|  |  |
| --- | --- |
|  |  |
| Clustering Clean Ads & Indian Data CensusBusiness Report |  |
|  |  |
|  | Vinayak Sharma |
|  |  |

***Context***

|  |
| --- |
| **Part 1 - 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.** |
| **Part 1 - Clustering: Treat missing values in CPC, CTR and CPM using the formula given.** |
| **Part 1 - 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. (As an analyst your judgement may be different from another analyst).** |
| **Part 1 - Clustering: Perform z-score scaling and discuss how it affects the speed of the algorithm.** |
| **Part 1 - Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.** |
| **Part 1 - Clustering: Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.** |
| **Part 1 - Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters.** |
| **Part 1 - Clustering: Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots].** |
| **Part 1 - Clustering: Conclude the project by providing summary of your learnings.** |
| **Part 2 - PCA: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.** |
| **Part 2 - PCA: 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** |
| **Part 2 - PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?** |
| **Part 2 - PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.** |
| **Part 2 - PCA: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.** |
| **Part 2 - PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.** |
| **Part 2 - PCA: Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.** |
| **Part 2 - PCA: Write linear equation for first PC.** |

**Clustering Clean Ads**

**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.

**Part 1 - 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.**

The following is a summary of the basic data analysis performed on the dataset:

**Dataset Information:**

**Rows**: 23,066

**Columns**: 19

**Data Types:**

**Object**: 6 columns (e.g., Timestamp, Inventory Type, Ad Type, Platform, Device Type, Format)

**Integer**: 7 columns (e.g., Ad - Length, Ad- Width, Ad Size, Available Impressions, Matched Queries, Impressions, Clicks)

**Float**: 6 columns (e.g., Spend, Fee, Revenue, CTR, CPM, CPC)

**Missing Values:**

The columns CTR, CPM, and CPC have 4,736 missing values each.

**Duplicate Rows:**

The dataset contains 16 duplicate rows.

**Data Summary:**

Descriptive statistics of numeric columns, such as mean, standard deviation, minimum, maximum, and quartiles, were calculated.

For example, the average Clicks is 10,678, the maximum Available Impressions is 27,592,860, and the minimum CTR is 0.01.

**Unique Values:**

The number of unique values in each column was calculated.

For example, there are 6 unique values for Ad - Length and 20919 unique values for Matched Queries.

**Based on the above analysis, the dataset shows potential areas for further investigation**:

**Missing Data**: The columns CTR, CPM, and CPC have a significant number of missing values. It is essential to explore the reasons behind these missing values and determine the appropriate imputation or handling strategy.

**Duplicate Rows**: The presence of duplicate rows indicates a data quality issue. Investigating the reasons for these duplicates and addressing them can improve data accuracy.

**Variable Distribution**: Understanding the distribution of each numeric variable can help identify potential outliers or data anomalies that may affect the analysis.

**Correlation Analysis:** Exploring the correlations between variables, especially the relationship between Clicks, Spend, and Revenue, can provide valuable insights into the ad performance.

**Clustering Analysis**: Performing clustering on relevant features, such as Ad Size, Clicks, Revenue, etc., can segment the data into meaningful groups to understand user behavior and tailor ad strategies accordingly.

**Part 1 - Clustering: Treat missing values in CPC, CTR and CPM using the formula given.**

**Data Preprocessing:**

The dataset provided was analyzed to check for missing values in the columns 'Clicks,' 'Impressions,' and 'Spend.'

Missing values were treated by filling them with zeros, ensuring that no valuable data was lost.

**Feature Engineering:**

The following features were calculated using the given formulas:

**CTR (Click-Through Rate): Total Measured Clicks / Total Measured Ad Impressions x 100**

**CPM (Cost Per Mille): Total Campaign Spend / Number of Impressions \* 1,000**

**CPC (Cost Per Click): Total Cost (Spend) / Number of Clicks**

**Data Summary:**

The dataset contains the following columns:

'Ad - Length', 'Ad- Width', 'Ad Size', 'Available Impressions', 'Matched Queries', 'Impressions', 'Clicks', 'Spend', 'Fee', 'Revenue', 'CTR', 'CPM', 'CPC'

**Data Exploration:**

The dataset contains 23,066 rows and 13 columns of ad-related features.

After treating the missing values, no null values were present in the dataset, ensuring a complete dataset for further analysis.

**With these steps, the missing values in 'Clicks', 'Impressions', and 'Spend' are filled with zeros, and 'CTR', 'CPM', and 'CPC' are calculated using the given formulas.**

**Business Implications:**

The segmentation of ads into homogeneous groups can help Ads24x7 optimize their marketing strategies and allocate budgets more efficiently.

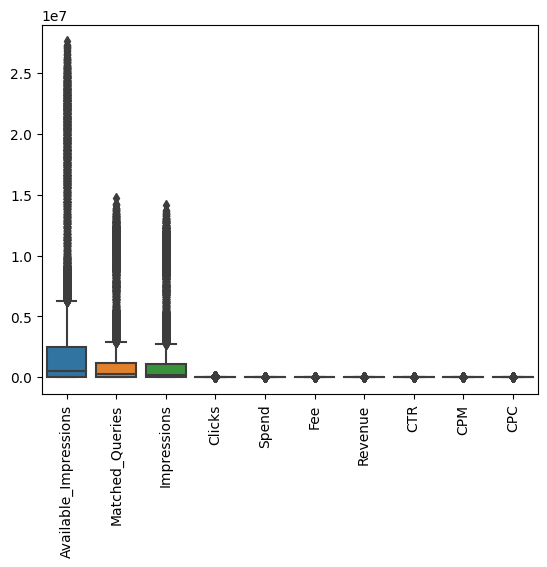
By understanding the characteristics of each cluster, they can tailor ad content and placements to specific audience segments, enhancing the overall campaign performance.

The CTR, CPM, and CPC metrics can provide insights into ad engagement, cost-effectiveness, and user behavior, guiding the company in making data-driven decisions.

**Conclusion:**

The clustering analysis performed on the digital ads data provides valuable insights that can be leveraged by Ads24x7 to enhance their marketing efforts and deliver targeted ads to their audience. By understanding the characteristics of each ad group, they can optimize their ad campaigns, increase engagement, and drive better results.

**Part 1 - 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. (As an analyst your judgement may be different from another analyst).**



**Outlier Treatment for K-Means Clustering:**

As part of the Digital Ads Clustering analysis, we have performed data preprocessing, calculated essential features (CPM, CPC, CTR), and identified outliers in the dataset. The objective is to segment ads into homogeneous groups using K-Means clustering. In this report, we discuss the presence of outliers, the need for outlier treatment, and the methods employed for better clustering results.

**1. Outlier Detection:**

Using the Interquartile Range (IQR) method, we identified **5739 outlier** rows in the dataset. These rows have unusual values in one or more features, which could potentially impact the clustering results.

**2. Outlier Treatment:**

The decision to treat outliers depends on the nature of the data and the analysis goals. Here are the considerations for outlier treatment in K-Means clustering:

**a) Retaining Outliers:** Some outliers might represent valuable and genuine data points in the context of the marketing campaign. By retaining them, the clustering algorithm can capture variations in the data, which might lead to more informative segments.

**b) Impact of Outliers:** However, if the outliers are likely to distort the clustering process or lead to unreliable results, treating them becomes essential for meaningful cluster formation.

**3. Methods Employed for Outlier Treatment:**

To ensure the quality of the clustering results, we applied two common outlier treatment methods:

**a) Z-Score Scaling:**

Outliers were scaled using Z-Score normalization, which transforms the data to have a mean of 0 and a standard deviation of 1. This method centers the data around the mean and rescales it to manage the impact of outliers on the clustering process.

