ML-1 Project

(Coded)

DSBA

By:

E. AuroRajashri

List of Content

PART-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 Excel File.

Perform the following in given order:

- Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.
- Treat missing values in CPC, CTR and CPM using the formula given. You may refer to the Bank_KMeans Solution File to understand the coding behind treating the missing values using a specific formula. You have to basically create an user defined function and then call the function for imputing.
- Check if there are any outliers.
- Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an analyst your judgement may be different from another analyst).
- Perform z-score scaling and discuss how it affects the speed of the algorithm.
- Perform clustering and do the following:
- Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.
- Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.
- Print silhouette scores for up to 10 clusters and identify optimum number of clusters.
- Profile the ads based on optimum number of clusters using silhouette score and your domain understanding
  [Hint: Group the data by clusters and take sum or mean to identify trends in clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots.]
- Conclude the project by providing summary of your learnings.

1.1 Clustering: Define the problem and perform Exploratory Data Analysis

1.1.1 Problem Definition

- Imported necessary libraries like NumPy, Pandas,matplotlib,seaborn.
- Loaded the given dataset to dataframe df

	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

Fig 1: Dataset Head rows

1.1.2 Check shape, Data types, statistical summary

- Dataset has shape of 23066 rows and 19 columns. And it has 6 float datatypes ,7 integer datatypes and 6 object datatypes.

```
(23066, 19)
```

```
<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
```

Fig 2: Dataset Info

- Below is the dataset statistical Summary

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Rev
count	23066.000000	23066.000000	23066.000000	2.306600e+04	2.306600e+04	2.306600e+04	23066.000000	23066.000000	23066.000000	23066.00
mean	385.163097	337.896037	96674.468048	2.432044e+06	1.295099e+06	1.241520e+06	10678.518816	2706.625689	0.335123	1924.25
std	233.651434	203.092885	61538.329557	4.742888e+06	2.512970e+06	2.429400e+06	17353.409363	4067.927273	0.031963	3105.23
min	120.000000	70.000000	33600.000000	1.000000e+00	1.000000e+00	1.000000e+00	1.000000	0.000000	0.210000	0.00
25%	120.000000	250.000000	72000.000000	3.367225e+04	1.828250e+04	7.990500e+03	710.000000	85.180000	0.330000	55.36
50%	300.000000	300.000000	72000.000000	4.837710e+05	2.580875e+05	2.252900e+05	4425.000000	1425.125000	0.350000	926.33
75%	720.000000	600.000000	84000.000000	2.527712e+06	1.180700e+06	1.112428e+06	12793.750000	3121.400000	0.350000	2091.33
max	728.000000	600.000000	216000.000000	2.759286e+07	1.470202e+07	1.419477e+07	143049.000000	26931.870000	0.350000	21276.18

Fig 3: Dataset Statistical Summary

- There are no duplicates in the dataset.

<u>1.1.3 Univariate analysis and Bivariate analysis</u>

- **<u>Categorical Variables</u>**

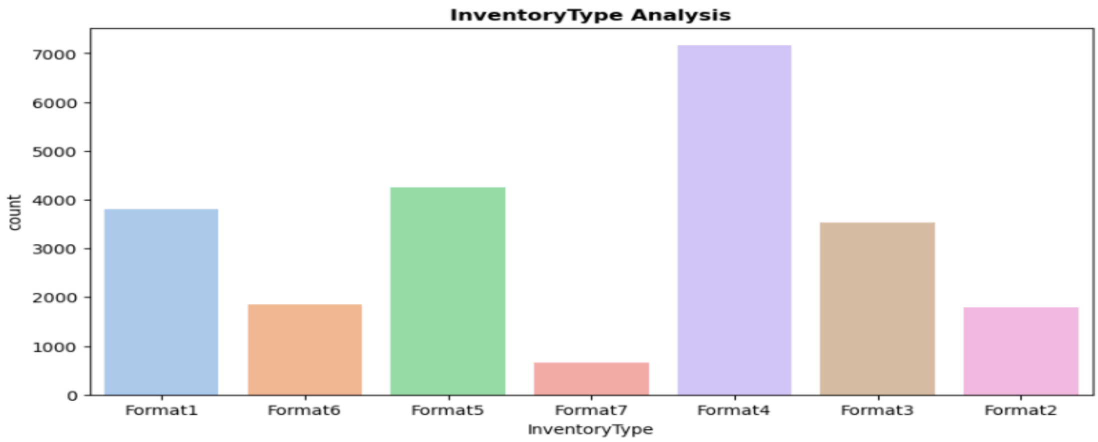


Fig 4: Inventory Type Analysis

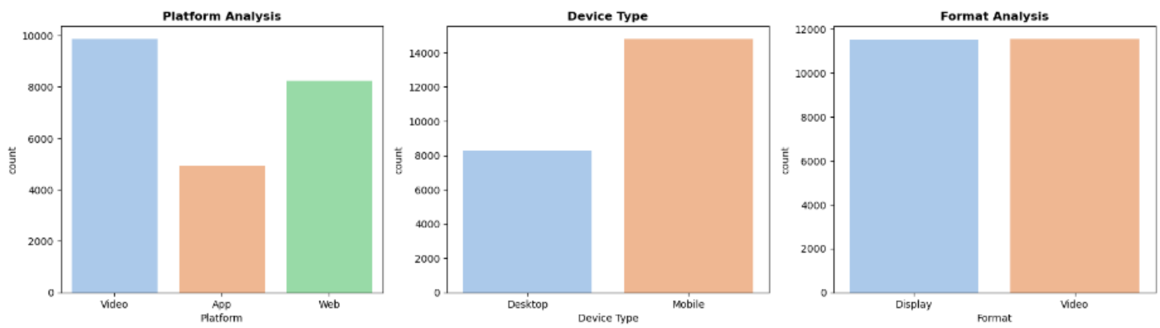


Fig 5: Platform, Device type and Format Analysis

1. Format 4 is the most used Inventory type, followed by Format 5 and Format 1.
2. The most preferred platform is Video, followed by Web and App.
3. Mobile is the top preferred Device type than Desktop.
4. Choice of Display and Video format are almost the same.

- **<u>Numerical Variables</u>**

1. The median of the spend lies between 1000 to 2000

2. There is a high frequency of data points with a low number of impressions, peaking at around 0.5 million impressions. This suggests that most of the data points have a low number of impressions.

3. KDE confirms the skewness of the distribution towards lower impression counts.

4. right-skewed distribution of clicks, where the most of the data points are clustered at the lower end of the click range, suggesting that lower click counts are more common.

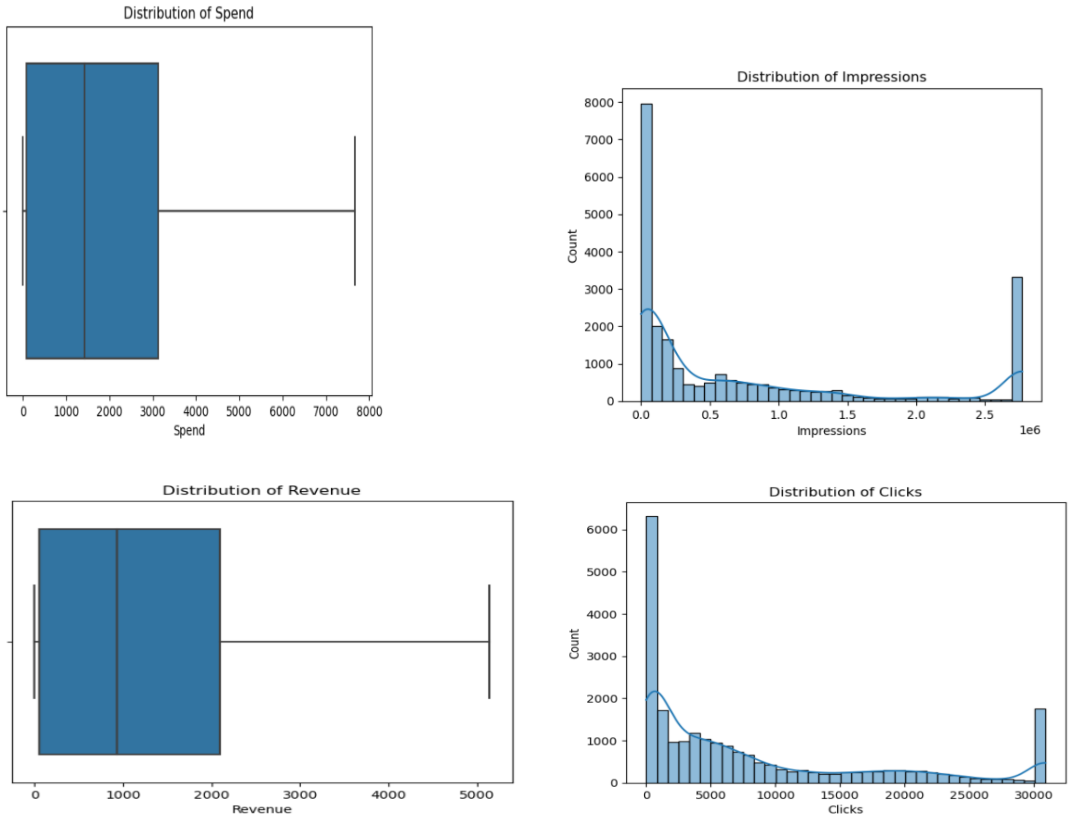


Fig 6: Spend,Impression,Revenue and clicks distribution

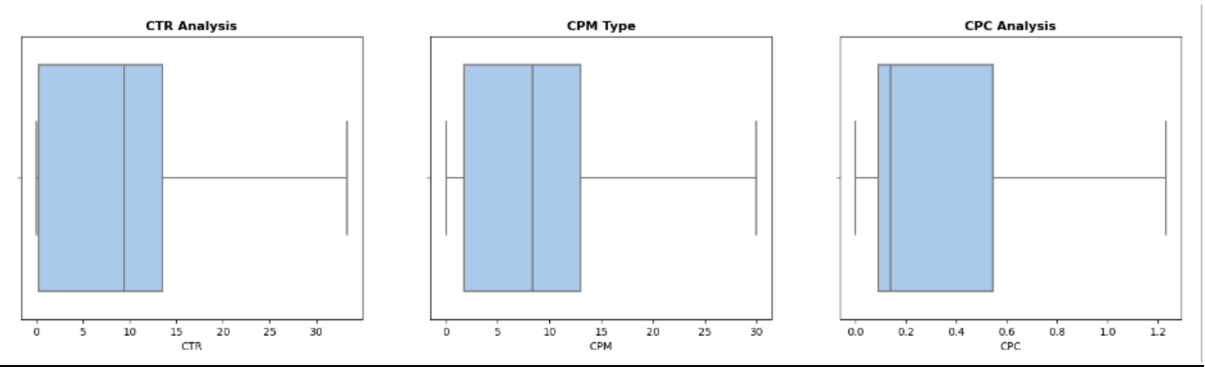


Fig 7: CTR,CPM,CPC distribution

- **<u>Relationship between Numerical Variables</u>**

  Based on Pair plot:

