Data Mining Project
Business Report

Yedupati Venkata Yamini

Table of Contents

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:

1.1 Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc…….. 5

1.2 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. …….7

1.3 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) …….8

1.4 Perform z-score scaling and discuss how it affects the speed of the algorithm……10

1.5 Perform clustering and do the following: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance……. 11

1.6 Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm…….13

1.7 Print silhouette scores for up to 10 clusters and identify optimum number of clusters…. 14

1.8 Profile the ads based on optimum number of clusters using silhouette score and your domain understanding…… 14

[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.]

1.9 Conclude the project by providing summary of your learnings……. 18

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 Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc…… 20

2.2 Perform detailed Exploratory analysis by creating certain questions like (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio? (Example Questions). 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 …… 25

2.3 We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary? ……. 29

2.4 Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment……. 29

2.5 Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector……. 30

2.6 Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot……. 33

2.7 Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables…… 38

2.8 Write linear equation for first PC……... 43

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.

Initial analysis of the data:

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

-> The data set has 23066 rows and 19 columns, which means we have 23066 observations and 19 features in the dataset.

-> There are no duplicated rows in the dataset.

-> Basic information about the data in the dataset looks as below:

```
<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)
```

Insights:

1. Timestamp ideally should have been datetime which is object instead. Need to change datatype of timestamp to datetime if required in further analysis

2. All other data and their datatypes seem to be as aligning to business understanding.

3. There are 4736 null values each in CTR, CPM and CPC columns which need to be filled using the formulae given in the problem statement.

Viewing the summary of data:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Timestamp	23066	2018	2020-11-13-22	13	NaN	NaN	NaN	NaN	NaN	NaN	NaN
InventoryType	23066	7	Format4	7165	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Ad - Length	23066.0	NaN	NaN	NaN	385.163097	233.651434	120.0	120.0	300.0	720.0	728.0
Ad- Width	23066.0	NaN	NaN	NaN	337.896037	203.092885	70.0	250.0	300.0	600.0	600.0
Ad Size	23066.0	NaN	NaN	NaN	96674.468048	61538.329557	33600.0	72000.0	72000.0	84000.0	216000.0
Ad Type	23066	14	Inter224	1658	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Platform	23066	3	Video	9873	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Device Type	23066	2	Mobile	14806	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Format	23066	2	Video	11552	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Available_Impressions	23066.0	NaN	NaN	NaN	2432043.665872	4742887.764666	1.0	33672.25	483771.0	2527711.75	27592861.0
Matched_Queries	23066.0	NaN	NaN	NaN	1295099.143241	2512969.861258	1.0	18282.5	258087.5	1180700.0	14702025.0
Impressions	23066.0	NaN	NaN	NaN	1241519.518859	2429399.961091	1.0	7990.5	225290.0	1112428.5	14194774.0
Clicks	23066.0	NaN	NaN	NaN	10678.518816	17353.409363	1.0	710.0	4425.0	12793.75	143049.0
Spend	23066.0	NaN	NaN	NaN	2706.625689	4067.927273	0.0	85.18	1425.125	3121.4	26931.87
Fee	23066.0	NaN	NaN	NaN	0.335123	0.031963	0.21	0.33	0.35	0.35	0.35
Revenue	23066.0	NaN	NaN	NaN	1924.252331	3105.23841	0.0	55.365375	926.335	2091.33815	21276.18
CTR	18330.0	NaN	NaN	NaN	0.073661	0.07516	0.0001	0.0026	0.08255	0.13	1.0
CPM	18330.0	NaN	NaN	NaN	7.672045	6.481391	0.0	1.71	7.66	12.51	81.56
CPC	18330.0	NaN	NaN	NaN	0.351061	0.343334	0.0	0.09	0.16	0.57	7.26

Insights:

1. Ad-size feature/variable is dependent on Ad-length and Ad-width, a product of the two variables. Mean ad-size is 72000 square units.

2. Matched queries variable is around half of the available impressions variable.

3. Top values of AdType, Platform, DeviceType and Format are Inter224, Video, Mobile, Video respectively.

4. There is considerable difference between 25% of revenue, 50% of revenue and 75% of revenue and the max revenue values.

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

-> 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.

Treating the missing values of 3 columns using above 3 formulas, the summary of the resultant dataset would be:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Timestamp	23066	2018	2020-11-13-22	13	NaN	NaN	NaN	NaN	NaN	NaN	NaN
InventoryType	23066	7	Format4	7165	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Ad - Length	23066.0	NaN	NaN	NaN	385.163097	233.651434	120.0	120.0	300.0	720.0	728.0
Ad- Width	23066.0	NaN	NaN	NaN	337.896037	203.092885	70.0	250.0	300.0	600.0	600.0
Ad Size	23066.0	NaN	NaN	NaN	96674.468048	61538.329557	33600.0	72000.0	72000.0	84000.0	216000.0
Ad Type	23066	14	Inter224	1658	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Platform	23066	3	Video	9873	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Device Type	23066	2	Mobile	14806	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Format	23066	2	Video	11552	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Available_Impressions	23066.0	NaN	NaN	NaN	2432043.665872	4742887.764666	1.0	33672.25	483771.0	2527711.75	27592861.0
Matched_Queries	23066.0	NaN	NaN	NaN	1295099.143241	2512969.861258	1.0	18282.5	258087.5	1180700.0	14702025.0
Impressions	23066.0	NaN	NaN	NaN	1241519.518859	2429399.961091	1.0	7990.5	225290.0	1112428.5	14194774.0
Clicks	23066.0	NaN	NaN	NaN	10678.518816	17353.409363	1.0	710.0	4425.0	12793.75	143049.0
Spend	23066.0	NaN	NaN	NaN	2706.625689	4067.927273	0.0	85.18	1425.125	3121.4	26931.87
Fee	23066.0	NaN	NaN	NaN	0.335123	0.031963	0.21	0.33	0.35	0.35	0.35
Revenue	23066.0	NaN	NaN	NaN	1924.252331	3105.23841	0.0	55.365375	926.335	2091.33815	21276.18
CTR	23066.0	NaN	NaN	NaN	2.614863	7.853405	0.0001	0.0034	0.11265	0.183778	200.0
CPM	23066.0	NaN	NaN	NaN	8.39673	9.057082	0.0	1.75	8.370742	13.04	715.0
CPC	23066.0	NaN	NaN	NaN	0.336652	0.341231	0.0	0.09	0.14	0.55	7.26

Insights on the last 3 observations, CTR, CPM, CPC:

1. Median value of CPM is 8.3. 75% value is 13.04 whereas the max value is 715.0. Spend per impressions is around tens for75% of the ads but the max value 715 indicates that spend is pretty much higher than impressions for some ads.

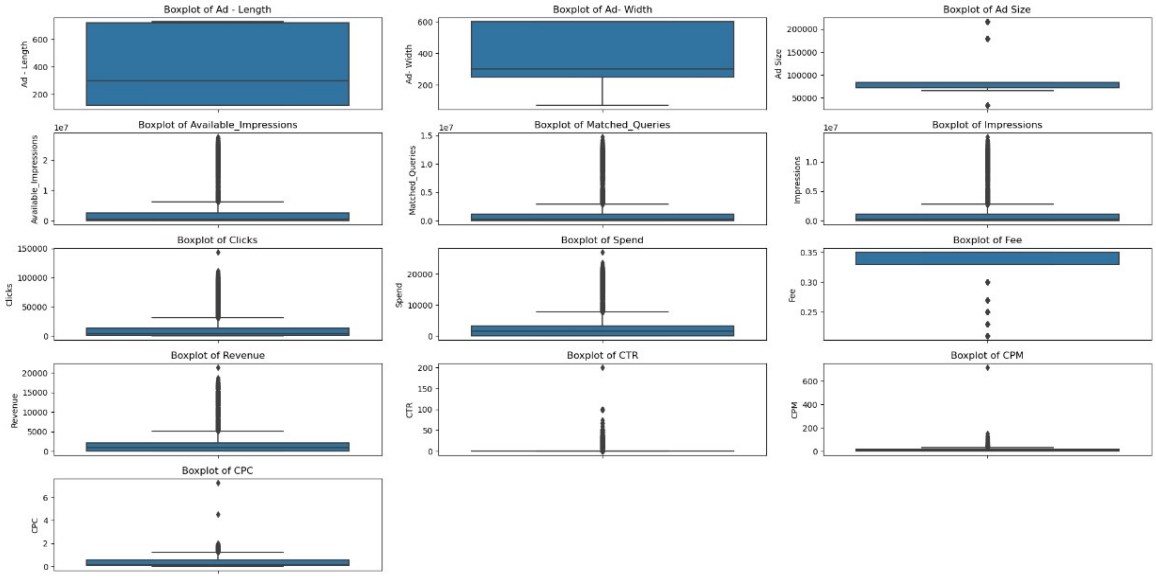
2. Median value for CTR is 0.1126 and max value is 200. Most of the times, clicks are less than ad impressions. For certain ads, clicks are greater than impressions.

3. Median value for CPC is 0.14 whereas max value is 7.26. Spend is less than clicks for considerable number of ads but for certain ads expenditure is more but retrieved less clicks.

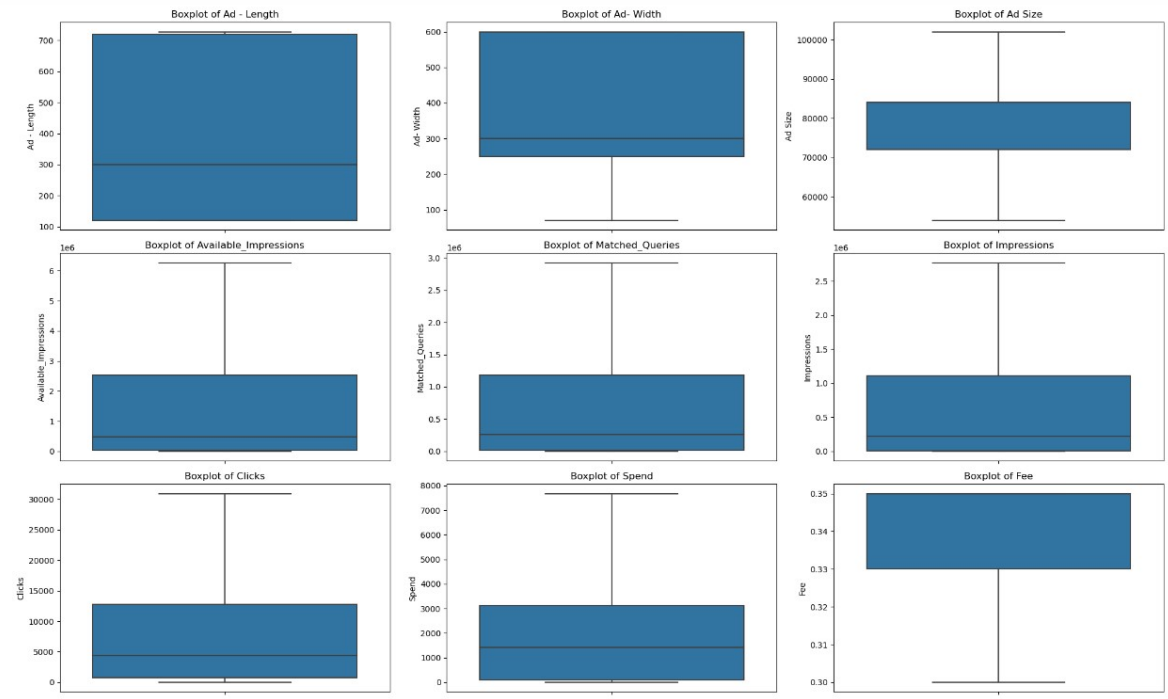
1.3 Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an analyst your judgement may be different from another analyst).

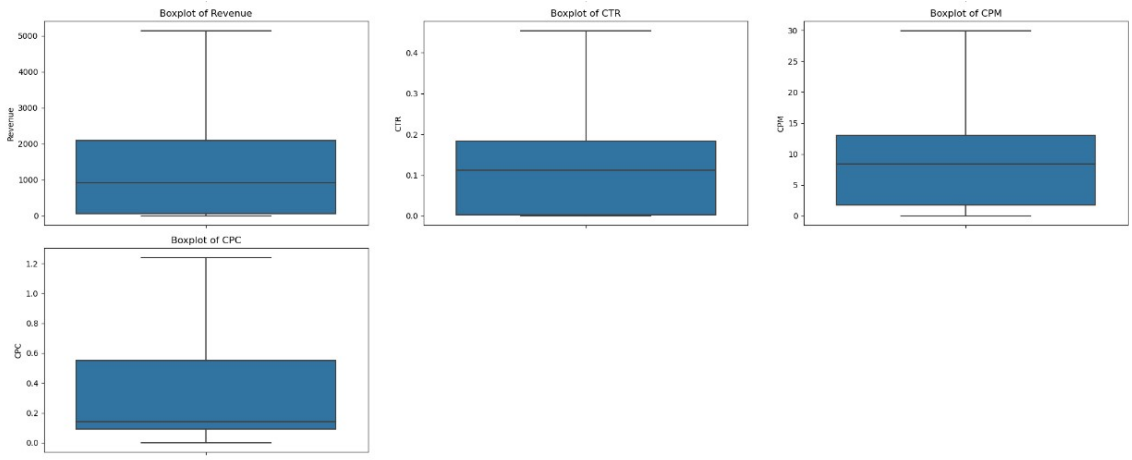
Ans: Below is the grid showing boxplots of all numeric variables present in the dataset:

From the box plots, it is evident that Ad-size, Available_impressions, Matched_Queries, Impressions, Clicks, Spend, Fee, Revenue, CTR, CPM, CPC have outliers.



Treating outliers is necessary for k-means clustering as it is a distance-based algorithm and not treating outliers and considering them may lead to undesired distances. If there are outliers, it's better to treat them. I choose to cap the outliers with lower range and upper range values i.e., Q3+1.5*IQR, Q1 - 1.5*IQR





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

Performed z-score scaling using StandardScaler which normalizes and standardizes the data i.e., the mean of all values of an observation tends to 0 and standard deviation of all values of an observation tend to 1.

Below is the summary of the scaled data. Data in mean and std columns show that the mean and standard deviations of all the variables are close to 0 and 1 respectively.