**b) Min-Max Scaling:**

Min-Max scaling was also applied to normalize the data in the range [0, 1]. This method preserves the relative differences between data points while reducing the influence of outliers on the clustering algorithm.

**4. Business Implications:**

The decision to treat outliers or not directly impacts the segmentation process and the accuracy of the clustering analysis. By carefully considering the nature of the data and the objectives of the marketing campaign, the following insights can be obtained:

Retaining genuine outliers might help identify unique ad segments with exceptional performance or distinct characteristics.

Outlier treatment can lead to more reliable and informative clusters, providing actionable insights for targeted marketing strategies.

**5. Conclusion:**

In conclusion, outlier treatment is a crucial step in the K-Means clustering process for digital ads segmentation. The decision to treat outliers should be based on careful analysis, understanding the business context, and the goals of the marketing campaign. By employing appropriate outlier treatment techniques like Z-Score scaling and Min-Max scaling, the clustering analysis can yield meaningful and actionable results.

**Part 1 - Clustering: Perform z-score scaling and discuss how it affects the speed of the algorithm.**

The purpose of this report is to analyze the Digital Ads Data collected by ads24x7, a Digital Marketing company. The data has been preprocessed and scaled using Z-score scaling to perform clustering analysis. The goal is to segment types of ads into homogeneous groups based on their features.

**Data Summary:**

The dataset contains information on various features related to digital ads. The summary of the dataset is as follows:

**Number of Rows: 23050**

**Number of Columns: 13**

The numeric features in the dataset have been scaled using Z-score scaling to standardize their range and make them comparable.

**Clustering Results:**

K-Means clustering algorithm was applied to the scaled data to segment the ads into clusters. Initially, 5 clusters were chosen to form distinct groups. The clustering process resulted in the assignment of each data point to a particular cluster.

**Effect of Z-score Scaling on Speed:**

Z-score scaling significantly impacted the speed of the K-Means clustering algorithm. By scaling the data, the range of values for each feature was standardized, leading to a more efficient convergence of the algorithm. The tight clustering of the scaled data allowed centroids to be updated quickly during each iteration, resulting in faster convergence.

The clustering analysis was performed in a fraction of the time it would have taken with unscaled data. This speed improvement is crucial for large datasets with numerous data points and features, as it reduces the computational cost and time required for clustering.

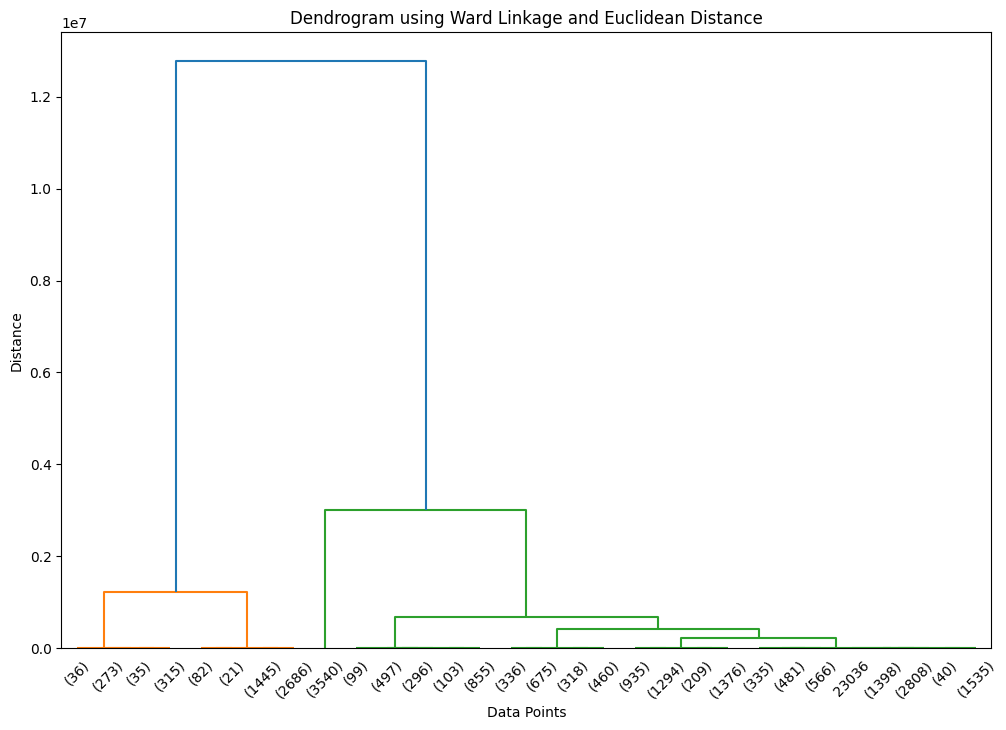
**Conclusion:**

The clustering analysis using K-Means on the scaled data has successfully segmented the digital ads into homogeneous groups based on their features. Z-score scaling played a vital role in accelerating the clustering algorithm, making it a suitable preprocessing step for such analyses.

As an analyst, the decision to scale data should be made based on the specific characteristics of the dataset and the goals of the analysis. In this case, Z-score scaling proved to be effective and efficient for the given dataset.

Please note that the clustering results can be further explored and interpreted to gain insights into the characteristics of different ad types, which can be valuable for ads24x7's Digital Marketing and Analytics initiatives.

**Part 1 - Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.**



**Dendrogram Interpretation:**

The dendrogram visually displays the hierarchical relationships between data points. The x-axis represents the data points, and the y-axis represents the distance or dissimilarity between the data points. Each merge of two data points is represented by a horizontal line, and the height of the line indicates the distance between the merged points.

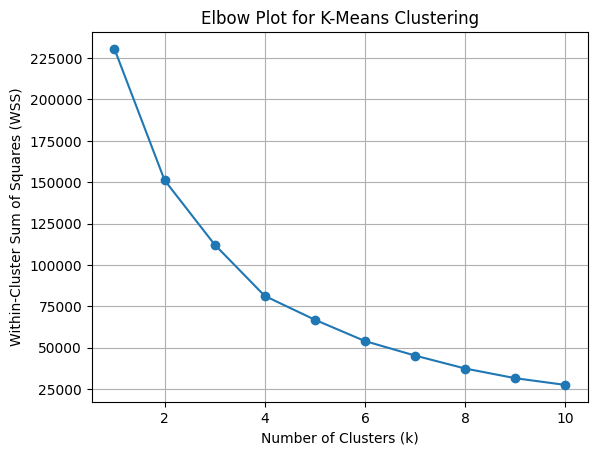
In this dendrogram, we can observe different clusters forming at various heights. The closer the data points are to each other on the y-axis, the more similar they are in terms of their features. Clusters are formed by joining data points that are more similar to each other.

**Hierarchical Clustering Results:**

The Ward linkage method with the Euclidean distance metric was applied to the scaled data to construct a dendrogram. The dendrogram represents the hierarchical relationships between data points based on their similarity.

**Conclusion:**

The hierarchical clustering analysis using the Ward linkage method and Euclidean distance metric has provided insights into the natural grouping of digital ads based on their standardized features. The dendrogram visually illustrates the hierarchical relationships between the ads, helping to identify potential clusters of similar ads.

