Business report on Machine learning

Table of Contents with list of figures

Problem Statement:

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.

Solution:

EDA:

Observations:

- There are 23066 observations and 19 columns in the data.
- All columns except CTR,CPC,CPM have 23066 non-null values i.e. there are no missing values.
- Ad-Type, Inventory type, Platform and device type are columns that are numerical.
- Everything looks great, lets move ahead to check duplicates.
- No Duplicates
- There are 4736 null values for CPM,CTR and CPC.
- After Summary Statistics we can observe

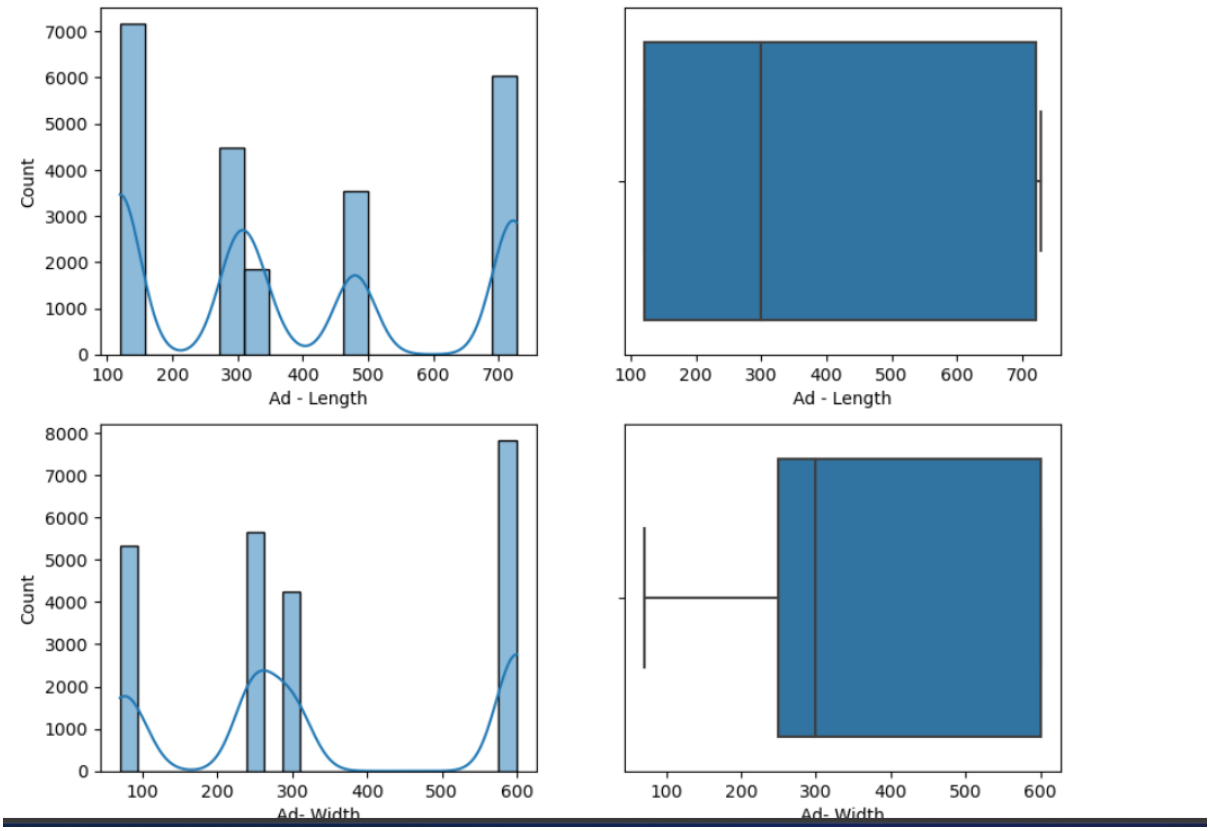
```
1 data_df.describe().T
```

	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	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

- The Length of Ad ranges from 120 to 728. The average length of Ads is 3.85.
- The Average revenue generated is 1.92.
- The Impressions made is 1 to 14194.
- The average spend is 2.7 and max spend is 26931.

