**Business Report**

**Digital Ads**

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 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:

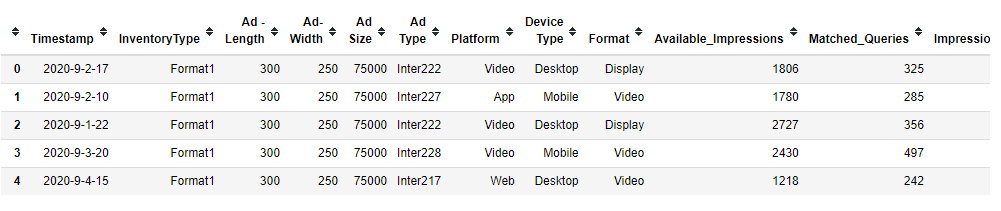
CPM = (Total Campaign Spend / Number of Impressions) \* 1,000

CPC = Total Cost (spend) / Number of Clicks

CTR = Total Measured Clicks / Total Measured Ad Impressions x 100

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

|  |  |  |
| --- | --- | --- |
| Data Dictionary | | |
|  |  |  |
|  |  |  |
| **Sl. No** | **Column Name** | **Column Description** |
| 1 | Timestamp | The Timestamp of the particular Advertisement. |
| 2 | InventoryType | The Inventory Type of the particular Advertisement. Format 1 to 7. This is a Categorical Variable. |
| 3 | Ad - Length | The Length Dimension of the particular Adverstisement. |
| 4 | Ad- Width | The Width Dimension of the particular Advertisement. |
| 5 | Ad Size | The Overall Size of the particular Advertisement. Length\*Width. |
| 6 | Ad Type | The type of the particular Advertisement. This is a Categorical Variable. |
| 7 | Platform | The platform in which the particular Advertisement is displayed. Web, Video or App. This is a Categorical Variable. |
| 8 | Device Type | The type of the device which supports the partciular Advertisement. This is a Categorical Variable. |
| 9 | Format | The Format in which the Advertisement is displayed. This is a Categorical Variable. |
| 10 | Available\_Impressions | How often the particular Advertisement is shown. An impression is counted each time an Advertisement is shown on a search result page or other site on a Network. |
| 11 | Matched\_Queries | Matched search queries data is pulled from Advertising Platform and consists of the exact searches typed into the search Engine that generated clicks for the particular Advertisement. |
| 12 | Impressions | The impression count of the particular Advertisement out of the total available impressions. |
| 13 | Clicks | It is a marketing metric that counts the number of times users have clicked on the particular advertisement to reach an online property. |
| 14 | Spend | It is the amount of money spent on specific ad variations within a specific campaign or ad set. This metric helps regulate ad performance. |
| 15 | Fee | The percentage of the Advertising Fees payable by Franchise Entities. |
| 16 | Revenue | It is the income that has been earned from the particular advertisement. |
| 17 | CTR | CTR stands for "Click through rate". CTR is the number of clicks that your ad receives divided by the number of times your ad is shown. Formula used here is CTR = Total Measured Clicks / Total Measured Ad Impressions x 100. Note that the Total Measured Clicks refers to the 'Clicks' Column and the Total Measured Ad Impressions refers to the 'Impressions' Column. |
| 18 | CPM | CPM stands for "cost per 1000 impressions." Formula used here is CPM = (Total Campaign Spend / Number of Impressions) \* 1,000. Note that the Total Campaign Spend refers to the 'Spend' Column and the Number of Impressions refers to the 'Impressions' Column. |
| 19 | CPC | CPC stands for "Cost-per-click". Cost-per-click (CPC) bidding means that you pay for each click on your ads. The Formula used here is CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column and the Number of Clicks refers to the 'Clicks' Column. |