1) Ad Length and Ad Width: There is a clear upward trend indicating that as the ad length increases, the ad width tends to increase as well
2) Ad Length and Ad Size: Since ad size is likely a function of ad length and ad width, it's not surprising to see a positive correlation here, with larger ad lengths contributing to larger overall ad size
3) Ad Width and Ad Size: Similar to ad length, as the ad width increases, the ad size also increases, showing a positive correlation
4) Impressions and Clicks: There is a positive correlation, as more impressions typically lead to more clicks
5) Impressions and Revenue: The scatter plot suggests that higher impressions are associated with higher revenue, indicating a positive correlation
6) Clicks and Revenue: This scatter plot also shows a positive correlation, where more clicks are associated with higher revenue

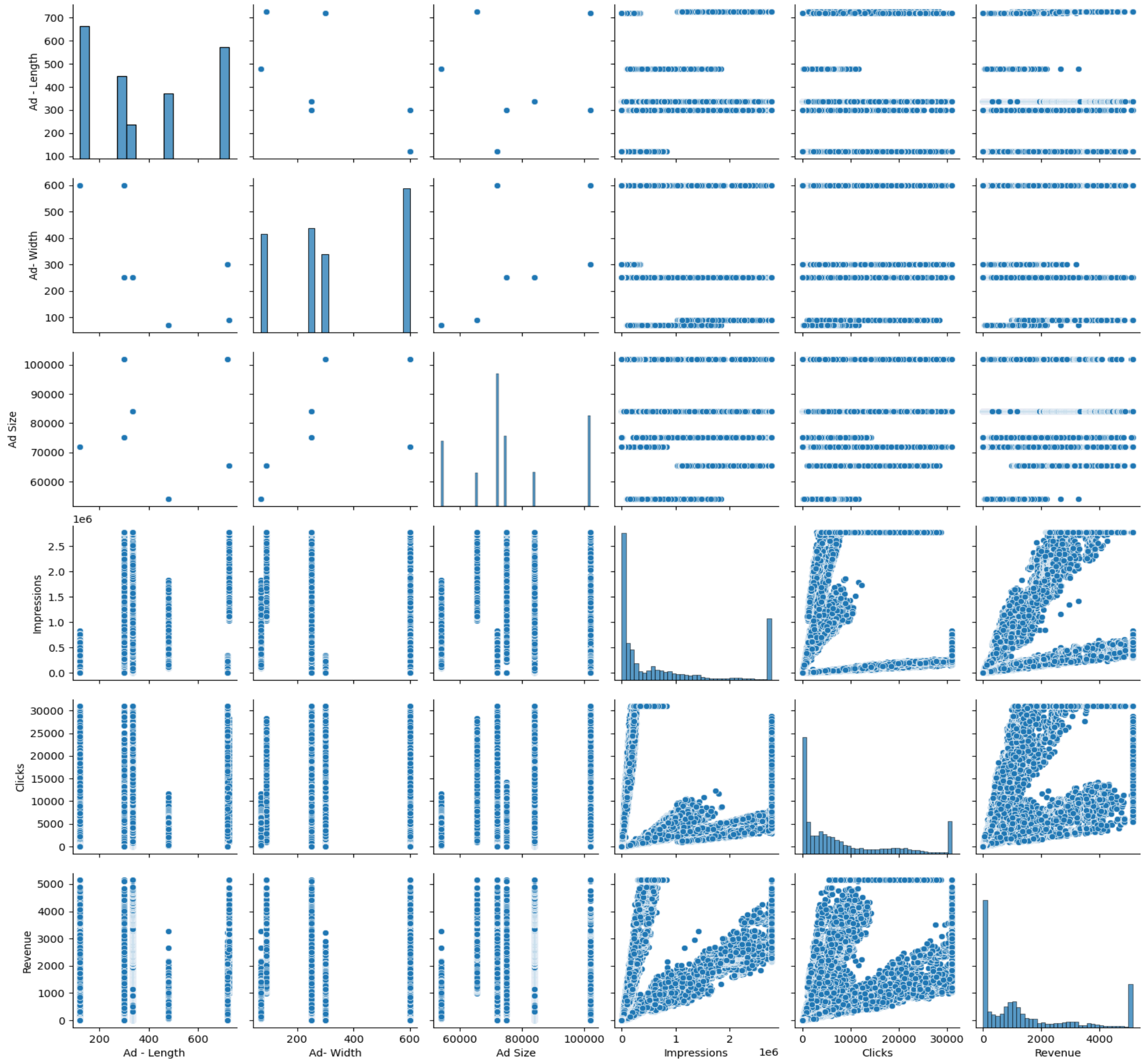


Fig 8: Pair plot of numeric variables

7) From above, we could see that, mobile transactions are generally of lower value compared to desktop transactions. This is because most significant peak occurring much earlier than in the desktop distribution

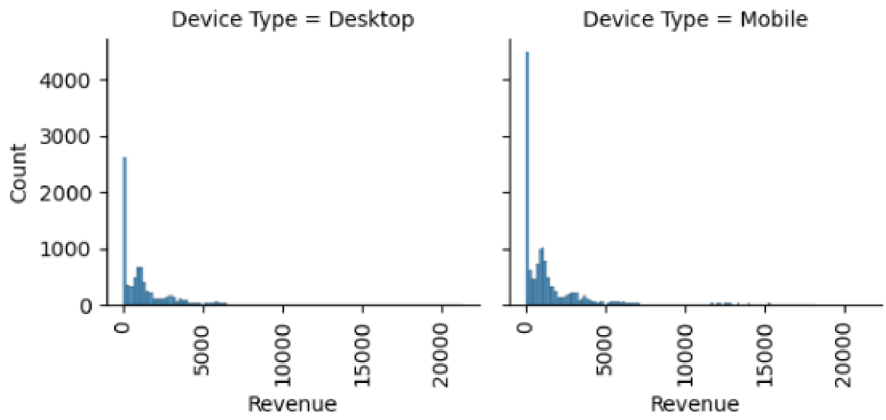


Fig 9: Revenue based on Device Type

8) The median spending on the App platform is higher than that on the Video platform but lower than on the Web platform.
9) Web being the platform where users tend to spend the most, followed by App, and then Video.

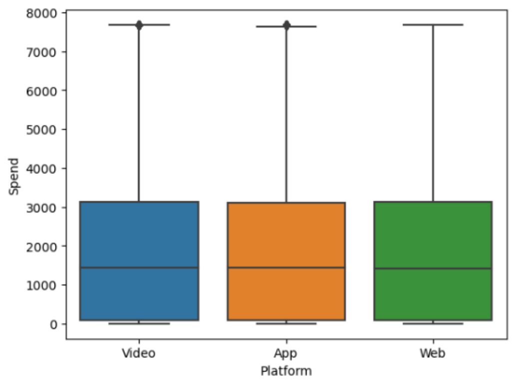


Fig 10: Spend based on Platform

1.2 Clustering: Data Preprocessing

1.2.1 Missing value check and treatment

- There is missing values in CTR ,CPM,CPC of 4736 each as shown below

```
Timestamp                 0
InventoryType             0
Ad - Length               0
Ad- Width                 0
Ad Size                   0
Ad Type                   0
Platform                  0
Device Type               0
Format                    0
Available_Impressions     0
Matched_Queries           0
Impressions               0
Clicks                    0
Spend                     0
Fee                       0
Revenue                   0
CTR                    4736
CPM                    4736
CPC                    4736
dtype: int64
```

Fig 11: Missing values in the dataset

- Imputed by the following formula and we could see there is no null post that.

```python
#creating user defined function
def calculate_cpc(x):
    spend=df.Spend
    clicks=df.Clicks
    cpc = (spend/clicks)
    return cpc

def calculate_ctr(x):
    impressions = df.Impressions
    clicks=df.Clicks
    ctr = (clicks/(impressions)*100)
    return ctr

def calculate_cpm(x):
    spend=df.Spend
    impressions = df.Impressions
    cpm = (spend/impressions)*1000
    return cpm
```

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

Fig 12: Formulae and Post imputation, Missing values in the dataset

1.2.2 <u>Outlier Treatment</u>

- We could see there is outlier in all features except Ad_length and Ad_width. Treated by IQR method.