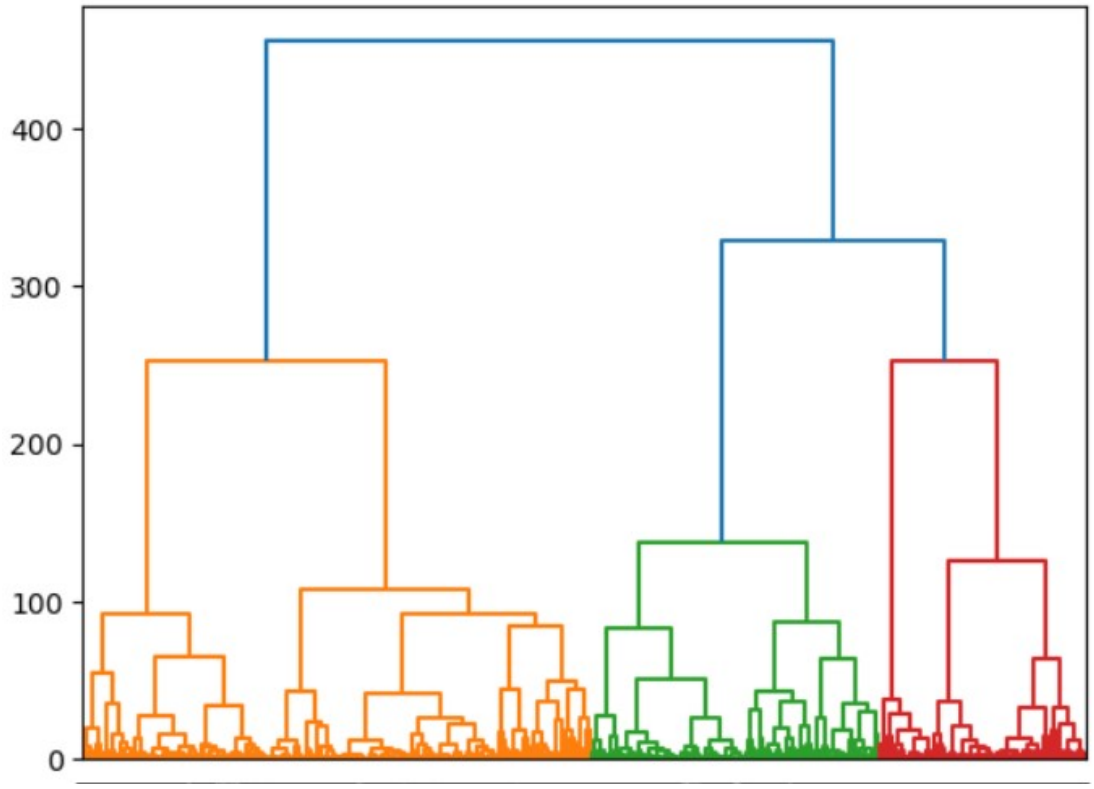
	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	-4.030447e-15	1.000022	-1.134891	-1.134891	-0.364496	1.433093	1.467332
Ad- Width	23066.0	5.390161e-15	1.000022	-1.319110	-0.432797	-0.186599	1.290590	1.290590
Ad Size	23066.0	-4.156304e-15	1.000022	-1.467840	-0.297564	-0.297564	0.482620	1.652896
Available_Impressions	23066.0	-3.617510e-15	1.000022	-0.756182	-0.740341	-0.528577	0.433059	2.193158
Matched_Queries	23066.0	1.341008e-15	1.000022	-0.779265	-0.761447	-0.527722	0.371498	2.070914
Impressions	23066.0	-1.224345e-15	1.000022	-0.768806	-0.760655	-0.538975	0.366051	2.056111
Clicks	23066.0	1.960656e-15	1.000022	-0.867488	-0.793438	-0.405431	0.468629	2.361729
Spend	23066.0	1.250852e-15	1.000022	-0.893170	-0.858046	-0.305523	0.393932	2.271900
Fee	23066.0	-2.322121e-14	1.000022	-2.222416	-0.567532	0.535724	0.535724	0.535724
Revenue	23066.0	3.136228e-15	1.000022	-0.880093	-0.846474	-0.317607	0.389803	2.244218
CTR	23066.0	-2.223858e-14	1.000022	-0.910603	-0.889261	-0.182714	0.277286	2.027108
CPM	23066.0	-6.707353e-16	1.000022	-1.194562	-0.940216	0.022045	0.700677	3.162016
CPC	23066.0	2.787153e-15	1.000022	-1.041140	-0.757396	-0.599760	0.692853	2.868227

Scaling brings the magnitudes of all the observations to a same comparable scale. This makes it easy for the algorithm to compute the distances and hence the speed of the algorithm is increased.

1.5 Perform clustering and do the following: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.

Performing clustering based on ward's linkage method for the above scaled dataset, the resultant dendrogram looks like below:

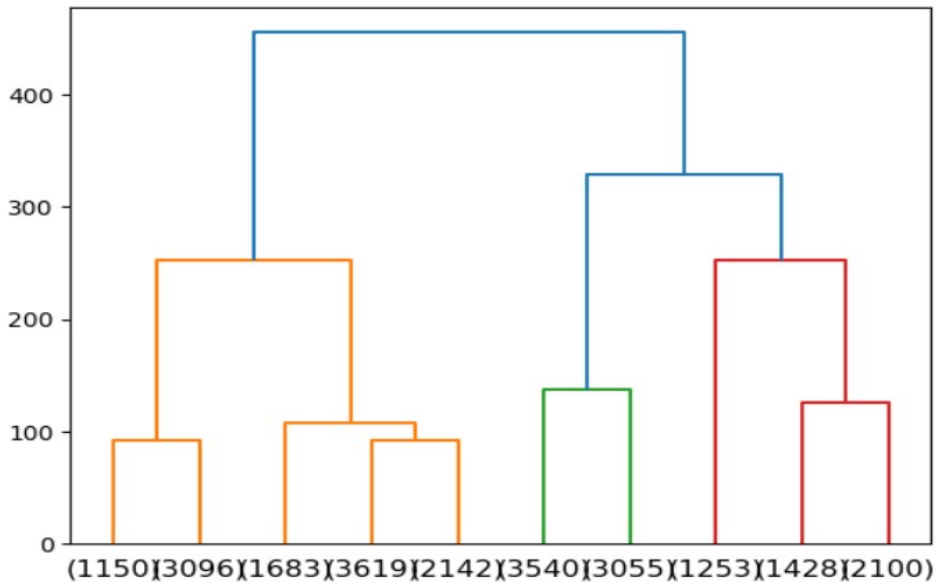
Dendrogram



Since the above dendrogram shows all possible clusters, we can choose ideal number of clusters by assigning a cutoff and filtering the above dendrogram based on the cutoff

Let's consider the cutoff to be 10 i.e., we are forming 10 clusters.

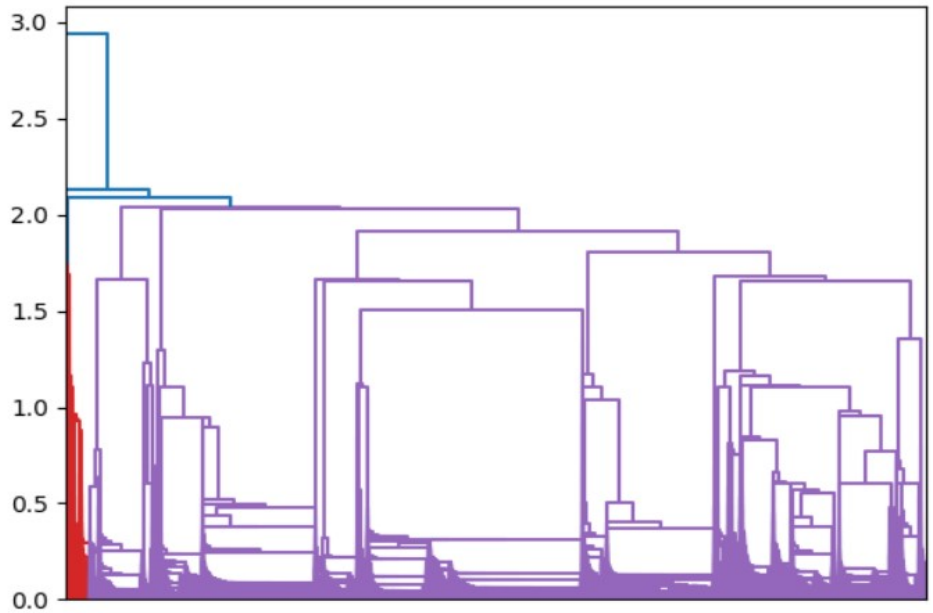
The above dendrogram with 10 clusters looks like below:



The above dendrogram describes the following:

1. Y-axis values of the dendrogram represents distances between each cluster
2. Values in X-axis describe the number of observations in each cluster.

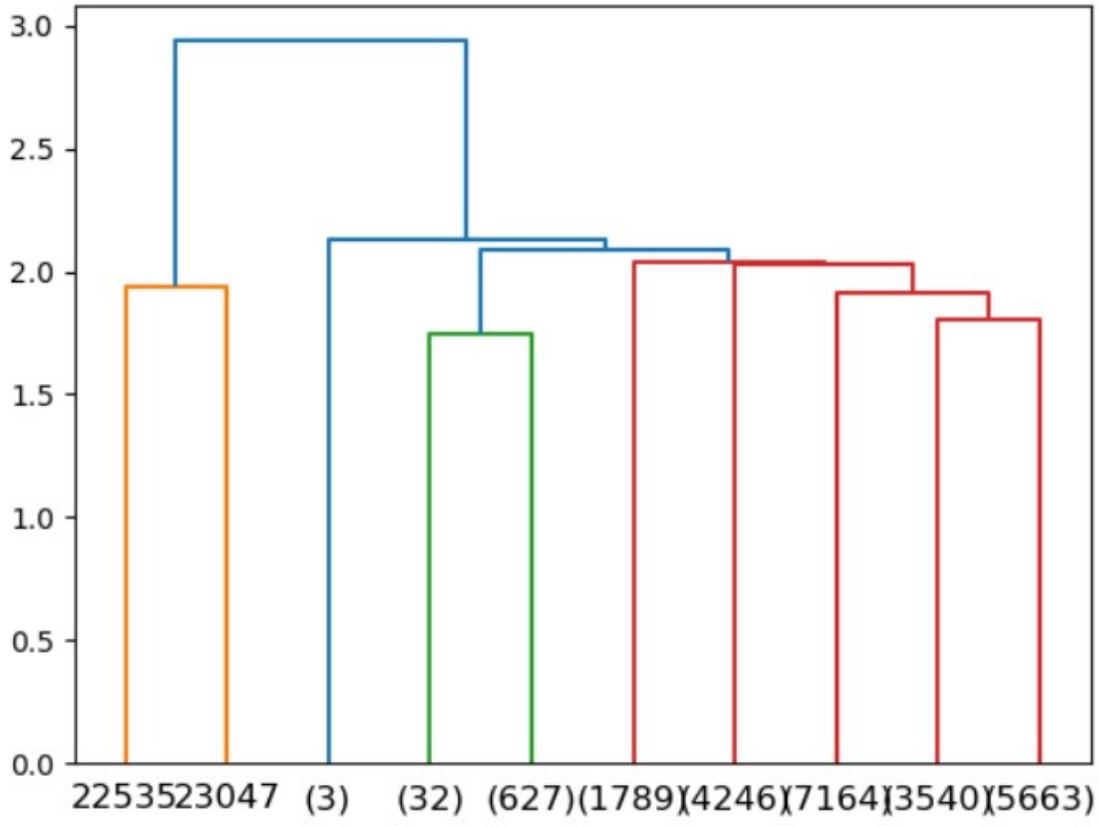
Hierarchical dendrogram using euclidean distance:



Since the above dendrogram shows all possible clusters, we can choose ideal number of clusters by assigning a cutoff and filtering the above dendrogram based on the cutoff

Let's consider the cutoff to be 10 i.e., we are forming 10 clusters.

The above dendrogram with 10 clusters looks like below:



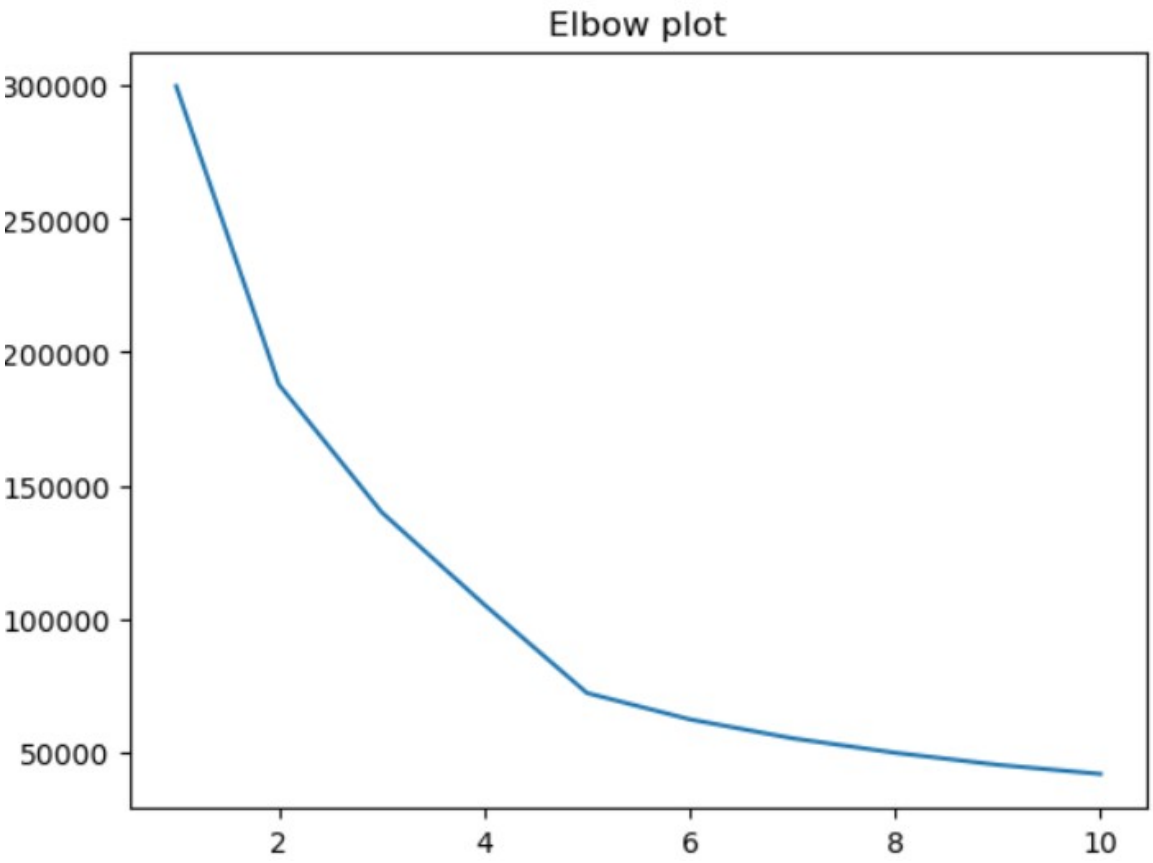
The above dendrogram describes the following:

1. Y-axis values of the dendrogram represents distances between each cluster

2. Values in X-axis describe the number of observations in each cluster.

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

Below is the elbow plot of the inertia values for the 10 clusters.



From the elbow plot it is evident that the sharpest drop is seen at X = 5, that indicates the ideal number of clusters we can choose for k-means clustering is 5.