A screenshot of a computer

Description automatically generated

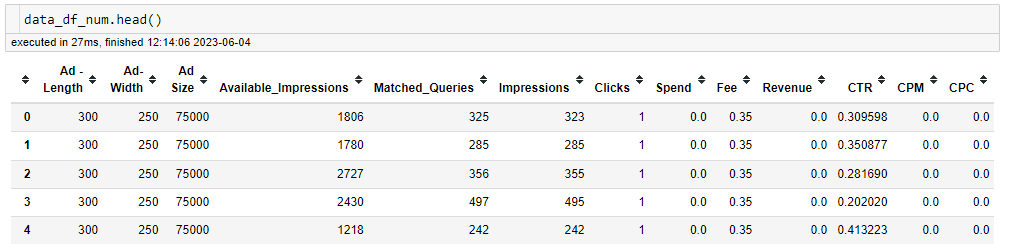
A screenshot of a computer

Description automatically generated with medium confidence

* Dataset provided in CSV format containing 23066 in rows and 19 columns.
* Data type in 6 float, 7 integer and 6 object
* Null values were found in the column Name: CTR, CPM and CPC
* There are 4736 missing values in CTR, CPM and CPC
* No Duplicate values found in the dataset

For further analysis, we bifurcate the data into Numerical column and Categorical column.

Categorical variables(data\_df\_cat) have Timestamp, Inventory type, Ad type, Platform, Device type and format categories, Remaining 13 variables are into Numerical column(data\_df\_num).



A screenshot of a computer

Description automatically generated with medium confidence

**Clustering: Treat missing values in CPC, CTR and CPM using the formula given**

Ad-Company has provided the formula to replace the missing values.

We have treated the null values in CPC using Two ways-

1. CPC = Total cost(spend)/Number of clicks, where clicks > 0
2. 0, where clicks = 0

We have treated the null values in CPM using Two ways-

1. CPM = (Total campaign spend/Number of impressions) x 1000 , where Ad impressions > 0
2. 0, where Ad impressions = 0

We have treated the null values in CTR using Two ways-

1. CTR = Total measured clicks/ Total Measured Ad impressions x 100, where Ad Impressions >0
2. 0, where Ad impressions = 0

**Clustering: 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.**

K-Means clustering is an unsupervised learning algorithm which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest centroid. The algorithm aims to minimize the squared Euclidean distances between the observation and the centroid of cluster to which it belongs.

A picture containing diagram, rectangle, plan, parallel

Description automatically generated

But sometime K-Means algorithm does not give best results. It is sensitive to outliers. An outlier is a point which is different from the rest of data points.

Hence, we will be removing outliers considering all the above mentioned factors.

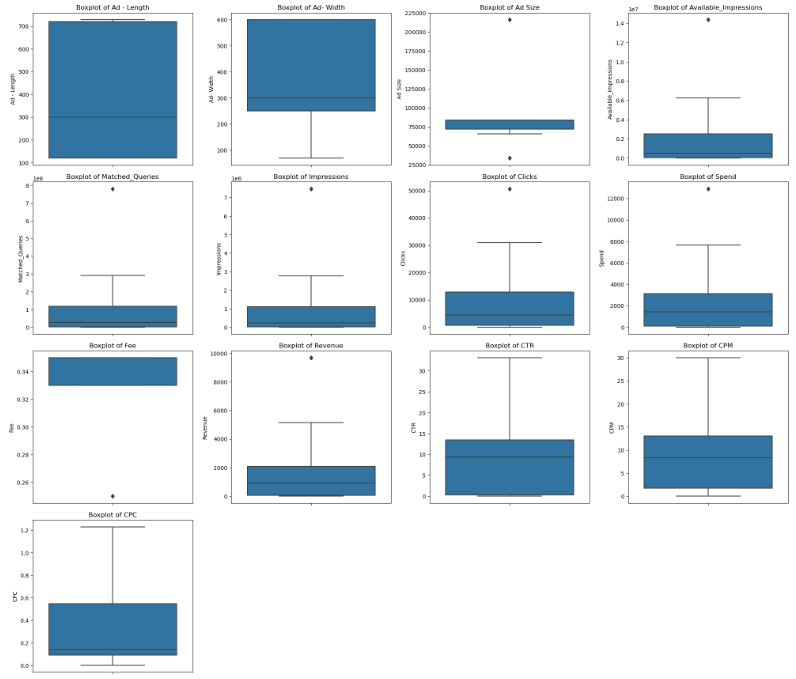
We have defined a function remove\_outlier, for the higher outliers we will treat it to get it at 95th percentile value and lower outliers will be treated to get at 25th percentile value.

Here we must note that the lower quartile’Q25’ is median of first half of data. The upper quartile ‘Q75’ is median of second half of data. The interquartile range ‘IQR’ is difference of Q75 and Q25. An outliers is a point that is greater than (Q75 +1.5 \* IQR) or lesser than (Q25 – 1.5 \* IQR).

