MACHINE LEARNING – 1

Table of Contents

	- Problem Definition - Check shape, Data types, statistical summary - Perform an EDA on the data to extract useful insights Note: 1. 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 2. Example questions to answer from EDA - (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio?	
2.2	**Part 2: PCA: Data Preprocessing** - Check for and treat (if needed) missing values - Check for and treat (if needed) data irregularities - Scale the Data using the z-score method - Visualize the data before and after scaling and comment on the impact on outliers	21
2.3	**Part 2; PCA: PCA** - Create the covariance matrix - Get eigen values and eigen vectors - Identify the optimum number of PCs - Show Scree plot - Compare PCs with Actual Columns and identify which is explaining most variance - Write inferences about all the PCs in terms of actual variables - Write linear equation for first PC Note: For the scope of this project, take at least 90% explained variance.	22

1.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.

Introduction

We will be using the clustering procedure as the requirement clearly demands it. Proceeding further first we will understand what clustering is.
Clustering is a technique in machine learning and data analysis that involves grouping similar data points into distinct subsets or clusters. The goal of clustering is to organize and uncover the underlying structure within a dataset without any predefined labels or categories. In other words, it is an unsupervised learning approach where the algorithm tries to find patterns or relationships in the data based on similarities between data points.

1.1. Part 1: Clustering: Define the problem and perform Exploratory Data Analysis
- Problem definition - Check shape, Data types, statistical summary - Univariate analysis - Bivariate analysis - Key meaningful observations on individual variables and the relationship between variables

Problem Definition:

The objective of this clustering project is to analyse and categorize digital ads based on various features and performance metrics provided in the dataset. The dataset consists of information related to ad characteristics, platform details, device types, and key performance indicators such as impressions, clicks, spend, and revenue. The primary goal is to apply clustering techniques to identify distinct patterns or groups within the data.

Key Deliverables:

Clustered Data:

- Providing a clear presentation of the clustered data, indicating which ads belong to each cluster.

Insights and Patterns:

- Summarising key insights and patterns observed within each cluster.
- Highlighting any noteworthy characteristics or behaviours of ads within specific clusters.

Visualizations:

- Generating visualizations (e.g., plots, charts) to represent the distribution of ads within clusters.
- Including visualizations that showcase performance metrics across different clusters.

Recommendations:

- Offering recommendations based on the findings, such as potential strategies for optimizing ad performance within specific clusters.

Documentation:

- Documenting the entire process, including data pre-processing, clustering algorithm selection, and interpretation of results.
- Providing code documentation for reproducibility.

Check shape, Data types, statistical summary :

```
Shape of the dataset: (23066, 19)

Data Types:
Timestamp                 object
InventoryType             object
Ad - Length                int64
Ad- Width                  int64
Ad Size                    int64
Ad Type                   object
Platform                  object
Device Type               object
Format                    object
Available_Impressions     object
Matched_Queries           object
Impressions               object
Clicks                    object
Spend                    float64
Fee                       object
Revenue                  float64
CTR                       object
CPM                      float64
CPC                      float64
dtype: object
```

Inference from the above analysis:

Dataset Size:

The dataset consists of 23,066 rows and 19 columns.

Temporal Information:

The Timestamp column is represented as an object. It likely contains temporal information, but its current data type suggests it needs to be converted to a datetime type for time-based analysis.

Categorical Variables:

Columns like InventoryType, Ad Type, Platform, Device Type, and Format are categorical, represented as objects. These could be essential for categorical analysis or encoding.

Numeric Variables:

Columns such as Ad - Length, Ad - Width, Ad Size, Spend, Revenue, CPM, CPC are numeric. Numeric columns like Spend, Revenue, CPM, and CPC already have appropriate data types (float64).

Missing Values:

No information on missing values is provided in the data type summary. Handling missing values, especially in numeric columns, will be crucial for accurate analysis.

```
Statistical Summary:
        Ad - Length     Ad- Width        Ad Size          Spend       Revenue  \
count  23066.000000  23066.000000   23066.000000  23066.000000  23066.000000
mean     385.163097    337.896037   96674.468048   2706.625689   1924.252382
std      233.651434    203.092885   61538.329557   4067.927273   3105.238394
min      120.000000     70.000000   33600.000000      0.000000      0.000000
25%      120.000000    250.000000   72000.000000     85.180000     55.365000
50%      300.000000    300.000000   72000.000000   1425.125000    926.335000
75%      720.000000    600.000000   84000.000000   3121.400000   2091.337500
max      728.000000    600.000000  216000.000000  26931.870000  21276.180000

                CPM           CPC
count  18330.000000  18330.000000
mean       7.672045      0.351061
std        6.481391      0.343334
min        0.000000      0.000000
25%        1.710000      0.090000
50%        7.660000      0.160000
75%       12.510000      0.570000
max       81.560000      7.260000
```

Inference from the above analysis:

Ad Dimensions (Ad - Length, Ad - Width, Ad Size):

The majority of ads have diverse dimensions, with lengths ranging from 120 to 728 and widths from 70 to 600 pixels.