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

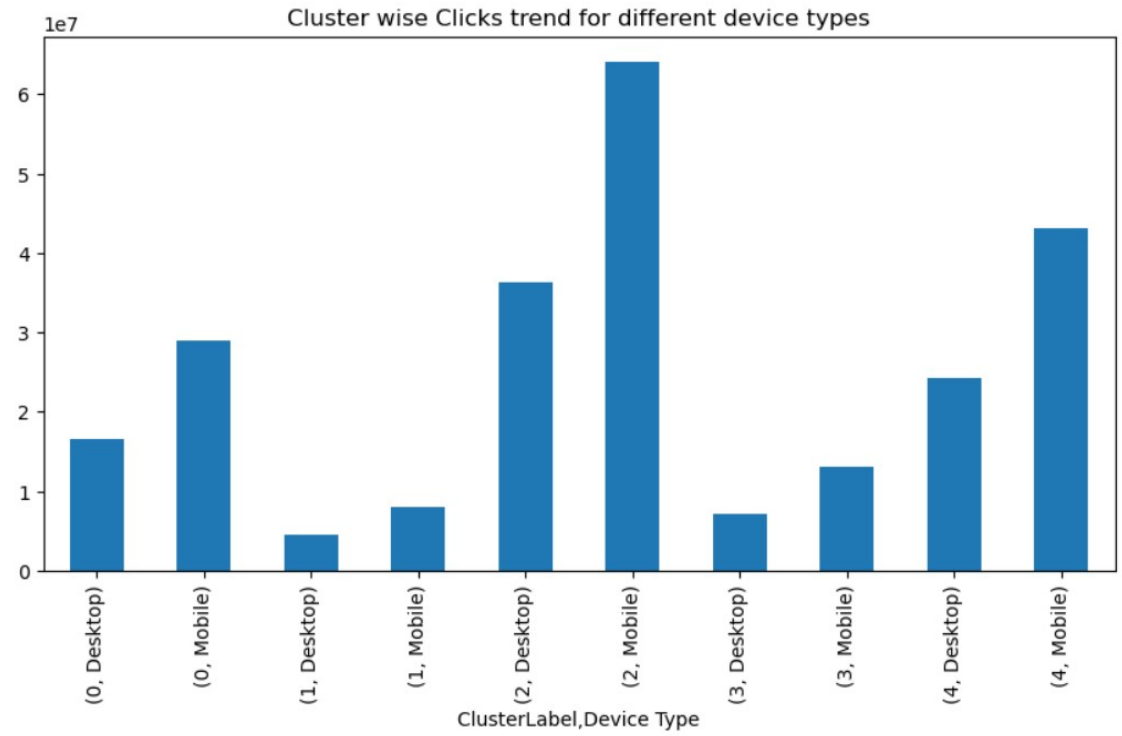
The silhouette scores for 2 to 10 clusters are

[0.4032032120085151, 0.34546476709156715, 0.41284225649057377, 0.48020783078233054, 0.47613811534407546, 0.4688486998487128, 0.4393009079307379, 0.41411135779630426, 0.40988932630125674]

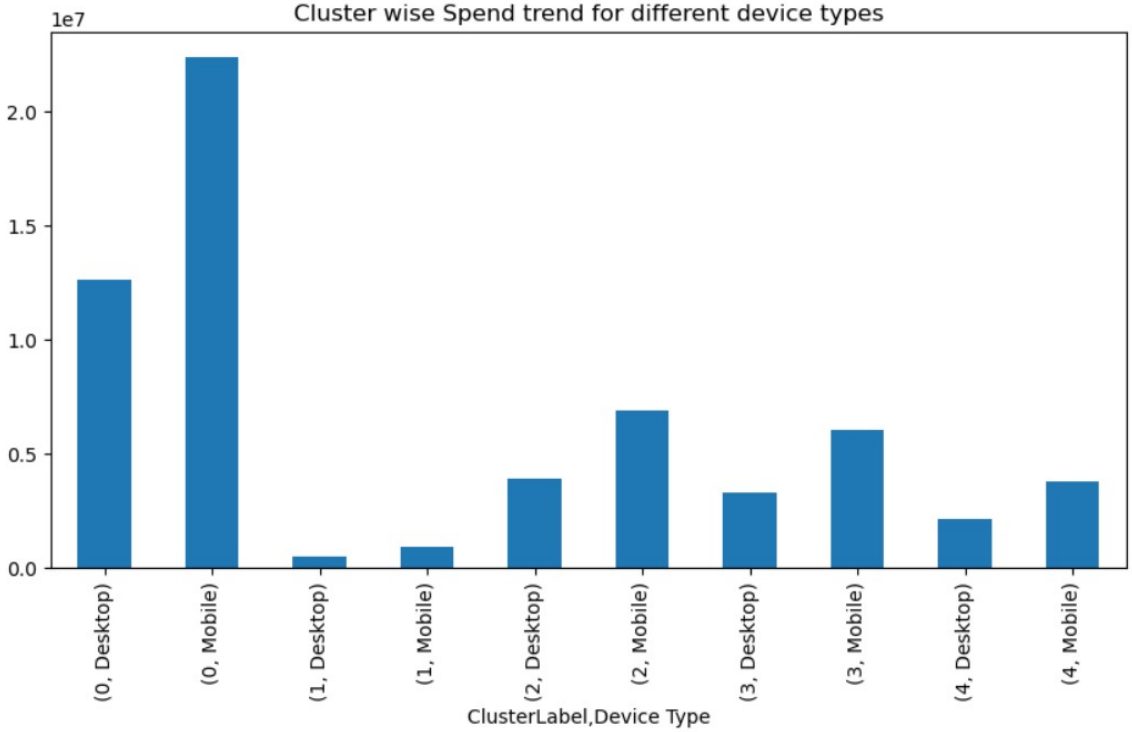
The number of clusters with highest silhouette score is the ideal number of clusters to choose. Higher the silhouette score, higher is the indication that the clusters are well divided. In this case it is 5. So, we choose the number of clusters to be 5

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

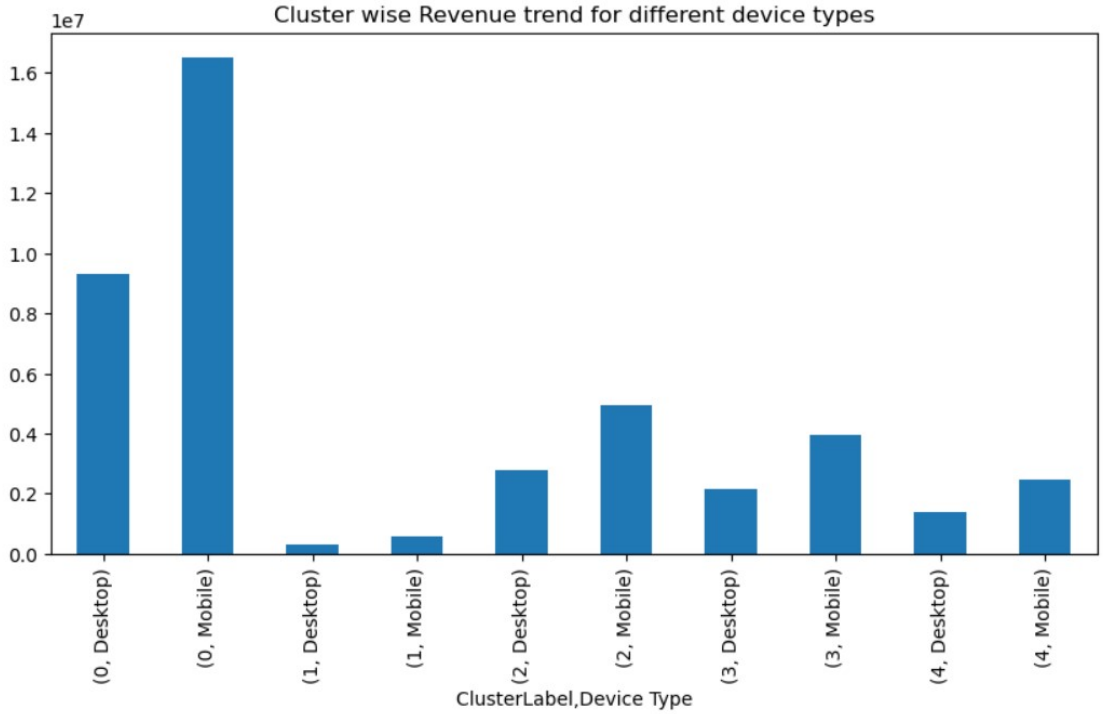
Ans: From the below graph we can identify that cluster 2 on device type mobile has clocked highest number of clicks followed by cluster 4 of type mobile. Other features of this cluster like ad_type, ad_size and inventory can be analyzed to understand the motivating factors for customers to click ads in this cluster.



From the below graph we can identify that cluster 0 on device type mobile has clocked highest number of expenditures both on desktop and mobile. Other features of this cluster like ad_type, ad_size and inventory can be analyzed to understand the motivating factors for customers to click ads in this cluster.

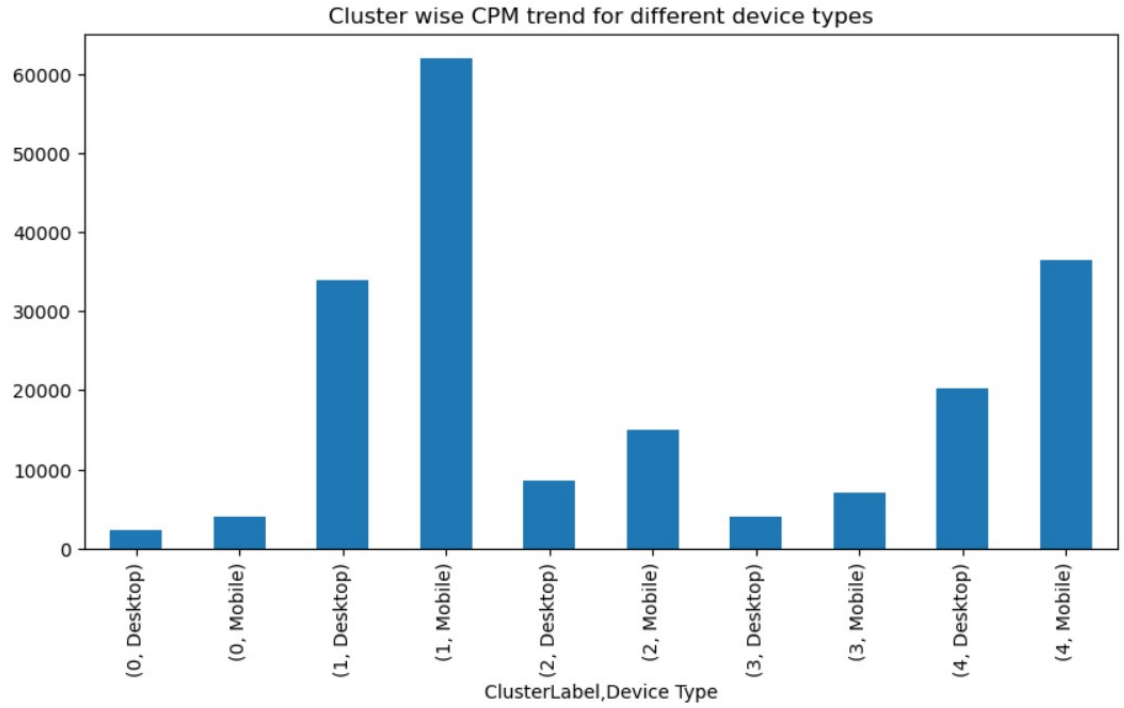


From the below graph we can identify that cluster 0 has clocked highest number of revenues both on desktop and mobile. Other features of this cluster like ad_type, ad_size and inventory can be analyzed to understand the factors why the revenue has increased for the ads belonging to these clusters.

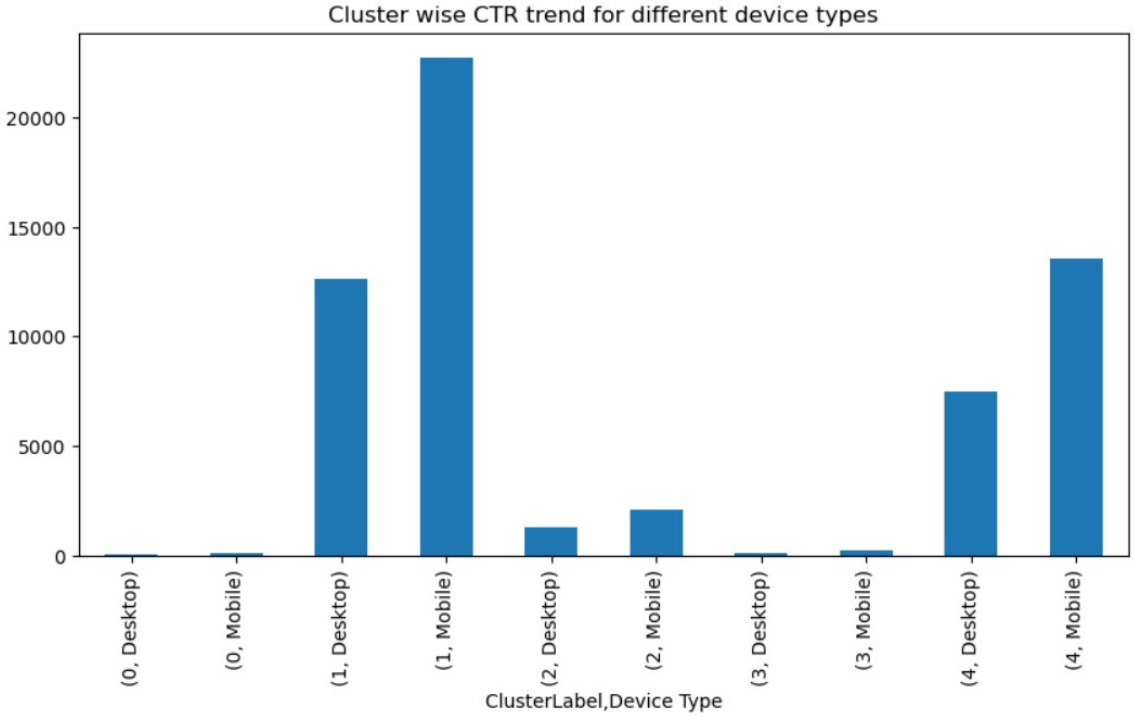


From the above two graphs it is evident that expenditure and revenue are both more for cluster 0 which indicates the ads in this cluster are yielding good results but with more expenditure. Whereas from the ads belonging to cluster 1, the revenue is very low. This area can be further researched to come up with more targeted marketing campaigns and attract customers to click on the ads and this increase revenue.

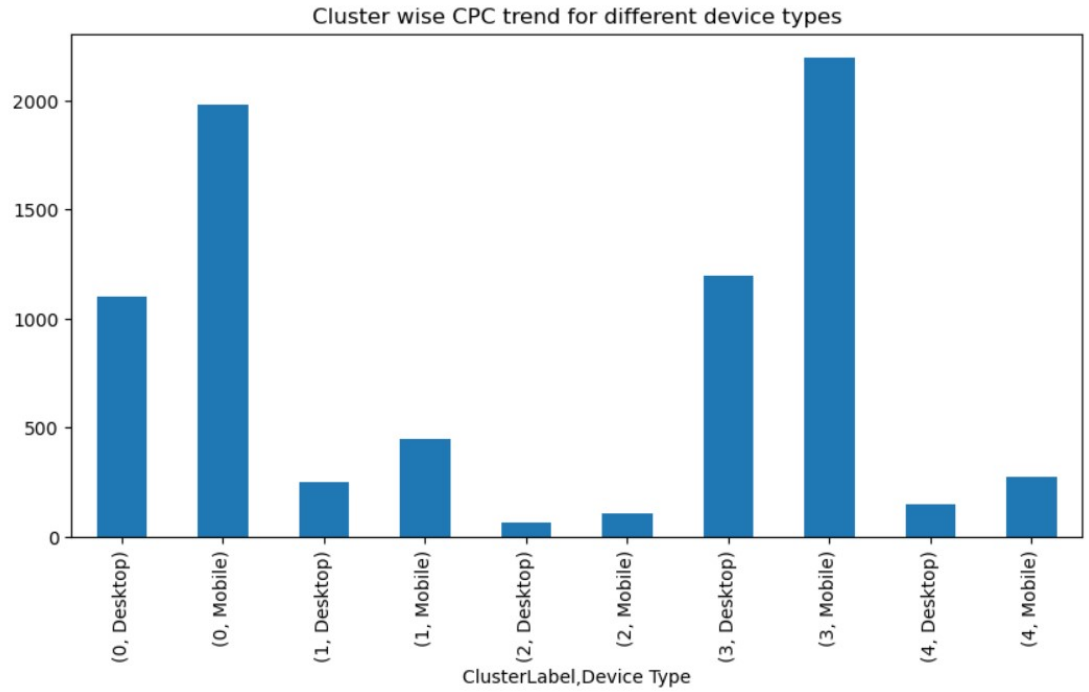
From the below graph we can identify that cluster 1 has clocked highest CPM value both on desktop and mobile. This indicates that Spend column to Impressions column is more. Expenditure is being more than the number of impressions on website for the ads in this cluster. This cluster needs analysis on why the impressions are low even though expenditure is high and should be rectified accordingly.



From the below graph we can identify that cluster 1 has clocked highest CTR value both on desktop and mobile. It is followed by cluster 4, so they are performing well every time impressions occur on websites. Whereas cluster 0,3 followed by 2 have least click through rate. This indicates that clicks are very less compared to impressions being visible on sites. These clusters need to be further analyzed to make ads more attractive and improve click through rate.