<u>Before Outlier treatment:</u>

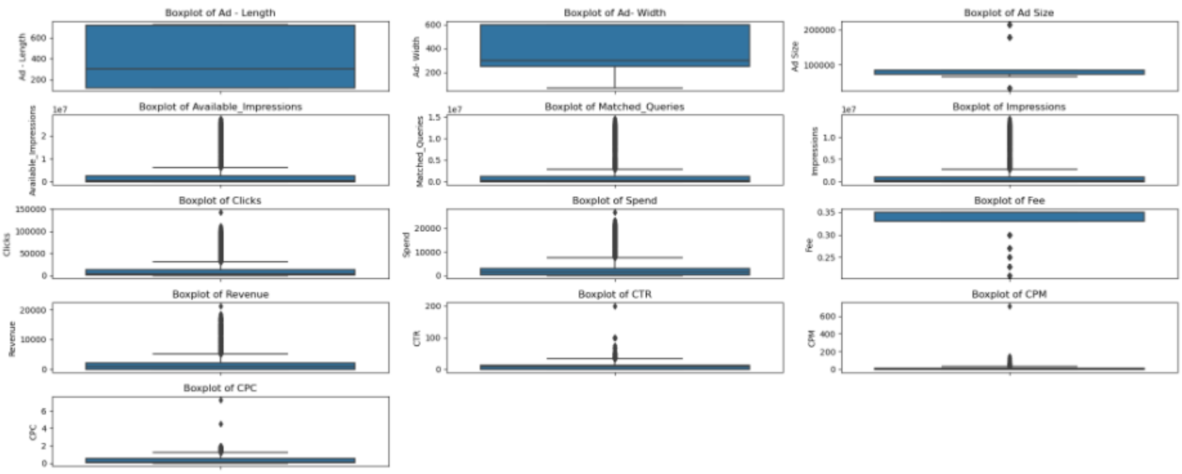


Fig 13: Before Outlier treament

<u>After Outlier treatment:</u>

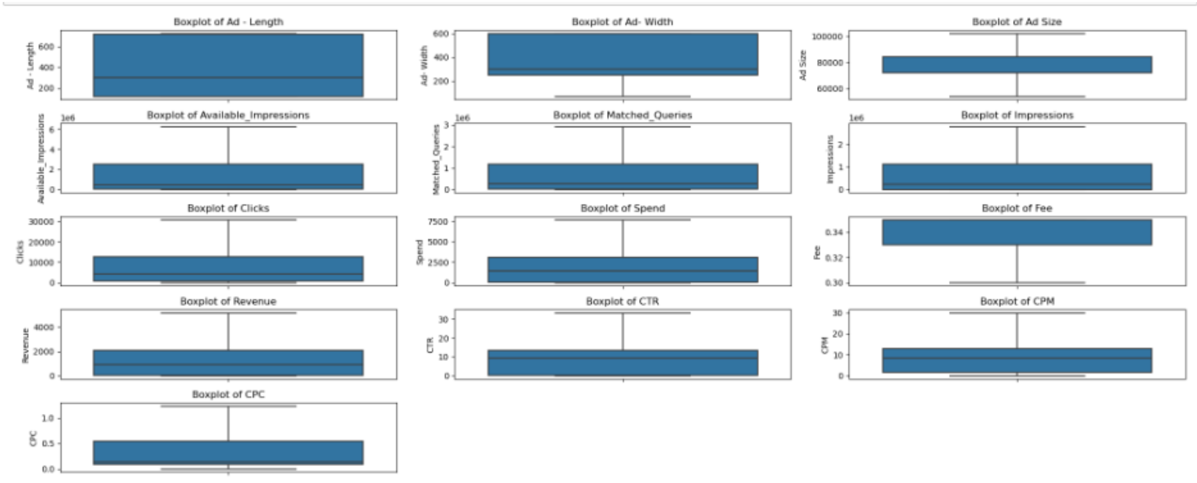


Fig 14: After Outlier treatment

1.2.3 <u>**z-score scaling**</u>

- From the below, we could see there is different scale units among the features. So, there is a need for scaling. Z-score scaling is done.

<u>Before scaling:</u>

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Rev
count	23066.000000	23066.000000	23066.000000	2.306600e+04	2.306600e+04	2.306600e+04	23066.000000	23066.000000	23066.000000	23066.00
mean	385.163097	337.896037	96674.468048	2.432044e+06	1.295099e+06	1.241520e+06	10678.518816	2706.625689	0.335123	1924.25
std	233.651434	203.092885	61538.329557	4.742888e+06	2.512970e+06	2.429400e+06	17353.409363	4067.927273	0.031963	3105.23
min	120.000000	70.000000	33600.000000	1.000000e+00	1.000000e+00	1.000000e+00	1.000000	0.000000	0.210000	0.00
25%	120.000000	250.000000	72000.000000	3.367225e+04	1.828250e+04	7.990500e+03	710.000000	85.180000	0.330000	55.36
50%	300.000000	300.000000	72000.000000	4.837710e+05	2.580875e+05	2.252900e+05	4425.000000	1425.125000	0.350000	926.33
75%	720.000000	600.000000	84000.000000	2.527712e+06	1.180700e+06	1.112428e+06	12793.750000	3121.400000	0.350000	2091.33
max	728.000000	600.000000	216000.000000	2.759286e+07	1.470202e+07	1.419477e+07	143049.000000	26931.870000	0.350000	21276.18

Fig 15: Before scaling

<u>After scaling:</u>

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Rev
count	2.306600e+04	2.306600e+04	2.306600e+04	2.306600e+04	2.306600e+04	23066.000000	2.306600e+04	2.306600e+04	2.306600e+04	2.30660(
mean	1.281478e-16	-1.182903e-16	3.055833e-16	9.857525e-18	1.971505e-17	0.000000	-1.182903e-16	-9.857525e-17	1.143473e-15	3.94301
std	1.000022e+00	1.000022e+00	1.000022e+00	1.000022e+00	1.000022e+00	1.000022	1.000022e+00	1.000022e+00	1.000022e+00	1.00002;
min	-1.134891e+00	-1.319110e+00	-1.467840e+00	-7.561823e-01	-7.792648e-01	-0.768806	-8.674882e-01	-8.931702e-01	-2.222416e+00	-8.800
25%	-1.134891e+00	-4.327968e-01	-2.975645e-01	-7.403406e-01	-7.614468e-01	-0.760655	-7.934379e-01	-8.580464e-01	-5.675316e-01	-8.464
50%	-3.644957e-01	-1.865987e-01	-2.975645e-01	-5.285774e-01	-5.277221e-01	-0.538975	-4.054310e-01	-3.055230e-01	5.357244e-01	-3.176
75%	1.433093e+00	1.290590e+00	4.826195e-01	4.330590e-01	3.714976e-01	0.366051	4.686290e-01	3.939323e-01	5.357244e-01	3.89802
max	1.467332e+00	1.290590e+00	1.652896e+00	2.193158e+00	2.070914e+00	2.056111	2.361729e+00	2.271900e+00	5.357244e-01	2.24421ε

Fig 16: after scaling

1.3 Clustering: Hierarchical Clustering

1.3.1 Construct a dendrogram using Ward linkage and Euclidean distance

- Imported dendrogram, linkage from scipy.cluster.hierarchy.
- By 'Ward' method and 'euclidean' metric, constructed the below dendrogram by truncating to the last 10 clustering.

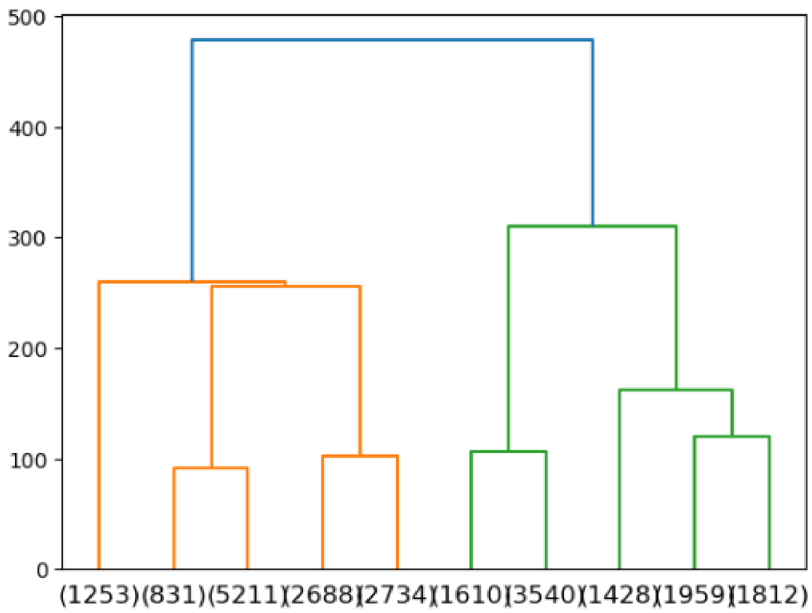


Fig 17: Truncated Dendrogram

1.3.2 Identify the optimum number of Clusters

- By the above dendrogram, we could see '5' could be the optimal number of clusters could be formed.
- fCluster are applied and the output is as below. And the column is added to original df.

```
array([4, 4, 4, ..., 3, 2, 3], dtype=int32)
```

Fig 18: Hierarchical cluster

Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	clusters
5000	Inter222	Video	Desktop	Display	1806	325	323	1	0.00	0.35	0.0000	0.309598	0.0	0.00	4
5000	Inter227	App	Mobile	Video	1780	285	285	1	0.00	0.35	0.0000	0.350877	0.0	0.00	4
5000	Inter222	Video	Desktop	Display	2727	356	355	1	0.00	0.35	0.0000	0.281690	0.0	0.00	4
5000	Inter228	Video	Mobile	Video	2430	497	495	1	0.00	0.35	0.0000	0.202020	0.0	0.00	4
5000	Inter217	Web	Desktop	Video	1218	242	242	1	0.00	0.35	0.0000	0.413223	0.0	0.00	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5000	Inter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	100.000000	70.0	0.07	3
5000	Inter224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.0260	50.000000	20.0	0.04	3
5000	Inter218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	100.000000	50.0	0.05	3
2000	inter230	Video	Mobile	Video	7	1	1	1	0.07	0.35	0.0455	100.000000	70.0	0.07	2
5000	Inter221	App	Mobile	Video	2	2	2	1	0.09	0.35	0.0585	50.000000	45.0	0.09	3

Fig 19: Hierarchical cluster added to original df

1.4 Clustering: K-means Clustering

1.4.1 Apply K-means Clustering

- Imported Kmeans, silhouette score and Silhouette samples library
- Below is the top 10 K means inertia.

```
[299858.00000000006,
 183349.1020288607,
 130878.34788742856,
 95573.8329226824,
 61539.18998404851,
 51676.89681600459,
 44598.262116139085,
 39597.84594043495,
 36061.729559138075,
 32998.39641381086]
```

Fig 20: K means Inertia

- It could be clear about the difference between inertia using elbow curve plot.