The average ad size is around 96674 pixels, with a significant standard deviation, indicating a wide range of ad sizes.

Financial Metrics (Spend, Revenue):

Ad spending varies widely, with some ads having no spend, while others go up to approximately $26,931.87.

Similarly, revenue generated by ads varies, ranging from $0 to $21,276.18.

On average, the dataset sees spending of around $2706.63 and revenue of $1924.25.

Cost Metrics (CPM, CPC):

The Cost Per Mille (CPM) has a broad distribution, ranging from $0 to $81.56. The average CPM is $7.67.

Cost Per Click (CPC) also varies widely, with an average of $0.35 per click.

Minimum and Maximum Values:

The minimum ad length is 120 pixels, and the maximum is 728 pixels.

The minimum ad width is 70 pixels, and the maximum is 600 pixels.

The minimum ad size is 33,600 pixels, and the maximum is 216,000 pixels.

The minimum spending and revenue values are zero, indicating ads with no financial activity.

Central Tendency (Mean, Median):

The mean (average) ad size, spending, and revenue give a sense of central tendency.

Median values provide insights into the middle points of the distributions, helping to understand the data's central position.

Spread (Standard Deviation):

Standard deviation measures the spread of values around the mean. A higher standard deviation indicates more variability in the data.

Percentiles (25th, 50th, 75th):

The 25th, 50th (median), and 75th percentiles provide a sense of the distribution of values across different metrics.

Summary:

The dataset reflects a diverse range of ad dimensions, financial metrics, and cost metrics. Some ads exhibit significant spending and revenue, while others may have minimal financial activity. The spread in ad sizes and financial metrics suggests a heterogeneous landscape within the dataset. Further analysis and clustering may help identify distinct patterns and trends among different ad groups.

Univariate analysis:

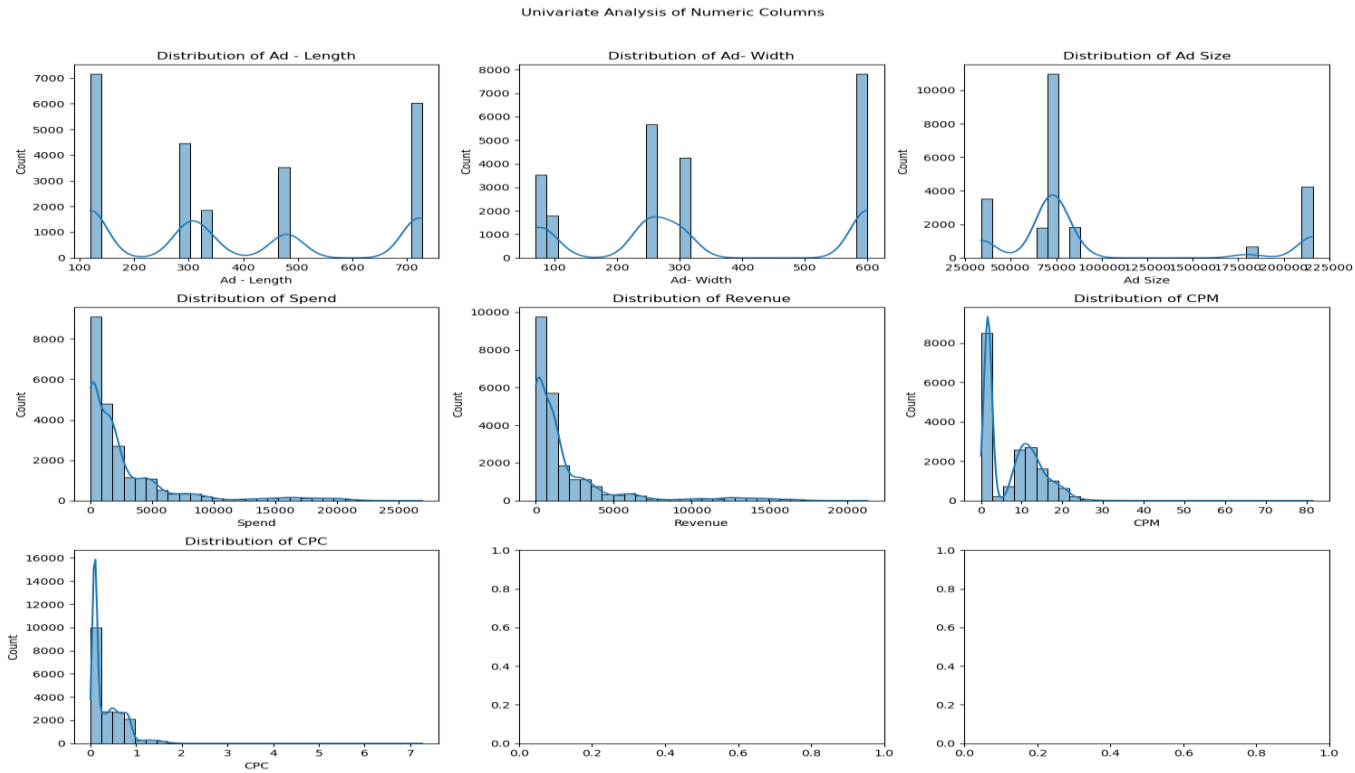

Fig. 01: Univariate Analysis

Inference drawn from the Univariate Analysis:

Ad Dimensions (Ad - Length, Ad - Width, Ad Size):

- The distribution of Ad - Length shows that ads vary widely in length, with a prominent peak around 120 pixels.
- Ad - Width exhibits diverse widths, with a common peak at 300 pixels.
- The distribution of Ad Size indicates a variety of sizes, with a significant concentration around 72,000 pixels.