Lower bound =Q25 – (1.5 x IQR)

Upper bound = Q75 + (1.5 x IQR)

Where the values are smaller than lower bound will get replaced by Q5 and where the values are bigger than upper bound will get replaced by Q95.



**Q4- Perform z-score scaling and discuss how it affects the speed of the algorithm.**

Scaling data after preprocessing replacing with non value with median with Z-Score.

A screenshot of a computer

Description automatically generated with low confidence

We observe in Ad-length, Ad\_size, Click, Spend fee columns are in hundred, where as Available Impressions, Matched queries Impressions column are very small digit (<5). So we scale the data so that are algorithm that we apply give balanced insights and they do not get influenced by only a few columns.

We split the data source into two separate files as Category(data\_df\_cat) and Numberic(data\_df\_num).

To scale the data we import Zscore from scipy.stats and we apply the Zscore to the data set (data\_df\_num) and the store the scaled variables in scaled\_df data set.

Scaling increases the speed of algorithm since the scale of all the variables becomes same.

**Q5- Perform Hierarchical by conducting a Dendogram using WARD and Euclidean distance.**

**Methods of Clustering**

There are two primary approaches to clustering; namely **Hierarchical or Agglomerative** clustering and **K-Means** clustering. In hierarchical clustering, the closest points are combined in a pairwise manner to form the clusters. It is an iterative procedure, where at every step of iteration, two points or two cluters are combined to form a bigger cluster. It is an iterative procedure, where at every step of the iteration, two points or two clusters are combined to form a bigger cluster. At the end of the process all points are combined into a single cluster. The number of clusters is not pre-determined.

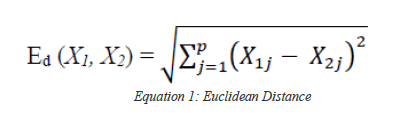
In the k-means clustering process, k denotes the number of clusters, which has to be pre-determined. Once k is fixed, the observations are allocated to one and only one cluster so that the closest points belong to one cluster. The cluster size is not controlled.

**Measures of Distance**

Since the clustering process is completely controlled by the distance between two points and distance between two clusters, it is paramount that the concept of distance is clear before we move on to the actual clustering process.

Clustering works only when the observations are multivariate. For univariate observations the concept of distance is trivial.

**Euclidean distance:** Between two obseervations X1 and X2, each of dimension p, Euclidean distance is defined as



**Ward’s Linkage:**  This is an alternative Method of clustering, also known as minimum variance clustering method. In that sense it is similar to the k-means clustering, an iterative method where after every merge, the distance are updated successively. Ward’s method often creates compact and even-sized clusters. However, the drawback is, it may result in less than optimal clusters.

**Hierarchical clustering**

Hierarchical clustering is a method of cluster analysis which seeks a build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

* Agglomerative : This is ‘Bottom-up’ approach: each observation start in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
* Divisive: This is a ‘Top-down’ approach: all observations start in one cluster, and splits are performed recursively as one move down the hierarchy.

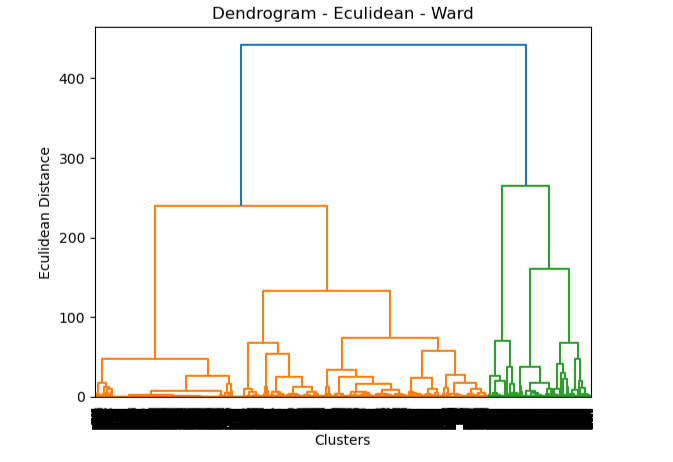
Using Scipy Cluster Hierarchy module to use dendogram and linkage function to processing Clusturing the Data.

After scaling we perform statistical significance of correlations using Batlett Spereicty. The P\_value as 0 which means that there is significant correlation in the dataset.