1.4.2 Plot the Elbow curve

- From the below elbow curve, we could see there is sudden drop of inertia from 1 to 5. Post 5, there is slow and smooth drop. So, 5 clusters would be the optimal number.

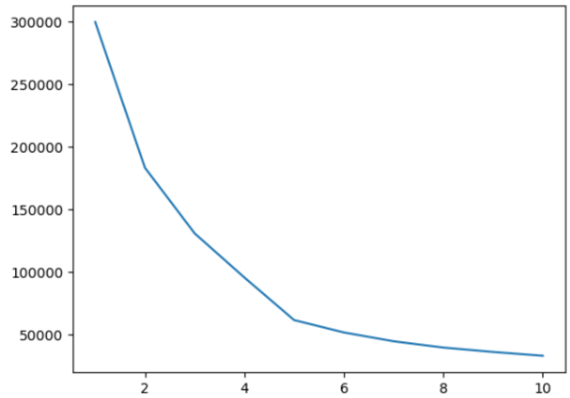


Fig 21: Elbow Curve

1.4.3 Check Silhouette Scores and Figure out the appropriate number of clusters

- To evaluate the model, silhouette score is used. As it is around 0.52 for 5 clusters, which is positive. It means clusters are very well separated.

```
0.5240956940501831
```

Fig 22: Silhouette score

- Below is the silhouette width which is positive as well meaning the mapping is correct to its centroid.

```
array([0.14263751, 0.14200708, 0.14309186, ..., 0.12833615, 0.38595215,
       0.12840723])
```

Fig 23: Silhouette width

- Silhouette width is added to the data frame as shown below.

Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	...	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	clusters	Clus_kmeans	sil_width
75000	Inter222	Video	Desktop	Display	1806	...	1	0.00	0.35	0.0000	0.309598	0.0	0.00	4	0	0.142638
75000	Inter227	App	Mobile	Video	1780	...	1	0.00	0.35	0.0000	0.350877	0.0	0.00	4	0	0.142007
75000	Inter222	Video	Desktop	Display	2727	...	1	0.00	0.35	0.0000	0.281690	0.0	0.00	4	0	0.143092
75000	Inter228	Video	Mobile	Video	2430	...	1	0.00	0.35	0.0000	0.202020	0.0	0.00	4	0	0.144273
75000	Inter217	Web	Desktop	Video	1218	...	1	0.00	0.35	0.0000	0.413223	0.0	0.00	4	0	0.141021
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
216000	Inter220	Web	Mobile	Video	1	...	1	0.07	0.35	0.0455	100.000000	70.0	0.07	3	1	0.128446
216000	Inter224	Web	Desktop	Video	3	...	1	0.04	0.35	0.0260	50.000000	20.0	0.04	3	1	0.196818
216000	Inter218	App	Mobile	Video	2	...	1	0.05	0.35	0.0325	100.000000	50.0	0.05	3	1	0.128336
72000	inter230	Video	Mobile	Video	7	...	1	0.07	0.35	0.0455	100.000000	70.0	0.07	2	3	0.385952
216000	Inter221	App	Mobile	Video	2	...	1	0.09	0.35	0.0585	50.000000	45.0	0.09	3	1	0.128407

Fig 24: Silhouette width added to df

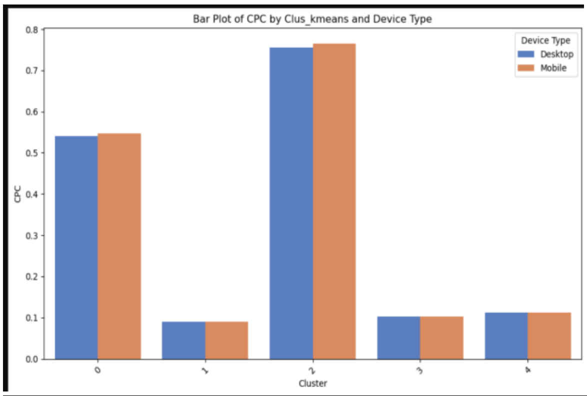
1.4.4 Cluster Profiling

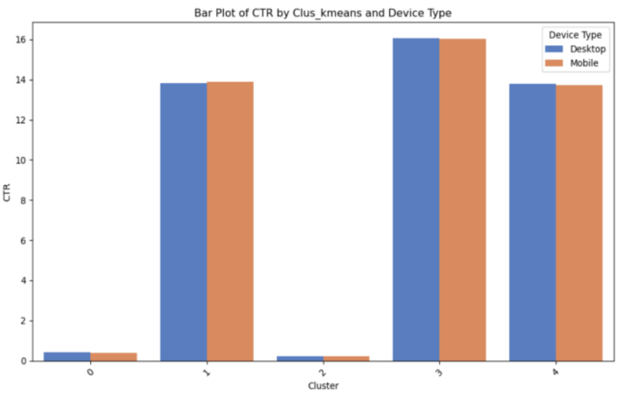
- Data are grouped by kmeans cluster and taken mean for the variables as shown below.

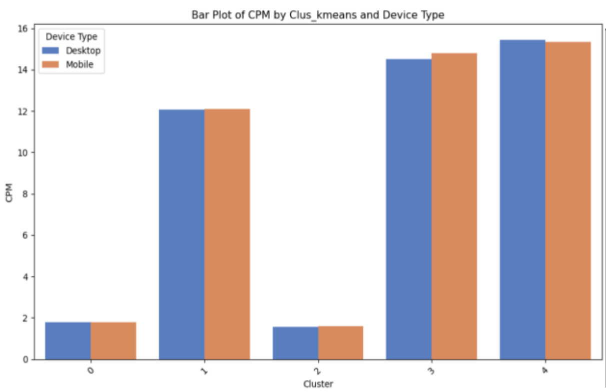
Clus_kmeans	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue
0	421.696255	152.001594	55008.841434	1.810314e+06	8.642623e+05	8.262209e+05	3263.131952	1500.090563	0.349264	977.424163
1	683.825492	303.785287	206160.821215	2.513465e+05	1.375509e+05	1.167714e+05	14406.540205	1252.285569	0.349538	815.541831
2	465.781944	199.148989	75176.566354	1.038821e+07	5.625808e+06	5.447310e+06	11245.754810	8646.647997	0.290439	6373.659814
3	143.280809	572.103004	76597.026364	3.209356e+04	1.962406e+04	1.349204e+04	1914.448804	209.162609	0.349988	135.993379
4	141.454782	572.446324	75614.834092	8.063284e+05	5.668641e+05	4.781485e+05	65315.176318	6990.360898	0.288302	5017.538285

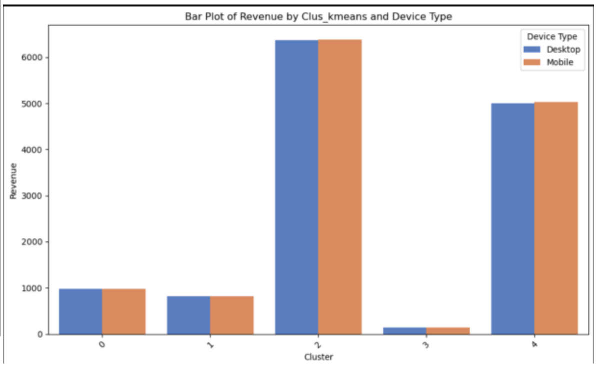
Fig 25: Data grouped by clusters

- As per rubric, plotted the bar graph of the above tabulation based on device type.









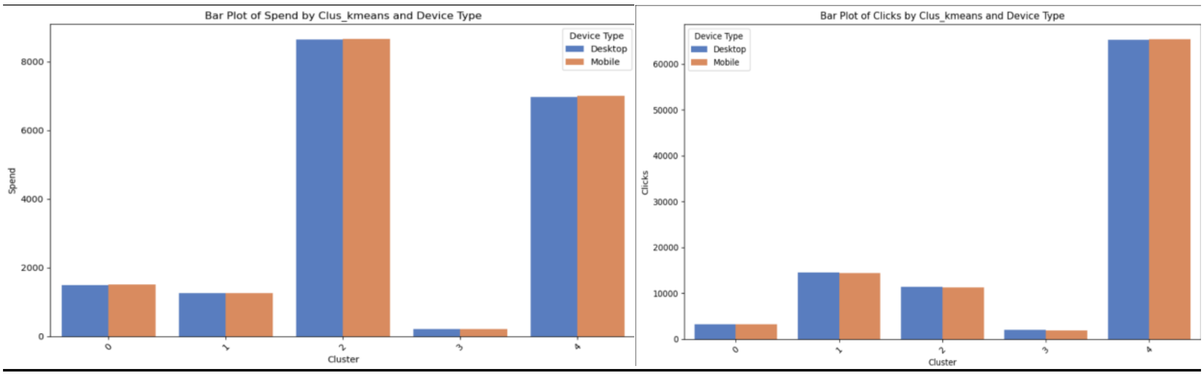


Fig 26: Data grouped by clusters by device type as hue

1.5 Clustering: Actionable Insights & Recommendations

- Based on above analysis, Clusters can be grouped into High spending, medium spending and low spending
- Cluster 2 and 4 are high spending, cluster 0 and 1 are medium spending and cluster 3 are low spending
- The cluster with the lowest Cost-Per-Click (CPC) is 'Cluster 0'. This indicates that, among the analysed clusters, Cluster 0 represents the most cost-effective segment for digital advertising, with the lowest average cost incurred per click on advertisements.
- Cluster 4 has the highest CTR for both Desktop and Mobile devices, making it the best performing cluster among the five presented in terms of CTR.
- Based on the analysis, the cluster with the highest revenue is Cluster 3. This suggests that Cluster 3 would be the best for revenue among the clusters analysed.
- Cluster 1 has the highest spend for both Desktop and Mobile devices, indicating it is the best performing cluster in terms of spend.
- Clusters 1 and 3 show a significant difference in ad dimensions (length and width) and their performance metrics. Consider testing different ad sizes to find the most effective dimensions for engagement and clicks
- Cluster 4 has a high number of clicks and a substantial revenue figure. Analyze the characteristics of ads in this cluster to understand what makes them successful and replicate these features in other ads.
- Clusters 0 and 2 have a lower spend-to-revenue ratio compared to others. Evaluate the ROI of each cluster and adjust your ad spend accordingly to maximize profitability.