Univariant Analysis

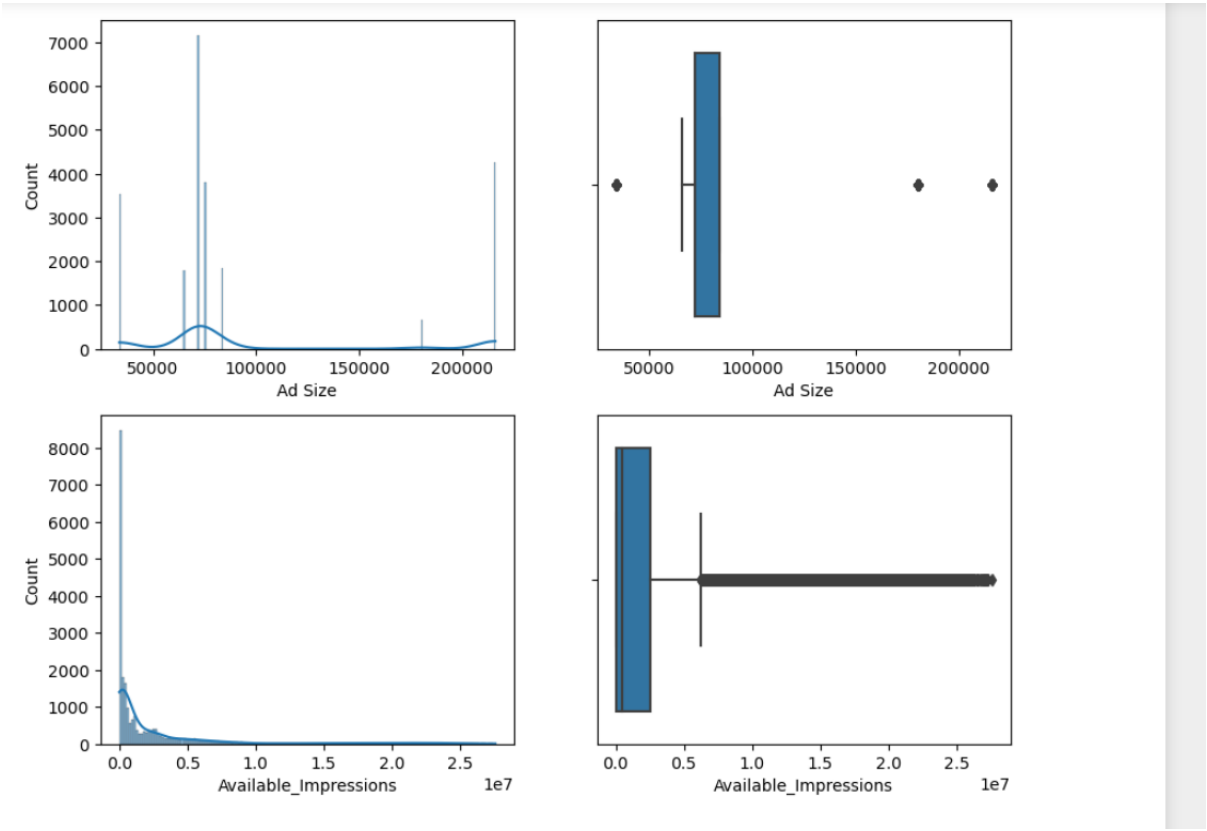
Fig 1.1



Observations

The Ad- lenghth does not have outliers. The max value is 728 and median value is 300. It is not distributed normally. The Ad-Width is left skewed and no outliers. The max value is 600 and median value is 300.

Fig 1.2



Observations

The Ad size and available impression both have outliers treatment with boxplots.

Fig 1.3

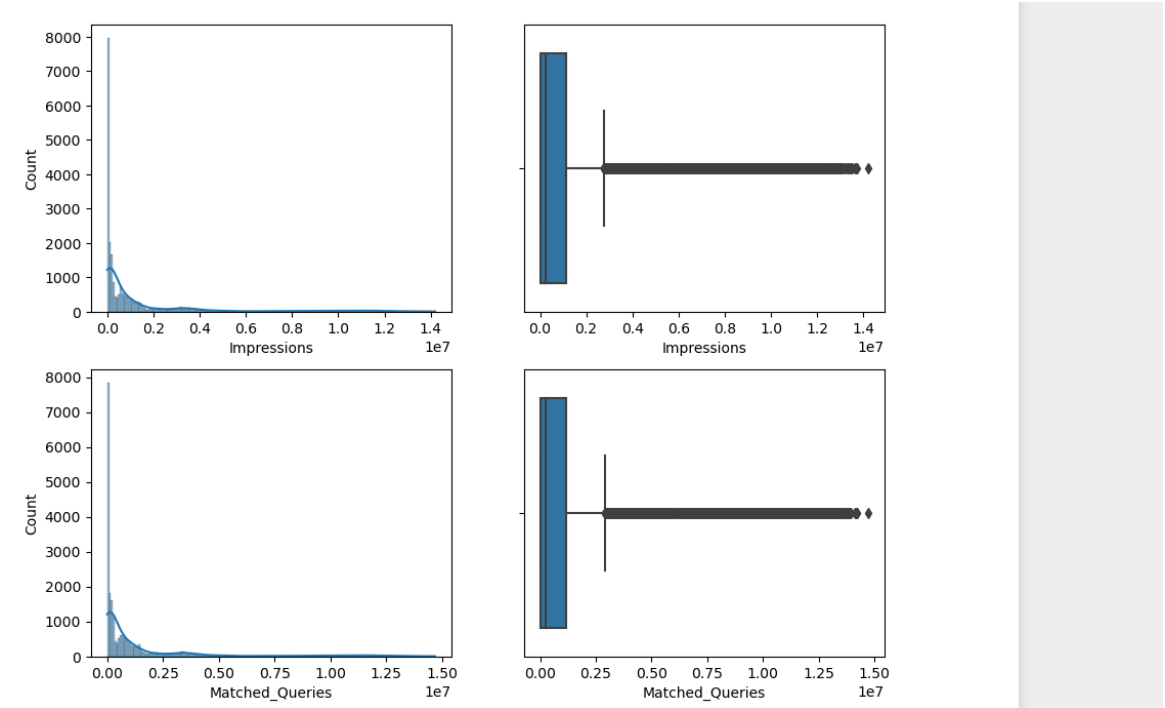
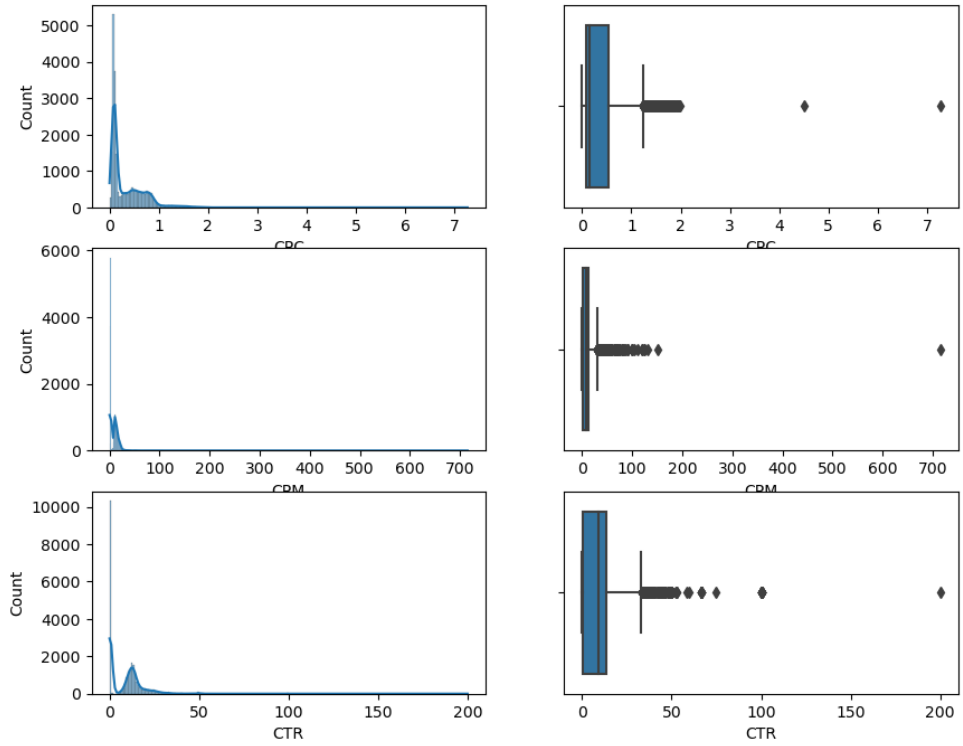