For performing hierarchical clustering wer import dendogram, linkage from scipy.cluster.hierarchy.

We use method **Ward** and the default distance taken is Euclidean distance.

**Dendrogram**



Since the clusters are not visible properly we use truncate method with number of clusters as 10 for better clarity.

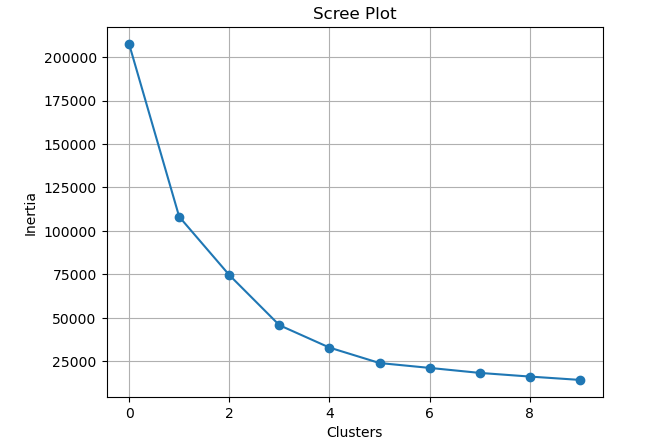
A picture containing text, diagram, plan, rectangle

Description automatically generated

**Q6 – Make Elbow plot (upto n=10) and identify optimum number of clusters for K-Mean algorithm.**

To make elbow plot we import KMeans for sklearn.cluster and get the **wss** values.

Using plt.plot we get below elbow plot for n= 10



Optimum number of clusters for K-Mean algorithm will be 3 since we can see between point 1 and point 2, point 2 and point 3

There is significant difference in the distance on Y axis but after that there is no significant difference in the distance points are very close to each other.

**Q7- Print silhouette scores for up to 10 clusters and idendity optimum number of clusters.**

To print silhouette soures we import silhouette\_samples, silhouette\_score from sklearn.metrics.

Silhouette score = 0.610 (which is positive – indicates that we have done clustering properly)

Optimum number of clusters that we take will be 3 since we can see on the elbow plot- between point 1 and 2, 2 and 3 there is significant difference in the distance on y axis but after that there is no significant difference in the distance, point are very close to each other.

**Q8- 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, & CTC based on Device Type. Make bar plots]**

A picture containing text, screenshot, font, number

Description automatically generated

A picture containing text, screenshot, font, line

Description automatically generated

A picture containing text, screenshot, diagram, line

Description automatically generated

A picture containing text, screenshot, line, diagram

Description automatically generated

A picture containing text, screenshot, line, diagram

Description automatically generated

A picture containing text, screenshot, diagram, line

Description automatically generated

A picture containing text, screenshot, line, diagram

Description automatically generated

To Identify the optimum number of cluster for k-means algorithm, silhouette scores provide this information .

Summary of Learnings:

A screenshot of a computer

Description automatically generated with low confidence

Cluster 0: High Ad Size with High Clicks with Medium Spend with High Fee generated Medium Revenue

Cluster 1: Low Ad Size with Low Clicks with High Spend with Medium Fee generated High Revenue

Cluster 2: Medium Ad Size with Medium Click with Low Spend with Low Fee generated Low Revenue

**Principle Component Analisys ( PCA)**

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.

**Principal Components:**

The concept of principal components is quite intuitive. Instead of dealing with a large number of possibly correlated variable, principal components are constructed as suitable linear combination of the observe variables such that the components have two important properties:

* The principal components(PCs) carry the total variance present in the data
* The PCs are orthogonal, i.e, uncorrelated, to one another

Information contents is the data is determined by variance of the attributes. A random variable whose variance is 0, is completely non-informative because for each unit this variable has the same value; in other words, this is a constant. Reduction of dimension involves sacrificing certain amount of variance. A balance must be struck so that significant reduction in the number of dimensions is achieved by sacrificing the least possible amount of variance.