From the below graph we can identify that cluster 0 and 3 on device type mobile and desktop has clocked highest CPC. Other features of this cluster like ad_type, ad_size and inventory can be analyzed to understand the motivating factors for customers to click ads in this cluster to see the click rate is same even though spend can be optimized.



1.9 Conclude the project by providing summary of your learnings.

The given clean ads data has been divided into 5 clusters based on silhouette scores obtained.

Here are the some of the statistics of data in each cluster.

Clusters /Features	**Most occurring inventories**	**Avg ad size**	**Avg number of clicks**	**Avg spend**	**Avg revenue**	**Avg CTR**	**Avg CPM**	**Avg CPC**
0	Format1, Format 2	75205	11253	8653	6378	0.034	1.57	0.76
1	Format4, Format5	77139	1888	210	136	5.32	14.44	0.10
2	Format4, Format6	75680	65260	6985	5013	2.2	15.39	0.11
3	Format1, Format3	55325	3304	1524	993	0.06	1.8	0.55
4	Format5	206058	14363	1254	816	4.4	12	0.09

Inferences:

1. Average number of clicks are highest in cluster2 and cluster4 and least in cluster1.
2. Average revenue is higher in cluster 0 even though it does not have the highest number of clicks.
3. Click through ratio is highest in cluster1. From this we can conclude that attempting to make more impressions and adding attractive offers for ads in cluster 1 can increase the chance of improving CTR and revenue.
4. Click through ratio is least in cluster3. Marketing ad strategies for items and ads belonging to this cluster need to be re-visited and improved.
5. Overall performance of ads in clusters 0,2 are good but the spend to revenue trend for the clusters seem to be suggesting a scope of improvement in terms of marketing campaigns.

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

-> Dataset has been loaded. Head and tail show the first and last rows of the dataset as expected.

-> The data set has 640 rows and 61 columns, which means we have 640 observations and 61 features in the dataset.

-> There are no duplicated rows in the dataset.

-> Basic information about the data in the dataset has been analyzed. There are two variables State.Code and Dist.Code that make more sense to be categorical as we have finite number of them. Converted the datatypes for these two columns to categorical.

Resultant info of the dataset looks like below:

```
RangeIndex: 640 entries, 0 to 639
Data columns (total 61 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   State Code      640 non-null    object
 1   Dist.Code       640 non-null    object
 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(57), object(4)
```

Insights:

-> There are 57 integer columns and 4 object columns

-> None of the columns have null values.

-> According to the description of the data, all other variables are counts of population and they all are integers which also makes sense.

Viewing the summary of data:

	count	mean	std	min	25%	50%	75%	max
No_HH	640.0	51222.871875	48135.405475	350.0	19484.00	35837.0	68892.00	310450.0
TOT_M	640.0	79940.576563	73384.511114	391.0	30228.00	58339.0	107918.50	485417.0
TOT_F	640.0	122372.084375	113600.717282	698.0	46517.75	87724.5	164251.75	750392.0
M_06	640.0	12309.098438	11500.906881	56.0	4733.75	9159.0	16520.25	96223.0
F_06	640.0	11942.300000	11326.294567	56.0	4672.25	8663.0	15902.25	95129.0
M_SC	640.0	13820.946875	14426.373130	0.0	3466.25	9591.5	19429.75	103307.0
F_SC	640.0	20778.392188	21727.887713	0.0	5603.25	13709.0	29180.00	156429.0
M_ST	640.0	6191.807813	9912.668948	0.0	293.75	2333.5	7658.00	96785.0
F_ST	640.0	10155.640625	15875.701488	0.0	429.50	3834.5	12480.25	130119.0
M_LIT	640.0	57967.979688	55910.282466	286.0	21298.00	42693.5	77989.50	403261.0
F_LIT	640.0	66359.565625	75037.860207	371.0	20932.00	43796.5	84799.75	571140.0
M_ILL	640.0	21972.596875	19825.605268	105.0	8590.00	15767.5	29512.50	105961.0
F_ILL	640.0	56012.518750	47116.693769	327.0	22367.00	42386.0	78471.00	254160.0
TOT_WORK_M	640.0	37992.407813	36419.537491	100.0	13753.50	27936.5	50226.75	269422.0
TOT_WORK_F	640.0	41295.760938	37192.360943	357.0	16097.75	30588.5	53234.25	257848.0
MAINWORK_M	640.0	30204.446875	31480.915680	65.0	9787.00	21250.5	40119.00	247911.0
MAINWORK_F	640.0	28198.846875	29998.262689	240.0	9502.25	18484.0	35063.25	226166.0
MAIN_CL_M	640.0	5424.342188	4739.161969	0.0	2023.50	4160.5	7695.00	29113.0
MAIN_CL_F	640.0	5486.042188	5326.362728	0.0	1920.25	3908.5	7286.25	36193.0
MAIN_AL_M	640.0	5849.109375	6399.507966	0.0	1070.25	3936.5	8067.25	40843.0
MAIN_AL_F	640.0	8925.995312	12864.287584	0.0	1408.75	3933.5	10617.50	87945.0
MAIN_HH_M	640.0	883.893750	1278.642345	0.0	187.50	498.5	1099.25	16429.0

MAIN_HH_F	640.0	1380.773438	3179.414449	0.0	248.75	540.5	1435.75	45979.0
MAIN_OT_M	640.0	18047.101562	26068.480886	36.0	3997.50	9598.0	21249.50	240855.0
MAIN_OT_F	640.0	12406.035938	18972.202369	153.0	3142.50	6380.5	14368.25	209355.0
MARGWORK_M	640.0	7787.960938	7410.791691	35.0	2937.50	5627.0	9800.25	47553.0
MARGWORK_F	640.0	13096.914062	10996.474528	117.0	5424.50	10175.0	18879.25	66915.0
MARG_CL_M	640.0	1040.737500	1311.546847	0.0	311.75	606.5	1281.00	13201.0
MARG_CL_F	640.0	2307.682813	3564.626095	0.0	630.25	1226.0	2659.25	44324.0
MARG_AL_M	640.0	3304.326562	3781.555707	0.0	873.50	2062.0	4300.75	23719.0
MARG_AL_F	640.0	6463.281250	6773.876298	0.0	1402.50	4020.5	9089.25	45301.0
MARG_HH_M	640.0	316.742188	462.661891	0.0	71.75	166.0	356.50	4298.0
MARG_HH_F	640.0	786.626562	1198.718213	0.0	171.75	429.0	962.50	15448.0
MARG_OT_M	640.0	3126.154687	3609.391821	7.0	935.50	2036.0	3985.25	24728.0
MARG_OT_F	640.0	3539.323438	4115.191314	19.0	1071.75	2349.5	4400.50	36377.0
MARGWORK_3_6_M	640.0	41948.168750	39045.316918	291.0	16208.25	30315.0	57218.75	300937.0
MARGWORK_3_6_F	640.0	81076.323438	82970.406216	341.0	26619.50	56793.0	107924.00	676450.0
MARG_CL_3_6_M	640.0	6394.987500	6019.806644	27.0	2372.00	4630.0	8167.00	39106.0
MARG_CL_3_6_F	640.0	10339.864063	8467.473429	85.0	4351.50	8295.0	15102.00	50065.0
MARG_AL_3_6_M	640.0	789.848438	905.639279	0.0	235.50	480.5	986.00	7426.0
MARG_AL_3_6_F	640.0	1749.584375	2496.541514	0.0	497.25	985.5	2059.00	27171.0
MARG_HH_3_6_M	640.0	2743.635938	3059.586387	0.0	718.75	1714.5	3702.25	19343.0
MARG_HH_3_6_F	640.0	5169.850000	5335.640960	0.0	1113.75	3294.0	7502.25	36253.0
MARG_OT_3_6_M	640.0	245.362500	358.728567	0.0	58.00	129.5	276.00	3535.0
MARG_OT_3_6_F	640.0	585.884375	900.025817	0.0	127.75	320.5	719.25	12094.0
MARGWORK_0_3_M	640.0	2616.140625	3036.964381	7.0	755.00	1681.5	3320.25	20648.0
MARGWORK_0_3_F	640.0	2834.545312	3327.836932	14.0	833.50	1834.5	3610.50	25844.0
MARG_CL_0_3_M	640.0	1392.973438	1489.707052	4.0	489.50	949.0	1714.00	9875.0
MARG_CL_0_3_F	640.0	2757.050000	2788.776676	30.0	957.25	1928.0	3599.75	21611.0
MARG_AL_0_3_M	640.0	250.889062	453.336594	0.0	47.00	114.5	270.75	5775.0
MARG_AL_0_3_F	640.0	558.098438	1117.642748	0.0	109.00	247.5	568.75	17153.0
MARG_HH_0_3_M	640.0	560.690625	762.578991	0.0	136.50	308.0	642.00	6116.0
MARG_HH_0_3_F	640.0	1293.431250	1585.377936	0.0	298.00	717.0	1710.75	13714.0
MARG_OT_0_3_M	640.0	71.379688	107.897627	0.0	14.00	35.0	79.00	895.0
MARG_OT_0_3_F	640.0	200.742188	309.740854	0.0	43.00	113.0	240.00	3354.0
NON_WORK_M	640.0	510.014063	610.603187	0.0	161.00	326.0	604.50	6456.0
NON_WORK_F	640.0	704.778125	910.209225	5.0	220.50	464.5	853.50	10533.0

Insights:

-> Average of total number of females is greater than average of total number of male.

-> Male and female literates are higher than male and female illiterates.

-> Main and marginal workers both and female data has been provided for all the states and districts.

2.2 Perform detailed Exploratory analysis by creating certain questions like (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio? (Example Questions). 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

TOT_M - Total population male, TOT_F - total population female, F_LIT - Literate population female, F_ILL - illiterate female, TOT_WORK_F - total worker population female

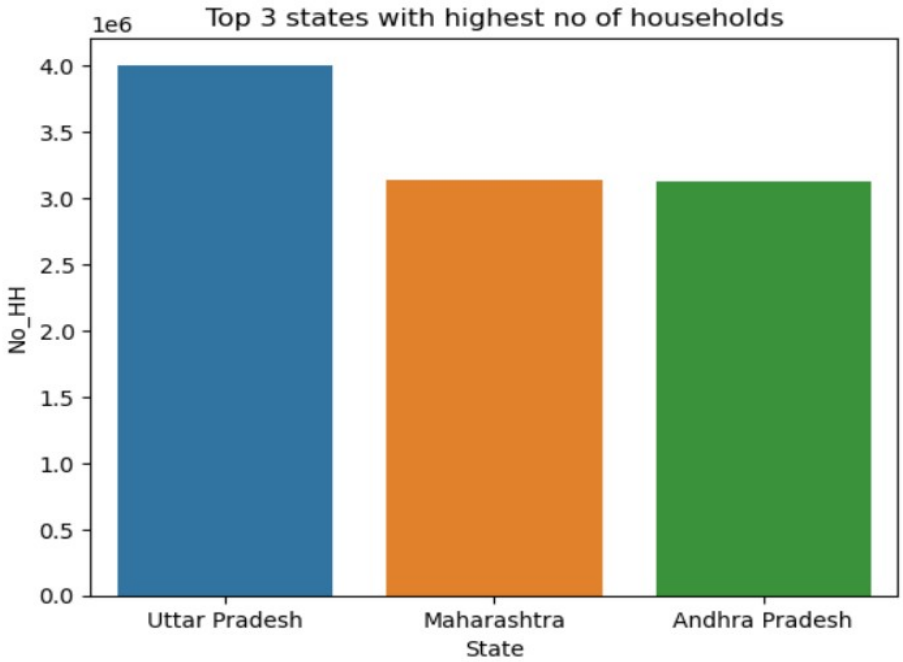
Univariate analysis:

Checked the box plots of all numeric columns

There are outliers in some values and that could be due to the heavy populations in some states.

Bivariate analysis done based on some framed questions below.

1. Which is the state with highest number of households



Shown in the graph are the top 3 states with highest number of households:

1. Uttar Pradesh

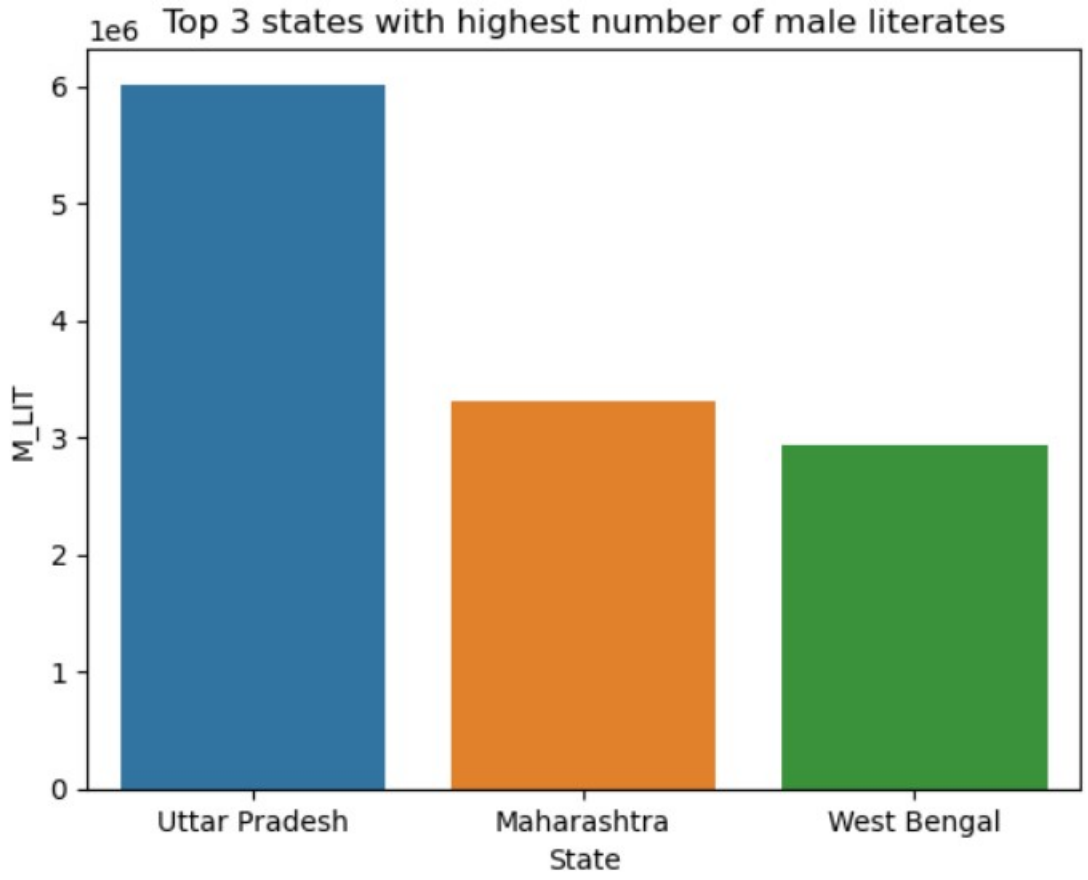
2. Maharashtra

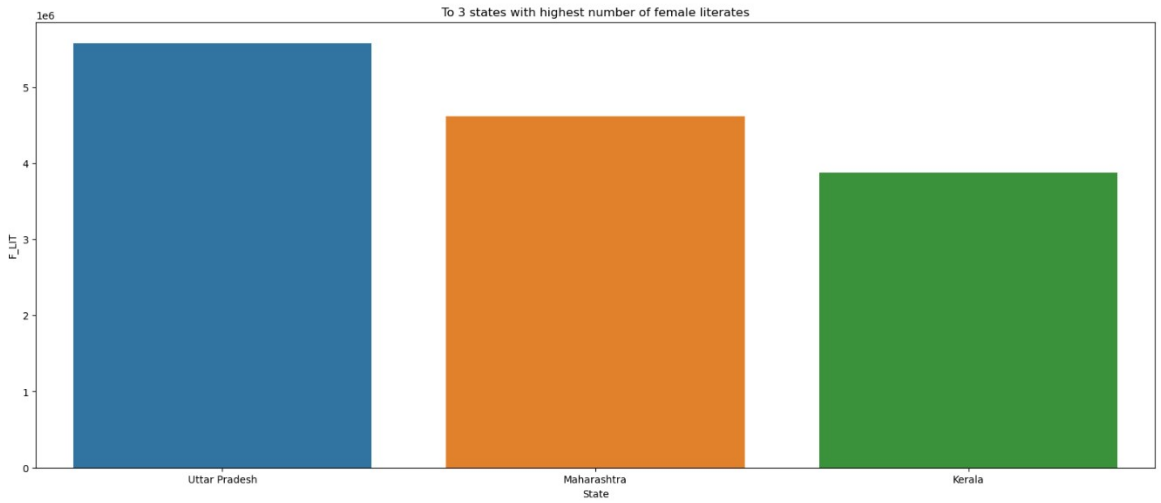
3. Andhra Pradesh with Maharashtra and Andhra Pradesh having almost same number of households.