Fig 1.4



Observations:

The CPC and CPM and CTR all have null values we have treated with the formula method. Post which all the nulls are removed.

Later could see that outliners which are treated with box plot techniques.

Categorical variable analysis univariant.

Observations:

1. There are few categorical variables each of which are visualized using count plot.
2. In the Inventory Type the format 4 is seen most peak reaching 7000
3. The Ad type is having all levels same.
4. Device type is mostly used is mobile. Around 14000 of the people uses mobile when compared to desktop.
5. Both the display and video have same level.

Fig 1.6

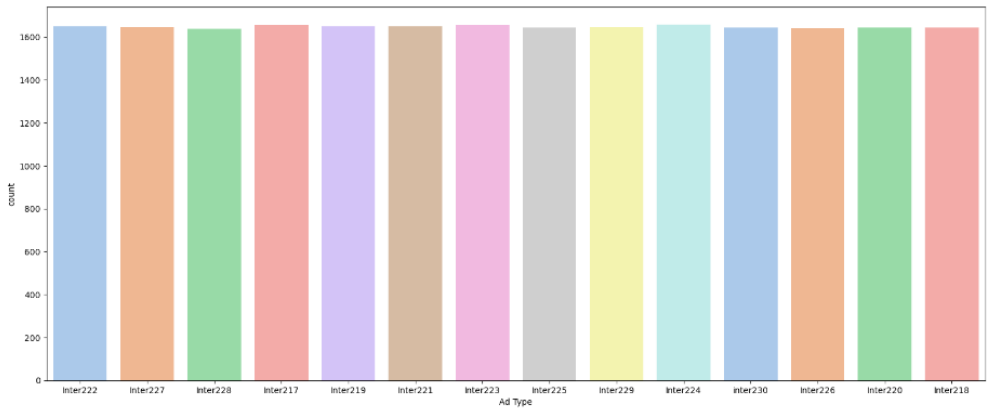


Fig 1.7

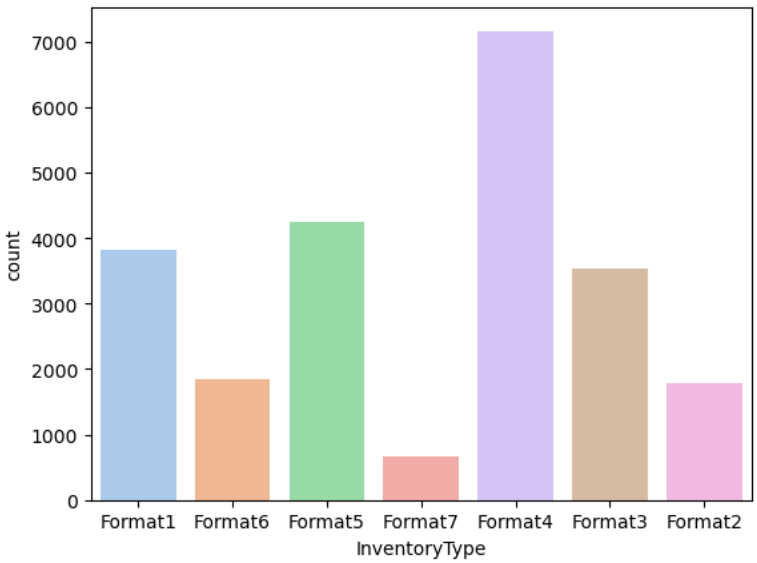


Fig 1.8

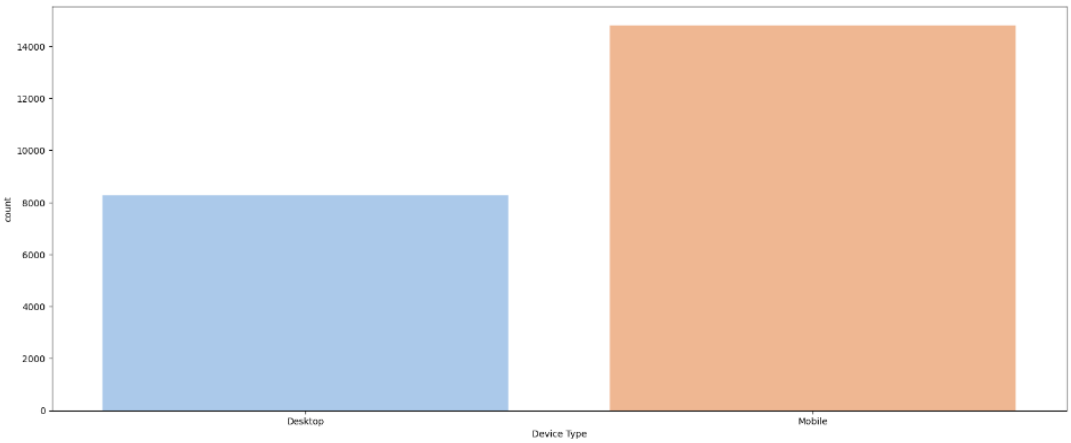
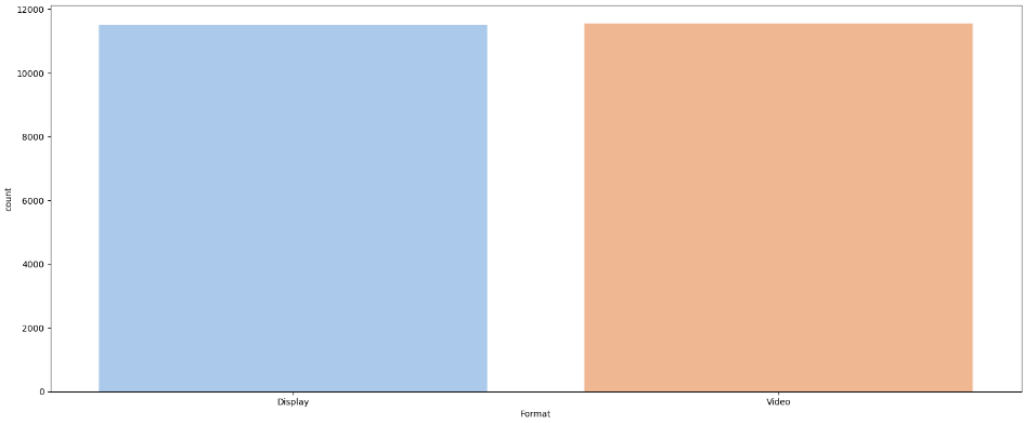


