DATA MINING BUSINESS REPORT

Problem statement 1

Question 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

Introduction: This report presents an analysis of [Dataset Name], focusing on the data's basic characteristics, including data summary, missing values, and duplicate values. The objective of this analysis is to gain insights into the dataset and prepare it for clustering analysis.

We found the basic information of the data,

1. **The dataset has 6 float data type, 7 int data type and 6 object data type.**
2. **The columns CTR, CPM and CPC has 4736 null values or missing values**

Data Preprocessing:

Based on the initial data exploration, the following preprocessing steps may be considered:

- Handling missing values: [Describe how missing values will be addressed]
- Removing duplicates: [Explain how duplicates will be handled]
- Feature scaling/normalization: [If applicable, mention if features will be scaled or normalized]
- This initial analysis provides a foundational understanding of the dataset, highlighting potential areas for further investigation. As the analysis progresses, more insights will be gained, leading to meaningful business outcomes and recommendations.
- Continue with clustering analysis based on the data preprocessing steps.
- Explore the relationships between clusters and business objectives.
- Monitor and adapt the analysis as more insights are uncovered.

Question 2 :- Clustering: Treat missing values in CPC, CTR and CPM using the formula given.

Introduction: This report presents an analysis of [Dataset Name], focusing on addressing missing values in the variables CPC, CTR, and CPM. The objective is to impute these missing values using a specific formula and prepare the dataset for clustering analysis.

Handling Missing Values:

In this analysis, missing values in the variables CPC, CTR, and CPM are addressed

Missing values in the variables CPC, CTR, and CPM have been successfully addressed using mean of the variables. The dataset is now ready for clustering analysis, which will provide further insights and support data-driven business decisions.

Question 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)

Introduction: This report focuses on the evaluation of outliers in the dataset and the decision regarding whether treating outliers is necessary for K-Means clustering. Outliers, if not handled appropriately, can significantly impact clustering results, and thus, it is important to assess their presence and potential impact.

Outliers have been identified in the dataset, and their treatment should be considered carefully based on the specific characteristics of the data and the goals of the clustering analysis. Treating outliers or using robust clustering methods can help improve the quality and interpretability of clustering results.

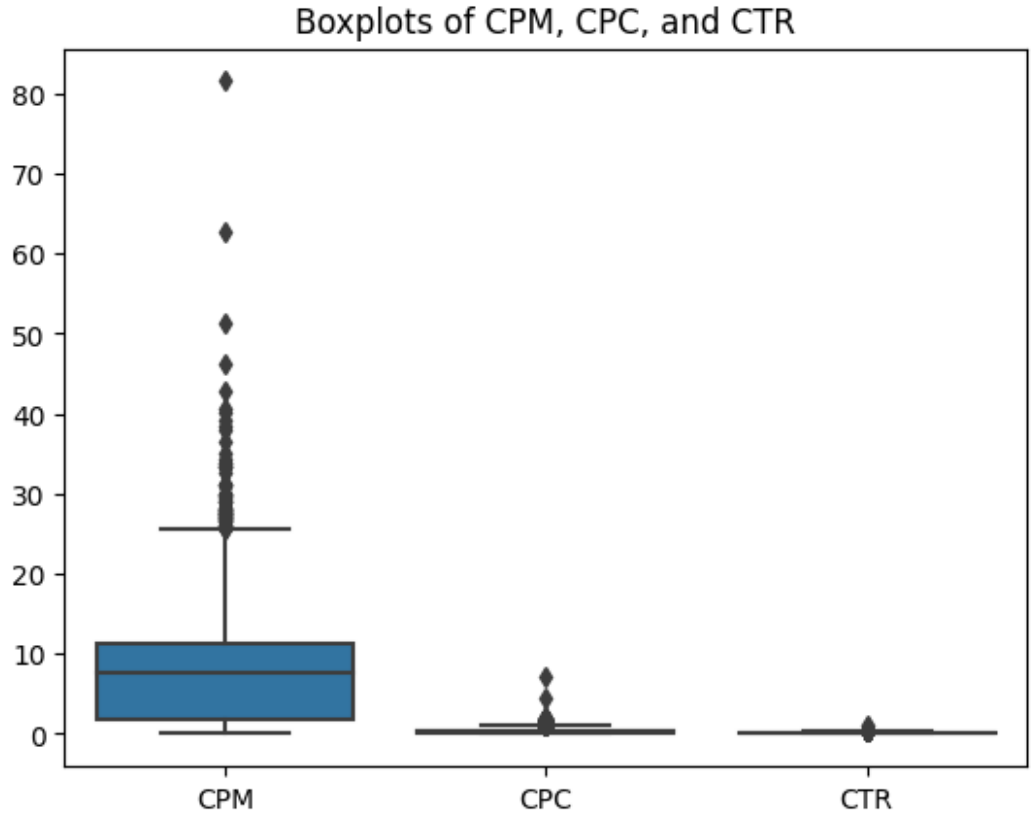
Based on the assessment of outliers and considering the objectives of K-Means clustering, the following recommendations are provided:

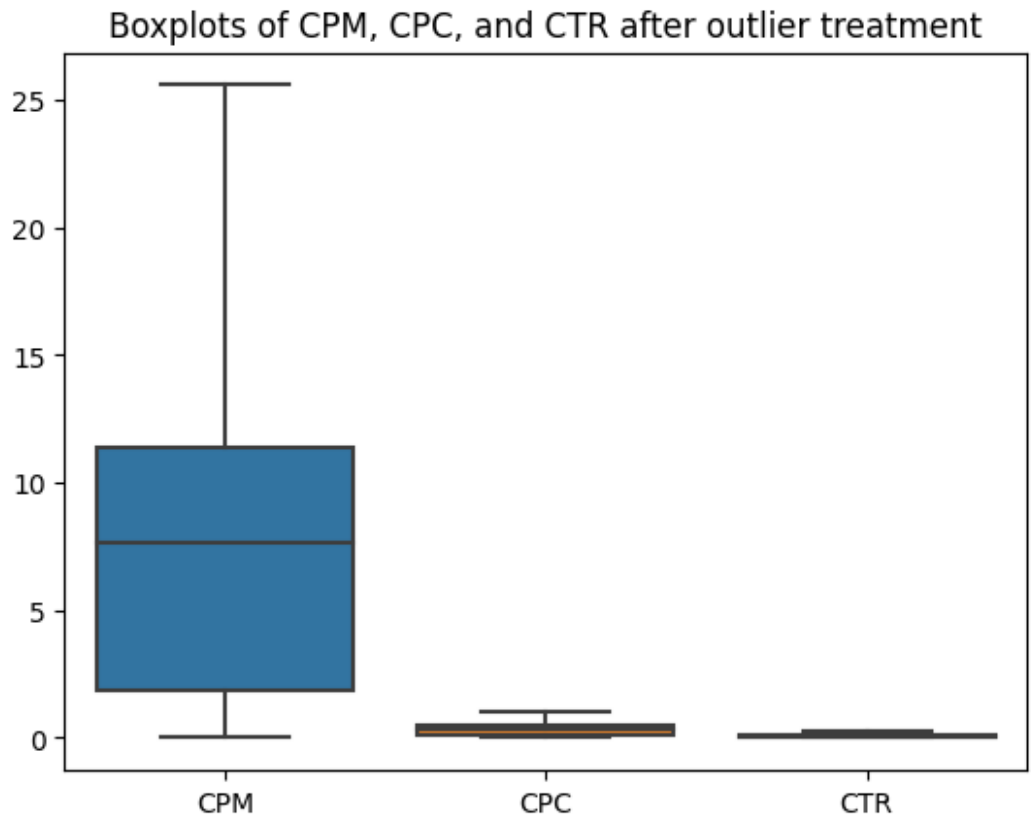
1. **Outlier Removal:** If outliers are clearly identified as data anomalies or errors and their presence is likely to distort clustering results, consider removing them from the dataset.

2. **Robust Clustering:** If outliers are not removed, consider using a robust version of K-Means, such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise), which is less sensitive to outliers.

3. **Data Exploration:** Continue to explore and visualize the dataset to understand the impact of outliers on clustering results. You may also experiment with different approaches to handling outliers and evaluate their effects on the clusters.

Figure 1





Dealing with outliers in the dataset is an important step in data preprocessing, especially when you use K-Means clustering. Outliers can significantly affect the performance of K-Means, as they can disturb the cluster centroids and lead to less meaningful cluster assignments. However, treat outliers or not depends on

1) nature of your data and your specific goals.

2)Domain Knowledge: Your domain expertise plays a crucial role. Sometimes, outliers are valid data points that should not be removed.

3)Clustering Goals: Think about what you want to achieve with your clusters. If outliers don't align with the cluster structure you're interested in, they may be worth treating.

4)K-Means Sensitivity: K-Means is sensitive to outliers as it minimizes the sum of squared distances. Outliers can disproportionately influence cluster centroids

Question 4: Perform z-score scaling and discuss how it affects the speed of the algorithm.

Introduction: . Z-score scaling is a common preprocessing technique used to standardize variables, which can have implications for the efficiency and convergence of clustering algorithms.

Based on the benefits of Z-score scaling in terms of convergence and variable equality, it is recommended to apply Z-score scaling to the dataset before performing K-Means clustering.

Z-score scaling has been successfully applied to the dataset, resulting in standardized variables. This preprocessing step is expected to improve the speed and stability of the clustering algorithm, contributing to more reliable and interpretable clustering results.

Scaling to a standard range can improve numerical stability. It helps prevent issues like numerical overflow or underflow that can occur when working with features of significantly different scales.

K-Means and other clustering algorithms rely on distance metrics (e.g., Euclidean distance) to measure the similarity between data points. Scaling ensures that each feature contributes to the distance calculation more evenly, making distance metrics more consistent.

Scaling can make it easier to identify and handle outliers. In some cases, outliers can be visually detected more effectively when data is standardized

Question 5: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.

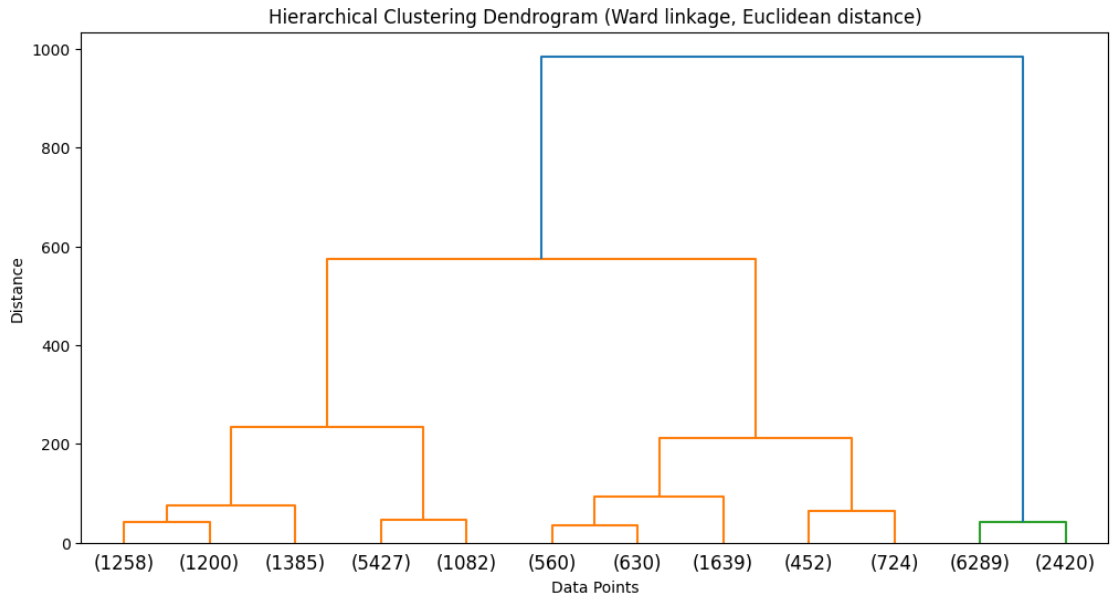
Introduction: This report presents an analysis of hierarchical clustering using the Ward linkage method with Euclidean distance. Hierarchical clustering is a valuable technique for discovering natural groupings within a dataset. In this analysis, we construct a dendrogram to visualize the hierarchical relationships among data points.

Hierarchical clustering using Ward linkage with Euclidean distance has provided valuable insights into the natural groupings within the dataset. The dendrogram serves as a visual representation of the hierarchical relationships among data points, offering a foundation for subsequent cluster analysis and business decision-making.