2. Top 3 states with most literate people either male/female

For male top3 states having high number of literate people are Uttar Pradesh, Maharashtra and West Bengal. Highest number of female literates are in the states Uttar Pradesh, Maharashtra, Kerala.

Below are the reference charts:

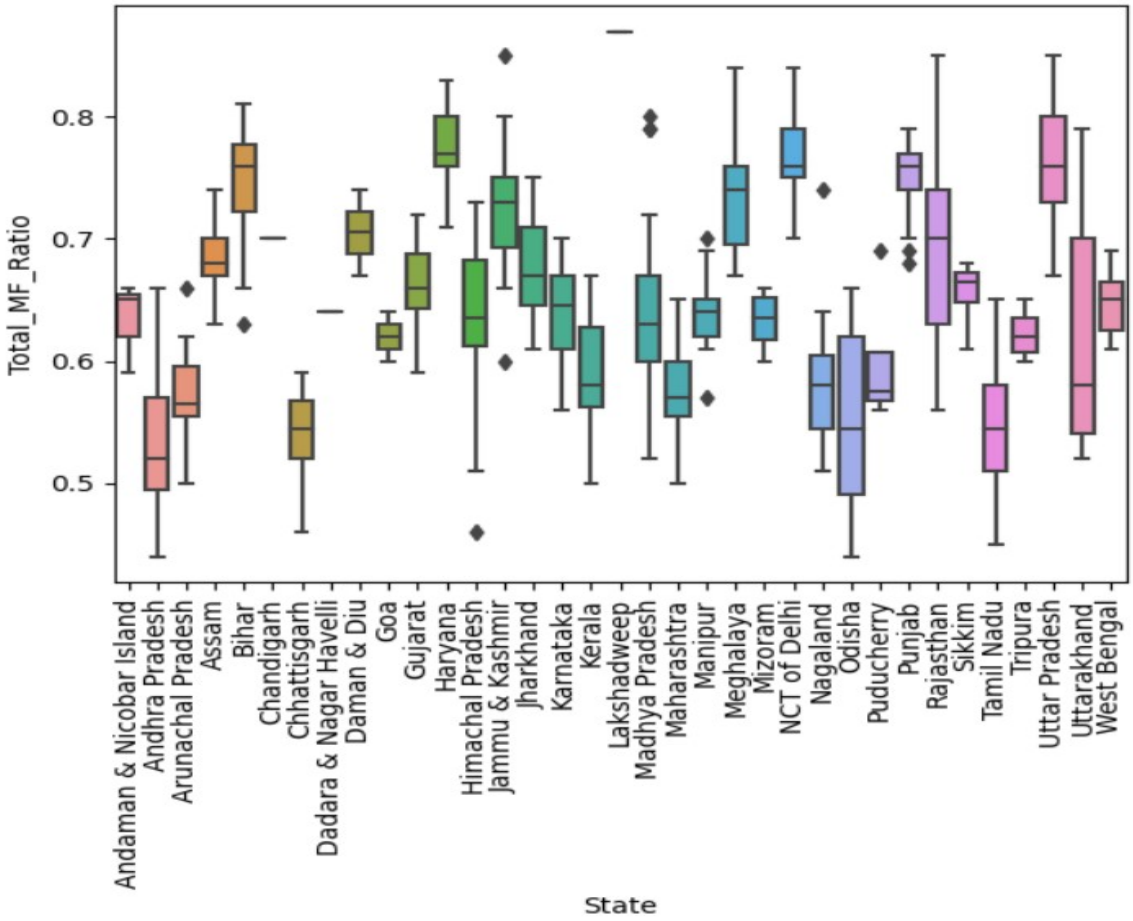




3. State with highest number and least number of illiterate people.

Top 3 states with highest number of illiterates are Uttar Pradesh, Bihar, Andhra Pradesh with Uttar Pradesh having a count around 94 lakh and Bihar, Andhra Pradesh having a count of around 42-48 lakh

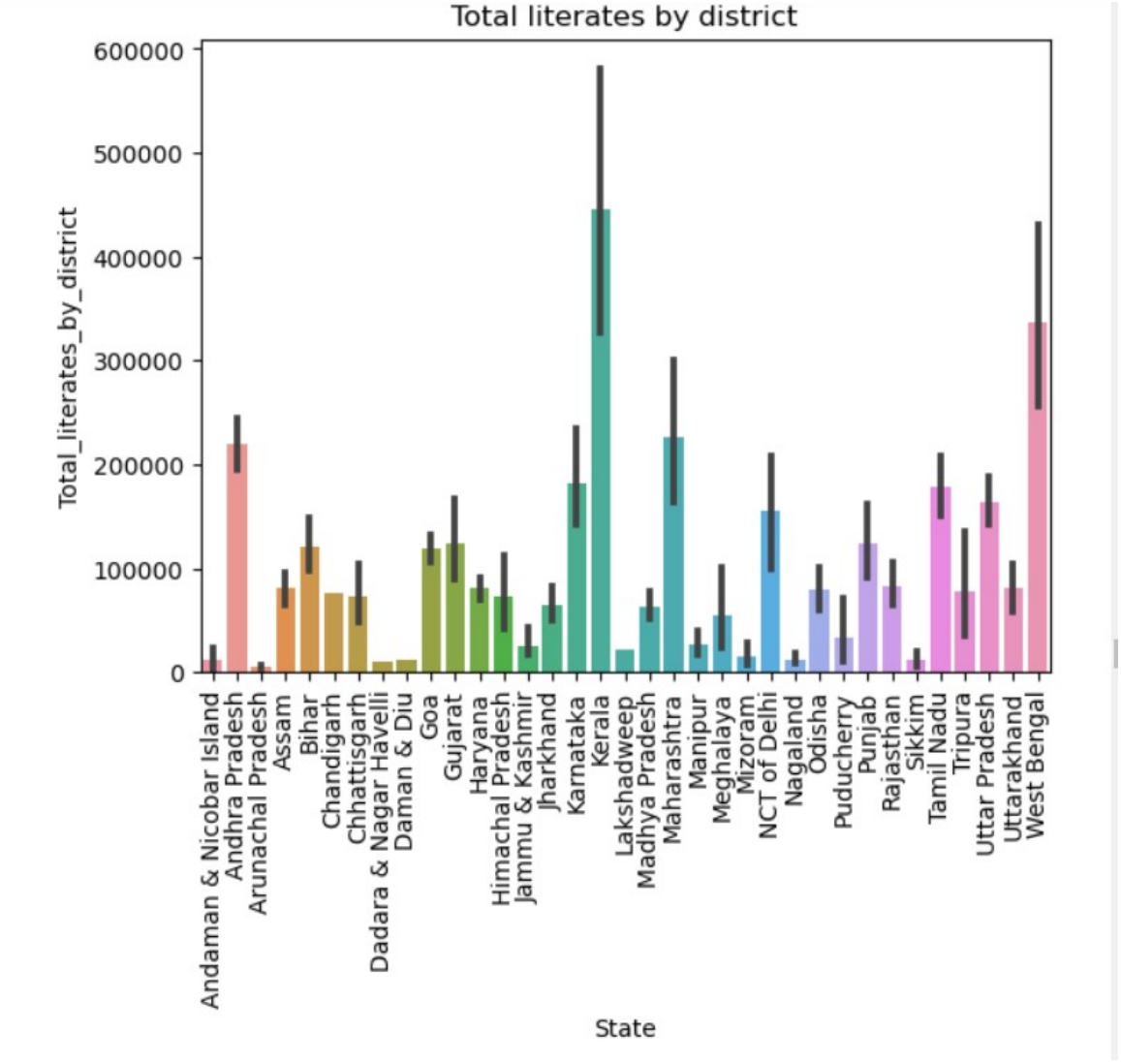
4. Ratio of male to female across different states.



The above picture shows male to female ratio trend across different districts in each state. From the above chart it is evident that median male to female ratio is high for Haryana and least for Andhra Pradesh.

Dadra and Nagar haveli is the state/UT with least male/female ratio. Uttar Pradesh is the state with highest male to female ratio.

5. District with the highest number of literates.



Districts in Kerala have the highest number of literates.

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

- Treating outliers for this case is necessary. For PCA we calculate co-variances of each variable to find principal components. If data contains outliers, variances are going to be skewed by outliers and as a result the covariances are also going to be skewed which might not result in correct principal co-ordinates.

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

There are outliers in box plots of most of the features before scaling

Applied z-score scaling on all the features.

There are outliers in box plot of all the scaled features after s-score scaling.

Scaling does not have impact on the outliers as it is not reducing the magnitude the outliers but bringing all the values to a comparable scale.

Above are boxplots before and after scaling the data and they show the outliers have not been treated before and after scaling

Therefore, treating outliers by using IQR method and then scaled the data.

After outlier treatment and scaling, there are no outliers present in the data and the data is scaled.

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

-> Removed outliers using IQR method by capping the outliers to max and min values if they are above the values respectively.

-> Performed scaling on the data using StandardScaler from sklearn library.

After data pre-processing the summary of the data looks like below:

	count	mean	std	min	25%	50%	75%	max
No_HH	640.0	-1.516148e-16	1.000782	-1.226295	-0.739143	-0.322796	0.518785	2.405677
TOT_M	640.0	1.457168e-17	1.000782	-1.256930	-0.761190	-0.294128	0.529633	2.465868
TOT_F	640.0	8.881784e-17	1.000782	-1.253026	-0.755432	-0.307934	0.523139	2.440995
M_06	640.0	3.295975e-17	1.000782	-1.252604	-0.746705	-0.268114	0.528005	2.440070
F_06	640.0	-6.626644e-17	1.000782	-1.245270	-0.731026	-0.286462	0.519980	2.396488
M_SC	640.0	-5.342948e-17	1.000782	-1.080447	-0.796150	-0.293766	0.513154	2.477110
F_SC	640.0	8.673617e-17	1.000782	-1.079963	-0.773791	-0.330877	0.514489	2.446907
M_ST	640.0	-9.037909e-17	1.000782	-0.842834	-0.793989	-0.454820	0.430539	2.267331
F_ST	640.0	-1.144917e-16	1.000782	-0.833741	-0.790834	-0.450670	0.413052	2.218881
M_LIT	640.0	-2.040035e-16	1.000782	-1.238527	-0.758902	-0.270522	0.535153	2.476235
F_LIT	640.0	9.003215e-17	1.000782	-1.185760	-0.774548	-0.317268	0.502780	2.418772
M_ILL	640.0	7.910339e-17	1.000782	-1.224367	-0.727556	-0.307301	0.497493	2.335065
F_ILL	640.0	-1.075529e-16	1.000782	-1.253898	-0.746433	-0.285501	0.545346	2.483015
TOT_WORK_M	640.0	-5.447032e-17	1.000782	-1.243185	-0.767497	-0.273362	0.503230	2.409321
TOT_WORK_F	640.0	3.837208e-16	1.000782	-1.266428	-0.753659	-0.281609	0.456096	2.270728
MAINWORK_M	640.0	7.355228e-17	1.000782	-1.180164	-0.770112	-0.286607	0.509224	2.428228
MAINWORK_F	640.0	3.032297e-16	1.000782	-1.172300	-0.743237	-0.327168	0.440847	2.216973
MAIN_CL_M	640.0	-1.422473e-16	1.000782	-1.261769	-0.775444	-0.261840	0.587637	2.632257
MAIN_CL_F	640.0	-2.844947e-17	1.000782	-1.236486	-0.776871	-0.300979	0.507492	2.434035
MAIN_AL_M	640.0	8.569534e-17	1.000782	-1.023486	-0.826055	-0.297310	0.464700	2.400832
MAIN_AL_F	640.0	4.022824e-16	1.000782	-0.916707	-0.740481	-0.424652	0.411472	2.139401
MAIN_HH_M	640.0	-7.927686e-17	1.000782	-1.023258	-0.767957	-0.344499	0.473485	2.335648

MAIN_HH_F	640.0	-1.080733e-16	1.000782	-0.981052	-0.734363	-0.445030	0.442805	2.208556
MAIN_OT_M	640.0	3.805983e-16	1.000782	-1.048469	-0.767145	-0.369429	0.457996	2.295708
MAIN_OT_F	640.0	-1.400789e-15	1.000782	-1.059051	-0.742233	-0.399080	0.447437	2.231943
MARGWORK_M	640.0	2.046974e-17	1.000782	-1.256152	-0.748763	-0.278609	0.450920	2.250446
MARGWORK_F	640.0	3.157197e-17	1.000782	-1.278163	-0.742281	-0.262637	0.616205	2.653933
MARG_CL_M	640.0	-1.800643e-16	1.000782	-1.108205	-0.724974	-0.362641	0.466515	2.253748
MARG_CL_F	640.0	2.886580e-16	1.000782	-1.101241	-0.729856	-0.378801	0.465766	2.259199
MARG_AL_M	640.0	1.099815e-16	1.000782	-1.067515	-0.754367	-0.328293	0.474293	2.317285
MARG_AL_F	640.0	-4.961309e-17	1.000782	-1.041643	-0.806518	-0.367616	0.482147	2.415144
MARG_HH_M	640.0	6.973588e-17	1.000782	-1.054877	-0.756972	-0.365648	0.425304	2.198718
MARG_HH_F	640.0	-1.075529e-17	1.000782	-1.054472	-0.777677	-0.363090	0.496706	2.408281
MARG_OT_M	640.0	-2.203099e-17	1.000782	-1.130894	-0.751919	-0.302742	0.492860	2.360028
MARG_OT_F	640.0	-5.776629e-17	1.000782	-1.174117	-0.774904	-0.290369	0.487390	2.380830
MARGWORK_3_6_M	640.0	5.898060e-18	1.000782	-1.249662	-0.745760	-0.299174	0.552535	2.499977
MARGWORK_3_6_F	640.0	1.908196e-16	1.000782	-1.198017	-0.778099	-0.295941	0.521108	2.469918
MARG_CL_3_6_M	640.0	-9.610368e-17	1.000782	-1.239572	-0.751722	-0.281972	0.453859	2.262230
MARG_CL_3_6_F	640.0	-5.932754e-17	1.000782	-1.285196	-0.740578	-0.237191	0.631721	2.690171
MARG_AL_3_6_M	640.0	9.783840e-17	1.000782	-1.124541	-0.748979	-0.358267	0.447876	2.243159
MARG_AL_3_6_F	640.0	-7.285839e-18	1.000782	-1.118092	-0.734747	-0.358342	0.469250	2.275247
MARG_HH_3_6_M	640.0	-7.945034e-17	1.000782	-1.056751	-0.755678	-0.338575	0.494061	2.368671
MARG_HH_3_6_F	640.0	9.020562e-18	1.000782	-1.035839	-0.806140	-0.356487	0.511418	2.487755
MARG_OT_3_6_M	640.0	6.262352e-17	1.000782	-1.057739	-0.746484	-0.362782	0.423406	2.178241
MARG_OT_3_6_F	640.0	2.449430e-16	1.000782	-1.047079	-0.773017	-0.359511	0.495925	2.399340
MARGWORK_0_3_M	640.0	-6.904199e-17	1.000782	-1.124879	-0.761067	-0.310435	0.486621	2.358151
MARGWORK_0_3_F	640.0	2.414735e-16	1.000782	-1.152416	-0.773006	-0.309564	0.512685	2.441220
MARG_CL_0_3_M	640.0	1.280226e-16	1.000782	-1.240654	-0.754702	-0.294774	0.470938	2.309398
MARG_CL_0_3_F	640.0	-2.393918e-17	1.000782	-1.203773	-0.761567	-0.298617	0.498641	2.388954
MARG_AL_0_3_M	640.0	2.411266e-17	1.000782	-1.005714	-0.754026	-0.392558	0.444174	2.241473
MARG_AL_0_3_F	640.0	-3.396589e-16	1.000782	-1.028083	-0.749969	-0.396587	0.423082	2.182658
MARG_HH_0_3_M	640.0	3.747003e-17	1.000782	-1.070466	-0.750507	-0.348507	0.434396	2.211749
MARG_HH_0_3_F	640.0	-7.598089e-17	1.000782	-1.014472	-0.753386	-0.386289	0.484362	2.340985
MARG_OT_0_3_M	640.0	5.030698e-18	1.000782	-1.026779	-0.771544	-0.388692	0.413475	2.191004
MARG_OT_0_3_F	640.0	-2.005340e-16	1.000782	-1.051594	-0.776204	-0.327896	0.485464	2.377967
NON_WORK_M	640.0	2.949030e-18	1.000782	-1.187845	-0.757044	-0.315540	0.429666	2.209731
NON_WORK_F	640.0	-3.191891e-17	1.000782	-1.184337	-0.762130	-0.284087	0.478041	2.338298

-> Check for presence of correlations using either pairplot or heatmap.