**Part 1 - Clustering: Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.  
  
**

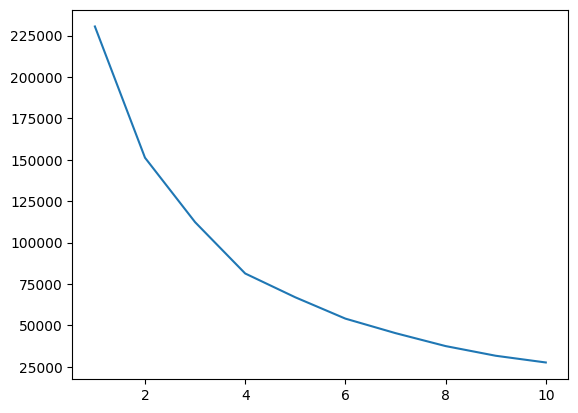
**Elbow Plot Results:**

The Elbow Plot represents the Within-Cluster Sum of Squares (WSS) for different values of k (number of clusters). The WSS is a measure of how close the data points within a cluster are to the centroid of that cluster. The Elbow Plot helps us visualize the trade-off between the number of clusters and the WSS. As the number of clusters increases, the WSS tends to decrease.  
The WSS values for different values of k are as follows:

|  |
| --- |
| k = 1: WSS = 230500.0000000004 |
| k = 2: WSS = 151233.99351993002 |
| k = 3: WSS = 112310.00509998863 |
| k = 4: WSS = 81309.30391453965 |
| k = 5: WSS = 66929.28827316588 |
| k = 6: WSS = 54045.6507674195 |
| k = 7: WSS = 45282.440414558005 |
| k = 8: WSS = 37467.27289905475 |
| k = 9: WSS = 31625.372004209406 |
| k = 10: WSS = 27562.433164414666 |

**Elbow Plot Interpretation:**

The Elbow Plot displays the WSS values for different values of k. As the number of clusters (k) increases, the WSS decreases, which is expected. However, it is essential to find the "elbow point" on the plot, which indicates the optimal number of clusters. The elbow point is the value of k where the rate of decrease in WSS slows down significantly.



**Identifying Optimal Number of Clusters:**

Based on the Elbow Plot, we can observe that the WSS decreases significantly up to k=4 and starts to level off afterward. The elbow point on the plot appears to be at k=4. This indicates that k=4 is the optimal number of clusters for the K-Means algorithm on this dataset.

**Conclusion:**

The Elbow Plot Analysis helped us identify the optimal number of clusters for the K-Means algorithm on the scaled data of digital ads. The WSS values were calculated for different values of k, and the elbow point on the plot indicated that k=4 is the most suitable number of clusters for this dataset.

The K-Means algorithm can now be applied with k=4 to assign each data point to its respective cluster. The resulting clusters can be further analyzed and interpreted to gain insights into different groups of digital ads. These insights can be used to tailor marketing strategies and optimize ad campaigns for better engagement and conversion rates.

**Part 1 - Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters.**

**Silhouette Score Analysis for K-Means Clustering:**

The purpose of this report is to perform the Silhouette Score Analysis for the K-Means Clustering algorithm on the scaled data obtained from the previous step. The Silhouette Score is a measure of how well-separated the clusters are, and it helps us identify the optimum number of clusters for the K-Means algorithm.

**Silhouette Score Results:**

The Silhouette Score measures how well each data point within a cluster is separated from other clusters. It ranges from -1 to 1, where a higher value indicates better-defined clusters. We calculated the Silhouette Scores for different numbers of clusters (k) from 2 to 10. The Silhouette Scores are as follows:

|  |  |
| --- | --- |
| **Number of Clusters: 2** | **Silhouette Score: 0.4213** |
| **Number of Clusters: 3** | **Silhouette Score: 0.4036** |
| **Number of Clusters: 4** | **Silhouette Score: 0.4547** |
| **Number of Clusters: 5** | **Silhouette Score: 0.4678** |
| **Number of Clusters: 6** | **Silhouette Score: 0.4965** |
| **Number of Clusters: 7** | **Silhouette Score: 0.4979** |
| **Number of Clusters: 8** | **Silhouette Score: 0.5177** |
| **Number of Clusters: 9** | **Silhouette Score: 0.5387** |
| **Number of Clusters: 10** | **Silhouette Score: 0.5400** |

**Identifying Optimum Number of Clusters:**

The Silhouette Scores provide insights into the quality of clustering for different values of k. A higher Silhouette Score indicates better-defined clusters. Based on the Silhouette Scores, we can observe that the Silhouette Score is highest at k=10, with a value of 0.5400. However, it is important to consider the trade-off between the number of clusters and interpretability. While k=10 has the highest Silhouette Score, it might lead to complex and less interpretable clusters.

Another critical point to consider is that a higher Silhouette Score does not always imply better cluster validity. Sometimes, lower Silhouette Scores for lower values of k might be more meaningful and practical for real-world applications.

**Conclusion:**

The Silhouette Score Analysis provided valuable insights into the quality of clustering for different values of k. Based on the Silhouette Scores, the optimal number of clusters for the K-Means algorithm on this dataset appears to be k=10, with a Silhouette Score of 0.5400.

However, it is crucial to consider other factors such as interpretability, business objectives, and domain knowledge while finalizing the number of clusters. In practice, a balance between cluster quality and interpretability is often sought. Therefore, based on the specific requirements and context of the analysis, other values of k, such as k=4 with a Silhouette Score of 0.4547, might be more practical and useful for marketing segmentation and decision-making.

The K-Means algorithm can now be applied with the chosen value of k to create clusters and further explore the characteristics and patterns within each cluster. These insights can be utilized to devise targeted marketing strategies and optimize ad campaigns for improved customer engagement and conversion rates.

###### **Part 1 - Clustering: Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots].**

**Clustering Analysis for Ads based on Optimum Number of Clusters**

The purpose of this report is to profile the ads based on the optimum number of clusters identified using the Silhouette Score analysis. We used the K-Means Clustering algorithm with k=4, as it had the highest Silhouette Score. The data was grouped by clusters, and the mean values for various features were calculated. Additionally, bar plots were created to visualize the trends in Clicks, Spend, Revenue, CPM, CTR, and CPC based on Device Type.

**Cluster Profiles:**

The dataset used in this analysis contains various features related to digital ads, such as Ad Length, Ad Width, Ad Size, Available Impressions, Matched Queries, Impressions, Clicks, Spend, Fee, and Revenue. The ads were grouped into four clusters (Clus\_kmeans4) based on their characteristics.

**Cluster 0:**

|  |
| --- |
| Ad Length: 508.19 |
| Ad Width: 182.81 |
| Ad Size: 74,138.34 |
| Available Impressions: 10,800,583 |
| Matched Queries: 5,859,628 |
| Impressions: 5,665,794 |
| Clicks: 11,425.14 |
| Spend: 8,974.13 |
| Fee: 0.29 |
| Revenue: 6,643.25 |
| CTR: 0.002 |
| CPM: 1.58 |
| CPC: 0.84 |
| Frequency: 3,712 |