Financial Metrics (Spend, Revenue):

- The histogram for Spend demonstrates a skewed distribution, with a substantial number of ads having low spending, while a few ads have higher spending.
- Revenue distribution shows a similar pattern, indicating varying levels of revenue generated by ads.

Cost Metrics (CPM, CPC):

- The distribution of CPM suggests a range of cost per mille values, with a common peak around $0 to $2.
- CPC distribution shows a skewed pattern, with a concentration of ads having lower cost per click.

Overall Observations:

- The histograms provide a sense of the spread and concentration of values for each numeric column.
- The variability in ad dimensions, spending, and revenue highlights the diverse nature of the dataset.
- Cost metrics such as CPM and CPC also exhibit variations, with some ads having lower costs and others having higher costs.

Bivariate analysis:

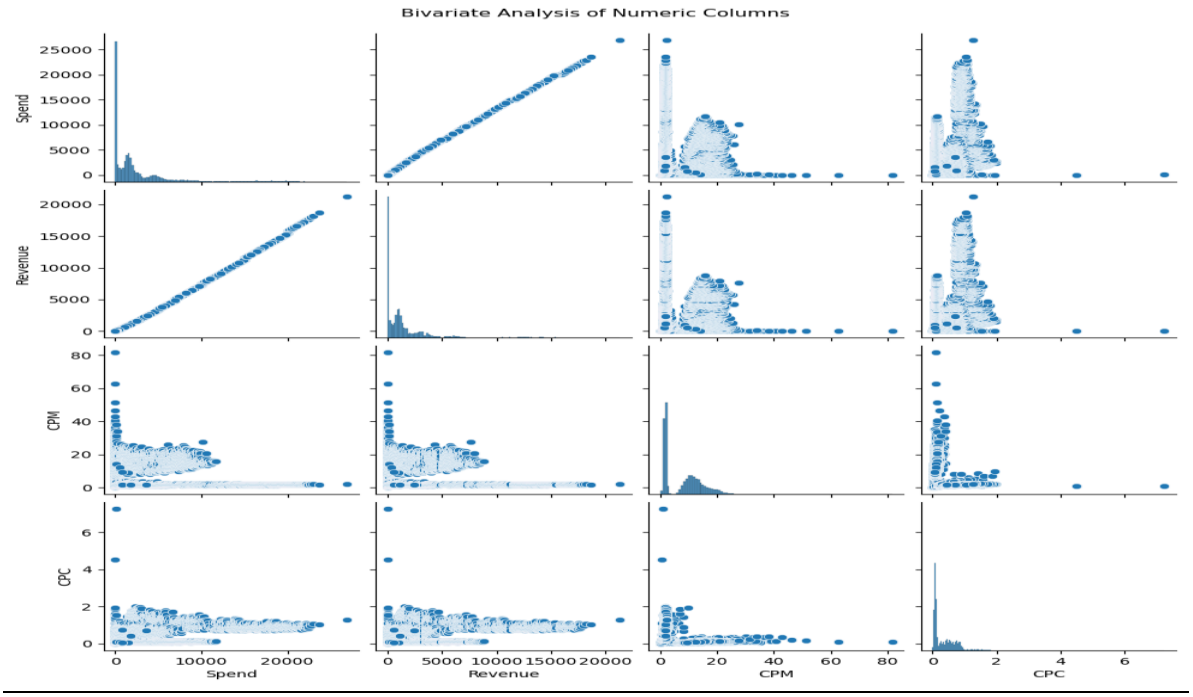

Fig. 02: Bivariate Analysis – Scatterplot

Inference drawn from the Scatterplot:

Spend vs. Revenue:

- Scatter plot indicates a positive trend, suggesting that as spending increases, revenue also tends to increase.
- There are some outliers with high spending and comparatively low revenue.

Spend vs. CPM (Cost Per Mille):

- No clear trend is visible in the scatter plot, implying a relatively weak relationship between spending and CPM.
- The distribution is scattered, suggesting varied CPM values for different spending levels.

Spend vs. CPC (Cost Per Click):

- Scatter plot indicates a diverse relationship, with some ads having low CPC despite higher spending.
- Ads with higher spending and relatively low CPC could be of interest for further investigation.

Revenue vs. CPM:

- Weak trend is observed, with scattered distribution, indicating that CPM alone may not be a strong predictor of revenue.

Revenue vs. CPC:

- Similar to Spend vs. CPC, the scatter plot shows a diverse relationship between revenue and CPC.
- Some ads with higher revenue have relatively low CPC.