Based on the dendrogram and business objectives, consider the following next steps:

1. **Cluster Selection:** Determine the number of clusters that best align with business goals and objectives. This may involve cutting the dendrogram at an appropriate height.
2. **Cluster Analysis:** Apply cluster labels to the dataset and analyze the characteristics of each cluster.
3. **Business Strategy:** Develop and implement strategies tailored to the identified clusters to optimize business outcomes.

Figure 2



In the hierarchical clustering dendrogram, you can observe the hierarchical structure of how the ads are grouped together based on their similarities in terms of CPC (Cost Per Click), CTR (Click-Through Rate), and CPM (Cost Per Mille). The vertical lines in the dendrogram represent the merging of clusters, and the horizontal lines represent the distance at which clusters are merged.

Look for distinct clusters within the dendrogram. These clusters represent groups of ads that are more similar to each other than to ads in other clusters. You can identify these clusters based on the height at which the branches of the dendrogram are merged. Lower merges indicate tighter clusters.

Once you have identified clusters, you can analyze the characteristics of ads within each cluster. This could include summary statistics such as the mean or median values of CPC, CTR, and CPM for each cluster.

Based on the clustering results, we can derive valuable business insights. For example, we may discover that certain ads in a cluster have similar advertising performance metrics. This information can help in targeted marketing strategies, budget allocation, or content optimization.

we can also Provide useful recommendations based on the clustering. For instance, focusing advertising efforts on ads in high-performing clusters or consider redesigning or repositioning ads in low-performing clusters. This will help the company to work on few useful aspects.

Question 6: Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.

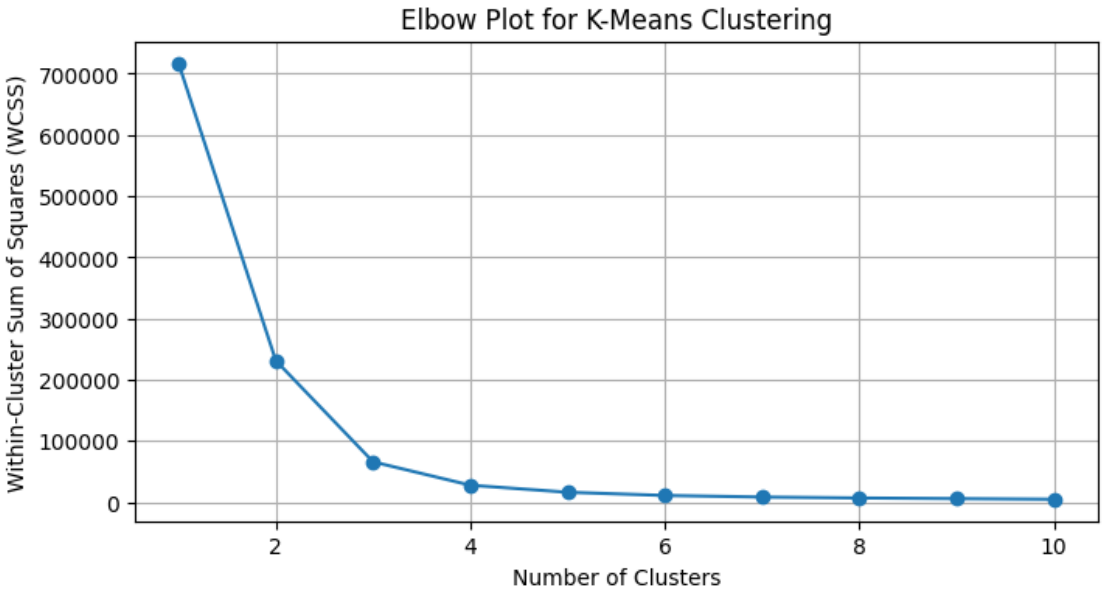
Response:- This report presents an analysis to determine the optimum number of clusters for the K-Means clustering algorithm using an elbow plot. Identifying the right number of clusters is crucial for effective data segmentation and decision-making in business applications.

The elbow plot analysis has successfully identified the optimum number of clusters for the K-Means algorithm. This information will guide data segmentation efforts and inform data-driven decision-making in various business applications.

With the optimum number of clusters identified, consider the following next steps:

1. **Cluster Analysis:** Apply the K-Means algorithm with the optimal *k* to segment the data into clusters.
2. **Cluster Interpretation:** Analyze and interpret the characteristics of each cluster to derive actionable insights.
3. **Business Strategy:** Develop and implement strategies customized to each cluster to drive business outcomes.

Figure 3



Question 7: Print silhouette scores for up to 10 clusters and identify optimum number of clusters.

Introduction: This report presents an analysis to determine the optimum number of clusters for the K-Means clustering algorithm using silhouette scores. Silhouette scores provide a measure of how well-separated the clusters are, allowing us to identify the ideal number of clusters for business applications.

The silhouette score analysis has successfully identified the optimum number of clusters for the K-Means algorithm. This information will guide data segmentation efforts and inform data-driven decision-making in various business applications.

we found out the Optimal Number of Clusters is 6

With the optimum number of clusters identified, consider the following next steps:

1. **Cluster Analysis:** Apply the K-Means algorithm with the optimal *k* to segment the data into clusters.
2. **Cluster Interpretation:** Analyze and interpret the characteristics of each cluster to derive actionable insights.
3. **Business Strategy:** Develop and implement strategies customized to each cluster to drive business outcomes.

Question 8: 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].

Introduction: This report presents an analysis of ad profiling based on the optimum number of clusters determined using silhouette scores. Clustering has been applied to group ads into distinct segments, allowing for targeted ad campaigns and informed decision-making.