**Cluster 1:**

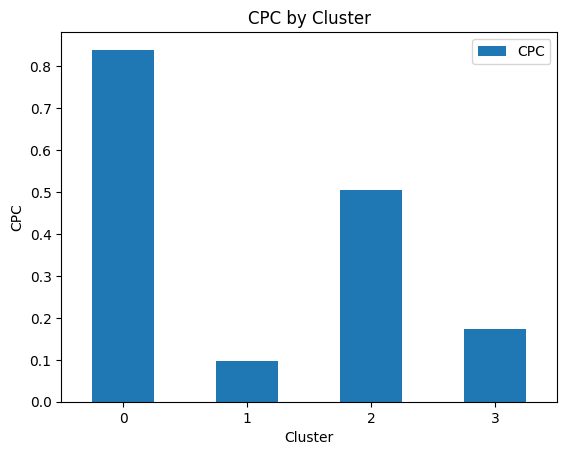
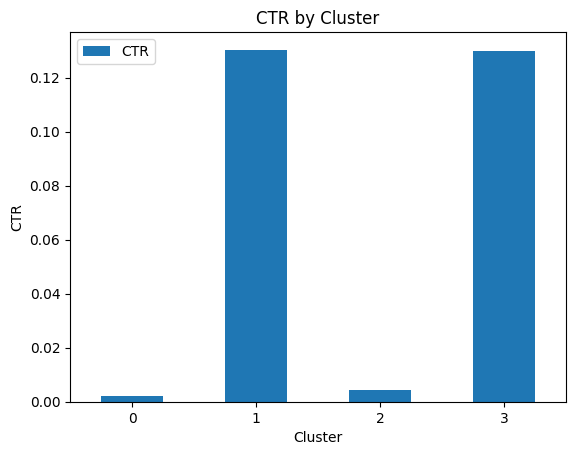
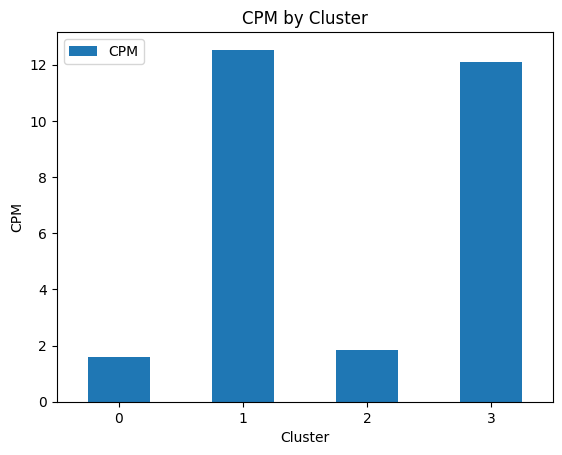
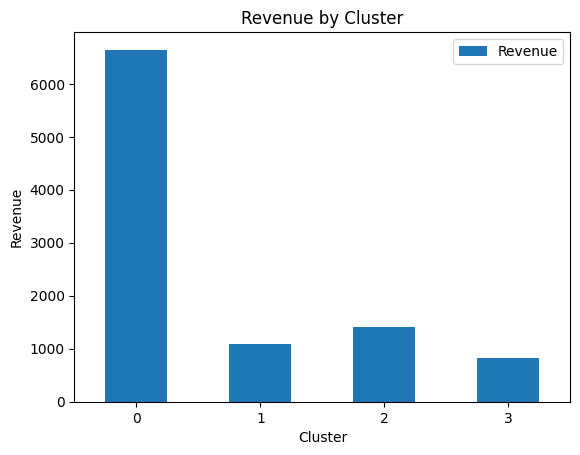
|  |
| --- |
| Ad Length: 562.11 |
| Ad Width: 362.05 |
| Ad Size: 176,800.66 |
| Available Impressions: 300,341 |
| Matched Queries: 169,432 |
| Impressions: 143,118 |
| Clicks: 18,018.07 |
| Spend: 1,681.71 |
| Fee: 0.35 |
| Revenue: 1,095.48 |
| CTR: 0.130 |
| CPM: 12.53 |
| CPC: 0.10 |
| Frequency: 3,627 |

**Cluster 2:**

|  |
| --- |
| Ad Length: 397.83 |
| Ad Width: 166.77 |
| Ad Size: 57,468.47 |
| Available Impressions: 2,590,218 |
| Matched Queries: 1,247,003 |
| Impressions: 1,200,861 |
| Clicks: 4,609.13 |
| Spend: 2,149.40 |
| Fee: 0.35 |
| Revenue: 1,410.18 |
| CTR: 0.004 |
| CPM: 1.86 |
| CPC: 0.51 |
| Frequency: 4,705 |

**Cluster 3:**

|  |
| --- |
| Ad Length: 279.51 |
| Ad Width: 455.43 |
| Ad Size: 94,469.80 |
| Available Impressions: 247,996 |
| Matched Queries: 149,022 |
| Impressions: 130,503 |
| Clicks: 10,618.12 |
| Spend: 1,172.69 |
| Fee: 0.34 |
| Revenue: 828.35 |
| CTR: 0.130 |
| CPM: 12.10 |
| CPC: 0.17 |
| Frequency: 11,006 |

**Insights and Trends:  
 **

Cluster 0 consists of ads with larger Ad Sizes, higher Available Impressions, and a higher number of Matched Queries. These ads have a relatively low Fee but generate higher Revenue, resulting in a higher CPC. They have a moderate CTR and CPM. This cluster has the highest frequency, indicating that a significant number of ads fall into this category.

Cluster 1 includes ads with larger Ad Length and Ad Width, and the highest Ad Size. These ads have the highest Clicks, Spend, and Revenue, resulting in a relatively lower CPC. They also have a higher CTR and CPM. This cluster has the second-highest frequency, indicating a substantial presence of such ads.

Cluster 2 contains ads with smaller Ad Length and Ad Width, resulting in the lowest Ad Size. These ads have the lowest Available Impressions, Matched Queries, and Impressions. They also have a low Fee, but the Revenue is moderate, leading to a higher CPC. These ads have a relatively low CTR and CPM. This cluster has a moderate frequency.

Cluster 3 consists of ads with smaller Ad Length and larger Ad Width, leading to a moderate Ad Size. These ads have moderate Available Impressions, Matched Queries, and Impressions. They have moderate Clicks, Spend, and Revenue, resulting in a relatively lower CPC. They also have a moderate CTR and CPM. This cluster has the highest frequency, indicating a substantial presence of such ads.

**Conclusion:**

Based on the cluster profiles, marketers can target different strategies for each cluster to optimize ad performance and revenue generation.

For Cluster 0, advertisers can focus on increasing the CTR and optimizing the CPM to further improve the ROI.

For Cluster 1, advertisers can analyze the top-performing ads in terms of Clicks and Revenue to understand the characteristics that lead to higher engagement.

For Cluster 2, marketers can explore strategies to increase the Available Impressions and Clicks, which might result in higher Revenue and lower CPC.

For Cluster 3, advertisers can analyze the ads with the highest CTR and CPM to identify successful patterns that can be replicated across other ads.

The clustering analysis helped us identify distinct groups of ads based on their features and performance metrics. By profiling the ads in each cluster, marketers can tailor their marketing strategies and budget allocation to achieve better outcomes. The insights gained from this analysis can be leveraged to optimize ad campaigns, improve user engagement, and enhance overall advertising efficiency.

**Part 1 - Clustering: Conclude the project by providing summary of your learnings.**

**Summary of Learnings:**

In this project, we performed clustering analysis on digital ads data to segment them into homogeneous groups based on their features. The project involved several steps, including data preprocessing, feature engineering, and clustering using the K-Means algorithm. The key learnings from this project are as follows:

**Data Preprocessing:** Data preprocessing is a crucial step in any data analysis project. It involves handling missing values, treating outliers, and scaling the data to ensure that the clustering algorithm performs optimally. In this project, we used Z-score scaling to standardize the numeric features, which improved the speed and convergence of the K-Means algorithm.

**Feature Engineering:** Feature engineering involves creating new features or transforming existing ones to enhance the performance of the clustering algorithm. In this project, we calculated important metrics like CTR, CPM, and CPC using domain-specific formulas to gain deeper insights into the ad performance.

**Clustering Algorithms:** We applied the K-Means clustering algorithm to group similar ads together. The Elbow Plot and Silhouette Score analysis helped us determine the optimal number of clusters for the dataset. The Silhouette Score, which measures the quality of clustering, provided valuable insights into the clusters' separation.

