MACHINE LEARNING -1
PROJECT

BUSINESS REPORT

Prepared By : Sahid
COURSE : DSBA
DATE : 30-06-2024

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 Project Overview:

In the digital age, data-driven decision-making is paramount for the success of marketing strategies. Automatically segmenting ads and understanding the demographics in the context of India's census data can provide actionable insights leading to improved marketing effectiveness and societal understanding.

Problem 1: Clustering Analysis for Digital Marketing Ads

Ads24x7, a burgeoning digital marketing firm, has recently secured $10 million in seed funding to expand its analytics capabilities. To optimize their ad campaigns, the company's Marketing Intelligence team has collected extensive data on various ad features. The objective of this segment of the project is to perform exploratory data analysis (EDA) and employ clustering algorithms to segment different types of ads into homogeneous groups. This segmentation will help ads24x7 better target their audience and allocate their marketing budget more effectively.

Problem 2: Principal Component Analysis on Census Data

The Census of India 2011 is one of the most comprehensive demographic datasets available, providing intricate details on various population metrics. However, the high dimensionality of the data poses a challenge for extracting meaningful insights. In this part of the project, we aim to utilize Principal Component Analysis (PCA) to reduce the complexity of the dataset while retaining the most critical information. This analysis will help in identifying the key components that explain the majority of the variance in the data, facilitating a better understanding of demographics, especially for female-headed households.

Objectives:

1. **For Problem 1:**
    - Conduct EDA to understand the underlying patterns and distributions in the ad dataset.
    - Apply hierarchical and K-means clustering to segment ads into meaningful clusters.
    - Provide actionable insights and recommendations to optimize digital marketing efforts.

2. **For Problem 2:**
    - Perform EDA to extract useful demographic insights from the census data.
    - Implement PCA to determine the most significant components driving the variance.

- Interpret the principal components and construct linear combinations representing the original variables.

Report Structure:

The report is structured into two main parts, each addressing a distinct problem statement. The first part focuses on the clustering analysis of digital marketing ads, while the second part delves into PCA of census data. Each part includes detailed EDA, data preprocessing steps, a description of the applied methodologies, and the resulting insights and recommendations.

2. Problem 1: Clustering Analysis

2.1 Problem Definition and EDA

Problem Definition

Context:

Ads24x7, a leading digital marketing company, has recently secured $10 million in seed funding to enhance its capabilities in marketing analytics. As part of their expansion, they've collected extensive data from their Marketing Intelligence team to better understand and segment their ads. The goal is to categorize these advertisements into homogeneous groups based on their features. Effective segmentation will allow Ads24x7 to optimize their marketing strategies, allocate budgets efficiently, and target the right audience segments with tailored ad content.

Problem Statement:

The objective of this analysis is to leverage clustering algorithms to segment the different types of advertisements into homogeneous groups based on their inherent features. By doing so, Ads24x7 aims to:

1. Identify distinct clusters of advertisements.
2. Understand the characteristics and performance metrics of each cluster.
3. Provide actionable insights to refine ad targeting and improve overall marketing effectiveness.

Exploratory Data Analysis (EDA):

Data Description:

The dataset comprises various features related to the ads, which are crucial for the clustering analysis. Here's a brief description of these features:

- Timestamp: The specific time when the advertisement data was recorded.
- InventoryType: Categorical variable representing different inventory types (Format 1 to 7).
- Ad-Length: The length dimension of the advertisement.
- Ad-Width: The width dimension of the advertisement.
- Ad Size: Product of `Ad-Length` and `Ad-Width`.
- Ad Type: Categorical variable representing different types of advertisements.
- Platform: The platform where the advertisement is displayed (Web, Video, or App).
- Device Type: The type of device supporting the advertisement (Categorical).
- Format: The format in which the advertisement is displayed (Categorical).
- Available_Impressions: Total number of times the advertisement could have been shown.
- Matched_Queries: Number of exact search queries that matched the ad criteria.
- Impressions: Actual count of how many times the ad was shown.
- Clicks: Number of times the ad was clicked by users.
- Spend: Total amount of money spent on the advertisement.
- Fee: Percentage of advertising fees payable.
- Revenue: Income generated from the advertisement.
- CTR (Click-Through Rate): Calculated as (Clicks / Impressions) * 100.
- CPM (Cost Per Thousand Impressions): Calculated as (Total Spend / Impressions) * 1,000.
- CPC (Cost Per Click): Calculated as (Total Spend / Clicks).

Data Overview :

The dataset has 23066 rows and 19 columns. It is always a good practice to view a sample of the rows. A simple way to do that is to use head() function.

Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
23061	2020-9-13-7	Format5	720	300	216000	Inter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	NaN	NaN
23062	2020-11-2-7	Format5	720	300	216000	Inter224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.026	NaN	NaN
23063	2020-9-14-22	Format5	720	300	216000	Inter218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	NaN	NaN
23064	2020-11-18-2	Format4	120	600	72000	inter230	Video	Mobile	Video	7	1	1	1	0.07	0.35	0.0455	NaN	NaN
23065	2020-9-14-0	Format5	720	300	216000	Inter221	App	Mobile	Video	2	2	2	1	0.09	0.35	0.0585	NaN	NaN

(Table 1: Top five rows of the dataset for Problem1)

Descriptive Statistics :

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	CLUSTER
count	23066	23066	23066	2.31E+04	2.31E+04	2.31E+04	23066	23066	23066	23066	23066	23066	23066	23066
mean	385.163097	337.896037	96674.46805	2.43E+06	1.30E+06	1.24E+06	10678.51882	2706.625689	0.335123	1924.252331	2.614863	8.39673	0.336652	0.899723
std	233.651434	203.092885	61538.32956	4.74E+06	2.51E+06	2.43E+06	17353.40936	4067.927273	0.031963	3105.23841	7.853405	9.057082	0.341231	1.247952
min	120	70	33600	1.00E+00	1.00E+00	1.00E+00	1	0	0.21	0	0.0001	0	0	0
25%	120	250	72000	3.37E+04	1.83E+04	7.99E+03	710	85.18	0.33	55.365375	0.0034	1.75	0.09	0
50%	300	300	72000	4.84E+05	2.58E+05	2.25E+05	4425	1425.125	0.35	926.335	0.11265	8.370742	0.14	1
75%	720	600	84000	2.53E+06	1.18E+06	1.11E+06	12793.75	3121.4	0.35	2091.33815	0.183778	13.04	0.55	1
max	728	600	216000	2.76E+07	1.47E+07	1.42E+07	143049	26931.87	0.35	21276.18	200	715	7.26	4

(Table 2: Descriptive statistics of the dataset for Problem1)

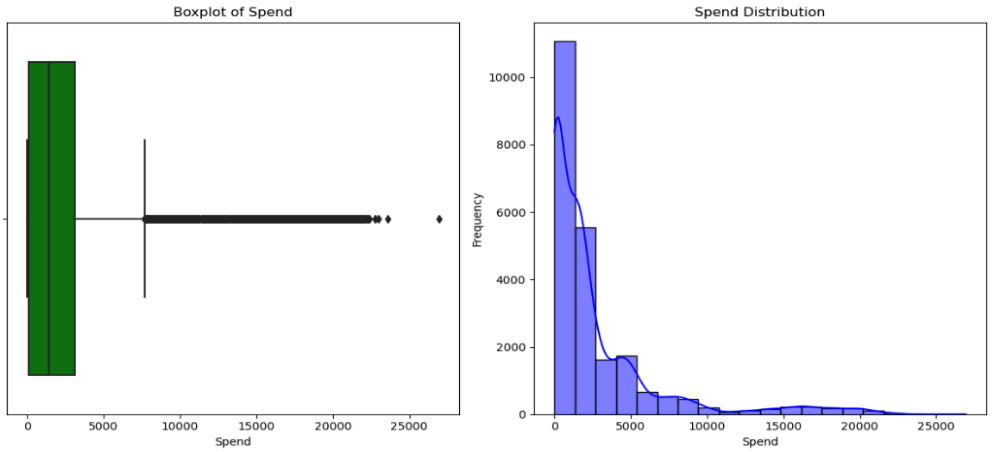
Observations:

- The data includes a wide range of ad sizes, with the mean ad size being 96,674.47 square units.
- The number of available impressions, matched queries, and impressions all show significant variation, indicated by high standard deviations.
- The click-through rate (CTR) and cost metrics (CPM, CPC) have substantial variability, suggesting different ad performance levels.
- The dataset is segmented into clusters, with cluster values ranging from 0 to 4, indicating the presence of 5 distinct clusters in the data.