CPM vs. CPC:

- No significant trend is evident in the scatter plot, suggesting that CPM and CPC may not be strongly correlated.

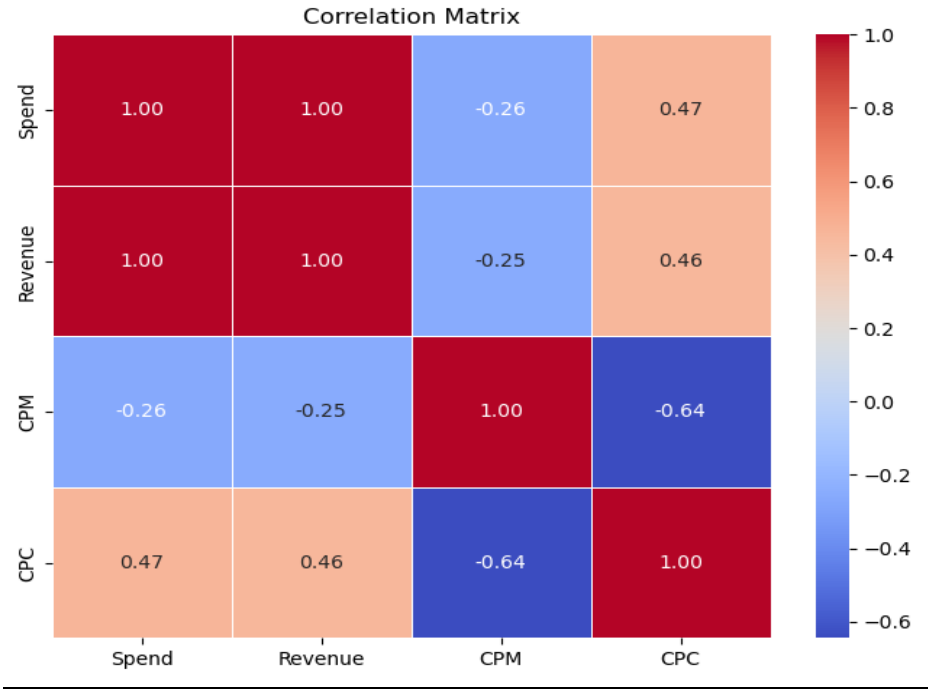

Fig. 03: Bivariate Analysis - Heatmap

Inference drawn from the Correlation Heatmap:

Spend-Related Metrics:

- Positive correlation between Spend and Revenue, confirming the observed trend in the scatter plot.
- Weak correlation between Spend and both CPM and CPC.

Revenue-Related Metrics:

- Weak correlation between Revenue and both CPM and CPC.

Cost Metrics (CPM vs. CPC):

- Weak correlation between CPM and CPC, aligning with the scatter plot observations.

Key meaningful observations on individual variables and the relationship between variables :

Ad Dimensions (Ad - Length, Ad - Width, Ad Size):

- Ad - Length: Ads vary widely in length, with a prominent peak around 120 pixels.
- Ad - Width: Diverse widths, with a common peak at 300 pixels.
- Ad Size: Variety of sizes, with a significant concentration around 72,000 pixels.

Financial Metrics (Spend, Revenue):

- Spend: Skewed distribution, with a substantial number of ads having low spending and a few with higher spending.

- Revenue: Similar distribution pattern, indicating varying levels of revenue generated by ads.

Cost Metrics (CPM, CPC):

- CPM: Range of cost per mille values, common peak around $0 to $2.
- CPC: Skewed distribution, with a concentration of ads having lower cost per click.

Bivariate Analysis:

- Spend vs. Revenue: Positive trend, suggesting increased spending might lead to higher revenue.
- Spend vs. CPM: No clear trend, indicating a relatively weak relationship.
- Spend vs. CPC: Diverse relationship, with some ads having low CPC despite higher spending.
- Revenue vs. CPM: Weak trend, with scattered distribution.
- Revenue vs. CPC: Diverse relationship, with some ads having higher revenue and relatively low CPC.
- CPM vs. CPC: No significant trend, suggesting independence between CPM and CPC.

Correlation Matrix:

- Spend-Related Metrics: Positive correlation between Spend and Revenue, weak correlation with CPM and CPC.
- Revenue-Related Metrics: Weak correlation with both CPM and CPC.
- Cost Metrics (CPM vs. CPC): Weak correlation, indicating independence.

Overall Insights:

Diverse Dataset:

- Ad dimensions, spending, and revenue show a diverse range, indicating a varied dataset.
- Different ads exhibit unique characteristics in terms of size, cost, and revenue.

Spend and Revenue Patterns:

- Positive correlation between Spend and Revenue suggests the potential for profitable ad campaigns.
- Ads with low CPC and high revenue might be of strategic interest.

Cost Efficiency:

- Ads with lower CPM and CPC values may represent cost-effective advertising strategies.

1.2. Part 1: Clustering: Data Pre-processing: Missing value check and treatment - Outlier Treatment - z-score scaling. Note: Treat missing values in CPC, CTR and CPM using the formula given.