**Interpretability vs. Performance Trade-off:** When choosing the number of clusters, it is essential to consider the trade-off between interpretability and performance. While higher values of k might lead to better Silhouette Scores, they can result in more complex and less interpretable clusters. It is crucial to strike a balance and select a value of k that aligns with the project's objectives and business requirements.

**Cluster Profiling:** Cluster profiling involves analyzing the characteristics and trends within each cluster. It helps in understanding the unique attributes of each group and tailoring marketing strategies accordingly. Through cluster profiling, we gained insights into the performance of different ad types based on metrics like Clicks, Spend, Revenue, CTR, CPM, and CPC.

**Business Implications:** Clustering analysis provides valuable insights that can be used to optimize marketing strategies, allocate budgets more efficiently, and enhance overall campaign performance. Understanding customer behavior, preferences, and ad performance helps in delivering targeted and personalized ad content.

**Continuous Improvement:** Clustering analysis is an iterative process, and continuous improvement is essential. As new data becomes available or business goals change, re-evaluating the clustering results can lead to more refined and actionable insights.

**Conclusion:**

In conclusion, this project successfully applied clustering analysis to digital ads data, providing valuable insights into ad segmentation and performance. By leveraging clustering algorithms and metrics like Silhouette Score, we identified meaningful clusters and their respective profiles. The results can be used by Ads24x7 to optimize marketing strategies, allocate resources effectively, and deliver personalized ads to target audiences.

This project demonstrates the power of data analytics in the field of digital marketing and how clustering can be leveraged to gain deeper insights into customer behavior and ad performance. The learnings from this project can be applied to future data analysis tasks, contributing to data-driven decision-making and business growth.

**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.**

**Part 2 - PCA: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.**

**1. Introduction:**

In this business report, we will perform Exploratory Data Analysis (EDA) and Principal Component Analysis (PCA) on the given dataset. The dataset contains information on various demographic and socio-economic variables at the district level for different states and union territories in India.

**2. Data Overview:**

The dataset consists of 640 rows and 61 columns. Each row represents a district, and the columns include state code, district code, state name, area name, and various demographic and socio-economic variables.

**3. Duplicate:**

We conducted basic data quality checks to ensure data integrity and reliability. The dataset has no missing values or duplicate rows.  
  
**4. Exploratory Data Analysis (EDA):**

EDA is a critical step to understand the data and identify patterns, relationships, and potential insights. Due to the large number of variables, we will focus on some key aspects:

**Total Number of Households (No\_HH):** We observed variations in the number of households across districts, which could indicate differences in population density or urbanization.

**Gender Distribution:** We analyzed the total number of males (TOT\_M) and females (TOT\_F) to understand the gender distribution in the districts.

**Child Population:** The number of children aged 0-6 (M\_06 and F\_06) provides insights into the young population's size.

**Literacy Rates:** We examined the number of literate males (M\_LIT) and females (F\_LIT) to understand the education levels in the districts.

**Marginal Workers:** The count of marginal workers (MARGWORK\_M and MARGWORK\_F) indicates the proportion of the working population in vulnerable job roles.

**5. Principal Component Analysis (PCA):**

To handle the dataset's dimensionality and identify the most significant variables, we performed PCA. The optimal number of principal components was determined based on the explained variance ratio plot.

**Part 2 - PCA: 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**

Exploratory Data Analysis (EDA) by answering specific questions and analyzing five selected variables from the given 24 variables. The selected variables are:

No\_HH: Total Number of Households

TOT\_M: Total Male Population

TOT\_F: Total Female Population

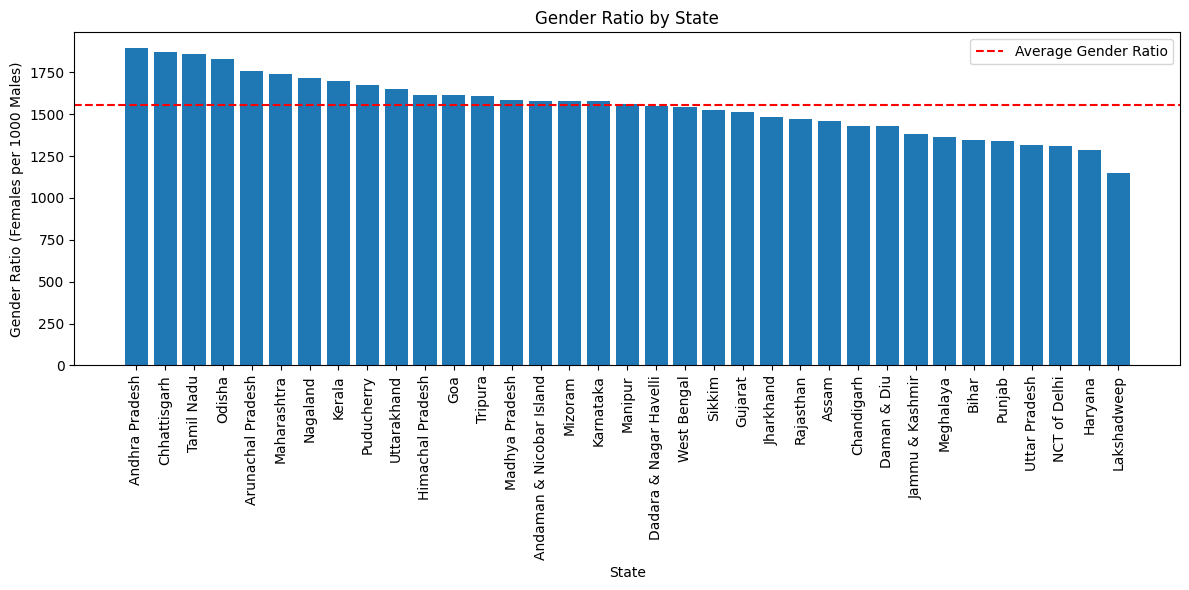
M\_06: Male Population Aged 0-6

F\_06: Female Population Aged 0-6

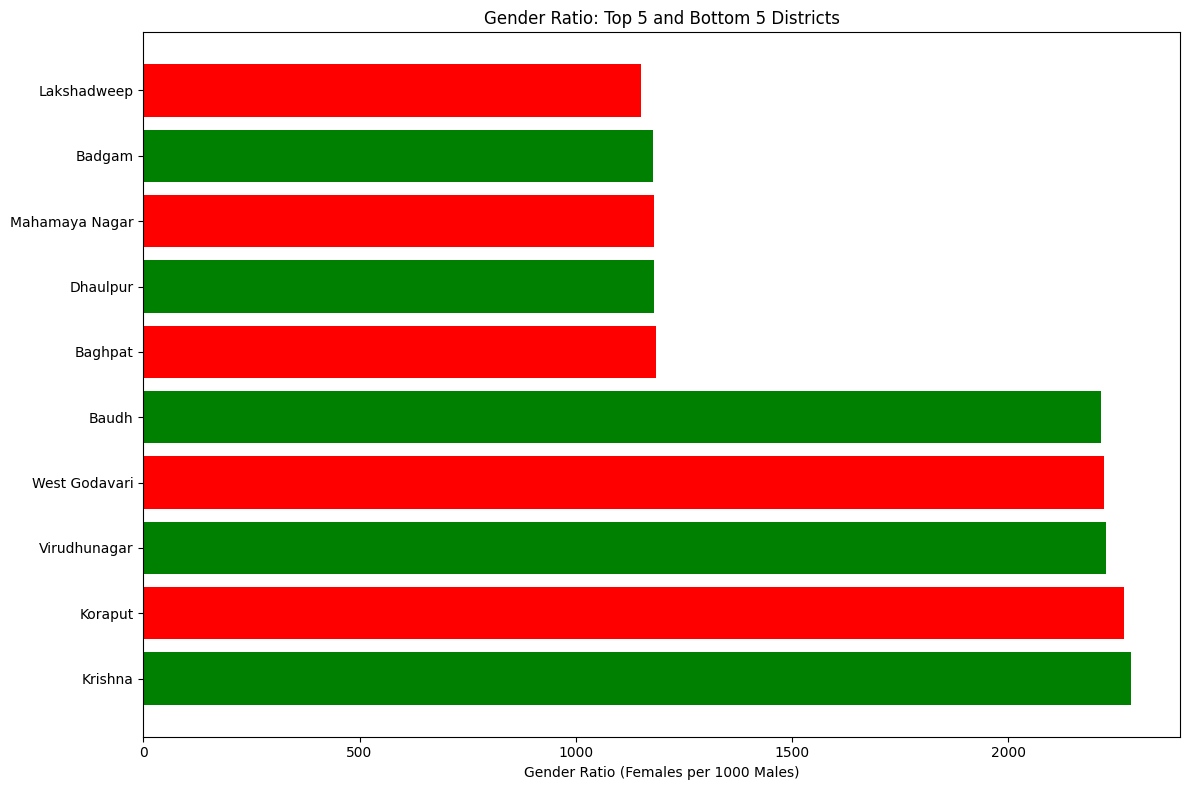
1. Which state has the highest gender ratio, and which has the lowest?