Univariate Analysis

For performing Univariate analysis we will take a look at the Boxplots and Histograms to get a better understanding of the distributions.

Observations on Spend



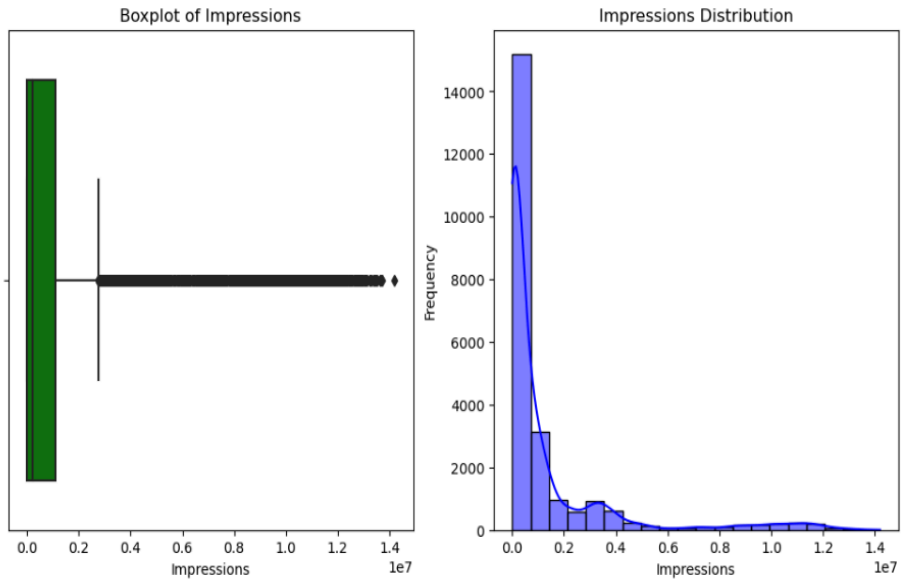
(Fig1 : Plot showing the boxplot and distribution for Spend)

Observations:

● The Distribution of Spend is right skewed.

● There are a lot of outliers in this variable.

Observations on Impressions



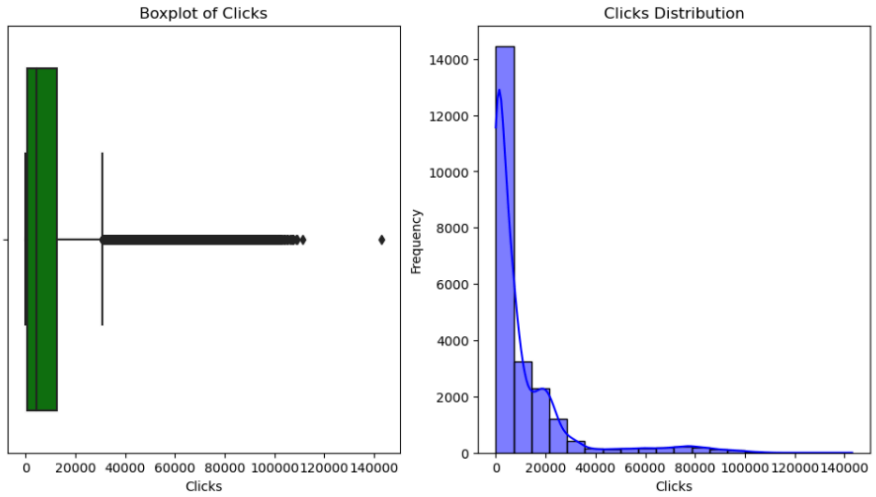
(Fig2 : Plot showing the boxplot and distribution for Impressions)

Observations:

● The Distribution of Impressions is right skewed.

- There are a lot of outliers in this variable.

Observations on Clicks



(Fig 3 : Plot showing the boxplot and distribution for Clicks)

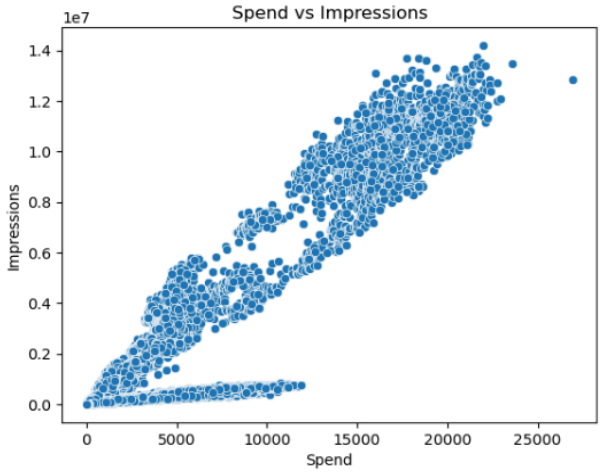
Observations:

- The Distribution of Clicks is right skewed.
- There are a lot of outliers in this variable.

Bivariate analysis

Observations on Spend vs Impressions

(Fig 4 : Plot showing the scatterplot of distribution for Observations on Spend vs Impressions)



Observations:

Positive Correlation:

There is a strong positive correlation between Spend and Impressions, indicating that higher spending generally leads to a higher number of impressions. This is evidenced by the upward-sloping trend in the plot.

Clusters of Points:

The plot shows some clustering patterns, particularly at lower spend and impression values. This may indicate different strategies or performance levels across different ad campaigns.

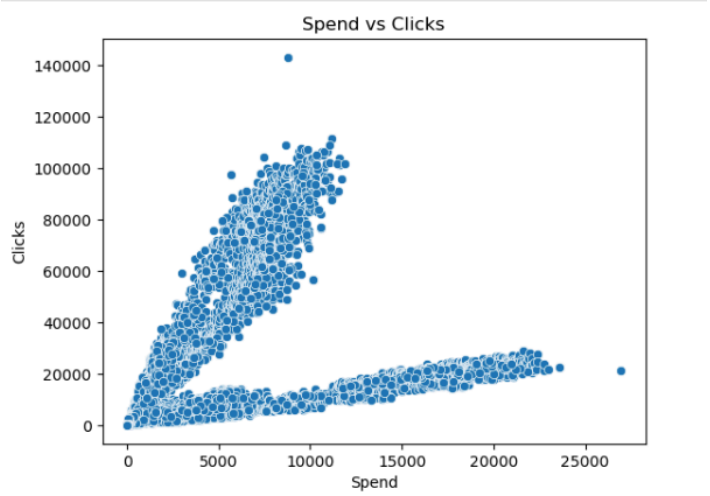
Outliers:

There are a few outliers where impressions are very high relative to the spend. These points could represent exceptionally effective ads or special campaigns.

Conclusion:

The scatter plot confirms the expected relationship between spend and impressions, reinforcing that increased budget allocation typically results in greater ad exposure. Further analysis could focus on identifying factors that influence the efficiency of spend, aiming to maximize impressions per unit of spend.

Observations on Spend vs Clicks



(Fig 5 : Plot showing the scatterplot of distribution for Observations on Spend vs Clicks)

Observations

Positive Correlation:

The scatter plot indicates a positive correlation between Spend and Clicks, meaning that as spending increases, the number of clicks also tends to increase.

Clusters of Points: One cluster shows a rapid increase in clicks with lower spending. , Another cluster shows a more linear relationship where clicks increase steadily with spending.

Outliers:

There are some outliers where clicks are significantly higher for certain spend values, indicating highly effective ads. Similarly, there are points where spend is high, but clicks are comparatively low, suggesting less effective campaigns.

Performance Variation:

The spread of points suggests variability in the performance of ad campaigns. Some ads achieve a higher number of clicks with less spending, indicating better performance or targeting.

Conclusion: This scatter plot reinforces the relationship between ad spend and the resulting clicks. While increased spend generally leads to more clicks, the variability and presence of outliers highlight the importance of optimizing ad campaigns for efficiency. Further analysis could involve identifying factors that contribute to higher click rates for lower spending to improve overall ad performance.

Observations on Impressions vs Clicks

(Fig 6 : Plot showing the scatterplot of distribution for Observations on impression vs Clicks)

Observations

Positive Correlation: The scatter plot shows a positive correlation between Impressions and Clicks, where an increase in impressions generally leads to an increase in clicks.

Clusters of Points: One cluster shows a rapid increase in clicks with lower impressions. , Another

cluster shows a more gradual increase in clicks with higher impressions.

Outliers: There is a noticeable outlier with a very high number of clicks but relatively lower impressions. This could indicate an exceptionally high click-through rate (CTR) for a specific ad.