**Q1:- Read The data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.**

**FEATURES AND THEIR DISCRIPTION**

The Dataset which provided with 640 Rows and 61 Columns

|  |  |
| --- | --- |
| Name | Description |
| State | State Code |
| District | District Code |
| Name | Name |
| TRU1 | Area Name |
| No\_HH | No of Household |
| TOT\_M | Total population Male |
| TOT\_F | Total population Female |
| M\_06 | Population in the age group 0-6 Male |
| F\_06 | Population in the age group 0-6 Female |
| M\_SC | Scheduled Castes population Male |
| F\_SC | Scheduled Castes population Female |
| M\_ST | Scheduled Tribes population Male |
| F\_ST | Scheduled Tribes population Female |
| M\_LIT | Literates population Male |
| F\_LIT | Literates population Female |
| M\_ILL | Illiterate Male |
| F\_ILL | Illiterate Female |
| TOT\_WORK\_M | Total Worker Population Male |
| TOT\_WORK\_F | Total Worker Population Female |
| MAINWORK\_M | Main Working Population Male |
| MAINWORK\_F | Main Working Population Female |
| MAIN\_CL\_M | Main Cultivator Population Male |
| MAIN\_CL\_F | Main Cultivator Population Female |
| MAIN\_AL\_M | Main Agricultural Labourers Population Male |
| MAIN\_AL\_F | Main Agricultural Labourers Population Female |
| MAIN\_HH\_M | Main Household Industries Population Male |
| MAIN\_HH\_F | Main Household Industries Population Female |
| MAIN\_OT\_M | Main Other Workers Population Male |
| MAIN\_OT\_F | Main Other Workers Population Female |
| MARGWORK\_M | Marginal Worker Population Male |
| MARGWORK\_F | Marginal Worker Population Female |
| MARG\_CL\_M | Marginal Cultivator Population Male |
| MARG\_CL\_F | Marginal Cultivator Population Female |
| MARG\_AL\_M | Marginal Agriculture Labourers Population Male |
| MARG\_AL\_F | Marginal Agriculture Labourers Population Female |
| MARG\_HH\_M | Marginal Household Industries Population Male |
| MARG\_HH\_F | Marginal Household Industries Population Female |
| MARG\_OT\_M | Marginal Other Workers Population Male |
| MARG\_OT\_F | Marginal Other Workers Population Female |
| MARGWORK\_3\_6\_M | Marginal Worker Population 3-6 Male |
| MARGWORK\_3\_6\_F | Marginal Worker Population 3-6 Female |
| MARG\_CL\_3\_6\_M | Marginal Cultivator Population 3-6 Male |
| MARG\_CL\_3\_6\_F | Marginal Cultivator Population 3-6 Female |
| MARG\_AL\_3\_6\_M | Marginal Agriculture Labourers Population 3-6 Male |
| MARG\_AL\_3\_6\_F | Marginal Agriculture Labourers Population 3-6 Female |
| MARG\_HH\_3\_6\_M | Marginal Household Industries Population 3-6 Male |
| MARG\_HH\_3\_6\_F | Marginal Household Industries Population 3-6 Female |
| MARG\_OT\_3\_6\_M | Marginal Other Workers Population Person 3-6 Male |
| MARG\_OT\_3\_6\_F | Marginal Other Workers Population Person 3-6 Female |
| MARGWORK\_0\_3\_M | Marginal Worker Population 0-3 Male |
| MARGWORK\_0\_3\_F | Marginal Worker Population 0-3 Female |
| MARG\_CL\_0\_3\_M | Marginal Cultivator Population 0-3 Male |
| MARG\_CL\_0\_3\_F | Marginal Cultivator Population 0-3 Female |
| MARG\_AL\_0\_3\_M | Marginal Agriculture Labourers Population 0-3 Male |
| MARG\_AL\_0\_3\_F | Marginal Agriculture Labourers Population 0-3 Female |
| MARG\_HH\_0\_3\_M | Marginal Household Industries Population 0-3 Male |
| MARG\_HH\_0\_3\_F | Marginal Household Industries Population 0-3 Female |
| MARG\_OT\_0\_3\_M | Marginal Other Workers Population 0-3 Male |
| MARG\_OT\_0\_3\_F | Marginal Other Workers Population 0-3 Female |
| NON\_WORK\_M | Non Working Population Male |
| NON\_WORK\_F | Non Working Population Female |

There are No Duplicated rows.

Basic information about the dataset.