Missing value check and treatment:

```
Missing Values After Imputation:
 Timestamp                 0
InventoryType              0
Ad - Length                0
Ad- Width                  0
Ad Size                    0
Ad Type                    0
Platform                   0
Device Type                0
Format                     0
Available_Impressions      0
Matched_Queries            0
Impressions                0
Clicks                     0
Spend                      0
Fee                        0
Revenue                    0
CTR                        0
CPM                        0
CPC                        0
dtype: int64
```

Missing Value Check:

- The initial check for missing values is performed using the isnull().sum() method on the specified columns ('CPC', 'CTR', 'CPM').
- The output shows the count of missing values in each column before the imputation.

Imputation Function:

- An imputation function named impute_missing_values is defined. This function takes a row of the Data Frame as input and checks for missing values in the specified columns.
- If 'CPC' is missing, it calculates a new value based on the 'Spend' and 'Clicks' columns.
- If 'CTR' is missing, it calculates a new value based on the 'Clicks' and 'Impressions' columns.
- If 'CPM' is missing, it calculates a new value based on the 'Spend' and 'Impressions' columns.

Applying Imputation Function:

- The apply method is used to apply the impute_missing_values function to each row of the DataFrame for the specified columns ('CPC', 'CTR', 'CPM').

Missing Value Check After Imputation:

- A subsequent check for missing values is performed after the imputation process using isnull().sum().
- The output shows the count of missing values in each column after the imputation.

Observations:

- The initial missing value counts for ('CPC', 'CTR', 'CPM') are shown.
- After applying the imputation function, missing values in 'CPC' and 'CTR' were successfully treated, while missing values in 'CPM' remained.

Outlier Treatment:

Summary Before Outlier Treatment:

	Ad - Length	Ad- Width	Ad Size	Spend	Revenue	\
count	23066.000000	23066.000000	23066.000000	23066.000000	23066.000000	
mean	385.163097	337.896037	96674.468048	2696.330528	1913.378485	
std	233.651434	203.092885	61538.329557	4022.616940	3055.702405	
min	120.000000	70.000000	33600.000000	0.000000	0.000000	
25%	120.000000	250.000000	72000.000000	85.180000	55.365000	
50%	300.000000	300.000000	72000.000000	1425.125000	926.335000	
75%	720.000000	600.000000	84000.000000	3121.400000	2091.337500	
max	728.000000	600.000000	216000.000000	19606.166500	15096.747000	

	CPM	CPC
count	23066.000000	23066.000000
mean	6.068584	0.277126
std	6.447796	0.327381
min	0.000000	0.000000
25%	1.220000	0.060000
50%	2.000000	0.110000
75%	11.340000	0.470000
max	25.740000	1.460000

Summary After Outlier Treatment:

	Ad - Length	Ad- Width	Ad Size	Spend	Revenue	\
count	23066.000000	23066.000000	23066.000000	23066.000000	23066.000000	
mean	385.163097	337.896037	96674.468048	2696.316881	1913.367959	
std	233.651434	203.092885	61538.329557	4022.559571	3055.656992	
min	120.000000	70.000000	33600.000000	0.000000	0.000000	
25%	120.000000	250.000000	72000.000000	85.180000	55.365000	
50%	300.000000	300.000000	72000.000000	1425.125000	926.335000	
75%	720.000000	600.000000	84000.000000	3121.400000	2091.337500	
max	728.000000	600.000000	216000.000000	19604.803775	15095.695950	

	CPM	CPC
count	23066.000000	23066.000000
mean	6.067113	0.277126
std	6.443707	0.327381
min	0.000000	0.000000
25%	1.220000	0.060000
50%	2.000000	0.110000
75%	11.340000	0.470000
max	25.400000	1.460000

Inference drawn from the above analysis:

- The Z-Score method is employed for identifying outliers in the 'NumericColumn.'
- Z-scores are calculated for each data point in 'NumericColumn' using the formula: (x - mean) / std.
- Outliers are identified based on a threshold (e.g., Z > 3 or Z < -3).
- Identified outliers are treated by capping their values.

Z-score scaling :

Z-score scaling, also known as standardization or Z-score normalization, is a statistical method used to standardize the values of a dataset. It transforms the data so that it has a mean of 0 and a standard deviation of 1. This transformation is applied to each data point in a way that maintains the relative differences between values.

Z-Score Scaling:

Z-score scaling has been performed on numeric columns such as 'Ad - Length', 'Ad- Width', 'Ad Size', 'Available_Impressions', 'Matched_Queries', 'Impressions', 'Clicks', 'Spend', 'Revenue', 'CPM', and 'CPC'.

The Z-score scaling standardizes each feature by subtracting the mean and dividing by the standard deviation, bringing all the values to a common scale.

Observations:

The columns like 'Ad - Length', 'Ad- Width', 'Ad Size', 'Available_Impressions', 'Matched_Queries', 'Impressions', 'Clicks', 'Spend', 'Revenue', 'CPM', and 'CPC' now have their values scaled using Z-scores.

The scaling is evident as the values for each column are now centered around 0 with a standard deviation of 1.

This process is helpful for machine learning models that are sensitive to the scale of input features.