Performance Variation: The plot reveals variability in the performance of ads. Some ads achieve high clicks with fewer impressions, suggesting effective targeting or high ad quality, while others require more impressions to achieve similar click numbers.

Conclusion: This scatter plot highlights the relationship between impressions and clicks, with an overall trend indicating that more impressions lead to more clicks. The presence of outliers and distinct clusters suggests variations in ad performance, pointing to opportunities for optimizing ad effectiveness and targeting.

2.2 Data Preprocessing

- **Handling Missing Values and Outliers**

CTR	4736
CPM	4736
CPC	4736

Missing Value Imputation

Description: Missing value imputation is a data preprocessing technique used to handle missing or null values in a dataset. In this case, the code is imputing missing values for the 'CPM', 'CPC', and 'CTR' columns of a DataFrame `df1`.

- **Outliers**

K-Means is sensitive to outliers because it uses the mean to calculate the centroid of each cluster. Outliers can pull the centroid away from the true center of the cluster, leading to poor clustering results. So we have decided to not treat the outliers as there are too many of them.

Scaling

Objective:

To scale the numerical features in the dataset to ensure that they are on a similar scale, which is a critical step for distance-based algorithms like K-means and hierarchical clustering. This helps in improving the performance and accuracy of clustering algorithms.

	CPM	CPC	CTR
0	-0.927110	-0.986603	-0.332572
1	-0.927110	-0.986603	-0.332521
2	-0.927110	-0.986603	-0.332610
3	-0.927110	-0.986603	-0.332712
4	-0.927110	-0.986603	-0.332444
...	...	...	...
23061	6.801816	-0.781459	12.400641
23062	1.281155	-0.869378	6.033837
23063	4.593551	-0.840072	12.400641
23064	6.801816	-0.781459	12.400641
23065	4.041485	-0.722846	6.033837

Methods and Approaches:

Choice of Scaling Method: Z-score Normalization

Reason: Z-score normalization (or standardization) transforms the data to have a mean of zero and a standard deviation of one. This centers the data and scales it based on the inherent spread, which ensures that all features have the same importance during clustering.

Implementation Steps:

(Fig 7 : Table showing scaled data)

Calculate the mean ((\mu)) and standard deviation ((\sigma)) for each numerical feature.

Transform the original values using the Z-score formula.

Observations:

For Ad-Length and Ad-Width: Before scaling, the data is centered around their means with a certain level of dispersion. After scaling, the data for each feature is centered at zero with a standard deviation of one.

Impact on Algorithms: Scaling ensures that no single feature dominates the distance measurement due to its scale, thus leading to more balanced clustering results.

Clustering Analysis

Hierarchical Clustering

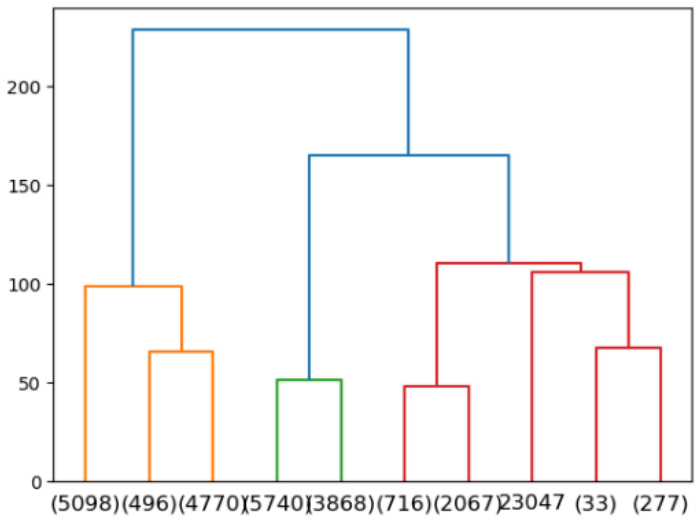
Objective: To identify the hierarchical structure and relationships among different advertisements and determine the optimal number of clusters.

Method:

- Linkage Method: Ward's linkage method, which minimizes the variance within each cluster.
- Distance Metric: Euclidean distance.

Steps:

1. Construct Dendrogram:
    - A visual representation showing how individual ads are merged into clusters.
    - The height of the dendrogram branches indicates the similarity between ads or clusters.
2. Optimal Number of Clusters:
    - Ideal number of clusters are 10
    - Look for the 'elbow' point, where the vertical distance between clusters is significantly higher, indicating a natural cut-off for clusters.



Optimal Number of Clusters are 10

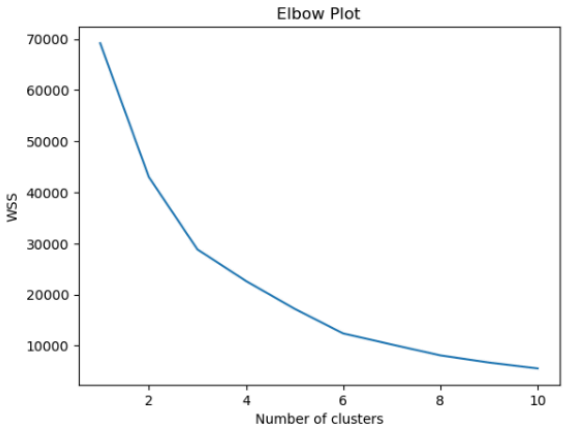
(Fig 8 : Dendrogram Plot)

K-means Clustering

Objective: To segment the advertisements into distinct clusters based on their features using K-means clustering.

Method:

- Algorithm: K-means clustering, which partitions the data into k clusters by minimizing the sum of squared distances between points and their respective cluster centroids.
- Initialization: Multiple runs with different centroid seeds to ensure robustness.



Elbow Curve: Plot the Within-Cluster-Sum of Squared Errors (WSS) against the number of clusters (k).

(Fig 9 : Elbow Plot)

Silhouette Scores: Top 10 Silhouette Scores

```
[0.5231204994359623,
 0.5733734624439143,
 0.5930413256947724,
 0.5965649588555965,
 0.574563273876748,
 0.5778925624534481,
 0.5898645726326806,
 0.549952572664179,
 0.5534277426173733]
```

(Fig 10 : elbow curve with WSS values)

Overall Range: The scores range from 0.5231 to 0.5966. All scores are positive, indicating that the clusters are generally well-defined.

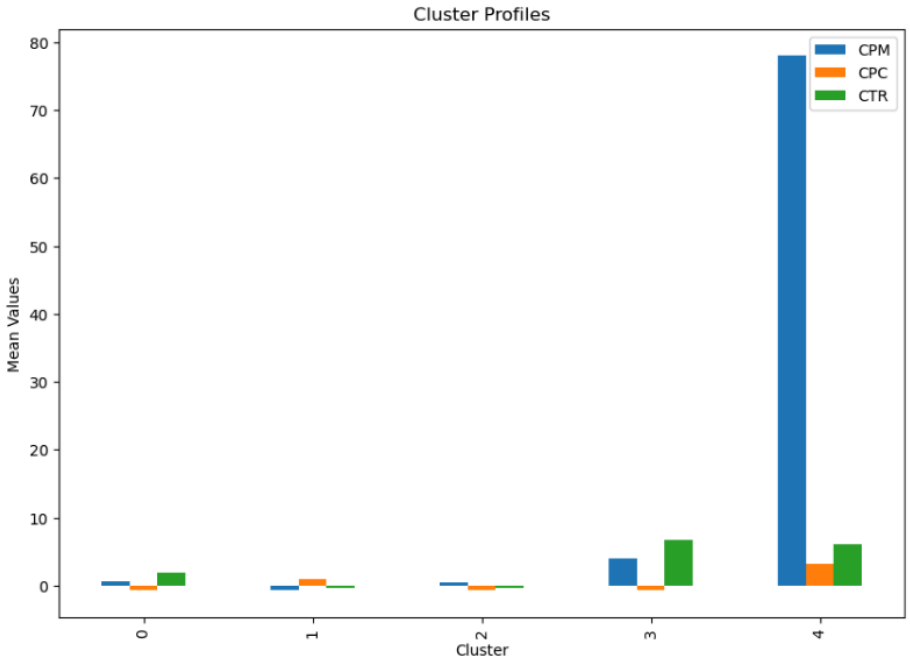
Highest Scores: The highest silhouette score is 0.5966, suggesting that this particular clustering configuration has the best-defined clusters with the highest separation between them.

Consistent Performance: Most of the scores are clustered around the 0.55 to 0.59 range, indicating consistent and reasonably good cluster separation across different clustering configurations.

Lowest Scores: The lowest score is 0.5231, which, while still positive, indicates relatively lower separation compared to the highest score but still suggests reasonable clustering quality.