A screenshot of a computer

Description automatically generated with medium confidence

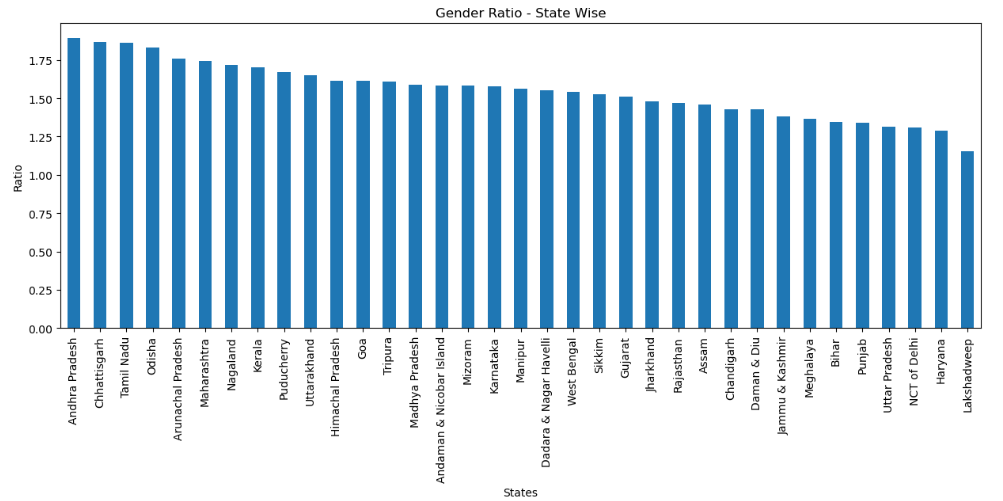
Drop the colums StateCode, Distcode, State and Area Name since the categorical variables and we analyze the numerical columns stored in the dataset.

**Q2:- Perform detailed Exploratory analysis by creating certain questions like :**

1. **Which state has the highest gender ratio and which has the lowest?**

Andhra Pradesh state has the highest female to male ratio(1.89) . The Union Territory of Laksadweep has the lowest dender ration of 1.15.

Haryana state have the lowest gender ratio.

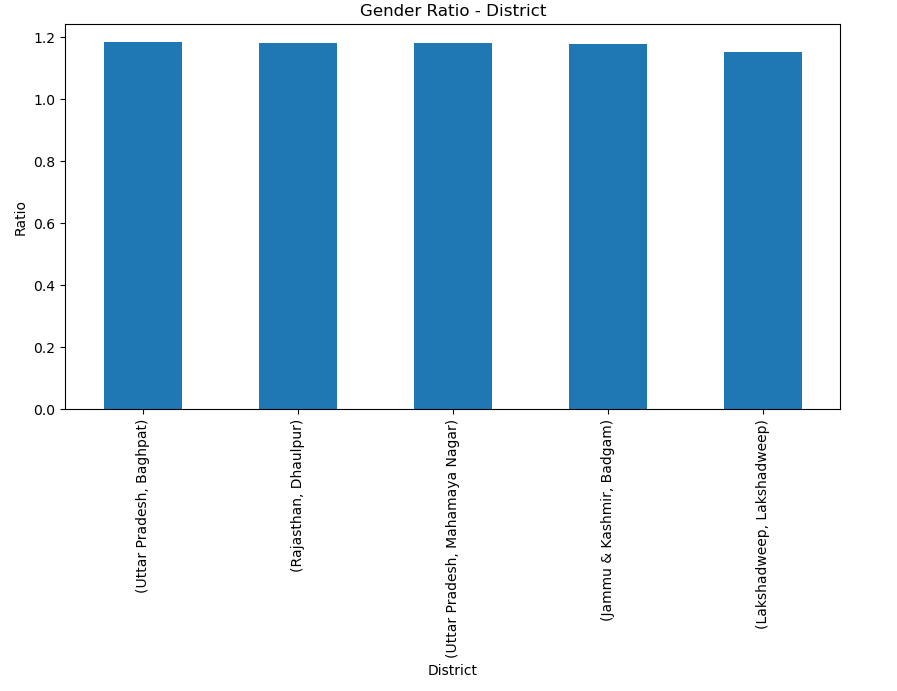


**Which district has the highest and lowest gender ratio?**

* Krishna District of Andhra Pradesh has the highest Female to Male ratio of 2.28
* Badgam District of Jammu & Kashmir has the lowest female to male ratio of 1.17