The correlation map indicates there are correlations that are significant between considerable amount of variables which gives us a hint that PCA can be performed on the given variables.

-> Check statistically if PCA can be performed on the given dataset by using bartlett sphericity test.

- Null hypothesis: correlations are not significant

- Alternate hypothesis: correlations are significant.

p_value obtained for the dataset is 0 and it is less than 0.05. It indicates to reject null hypothesis that correlations are not significant.

Conclusion: Correlations are significant to perform PCA.

-> Confirm adequacy of sample size using kmo test

- Value obtained for given sample size in kmo test is 0.93. High kmo value indicates good adequacy of sample size.

Conclusion: Sample size is adequate.

Applying initial PCA with principal components same as the number of features:

Obtained extracted principal components (also called eigen vectors):

Obtained extracted eigen values:

array([3.56488638e+01, 7.64357559e+00, 3.76919551e+00, 2.77722349e+00,
1.90694892e+00, 1.15490310e+00, 9.87726707e-01, 4.64629906e-01,
3.96708513e-01, 3.22346888e-01, 2.73207369e-01, 2.35647574e-01,
1.81401107e-01, 1.69243770e-01, 1.38592325e-01, 1.31505852e-01,
1.03809666e-01, 9.55333831e-02, 8.58580407e-02, 8.09138742e-02,
6.60179067e-02, 6.30797999e-02, 4.82756124e-02, 4.59506197e-02,
4.37747566e-02, 3.19339710e-02, 2.86194563e-02, 2.75481445e-02,
2.34340044e-02, 2.20296816e-02, 1.87487040e-02, 1.59004895e-02,
1.39957919e-02, 1.18916465e-02, 1.11133495e-02, 9.07842645e-03,
7.25127869e-03, 6.27213692e-03, 4.95541908e-03, 4.60667097e-03,
3.45902033e-03, 2.18408510e-03, 2.13514664e-03, 1.92111328e-03,
1.43840980e-03, 1.09968912e-03, 9.65752052e-04, 8.62630267e-04,
6.51634478e-04, 5.76658846e-04, 4.35790607e-04, 3.70037468e-04,
3.06660171e-04, 2.07854170e-04, 1.38286484e-04, 8.97034441e-05,
4.61745385e-05])

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

-> Explained variance ratio: which is the contribution of each principal component to covering complete randomness of data.

[6.24441446e-01, 1.33888289e-01, 6.60229147e-02, 4.86470891e-02,
3.34029704e-02, 2.02297994e-02, 1.73014629e-02, 8.13866529e-03,
6.94892379e-03, 5.64637229e-03, 4.78562250e-03, 4.12770833e-03,
3.17750294e-03, 2.96454958e-03, 2.42764517e-03, 2.30351534e-03,
1.81837655e-03, 1.67340548e-03, 1.50392785e-03, 1.41732362e-03,
1.15639919e-03, 1.10493400e-03, 8.45617224e-04, 8.04891611e-04,
7.66778221e-04, 5.59369722e-04, 5.01311201e-04, 4.82545623e-04,
4.10480504e-04, 3.85881758e-04, 3.28410688e-04, 2.78520087e-04,
2.45156553e-04, 2.08299401e-04, 1.94666401e-04, 1.59021779e-04,
1.27016642e-04, 1.09865556e-04, 8.68013375e-05, 8.06925096e-05,
6.05897475e-05, 3.82574118e-05, 3.74001838e-05, 3.36510796e-05,
2.51958296e-05, 1.92626466e-05, 1.69165450e-05, 1.51102177e-05,
1.14143210e-05, 1.01010143e-05, 7.63350323e-06, 6.48174183e-06,
5.37159674e-06, 3.64086663e-06, 2.42228792e-06, 1.57128566e-06,
8.08813873e-07]

-> Cumulative variance ratio: Cumulative sum of explained variance ratios displayed in the array above.

[0.62444145, 0.75832974, 0.82435265, 0.87299974, 0.90640271,
0.92663251, 0.94393397, 0.95207264, 0.95902156, 0.96466793,
0.96945356, 0.97358126, 0.97675877, 0.97972332, 0.98215096,
0.98445448, 0.98627285, 0.98794626, 0.98945019, 0.99086751,
0.99202391, 0.99312884, 0.99397446, 0.99477935, 0.99554613,
0.9961055 , 0.99660681, 0.99708936, 0.99749984, 0.99788572,
0.99821413, 0.99849265, 0.99873781, 0.99894611, 0.99914077,

0.99929979, 0.99942681, 0.99953668, 0.99962348, 0.99970417,

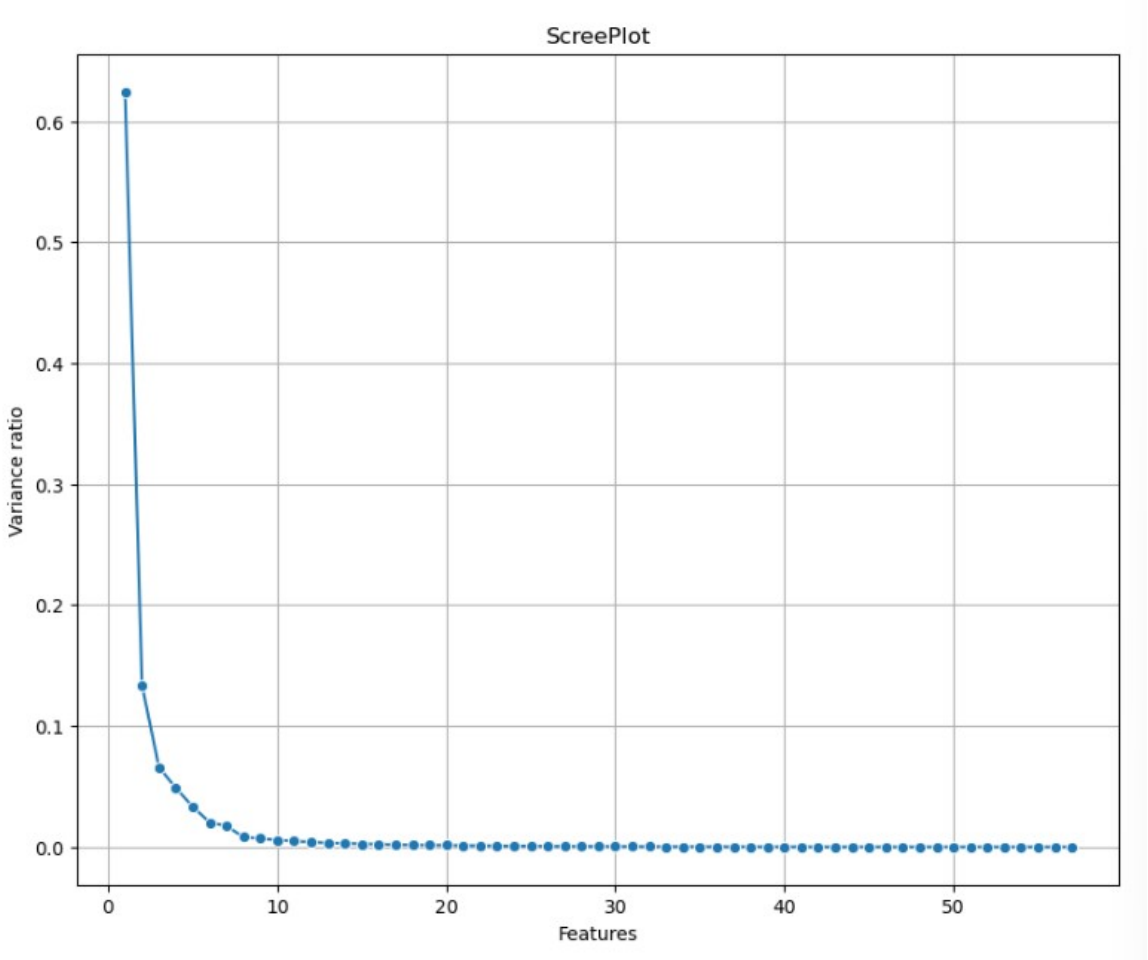
0.99976476, 0.99980302, 0.99984042, 0.99987407, 0.99989927,

0.99991853, 0.99993544, 0.99995055, 0.99996197, 0.99997207,

0.9999797 , 0.99998619, 0.99999156, 0.9999952 , 0.99999762,

0.99999919, 1.   ]

Scree plot for the explained variance ratios calculated:



To cover at least 90% of total variance we can consider the first 10 components which will cover almost 96% of the variance in the data for us.

The table with principal components considered is as shown below:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
No_HH	0.149222	-0.115487	0.101528	0.076814	-0.012090	0.082558	0.106896	-0.099501	0.026072	0.068128
TOT_M	0.159169	-0.080239	-0.038662	0.052976	-0.042344	0.073667	-0.124085	-0.108871	0.032857	-0.048419
TOT_F	0.158209	-0.093718	0.028959	0.070022	-0.022927	0.082812	-0.010291	-0.115268	0.036390	-0.022464
M_06	0.156340	-0.020341	-0.074419	0.028520	-0.080339	0.092379	-0.200807	-0.132954	0.138427	-0.157253
F_06	0.156814	-0.014310	-0.068223	0.016398	-0.078326	0.080010	-0.203412	-0.139352	0.165736	-0.145036
M_SC	0.143350	-0.079667	-0.037619	0.010210	-0.167893	0.050969	-0.040399	0.189160	-0.531730	-0.098450
F_SC	0.143537	-0.087098	0.021350	0.016244	-0.158092	0.054568	0.053990	0.177359	-0.515058	-0.065844
M_ST	0.018849	0.069101	0.323827	0.091143	0.418412	-0.231809	-0.355238	-0.071636	-0.113014	-0.008383
F_ST	0.017878	0.067316	0.338705	0.079554	0.415965	-0.214542	-0.327677	-0.078392	-0.136032	-0.028623
M_LIT	0.155152	-0.105986	-0.032107	0.089187	-0.014033	0.081378	-0.067062	-0.102886	-0.017442	0.000585
F_LIT	0.145450	-0.133234	-0.005133	0.125412	0.029084	0.102207	0.013492	-0.127069	0.000977	0.123405
M_ILL	0.154551	-0.009460	-0.047054	-0.034665	-0.104073	0.037957	-0.243097	-0.091042	0.129506	-0.155151
F_ILL	0.158283	-0.021793	0.079345	-0.010578	-0.110332	0.013986	-0.036988	-0.053630	0.030322	-0.148283
TOT_WORK_M	0.154076	-0.120912	-0.001116	0.069046	-0.023104	0.035802	-0.085403	-0.045795	-0.024613	0.090111
TOT_WORK_F	0.142530	-0.076003	0.194130	0.111057	-0.018931	-0.016587	0.174258	-0.068363	0.072895	-0.014482
MAINWORK_M	0.141932	-0.166700	0.019821	0.100188	-0.043225	0.018054	-0.087326	-0.052081	-0.051224	0.124996
MAINWORK_F	0.125732	-0.142250	0.209976	0.133013	-0.054674	-0.051951	0.149036	-0.077194	0.097198	0.062322
MAIN_CL_M	0.111692	0.042552	0.033131	0.078851	-0.303376	-0.293504	-0.288790	0.425831	-0.021079	0.210613
MAIN_CL_F	0.083035	0.095893	0.188822	0.265022	-0.257925	-0.269914	0.026294	0.197728	0.206008	-0.308022
MAIN_AL_M	0.119291	-0.053342	0.225831	-0.121379	-0.253131	-0.023336	-0.110701	0.036290	0.104113	0.382113
MAIN_AL_F	0.090089	-0.072467	0.356566	-0.020989	-0.199220	-0.056558	0.125689	0.050167	0.166162	0.203850
MAIN_HH_M	0.141850	-0.101835	-0.102202	-0.021969	-0.060812	-0.142869	-0.064688	-0.117888	-0.277749	-0.212708