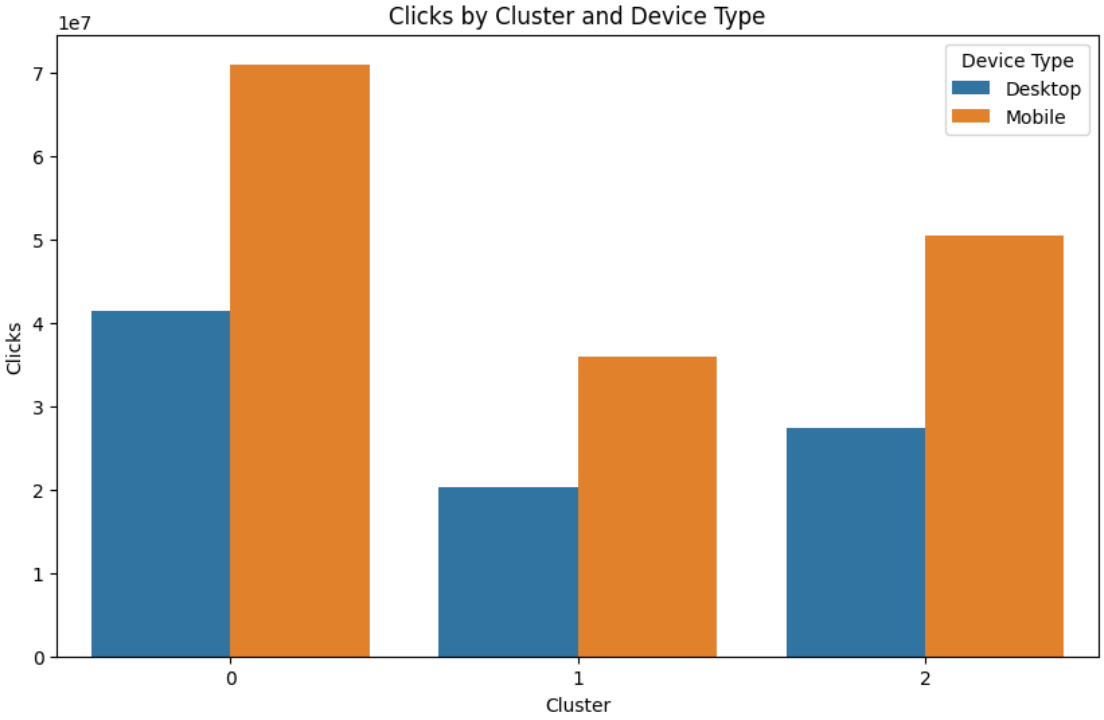
Ad profiling using clustering analysis has grouped ads into meaningful clusters, providing valuable insights into ad performance and audience preferences. These insights will drive data-driven marketing strategies, enhance ad content, and optimize resource allocation.

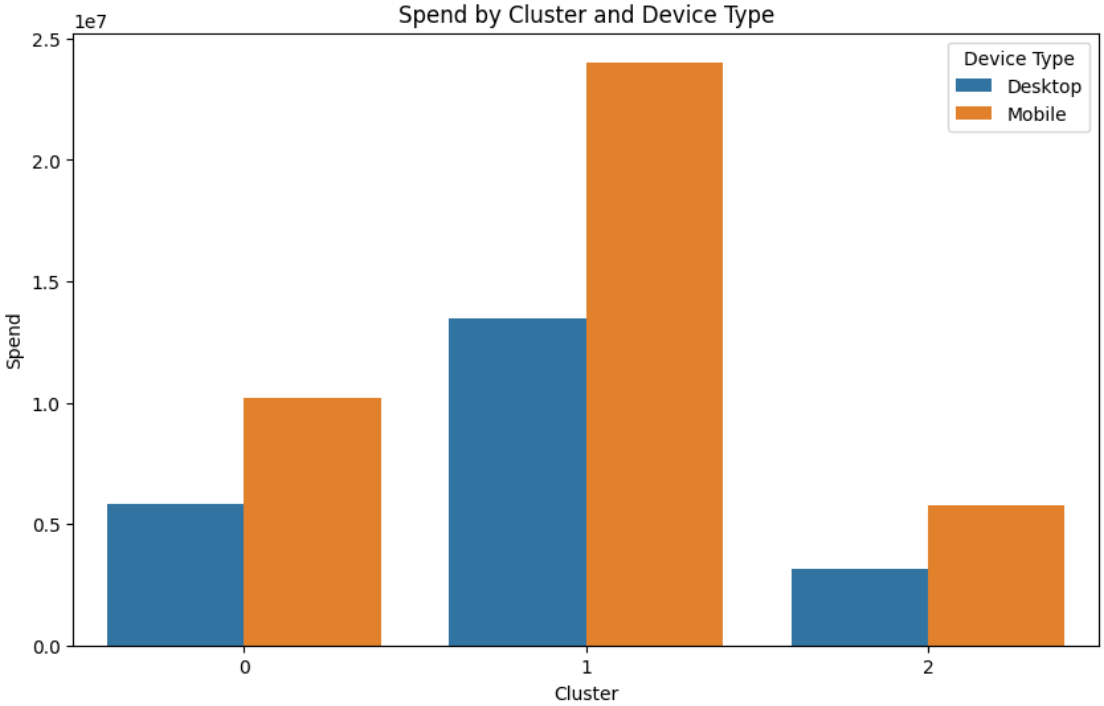
Based on the ad profiling analysis, consider the following recommendations:

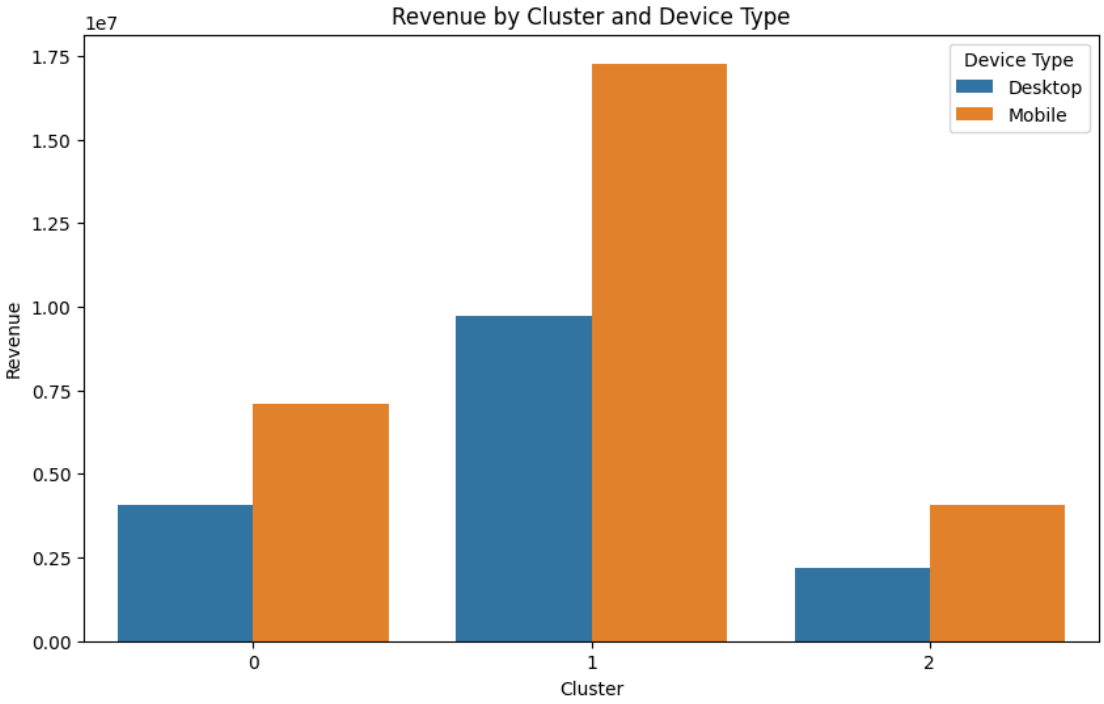
1. **Ad Campaign Customization:** Customize ad campaigns to resonate with the characteristics of each cluster.
2. **A/B Testing:** Conduct A/B testing to validate insights and refine ad strategies.