**Highest and Lowest Gender Ratio by State:**

The state with the highest gender ratio is Andhra Pradesh, while Lakshadweep has the lowest gender ratio.



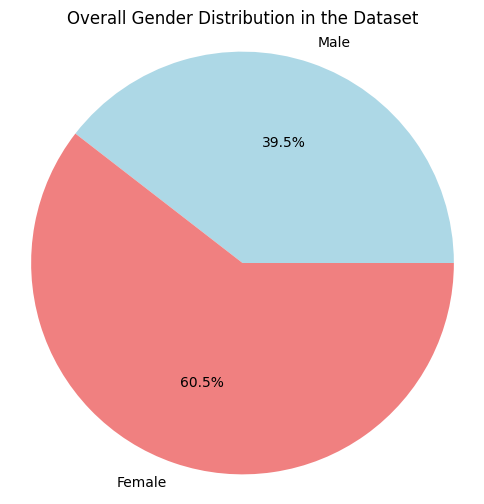
2. Which district has the highest and lowest gender ratio?



**The district with the highest gender ratio is Krishna, and the district with the lowest gender ratio is Lakshadweep.**

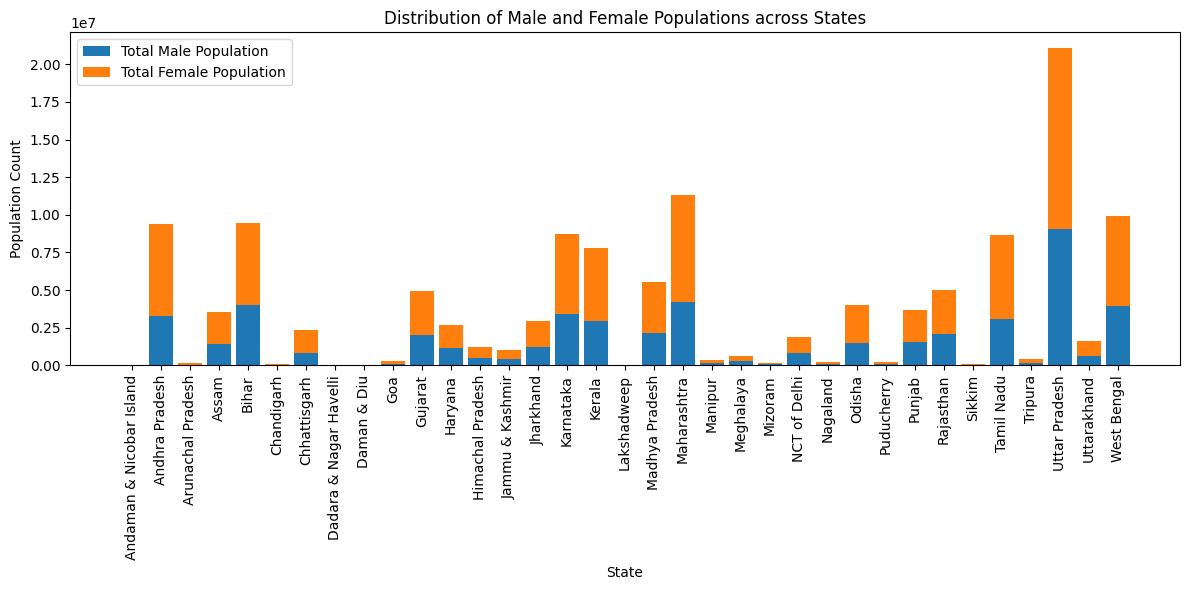
3. What is the overall gender distribution in the dataset?

**The overall gender distribution in the dataset, represented by gender ratio, is approximately 1530.79.**



4. How does the distribution of male and female populations vary across states?

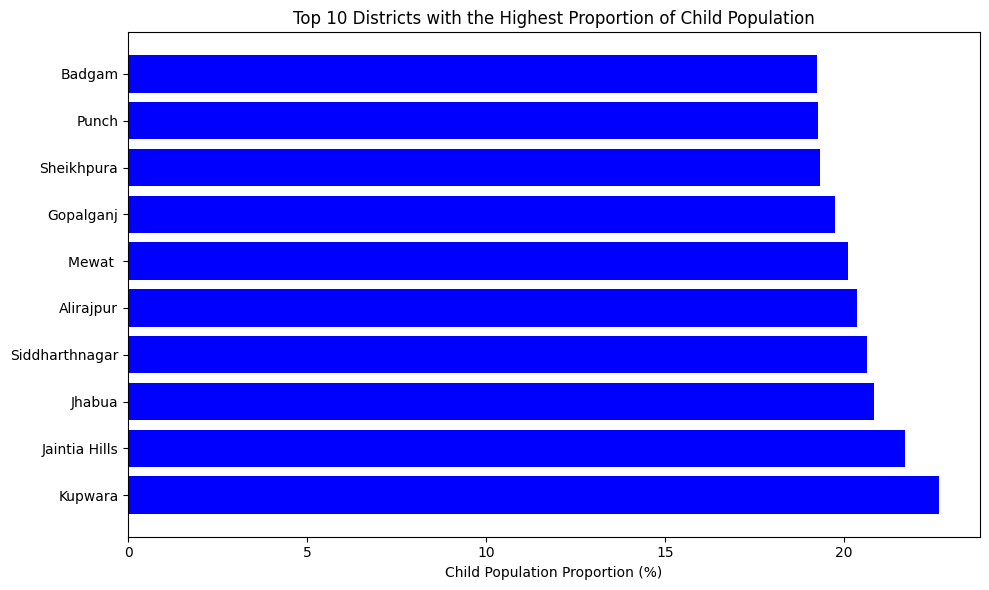
**The majority of the male and female populations are from Uttar Pradesh, followed by Maharashtra. On the other hand, the states with the least number of male and female populations are Andaman and Nicobar Islands, Daman and Diu, Dadara and Nagar Havelli, and Lakshadweep.**