Conclusion:

The silhouette scores indicate a good overall performance of the clustering algorithm, with all scores being above 0.5. This suggests that the clusters are fairly well-separated and well-defined, though there is some variability in cluster quality.



(Fig 11 :figure showing cluster profiles)

Actionable Insights & Recommendations

Cluster 0:

CPM (Cost Per Thousand Impressions): Slightly above average

CPC (Cost Per Click): Below average

CTR (Click-Through Rate): Significantly above average

Insight: These ads have a relatively high engagement rate at a lower cost per click. They are cost-efficient in driving clicks.

Cluster 1:

CPM: Below average

CPC: Above average

CTR: Below average

Insight: These ads are cheaper per thousand impressions but more expensive per click and have a lower engagement rate. They are less effective in generating clicks.

Cluster 2:

CPM: Slightly above average

CPC: Below average

CTR: Below average

Insight: These ads have average costs with below-average engagement rates. They may need optimization to improve performance.

Cluster 3:

CPM: Significantly above average

CPC: Below average

CTR: Significantly above average

Insight: These ads have high costs but also very high engagement rates, making them highly effective despite the high CPM.

Cluster 4:

CPM: Extremely high

CPC: Very high

CTR: Very high

Insight: These ads are extremely expensive but also drive high engagement. These could be premium ad placements that yield strong results despite their high cost.

Recommendations

Optimize Ads in Cluster 0:

These ads are performing well with high engagement rates and low cost per click. Maintain these ads and use them as a benchmark for other campaigns.

Replicate the strategies used for these ads in other clusters to improve overall performance.

Improve Ads in Cluster 1:

These ads have low engagement and high cost per click. Consider revising the ad creatives, target audience, or overall strategy.

Test different variations and optimize the elements that are underperforming.

Monitor Ads in Cluster 2:

These ads are average in cost and engagement. Regularly monitor these ads and make incremental changes to improve their performance. Conduct A/B testing to identify effective changes.

Leverage High-Performing Ads in Cluster 3:

These ads are very effective in driving engagement. Ensure that the high cost is justified by the return on investment.

Use these ads as a model for creating new high-engagement ads.

Evaluate Ads in Cluster 4:

These ads are extremely expensive but also drive significant engagement. Assess whether the high cost is sustainable and justified.

Negotiate better rates or optimize placements to reduce CPM while maintaining high CTR.

Additional Actionable Steps

Segment by Device Type: Further analyze the performance of ads within each cluster by device type (e.g., mobile, desktop) to understand which devices drive the most engagement. This can help optimize ad spend accordingly.

Continuous Monitoring and Adjustment: Regularly monitor the performance of ads within each cluster. Adjust strategies based on real-time data to ensure optimal performance.

Budget Reallocation: Allocate more budget to ads in Clusters 0 and 3, which are performing well, and reduce the budget for ads in Clusters 1 and 2 until improvements are identified. Consider whether the high costs of Cluster 4 are justified and if so, maintain or slightly reduce the budget for these premium ads.

3. Problem 2: PCA Analysis

3.1 Problem Definition and EDA

Problem Definition

Context:

The 2011 Census of India is one of the most comprehensive and systematic records of the country's demographic data. It includes detailed information on various aspects like population, literacy, workforce distribution, and more. Specifically, the dataset for this project pertains to female-headed households, excluding institutional households, at the district level across India. The richness and diversity of this data present a significant challenge: with so many variables, it becomes difficult to comprehend and extract meaningful insights directly.

Problem Statement:

The primary objective of this analysis is to perform a detailed exploratory data analysis (EDA) and utilize Principal Component Analysis (PCA) to reduce the dimensionality of the dataset. By identifying the most important principal components, we can explain the maximum variance in the data with fewer variables. This simplification aims to:

1. Facilitate better understanding and visualization of the data.
2. Uncover underlying patterns and relationships.
3. Make it easier to draw actionable insights for policy-making and resource allocation.

Exploratory Data Analysis (EDA):

1. Categorical Variables:
    - State: Coded representation of different states and union territories.
    - District: Coded representation of various districts.
2. Demographic Variables:
    - Variables like `TOT_M`, `TOT_F`, `M_06`, `F_06` provide a breakdown of the population based on gender and age group.
    - Literacy and Illiteracy rates among males (`M_LIT`, `M_ILL`) and females (`F_LIT`, `F_ILL`).
3. Workforce Variables:
    - `TOT_WORK_M`, `TOT_WORK_F` represent the total working population segmented by gender.
    - Detailed breakdown of occupations such as cultivators, agricultural laborers, household industry workers, and other workers for both genders.
4. Marginal Workers:
    - Segmented across different time periods (e.g., worked less than 6 months, 3-6 months, and less than 3 months) for both genders.
5. Non-Working Population:
    - `NON_WORK_M` and `NON_WORK_F` representing non-working males and females in female-headed households.

Data Overview :

The dataset has 640 rows and 61 columns. It is always a good practice to view a sample of the rows. A simple way to do that is to use head() function.

State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	MARG_CL_	MARG_CL_	MARG_AL_	MARG_AL_	MARG_HH_	MARG_HH_	MARG_OT_	MARG_OT_	NON_WORI	NON_WORK_F
0	1	1	Jammu & Ka	Kupwara	7707	23388	29796	5862	6196	...	1150	749	180	237	680	252	32	46	258	214
1	1	2	Jammu & Ka	Badgam	6218	19585	23102	4482	3733	...	525	715	123	229	186	148	76	178	140	160
2	1	3	Jammu & Ka	Leh(Ladakh	4452	6546	10964	1082	1018	...	114	188	44	89	3	34	0	4	67	61
3	1	4	Jammu & Ka	Kargil	1320	2784	4206	563	677	...	194	247	61	128	13	50	4	10	116	59
4	1	5	Jammu & Ka	Punch	11654	20591	29981	5157	4587	...	874	1928	465	1043	205	302	24	105	180	478

(Table 3: Top five rows of the dataset for Problem2)

Descriptive Statistics :

	State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
count	640	640	640	640	640	640	640	640	640	640	...	640	640	640	640	640	640	640	640	640	640
mean	17.114062	320.5	51222.87188	79540.57656	122372.0844	12306.09844	11942.3	13820.54688	20778.39219	6151.807813	...	1392.573438	2757.05	250.885062	556.038438	560.690625	1283.45125	71.375688	200.742188	510.014063	704.778125
std	9.426486	184.896367	48135.40548	73384.51111	113600.7173	11500.90688	11326.29457	14426.37313	21727.88771	9512.668948	...	1468.707052	2788.776676	453.336594	1117.642748	762.578991	1585.377936	107.857627	309.740854	610.603187	910.209225
min	1	1	350	391	698	56	56	0	0	0	...	4	30	0	0	0	0	0	0	0	5
25%	9	160.75	15484	30228	46517.75	4733.75	4672.25	3466.25	5603.25	293.75	...	469.5	957.25	47	109	136.5	298	14	43	161	220.5
50%	18	320.5	35837	58339	87724.5	9159	8663	9591.5	13709	2333.5	...	949	1928	114.5	247.5	308	717	35	113	326	464.5
75%	24	480.25	68892	107918.5	164251.75	16520.25	15902.25	19429.75	29160	7658	...	1714	3559.75	270.75	568.75	642	1710.75	79	240	604.5	853.5
max	35	640	310450	465417	750352	96223	95129	103307	156429	96785	...	9875	21611	5775	17153	6116	13714	895	3354	6456	10533

(Table 4: Descriptive statistics of the dataset for Problem 2)

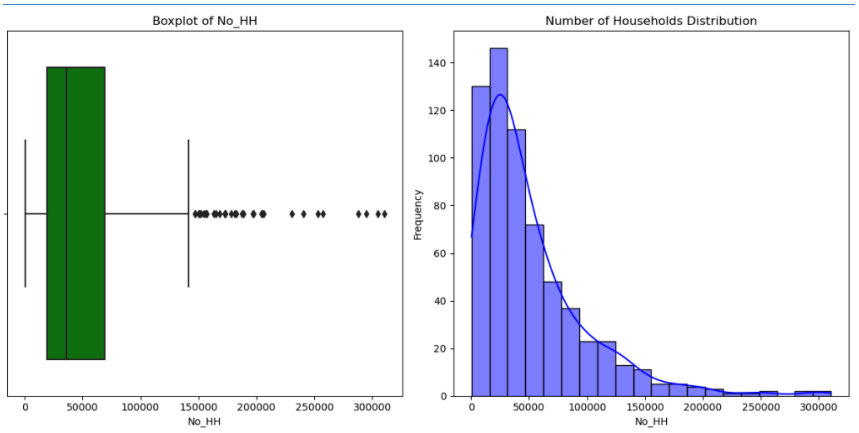
Key Insights:

- **Population Distribution**: There's a wide variation in the population metrics across districts, with significant differences between the minimum and maximum values, indicating diverse demographic structures.
- **Scheduled Castes and Tribes**: The population of Scheduled Castes and Tribes also shows significant variability, reflecting the diverse socio-economic conditions across districts.
- **Children (0-6 years)**: The mean values for children aged 0-6 indicate a substantial young population in these households.
- **Non-Workers**: A substantial portion of the population is classified as non-workers, with a slightly higher number of non-working females compared to males.