PCA:

PART 2: PCA FH (FT): Primary census abstract for female headed households excluding institutional households (India & States/UTs - District Level), Scheduled tribes - 2011 PCA for Female Headed Household Excluding Institutional Household. The Indian Census has the reputation of being one of the best in the world. The first Census in India was conducted in the year 1872. This was conducted at different points of time in different parts of the country. In 1881 a Census was taken for the entire country simultaneously. Since then, Census has been conducted every ten years, without a break. Thus, the Census of India 2011 was the fifteenth in this unbroken series since 1872, the seventh after independence and the second census of the third millennium and twenty first century. The census has been uninterruptedly continued despite of several adversities like wars, epidemics, natural calamities, political unrest, etc. The Census of India is conducted under the provisions of the Census Act 1948 and the Census Rules, 1990. The Primary Census Abstract which is important publication of 2011 Census gives basic information on Area, Total Number of Households, Total Population, Scheduled Castes, Scheduled Tribes Population, Population in the age group 0-6, Literates, Main Workers and Marginal Workers classified by the four broad industrial categories, namely, (i) Cultivators, (ii) Agricultural Laborers, (iii) Household Industry Workers, and (iv) Other Workers and also Non-Workers. The characteristics of the Total Population include Scheduled Castes, Scheduled Tribes, Institutional and Houseless Population and are presented by sex and rural-urban residence. Census 2011 covered 35 States/Union Territories, 640 districts, 5,924 sub-districts, 7,935 Towns and 6,40,867 Villages.
The data collected has so many variables thus making it difficult to find useful details without using Data Science Techniques. You are tasked to perform detailed EDA and identify Optimum Principal Components that explains the most variance in data. Use Sklearn only.

- **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 PCA: Define the problem and perform Exploratory Data Analysis

2.1.1 Problem Definition - Check shape, Data types, statistical summary

- Exported necessary libraries like NumPy, Pandas, Seaborn
- Data is read using pd_excel and top 5 head rows are shown below.
- Dataset has 640 rows and 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_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

Fig 27: Data Head

	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	...	

8 rows × 59 columns

Fig 28: Data Statistical Summary

- Dataset has 59 numeric variable and 2 object variables. And there is no null value as shown below.
- There are no duplicates in the dataset
- According to the statistical summary, 50% row represents the median, shows that for many columns, the mean is higher than the median, indicating a right-skewed distribution.

```
<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
```

Fig 29: Data info

2.1.2 Perform an EDA on the data to extract useful insights

- As per below graph, Uttar Pradesh, Madhya Pradesh and Bihar has higher number of area name.
- Considered these 5 variables for EDA: State,LIT_F,LIT_M,TOT_M,TOT_F

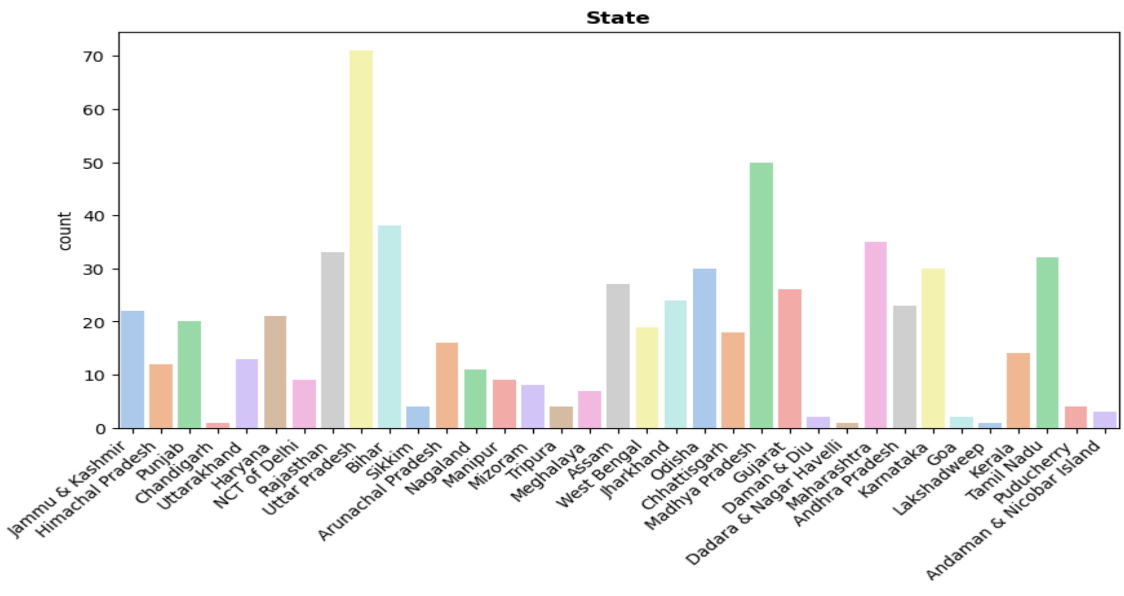


Fig 30: State of India

- Kerala has highest Female and male population, Followed by West Bengal as shown in below graphs

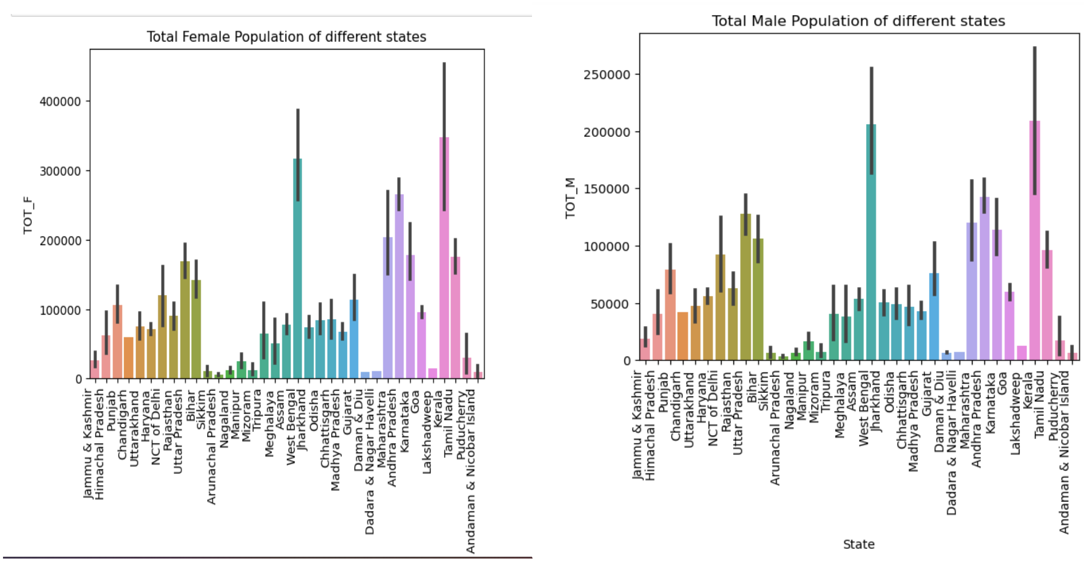
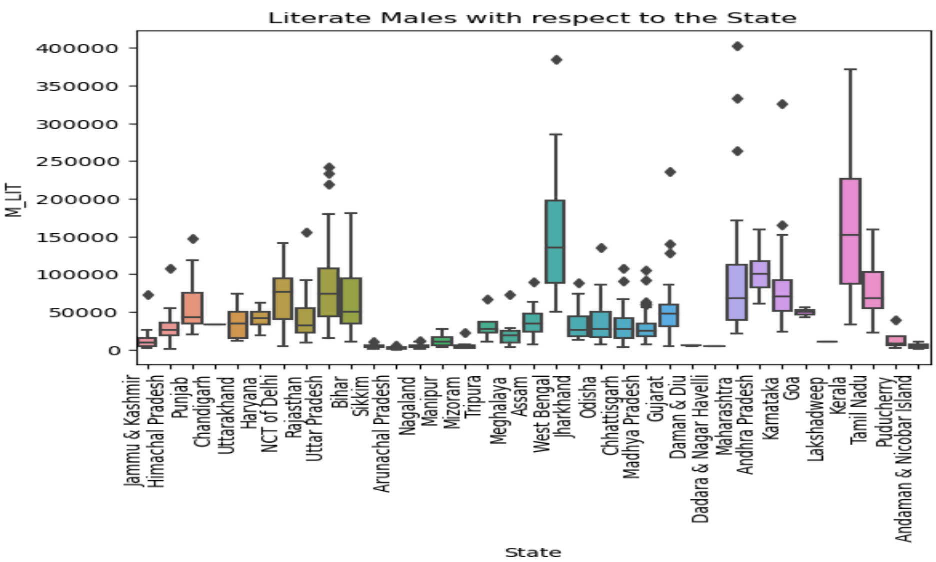


Fig 31: Total Male and Female Population of different states

- Kerala, Maharashtra, and Tamil Nadu were noted for having higher median values which means high number of literate males
- Kerala have highest literate females and Bihar has lowest Literate females.