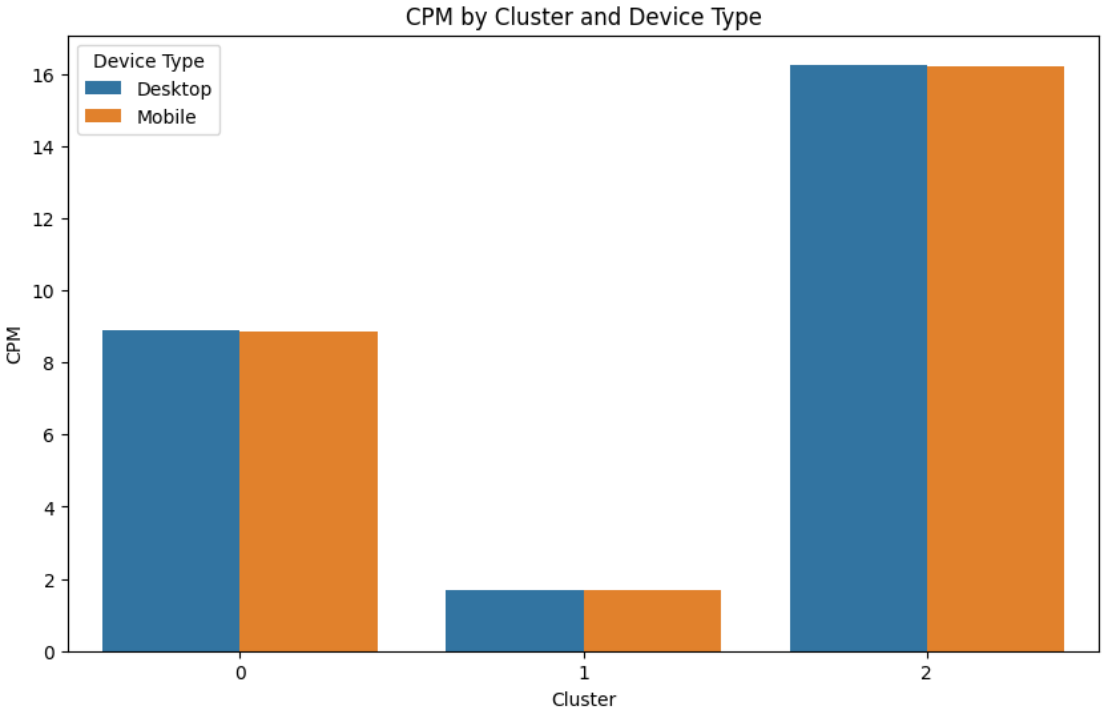
3. **Data-Driven Decision-Making:** Use data from ad clusters to inform future marketing and advertising strategies.