**A picture containing text, screenshot, parallel, diagram

Description automatically generated**

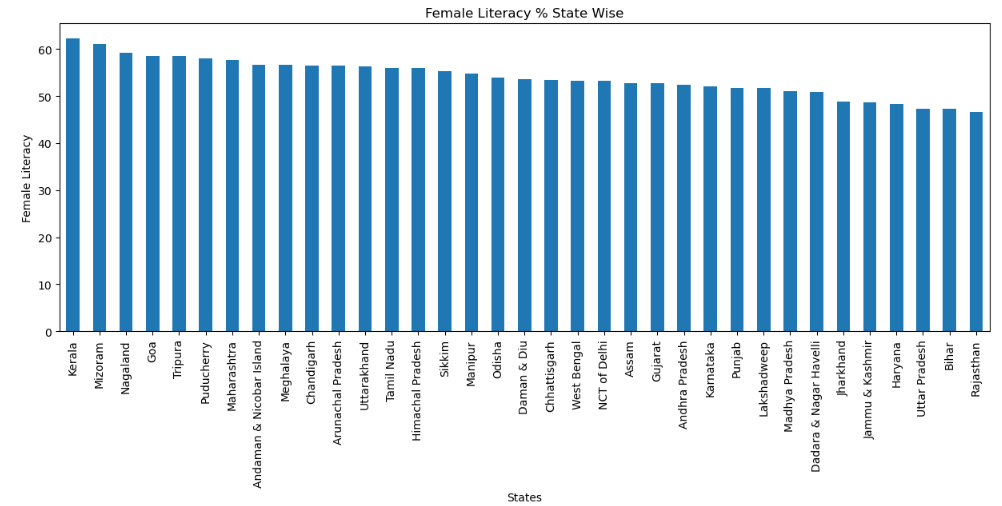
****

1. **Literacy**

**Female Literacy rate is defined as the**

Number of literate females/Total Literate Population \* 100

Kerala State is at the top in Literacy while Rajasthan at the bottom



**3. Non-working population**

Uttar Pradesh has the most ‘non-working’ population. Kerala has most ‘non-working’ female population after Uttar Pradesh.

Daman & Diu and Dadra Nagar Haveli have the lowest number of non-working population for both Females and Males.

A picture containing text, screenshot, diagram, plot

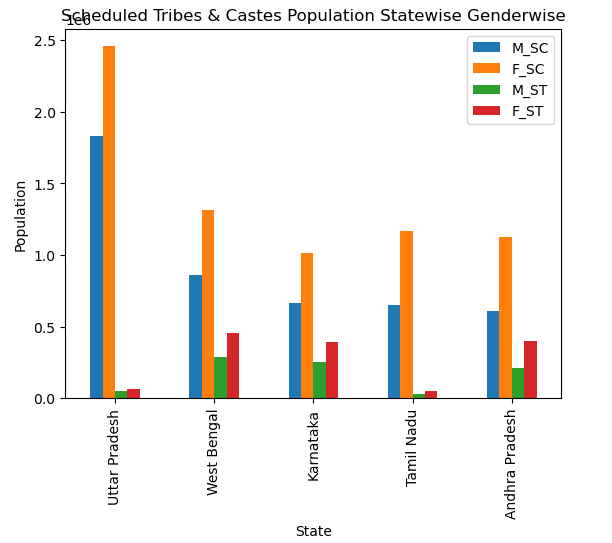
Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

**4. Statewise SC/ST population by gender**

Uttar Pradesh has the highest number of SC/ST population. It is also observed that SC population is significantly higher that ST population. It is also noted that there are more SC Females than Males.



### **State has highest & lowest gender ratio**

* Andhra Pradesh state has the highest Female to Male Ratio(1.89), while the Union Teritory Lakshadweep has the lowest gender ratio of 1.15. Among the States, Haryana have the lowest gender ratio of 1.28(Female to Male).

### **District has highest & lowest gender ratio**

* Krishna District of Andhra Pradesh has the highest Female to Male Ratio of 2.28, Badgam District of Jammu & Kashmir has the lowest Female to Male Ratio of 1.17

### **Literacy**

* Kerala is at the top while Rajasthan & Bihar at the bottom.
* Non-Working Population - Uttar Pradesh has the most 'non-working' population. Kerala has most 'non-working' female population. Daman & Diu and Dadra Nagar Haveli have the lowest number of 'non-working' population for both Female & Male

### **Statewise SC/ST population by gender**

* Uttar Pradesh has the highest number of SC/ST population. It is also observed that SC population is significantly higher than ST population.
* There are more SC Females than Males

###### 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

**A picture containing diagram, text, plan, map

Description automatically generated**

**Data are Right skew and present of Outliers**

**A screenshot of a computer screen

Description automatically generated with low confidence**

**Q3:- We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?**

Yes, removing outliers should be treated before doing PCA, since PCA is sensitive to outliers which may lead to inefficient results if not treated otherwise. The dataset has a lot of outliers which needs to be removed. This is traced through Boxplot.