Fig 1.9

```
Out[170]: <Axes: xlabel='Format', ylabel='count'>
```



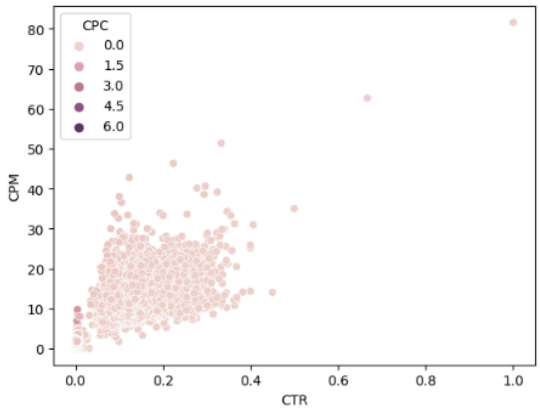
Bivariant Analysis

1. Could see the CTR and CPM is +ve correlated.
2. Ad-Length and Ad-width are not even related. There is no increase in positive way. Its not distributed.

Fig 2.0

```
In [178]:
1 # numeric vs numeric
2 sns.scatterplot(data=data_df_num, x='CTR', y='CPM', hue='CPC')

Out[178]: <Axes: xlabel='CTR', ylabel='CPM'>
```




```
In [174]:
1 sns.scatterplot(data=data_df_num, x='Ad - Length', y='Ad- Width', hue='Ad Size')

Out[174]: <Axes: xlabel='Ad - Length', ylabel='Ad- Width'>
```

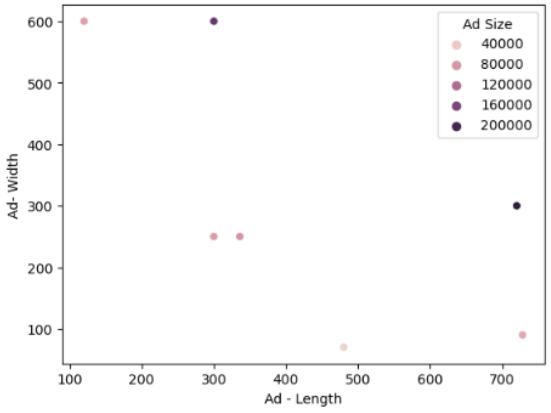
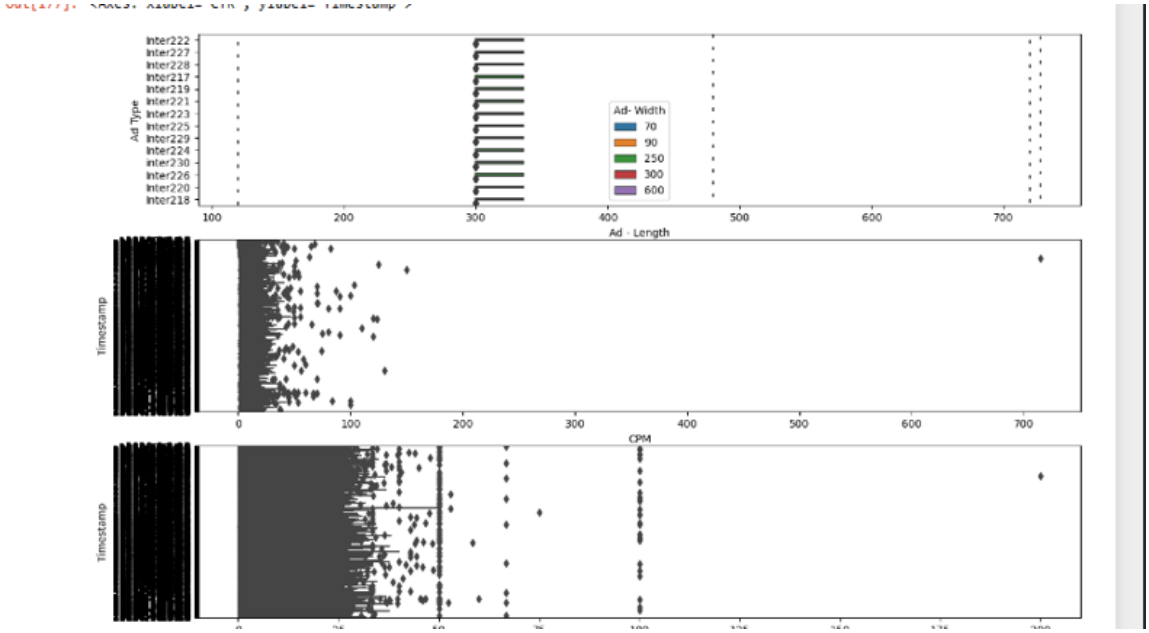