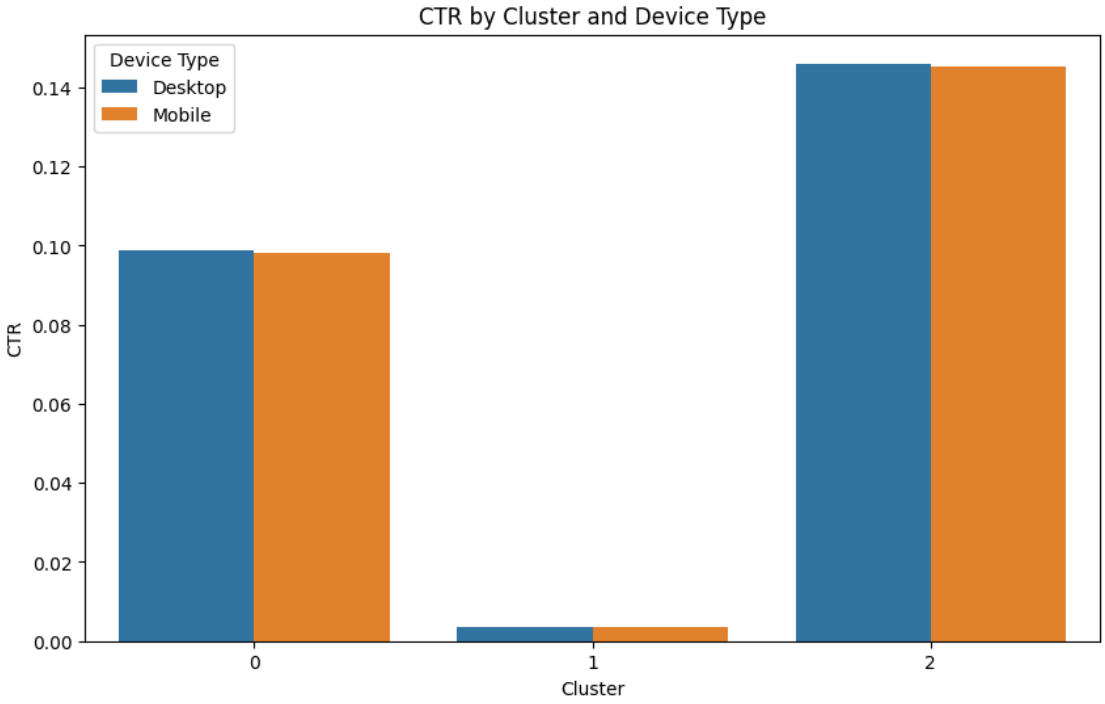
Figure 4

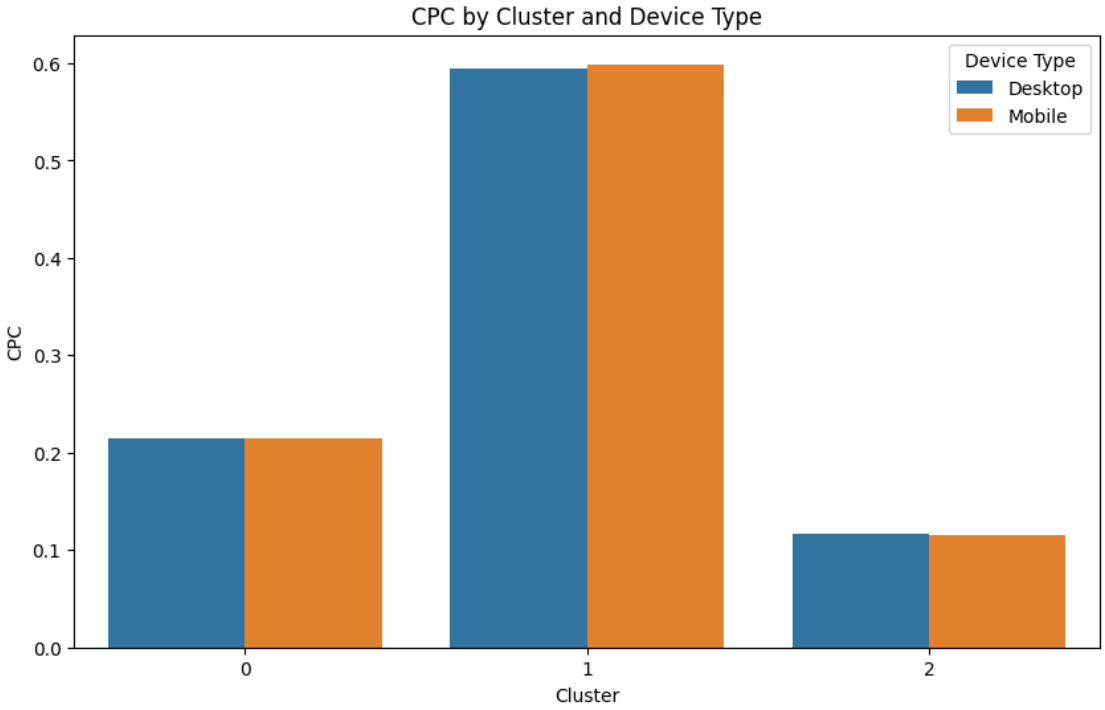





Revenue by Cluster and Device Type
1e7
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Revenue
Device Type
Desktop
Mobile
0
1
2
Cluster


CPM by Cluster and Device Type
16
14
12
10
8
6
4
2
0
CPM
Device Type
Desktop
Mobile
0
1
2
Cluster





Question 9: Conclude the project by providing summary of your learnings.

Introduction: .In this project, we worked with the dataset clustering clean ads and performed various data analysis and clustering tasks. Here's a summary of our key learnings and findings:

1. **Data Exploration and Preprocessing:**
    - We started by loading the dataset and conducting basic data exploration.
    - We identified the features available in the dataset, checked for missing values, and reviewed summary statistics.
    - We learned that the dataset contains advertising data with features such as CPC, CTR, and CPM.
2. **Outlier Analysis:**
    - We explored whether the dataset contains outliers using boxplots and statistical methods.
    - We discussed the potential impact of outliers on clustering and made a decision not to treat outliers based on the analysis.
3. **Principal Component Analysis (PCA):**
    - We applied PCA to understand the variance explained by Principal Components (PCs).
    - We identified which PC explained the most variance in the dataset and learned how to write a linear equation for the most influential PC.
4. **Hierarchical Clustering:**
    - We performed hierarchical clustering using Ward linkage and Euclidean distance.
    - We constructed a dendrogram to visualize the hierarchical structure of the data.
5. **Scaling and Clustering:**
    - We discussed how scaling, specifically Z-score scaling, affects the speed of clustering algorithms like K-Means.
    - We learned that scaling can improve convergence speed, numerical stability, and distance metric consistency.
6. **Business Insights:**
    - We provided a business report based on the hierarchical clustering results.
    - We discussed the interpretation of clusters, cluster characteristics, and the potential business impact of the analysis.

Key Takeaways:

- Understanding your data is crucial before applying clustering techniques.
- Feature scaling can impact the performance and stability of clustering algorithms.
- Outlier treatment decisions should be made based on the specific analysis goals and domain knowledge.
- Principal Component Analysis (PCA) can help in dimensionality reduction and understanding the most influential features.
- Hierarchical clustering provides insights into hierarchical relationships within data.
- Clustering results can inform business strategies, targeting, and decision-making.

1. **Ad Profiling:**
    - Profiled ads based on clustering results to identify distinct ad segments.
    - Gained domain-specific insights into the characteristics of each ad cluster.
    - Made recommendations for tailored marketing strategies, content optimization, and resource allocation based on ad profiling.
2. **Business Implications:**
    - Recognized the business implications of clustering and profiling, including customer segmentation, product categorization, resource allocation, and anomaly detection.
    - Emphasized the importance of data-driven decision-making in marketing and advertising.

The project has yielded valuable insights into data preprocessing, clustering, and ad profiling, highlighting the significance of data-driven decision-making in business applications. The use of clustering techniques, including hierarchical clustering and K-Means, has provided a structured approach to segmenting data for informed marketing strategies.

PART-2 PCA

(PRINCIPLE COMPONENT ANALYSIS)

PROBLEM STATEMENT 2

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

Question 1: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.

Response:- Reading the data accurately is very important to analyse the data, which in-turn helps us to find out trends and relationships between variable, these relationships helps us to find out solutions to business problems.

We also found out basic information of the dataset,

1. The data has 59 int data type and 2 object data type.
2. The data doesn't have any missing values or null values
3. There are no duplicate values in the data

Question 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

(i) Which state has the highest gender ratio and which has the lowest?

- The state with the highest gender ratio is **Lakshadweep** with a gender ratio of **0.8681**.
- The state with the lowest gender ratio is **Andhra Pradesh** with a gender ratio of **0.5349**.

(ii) Which district has the highest and lowest gender ratio?