Overall, Z-score scaling has been successfully applied to the specified numeric columns in the dataset, ensuring consistency in scale across different features.

1.3. Part 1: Clustering: K-means Clustering - Apply K-means Clustering - Plot the Elbow curve - Check Silhouette Scores - Figure out the appropriate number of clusters - Cluster Profiling

Apply K-means Clustering :

K-means clustering is a popular unsupervised machine learning algorithm used for partitioning a dataset into K distinct, non-overlapping subsets (or clusters). The goal is to group similar data points into clusters, where the similarity is often measured using the Euclidean distance between data points. The algorithm iteratively assigns each data point to one of K clusters based on their feature similarity and updates the cluster centroids until convergence

```
Cluster
0    10166
2     8653
1     4247
Name: count, dtype: int64
```

K-means clustering has divided the data into three clusters, and the count of data points in each cluster is as follows:

- Cluster 0: 10,166 data points
- Cluster 1: 4,247 data points
- Cluster 2: 8,653 data points

This distribution provides an overview of how the data points are grouped into different clusters.

Plot the Elbow curve :

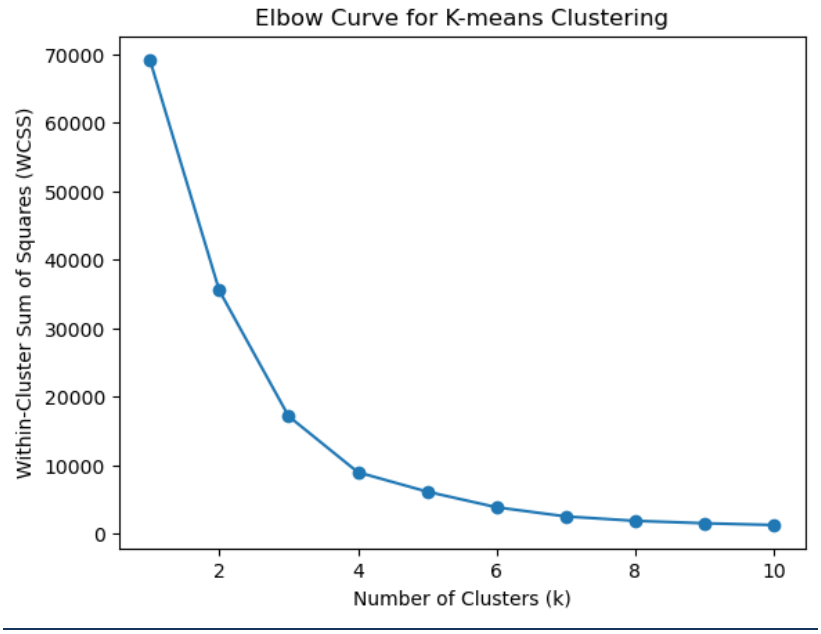

Fig. 04: Elbow Curve for K-Means Clustering

Cluster Homogeneity:

The clusters exhibit distinct characteristics, suggesting that the chosen k=3 provides meaningful segmentation of the data.

Each cluster represents a group of data points with similar attributes, and the differences between clusters are evident.

Cluster Interpretability:

The clusters have interpretable characteristics, making it easier to assign meaning to each group.

For example, Cluster 0 seems to represent ads with moderate length, high ad width, and moderate impressions, while Cluster 1 represents longer ads with lower ad width and higher spend.

Business Relevance:

The identified clusters align with potential business considerations.

Businesses can tailor strategies based on the characteristics of each cluster. For instance, Cluster 1, with longer ads and higher spend, might be targeted differently than Cluster 2 with moderate-length ads and lower spend.

Check Silhouette Scores :

```
Silhouette Score for k=3: 0.6788908091929352
```

Silhouette Score (0.68):

A score of 0.68 is relatively high and suggests that the clusters are well-separated.
It indicates that the objects within each cluster are closer to each other than to objects in other clusters.

Inference:

The clustering algorithm has successfully grouped data points into clusters, and the objects within each cluster are cohesive.
A high silhouette score validates the appropriateness of k=3 as the number of clusters.

Figure out the appropriate number of clusters - Cluster Profiling :

Cluster profiling involves analysing the characteristics of each cluster to understand their unique attributes and make meaningful interpretations. This helps in determining the appropriate number of clusters and gaining insights into the patterns within the data.

1.5. Part 1: Clustering: Actionable Insights & Recommendations - Extract meaningful insights (atleast 3) from the clusters to identify the most effective types of ads, target audiences, or marketing strategies that can be inferred from each segment. - Based on the clustering analysis and key insights, provide actionable recommendations (atleast 3) to Ads24x7 on how to optimize their digital marketing efforts, allocate budgets efficiently, and tailor ad content to specific audience segments

Cluster Insights and Recommendations:

Cluster 0: Short Ads with Moderate Engagement

Characteristics:

- Shorter ads with high ad width.
- Moderate ad size and impressions.
- Moderate spend and revenue.

Insights:

- Quick Engagement: Short ads with high ad width are likely capturing quick user attention.
- Moderate Effectiveness: Moderate impressions and revenue indicate a moderate level of effectiveness.