Univariate Analysis

For performing Univariate analysis we will take a look at the Boxplots and Histograms to get a better understanding of the distributions.

Observations on Households Distribution



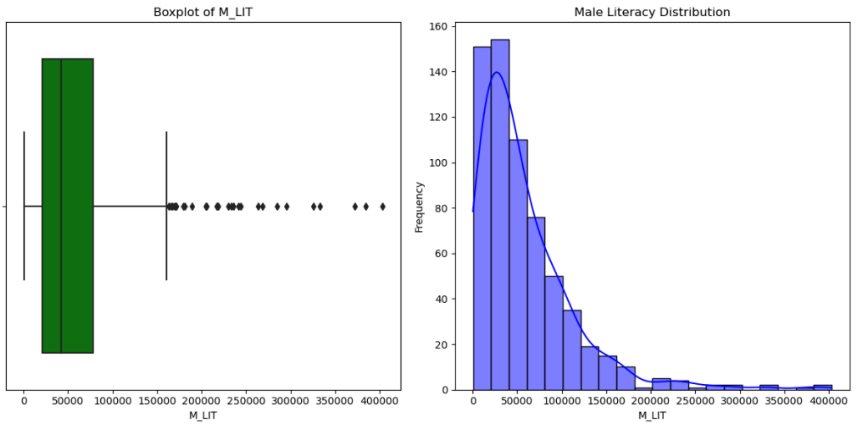
(Fig 12 : Plot showing the boxplot and distribution for Households Distribution)

Observations:

● The Distribution of Households Distribution is right skewed.

● There are a lot of outliers in this variable.

Observations on Male Literacy Distribution



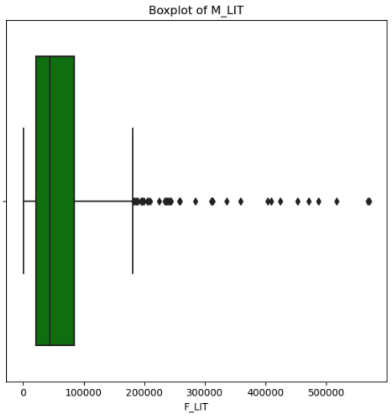
(Fig 13 : Plot showing the boxplot and distribution for Male Literacy Distribution)

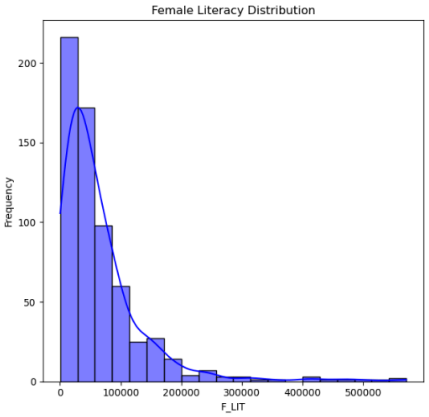
Observations:

● The Distribution of Male Literacy Distribution is right skewed.

● There are a lot of outliers in this variable.

Observations on Female Literacy Distribution





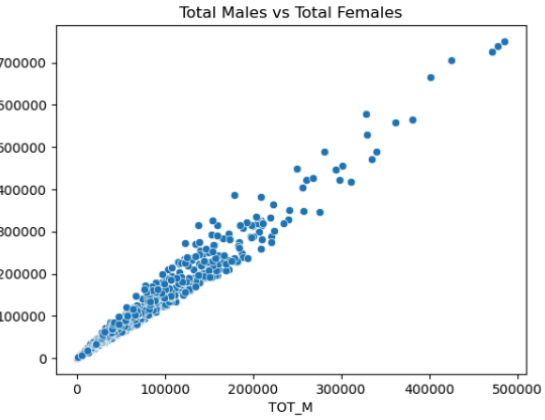
(Fig 14 : Plot showing the boxplot and distribution for Female Literacy Distribution)

Observations:

- The Distribution of Female Literacy Distribution is right skewed.
- There are a lot of outliers in this variable.

Bivariate analysis

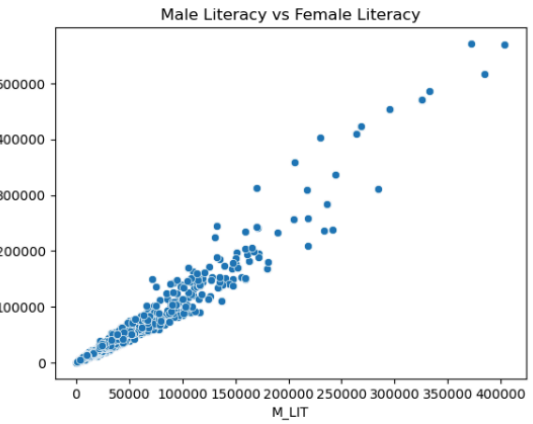
Total Males vs Total Females



(Fig 15. showing the scatterplot between Total Males vs Total Females)

Positive Correlation: The scatter plot shows a positive correlation between Total Males and Total Females

Male Literacy vs Female Literacy



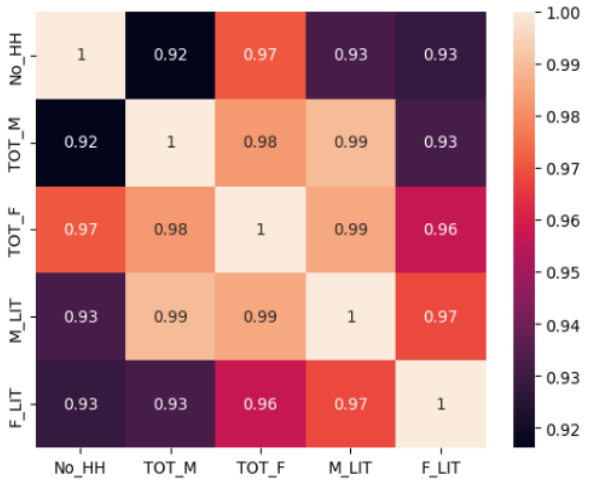
(Fig 16 showing the scatterplot between Male Literacy vs Female Literacy)

Positive Correlation: The scatter plot shows a positive correlation between Male Literacy vs Female Literacy and the literacy rate seems to almost the same including a few outliers

Multivariate Analysis

Correlation of Numerical Variables

(fig 17 showing heatmap of Correlation of numerical variables)



Interpretation:

- High Correlations: Most of the variables are highly correlated with each other, suggesting that regions with a higher number of households tend to have higher populations (both male and female) and higher literacy rates (both male and female).
- Socio-Demographic Trends: The strong positive correlations indicate that the increase in one demographic metric is generally associated with an increase in other related metrics. For example, a higher number of households is strongly associated with a higher total population and higher literacy rates.
- Literacy Correlations: The strong correlation between male and female literates (0.968) indicates that areas with high literacy among males also tend to have high literacy among females, pointing to a general trend of literacy improvement across both genders in these areas.

Example Questions

Which state has the highest and lowest gender ratio?

Ans : Andhra Pradesh Lakshadweep

To understand the gender dynamics within the dataset, a new variable called Gender_Ratio was calculated. This ratio is defined as the total number of females divided by the total number of males in each state. Here are the key findings from this analysis:

Gender_Ratio Calculation:

- The Gender_Ratio is a simple metric that helps to understand the balance between the number of females and males in each state. A ratio greater than 1 indicates more females than males, while a ratio less than 1 indicates more males than females.

States with the Highest and Lowest Gender Ratios:

- Highest Gender Ratio: The state with the highest gender ratio, indicating the largest proportion of females relative to males, is found to be Andhra Pradesh.
- Lowest Gender Ratio: Conversely, the state with the lowest gender ratio, indicating the smallest proportion of females relative to males, is [Lakshadweep].

Which district has the highest & lowest gender ratio?