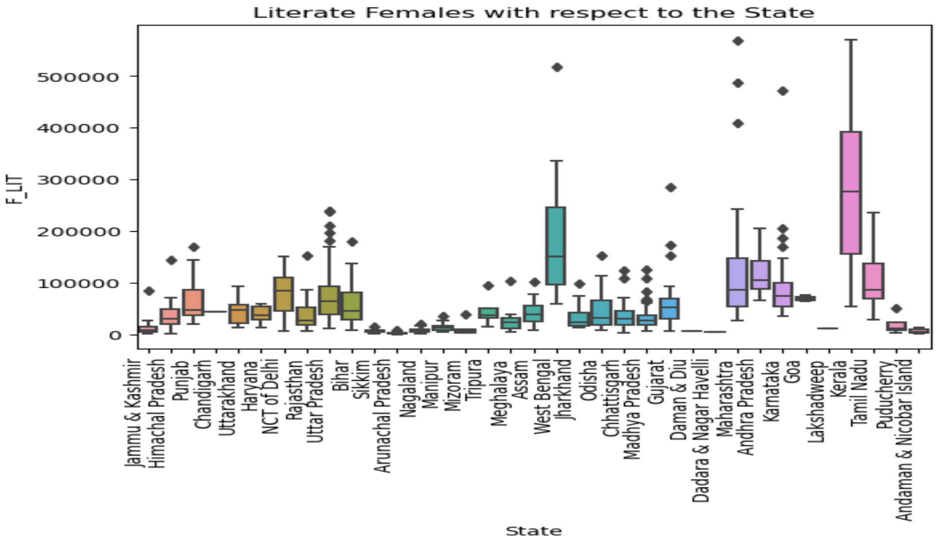


Fig 32: Total Male and Female Literates of different states

- Mumbai Suburban of Maharashtra has highest literate male, followed by North 24 parganas of West Bengal.
- Below is the top 5 literate male grouped by state and area.

```
State        Area Name                 TOT_M
Maharashtra  Mumbai Suburban           485417    403261
West Bengal  North Twenty Four Parganas 471482   384839
Kerala       Malappuram                477790    371829
Maharashtra  Thane                     424759    332986
Karnataka    Bangalore                 401545    325690
Name: M_LIT, dtype: int64
```

Fig 33: Top 5 Literate male grouped by state and area

2.2 PCA: Data Preprocessing

2.2.1 Check for and treat (if needed) missing values

- There are no null values as shown below.

```
State Code       0
Dist.Code        0
State            0
Area Name        0
No_HH            0
                ..
MARG_HH_0_3_F    0
MARG_OT_0_3_M    0
MARG_OT_0_3_F    0
NON_WORK_M       0
NON_WORK_F       0
Length: 61, dtype: int64
```

Fig 34: Null values in dataset

2.2.3 Scale the Data using the z-score method

Before scaling:

- Data is of different scalar units. To make the analysis better, scaling is necessary. Z-score technique is used.

	State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MAF
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	...	

8 rows × 59 columns

Fig 35: Before Scaling

After Scaling:

	State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MA
count	6.400000e+02	640.000000	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	...	
mean	8.881784e-17	0.000000	4.440892e-17	-8.881784e-17	-4.440892e-17	-5.551115e-17	6.661338e-17	5.551115e-18	-5.551115e-17	-4.440892e-17	...	
std	1.000782e+00	1.000782	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	...	
min	-1.710782e+00	-1.729347	-1.057697e+00	-1.084858e+00	-1.071906e+00	-1.066236e+00	-1.050264e+00	-9.587827e-01	-9.570486e-01	-6.251244e-01	...	
25%	-8.614460e-01	-0.864673	-6.598822e-01	-6.779559e-01	-6.682499e-01	-6.591892e-01	-6.423757e-01	-7.183230e-01	-6.989640e-01	-5.954674e-01	...	
50%	9.405736e-02	0.000000	-3.198873e-01	-2.945918e-01	-3.052330e-01	-2.741142e-01	-2.897563e-01	-2.934040e-01	-3.256148e-01	-3.895344e-01	...	
75%	7.310596e-01	0.864673	3.673585e-01	3.815493e-01	3.689451e-01	3.664446e-01	3.498980e-01	3.890923e-01	3.869764e-01	1.480266e-01	...	
max	1.898897e+00	1.729347	5.389586e+00	5.529690e+00	5.532633e+00	7.301993e+00	7.350309e+00	6.207800e+00	6.248040e+00	9.146281e+00	...	

8 rows × 59 columns

Fig 36: After Scaling

2.2.4 Visualize the data before and after scaling and comment on the impact on outliers

Before Scaling: Outliers

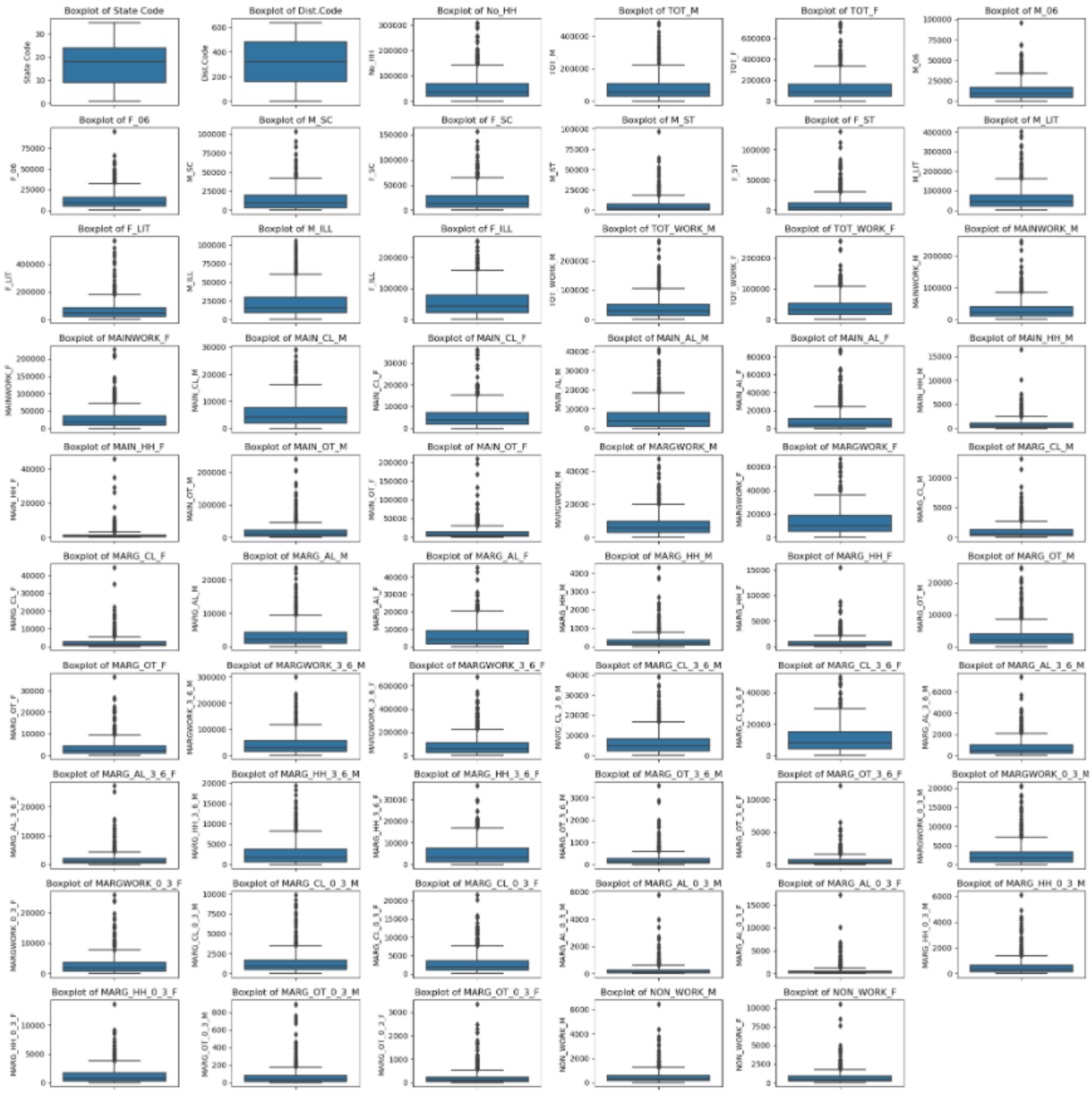


Fig 37: Before Scaling

<u>After Scaling: Outliers</u>

- There is no impact of scaling on the outliers. This could be seen by comparing two pair plot before and after scaling.