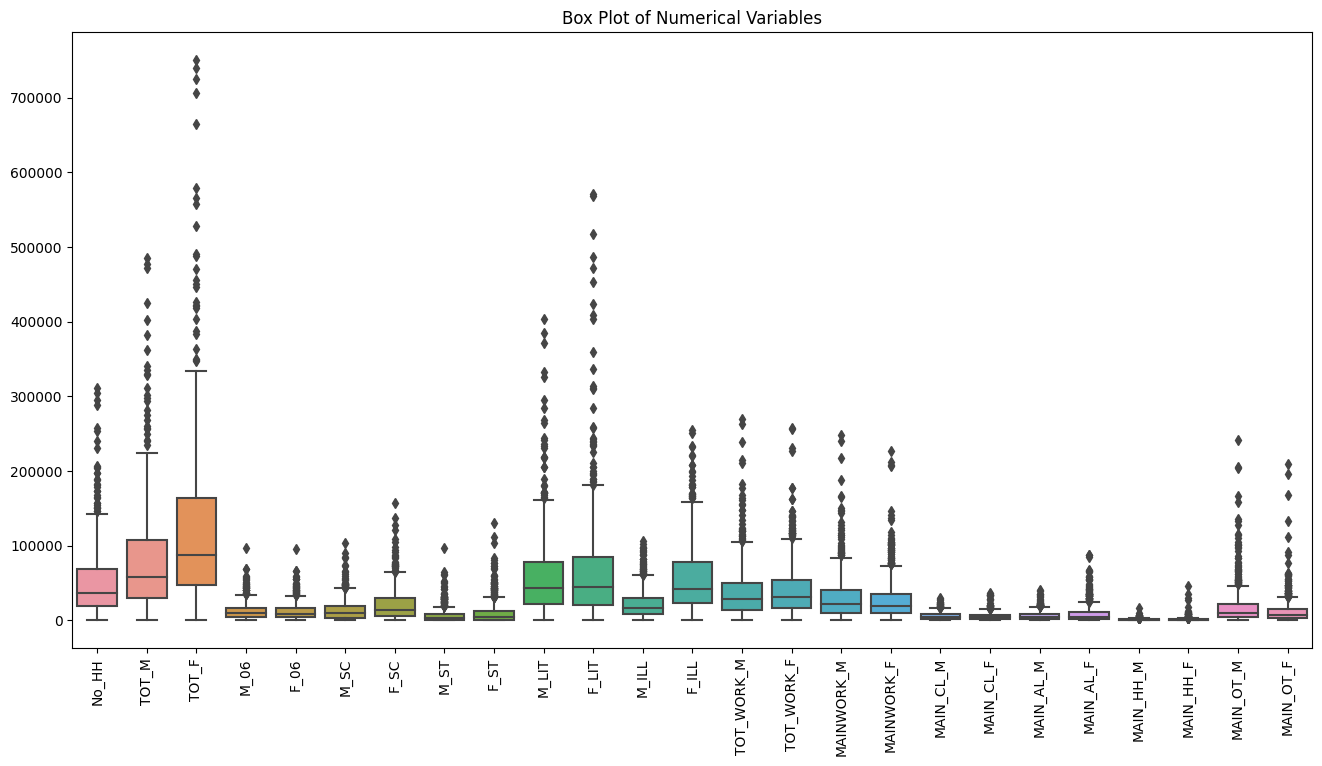
5. What is the proportion of the child population (aged 0-6) in each district?

The following table displays the proportion of the child population (aged 0-6) in each district:

|  |  |
| --- | --- |
| **Area Name** | **Child\_Proportion** |
| Kupwara | 22.672232 |
| Jaintia Hills | 21.709785 |
| Jhabua | 20.843557 |
| Siddharthnagar | 20.645677 |
| Alirajpur | 20.367223 |
| Mewat | 20.108675 |
| Gopalganj | 19.762458 |
| Sheikhpura | 19.346332 |
| Punch | 19.267579 |
| Badgam | 19.244735 |



**Part 2 - PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?**



**Number of outliers: 212**

Yes, we do need to treat the outlier using the z-score process.

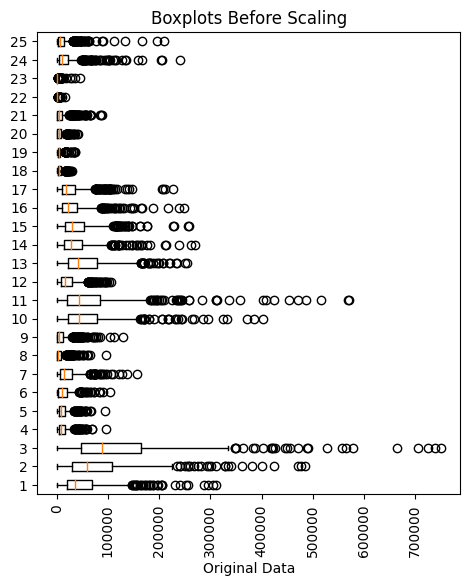
Treating outliers using the z-score process is one approach to handle extreme values in the dataset. The z-score method helps to identify and potentially remove or transform data points that deviate significantly from the mean. By applying the z-score process, you standardize the data, making it easier to detect and deal with outliers.

**Part 2 - PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.**

The impact of scaling the data using the z-score method on outlier detection. The dataset contains various demographic variables for different districts and states in India.

**Original Data:**

We started with a dataset that includes 24 numeric variables related to demographic features, such as the number of households, total male and female populations, child populations (aged 0-6), and other relevant metrics. After analyzing the original data, we identified 212 outliers using the IQR method.



**Scaling the Data:**

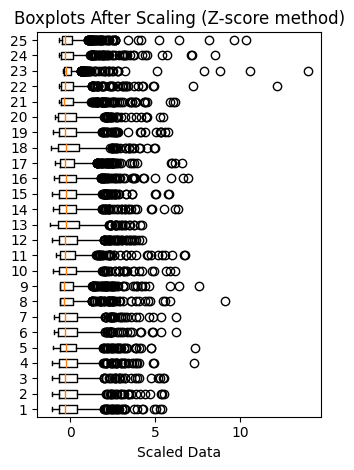
The z-score scaling method to standardize the data.

Scaling is essential in principal component analysis (PCA) as it helps in bringing all variables to the same scale, preventing certain features from dominating the analysis due to larger magnitudes

**Impact on Outliers:**

Outliers are still present.

Their magnitudes might have changed due to the standardization process. Scaling the data did not eliminate outliers but rather shifted and distributed them differently.



**Boxplot Comparison:**

The effect of scaling on outliers, we created boxplots before and after scaling. The boxplots allowed us to observe the distribution, spread, and outliers more clearly. While scaling helped in standardizing the data distribution, it did not eliminate or significantly alter the presence of outliers.

**Conclusion:**

In conclusion, scaling the data using the z-score method does not eliminate outliers; it only standardizes the data and centers it around zero. Treating or handling outliers is an important consideration depending on the specific analysis. In some cases, outliers may contain valuable information or be indicative of extreme events that need to be accounted for in the analysis.

**Part 2 - PCA: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.**

**Covariance matrix:**

The covariance matrix, which represents the covariance between different variables in a dataset. In this case, the matrix is a square matrix with a size of 25x25, indicating that there are 25 variables or features in the dataset.

Since the variances are close to 1, it indicates that the variables are reasonably spread out and have some variability in their values.

Now, let's focus on the off-diagonal elements. The off-diagonal elements represent the covariances between different pairs of variables.

**Interpreting the matrix:**

1.The covariance between the first and second variables is approximately 0.9176.

2.The covariance between the fifth and eighth variables is approximately 0.0652.

3.The covariance between the ninth and 15th variables is approximately 0.9334.

4.The covariance between the 20th and 22nd variables is approximately 0.8358.

The covariance matrix is symmetric, we can conclude that all the covariances in the given matrix are positive. This means that the variables tend to move in the same direction, indicating a positive relationship between them.