Ans :

The aim of this analysis is to identify the districts with the highest and lowest gender ratios, which is defined as the ratio of the total female population (TOT_F) to the total male population (TOT_M).

Formula Used: The gender ratio for each district is calculated using the formula: Gender Ratio = Total Female Population (TOT_F)/Total Male Population (TOT_M)

District with the Highest Gender Ratio:

- **District Code**: [547]
- **District Name**: [Krishna]
- This district has a gender ratio indicating a higher proportion of females to males.

District with the Lowest Gender Ratio:

- **District Code**: [587]
- **District Name**: [Lakshadweep]
- This district has a gender ratio indicating a lower proportion of females to males.

3.2 Data Preprocessing

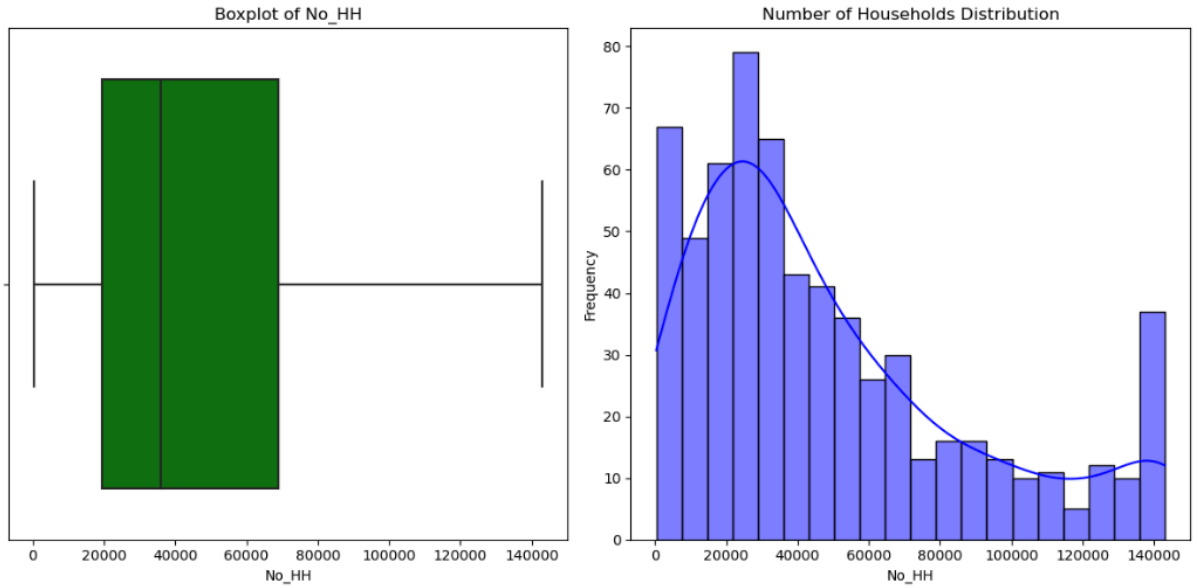
Check for missing values : There are no missing values

Treating outliers

Outliers can significantly impact the results of data analysis and modeling. To ensure robust and accurate analysis, we employed the Interquartile Range (IQR) method for outlier treatment. This method involves the following steps:

1. **Calculation of IQR**:
    - For each numerical column, the first quartile (Q1) and the third quartile (Q3) were calculated.
    - The Interquartile Range (IQR) was computed as the difference between Q3 and Q1.
2. **Determination of Bounds**:
    - The lower bound was defined as Q1−1.5×IQRQ1
    - The upper bound was defined as Q3+1.5×IQRQ3
3. **Treatment of Outliers**:
    - Any data point below the lower bound was replaced with the lower bound value.
    - Any data point above the upper bound was replaced with the upper bound value.

This method ensures that extreme outliers are adjusted to fall within a reasonable range, minimizing their impact on the analysis. By treating outliers using the IQR method, the dataset becomes more robust and reliable for further analysis, leading to more accurate insights and conclusions.



(Fig 18 : Box plots of Distribution Between the Number of Households Distribution)

included just one boxplot for reference

Scaling

State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
-1.710782	-1.728347	-1.038988	-0.874837	-0.937027	-0.824685	-0.561282	-1.080201	-1.079863	-0.51044	...	-0.093587	-0.860882	-0.0418	-0.423378	0.523468	-0.793688	-0.443385	-0.756991	-0.497493	-0.774865
-1.710782	-1.723834	-1.078898	-0.938023	-1.009723	-0.775832	-0.836657	-1.079873	-1.079835	-0.771833	...	-0.719169	-0.877096	-0.34704	-0.44379	-0.834478	-0.884805	0.368782	0.088381	-0.813235	-0.880662
-1.710782	-1.718821	-1.121858	-1.164865	-1.141539	-1.141842	-1.138104	-1.080201	-1.079835	0.122588	...	-1.130551	-1.128423	-0.770091	-0.800999	-1.063434	-0.964684	-1.026779	-1.025978	-1.008568	-1.074622
-1.710782	-1.713109	-1.201599	-1.217171	-1.21493	-1.197772	-1.176091	-1.080447	-1.079883	-0.399531	...	-1.050477	-1.100286	-0.679055	-0.701491	-1.039994	-0.970686	-0.953855	-0.98755	-0.877454	-1.078541
-1.710782	-1.707896	-0.938495	-0.921309	-0.935018	-0.700931	-0.740523	-1.078807	-1.07816	0.432534	...	-0.368844	-0.288817	1.484398	1.83313	-0.589942	-0.749882	-0.589234	-0.379131	-0.706204	-0.257837

(Table 5 showing scaled data for problem 2)

Objective:

To scale the numerical features in the dataset to ensure that they are on a similar scale, which is a critical step for distance-based algorithms like K-means and hierarchical clustering. This helps in improving the performance and accuracy of clustering algorithms.

Methods and Approaches:

Choice of Scaling Method: Z-score Normalization

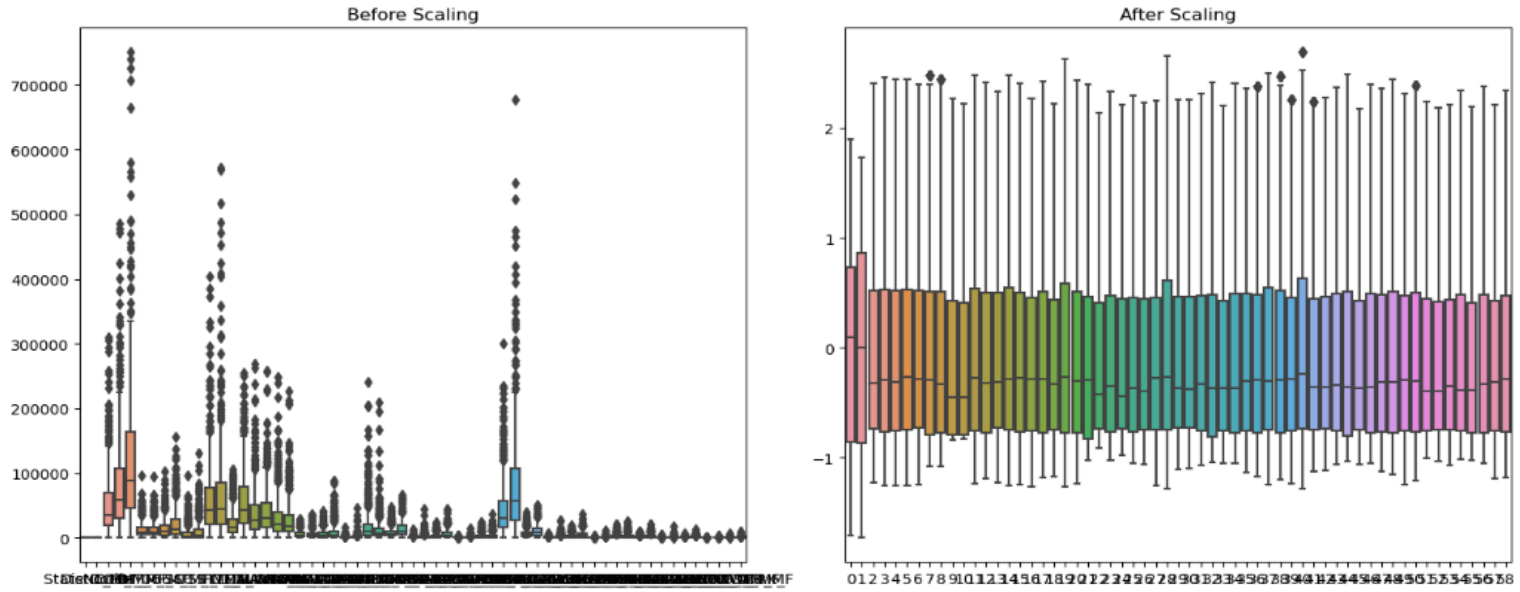
Reason: Z-score normalization (or standardization) transforms the data to have a mean of zero and a standard deviation of one. This centers the data and scales it based on the inherent spread, which ensures that all features have the same importance during clustering.