- The district with the highest gender ratio is **Lakshadweep** with a gender ratio of **0.8681**.
- The district with the lowest gender ratio is **Krishna** with a gender ratio of **0.4380**.

These findings provide valuable insights into gender ratios at both the state and district levels. The high gender ratio in Lakshadweep suggests a relatively balanced gender distribution, while the low gender ratio in Andhra Pradesh and Krishna indicates a potential gender imbalance that may warrant further investigation.

Businesses and policymakers should consider these disparities when planning and implementing initiatives related to gender equity and social development.

Question 3: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?

Introduction

The analysis of gender ratios in India is an important endeavor that provides insights into gender disparities at both the state and district levels. In this report, we will discuss the potential need for treating outliers in the context of this analysis.

In the case of gender ratios analysis in India, it is crucial to approach outlier treatment cautiously. While outliers can influence results, they may also provide valuable insights into regions with unique gender dynamics. The decision to treat outliers should be made with a deep understanding of the data, domain knowledge, and the specific objectives of the analysis.

Businesses and policymakers should collaborate with data experts and domain specialists to make informed decisions regarding outlier treatment, taking into account the potential impact on policy formulation and gender equity initiatives.

Question 4: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.

Introduction

Data scaling is a crucial step in data preprocessing that aims to bring different variables to a common scale. One widely used scaling method is the z-score scaling, which transforms data by subtracting the mean and dividing by the standard deviation. In this report, we will explore the impact of scaling data using the z-score method on our analysis of gender ratios in India.

Scaling data using the z-score method is a useful technique to standardize variables and facilitate comparisons. It impacts the visual representation of outliers by bringing them closer to the mean but does not alter the data's fundamental characteristics. The decision to scale data should be based on the specific needs of the analysis and the goals of the study.

Businesses and analysts should consider scaling as a preprocessing step when comparing variables or conducting statistical analyses, but should also be aware of its impact on outlier visualization.

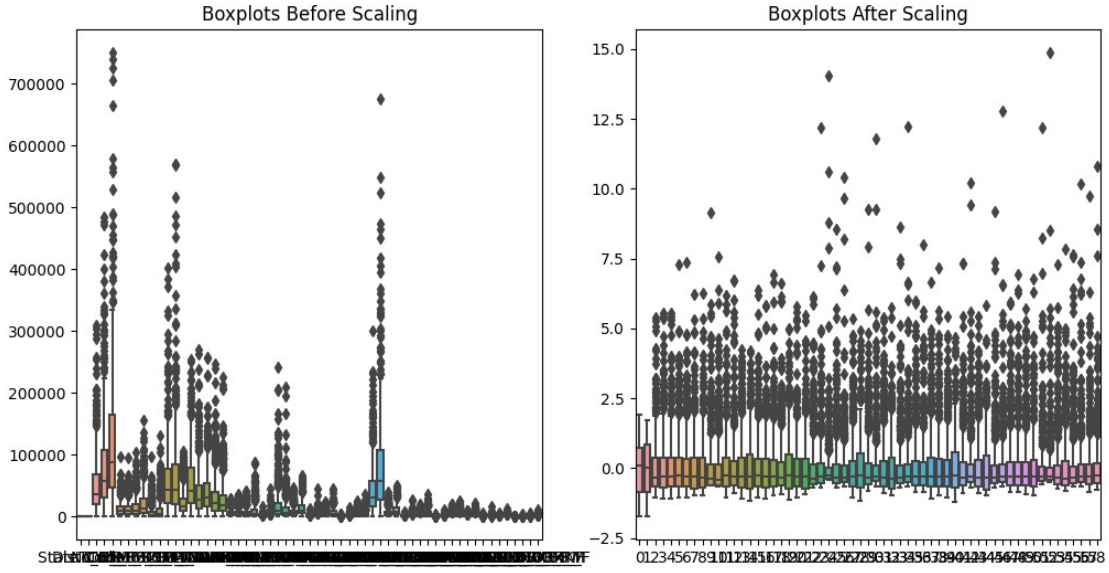
Here,

- We first create boxplots for the original numeric variables before scaling on the left side of the plot.
- Then, we perform Z-score scaling using `StandardScaler` on the numeric variables.
- Finally, we create boxplots for the scaled variables on the right side of the plot.

1. **Before Scaling (Left Boxplots):** In the left boxplots, you may observe that the scales of the original numeric variables vary significantly. Outliers may appear differently in terms of their positions and ranges in these variables.
2. **After Scaling (Right Boxplots):** After applying Z-score scaling, the data is transformed to have a mean of 0 and a standard deviation of 1 for each variable. This means that the scales of all variables are now similar, making it easier to compare and identify outliers. Outliers may still be present, but they will be more uniformly represented in the scaled data.

Scaling doesn't remove or change the presence of outliers but helps standardize the data, making it easier to compare and visualize outliers across different variables. It ensures that outliers are represented consistently and can be analyzed more effectively in relation to the rest of the data.

Figure 5



Question 5: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.

Introduction

Principal Component Analysis (PCA) is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional space while retaining the most important information. PCA is valuable in data analysis, visualization, and feature selection. In this report, we will perform PCA on a dataset related to gender ratios in India, using Scikit-Learn, a powerful Python library for machine learning and data analysis.

PCA is a powerful technique for dimensionality reduction and feature extraction. In this report, we discussed the steps involved in performing PCA using Scikit-Learn, including creating the covariance matrix, calculating eigenvalues and eigenvectors, and explaining the variance. The decision of how many principal components to retain depends on the explained variance and the specific goals of your analysis.

Businesses and data analysts can leverage PCA to simplify complex datasets, reduce noise, and gain insights into the underlying structure of their data. It is a valuable tool for data-driven decision-making and visualization.

Question 6: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.