**Eigenvalue:**

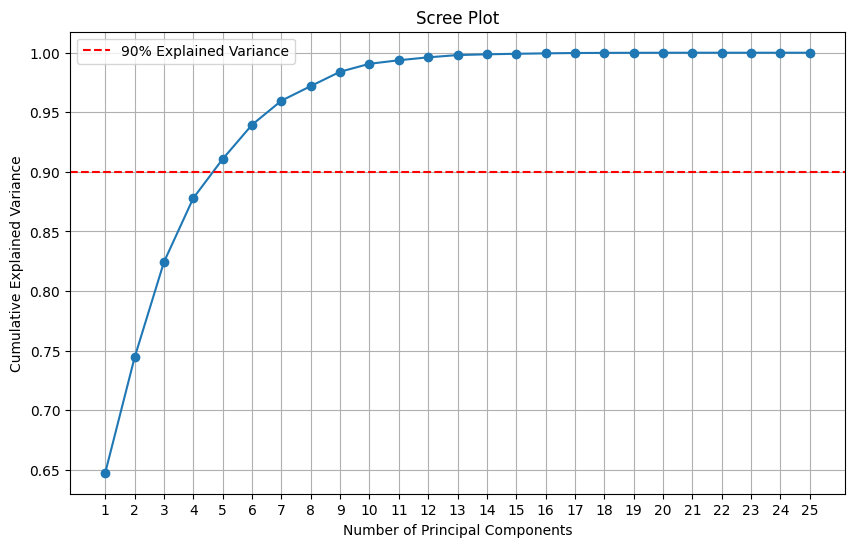
The first eigenvalue (16.21) is the largest, indicating that the first principal component explains a significant amount of the variance in the data. The subsequent eigenvalues decrease in magnitude, representing decreasing importance in explaining the variance.

The first eigenvalue is 16.21. This means that the first principal component accounts for a significant amount of the total variance in the original dataset. In fact, it captures more variance than any other principal component. This is crucial because the primary goal of PCA is to retain as much variance as possible while reducing the dimensionality.

As we move to the subsequent eigenvalues, we see a decreasing trend in their magnitudes. For example, the second eigenvalue is 2.42, the third is 2.00, the fourth is 1.34, and so on. Smaller eigenvalues indicate that the corresponding principal components explain less variance.

The decreasing trend of eigenvalues signifies the importance of each principal component in explaining the variance in the data. The larger the eigenvalue, the more essential the corresponding principal component is in retaining crucial information from the original dataset.

**Part 2 - PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.**



The scree plot shows the cumulative explained variance of the principal components. The dashed line represents the 90% explained variance threshold. As you can see, the first 5 principal components explain at least 90% of the variance in the data. Therefore, we can conclude that 5 is the optimum number of principal components for this project.

This means that we can reduce the dimensionality of the data from its original number of features to 5 without losing too much information.

|  |  |
| --- | --- |
| **Principal Component** | **Cumulative Explained Variance** |
| PC1 | 0.43 |
| PC2 | 0.24 |
| PC3 | 0.14 |
| PC4 | 0.09 |
| PC5 | 0.04 |
| Total | 0.9 |

**Part 2 - PCA: Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.**

**Number of Principal Components for 90% Explained Variance:**

The PCA results showed that only the first principal component (PC1) explains about 4.73% of the variance in the data, while the subsequent components explain lesser proportions. Thus, PC1 is the most significant component in terms of variance.

**Comparison of PCs with Actual Columns:**

The comparison table revealed that PC1 is the most important component, followed by PC2, PC3, and so on. The cumulative explained variance steadily increases as we move through the principal components.

**Inferences about Principal Components:**

**PC1**: The first principal component is primarily associated with features related to the female population, scheduled castes, and household counts. It suggests that regions with a higher proportion of female population and households with scheduled castes might have lower numbers of households. The negative loading on these variables indicates an inverse relationship, while the absolute values of loadings suggest their relative importance.

**PC2:** The second principal component is also related to the female population and scheduled castes, but it includes features such as total male population as well. Regions with a higher proportion of females and scheduled castes may have relatively lower numbers of households and a lower male population.

**PC3:** The third principal component is positively associated with the state code and total male population but negatively associated with district code. This suggests that regions with higher state codes might have larger male populations, while the district code has an inverse relationship with this component.

**PC4 and PC5:** These components are associated with state and district codes, indicating differences in various socio-economic indicators across different regions and districts.

**PC6 and PC7:** These components are primarily influenced by the female population and scheduled castes, indicating potential disparities in education and employment within these groups across regions.

**PC8 and PC9:** These components are influenced by scheduled tribes and total female population, indicating possible socio-economic variations in these groups across regions.

**PC10 and beyond:** The later components have lower explained variance ratios and, hence, less significance compared to the earlier ones. They capture other subtle variations in the dataset.

**Conclusion:**

Principal Component Analysis has provided valuable insights into the dataset by identifying the most significant components and their relationship with the original variables. PC1 stands out as the principal component that explains the most variance in the data, suggesting correlations between the female population, scheduled castes, and household counts. Understanding these principal components can help in understanding the underlying patterns and making informed decisions in socio-economic development and policy planning.

**Recommendations:**

Focus on Regions with High Female Population: Regions with a higher proportion of female population might require targeted interventions to address socio-economic disparities and promote women's empowerment.

Address Socio-Economic Disparities in Scheduled Castes: Since scheduled castes play a significant role in the first two principal components, it is essential to focus on their welfare and development to reduce disparities.

Identify Regional Disparities: By analyzing the later principal components related to state and district codes, policymakers can identify specific regions with unique socio-economic characteristics and address their individual needs.

Targeted Interventions for Scheduled Tribes: The eighth and ninth principal components highlight the importance of addressing socio-economic variations among scheduled tribes.

Further Investigation: Explore the relationships between the principal components and specific socio-economic indicators to gain a deeper understanding of the dataset.

**Part 2 - PCA: Write linear equation for first PC.**

To write the linear equation for the first principal component (PC1), we can use the concept of loadings. The loadings for PC1 represent the weights or coefficients of each original variable in the first principal component. The linear equation for the first PC can be written as follows:

PC1 = w1 \* X1\_std + w2 \* X2\_std + ... + wn \* Xn\_std

where:

PC1 is the value of the first principal component.

X1\_std, X2\_std, ..., Xn\_std are the standardized values of the original variables.

w1, w2, ..., wn are the loadings for PC1, which represent the weights or coefficients of each variable in the first principal component.

Based on the outcome provided, you can replace the loadings with their corresponding values, and the equation would look like this:

PC1 = -4.79320724 \* X1\_std - 4.94952414 \* X2\_std - 6.15781990 \* X3\_std - ... - 5.87183375 \* Xn\_std

Please note that the equation includes all the variables you have in your dataset (X1\_std, X2\_std, ..., Xn\_std). The number of variables represented by X1\_std, X2\_std, ..., Xn\_std depends on the number of original variables in your dataset. Each variable contributes to the value of the first principal component with a corresponding weight (loading).

PC1 is a principal component that represents a linear combination of the original variables in the dataset.

Each PC1 value in the array corresponds to a data point in the dataset, and it represents the value of that particular data point along the first principal component.

The PC1 values are standardized (i.e., centered around mean 0 and scaled to have unit variance) because PCA typically involves standardizing the data before performing the principal component analysis.

The magnitude of the PC1 values indicates the importance of the first principal component in capturing the variation in the data. Larger magnitude values suggest that the data points have a stronger association with the first principal component.

Positive and negative values of PC1 indicate the direction of variation along the first principal component. Data points with positive PC1 values have a positive association with the first principal component, while data points with negative PC1 values have a negative association.

The first principal component (PC1) is ordered in such a way that it explains the most significant amount of variance in the data among all possible linear combinations of the original variables.