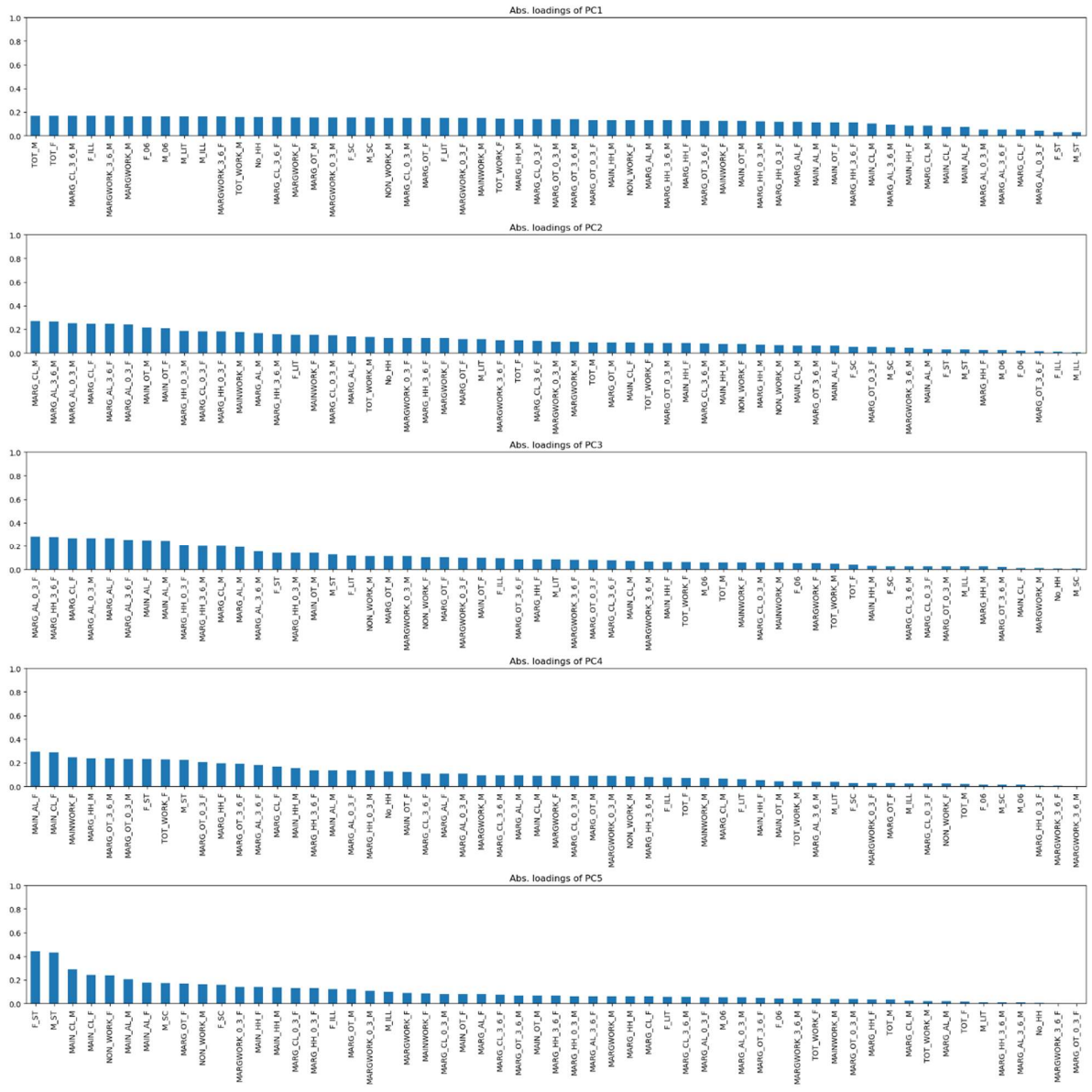


Figure 48: Comparison of PCs with Actual Variables

Linear equation for first PC is,

PC1=a1x1*a2x2*……anxn.