Fig 2.1

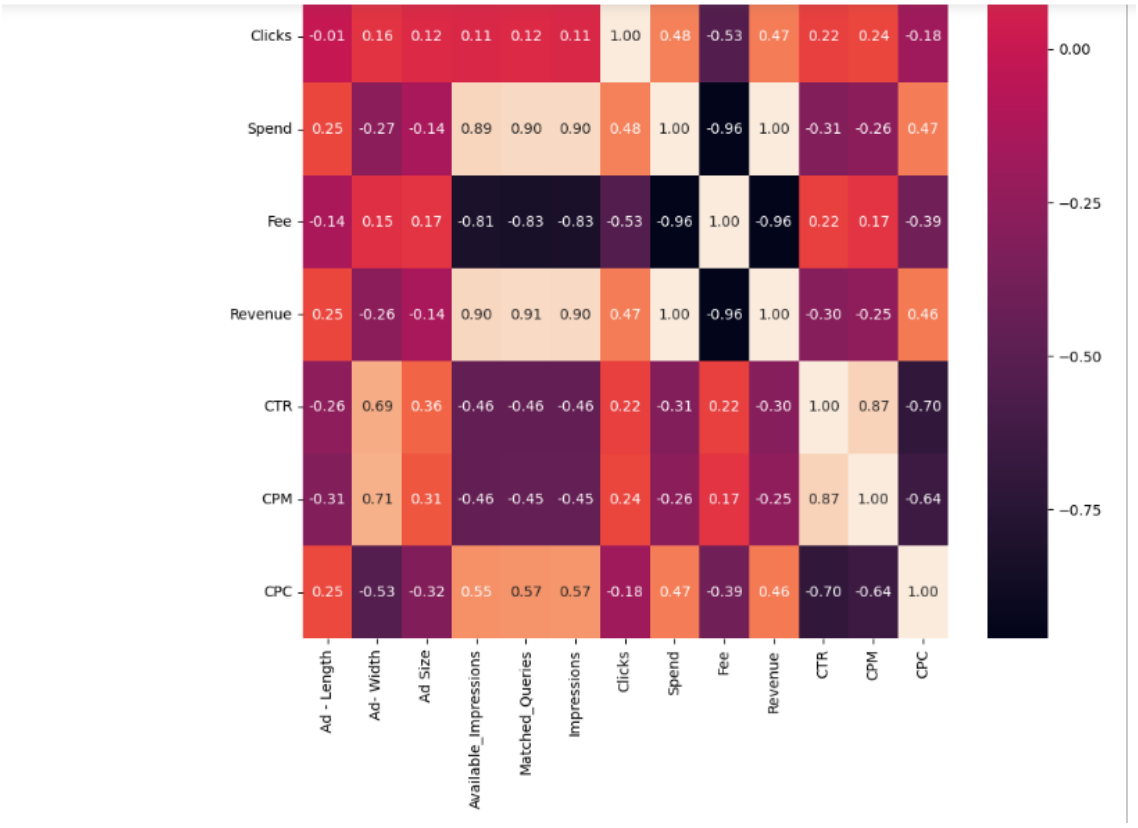
Categorical vs Numerical



Relation between all numeric variables

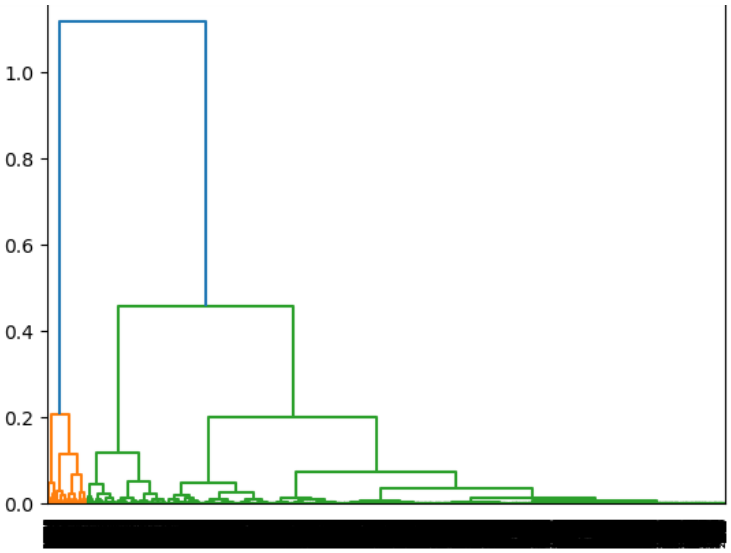
Heatmap shows the correlation between all the numeric variables.

Fig 2.2



Creating the Dendrogram

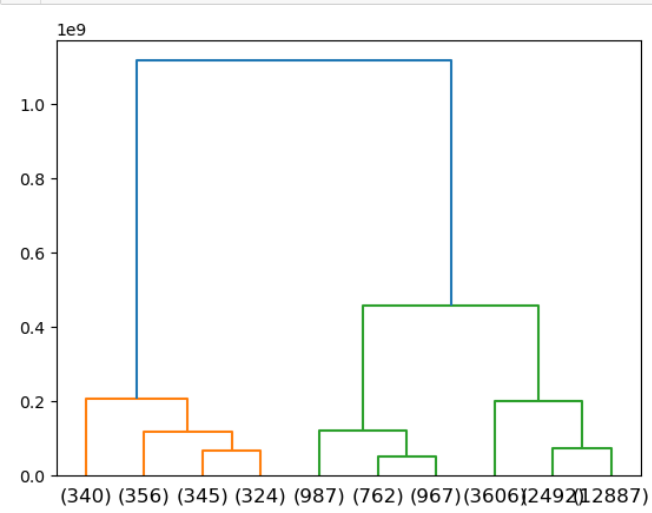
Fig 2.4



link_method_ward:

For creating dendogram using ward method.

Fig 2.5



Overall K means:

1. After the EDA analysis, the Z Score method is applied to enhances the data and set it to equal level. Otherwise, the mean min will vary widely.
2. K means clustering is applied on the scaled data.
3. Fetch the labels and within Cluster Sum of Squares.
4. In order to decide on no of clusters, calculate the number of clusters with k clusters. Using 1,2,3,4,5.. etc. Perform the fit and transform the scaled data.
5. Wss reduces as K keeps increasing.