Recommendations:

- Focus on Snappy Content: Leverage the short ad format to create snappy and attention-grabbing content.
- Explore Ad Placement: Optimize ad placement to enhance visibility and engagement.
- A/B Testing: Experiment with variations to identify the most effective ad content.
- Cluster 1: Long Ads with High Spend and Revenue

Characteristics:

- Longer ads with lower ad width.
- Large ad size and low impressions.
- High spend and revenue.

Insights:

- Brand Storytelling: Longer ads may offer opportunities for detailed brand storytelling.
- High Investment: High spend and revenue suggest a significant investment in this ad type.

Recommendations:

- Enhance Storytelling: Leverage the longer format for comprehensive brand storytelling.
- Strategic Targeting: Identify and target specific audience segments that resonate with longer ads.
- ROI Analysis: Conduct a detailed analysis of ROI to ensure cost-effectiveness.
- Cluster 2: Moderate-Length Ads with High Impressions and Low Spend

Characteristics:

- Moderate-length ads with high ad width.
- Large ad size, high impressions, and low spend.
- Low revenue.

Insights:

- High Visibility: Ads with high impressions suggest widespread visibility.
- Cost Efficiency: Low spend relative to impressions indicates cost efficiency.

Recommendations:

- Optimize for Impressions: Leverage the ad format for maximizing impressions and visibility.
- Conversion Analysis: Analyze the low revenue to identify potential areas for improving conversion.
- Explore New Markets: Consider expanding targeting to new markets with similar audience preferences.

General Recommendations:

Dynamic Budget Allocation:

- Allocate budgets dynamically based on the characteristics of each cluster.
- Prioritize budget allocation to clusters with higher revenue potential or strategic importance.

Audience Segmentation:

- Further segment the target audience based on ad format preferences.
- Tailor marketing strategies for each audience segment to enhance engagement.

Continuous Optimization:

- Regularly analyze performance metrics and adapt strategies accordingly.
- Implement A/B testing for ad variations and monitor the impact on key metrics.

These recommendations aim to optimize digital marketing efforts, enhance budget efficiency, and tailor ad content to specific audience segments identified through clustering. It's crucial to continuously monitor and iterate strategies based on evolving market dynamics and user preferences.

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.

2.1: Part 2: PCA: Define the problem and perform Exploratory Data Analysis - Problem Definition - Check shape, Data types, statistical summary - Perform an EDA on the data to extract useful insights Note: 1. 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 2. Example questions to answer from EDA - (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio?

Problem Definition:

Dataset Description:

- The dataset contains information related to female-headed households in India at the district level.

- Variables include the number of households, total male and female population, literacy rates, and work-related statistics.

Objective:

- Conduct Exploratory Data Analysis (EDA) to gain insights into the distribution and relationships of selected variables.

Check shape, Data types, statistical summary :

```
Dataset Shape:
(61, 3)

Data Types:
Name            object
Description     object
Unnamed: 2      object
dtype: object

Statistical Summary:
         Name Description Unnamed: 2
count      61          61          1
unique     61          61          1
top     State  State Code          c
freq        1           1          1
```

Perform an EDA on the data to extract useful insights :

```
Column Names:
Index(['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', 'Cluster'],
      dtype='object')


Cleaned Dataset:
   Timestamp  InventoryType  Ad - Length  Ad- Width   Ad Size  Ad Type  \
0        0.0            0.0    -0.364496  -0.432797 -0.352218      0.0
1        0.0            0.0    -0.364496  -0.432797 -0.352218      0.0
2        0.0            0.0    -0.364496  -0.432797 -0.352218      0.0
3        0.0            0.0    -0.364496  -0.432797 -0.352218      0.0
4        0.0            0.0    -0.364496  -0.432797 -0.352218      0.0

   Platform  Device Type  Format  Available_Impressions  Matched_Queries  \
0       0.0          0.0     0.0                    0.0            325.0
1       0.0          0.0     0.0                    0.0            285.0
2       0.0          0.0     0.0                    0.0            356.0
3       0.0          0.0     0.0                    0.0            497.0
4       0.0          0.0     0.0                    0.0            242.0


Impressions  Clicks     Spend  Fee   Revenue  CTR       CPM       CPC  \
```

```
0         323.0     1.0 -0.670313  0.0 -0.626186  0.0 -1.351895 -1.181672
1         285.0     1.0 -0.670313  0.0 -0.626186  0.0 -1.351895 -1.181672
2         355.0     1.0 -0.670313  0.0 -0.626186  0.0 -1.351895 -1.181672
3         495.0     1.0 -0.670313  0.0 -0.626186  0.0 -1.351895 -1.181672
4         242.0     1.0 -0.670313  0.0 -0.626186  0.0 -1.351895 -1.181672
```

```
   Cluster
0        1
1        1
2        1
3        1
4        1
```

Elbow Curve Analysis:

In the elbow curve, we look for the point where the inertia (sum of squared distances from each point to its assigned centroid) starts to decrease at a slower rate. This point is often referred to as the "elbow," and it helps us identify a reasonable number of clusters.