Implementation Steps:

Calculate the mean ((\mu)) and standard deviation ((\sigma)) for each numerical feature.

Transform the original values using the Z-score formula.

(Fig.19 Visualising the data before and after scaling)



We should consider handling outliers before performing PCA because they can distort the results.

By carefully handling outliers, we ensure that PCA gives a more accurate and meaningful representation of the data structure.

3.3 PCA Analysis

Objective:

To reduce the dimensionality of the dataset by transforming the original set of variables into a new set of uncorrelated variables (principal components) that capture the maximum variance in the data. This helps in simplifying the dataset, making it easier to interpret and visualize.

```
array([[1.00156495, 0.99457535, 0.40338248, ..., 0.05909443, 0.12308779,
        0.2447105 ],
       [0.99457535, 1.00156495, 0.39634859, ..., 0.06808479, 0.11066561,
        0.22599926],
       [0.40338248, 0.39634859, 1.00156495, ..., 0.65276151, 0.76840117,
        0.79788409],
       ...,
       [0.05909443, 0.06808479, 0.65276151, ..., 1.00156495, 0.76249106,
        0.72075284],
       [0.12308779, 0.11066561, 0.76840117, ..., 0.76249106, 1.00156495,
        0.90224595],
       [0.2447105 , 0.22599926, 0.79788409, ..., 0.72075284, 0.90224595,
        1.00156495]])
```

Key Observations

1. Variances: The diagonal elements (e.g., 1.00156495) represent the variances of the respective variables. These values provide insights into the variability within each variable.
2. Covariances: The off-diagonal elements (e.g., 0.99457535) indicate the covariance between pairs of variables. These values reveal how changes in one variable are associated with changes in another variable.

Insights and Implications

- Strong Correlations: Variables with high covariance values, such as the first and second variables (0.99457535), exhibit strong positive correlations. This suggests that these variables tend to move together, which is important for understanding market trends and dependencies.
- Weak Relationships: Some variables, like the first and 50th (0.05909443), have low covariance values, indicating weak or no significant relationship. This information is useful for diversification strategies, as weakly related variables can reduce overall risk.
- Risk Management: By identifying pairs of variables with high covariances, we can assess potential risks where variables might exhibit similar behaviors during market fluctuations.
- Strategic Decisions: Understanding these relationships aids in resource allocation, inventory management, and other operational decisions, ensuring that strategies are data-driven and well-informed.

Conclusion

The covariance matrix serves as a foundational tool for analyzing the relationships between variables in our dataset. By interpreting these covariances, we can make better strategic decisions, optimize our portfolios, and manage risks effectively.

Eigenvalues and Eigenvectors

To identify the principal components, we first compute the eigenvalues and eigenvectors from the covariance matrix. The eigenvalues and eigenvectors obtained are as follows:

Eigenvalues:

```
[3.57108475e+01, 7.98557733e+00, 4.50785903e+00, 2.77867519e+00,
 1.97472860e+00, 1.17776767e+00, 1.13039501e+00, 7.22103375e-01,
 4.64431676e-01, 3.46774532e-01, 3.05963732e-01, 2.68366978e-01,
 2.20811847e-01, 1.80278141e-01, 1.68296796e-01, 1.32409265e-01,
 1.29436740e-01, 1.03406138e-01, 9.55347371e-02, 8.58417456e-02,
 8.09066019e-02, 6.56476263e-02, 6.23708292e-02, 4.79008765e-02,
 4.56408231e-02, 4.38435424e-02, 3.10290046e-02, 2.86009130e-02,
 2.74987147e-02, 2.33916183e-02, 2.16432655e-02, 1.87723745e-02,
 1.56678899e-02, 1.40371782e-02, 1.18761437e-02, 1.11316049e-02,
 9.08077540e-03, 7.25913797e-03, 6.18691864e-03, 4.89879738e-03,
 4.55034891e-03, 4.24001897e-03, 3.26372660e-03, 2.18239672e-03,
 2.12902353e-03, 1.90742071e-03, 1.43490578e-03, 1.09833856e-03,
 9.62038195e-04, 8.56614567e-04, 6.51562449e-04, 5.76295579e-04,
 4.31846786e-04, 3.69015469e-04, 3.06582238e-04, 2.07171377e-04,
 1.38262667e-04, 8.95049889e-05, 4.60548161e-05])
```

Eigenvectors:

```
[ 3.03688860e-02,  3.03440259e-02,  1.49578258e-01, ...,
  1.41068150e-01,  1.47433908e-01,  1.42143173e-01],
[-1.72196615e-01, -1.69016895e-01, -1.19691823e-01, ...,
  4.13029333e-02, -3.75381832e-02, -3.86825793e-02],
[ 3.01346957e-01,  3.06417948e-01,  6.82434200e-02, ...,
 -7.35067283e-02, -1.12204882e-01, -1.79425675e-02],
...,
[ 1.13328651e-03, -9.20198117e-04,  8.13952594e-04, ...,
 -1.16253522e-02,  5.62399229e-02, -6.20621258e-03],
[ 2.52804397e-03, -1.82690047e-03, -6.49442371e-05, ...,
  1.42236570e-02, -7.70492389e-02, -8.36657342e-04],
[-2.13017946e-03,  1.60157621e-03, -4.00867509e-03, ...,
  2.09764678e-03,  5.52636585e-03,  1.52530297e-03]])
```

Key Observations

1. **Magnitude of Variance**: The eigenvalues indicate the magnitude of the variance captured by each principal component. The first few eigenvalues are significantly larger than the others, indicating that the first few principal components capture most of the variance in the dataset.
2. **Principal Components:** The eigenvectors corresponding to the largest eigenvalues determine the directions of the principal components. These principal components represent the new feature space that maximizes the variance captured from the original dataset.

3. **Insights and Implications**

- Dimensionality Reduction: By focusing on the principal components with the highest eigenvalues, we can reduce the dimensionality of our dataset while retaining most of the variance. This simplifies our analysis and makes it more computationally efficient.
- Data Visualization: The principal components can be used to visualize the data in a lower-dimensional space, making it easier to identify patterns and clusters.
- Feature Importance: The magnitude of the eigenvalues helps us understand the importance of each principal component. Components with higher eigenvalues are more important and capture more information from the data.

Identifying the Optimal Number of Principal Components (PCs)

Problem Statement

Our objective is to identify the optimal number of principal components (PCs) that capture at least 90% of the variance in our dataset. This helps in reducing the dimensionality of the data while retaining most of the important information.

Analysis

To determine the optimal number of PCs, we performed Principal Component Analysis (PCA) and calculated the cumulative variance explained by each principal component. The explained variance ratio indicates how much

variance each principal component captures, and the cumulative variance explained shows the total variance captured as more principal components are included.

Results

The cumulative variance explained by the principal components is as follows:

```
Cumulative Variance Explained in Percentage: [ 60.43  73.95  81.57  86.28  89.62  91.61  93.52  94.75  95.53  96.12
  96.64  97.09  97.46  97.77  98.05  98.28  98.5   98.67  98.83  98.98
  99.12  99.23  99.33  99.41  99.49  99.57  99.62  99.67  99.71  99.75
  99.79  99.82  99.85  99.87  99.89  99.91  99.93  99.94  99.95  99.96
  99.96  99.97  99.98  99.98  99.98  99.99  99.99  99.99  99.99 100.
 100.   100.   100.   100.   100.   100.   100.   100.   100.  ]
```

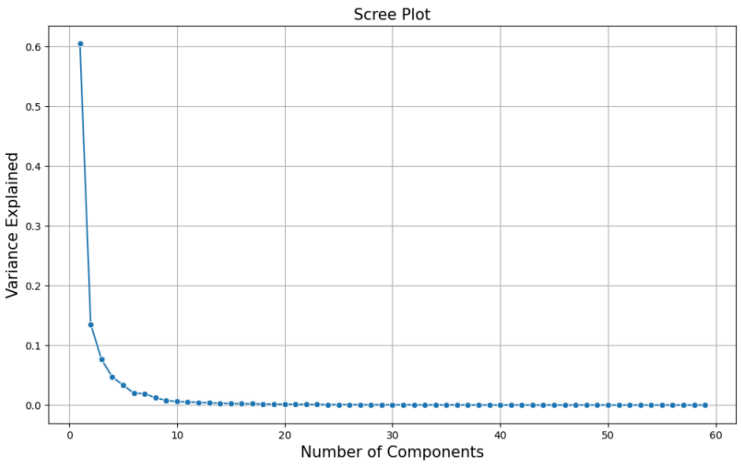
From this analysis, we observe the following:

- The first principal component captures 60.43% of the variance.
- The cumulative variance reaches 90% with the inclusion of the 6th principal component.
- Including more than 6 principal components continues to capture more variance, but the incremental benefit diminishes significantly after reaching 90%.