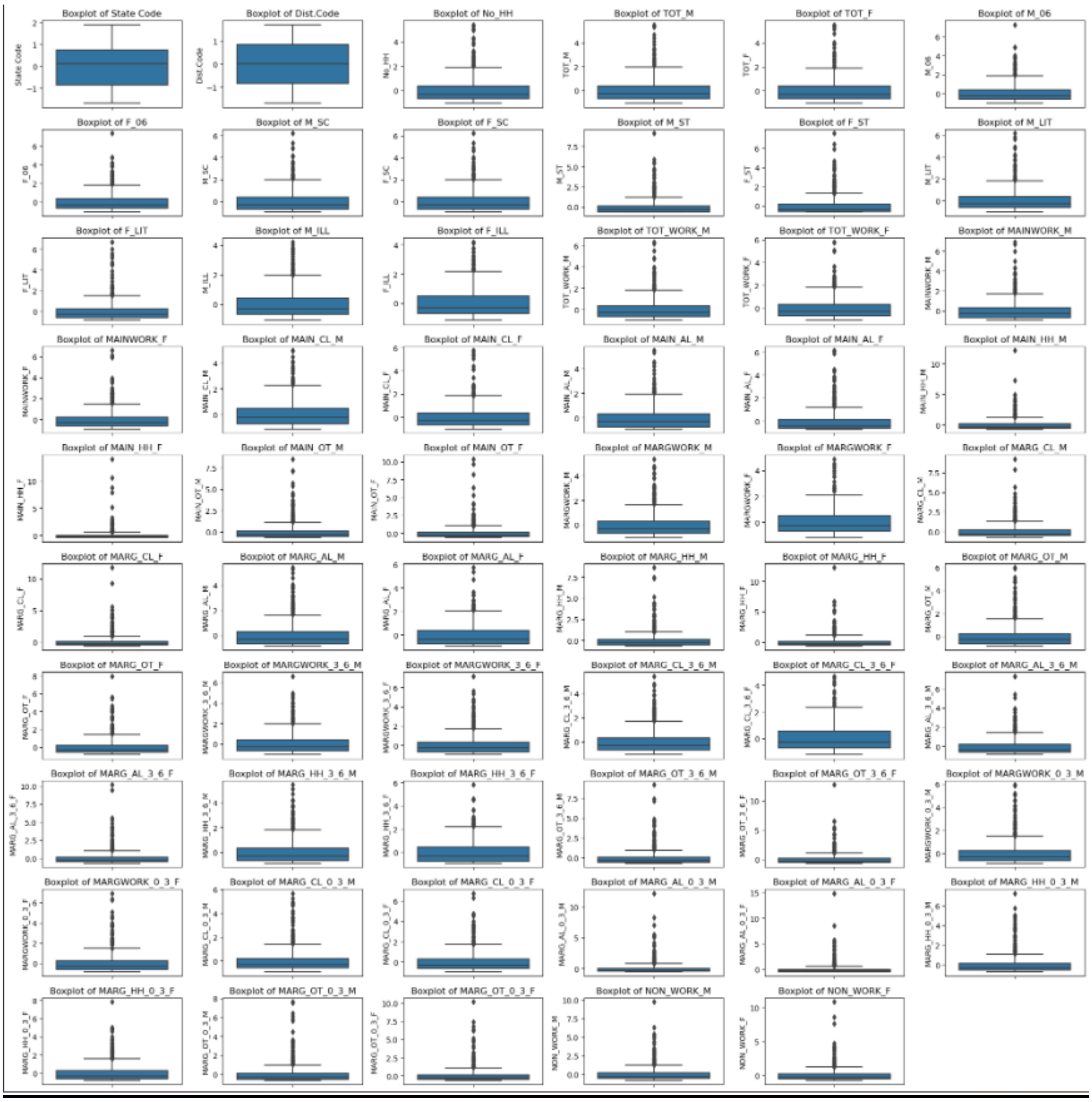


Fig 38: Before Scaling

2.3 PCA: PCA

2.3.1 Create the covariance matrix

- The variance and the relationship between different variables in the dataset are the co-variance matrix. It is shown as heatmap as below.

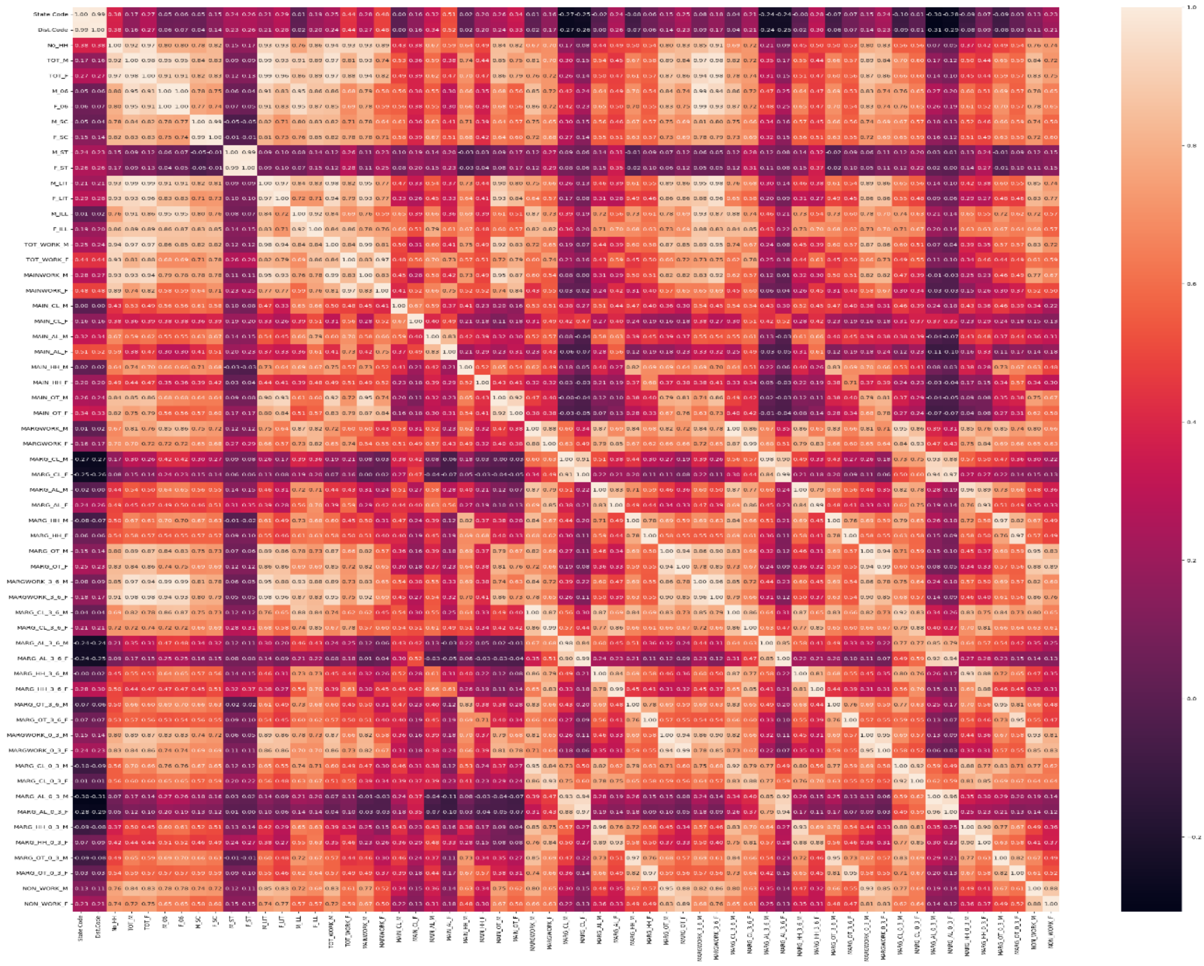


Fig 39: Heatmap of dataset

- Since there is correlation between the variables. To know the significance of correlation, Bartlett sphericity test conducted. If p-value < 0.05, then there is no significance of correlation. For the dataset given, its 0 so, we can proceed with PCA.

```
0.0
```

Fig 40: Bartlett Sphericity

- To confirm the sample adequacy, kmo model technique is used. Above 0.7 is good. For the given dataset, it is 0.8 which confirms the sample adequancy

```
0.8053442139018131
```

Fig 41: Kmo model

2.3.2 Get eigen values and eigen vectors

- No of eigen value = No of PC we have.
- Below is the eigen vectors of the data set.

```
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]]
```

Fig 42: Eigen vectors

```
array([3.18674263e+01, 8.18907061e+00, 4.54275124e+00, 3.84336785e+00,
       2.27105793e+00, 1.95992589e+00, 1.37548006e+00, 8.87342674e-01,
       7.19897963e-01, 6.14059555e-01, 4.94399686e-01, 4.24147991e-01,
       3.43932360e-01, 2.96118628e-01, 2.75961760e-01, 1.84995268e-01,
       1.28846861e-01, 1.11536962e-01, 1.03594789e-01, 9.73429345e-02,
       7.82132546e-02, 5.59614544e-02, 4.44214277e-02, 3.78654873e-02,
       2.96705436e-02, 2.70572400e-02, 2.34417688e-02, 1.43611558e-02,
       1.10964929e-02, 9.28775833e-03, 8.27176626e-03, 7.61344489e-03,
       5.02300148e-03, 4.49943614e-03, 2.51573519e-03, 1.06257176e-03,
       7.11882677e-04, 6.28474170e-30, 6.46518301e-31, 1.64432752e-31,
       1.64432752e-31, 1.64432752e-31, 1.64432752e-31, 1.64432752e-31,
       1.64432752e-31, 1.64432752e-31, 1.64432752e-31, 1.64432752e-31,
       1.64432752e-31, 1.64432752e-31, 1.64432752e-31, 1.64432752e-31,
       1.64432752e-31, 1.64432752e-31, 1.64432752e-31, 1.64432752e-31,
       1.64432752e-31, 1.64432752e-31, 1.64432752e-31])
```

Fig 43: Eigen value

- Below is the explained variance ratio

```
[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.  ]
```

Fig 44: Explained Variance ratio

2.3.3 Identify the optimum number of PCs

- As per rubric, 90% explained variance need to be considered. A
- As per below the cumulative variance in %, 6 PCA can be considered which covers 90% of variance.

```
Cumulative Variance Explained in Percentage: [ 53.93  67.79  75.47  81.98  85.82  89.14  91.47  92.97  94.19  95.22
  96.06  96.78  97.36  97.86  98.33  98.64  98.86  99.05  99.22  99.39
  99.52  99.62  99.69  99.76  99.81  99.85  99.89  99.92  99.93  99.95
  99.96  99.98  99.99  99.99 100.   100.   100.   100.   100.   100.
 100.   100.   100.   100.   100.   100.   100.   100.   100.   100.
 100.   100.   100.   100.   100.   100.   100.   100.   100.  ]
```

Fig 45: Explained Variance in %

2.3.4 Show Scree plot

- As per scree plot, Post 6 PC, the drop is slow. The optimal number of PC would be 6 which results in dimensionality reduction.