MAIN_HH_F	0.133880	-0.113257	0.021613	-0.045436	-0.023063	-0.318473	0.231188	-0.248439	-0.125225	0.008395
MAIN_OT_M	0.122762	-0.203602	-0.028144	0.147025	0.069907	0.071213	-0.007768	-0.077346	-0.111308	0.143569
MAIN_OT_F	0.116866	-0.205899	0.069034	0.155917	0.106774	0.033885	0.091291	-0.082668	-0.041279	0.160743
MARGWORK_M	0.156656	0.079039	-0.068685	-0.078572	0.065812	0.078655	-0.057223	0.038326	0.101418	0.038944
MARGWORK_F	0.148695	0.108813	0.104957	0.015788	0.077624	0.099156	0.152719	0.056840	0.044890	-0.143357
MARG_CL_M	0.088163	0.271522	-0.104745	0.157104	-0.018005	-0.032738	-0.002942	-0.059979	-0.007296	0.253900
MARG_CL_F	0.065160	0.275398	-0.036325	0.285024	-0.055152	-0.031787	0.063488	-0.035425	-0.012988	-0.090082
MARG_AL_M	0.127278	0.156579	0.070434	-0.250594	-0.047200	0.079748	-0.093442	0.016850	-0.018887	0.116877
MARG_AL_F	0.115888	0.135048	0.259987	-0.153798	-0.012643	0.117625	0.092224	0.032799	-0.051672	-0.148351
MARG_HH_M	0.145366	0.040974	-0.144347	-0.167540	0.005575	-0.169980	-0.055670	0.033630	0.046021	-0.103482
MARG_HH_F	0.142302	0.006685	-0.093838	-0.151469	0.043616	-0.319596	0.184005	-0.133218	-0.009779	0.029823
MARG_OT_M	0.150877	-0.073440	-0.131415	0.021195	0.145109	0.018233	-0.021393	0.178069	0.060844	0.009113
MARG_OT_F	0.148018	-0.088361	-0.053883	0.059961	0.190756	0.002409	0.099744	0.251917	0.081173	0.014296
MARGWORK_3_6_M	0.157908	-0.044044	-0.066877	0.039319	-0.059886	0.103377	-0.153180	-0.149834	0.089348	-0.155256
MARGWORK_3_6_F	0.155831	-0.092383	-0.058718	0.046130	-0.022476	0.117467	-0.098367	-0.121238	0.013202	-0.002065
MARG_CL_3_6_M	0.157640	0.066208	-0.060172	-0.091315	0.059078	0.072381	-0.064219	0.042557	0.118641	0.027123
MARG_CL_3_6_F	0.149501	0.089651	0.125792	0.018865	0.064349	0.070896	0.142888	0.072060	0.068382	-0.182673
MARG_AL_3_6_M	0.094785	0.261268	-0.096551	0.131591	-0.013887	-0.041377	-0.011264	-0.036642	0.037597	0.255107
MARG_AL_3_6_F	0.067158	0.266691	-0.018256	0.292845	-0.061019	-0.049367	0.059638	-0.012890	0.026899	-0.135939
MARG_HH_3_6_M	0.128184	0.149831	0.078194	-0.250337	-0.058665	0.073152	-0.095948	0.028983	-0.000808	0.108639
MARG_HH_3_6_F	0.113959	0.120648	0.283235	-0.143045	-0.025386	0.094867	0.089539	0.062946	-0.028435	-0.164034
MARG_OT_3_6_M	0.145108	0.036763	-0.142511	-0.166002	0.003315	-0.174634	-0.055483	0.032641	0.037624	-0.107435
MARG_OT_3_6_F	0.141029	-0.003685	-0.089356	-0.142599	0.041678	-0.343970	0.177354	-0.121278	-0.022472	0.016195
MARGWORK_0_3_M	0.150922	-0.077739	-0.130687	0.019887	0.132794	0.015826	-0.022591	0.166819	0.069807	0.007938
MARGWORK_0_3_F	0.147534	-0.101141	-0.058489	0.060087	0.170596	-0.004857	0.078573	0.222494	0.088683	0.006304
MARG_CL_0_3_M	0.142987	0.136839	-0.103565	-0.018223	0.094293	0.111045	-0.025902	0.018290	-0.004789	0.106539
MARG_CL_0_3_F	0.133784	0.166416	0.033423	0.005954	0.112351	0.185882	0.178500	-0.004079	-0.023975	0.008568
MARG_AL_0_3_M	0.062964	0.281881	-0.120293	0.208941	-0.018070	-0.004600	0.009473	-0.115859	-0.134042	0.180282
MARG_AL_0_3_F	0.056741	0.287541	-0.088097	0.240499	-0.036293	0.022023	0.066497	-0.095448	-0.134237	0.042605
MARG_HH_0_3_M	0.119102	0.182341	0.026176	-0.240416	0.016981	0.109387	-0.082858	-0.048639	-0.093377	0.160394
MARG_HH_0_3_F	0.113044	0.177112	0.164774	-0.189408	0.047538	0.189006	0.109686	-0.070174	-0.137705	-0.045505
MARG_OT_0_3_M	0.142140	0.052925	-0.144419	-0.167554	0.014187	-0.149690	-0.050786	0.038874	0.075369	-0.084102
MARG_OT_0_3_F	0.141370	0.035109	-0.102175	-0.169020	0.047504	-0.233858	0.194686	-0.151097	0.038136	0.085147
NON_WORK_M	0.147629	-0.049122	-0.126673	0.024036	0.191790	0.022904	-0.016338	0.232588	0.013636	0.033467
NON_WORK_F	0.142103	-0.039848	-0.028545	0.057402	0.249765	0.042833	0.175252	0.325837	0.050929	0.023739

-> Applying fit transform for the given pca model and the dataset with the extracted principal components gives us the final data frame with 640 rows having values calculated according to new principal components.

-> Below is the glimpse of the principal component-oriented dataset.

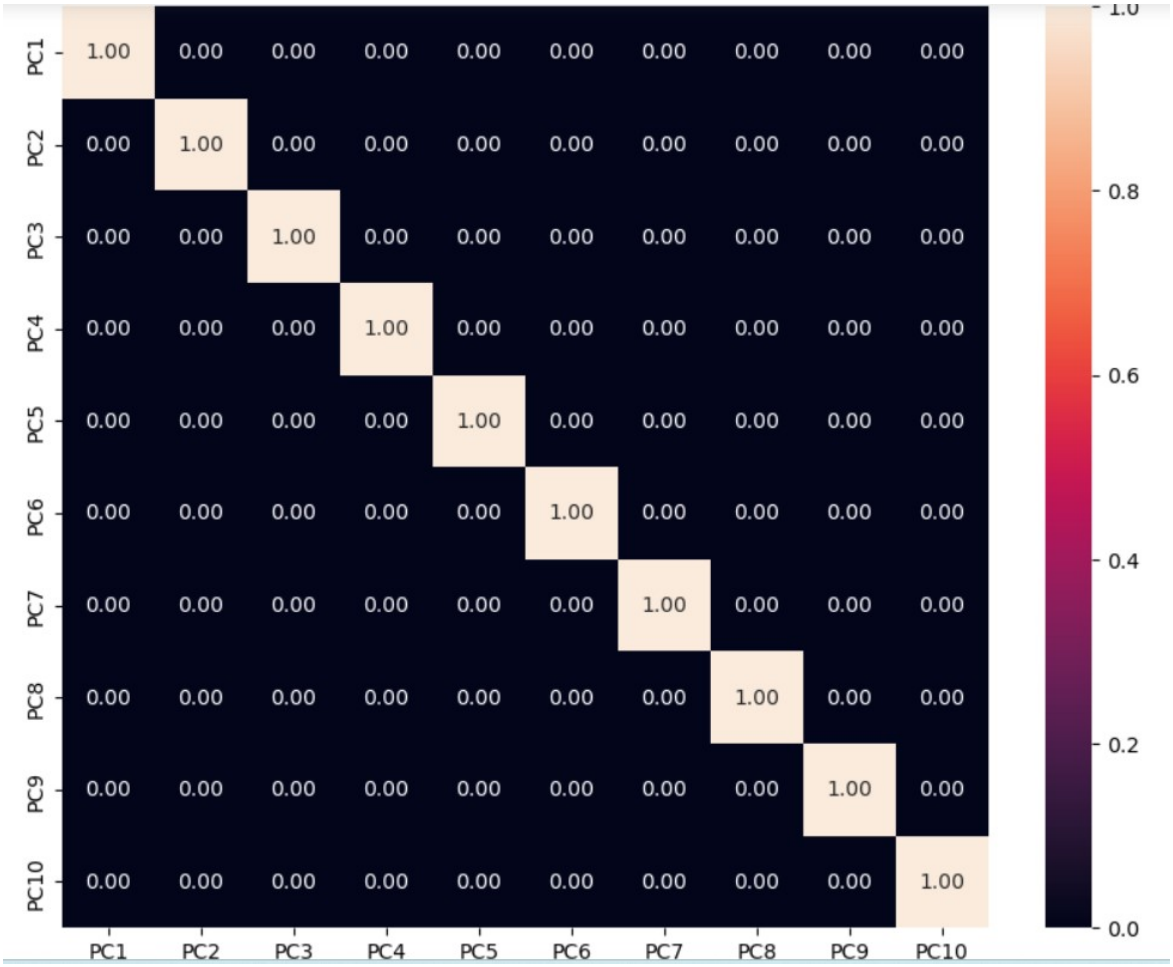
-> Viewing first 5 rows

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
0	-5.528161	0.430378	-1.473827	-1.278049	0.376358	0.536634	-0.379796	-0.261872	0.349283	0.338051
1	-5.492016	-0.106110	-2.015641	-1.750168	-0.006857	-1.006038	0.288553	-0.513547	0.317033	-0.389631
2	-7.474643	-0.217194	-0.247428	0.006079	0.556282	-0.150034	-0.190095	-0.118042	0.038825	-0.245781
3	-7.919737	-0.652311	-0.659220	-0.735550	0.272465	0.212952	0.110382	-0.038089	-0.014032	-0.117605
4	-5.175695	2.304059	-1.157327	1.060796	1.080249	-0.050521	0.073161	-0.677402	-0.236753	0.662862

-> Viewing last 5 rows

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
635	-7.946040	-1.302846	-0.819360	-0.871301	-0.101966	0.533559	0.236931	-0.036711	0.176688	-0.169134
636	-7.244719	-1.396520	-0.719443	-0.895934	-0.062742	0.552805	0.435032	0.010509	0.038017	-0.054548
637	-7.886268	-1.003537	-0.909285	-1.238009	0.146031	-0.046046	0.361532	0.040313	0.149283	-0.230975
638	-7.864260	-0.999338	-0.851569	-0.782561	-0.081681	0.441821	0.309963	0.127153	0.169276	-0.058645
639	-7.416226	-1.412143	-0.865921	-0.680528	0.096861	0.549783	0.277154	0.071644	0.270571	-0.030467

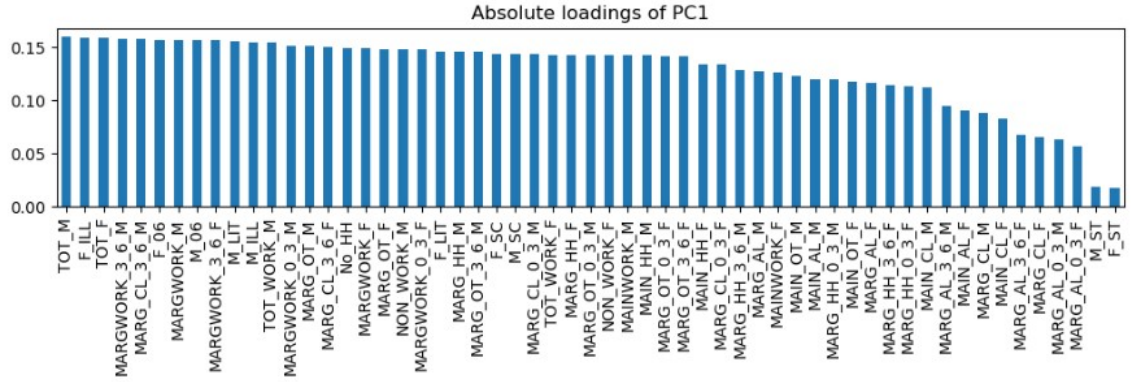
-> Final check to confirm if there is no more noise in the new principal component reduced dataset. Generate a heatmap on the new dataset



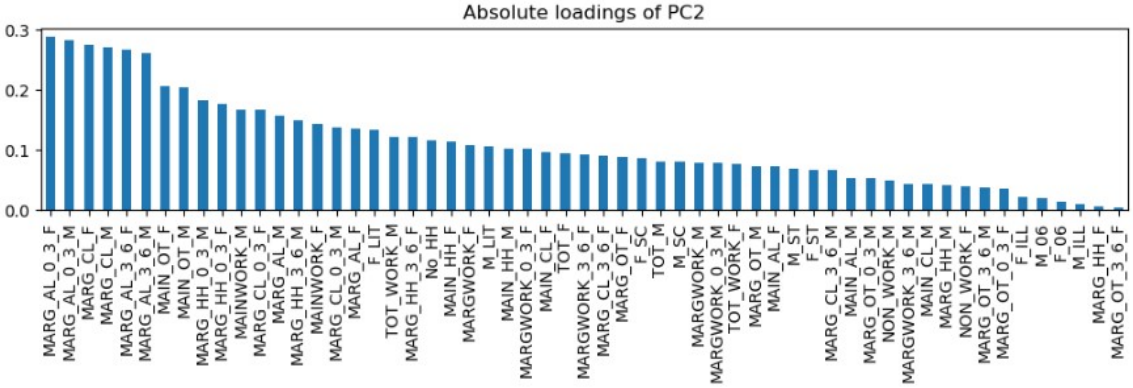
As we can see there are no more correlations among new features i.e., principal components which indicates that this new reduced dataset is good to be passed into a machine learning model.

2.7 Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.

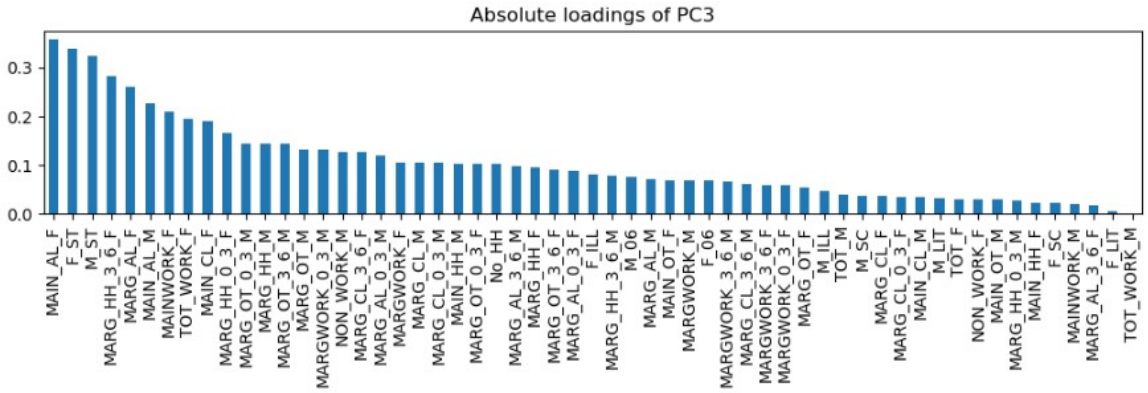
-> Below are the variations captured by each principal component from the actual columns.



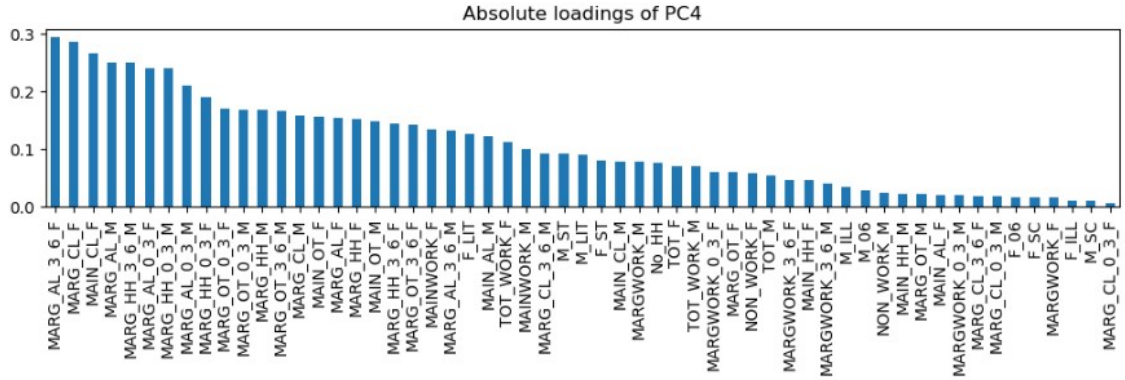
-> Principal component 1 is capturing high variance of considerably most of the actual columns with highest variance of TOT_M captured.