Conclusion

Based on the cumulative variance explained, the optimal number of principal components is 6, as this captures approximately 91.61% of the total variance. Using these 6 principal components allows us to reduce the dimensionality of the dataset while retaining most of the important information, thereby simplifying our analysis and enhancing computational efficiency.

Scree Plot to identify the number of components to be built



(Fig 20 showing scree plot)

The number of components can be decided based upon the explained variance. Here, it is decided to keep the number of components as 6 as the cumulative explained variance is around 90%

Explained Variance Ratio

The explained variance ratio for the first six principal components is as follows:

array([0.6 , 0.14, 0.08, 0.05, 0.03, 0.02])

Key Observations

1. **Principal Component 1 (PC1)**:
    - PC1 captures 60% of the total variance in the dataset.
    - This indicates that PC1 is the most significant component in terms of variance explained.
2. **Principal Component 2 (PC2)**:
    - PC2 captures an additional 14% of the variance.
    - Together with PC1, it explains 74% of the total variance (60% + 14%).
3. **Principal Component 3 (PC3)**:
    - PC3 captures 8% of the variance.
    - The cumulative variance explained by the first three PCs is 82% (60% + 14% + 8%).
4. **Principal Component 4 (PC4)**:
    - PC4 explains 5% of the variance.
    - The cumulative variance explained by the first four PCs is 87% (60% + 14% + 8% + 5%).
5. **Principal Component 5 (PC5)**:
    - PC5 captures 3% of the variance.
    - The cumulative variance explained by the first five PCs is 90% (60% + 14% + 8% + 5% + 3%).
6. **Principal Component 6 (PC6)**:
    - PC6 explains 2% of the variance.
    - The cumulative variance explained by the first six PCs is 92% (60% + 14% + 8% + 5% + 3% + 2%).

Conclusion

Based on the explained variance ratio, we can conclude:

- **PC1** is the most important principal component, capturing 60% of the total variance.
- The first **three principal components (PC1, PC2, and PC3)** together explain 82% of the variance.
- To capture at least 90% of the variance, the first **five principal components (PC1 to PC5)** are sufficient.
- **PC6** contributes an additional 2%, bringing the total variance explained to 92%.

Linear Equation for the First Principal Component (PC1)

Principal Component Equation

The equation for the first principal component (PC1) is a linear combination of the original variables, weighted by the coefficients from the first eigenvector. The general form is:

$$PC1 = x1 \cdot v1 + x2 \cdot v1,2 + ... + xn \cdot v1,n$$

where:

- xi represents the original variables.
- v1,i represents the coefficients (weights) from the first eigenvector.

```
'0.030 * State Code + 0.030 * Dist.Code + 0.150 * No_HH + 0.159 * TOT_M + 0.158 * TOT_F + 0.156 * M_06 + 0.156 * F_06 + 0.143 * M_SC + 0.143 * F_SC + 0.
019 * M_ST + 0.018 * F_ST + 0.155 * M_LIT + 0.146 * F_LIT + 0.154 * M_ILL + 0.158 * F_ILL + 0.154 * TOT_WORK_M + 0.143 * TOT_WORK_F + 0.142 * MAINWORK_M
+ 0.126 * MAINWORK_F + 0.111 * MAIN_CL_M + 0.083 * MAIN_CL_F + 0.120 * MAIN_AL_M + 0.091 * MAIN_AL_F + 0.142 * MAIN_HH_M + 0.134 * MAIN_HH_F + 0.123 * M
AIN_OT_M + 0.117 * MAIN_OT_F + 0.156 * MARGWORK_M + 0.149 * MARGWORK_F + 0.087 * MARG_CL_M + 0.064 * MARG_CL_F + 0.127 * MARG_AL_M + 0.116 * MARG_AL_F +
0.145 * MARG_HH_M + 0.142 * MARG_HH_F + 0.151 * MARG_OT_M + 0.148 * MARG_OT_F + 0.158 * MARGWORK_3_6_M + 0.156 * MARGWORK_3_6_F + 0.157 * MARG_CL_3_6_M
+ 0.149 * MARG_CL_3_6_F + 0.094 * MARG_AL_3_6_M + 0.066 * MARG_AL_3_6_F + 0.128 * MARG_HH_3_6_M + 0.114 * MARG_HH_3_6_F + 0.145 * MARG_OT_3_6_M + 0.141
* MARG_OT_3_6_F + 0.151 * MARGWORK_0_3_M + 0.148 * MARGWORK_0_3_F + 0.142 * MARG_CL_0_3_M + 0.133 * MARG_CL_0_3_F + 0.062 * MARG_AL_0_3_M + 0.056 * MARG
_AL_0_3_F + 0.119 * MARG_HH_0_3_M + 0.113 * MARG_HH_0_3_F + 0.142 * MARG_OT_0_3_M + 0.141 * MARG_OT_0_3_F + 0.147 * NON_WORK_M + 0.142 * NON_WORK_F'
```

Explanation

- PC1 is constructed from a combination of multiple demographic and employment-related variables.
- Each coefficient represents the weight or contribution of the corresponding variable to PC1.
- Variables with higher coefficients have a more significant impact on the principal component.
- The positive coefficients across variables indicate that an increase in these variables will increase the value of PC1.

This linear equation allows us to understand how different variables contribute to the overall variance captured by the first principal component.

Conclusion

- The Indian Census, renowned for its comprehensive and detailed nature, has been conducted every ten years since 1881.
- The 2011 Census was the fifteenth in the series since 1872, the seventh after independence, and the second of the third millennium.
- Despite challenges like wars, epidemics, and natural calamities, the Census has continued uninterrupted.

Key Points of the 2011 Census:

- Coverage: Included 35 States/Union Territories, 640 districts, 5,924 sub-districts, 7,935 towns, and 640,867 villages.
- Data Collected: The Primary Census Abstract (PCA) includes information on area, total households, total population, scheduled castes (SC), scheduled tribes (ST) population, population in the age group 0-6, literates, main workers, marginal workers, and non-workers.
- Categories of Workers: The data is categorized into four broad industrial categories: cultivators, agricultural laborers, household industry workers, and other workers.

Specific Data for Female Headed Households Excluding Institutional Households:

- Variables Analyzed: State Code, District Code, Number of Households (No_HH), Total Male Population (TOT_M), Total Female Population (TOT_F), Male Literates (M_LIT), Female Literates (F_LIT), Male Population in the Age Group 0-6 (M_06), etc.
- Principal Component Analysis (PCA):
    - PCA identifies the principal components (PCs) that explain the variance in the data.
    - Key Principal Components and their Loadings:
        - PC1: Strong positive loadings on No_HH, TOT_M, TOT_F, and M_06.
        - PC2: Negative loadings on most variables except TOT_F.
        - PC3: Positive loadings on No_HH and TOT_F.
    - Cumulative Variance Explained: The first few principal components explain a high percentage of the total variance, with the first six PCs explaining over 91.61% of the variance.

Correlation Analysis:

- High positive correlations between:

    - No_HH and TOT_M (0.916), TOT_F (0.971)
    - TOT_M and TOT_F (0.983), M_LIT (0.989)
    - TOT_F and M_LIT (0.985), F_LIT (0.957)
    - M_LIT and F_LIT (0.968)
- This indicates a strong relationship between the number of households, total population, and literacy rates among both males and females.

Eigenvectors:

- Eigenvectors represent the direction of the principal components in the multidimensional space.
- The first few eigenvectors have significant values indicating the variables that contribute most to each principal component.

Conclusion:

The 2011 Census data for female-headed households reveals strong interrelations among household count, total population, and literacy rates. Principal Component Analysis helps in reducing the dimensionality of the data while preserving most of the variance, and correlation analysis shows strong positive relationships among key demographic variables.

6. References

- **Previous jupyter notebooks**
- **Google**
- **Course Resources**