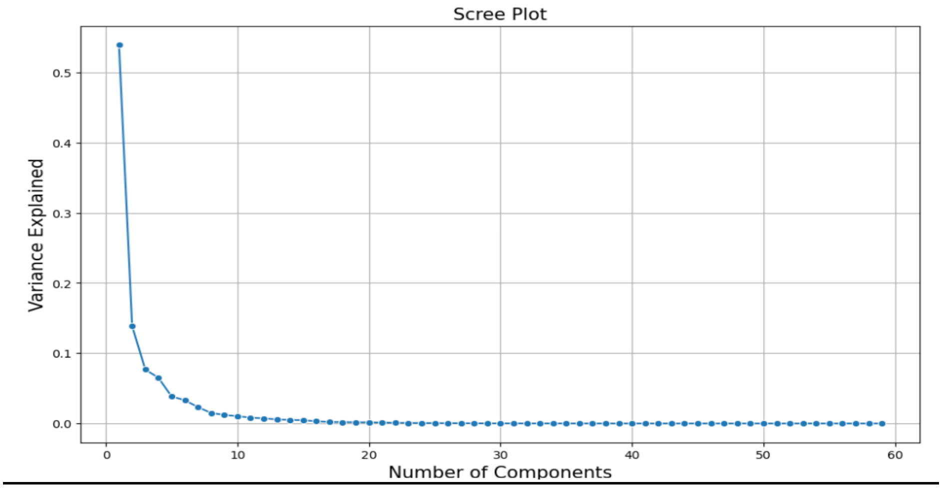


Fig 46: Scree Plot

- Post confirming the no of PC, PCA getting applied to all the features and shape becomes 640 rows and 6 columns as shown below

```
array([[-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],
       [-1.52, -1.94, -0.71, ..., -0.68, -0.42, -0.37],
       [ 0.09, -0.26,  0.15, ...,  0.91,  0.78,  0.85],
       [-0.61,  0.31, -0.02, ...,  0.55,  0.31,  0.25]])
```

```
: (640, 6)
```

Fig 47: Post applying PCA

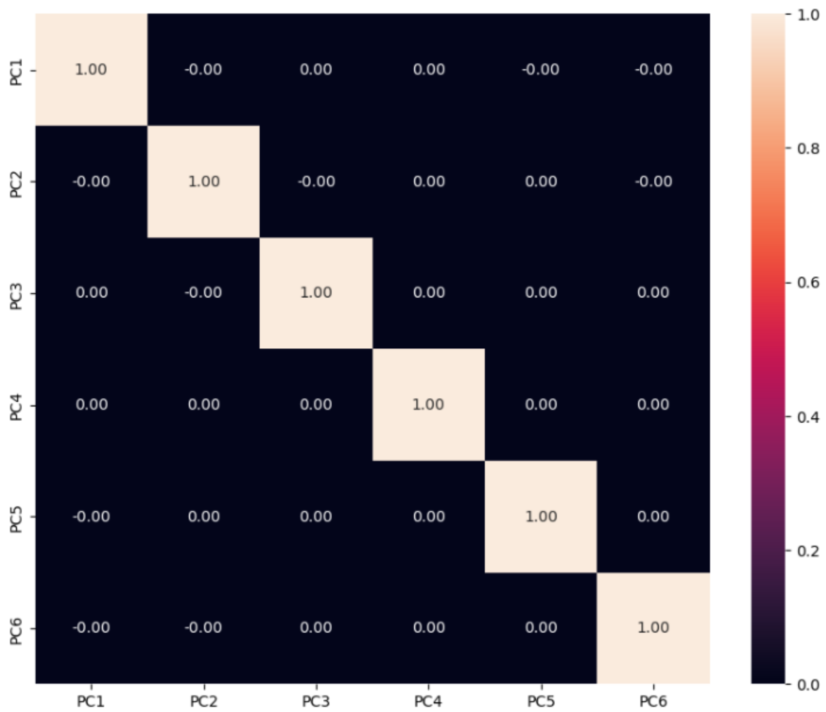


Fig 48: Correlation post PCA

2.3.5 Compare PCs with Actual Columns and identify which is explaining most variance

- PC1 has the highest absolute loading values compared to the other PCs.
- The first bar in the PC1 graph is the tallest among all the first bars in the other PC graphs, which suggests that PC1 accounts for the most variance within the data set.

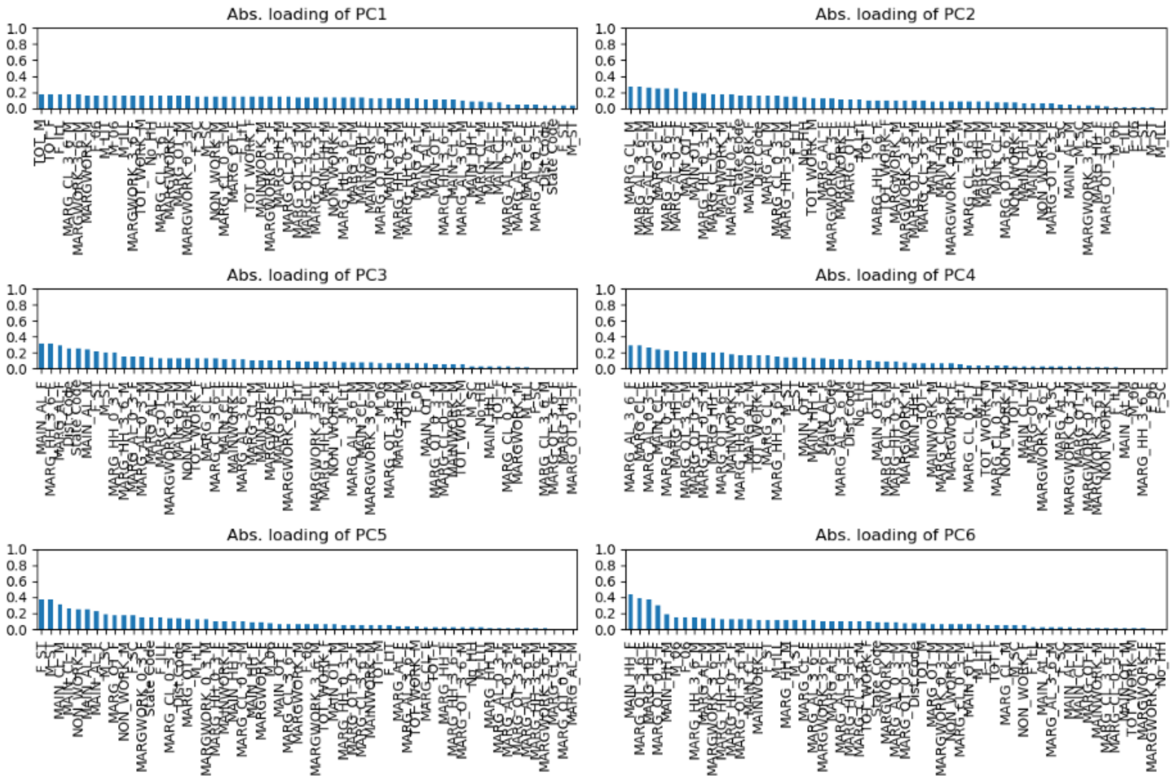


Fig 49: Absolute Loadings of PC's

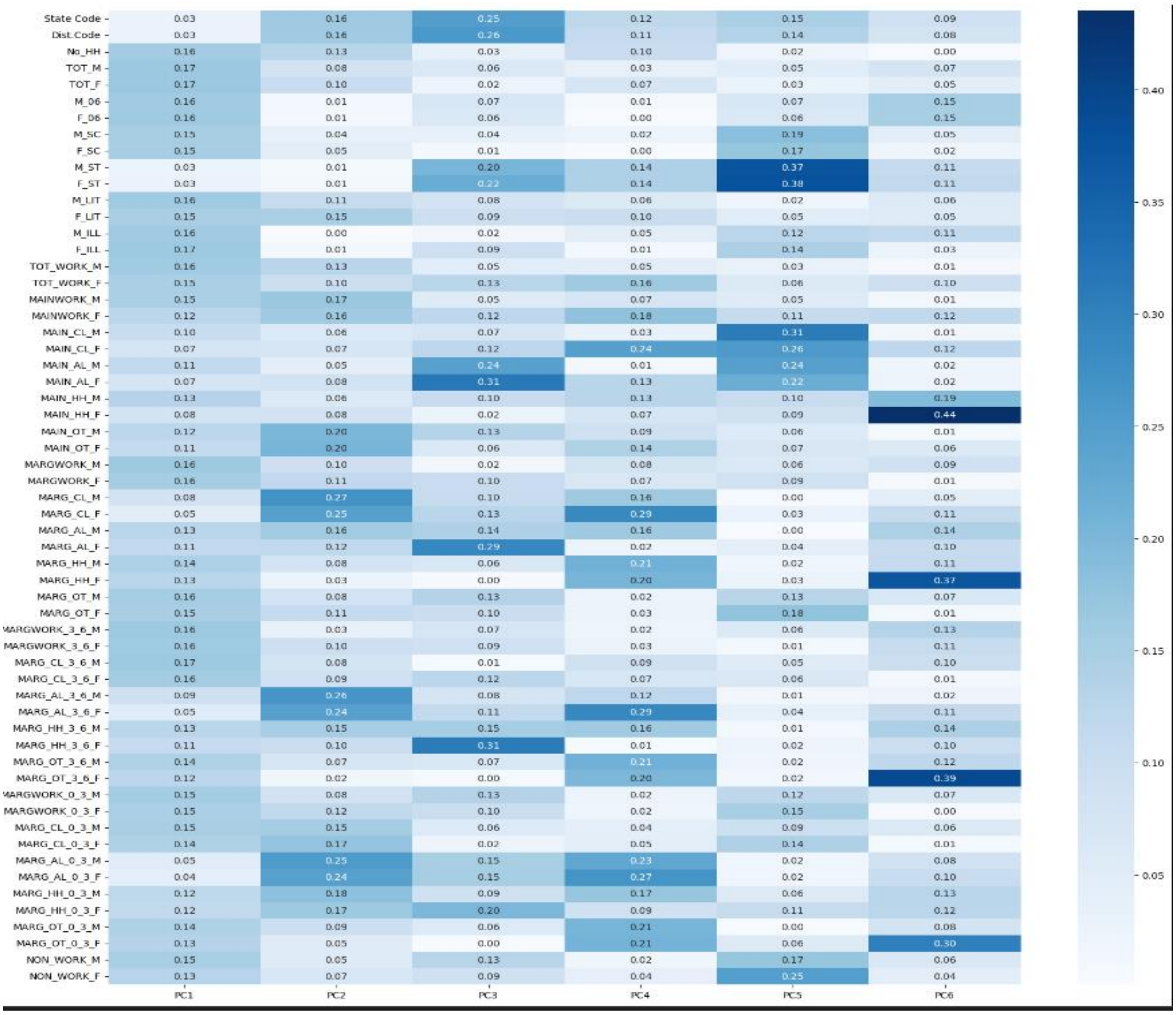


Fig 50: Correlations of PC's with original feature

2.3.7 Write linear equation for first PC

PC 1 = a1x1 + a2x2 + a3X3 + …….+ anxn