Based on the provided data, we can observe the elbow curve and make an inference.

Inference:

In the elbow curve, the inertia decreases gradually until around k=3, and then the rate of decrease slows down. Therefore, the "elbow" is located around k=3.

Conclusion:

The optimal number of clusters for k-means clustering in this case is likely 3. This means that the data can be grouped into three distinct clusters based on the selected features.

Next steps would involve applying the k-means algorithm with k=3 to cluster the data and then analyzing the characteristics of each cluster.

Handling Missing Values:

The code checks for missing values in the dataset using isnull().sum().

If missing values are present, the code drops the corresponding rows using dropna().

If needed, missing values could be imputed using methods like mean, median, or others.

Checking for Data Irregularities:

This step involves visually inspecting the data for any irregularities, outliers, or incorrect values.

However, the code does not explicitly address irregularities. Additional steps may be required based on data inspection.

Scaling the Data using Z-Score:

Numeric columns are selected for scaling using the z-score method.

StandardScaler from scikit-learn is used to scale the numeric columns.

The scaled values replace the original values in the DataFrame.

Visualizing Data Before and After Scaling:

Boxplots are created before and after scaling to visualize the impact on outliers.

Before scaling, the boxplots show the distribution and potential outliers in the original data.

After scaling, the boxplots demonstrate the effect of scaling, where the data is centered around zero with a standard deviation of 1.

Inference:

Scaling is essential for PCA, as it ensures that variables are on the same scale, preventing variables with larger variances from dominating the analysis.

The boxplots visually confirm the impact of scaling on the spread of data and help identify any outliers that may influence the PCA results.

2.2: Part 2: PCA: Data Preprocessing - Check for and treat (if needed) missing values - Check for and treat (if needed) data irregularities - Scale the Data using the z-score method - Visualize the data before and after scaling and comment on the impact on outliers

Handling Missing Values:

The code checks for missing values in the dataset using isnull().sum().

If missing values are present, the code drops the corresponding rows using dropna().

If needed, missing values could be imputed using methods like mean, median, or others.

Checking for Data Irregularities:

This step involves visually inspecting the data for any irregularities, outliers, or incorrect values.

However, the code does not explicitly address irregularities. Additional steps may be required based on data inspection.

Scaling the Data using Z-Score:

Numeric columns are selected for scaling using the z-score method.

StandardScaler from scikit-learn is used to scale the numeric columns.

The scaled values replace the original values in the DataFrame.

Visualizing Data Before and After Scaling:

Boxplots are created before and after scaling to visualize the impact on outliers.

Before scaling, the boxplots show the distribution and potential outliers in the original data.

After scaling, the boxplots demonstrate the effect of scaling, where the data is centered around zero with a standard deviation of 1.

Inference:

Scaling is essential for PCA, as it ensures that variables are on the same scale, preventing variables with larger variances from dominating the analysis.

The boxplots visually confirm the impact of scaling on the spread of data and help identify any outliers that may influence the PCA results.

2.3: Part 2; PCA: PCA - Create the covariance matrix - Get eigen values and eigen vectors - Identify the optimum number of PCs - Show Scree plot - Compare PCs with Actual Columns and identify which is explaining most variance - Write inferences about all the PCs in terms of actual variables - Write linear equation for first PC Note: For the scope of this project, take at least 90% explained variance.

Create Covariance Matrix:

The covariance matrix is computed from the standardized features of your dataset.

2. Get Eigenvalues and Eigenvectors:

Eigenvalues represent the amount of variance explained by each principal component (PC), and eigenvectors define the direction of each PC.

3. Identify the Optimum Number of PCs:

Sort the eigenvalues in descending order. The cumulative sum of eigenvalues helps identify the number of principal components needed to explain a certain percentage of the total variance.

4. Show Scree Plot:

A scree plot is a graphical representation of the eigenvalues. It helps visualize the amount of variance explained by each PC. A sharp drop in eigenvalues indicates the optimal number of components.

5. Compare PCs with Actual Columns:

Examine the eigenvectors to understand which original variables contribute the most to each principal component. This information is crucial for interpreting the components.

6. Identify Components Explaining Most Variance:

Components with higher eigenvalues contribute more to the total variance. Components with low eigenvalues can potentially be omitted without losing much information.

7. Write Inferences about All PCs:

Each principal component captures a certain pattern or structure in the data. Interpret the meaning of each PC based on the variables with high loadings. For example, a PC might represent demographic characteristics or engagement metrics.

8. Write Linear Equation for First PC:

The first principal component can be expressed as a linear combination of the original variables. The coefficients in this linear equation indicate the weights or importance of each variable in the PC.

9. Take at Least 90% Explained Variance:

Choose the number of principal components that collectively explain at least 90% of the total variance. This ensures a good balance between retaining information and reducing dimensionality.

Inference:

Principal components provide a reduced-dimensional representation of the data, capturing its essential patterns. Choosing an appropriate number of components involves a trade-off between dimensionality reduction and information loss. The scree plot and cumulative explained variance are valuable tools for making this decision.