6. After getting values on wss check how it has reduced or increased based on k mean cluster.
7. Silhouette scores are plotted to finally decide on no of cluster. As and when the score increases the cluster no can be decided. If the silhouette width lies in range between -1 to +1 then all the observations are correctly mapped to clusters.
8. Finalize the no of clusters.
9. Cluster profiling is a powerful tool for exploring and understanding the internal structure of datasets

Clustering: Actionable Insights & Recommendations

Cluster 0 - Medium size cluster with highest Ad-length and width. Total Campaign Spend is on medium side. The impressions are pretty low side. The CPM came with medium range with 12.09. The cost per click is very low. The Ads have reached with good amount of people.

Cluster 1 - Medium size cluster the spend to advertise their company is on the higher side. The revenue generated by the digital marketing is highest with 6373.6598. The Impression made by the advertisements is 2nd highest in the dataset. The CPC is 0.76 is on medium scale. CPM is very low the Ads need to get more money for thousand impressions.

Cluster 2 - It is 2nd largest cluster with max impressions made, it will generate a low revenue when compared to the amount of spend on advertising. It has reached max people but still the revenue is low. Hence, we could improve on the Ad's targeted audience. The CPC is again on low side. The CPM is 14.69 is good. The CTR is again on low side. The Ads need more views and clicks with more display of Ads in appropriate places.

Cluster 3 - It’s the largest cluster with least spent on advertising. Have reached good amount of people. But the revenue generates is very low it’s the least in the dataset. So should improve on the spend to quality of advertising and the company. The CPM is in good state. The Ads to be displayed should suit the age group to generate the max revenue.

Cluster 4 – It’s the smallest cluster. The Ad is shown and clicked by pretty good number. It has reached good amount of people. The CPC is 0.11 which is good for a small cluster size. The Revenue generated is 2nd highest 5017.53, The spend on the Ad is again 2nd highest in dataset. The CPM is also good for a small cluster. The CPM is highest. The Ads have reached 13.75.

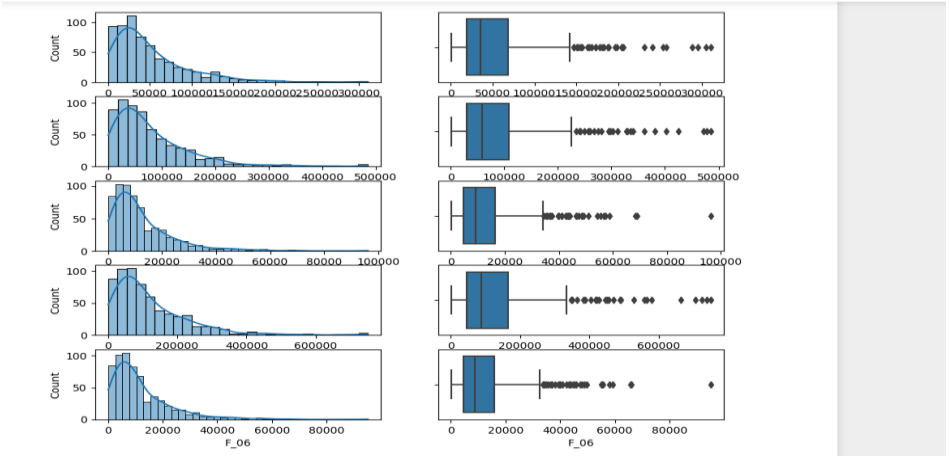
PCA:

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.

Solution:

For checking on outliers and univariant analysis.

Fig 1.1



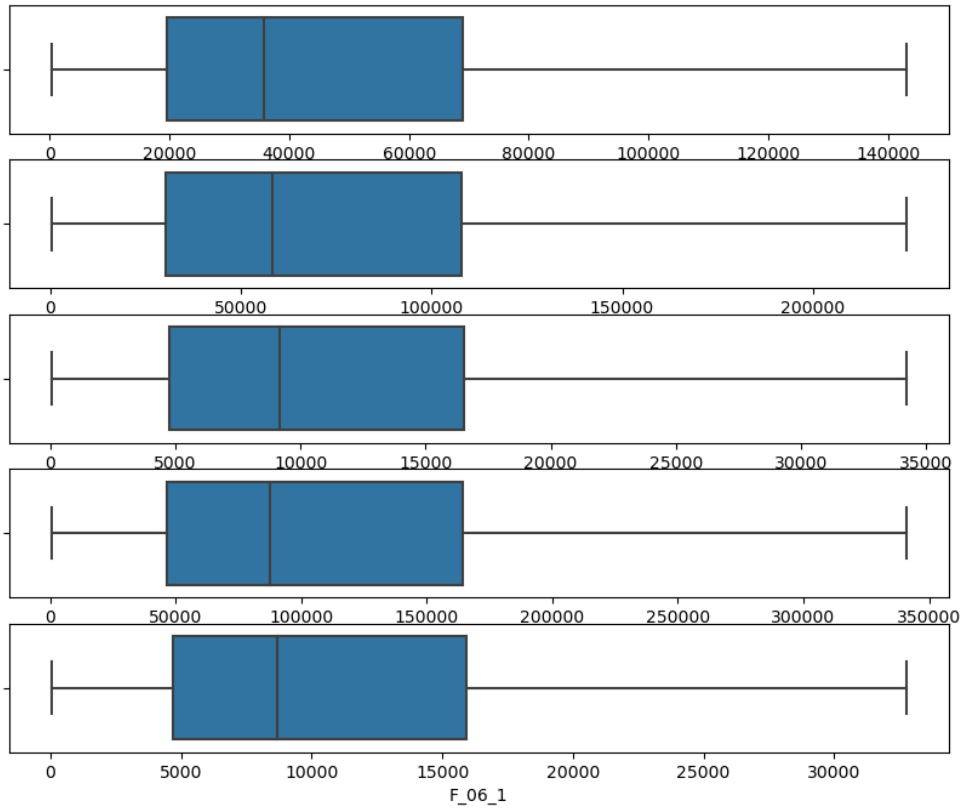
Observations:

1)There are total 640 records present.