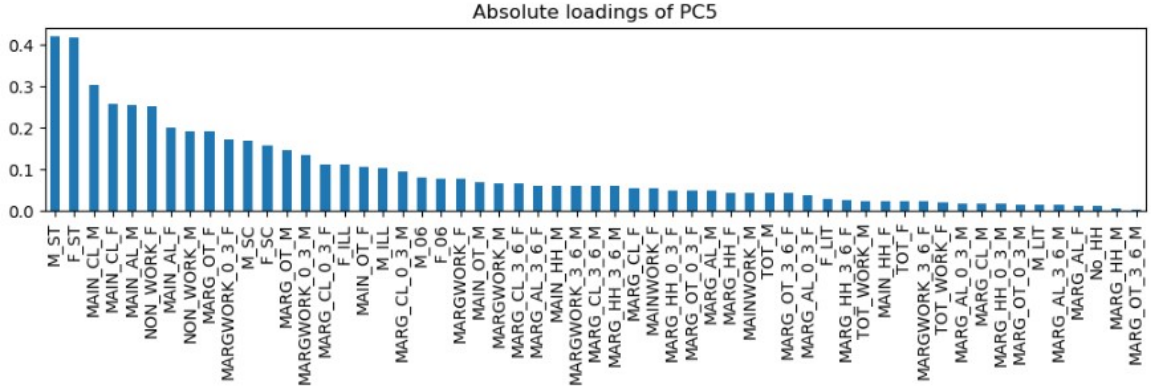
-> The major contribution for the principal component 2 came from columns MARG_AL_0_3_F, MARG_AL_0_3_M, MARG_CL_F, MARG_CL_M, MARG_AL_3_6_F, MARG_AL_3_6_M.



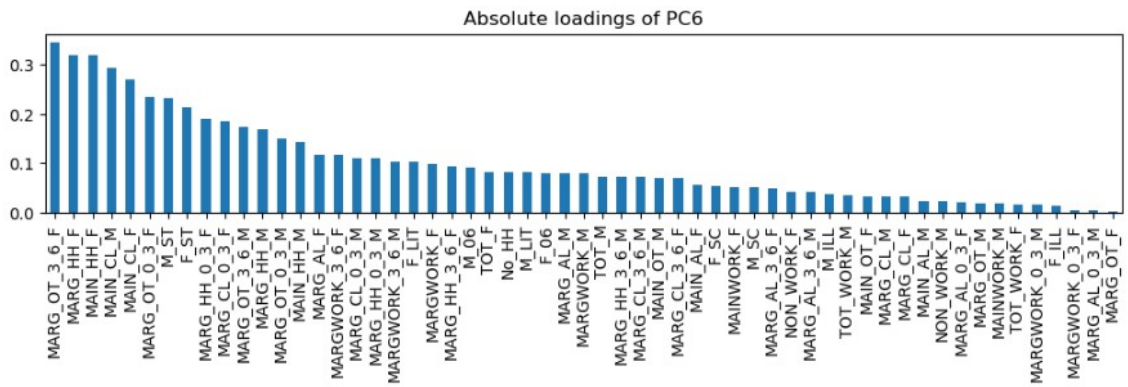
-> Principal component 3 is capturing next highest variance with variance of columns MAIN_AL_F, F_ST< M_ST< MARG_HH_3_6_F, MARG_AL_F most captured.



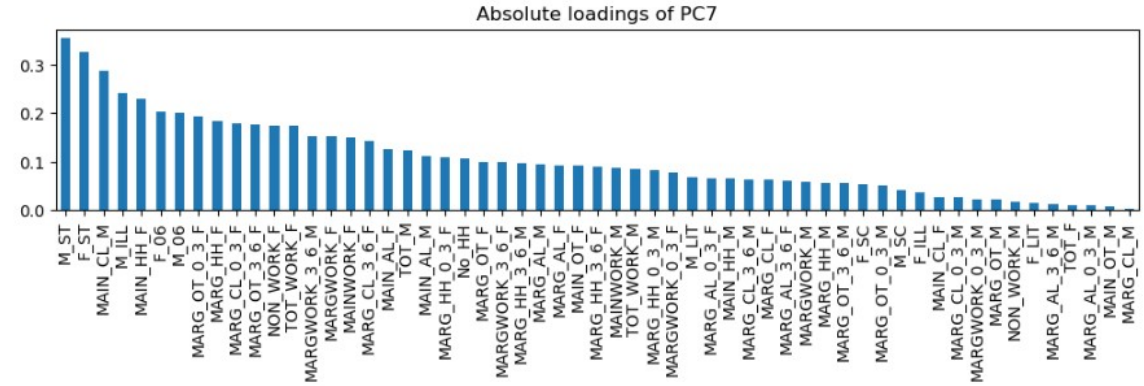
-> Principal component 4 is capturing next highest variance with variance of columns MARG_AL_3_6_F, MARG_CL_F, MAIN_CL_F, MARG_AL_M, MARG_HH_3_6_M, MARG_AL_0_3_F, MARG_HH_0_3_F most captured



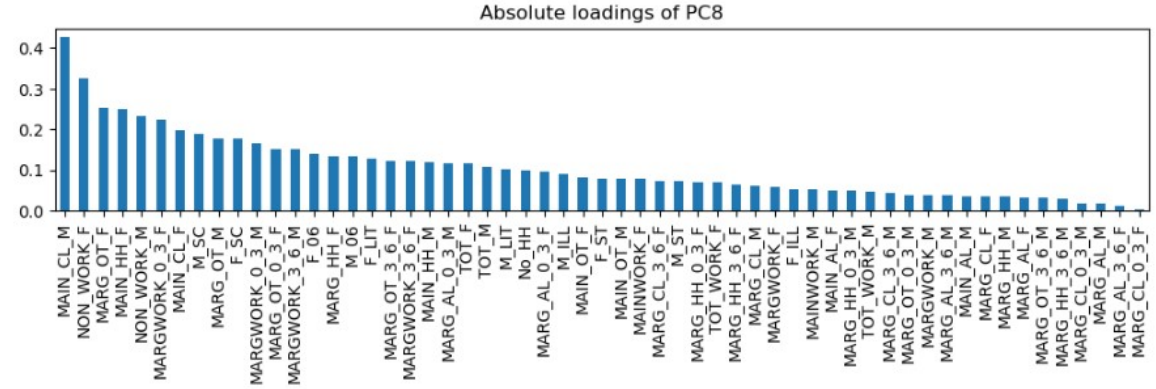
-> Principal component 5 is capturing next highest variance with variance of columns M_ST< F_ST< MAIN_CL_M, MAIN_CL_F most captured



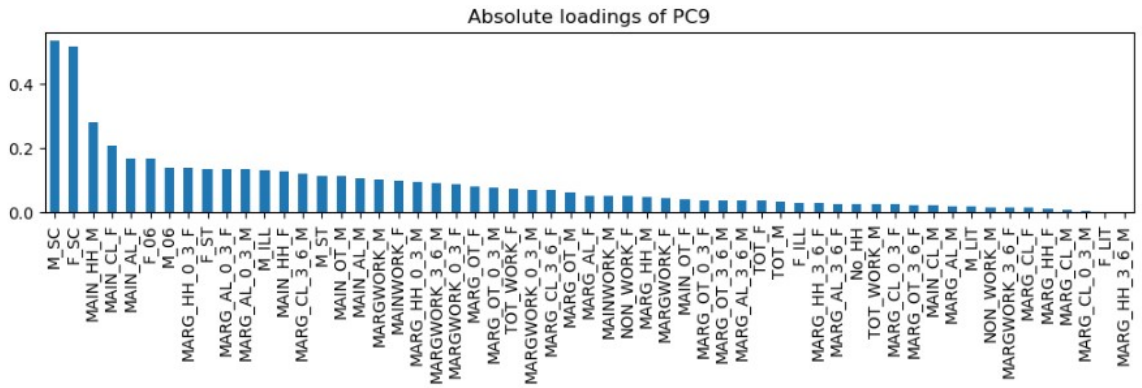
-> Principal component 6 is capturing next highest variance with variance of columns MARG_OT_3_6_F, MARG_HH_F, MAIN_HH_F, MAIN_CL_M, MAIN_CL_F most captured



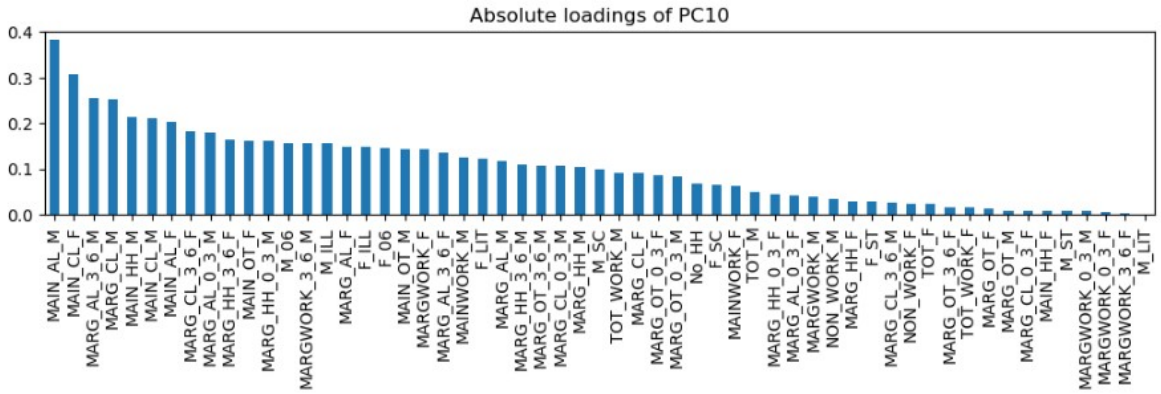
-> Principal component 7 is capturing next highest variance with variance of columns M_ST< F_ST< MAIN_CL_M, M_ILL most captured



-> Principal component 8 is capturing next highest variance with variance of columns MAIN_CL_M, NON_WOEK_F, MARG_OT_F, MAIN_HH_F most captured.



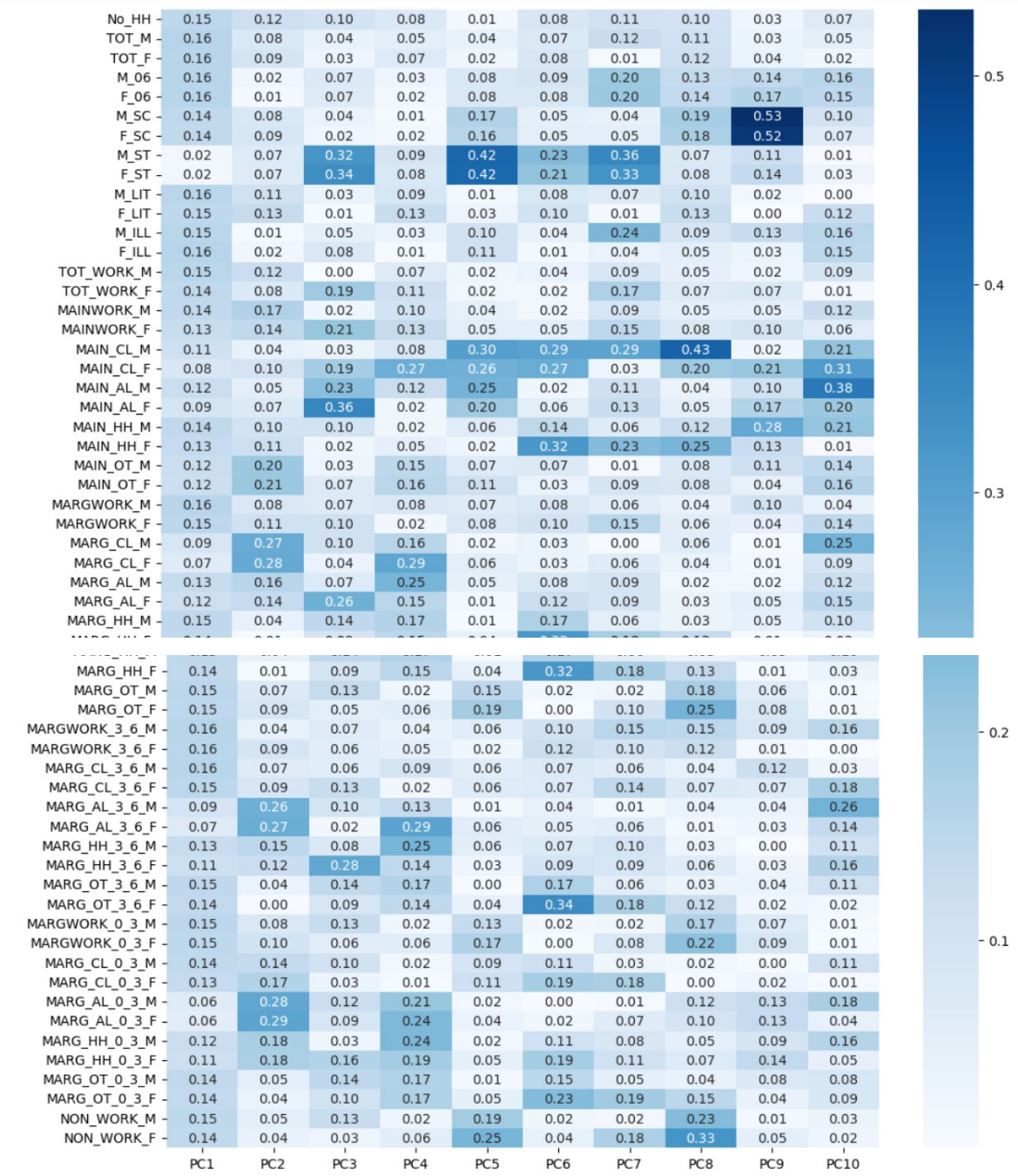
-> Principal component 9 is capturing next highest variance with variance of columns M_SC, F_SC, MAIN_HH_M, mAIN_CL_F, MAIN_AL_F most captured.



-> Principal component 10 is capturing next highest variance with variance of columns MAIN_AL_M, MAIN_CL_F, MARG_AL_3_6_M, MARG_CL_M most captured.

-> Above are the contributions of all the columns to each principal component that captured around 96% of the complete actual data.

-> The relation between each principal component and its corresponding column value contributions can also be viewed through heat map.



2.8 Write linear equation for first PC.

-> Linear equation for PC1 is as below:

(0.15) * No_HH + (0.16) * TOT_M + (0.16) * TOT_F + (0.16) * M_06 + (0.16) * F_06 + (0.14) * M_SC + (0.14) * F_SC + (0.02) * M_ST + (0.02) * F_ST + (0.16) * M_LIT + (0.15) * F_LIT + (0.15) * M_ILL + (0.16) * F_ILL + (0.15) * TOT_WORK_M + (0.14) * TOT_WORK_F + (0.14) * MAINWORK_M + (0.13) * MAINWORK_F + (0.11) * MAIN_CL_M + (0.08) * MAIN_CL_F + (0.12) * MAIN_AL_M + (0.09) * MAIN_AL_F +

(0.14) * MAIN_HH_M + (0.13) * MAIN_HH_F + (0.12) * MAIN_OT_M + (0.12) * MAIN_OT_F + (0.16) * MARGWORK_M + (0.15) * MARGWORK_F + (0.09) * MARG_CL_M + (0.07) * MARG_CL_F + (0.13) * MARG_AL_M + (0.12) * MARG_AL_F + (0.15) * MARG_HH_M + (0.14) * MARG_HH_F + (0.15) * MARG_OT_M + (0.15) * MARG_OT_F + (0.16) * MARGWORK_3_6_M + (0.16) * MARGWORK_3_6_F + (0.16) * MARG_CL_3_6_M + (0.15) * MARG_CL_3_6_F + (0.09) * MARG_AL_3_6_M + (0.07) * MARG_AL_3_6_F + (0.13) * MARG_HH_3_6_M + (0.11) * MARG_HH_3_6_F + (0.15) * MARG_OT_3_6_M + (0.14) * MARG_OT_3_6_F + (0.15) * MARGWORK_0_3_M + (0.15) * MARGWORK_0_3_F + (0.14) * MARG_CL_0_3_M + (0.13) * MARG_CL_0_3_F + (0.06) * MARG_AL_0_3_M + (0.06) * MARG_AL_0_3_F + (0.12) * MARG_HH_0_3_M + (0.11) * MARG_HH_0_3_F + (0.14) * MARG_OT_0_3_M + (0.14) * MARG_OT_0_3_F + (0.15) * NON_WORK_M + (0.14) * NON_WORK_F

(It is basically the first column in the above heatmap multiplied with its corresponding original feature.)