**Q4:- Scale the Data using z-score method. Does scaling have any impact on ouliers? Compare boxplot before and after scaling and comment.**

Scaling does not have an effect on ouliers, scaling only reduces the mangnitude

Before performing PCA we apply Z-score method scale the data since the numerical variable are highly different in magnitude.

We check the correlation of variable, and look at the pairplot. We see there is some correlation in the given variable.

**Perform Bartletts Test of Sphericity**

**Bartlett’s, test of sphericity test the hypothesis that the variables are uncorrelated in the population**

**H0**: All variable in the data are uncorrelated

HA: At least one pair of variables in the data are correlated

If the null hypothesis cannot be rejected, then PCA is not required

**KMO Test**

**The Kaiser-Meyer-Olkin(KMO) – measure of sampling adequacy(MSA) is an index used to examine how appropriate pca is**

Since we got kmo value = 0.936, which is greater than 0.7 this show us that we have adequate sample size to perform PCA.

After apply PCA taking all 57 features, we have 61 features(dropped a few feature).

For this we import PCA for sklear.decomposition and we put random state = 123 so that we get the same results everytime we perform the algorithms. And we performed pca.fit\_tranform to the scaled dataset.

**Covariance Matrix:**

A picture containing text, screenshot, font, number

Description automatically generated

**PCA Components(Eign Vectors):**

**A screenshot of a computer code

Description automatically generated with low confidence**

**Eigen Values:**

**A picture containing text, screenshot, font, document

Description automatically generated**

**Explained Variance Ratio:**

A screenshot of a computer code

Description automatically generated with low confidence

**Q5- Identify the optimum number of PCs(for the project, take at least 90% explained variance). Show Scree plot.**

**Cumulative explained variance ratio:**

**A picture containing text, screenshot, font, number

Description automatically generated**

**A graph with a number of components

Description automatically generated with low confidence**

**Although there are 57 observed variables the first 5 principal components can explain more than 90% of total variation, hence the optimum number of PC’s that we take is 5 pcs instead of original 57, thereby reducing the dimensions of a significant number.**

**We also look at the scree plot which has indices of the PCs on the X-axis and variances on the Y-axis. We can see that at point 5 there is a distinct break point and the line joining the variances(elbow point) becomes approximately horizontal.**

**Q6:-Compare PCs with Actual Columns and identiy which is explaining most variance. Write inference about all the Principal components in terms of actual variables.**

The Names of PC’s are:-

PC1:- Skilled\_unskilled\_workforce

PC2:-Marginalized\_working\_population

PC3:-Female\_workforce

PC4:-Household\_population

PC5:- SC/ST workforce

A screenshot of a graph

Description automatically generated with low confidence

A picture containing screenshot, line, plot

Description automatically generated

A picture containing screenshot, line, plot, diagram

Description automatically generated

A picture containing line, screenshot, plot, diagram

Description automatically generated

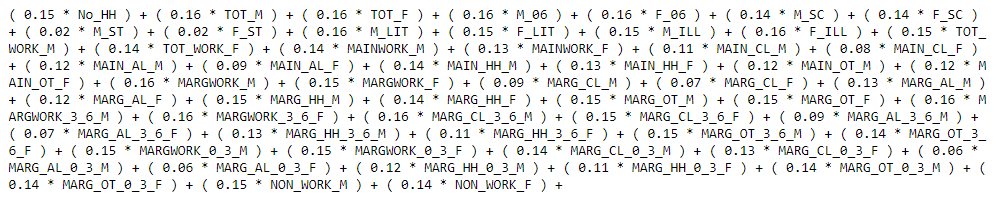
A screenshot of a graph

Description automatically generated with low confidence

A picture containing text, font, line, screenshot

Description automatically generated

**Q7- Write linear equation for first PC.**

****