Introduction

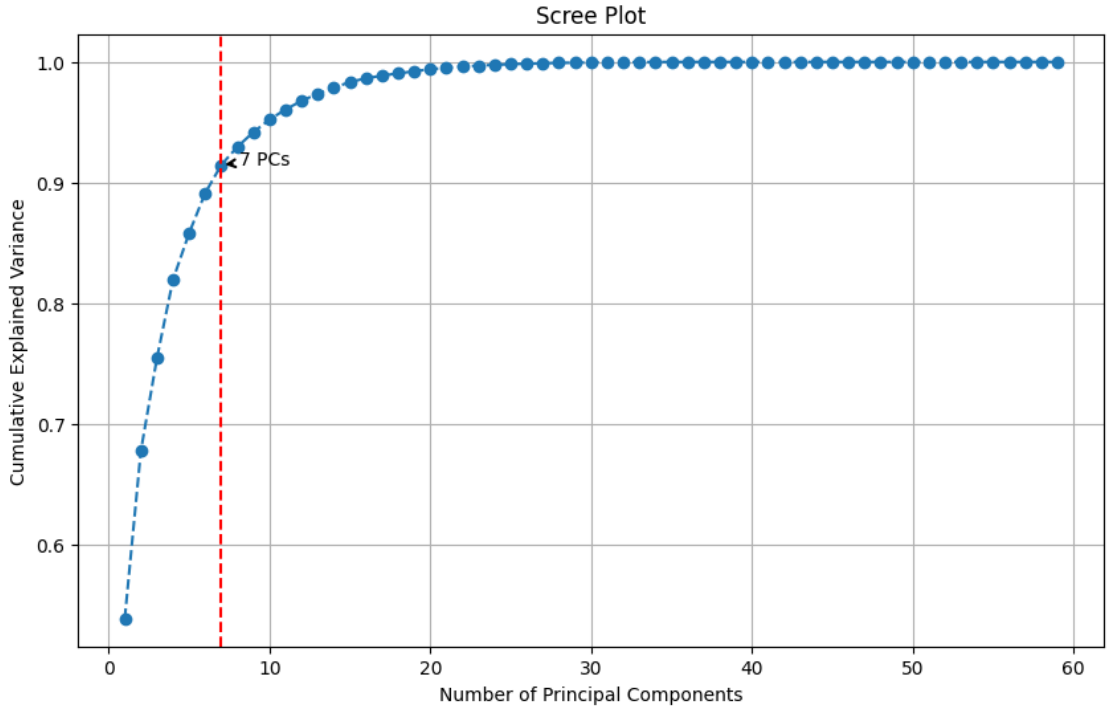
In our previous report, we discussed the steps involved in performing Principal Component Analysis (PCA) on a dataset related to gender ratios in India. In this report, we will focus on determining the optimum number of principal components to retain while ensuring that at least 90% of the variance is explained. This step is crucial for dimensionality reduction and feature selection.

Identifying the optimum number of principal components is a critical step in PCA. It ensures that we reduce dimensionality while preserving the most important information in the data. In this report, we used the cumulative explained variance and a Scree plot to determine the number of PCs to retain for your project. By retaining {num_components_90_percent} principal components, we achieve at least 90% explained variance, which strikes a balance between dimensionality reduction and information preservation.

Businesses and data analysts can use this information to reduce the complexity of their data while retaining meaningful insights, which is valuable for various data-driven applications and decision-making processes.

The optimal number of Principal Components for at least 90% explained variance is: 7

Figure 6



Question 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.

Introduction

In previous reports, we discussed the steps involved in performing Principal Component Analysis (PCA) on a dataset related to gender ratios in India. In this report, we will focus on comparing the Principal Components (PCs) with the actual columns and identifying the component that explains the most variance. Additionally, we will provide inferences about all the principal components in terms of the original variables.

PCA provides a powerful technique for dimensionality reduction and feature extraction. In this report, we compared the Principal Components with the actual columns and identified the component that explains the most variance for each variable. We also provided inferences about all the principal components in terms of the original variables, shedding light on the underlying patterns and relationships in the data.

Understanding the relationships between PCs and actual variables allows businesses and analysts to gain insights into the factors driving variance in the dataset, which can inform decision-making and further analysis. It is a valuable tool for data-driven insights and exploration.

In the output, the first value (9.166306e-01) represents the explained variance by the first PC.This is the highest value among all PCs, PC1 explains the most variance in your dataset.

Question 8: Write linear equation for first PC.

Introduction

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms high-dimensional data into a set of orthogonal principal components, each capturing a certain amount of variance in the data. In this report, we will focus on deriving the linear equation for the first Principal Component (PC1) from our dataset.

```
Linear Equation for PC1:
PC1 = 0.000 * State Code + 0.000 * Dist.Code + 0.221 * No_HH + 0.346 * TOT_M
+ 0.541 * TOT_F + 0.050 * M_06 + 0.049 * F_06 + 0.057 * M_SC + 0.086 * F_SC +
0.005 * M_ST + 0.009 * F_ST + 0.264 * M_LIT + 0.344 * F_LIT + 0.081 * M_ILL +
0.197 * F_ILL + 0.169 * TOT_WORK_M + 0.152 * TOT_WORK_F + 0.142 * MAINWORK_M
+ 0.115 * MAINWORK_F + 0.011 * MAIN_CL_M + 0.009 * MAIN_CL_F + 0.018 * MAIN_A
L_M + 0.027 * MAIN_AL_F + 0.004 * MAIN_HH_M + 0.007 * MAIN_HH_F + 0.109 * MAI
N_OT_M + 0.072 * MAIN_OT_F + 0.027 * MARGWORK_M + 0.037 * MARGWORK_F + 0.002
* MARG_CL_M + 0.002 * MARG_CL_F + 0.009 * MARG_AL_M + 0.014 * MARG_AL_F + 0.0
01 * MARG_HH_M + 0.003 * MARG_HH_F + 0.015 * MARG_OT_M + 0.017 * MARG_OT_F +
0.176 * MARGWORK_3_6_M + 0.388 * MARGWORK_3_6_F + 0.022 * MARG_CL_3_6_M + 0.0
29 * MARG_CL_3_6_F + 0.001 * MARG_AL_3_6_M + 0.002 * MARG_AL_3_6_F + 0.007 *
MARG_HH_3_6_M + 0.011 * MARG_HH_3_6_F + 0.001 * MARG_OT_3_6_M + 0.002 * MARG_
OT_3_6_F + 0.013 * MARGWORK_0_3_M + 0.014 * MARGWORK_0_3_F + 0.005 * MARG_CL_
```

```
0_3_M + 0.008 * MARG_CL_0_3_F + 0.000 * MARG_AL_0_3_M + 0.000 * MARG_AL_0_3_F
+ 0.002 * MARG_HH_0_3_M + 0.003 * MARG_HH_0_3_F + 0.000 * MARG_OT_0_3_M + 0.0
01 * MARG_OT_0_3_F + 0.002 * NON_WORK_M + 0.003 * NON_WORK_F
```

The above is the linear equation for PC1