2) Most of the variables have outliers. We have visualized few variables and plotted and treated with boxplots.

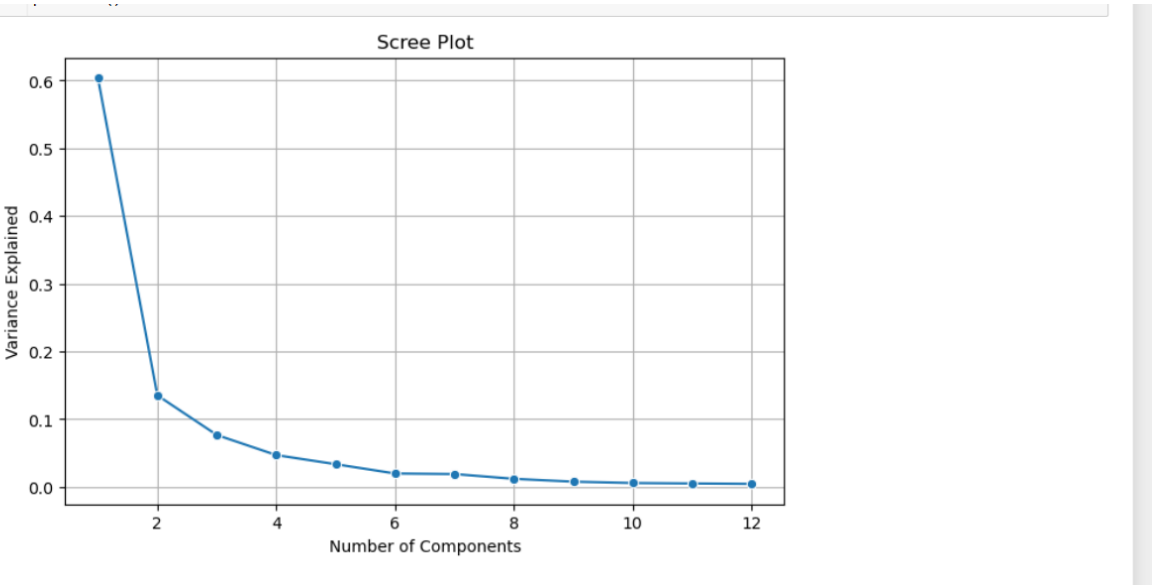
After outliner treatment:

Fig 1.2



Scree plot

Fig 1.3



The original features matter to each PC

Fig 1.4

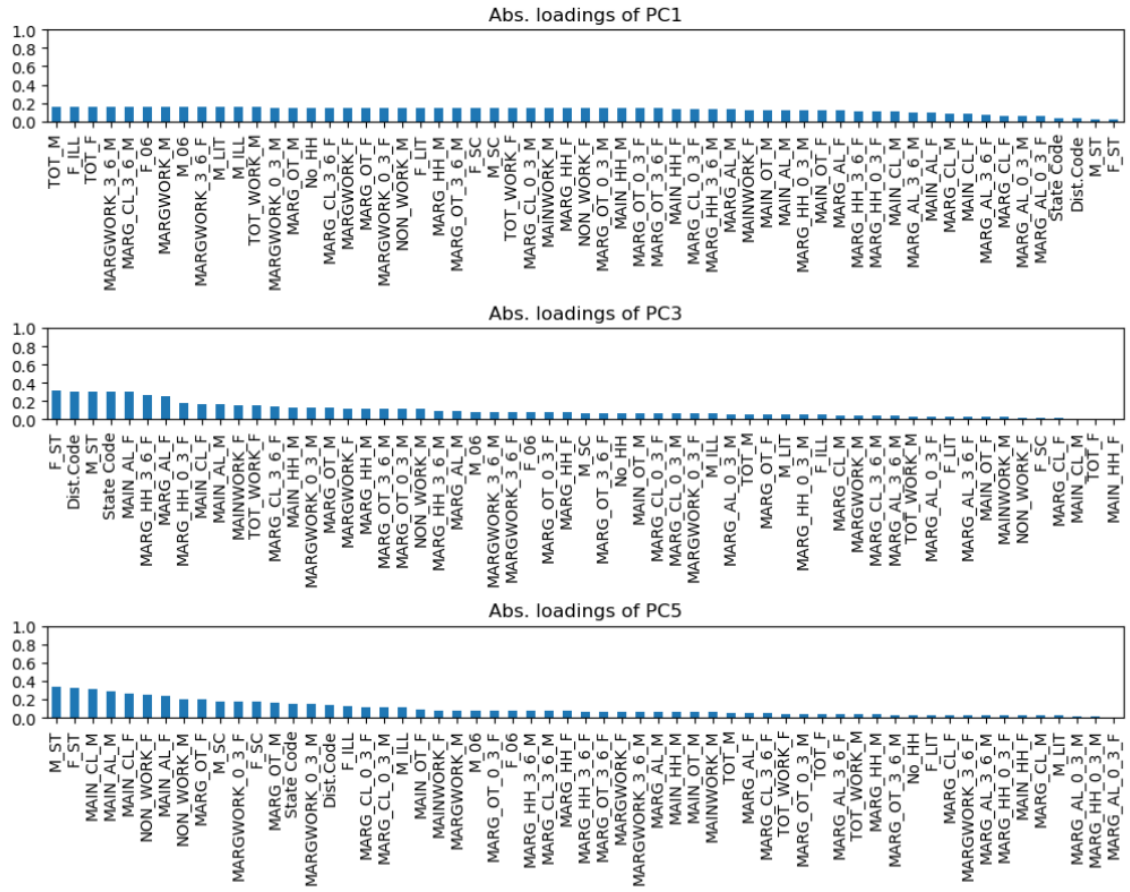
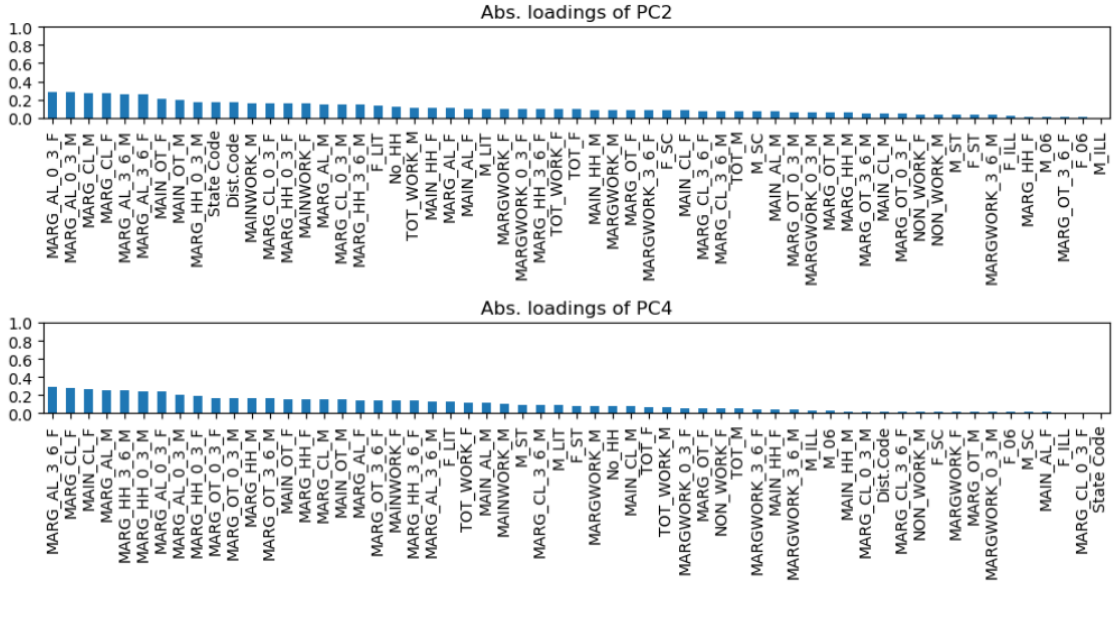
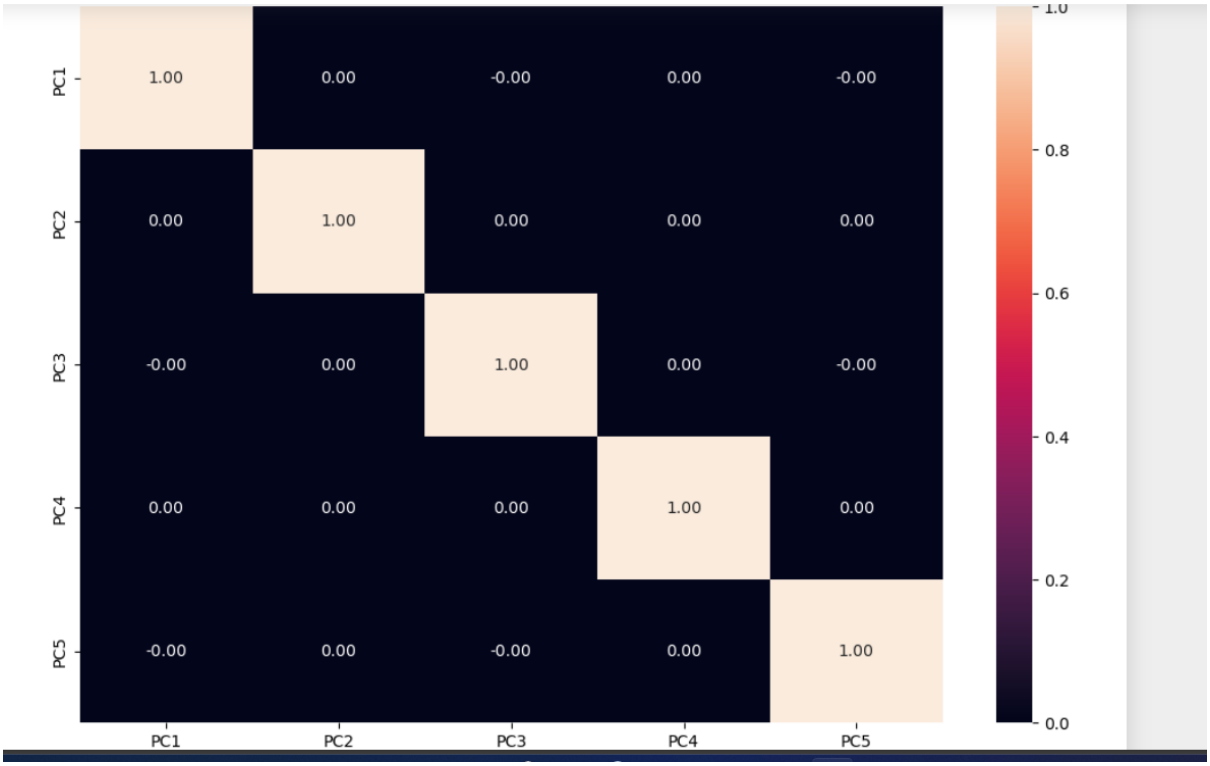


Fig 1.5



Check for presence of correlations among the PCs

Fig 1.6



After all the treatment we can see the only diagonal elements have 1 and rest are 0.

Overall PCA insights

Performed Z scaling technique and scaled data.

Before scaling data was not comparable the mean and median values where far. After Scaling the data seemed closer.

Later got the eigen values and eigen vectors.

The optimum number of PCs would be 5

Scree plot is depicted

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

Conclusion

With help of PCA we have been able to reduce 58 numeric features into 5 components which is able to explain 90% of variance in the data. With help of reduced components, we have been able to observe some patterns. Using some rules around business context. Using the components additional rules can be derived and analysed. Unsupervised learning like clustering can further be applied on the data to segment the customers based on the components created and further analysed.